European Community on Computational Methods in Applied Sciences

# ECCOMAS Congress 2016

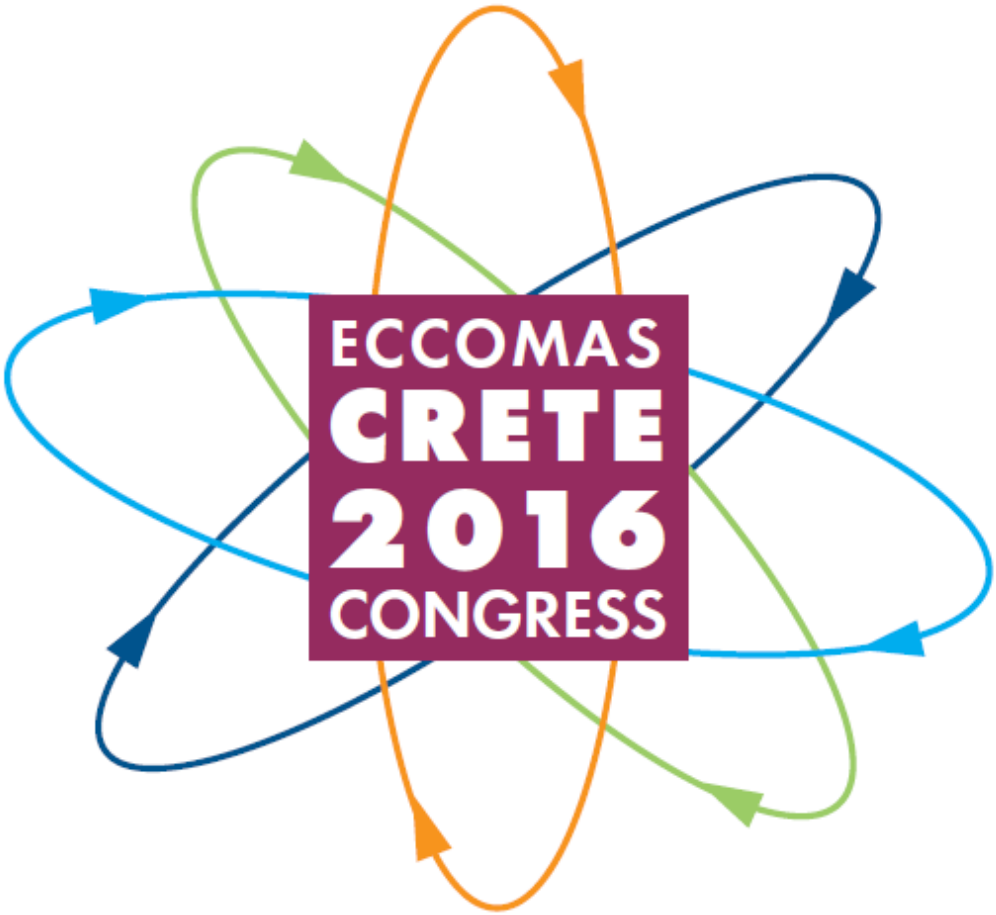
*VII European Congress*

*on Computational Methods in Applied Sciences and Engineering*

# PROCEEDINGS

## Volume IV

M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)



## ECCOMAS Congress 2016

Proceedings of the VII European Congress on Computational Methods
in Applied Sciences and Engineering
Held in Crete, Greece
5-10 June 2016

Edited by:

**M. Papadrakakis**
National Technical University of Athens, Greece

**V. Papadopoulos**
National Technical University of Athens, Greece

**G. Stefanou**
Aristotle University of Thessaloniki, Greece

**V. Plevris**
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

**A publication of:**

Institute of Structural Analysis and Antiseismic Research
School of Civil Engineering
National Technical University of Athens (NTUA)
Greece

**ECCOMAS Congress 2016**
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)

First Edition, September 2016

© The authors

ISBN: **978-618-82844-0-1**

# PREFACE

This volume contains the full-length papers presented at the **VII European Congress on Computational Methods in Applied Sciences and Engineering** (ECCOMAS Congress 2016) that was held on June 5-10, 2016 on the Crete Island, Greece.

The main objective of the quadrennial ECCOMAS Congresses is to provide a forum for presentation and discussion of state-of-the-art advances in computational methods in applied sciences and engineering, including basic methodologies, scientific developments and industrial applications and to serve as a platform for establishing links between research groups of academia and industry with common as well as complementary activities. About 2,200 papers were presented at the ECCOMAS Congress by authors from 53 countries around the world. This volume, consists of 667 full length accepted papers which will be indexed by SCOPUS database with access to the pdf file of the paper.

The ECCOMAS Congress 2016 is organized by the Institute of Structural Analysis and Antiseismic Research of the National Technical University of Athens under the support of the Greek Association for Computational Mechanics (GRACM), the Institute of Research and Development for Computational Methods in Engineering Sciences (ICMES) and the Computer Applications and Education in Engineering Sciences (CAEES).

The editor of this volume would like to thank all authors for their contributions. Special thanks go to the colleagues who contributed to the organization of the Minisymposia and to the reviewers who, with their work, contributed to the scientific quality of this e-book.

**M. Papadrakakis**
National Technical University of Athens, Greece

**V. Papadopoulos**
National Technical University of Athens, Greece

**G. Stefanou**
Aristotle University of Thessaloniki, Greece

**V. Plevris**
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

## ACKNOWLEDGEMENTS

The conference organizers acknowledge the support towards the organization of the 7th European Congress on Computational Methods in Applied Sciences and Engineering, to the following organizations:

- Greek Association for Computational Mechanics (GRACM)
- Institute of Structural Analysis and Antiseismic Research, National Technical University of Athens
- John Argyris Foundation
- Institute of Research and Development for Computational Methods in Engineering Sciences (ICMES)
- Computer Applications and Education in Engineering Sciences (CAEES)
- Region of Crete
- Municipality of Hersonissos

## Plenary Speakers

We would also like to thank the Plenary and Semi-Plenary Speakers for their help in the setting up of a high standard Scientific Programme.

**Plenary Speakers:** Alexandre Ern, Charbel Farhat, Gerhard A. Holzapfel, Antonio Huerta, Thomas J.R. Hughes, Alfio Quarteroni, Ole Sigmund, Theodosios P. Tasios

**Semi-Plenary Speakers:** Assyr Abdulle, Olivier Allix, Lourenco Beirao da Veiga, Andreas G. Boudouvis, Annalisa Buffa, Ramon Codina, Laura De Lorenzis, Bernard Geurts, Kyriakos Giannakoglou, Antonio J. Gil, Dan Givoli, Anthony Gravouil, George Karniadakis, Mats G. Larson, Kim Meow Liew, Wing Kam Liu, Federico Negri, Shinji Nishiwaki, Ursula Rasthofer, Ernst Rank, Cord-Christian Rossow, Giancarlo Sangalli, Jörg Schröder, Bert Sluys, Christian Soize, Roberto Verzicco, Barbara Wohlmuth

## SUMMARY

**Preface**..... iii

**Acknowledgements**..... iv

**Contents**..... xvii

# VOLUME I

**SEMI-PLENARY LECTURES** ..... 1

**MINISYMPOSIA**

**MS 102 - COMPUTATIONAL MODELS IN BIOMECHANICS AND MECHANOBIOLOGY** ..... 27
*Organized by Estefania Peña, Renato Natal Jorge, Miguel A. Martínez, Pedro S. Martins*

**MS 103 - MECHANICS OF BIOLOGICAL TISSUES** ..... 55
*Organized by Markus Böl, Gerhard A. Holzapfel*

**MS 105 - SIMULATION OF CARDIOVASCULAR PROCEDURES AND DEVICES** ..... 81
*Organized by Ferdinando Auricchio, Michele Conti, Simone Morganti, Alessandro Reali, Alessandro Veneziani*

**MS 106 - DIRECT AND INVERSE METHODS FOR CARDIOVASCULAR AND PULMONARY BIOMECHANICS** ..... 98
*Organized by Wolfgang A. Wall, C. Alberto Figueroa, Marek Behr*

**MS 108 - NUMERICAL METHODS FOR COUPLED PROBLEMS IN BIOMEDICAL APPLICATIONS** ..... 105
*Organized by Martina Bukac, Annalisa Quaini*

**MS 111 - POPULATION BALANCE MODELING: CURRENT STATUS, FUTURE PROSPECTS AND NOVEL APPLICATIONS FROM NANOPARTICLES' SYNTHESIS TO (LUNG) CANCER** ..... 120
*Organized by Georgios Lolas, Georgios Bourantas, Panagiotis Gavriliadis, Konstantinos Syrigos*

**MS 112 - ANEURYSMS: SOLID MECHANICS, FLUID MECHANICS, AND MECHANOBIOLOGY** ..... 137
*Organized by Christian J. Cyron, Sven Hirsch, Philippe Bijlenga, Roland C. Aydin, Anne M. Robertson, Gerhard A. Holzapfel*

**MS 113 - MATHEMATICAL AND NUMERICAL MODELING OF THE HEART** ..... 155
*Organized by Luca Dede', Luca Pavarino, Alfio Quarteroni*

**MS 115 - TUMOR GROWTH MODELING AND THE MECHANICAL ASPECTS OF CANCER** ..... 178
*Organized by Hector Gomez, Assad Oberai, Krishna Garikipati, Kristen Mills, Thomas J.R. Hughes*

**MS 201 - MICROSTRUCTURE-DRIVEN DEFORMATION AND FAILURE IN CRYSTALLINE MATERIALS** ..... 188
*Organized by Pilar Ariza, Lucia Nicola, Angelo Simone*

**MS 202 - CIVIL ENGINEERING MATERIALS AND STRUCTURES UNDER EXTREME LOADINGS** ..... 216
*Organized by Fabrice Gatuingt, Frédéric Dufour, Panagiotis Kotronis*

**MS 203 - COMPUTATIONAL METHODS FOR MODELLING INSTABILITIES IN SOLIDS & STRUCTURES** ..... 254
*Organized by Spyros A. Karamanos*

**MS 204 - IMPACT AND CRASH MECHANICS** ..... 273
*Organized by Manfred Bischoff, Fabian Duddeck*

**MS 205 - PROBABILISTIC APPROACH TO NUMERICAL SIMULATION OF FRACTURE** ..... 323
*Organized by Alexander V. Gerasimov*

**MS 301 - METHODS FOR CUT AND COMPOSITE MESHES: THEORY, ALGORITHMS AND APPLICATIONS** ..... 373
*Organized by Mats G. Larson, André Massing*

**MS 302 - MESH GENERATION AND ADAPTION** ..... 384
*Organized by Josep Sarrate, Xevi Roca, Rafael Montenegro, Eloi Ruiz*

**MS 303 - CURVED MESH GENERATION FOR HIGH-ORDER METHODS** ..... 417
*Organized by Xevi Roca, Josep Sarrate*

**MS 304 - COMPUTATIONAL MODELLING OF ADDITIVE PRODUCTION PROCESSES** ..... 469
*Organized by Dirk Hartmann, Stefan Kollmannsberger, Ernst Rank, Utz Wever*

**MS 305 - ADVANCED MESHING METHODS FOR INDUSTRIAL APPLICATIONS** ..... 480
*Organized by Frederic Alauzet, Thierry Coupez, Alain Dervieux, Adrien Loseille*

**MS 306 - LATTICE SPRING METHODS FOR LINEAR AND NONLINEAR CONTINUA** ..... 511
*Organized by Ioannis Doltsinis*

**MS 403 - PARTICLE-BASED METHODS IN FLUID MECHANICS** ..... 539
*Organized by Sergio Idelsohn, Eugenio Oñate*

**MS 404 - SIMULATION OF ENVIRONMENTAL FLOWS** ..... 650
*Organized by Pablo Ortiz, Piotr K. Smolarkiewicz, Joanna Szmelter*

**MS 405 - COMPUTATIONAL MODELING OF MULTIPHASE FLOWS: ADVANCED METHODS, INTERFACE PHENOMENA AND ENVIRONMENTAL APPLICATIONS** ..... 750
*Organized by Adeline Montlaur, Santiago Arias Calderón, Martin Kronbichler*

**MS 406 - ADVANCES IN COMPUTATIONAL METHODS FOR GAS-LIQUID TWO-PHASE FLOW** ..... 788
*Organized by Byeong Rog Shin, Takeo Kajishima*

**MS 408 - MANIPULATION AND CONTROL OF TURBULENT FLOW** ..... 831
*Organized by Markus Rütten, Christina Voß*

**MS 409 - CURRENT TRENDS IN MODELLING AND SIMULATION OF TURBULENT FLOWS** ..... 865
*Organized by Suad Jakirlić*

**MS 410 - COMPLEX FLUID FLOWS IN ENGINEERING: MODELLING, SIMULATION AND OPTIMIZATION** ..... 949
*Organized by Stefanie Elgeti, Philipp Knechtges*

**MS 411 - NON-NEWTONIAN HEAT AND FLUID FLOW SUBJECTED TO MAGNETIC FORCES** ..... 1061
*Organized by Laszlo Konozsy, Dimitris Drikakis*

**MS 412 - NUMERICAL METHODS FOR WAVES AND FLOWS IN COASTAL AND DEEP WATER HYDRODYNAMICS** ..... 1093
*Organized by Nina Shokina, Yuri Shokin, Leonid Chubarov, Gayaz Khakimzyanov, Vadym Aizinger, Denys Dutykh*

**MS 414 - NEW TRENDS IN NUMERICAL METHODS FOR MULTI-MATERIAL COMPRESSIBLE FLUID FLOWS** ..... 1139
*Organized by Andy Barlow, Michael Dumbser, Raphaël Loubère, Pierre-Henri Maire, Rob Rieben, Mikhail Shashkov, François Vilar*

**MS 415 - COMPUTATIONAL NON-NEWTONIAN FLUID MECHANICS** ..... 1205
*Organized by Georgios Georgiou, John Tsamopoulos*

**MS 501 - ALGORITHMIC ASPECTS OF HIGH-PERFORMANCE COMPUTING FOR MECHANICS AND PHYSICS** ..... 1237
*Organized by Santiago Badia, Victor Calo, Javier Principe*

**MS 503 - HPC-BASED SIMULATIONS FOR THE ENGINEERING REALM AND INDUSTRIAL APPLICATIONS** ..... 1277
*Organized by Makoto Tsubokura, Mariano Vázquez, Takayuki Aoki*

**MS 504 - NUMERICAL METHODS AND TOOLS FOR KEY EXASCALE COMPUTING CHALLENGES IN ENGINEERING AND APPLIED SCIENCES** ..... 1352
*Organized by Eugenio Oñate, Manolis Papadrakakis, Peter Wriggers*

**MS 505 - INTERACTIVE SIMULATIONS IN COMPUTATIONAL ENGINEERING** ..... 1392
*Organized by Adrian Harwood, Petra Wenisch*

**MS 506 - ACCURACY AND EFFICIENCY OF APPROXIMATE COMPUTATIONS IN SCIENCE AND ENGINEERING** ..... 1409
*Organized by Aram Soroushian*

**MS 601 - SHOCK WAVE-BOUNDARY LAYER INTERACTION AND ITS CONTROL** ..... 1455
*Organized by Piotr Doerffer, George Barakos*

**MS 602 - INNOVATIVE METHODS FOR FLUID-STRUCTURE-INTERACTION** ..... 1498
*Organized by E. Harald van Brummelen, Roger Ohayon, Trond Kvamsdal*

**MS 603 - COMPUTATIONAL METHODS IN FLUID-STRUCTURE INTERACTION WITH IMPACT ON INDUSTRIAL APPLICATIONS** ..... 1582
*Organized by Elisabeth Longatte, Yannick Hoarau, Marianna Braza*

**MS 605 - FRICTIONAL CONTACTS WITH LUBRICATION – BASICS AND APPLICATIONS** ..... 1623
*Organized by Michael Müller, Thomas Hagemann*

**MS 606 - COMPUTATIONAL MODELING OF HYDRAULIC FRACTURING** ..... 1637
*Organized by Gianluca Cusatis, Gilles Pijaudier-Cabot, Günther Meschke*

**MS 607 - ADVANCES IN COMPUTATIONAL METHODS FOR LIQUID-VAPOR FLOWS WITH PHASE TRANSFER PROCESSES** ..... 1672
*Organized by Rémi Abgrall, Pietro M. Congedo, Tore Flåtten, Bernhard Müller, Marica Pelanti, Maria Giovanna Rodio*

**MS 608 - ADVANCES IN TIME INTEGRATION FOR SOLID, FLUID AND COUPLED SYSTEMS** ..... 1709
*Organized by Ilinca Stanciulescu, Peter Betsch*

**MS 609 - ADVANCED COMPUTATIONAL MODELING OF BATTERIES AND FUEL CELLS** ..... 1838
*Organized by Edwin Knobbe, Wolfgang A. Wall*

**MS 610 - NUMERICAL METHODS TO STUDY THE CONTACT MECHANICS OF DRY, ADHESIVE AND LUBRICATED ROUGH SURFACES** ..... 1850
*Organized by Carmine Putignano, Daniele Dini*

**MS 612 - NUMERICAL SIMULATIONS FOR SMART-CITY APPLICATIONS** ..... 1876
*Organized by Julien Waeytens, Rachida Chakir*

**MS 615 - COMPUTATIONAL MODELS IN MAGNETOHYDRODYNAMICS** ..... 1886
*Organized by Oleg Zikanov*

**MS 701 - ADVANCED MATERIALS: COMPUTATIONAL ANALYSIS OF PROPERTIES AND PERFORMANCE** ..... 1893
*Organized by Vadim V. Silberschmidt, Valery P. Matveenko*

**MS 702 - MODELING OF NANOFILLED COMPOSITES** ..... 1971
*Organized by Konstantinos I. Tserpes*

**MS 705 - IDENTIFICATION OF MATERIAL MODELS** ..... 2003
*Organized by Danuta Szeliga, Wacław Kuś, Tadeusz Burczyński, Jan Kusiak*

**MS 706 - MODELING OF FIBER-BASED STRUCTURES - TEXTILES AND TEXTILE REINFORCED COMPOSITES** ..... 2010
*Organized by Yordan Kyosev, Philippe Boisse, Nahiene Hamila, Damien Durville*

**MS 707 - MICROMECHANICAL MODELLING: COMPETITION BETWEEN ANALYTICAL AND NUMERICAL METHODS** ..... 2039
*Organized by Siegfried Schmauder, Vera Petrova*

**MS 711 - FOURIER-BASED METHODS FOR COMPUTING THE BEHAVIOR OF HETEROGENEOUS MATERIALS DEVELOPMENTS, EXTENSIONS AND APPLICATIONS** ..... 2099
*Organized by Lionel Gélébart, Hervé Moulinec, Franz Roters, François Willot*

**MS 712 - SMART MATERIAL SYSTEMS AND STRUCTURES** ..... 2110
*Organized by Mieczysław Kuczma, Pavel Krejči, Jörg Schröder, Georgios E. Stavroulakis, Gwidon Szefer*

**MS 713 - MICROSTRUCTURE-BASED MODELLING OF HETEROGENEOUS MATERIALS** ..... 2134
*Organized by Jan Zeman, Jan Novak, Guillermo Díaz*

**MS 714 - STRENGTH, FATIGUE AND STABILITY OF COMPOSITE STRUCTURES** ..... 2182
*Organized by Raimund Rolfes, Martin Ruess, Kai-Uwe Schröder*

## VOLUME II

**MINISYMPOSIA**

**MS 715 - COMPUTATIONAL ANALYSIS OF COMPOSITE STRUCTURES** ..... 2210
*Organized by Efstathios E. Theotokoglou*

**MS 801 - MULTISCALE COMPUTATIONAL HOMOGENIZATION FOR BRIDGING SCALES IN THE MECHANICS AND PHYSICS OF COMPLEX MATERIALS** ..... 2266
*Organized by Julien Yvonnet, Kenjiro Terada, Peter Wriggers, Marc Geers*

**MS 802 - ADVANCES IN THE MODELLING OF MULTI-SCALE, MULTI-PHYSICS AND MULTI-UNCERTAINTY PROBLEMS** ..... 2323
*Organized by Francisco M. Andrade Pires, Chengfeng Li*

**MS 805 - ADVANCED MULTI-PHYSICS AND MULTI-SCALE TECHNIQUES FOR MODELING INELASTIC PROCESSES IN SOLIDS: DAMAGE, FRACTURE AND CONTACT MECHANICS** ..... 2353
*Organized by Mauro Corrado, Marco Paggi, José Reinoso*

**MS 806 - MULTISCALE MODELLING OF MATERIALS AND STRUCTURES** ..... 2363
*Organized by Tadeusz Burczyński, Xavier Oliver, Maciej Pietrzyk, Alfredo Huespe*

**MS 807 - ADVANCED COMPUTATIONAL STRATEGIES FOR MODELLING, SIMULATION AND CHARACTERISATION OF MULTI-SCALE HETEROGENEOUS MATERIALS** ..... 2449
*Organized by Stéphane Bordas, Daniel Dias-da-Costa, Fabrice Pierron, Timon Rabczuk, Pierre Kerfriden, Pascal Lava*

**MS 808 - MULTISCALE AND MULTIPHYSICS MODELING OF CEMENTITIOUS MATERIALS** ..... 2460
*Organized by Jörg F. Unger, Thomas Titscher*

**MS 809 - MULTISCALE STOCHASTIC FINITE ELEMENT METHODS** ..... 2471
*Organized by George Stefanou, Xi Frank Xu, Yu Ching Wu*

**MS 811 - MULTISCALE MODELING OF CONCRETE AND CONCRETE STRUCTURES** ..... 2501
*Organized by Herbert Mang, Yong Yuan*

**MS 901 - ISOGEOMETRIC METHODS** ..... 2514
*Organized by Yuri Bazilevs, David J. Benson, Rene De Borst, Thomas J.R. Hughes, Trond Kvamsdal, Alessandro Reali, Giancarlo Sangalli, Clemens V. Verhoosel*

**MS 902 - INNOVATIVE NUMERICAL APPROACHES FOR MULTI-PHYSICS PROBLEMS** ..... 2558
*Organized by Anna Pandolfi, Laurent Stainier, Kerstin Weinberg*

**MS 903 - ADVANCES IN FICTITIOUS DOMAIN METHODS FOR SOLID MECHANICS** ..... 2566
*Organized by Alexander Düster, Ernst Rank, Stefan Kollmannsberger, Andreas Schröder*

**MS 905 - DISCONTINUOUS GALERKIN METHODS: NEW TRENDS AND APPLICATIONS** ..... 2614
*Organized by Sonia Fernández-Méndez, Nicoletta Franchina*

**MS 906 - MATHEMATICAL ADVANCES IN ISOGEOMETRIC ANALYSIS** ..... 2715
*Organized by Annalisa Buffa, John A. Evans, Thomas J.R. Hughes, Giancarlo Sangalli*

**MS 907 - REGULARIZED ENRICHED APPROXIMATIONS AND QUADRATURE FOR DISCONTINUITIES, SINGULARITIES AND CONTINUOUS-DISCONTINUOUS TRANSITION** ..... 2726
*Organized by Elena Benvenuti, Giulio Ventura, José M.A. César de Sá*

**MS 908 - VERIFICATION AND VALIDATION OF STRUCTURAL MECHANICS SIMULATION MODELS** ..... 2734
*Organized by George Lampeas*

**MS 910 - HIGH ORDER CFD METHODS: CONCLUSIONS AND OUTLOOK** ..... 2843
*Organized by Koen Hillewaert, John Ekaterinaris, Peter Vincent, Norbert Kroll, Norbert Huynh, Z.J. Wang*

**MS 911 - NUMERICAL METHODS IN THE MECHANICS OF GENERALIZED CONTINUA** ..... 2869
*Organized by Elena Atroshchenko, Jack S. Hale, George Bourantas, Stéphane P.A Bordas*

**MS 912 - HIGH-ORDER METHODS, SENSITIVITY ANALYSIS AND ADAPTATION FOR THE NAVIER STOKES EQUATIONS** ..... 2883
*Organized by Vincent Couaillier, Rémi Abgrall, Eusebio Valero*

**MS 913 - HIGH-ORDER METHODS FOR POLYGONAL AND POLYHEDRAL MESHES** ..... 2930
*Organized by Lourenço Beirao da Veiga, Franco Brezzi, Donatella Marini, Alessandro Russo*

**MS 916 - DIRECT METHODS FOR LIMIT STATES OF MATERIALS AND STRUCTURES** ..... 2971
*Organized by Konstantinos V. Spiliopoulos, Dieter Weichert*

**MS 917 - MESOSCOPIC METHODS FOR COMPLEX FLUIDS AND SOFT MATTER** ..... 3026
*Organized by Zhen Li, Wenxiao Pan, Igor V. Pivkin*

**MS 918 - COMPUTER ALGEBRA SYSTEMS IN MODELLING STATIC AND DYNAMIC PROBLEMS IN MECHANICS OF SOLIDS** ..... 3045
*Organized by Alexander V. Matrosov, Dmitriy P. Goloskokov*

**MS 919 - RECENT ADVANCES IN NUMERICAL SIMULATION AND ANALYSIS OF KINETIC MODELS** ..... 3113
*Organized by E. Harald van Brummelen, Manuel Torrilhon*

**MS 921 - RECENT ADVANCES IN BOUNDARY ELEMENT METHODS** ..... 3139
*Organized by Gernot Beer, Luiz Wrobel, Martin Schan*

**MS 922 - HIGH-ORDER METHODS FOR ELASTIC WAVES AND THEIR APPLICATION** ..... 3211
*Organized by Thomas Hagstrom, Daniel Appelo*

**MS 923 - NOVEL DISCRETIZATION METHODS – MATHEMATICAL AND MECHANICAL ASPECTS** ..... 3224
*Organized by Jörg Schröder, Peter Wriggers, Ferdinando Auricchio, Carsten Carstensen*

**MS 1001 - STRUCTURAL AND MULTIDISCIPLINARY OPTIMIZATION** ..... 3239
*Organized by J.F. Aguilar Madeira, Helder C. Rodrigues*

**MS 1002 - EVOLUTIONARY ALGORITHMS AND METAHEURISTICS IN CIVIL ENGINEERING AND CONSTRUCTION MANAGEMENT** ..... 3473
*Organized by Jorge Magalhaes-Mendes, David Greiner*

**MS 1003 - ADVANCES IN DESIGN OPTIMIZATION OF STRUCTURES AND MATERIALS** ..... 3534
*Organized by Zhen Luo, Zhan Kang*

**MS 1004 - AERODYNAMIC STRATEGIES FOR THE GLOBAL OPTIMIZATION OF FLYING CONFIGURATIONS IN SUPERSONIC FLOW** ..... 3596
*Organized by Adriana Nastase, Catalin Nae*

**MS 1005 - MONITORING AND CONTROL OF STRUCTURES** ..... 3630
*Organized by Resat Oyguc*

**MS 1006 - PARAMETER IDENTIFICATION IN SOLID MECHANICS** ..... 3652
*Organized by A. Gil Andrade-Campos, Marco Rossi, Sandrine Thuillier, Franck Toussaint, Marta C. Oliveira*

**MS 1007 - ADDITIVE MANUFACTURING AND OPTIMIZATION** ..... 3689
*Organized by Ekkehard Ramm, Ole Sigmund, Pierre Duysinx, Wing Kam Liu*

**MS 1008 - ULTRASONIC GUIDED WAVES TESTING AND MONITORING** ..... 3753
*Organized by Yaacoubi Slah, Nico Declercq*

**MS 1009 - ADJOINT METHODS FOR STEADY & UNSTEADY OPTIMIZATION** ..... 3770
*Organized by Kyriakos C. Giannakoglou, Jens Dominik Mueller*

**MS 1010 - INVERSE PROBLEMS, DESIGN AND OPTIMIZATION** ..... 4061
*Organized by Marcelo J. Colaço, Helcio R. B. Orlande, George S. Dulikravich*

**MS 1011 - SURROGATE-ASSISTED EVOLUTIONARY ALGORITHMS IN AERODYNAMIC DESIGN/OPTIMIZATION** ..... 4082
*Organized by Varvara Asouti, Esther Andrés, Emiliano Iuliano*

**MS 1013 - SOLUTION OF LARGE-SCALE INVERSE PROBLEMS** ..... 4164
*Organized by Clint Dawson, Steve Mattis, Troy Butler, Lindley Graham*

**MS 1014 - DESIGN OPTIMIZATION AND INVERSE PROBLEMS FOR WAVE PROPAGATION PROBLEMS** ..... 4178
*Organized by Martin Berggren*

**MS 1101 - REDUCED BASIS, POD AND PGD MODEL ORDER REDUCTION TECHNIQUES** ..... 4187
*Organized by*

**MS 1103 - MATHEMATICAL SURROGATE MODELLING IN ELECTROMAGNETICS** ..... 4284
*Organized by Francisco Chinesta, Elias Cueto, Antonio Huerta, Pierre Ladeveze, Gianluigi Rozza*

**MS 1201 - COMPUTATIONAL STRUCTURAL DYNAMICS** ..... 4351
*Organized by*
*Evangelos J. Sapountzakis, Andreas E. Kampitsis*

## VOLUME III

**MINISYMPOSIA**

**MS 1202 - ADVANCED BEAM MODELS** ..... 4445
*Organized by Dinar Camotim, Zuzana Dimitrovová, Rodrigo Gonçalves*

**MS 1203 - THE MODELS AND INVESTIGATIONS METHODS OF DYNAMICS OF THE SOLIDS SYSTEMS WITH DRY FRICTION** ..... 4553
*Organized by Alexey A. Kireenkov, Alexander V. Karapetyan*

**MS 1204 - NONLINEAR DYNAMICS OF ROTATING STRUCTURES** ..... 4611
*Organized by Evangeline Capiez-Lernout, Marc P. Mignolet, Christian Soize*

**MS 1206 - ADVANCES IN NUMERICAL METHODS FOR LINEAR AND NON-LINEAR DYNAMICS AND WAVE PROPAGATION** ..... 4674
*Organized by Alexander Idesman*

**MS 1207 - COMPUTATIONAL SIMULATION OF SMART STRUCTURES AND MATERIALS** ..... 4728
*Organized by Ruediger Schmidt, Kai-Uwe Schröder*

**MS 1208 - BIFURCATIONS AND STABILITY** ..... 4741
*Organized by Pekka Neittaanmäki, Nikolay Banichuk, Juha Jeronen, Tero Tuovinen*

**MS 1209 - DYNAMIC ANALYSIS OF BEAMS UNDER MOVING VEHICLES: APPLICATION TO RAILWAY TRACK MODELLING, DESIGN AND REHABILITATION** ..... 4780
*Organized by Fernardo Simões, Antonio Pinto da Costa*

**MS 1210 - ADVANCES IN MODELING AND ANALYSIS OF FGM STRUCTURES** ..... 4810
*Organized by Justin Murin, Stephan Kugler, Mehdi Aminbaghai*

**MS 1211 - COMPUTATIONAL STRATEGIES FOR STRUCTURAL ROBUSTNESS ASSESSMENT** ..... 4900
*Organized by Domenic Asprone, Fulvio Parisi*

**MS 1212 - DYNAMICS AND SEISMIC RESPONSE OF ROCKING AND SELF-CENTERING STRUCTURES**..... 4959
*Organized by Matthew DeJong, Elias Dimitrakopoulos, Michalis Fragiadakis*

**MS 1213 - INNOVATIVE STRUCTURAL SYSTEMS FOR SEISMIC RESISTANT BUILDINGS** ..... 5142
*Organized by Carlo Castiglioni*

**MS 1214 - HISTORIC MASONRY STRUCTURES: MODELLING, ASSESSMENT & RETROFIT** ..... 5172
*Organized by Panagiotis Asteris, Charilaos Maniatakis, Constantine Spyrakos*

**MS 1215 - NONLINEAR VIBRATIONS OF CONSERVATIVE AND NONCONSERVATIVE SYSTEMS: PHENOMENA AND ADVANCED NUMERICAL METHODS** ..... 5396
*Organized by Malte Krack, Ludovic Renson, Gaëtan Kerschen*

**MS 1216 - STRUCTURAL ANALYSIS AND VIBRATIONS** ..... 5444
*Organized by Diana V. Bambill, Carlos A. Rossit*

**MS 1217 - COMPUTATIONAL METHODS IN EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS** ..... 5464
*Organized by Vagelis Plevris, Georgia Kremmyda, Yasin Fahjan*

**MS 1218 - STABILITY AND CONTROL OF FLEXIBLE STRUCTURES** ..... 5646
*Organized by Ilaria Venanzi, Marco Lepidi*

**MS 1220 - DYNAMICS OF STRUCTURES SUBJECT TO SEISMIC EXCITATIONS** ..... 5712
*Organized by Michel Géradin, Evtim Zahariev, Evangelos J. Sapountzakis*

**MS 1221 - COMPUTATIONAL STRUCTURAL STABILITY** ..... 5768
*Organized by Herbert A. Mang, Yeon-Bin Yang*

**MS 1222 - INFLUENCE OF LIQUEFIABLE SOIL ON SINGLE AND CLOSELY CLUSTERED STRUCTURES** ..... 5817
*Organized by Nawawi Chouw, Rolly Orense, Tam Larkin*

**MS 1224 - INNOVATIVE SOLUTIONS FOR THE SEISMIC PROTECTION OF INDUSTRIAL BUILDINGS** ..... 5834
*Organized by Walter Salvatore, Carlo Castiglioni, Francesco Morelli, Nikolaos Bakas*

**MS 1225 - SEISMIC PERFORMANCE ASSESSMENT OF STRUCTURES AND SEISMIC RISK MITIGATION STRATEGIES** ..... 5934
*Organized by Marco Vona*

**MS 1301 - THE STOCHASTIC COMPUTER METHODS IN MECHANICS** ..... 6029
*Organized by Marcin Kamiński, Takahiko Kurahashi*

**MS 1303 - ANALYSIS AND DESIGN OF SAFETY CRITICAL SYSTEMS UNDER UNCERTAINTY** ..... 6091
*Organized by Edoardo Patelli, Michael Beer, Matteo Broggi, Francisco Alejandro Díaz De la O*

**MS 1304 - STOCHASTIC MODELING AND IDENTIFICATION OF UNCERTAINTIES IN COMPUTATIONAL MECHANICS** ..... 6136
*Organized by Johann Guilleminot, Maarten Arnst, Christian Soize*

**MS 1306 - ERCOFTAC SIG-45: UNCERTAINTY QUANTIFICATION IN CFD AND FLUID STRUCTURE INTERACTION** ..... 6196
*Organized by Didier Lucor, Sunetra Sarkar*

**MS 1307 - NON-INTRUSIVE SURROGATE MODELS FOR UNCERTAINTY QUANTIFICATION IN HIGH DIMENSIONS** ..... 6250
*Organized by Bruno Sudret, Eleni Chatzi, Jean-Marc Bourinet, Nicolas Gayton*

**MS 1308 - MODELLING AND INVERSE METHODS IN NONLINEAR DYNAMICAL SYSTEMS** ..... 6316
*Organized by Sotirios Natsiavas, Costas Papadimitriou, Eleni Chatzi, Dimitrios Giagopoulos*

**MS 1309 - SCALABLE MULTI-FIDELITY MODELING FOR DESIGN, UNCERTAINTY QUANTIFICATION, AND INVERSE PROBLEMS** ..... 6383
*Organized by Paris Perdikaris, George Em. Karniadakis*

**MS 1310 - COMPUTATIONAL METHODS FOR THE SOLUTION OF STOCHASTIC DIFFERENTIAL EQUATIONS** ..... 6398
*Organized by Jianbing Chen, Ioannis Kougioumtzoglou, Vissarion Papadopoulos*

**MS 1401 - TOICA: THERMAL OVERALL INTEGRATED CONCEPT AIRCRAFT** ..... 6420
*Organized by Pierre Arbez, Jean-Claude Dunyach*

**STS 3 - INNOVATIVE DESIGN OPTIMIZATION TOOLS LINKED TO INDUSTRIAL AERONAUTICAL APPLICATIONS: TARGETING GREENER PERFORMANCES** ..... 6471
*Organized by Jacques Periaux, Gabriel Bugeda*

**STS 5 - TRANSITION LOCATION EFFECT ON SHOCK WAVE BOUNDARY LAYER INTERACTION** ..... 6485
*Organized by Piotr Doerffer, Pawel Flaszynski*

**STS 7 - MORPHING TECHNOLOGIES FOR AIRCRAFT WINGS** ..... 6496
*Organized by Hans Peter Monner*

**STS 9 - ADVANCED WING HIGH-LIFT SYSTEMS** ..... 6505
*Organized by Jochen Wild*

**YOUNG INVESTIGATORS MINISYMPOSIUM** ..... 6515
*Organized by Jaan-Willem Simon, Alexander Popp, Joan Baiges*

**ECCOMAS OLYMPIAD MINISYMPOSIUM** ..... 6526

## VOLUME IV

**CONTRIBUTED SESSIONS**

**CS 110 - NUMERICAL MODELS IN BIOMECHANICS** ..... 6539

**CS 210 - NUMERICAL SIMULATION OF COMPOSITE MATERIALS** ..... 6637

**CS 211 - CRACK PROPAGATION** ..... 6688

**CS 212 - NUMERICAL MODELING OF DAMAGE, FAILURE AND FRACTURE** ..... 6699

**CS 230 - DYNAMIC FAILURE AND FRACTURE** ..... 6799

**CS 310 - CAD, CAM AND CAE** ..... 6847

**CS 320 - GRID GENERATION AND ADAPTIVE TECHNIQUES** ..... 6911

**CS 410 - COMPUTATIONAL FLUID MECHANICS** ..... 6955

**CS 420 - MULTI-PHASE AND CHEMICALLY REACTING FLOWS** ..... 7138

**CS 450 - NUMERICAL METHODS AND CONVERGENCE ACCELERATION IN CFD** ..... 7274

**CS 460 - UNSTEADY FLOW COMPUTATION** ..... 7390

**CS 500 - HIGH PERFORMANCE COMPUTING** ..... 7572

**CS 610 - AERO-ACOUSTICS** ..... 7589

**CS 620 - COMPUTATIONAL CONTACT MECHANICS** ..... 7628

**CS 630 - SIMULATION OF FLUID-STRUCTURE INTERACTION** ..... 7640

**CS 720 - COMPUTATIONAL MATERIALS SCIENCE** ..... 7711

**CS 750 - COMPUTATIONAL MODELING OF COMPOSITES** ..... 7765

**CS 751 - SMART MATERIALS AND STRUCTURES** ..... 7853

**CS 830 - COMPUTATIONAL NANOTECHNOLOGY** ..... 7875

**CS 840 - MULTI-SCALE COMPUTATIONAL METHODS FOR SOLIDS AND FLUIDS** ..... 7886

**CS 930 - HIGH-ORDER DISCRETIZATION METHODS** ..... 7956

**CS 940 - EXTENDED DISCRETIZATION METHODS** ..... 8066

**CS 960 - MESHLESS METHODS** ..... 8115

**CS 980 - NUMERICAL AND SYMBOLIC COMPUTATION** ..... 8162

**CS 990 - PARTICLE-BASED METHODS** ..... 8196

**CS 1010 - COMPUTATIONAL INVERSE PROBLEMS AND OPTIMIZATION** ..... 8247

**CS 1020 - EVOLUTIONARY AND DETERMINISTIC METHODS FOR DESIGN, OPTIMIZATION AND CONTROL** ..... 8391

**CS 1100 - REDUCTION ORDER METHODS** ..... 8539

**CS 1200 - STRUCTURAL DYNAMICS** ..... 8597

**CS 1201 - COMPUTATIONAL SOIL MECHANICS** ..... 8708

**CS 1202 - STRUCTURAL ANALYSIS AND MULTIBODY DYNAMICS** ..... 8739

**CS 1300 - UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION** ..... 8769

# CONTENTS

## Semi-plenary Lectures

NONPARAMETRIC PROBABILISTIC APPROACH OF MODEL UNCERTAINTIES INTRODUCED BY A PROJECTION-BASED NONLINEAR REDUCED-ORDER MODEL ..... 1
*Christian Soize, Charbel Farhat*

## Minisymposia

**100 BIOLOGICAL SYSTEMS**

**MS 102: COMPUTATIONAL MODELS IN BIOMECHANICS AND MECHANOBIOLOGY**

THE INFLUENCE OF MECHANICAL STIMULUS ON NUTRIENT TRANSPORT AND CELL GROWTH IN ENGINEERED CARTILAGE: A FINITE ELEMENT APPROACH ..... 27
*S. Cortez, A. Completo, J.L. Alves*

HYBRID CELL CENTRED/VERTEX MODEL FOR LARGE TISSUE DEFORMATIONS ..... 37
*José J. Muñoz, Payman Mosaffa, Yanlan Mao, Rob Tetley, Nina Asadipour, Antonio Rodríguez-Ferran*

SENSITIVITY OF TEMPERATURE FIELD IN THE SYSTEM PROTECTIVE CLOTHING - FOREARM WITH RESPECT TO PERTURBATIONS OF EXTERNAL HEATING CONDITIONS ..... 46
*Bohdan Mochnacki, Mariusz Ciesielski*

**MS 103: MECHANICS OF BIOLOGICAL TISSUES**

INFLUENCE OF MUSCLE COMPRESSION ON DYNAMIC MUSCLE PERFORMANCE ..... 55
*Tobias Siebert, Norman Stutzig, Olaf Till, Christian Rode*

A 3D ELECTROMECHANICAL FEM-BASED MODEL FOR CARDIAC TISSUE ..... 63
*Minh Tuan Duong, Alexander Jung, Ralf Frotscher, Manfred Staat*

BIOMECHANICAL PROPERTIES OF CORNEAL EXTRACELLULAR MATRIX USING EXPERIMENTAL AND COMPUTATIONAL METHODS ..... 74
*Hamed Hatami-Marbini*

**MS 105: SIMULATION OF CARDIOVASCULAR PROCEDURES AND DEVICES**

MODELLING OF PATIENT-SPECIFIC CASES OF ATHEROSCLEROSIS IN CAROTID ARTERIES ..... 81
*Timur Gamilov, Roman Pryamonosov, Sergey Simakov*

PERSONALIZED COMPUTATION OF FRACTIONAL FLOW RESERVE IN CASE OF TWO CONSECUTIVE STENOSES ... 90
*Yuri Vassilevski, Timur Gamilov, Philip Kopylov*

**MS 106: DIRECT AND INVERSE METHODS FOR CARDIOVASCULAR AND PULMONARY BIOMECHANICS**

BLOOD FLOW IN THE COMMON CAROTID ARTERY WITH STENOSIS ..... 98
*Helena Henriques, Luisa Sousa, Catarina Castro, Carlos António, Rosa Santos, Pedro Castro, Elsa Azevedo*

**MS 108: NUMERICAL METHODS FOR COUPLED PROBLEMS IN BIOMEDICAL APPLICATIONS**
A FENICS-HPC FRAMEWORK FOR MULTI-COMPARTMENT BLOCH-TORREY MODELS ..... 105
*Dang Van Nguyen, Johan Jansson, Johan Hoffman*

**MS 111: POPULATION BALANCE MODELING: CURRENT STATUS, FUTURE PROSPECTS AND NOVEL APPLICATIONS FROM NANOPARTICLES' SYNTHESIS TO (LUNG) CANCER**
THERMODYNAMICS OF IRREVERSIBLE AGGREGATION ..... 120
*Themis Matsoukas*

**MS 112: ANEURYSMS: SOLID MECHANICS, FLUID MECHANICS, AND MECHANOBIOLOGY**
RELATIVE ROLES OF MECHANICS AND BIOCHEMISTRY IN THE INITIATION AND PROGRESSION OF CEREBRAL ANEURYSM THROMBOSIS ..... 137
*Malebogo Ngoepe, Yiannis Ventikos*

TECHNIQUES TO INTEGRATE PATIENT-SPECIFIC SIMULATION OF ANEURYSMAL BLOOD FLOW INTO THE CLINICAL WORKFLOW ..... 147
*Gábor Závodszky, Roland Joó-Kovács, György Paál, István Szikora*

**MS 113: MATHEMATICAL AND NUMERICAL MODELING OF THE HEART**
PROXIMAL ISOVELOCITY SURFACE FOR DIFFERENT MITRAL VALVE HOLE GEOMETRIES ..... 155
*Alexandre This, Hernán G. Morales, Odile Bonnefous*

COMPARISON OF TETRAHEDRAL AND HEXAHEDRAL MESHES FOR FINITE ELEMENT SIMULATION OF CARDIAC ELECTRO-MECHANICS ..... 164
*Bernardo Lino De Oliveira, Joakim Sundnes*

**MS 115: TUMOR GROWTH MODELING AND THE MECHANICAL ASPECTS OF CANCER**
A PARAMETRIC STUDY OF A MULTIPHASE POROUS MEDIA MODEL FOR TUMOR SPHEROIDS AND ENVIRONMENT INTERACTIONS ..... 178
*P. Mascheroni, D.P. Boso, C. Stigliano, M. Carfagna, L. Preziosi, P. Decuzzi, B.A. Schrefler*

**200 DAMAGE, FRACTURE AND FAILURE**

**MS 201: MICROSTRUCTURE-DRIVEN DEFORMATION AND FAILURE IN CRYSTALLINE MATERIALS**
STABILITY OF ASYMMETRIC GRAIN BOUNDARIES IN GRAPHENE ..... 188
*J.P. Mendez, F. Arca, M.P. Ariza*

DEFORMATION-DIFFUSION COUPLED ANALYSIS OF LONG-TERM HYDROGEN DIFFUSION IN NANOFILMS ..... 197
*Xingsheng Sun, Pilar Ariza, Kevin Wang*

MESH OBJECTIVE DAMAGE MODELING OF DUCTILE FRACTURE AT VISCO-PLASTIC CONTINUUM RESPONSE ... 209
*Senad Razanica, Ragnar Larsson, B. Lennart Josefson*

**MS 202: CIVIL ENGINEERING MATERIALS AND STRUCTURES UNDER EXTREME LOADINGS**

EQUIVALENT SDOF MODEL FOR ESTIMATING BLAST-INDUCED DYNAMIC REACTIONS OF EQUILATERAL TRIANGULAR HARDENED WALL ELEMENTS ..... 216
*Sebastian Mendes, Liling Cao, Douglas Heinze, Elisabeth Malsch*

MODELLING STRATEGIES OF PRESTRESSING TENDONS AND REINFORCEMENT BARS IN CONCRETE STRUCTURES ..... 234
*Antoine Llau, Ludovic Jason, Frédéric Dufour, Julien Baroth*

MODELING AND SIMULATION FOR AN OPTIMIZED DESIGN OF A DYNAMIC BEND TEST. ..... 242
*Rana Akiki, Fabrice Gatuingt, Cédric Giry, Nicolas Schmitt, Lavinia Stéfan*

**MS 203: COMPUTATIONAL METHODS FOR MODELLING INSTABILITIES IN SOLIDS & STRUCTURES**

ON STABILITY LOSS OF A THIN-WALLED SPHERICAL SHELL SUBJECTED TO EXTERNAL PRESSURE AND INTERNAL HOMOGENEOUS CORROSION ..... 254
*Emmanuel Gutman, Rudolf Bergman, Semyon Levitski*

EFFICIENCY OF DAMAGE-PLASTICITY MODELS IN CAPTURING COMPACTION-EXPANSION TRANSITION OF CONCRETE UNDER DIFFERENT COMPRESSION LOADING CONDITIONS ..... 261
*Reza Mousavi, Masoud D. Champiri, Kaspar J. Willam*

**MS 204: IMPACT AND CRASH MECHANICS**

A SIMPLIFIED MODEL OF A STEEL COLUMN SUBJECTED TO IMPACT ..... 273
*Piseth Heng, Mohammed Hjiaj, Jean-Marc Battini*

MATERIAL MODELLING OF CAST ALUMINIUM BY APPLICATION OF THE WILKINS DAMAGE MODEL ..... 287
*Christian Mühlstätter, Matthias Hartmann*

FINITE ELEMENT SIMULATION OF CRACK PROPAGATION AND DELAMINATION IN LAYERED SHELLS DUE TO BLADE CUTTING ..... 295
*Federica Confalonieri, Aldo Ghisi, Umberto Perego*

EVOLUTIONARY LEVEL SET METHOD FOR CRASHWORTHINESS TOPOLOGY OPTIMIZATION ..... 309
*Mariusz Bujny, Nikola Aulig, Markus Olhofer, Fabian Duddeck*

**MS 205: PROBABILISTIC APPROACH TO NUMERICAL SIMULATION OF FRACTURE**

NUMERICAL SIMULATION OF HIGH-VELOCITY INTERACTION ELEMENTS WITH THIN-WALLED STRUCTURES ... 323
*Alexander V. Gerasimov, Sergey V. Pashkov*

NUMERICAL SIMULATION OF MULTILAYER COMPOSITES FAILURE UNDER DYNAMIC LOADING ..... 334
*Sergey A. Zelepugin, Aleksey S. Zelepugin, Vladimir F. Tolkachev*

SIMULATION OF MECHANICAL PROPERTIES OF CERAMIC PARTS PRODUCED BY ADDITIVE TECHNOLOGIES IN WIDE RANGE OF LOADING RATES ..... 341
*Vladimir A. Skripnyak, Evgeniya Skripnyak, Vladimir V. Skripnyak, Irina K. Vaganova*

FRACTURE OF THIN METAL SHEETS WITH DISTRIBUTION OF GRAIN SIZES IN THE LAYERS ..... 355
*Natalia V. Skripnyak, Vladimir V. Skripnyak, Vladimir A. Skripnyak*

BRITTLE OR QUASI-BRITTLE FRACTURE OF CERAMIC NANOCOMPOSITES UNDER DYNAMIC LOADING ..... 366
*Evgeniya G. Skripnyak, Vladimir A. Skripnyak, Vladimir V. Skripnyak, Natalia V. Skripnyak, Irina K. Vaganova*

**300 DISCRETIZATION METHODS, GRID, MESH AND SOLID GENERATION**

**MS 301: METHODS FOR CUT AND COMPOSITE MESHES: THEORY, ALGORITHMS AND APPLICATIONS**
REMOVING THE STABILIZATION PARAMETER IN FITTED AND UNFITTED SYMMETRIC NITSCHE FORMULATIONS ..... 373
*Christoph Lehrenfeld*

**MS 302: MESH GENERATION AND ADAPTION**
A SEMI-UNSTRUCTURED TURBOMACHINERY MESHING LIBRARY WITH FOCUS ON MODELING OF SPECIFIC GEOMETRICAL FEATURES ..... 384
*Marco Stelldinger, Thomas Giersch, Felix Figaschewsky, Arnold Kühhorn*

CARTESIAN MESHES WITH DYNAMIC LOCAL WAVELET-BASED REFINEMENT FOR FLOW AROUND BODY PROBLEMS. ..... 402
*Kirill Merkulov, Igor Menshov, Andrew Plenkin*

A MESH GENERATION METHOD FOR HISTORICAL MONUMENTAL BUILDINGS: AN INNOVATIVE APPROACH ... 409
*Giovanni Castellazzi, Antonio Maria D'Altri, Stefano De Miranda, Francesco Ubertini, Gabriele Bitelli, Alessandro Lambertini, Ilenia Selvaggi, Antonio Tralli*

**MS 303: CURVED MESH GENERATION FOR HIGH-ORDER METHODS**
HIGH-ORDER CURVED MESH GENERATION BY USING A FINE LINEAR TARGET MESH ..... 417
*Verena Schmid, Hadrien Beriot, Onur Atak, Gwenael Gabard*

AUTOMATIC GENERATION OF 3D UNSTRUCTURED HIGH-ORDER CURVILINEAR MESHE ..... 428
*Michael Turner, David Moxey, Spencerj.Sherwin, Joaquim Peiro*

MESH GENERATION FOR THE 2D NURBS-ENHANCED FINITE ELEMENT METHOD ..... 444
*Ruben Sevilla, Luke Rees, Oubay Hassan*

IMAGE SEGMENTATION TECHNIQUES FOR BIOMEDICAL MODELING: ELECTROPHYSIOLOGY AND HEMODYNAMICS ..... 454
*Alexander Danilov, Roman Pryamonosov, Alexandra Yurova*

A CONSISTENTLY LINEARISED SOLID MECHANICS BASED MESH DEFORMATION TECHNIQUE FOR HIGH ORDER CURVED ELEMENTS ..... 462
*Roman Poya, Ruben Sevilla, Antonio Gil*

**MS 304: COMPUTATIONAL MODELLING OF ADDITIVE PRODUCTION PROCESSES**
VALIDATION OF MODELLING ASSUMPTIONS FOR THE BUILDUP SIMULATION OF LASER BEAM MELTING ON THE BASIS OF THE RESIDUAL STRESS DISTRIBUTION ..... 469
*Fabian Bayerlein, Christian Zeller, Martin Wunderer, Johannes Weirather, Michael F. Zäh, Markus Schmid, Georg Schlick, Roland Hessert, Michael Hofmann, Thomas Uihlein*

**MS 305: ADVANCED MESHING METHODS FOR INDUSTRIAL APPLICATIONS**
MESHING OCEAN DOMAINS FOR COASTAL ENGINEERING APPLICATIONS ..... 480
*Alexandros Avdis, Christian Jacobs, Simon Mouradian, Jon Hill, Matthew Piggott*

HIGH-ORDER TRACKING METHOD FOR THE SIMULATION OF BURNING FRONTS IN 3D ..... 493
*Denis Gueyffier, Bastien Andrieu*

SINGULAR MESH GENERATION FROM MULTIPLE OVERSET MESHES: A TOOL FOR INDUSTRIAL APPLICATIONS...501
*Gennaro Abbruzzese, Marta Cordero Gracia, Mariola Gómez López, Nuno Vinha*

**MS 306: LATTICE SPRING METHODS FOR LINEAR AND NONLINEAR CONTINUA**
GENERATING TRIANGULAR LATTICES FOR SURFACES WITH IRREGULAR BOUNDARY ..... 511
*Tatiana Sá Marques, Vitor Dias Da Silva*

WAVE TRANSMISSION THROUGH NONLINEAR IMPACTING METAMATERIAL UNIT ..... 526
*Arnab Banerjee, Raj Das, Emilio Calius*

**400 FLOW PROBLEMS**

**MS 403: PARTICLE-BASED METHODS IN FLUID MECHANICS**
A PARTICLE FINITE ELEMENT METHOD FOR MACHINING SIMULATIONS ..... 539
*Juan Manuel Rodríguez Prieto, Pär Jonsén, Ales Svoboda*

DYNAMICS OF DUST PARTICLES NEAR THE ROUND BELL SUCTION INLET ..... 554
*Olga Averkova, Konstantin Logachev, Artur Logachev, Elena Tolmacheva*

RECOVERY OF DIFFERENTIATION/INTEGRATION COMPATIBILITY OF MESHLESS OPERATORS VIA LOCAL ADAPTATION OF THE POINT CLOUD IN THE CONTEXT OF NODAL INTEGRATION ..... 568
*Gabriel Fougeron, Guillaume Pierrot, Denis Aubry*

MODELING OF NON-SPHERICAL, ELONGATED PARTICLES FOR INDUSTRIAL SUSPENSION FLOW SIMULATION ... 586
*Jakob Dominik Redlinger-Pohn, Lisa Maria König, Christoph Kloss, Christoph Goniva, Stefan Radl*

A LAGRANGIAN FINITE ELEMENT METHOD FOR THE SIMULATION OF 3D COMPRESSIBLE FLOWS ..... 600
*Massimiliano Cremonesi, Attilio Frangi*

A LAGRANGIAN PFEM APPROACH TO THE NUMERICAL SIMULATION OF 3D LARGE SCALE LANDSLIDES IMPINGING IN WATER RESERVOIRS ..... 608
*Massimiliano Cremonesi, Francesco Ferri, Umberto Perego*

A NUMERICAL INVESTIGATION OF FLOW DYNAMICS OVER A TRAPEZOIDAL SMOOTH OPEN CHANNEL ……..... 619
*Giacomo Viccione*

A NUMERICAL INVESTIGATION OF LIQUID IMPACT ON PLANAR SURFACES. …….. 627
*Giacomo Viccione, Vittorio Bovolin, Eugenio Pugliese Carratelli*

AERODYNAMICS OF AN UNSTEADY PARTICLE FLOW IN THE CHUTE …….. 638
*Olga Averkova, Ivan Logachev, Konstantin Logachev*

**MS 404: SIMULATION OF ENVIRONMENTAL FLOWS**
PROJECTION-BASED MODEL REDUCTION FOR FINITE ELEMENT APPROXIMATION OF SHALLOW WATER FLOWS …….. 650
*Matthew Farthing, Alexander Lozovskiy, Christopher Kees*

EFFICIENT TWO-DIMENSIONAL SIMULATION MODELS FOR HYDRAULIC AND MORPHODYNAMIC TRANSIENTS..670
*Pilar Garcia-Navarro, Javier Murillo, Mario Morales-Hernandez, Carmelo Juez, Asier Lacasta*

ADDRESSING THE CHALLENGES OF IMPLEMENTATION OF HIGH-ORDER FINITE-VOLUME SCHEMES FOR ATMOSPHERIC DYNAMICS ON UNSTRUCTURED MESHES …….. 684
*Panagiotis Tsoutsanis, Dimitris Drikakis*

MODELLING OF LONG WAVES (SEICHES) IN CASCAIS BAY …….. 709
*Vera Bras, Antonio Trigo-Teixeira*

A MESHLESS METHOD FOR SIMULATION OF PARTICLE-DRIVEN GRAVITY CURRENTS WITH OBSTACLES ……. 720
*Karel Kovářík, Jozef Mužík, Dana Sitányiová*

APPLICATION OF THE SHALLOW WATER EQUATIONS TO REAL FLOODING CASE …….. 735
*Hani Ali, Pierre-Yves Lagrée, Jose-Maria Fullana*

**MS 405: COMPUTATIONAL MODELING OF MULTIPHASE FLOWS: ADVANCED METHODS, INTERFACE PHENOMENA AND ENVIRONMENTAL APPLICATIONS**
A MULTIPHASE MODEL FOR THE NUMERICAL SIMULATION OF ICE-FORMATION IN SEA-WATER …….. 750
*Vanessa Covello, Antonella Abbà, Luca Bonaventura, Alessandro Della Rocca, Lorenzo Valdettaro*

A 3D CFD NUMERICAL STUDY OF THE BUBBLE GENERATION PROCESS INTO A BUBBLE T-JUNCTION GENERATOR AND ITS COMPARISON WITH EXPERIMENTAL DATA: PART I …….. 772
*Santiago Arias, Adeline Montlaur*

**MS 406: ADVANCES IN COMPUTATIONAL METHODS FOR GAS-LIQUID TWO-PHASE FLOW**
COUPLING OF FLUID STUCTURE INTRACTION SOLVER WITH A VOF METHOD FOR MULTIPHASE STRUCTURE INTERACTION …….. 788
*Daniele Cerroni, Roberto Da Via', Sandro Manservisi, Filippo Menghini*

A METHOD OF CHAINED ANALYTICAL WAVE STRUCTURES FOR LARGE SCALE STRATIFIED TWO-PHASE PIPE FLOWS …….. 806
*Andreas Holm Akselsen*

**MS 408: MANIPULATION AND CONTROL OF TURBULENT FLOW**
MODELING OF EXCITATION OF CONTROLLING DISTURBANCES IN SWEPT WING BOUNDARY LAYER BY MEANS OF PLASMA ACTUATORS ..... 831
*Sergey Chernyshev, Andrey Kiselev, Aleksandr Kuryachii*

NUMERICAL SIMULATION OF INTERACTION PROCESS BETWEEN DIELECTRIC BARRIER DISCHARGE AND DUCT FLOW ..... 844
*Pavel Semenev, Dmitriy Pudovikov, Pavel Toktaliev*

IMPLEMENTATION OF POD AND DMD METHODS IN APACHE SPARK FRAMEWORK FOR SIMULATION OF UNSTEADY TURBULENT FLOW IN THE MODEL COMBUSTOR ..... 857
*Mikhail Kalugin, Sergei Strijhak*

**MS 409: CURRENT TRENDS IN MODELLING AND SIMULATION OF TURBULENT FLOWS**
NUMERICAL VALIDATION OF A FOUR PARAMETER LOGARITHMIC TURBULENCE MODEL ..... 865
*Daniele Cerroni, Roberto Da Vià, Sandro Manservisi, Filippo Menghini*

LARGE-EDDY SIMULATION AND FAR FIELD ACOUSTICS OF A SUBSONIC HOT JET ..... 878
*O. Labbé*

ON THE APPLICATION OF MODERN TURBULENCE MODELS IN THE FLOW LAMINARIZATION PROBLEMS ..... 893
*Sergey Chernyshev, Aleksander Ivanov, Andrey Kiselev, Dmitriy Sboev, Leonid Teperin, Petr Vorotnikov, Valeriy Vozhdaev*

ASSESSMENT OF RANS AND DES METHODS FOR THE AHMED BODY ..... 912
*Emmanuel Guilmineau, Gan Bo Deng, Alban Leroyer, Patrick Queutey, Michel Visonneau, Jeroen Wackers*

A WAY TO IMPROVE JET MODELING WITHIN RANS EQUATION SYSTEM ..... 922
*Alexey Troshin*

A MODIFIED SSG/LLR-$\omega$ REYNOLDS STRESS MODEL FOR PREDICTING BLUFF BODY AERODYNAMICS ..... 936
*Csaba Klajbár, László Könözsy, Karl W. Jenkins*

**MS 410: COMPLEX FLUID FLOWS IN ENGINEERING: MODELLING, SIMULATION AND OPTIMIZATION**
NUMERICAL ANALYSIS OF RAREFIED GAS FLOWS USING THE ACADEMIC CFD CODE GALATEA ..... 949
*Angelos Klothakis, Georgios Lygidakis, Ioannis Nikolos*

TOPOLOGY OPTIMIZATION FOR FLUID FLOW EMPLOYING LOCAL OPTIMALITY CRITERIA ..... 967
*Philip Sarstedt, Gerhard Kachel, Jörg Ettrich, Karl Bühler*

NUMERICAL SIMULATION OF THE CONJUGATE HEAT TRANSFER IN THE COOLING SYSTEM OF THE COMBUSTION CHAMBERS OF THE AVIATION RAMJET ON THE ENDOTHERMIC FUELS ..... 979
*Vadim Volokhov, Pavel Toktaliev, Sergei Martynenko*

SENSITIVITY STUDY OF THE NUMERICAL SETUP FOR AN AUTOMATIC OPTIMIZATION PROCEDURE FOR A HYDRAULIC MACHINE ..... 992
*Alexander Tismer, Markus Schlipf, Stefan Riedelbauch*

ADVANCES IN GEOMETRICAL PARAMETRIZATION AND REDUCED ORDER MODELS AND METHODS FOR COMPUTATIONAL FLUID DYNAMICS PROBLEMS IN APPLIED SCIENCES AND ENGINEERING: OVERVIEW AND PERSPECTIVES ..... 1013
*F. Salmoiraghi, F. Ballarin, G. Corsi, A. Mola, M. Tezzele, G. Rozza*

NEAR WALL BEHAVIOR OF IMPLICIT LARGE EDDY SIMULATIONS ..... 1032
*Ioannis Kokkinakis, Dimitris Drikakis*

VALIDATION AND VERIFICATION OF A 2D LATTICE BOLTZMANN SOLVER FOR INCOMPRESSIBLE FLUID FLOW ..... 1046
*Tamás István Józsa, Máté Szőke, Tom-Robin Teschner, László Könözsy, Irene Moulitsas*

**MS 411: NON-NEWTONIAN HEAT AND FLUID FLOW SUBJECTED TO MAGNETIC FORCES**
VALIDATION OF A MAGNETO- AND FERRO-HYDRODYNAMIC MODEL FOR NON-ISOTHERMAL FLOWS IN CONJUNCTION WITH NEWTONIAN AND NON-NEWTONIAN FLUIDS ..... 1061
*László Könözsy, Pietro Scienza, Dimitris Drikakis*

NUMERICAL STUDY OF MAGNETIC PARTICLES CONCENTRATION IN BIOFLUID (BLOOD) UNDER THE INFLUENCE OF HIGH GRADIENT MAGNETIC FIELD IN MICROCHANNEL ..... 1084
*Vassilios Loukopoulos, George Bourantas, Dimitrios Labropoulos, Vassilios-Martin Nikiforidis, Stéphane Bordas, George Nikiforidis*

**MS 412: NUMERICAL METHODS FOR WAVES AND FLOWS IN COASTAL AND DEEP WATER HYDRODYNAMICS**
NEW ALGORITHM FOR NUMERICAL SIMULATION OF SURFACE WAVES WITHIN THE FRAMEWORK OF THE FULL NONLINEAR DISPERSIVE MODEL ..... 1093
*Zinaida Fedotova, Oleg Gusev, Gayaz Khakimzyanov*

NUMERICAL MODELLING OF SURFACE WAVES IN THE FRAMEWORK OF THE SHALLOW WATER MODEL ..... 1104
*Nina Shokina, Gayaz Khakimzyanov*

ON NUMERICAL METHODS FOR SOLVING RUN-UP PROBLEMS. COMPARATIVE ANALYSIS OF NUMERICAL ALGORITHMS AND NUMERICAL RESULTS. ..... 1127
*Leonid Chubarov, Alexandr Rychkov, Gayaz Khakimzyanov, Yurii Shokin*

**MS 414: NEW TRENDS IN NUMERICAL METHODS FOR MULTI-MATERIAL COMPRESSIBLE FLUID FLOWS**
EULERIAN CALCULATIONS OF MULTIMATERIAL FLOWS WITH SUB-CELL RECONSTRUCTION OF INTERFACES ... 1139
*Igor Menshov, Pavel Zakharov*

SOME ACOUSTIC-TRANSPORT SPLITTING SCHEMES FOR TWO-PHASE COMPRESSIBLE FLOWS ..... 1151
*Simon Peluchon, Gérard Gallice, Pierre-Henri Maire*

LAGRANGE-FLUX EULERIAN SCHEMES FOR COMPRESSIBLE MULTIMATERIAL FLOWS ..... 1165
*Florian De Vuyst, Thibault Gasc, Renaud Motte, Mathieu Peybernes, Raphael Poncet*

ISOTROPY–PRESERVING SLOPE LIMITERS FOR FINITE VOLUME METHODS ON SQUARE MESHES ..... 1179
*Jan Velechovsky, Marianne Francois*

BUILDING A MORE EFFICIENT LAGRANGE-REMAP SCHEME THANKS TO PERFORMANCE MODELING ..... 1191
*Thibault Gasc, Florian De Vuyst, Mathieu Peybernes, Raphaël Poncet, Renaud Motte*

**MS 415: COMPUTATIONAL NON-NEWTONIAN FLUID MECHANICS**
DEVELOPMENT OF CONFINED VISCOPLASTIC FLOWS WITH HETEROGENEOUS WALL SLIP ..... 1205
*Pandelitsa Panaseti, Maria Philippou, Zacharias Kountouriotis, Georgios Georgiou*

SQUEEZE FLOW OF VISCOPLASTIC BINGHAM MATERIAL ..... 1215
*Larisa Muravleva*

**500 HIGH PERFORMANCE COMPUTING**

**MS 501: ALGORITHMIC ASPECTS OF HIGH-PERFORMANCE COMPUTING FOR MECHANICS AND PHYSICS**
PSEUDOMULTIGRID GAUSS–SEIDEL METHOD FOR LARGE SCALE AND HIGH PERFORMANCE COMPUTING ..... 1237
*Sergey Martynenko, Vadim Volokhov, Pavel Toktaliev*

ALGORITHM FOR FAST SIMULATIONS OF SPACE-TIME FINITE ELEMENT METHOD ..... 1260
*Marcin Skotniczny, Anna Paszynska, Maciej Paszynski*

HYBRID PARALLELISATION OF AN ALGORITHMICALLY DIFFERENTIATED ADJOINT SOLVER ..... 1268
*Pavanakumar Mohanamuraly, Jan Christian Huckelheim, Jens-Dominik Mueller*

**MS 503: HPC-BASED SIMULATIONS FOR THE ENGINEERING REALM AND INDUSTRIAL APPLICATIONS**
EFFICIENT PARALLEL GEOMETRY DISTRIBUTION FOR THE SIMULATION OF COMPLEX FLOWS ..... 1277
*Andreas Lintermann*

LARGE-SCALE FREE-SURFACE FLOW SIMULATION USING LATTICE BOLTZMANN METHOD ON MULTI-GPU CLUSTERS ..... 1294
*Naoyuki Onodera, Kunihide Ohashi,*

LARGE-SCALE LES ANALYSIS OF AUTOMOTIVE ENGINE COOLING FAN ..... 1305
*Yuji Kobayashi, Kenji Yoshida, Itsuhei Kohri, Masaharu Sakai, Hideo Asano*

DIRECT AEROACOUSTIC SIMULATION RELATED WITH MODE CHANGE IN A RECORDER ..... 1315
*Hiroshi Yokoyama, Ryoma Hamasuna, Akira Miki, Hirofumi Onitsuka, Akiyoshi Iida*

NUMERICAL ANALYSIS OF RADIO FREQUENCY BLACKOUT FOR ATMOSPHERIC REENTRY VEHICLE USING CFD-CEM COMBINED METHOD ..... 1325
*Yusuke Takahashi, Reo Nakasato, Nobuyuki Oshima*

DIRECT FEM LARGE SCALE COMPUTATION OF TURBULENT MULTIPHASE FLOW IN URBAN WATER SYSTEMS AND MARINE ENERGY ..... 1339
*Ezhilmathi Krishnasamy, Johan Hoffman, Johan Jansson*

**MS 504: NUMERICAL METHODS AND TOOLS FOR KEY EXASCALE COMPUTING CHALLENGES IN ENGINEERING AND APPLIED SCIENCES**
KINETIC MODELS AND ALGORITHMS FOR SOLUTION OF THE MAGNETOGASDYNAMIC PROBLEMS ON THE MODERN SUPERCOMPUTING SYSTEMS ..... 1352
*Boris Chetverushkin, Nicola D'Ascenzo, A. Saveliev, Valeri Saveliev*

ERROR ANALYSIS AND QUANTIFICATION IN NEURON SIMULATIONS ..... 1366
*Francesco Casalegno, Francesco Cremonesi, Stuart Yates, Michael L. Hines, Felix Schürmann, Fabien Delalondre*

EVALUATION OF SPARSE LINEAR ALGEBRA OPERATIONS IN TRILINOS ..... 1381
*Mohammad Siahatgar, Gabriele Von Voigt*

**MS 505: INTERACTIVE SIMULATIONS IN COMPUTATIONAL ENGINEERING**
AUTOMATIC VISUALIZATION AND CONTROL OF ARBITRARY NUMERICAL SIMULATIONS ..... 1392
*Jan P. Springer, Helen Wright*

**MS 506: ACCURACY AND EFFICIENCY OF APPROXIMATE COMPUTATIONS IN SCIENCE AND ENGINEERING**
DIRECT TIME-DOMAIN INTEGRATION APPROACH FOR VISCOELASTIC SYSTEMS INVOLVING VARIOUS DAMPING MODELS ..... 1409
*Zhe Ding, Li Li, Yujin Hu*

ON THE PERFORMANCE OF A COMPUTATIONAL COST REDUCTION TECHNIQUE APPLIED TO COOLING TOWERS TRANSIENT ANALYSIS ..... 1421
*Aram Soroushian, Amin Jahani Mehrnoosh, Yalda Zarabimanesh, Mohammad Hadi Ghondaghsaz, Ali Baiani, Ali Zakizade*

A DIFFERENT LOOK AT THE RICHARDSON EXTRAPOLATION LEADING TO A NEW PROPOSITION ..... 1432
*Aram Soroushian*

ON THE FREQUENCY CONTENT OF ERRORS ORIGINATED IN A TIME INTEGRATION COMPUTATIONAL COST REDUCTION TECHNIQUE ..... 1443
*Aram Soroushian, Mahmood Hosseini, Seyed Mohammad Khalkhali*

**600 INTERDISCIPLINARY COUPLED AND CONTACT PROBLEMS**

**MS 601: SHOCK WAVE-BOUNDARY LAYER INTERACTION AND ITS CONTROL**
NUMERICAL SIMULATION OF TRANSONIC BUFFET AND ITS CONTROL USING TANGENTIAL JET BLOWING ..... 1455
*Ksenia Abramova, Kamil Khairullin, Alexander Ryzhov, Vitaly Soudakov*

SHOCK WAVE BOUNDARY LAYER INTERACTION IN PROFILE CASCADES REPRESENTING ROTOR BLADINGS OF THE LAST STAGE OF LARGE OUTPUT STEAM TURBINES ..... 1464
*Martin Luxa, David Simurda, Jaromir Prihoda, Jana Vachova*

SHOCK WAVE BOUNDARY LAYER INTERACTION INVESTIGATION: CONTROL BY ROD VORTEX GENERATORS AND THEIR APPLICATION ON HELICOPTER ROTOR BLADES ..... 1474
*Fernando Tejero Embuena, Piotr Doerffer, Pawel Flaszynski*

ANALYSIS OF EFFECTS OF SHAPE AND LOCATION OF MICRO - TURBULATORS ON UNSTEADY SHOCKWAVE - BOUNDARY LAYER INTERACTION IN TRANSONIC FLOW ..... 1484
*Janusz Sznajder, Tomasz Kwiatkowski*

**MS 602: INNOVATIVE METHODS FOR FLUID-STRUCTURE-INTERACTION**
ASSESSMENT OF THE FLUID-STRUCTURE INTERACTION CAPABILITIES FOR AERONAUTICAL APPLICATIONS OF THE OPEN-SOURCE SOLVER SU2. ..... 1498
*Ruben Sanchez, Heather L. Kline, David Thomas, Anil Variyar, Marcello Righi, Thomas D. Economon, Juan J. Alonso, Rafael Palacios, Grigorios Dimitriadis, Vincent Terrapon*

CONVERGENCE ANALYSIS OF COUPLING ITERATIONS FOR THE UNSTEADY TRANSMISSION PROBLEM WITH MIXED DISCRETIZATIONS ..... 1530
*Azahar Monge, Philipp Birken*

USING IMPROVED RADIAL BASIS FUNCTIONS METHODS FOR FLUID-STRUCTURE COUPLING AND MESH DEFORMATION ..... 1545
*Giorgos Strofylas, Georgios Mazanakis, Sotirios Sarakinos, Georgios Lygidakis, Ioannis Nikolos*

A MONOLITHIC FLUID-STRUCTURE INTERACTION METHOD, APPLICATION TO A PISTON PROBLEM ..... 1564
*Felix Ischinger, Martijn Anthonissen, Barry Koren*

**MS 603: COMPUTATIONAL METHODS IN FLUID-STRUCTURE INTERACTION WITH IMPACT ON INDUSTRIAL APPLICATIONS**
SLAMMING IMPACT SIMULATION OF 2D WATER ENTRY FOR RIGID STRUCTURES ..... 1582
*Omar Hashim Hassoon, Mostapha Tarfaoui, A. El Malki*

A MESH MORPHING BASED FSI METHOD USED IN AERONAUTICAL OPTIMIZATION APPLICATIONS ..... 1594
*Matej Andrejašič, David Eržen, Emiliano Costa, Stefano Porziani, Marco Evangelos Biancolini, Corrado Groth*

A PATH-CONSERVATIVE OSHER-TYPE SCHEME FOR AXIALLY SYMMETRIC COMPRESSIBLE FLOWS IN FLEXIBLE TUBES ..... 1613
*Julia Leibinger, Michael Dumbser, Uwe Iben*

**MS 605: FRICTIONAL CONTACTS WITH LUBRICATION – BASICS AND APPLICATIONS**
LIMITING SHEAR STRESS FORMULATION FOR TEHL SIMULATION ..... 1623
*Thomas Lohner, Klaus Michaelis, Karsten Stahl*

**MS 606: COMPUTATIONAL MODELING OF HYDRAULIC FRACTURING**
THE XFEM FOR A SIMPLIFIED MODEL IN HYDRAULIC FRACTURING ..... 1637
*Markus Schätzer, Thomas-Peter Fries*

SENSITIVITY ANALYSIS OF HYDRAULIC FRACTURING USING AN EXTENDED FINITE ELEMENT METHOD FOR THE PKN MODEL ..... 1647
*Hasini Garikapati, Clemens V. Verhoosel, Harald Van Brummelen, Pedro Diez*

NUMERICAL MODELLING OF COUPLED HYDRO-THERMAL PROCESSES OF THE SOULTZ HETEROGENEOUS GEOTHERMAL SYSTEM ..... 1659
*Musa Dahiru Aliyu, Hua-Peng Chen*

**MS 607: ADVANCES IN COMPUTATIONAL METHODS FOR LIQUID-VAPOR FLOWS WITH PHASE TRANSFER PROCESSES**
A LOW DIFFUSIVE APPROACH FOR TWO PHASE FLOWS ..... 1672
*Alberto Beccantini*

A LARGE TIME STEP ROE SCHEME APPLIED TO TWO-PHASE FLOW ..... 1693
*Sofia Lindqvist, Halvor Lund*

**MS 608:ADVANCES IN TIME INTEGRATION FOR SOLID, FLUID AND COUPLED SYSTEMS**
ENERGY CONSERVING TIME INTEGRATION BASED ON GALERKIN-VARIATIONAL INTEGRATORS WITH CONSTRAINTS ..... 1709
*Matthias Bartelt, Michael Groß*

MIXED INTEGRATORS FOR STRUCTURE-PRESERVING SIMULATIONS IN NONLINEAR STRUCTURAL DYNAMICS ..... 1736
*Alexander Janz, Peter Betsch, Christian Hesch*

ENERGY AND MOMENTUM CONSERVING VARIATIONAL BASED TIME INTEGRATION OF ANISOTROPIC HYPERELASTIC CONTINUA ..... 1765
*Michael Groß, Rajesh Ramesh, Julian Dietzsch*

STRUCTURE-PRESERVING OPTIMAL CONTROL OF DISCRETE MECHANICAL SYSTEMS ..... 1789
*Peter Betsch, Christian Becker*

VARIATIONAL INTEGRATORS OF MIXED ORDER FOR DYNAMICAL SYSTEMS WITH MULTIPLE TIME SCALES AND SPLIT POTENTIALS ..... 1818
*Theresa Wenger, Sina Ober-Blöbaum, Sigrid Leyendecker*

CO-SIMULATION OF LARGE MBD, FEA AND DEM SYSTEMS ..... 1832
*Jose Ortiz, Neil Macdonald*

**MS 609: ADVANCED COMPUTATIONAL MODELING OF BATTERIES AND FUEL CELLS**
STOCHASTIC 3D MODELING OF AMORPHOUS MICROSTRUCTURES - A POWERFUL TOOL FOR VIRTUAL MATERIALS TESTING ..... 1838
*Matthias Neumann, Volker Schmidt*

**MS 610: NUMERICAL METHODS TO STUDY THE CONTACT MECHANICS OF DRY, ADHESIVE AND LUBRICATED ROUGH SURFACES**
PERCOLATION PROPERTIES OF THE FREE VOLUME GENERATED BY TWO ROUGH SURFACES IN CONTACT ..... 1850
*Paolo Cinat, Marco Paggi, Claudia Borri*

A PARAMETRICALLY TIME-DEPENDENT METHODOLOGY FOR RECIPROCATING CONTACT MECHANICS BETWEEN VISCOELASTIC SOLIDS ..... 1856
*Carmine Putignano, Giuseppe Carbone, Daniele Dini*

STATISTICAL AND GEOMETRICAL CHARACTERISTICS OF RANDOMLY ROUGH SURFACES USED FOR CONTACT SIMULATIONS. ..... 1864
*Rafael Schouwenaars, Carlos G. Figueroa, Víctor Hugo Jacobo, Armando Ortiz*

**MS 612: NUMERICAL SIMULATIONS FOR SMART-CITY APPLICATIONS**
NUMERICAL OPTIMIZATION OF NEAR-ROAD VEGETATION BARRIERS ..... 1876
*Ludek Benes, Viktor Sip*

**MS 615: COMPUTATIONAL MODELS IN MAGNETOHYDRODYNAMICS**
LARGE EDDY SIMULATIONS OF MAGNETIC FIELD EFFECT ON TURBULENT FLOW IN A SQUARE DUCT ..... 1886
*Jie Mao, Kunlei Zhang, Zhongquan Tan, Ke Liu*

**700 MATERIALS**

**MS 701: ADVANCED MATERIALS: COMPUTATIONAL ANALYSIS OF PROPERTIES AND PERFORMANCE**
OPTIMAL DESIGN AND ADDITIVE MANUFACTURING OF NOVEL REINFORCING ELEMENTS FOR COMPOSITE MATERIALS ..... 1893
*Francesco Fabbrocino, Ilenia Farina, Ada Amendola, Luciano Feo, Fernando Fraternali*

A TEMPERATURE SENSITIVE CRYSTAL PLASTICITY MODEL FOR THE PREDICTION OF HIGH TEMPERATURE MECHANICAL BEHAVIOUR OF MULTI-PHASE TIAL ALLOY ..... 1909
*M. Umer Ilyas, M. Rizviul Kabir*

LAYOUT OPTIMIZATION OF PIEZOELECTRIC ELEMENTS WITH EXTERNAL ELECTRIC CIRCUITS IN SMART CONSTRUCTIONS BASED ON SOLUTION OF THE NATURAL VIBRATIONS PROBLEM ..... 1920
*Nataliia Iurlova, Valerii P. Matveenko, Dmitriy A. Oshmarin, Nataliya V. Sevodina, Maksim A. Yurlov*

DETERMINATION OF OPTIMAL PARAMETERS FOR A PASSIVE RL-CIRCUIT BY SOLVING THE PROBLEM ON NATURAL VIBRATIONS OF ELECTROELASTIC BODIES ..... 1930
*Maksim Yurlov, Valerii P. Matveenko, Dmitriy A. Oshmarin, Nataliya V. Sevodina, Nataliya A. Iurlova*

COMPUTATIONAL MODELLING OF MECHANICAL BEHAVIOR OF BIOLOGICAL TISSUES FOR BIOMEDICAL APPLICATIONS ..... 1940
*Simin Li, Yang Liu, Juan Du, Begum Zeybek, Lorenzo Zani, Mark P. Lewis, Vadim V. Silberschmidt*

ANALYSIS OF FIBRE BRAGG GRATINGS SENSORS OPTIMAL PLACEMENT FOR MONITORING OF DAMAGE PROPAGATION IN LAMINATE COMPOSITES ..... 1950
*Mikhail Tashkinov*

SEVEN DIFFERENT WAYS TO MODEL VISCOELASTICITY IN A GEOMETRICALLY EXACT SETTING ..... 1959
*Alexey Shutov*

**MS 702: MODELING OF NANOFILLED COMPOSITES**
SIMULATING ATOMIC FORCE MICROSCOPY FOR THE DETERMINATION OF THE ELASTIC PROPERTIES OF NANOPARTICLE REINFORCED EPOXY RESIN ..... 1971
*Johannes Fankhänel, Andreas Kempe, Raimund Rolfes*

MULTI-SCALE FINITE ELEMENT ANALYSIS OF GRAPHENE/POLYMER NANOCOMPOSITES ELECTRICAL PERFORMANCE ..... 1984
*Asimina Manta, Matthieu Gresil, Constantinos Soutis*

**MS 705: IDENTIFICATION OF MATERIAL MODELS**
THE BOOTSTRAP APPROACH TO THE STATISTICAL SIGNIFICANCE OF PARAMETERS IN RSM MODEL ..... 2003
*Jacek Pietraszek, Leszek Wojnar*

**MS 706: MODELING OF FIBER-BASED STRUCTURES - TEXTILES AND TEXTILE REINFORCED COMPOSITES**
INVERSE IDENTIFICATION OF THE BENDING STIFFNESS OF A BRAIDED POLYETHYLENE TWINE SUBJECT TO LARGE DEFORMATION: APPLICATION TO THE IDENTIFICATION OF THE MESH OPENING RIGIDITY OF FISHING NETS ..... 2010
*Barthélémy Morvan, Guilhem Bles, Nicolas Dumergue, Daniel Priour*

MULTI-SCALE FRAMEWORK FOR MODELLING CARBON FIBRE WEAVES BASED ON STOCHASTIC REINFORCEMENT GEOMETRY ..... 2029
*Andy Vanaerschot, Karen Soete, Stepan Lomov, Dirk Vandepitte*

**MS 707: MICROMECHANICAL MODELLING: COMPETITION BETWEEN ANALYTICAL AND NUMERICAL METHODS**
MULTISCALE MODELING OF THERMO-ELASTIC PROPERTIES OF MICROCRACKED MATERIAL ..... 2039
*Antonino Favata, Patrizia Trovalusci, Renato Masiani*

EFFECTIVE ELASTICITY COEFFICIENTS IN DRY POROUS MATERIALS. NUMERICAL AND SEMI-ANALYTICAL APPROACHES ..... 2050
*Vladimir Sladek, Bruno Musil, Jan Sladek*

INFLUENCE OF MICRO CRACKS ON EFFECTIVE MATERIAL PROPERTIES IN FIBER REINFORCED SMART COMPOSITE MATERIALS ..... 2073
*Michael Wünsche, Jan Sladek, Vladimir Sladek*

DEFECT ACCUMULATION IN NANOPOROUS WEAR-RESISTANT COATINGS UNDER COLLECTIVE RECRYSTALLIZATION. SIMULATION BY HYBRID CELLULAR AUTOMATON METHOD ..... 2080
*Dmitry Moiseenko, Pavel Maksimov, Sergey Panin, Viktor Panin*

**MS 711: FOURIER-BASED METHODS FOR COMPUTING THE BEHAVIOR OF HETEROGENEOUS MATERIALS DEVELOPMENTS, EXTENSIONS AND APPLICATIONS**
NONLINEAR COMPOSITE VOXELS AND FFT-BASED HOMOGENIZATION ..... 2099
*Matthias Kabel, Andreas Fink, Felix Ospald, Matti Schneider*

**MS 712: SMART MATERIAL SYSTEMS AND STRUCTURES**
A THERMODYNAMICALLY CONSISTENT FINITE STRAIN MICRO-SPHERE FRAMEWORK FOR PHASE-TRANSFORMATION ..... 2110
*Richard Ostwald, Thorsten Bartel, Andreas Menzel*

OPTIMAL CONTROL TUNNING IN SMART STRUCTURES WITH DELAMINATIONS ..... 2123
*Panos Koutsianitis, Amalia Moutsopoulou, Georgios Drosopoulos, Georgios Tairidis, Georgia Foutsitzi, Georgios Stavroulakis*

**MS 713: MICROSTRUCTURE-BASED MODELLING OF HETEROGENEOUS MATERIALS**
DEPENDENCE OF THE MECHANICAL PROPERTIES OF PENTAMODE MATERIALS ON THE LATTICE MICROSTRUCTURE ..... 2134
*Ada Amendola, Francesco Fabbrocino, Luciano Feo, Ferdinando Auricchio, Fernando Fraternali*

HOMOGENIZATION TECHNIQUES FOR ACCURATE AND APPROXIMATE ESTIMATES FOR OVERALL PROPERTIES OF MICROCRACKED VISCOELASTIC MASONRIES ..... 2151
*Amna Rekik, Alain Gasser*

APPLICATION OF POLAR ELASTICITY TO THE PROBLEM OF PURE BENDING OF A THICK PLATE ..... 2162
*Svitlana Fedorova, Jiří Burša*

ARTIFICIAL MICROSTRUCTURE GENERATOR FOR DUAL-PHASE STEELS ..... 2172
*Yuliang Hou, Alexandre Dumon, Pièrre Culière, Mohamed Rachik*

**MS 714: STRENGTH, FATIGUE AND STABILITY OF COMPOSITE STRUCTURES**
ON THE STRUCTURAL DESIGN OF IMPERFECTION SENSITIVE LAMINATED COMPOSITE SHELL STRUCTURES SUBJECTED TO AXIAL COMPRESSION ..... 2182
*Linus Friedrich, Pawel Lyssakow, Garth Pearce, Martin Ruess, Chiara Bisagni, Kai-Uwe Schröder*

STRENGTH AND WEIGHT EQUIVALENT SUBSTITUTION OF LARGE SANDWICH PANELS BY MONOLITHIC CFRP STRUCTURES ..... 2190
*Martin Meindlhumer, Martin Schagerl*

ROBUST INTEGRAL COMPOSITE AIRCRAFT STRUCTURES ..... 2202
*Tamas Havar*

**MS 715: COMPUTATIONAL ANALYSIS OF COMPOSITE STRUCTURES**
MINIMAL MASS DESIGN OF STRENGTHENING TECHNIQUES FOR PLANAR AND CURVED MASONRY STRUCTURES ..... 2210
*Gerardo Carpentieri, Francesco Fabbrocino, Mariella De Piano, Valentino Paolo Berardi, Luciano Feo, Fernando Fraternali*

MODELING OF FIBER-REINFORCED PLASTICS TAKING INTO ACCOUNT THE MANUFACTURING PROCESS …….. 2220
*Cherry Ann Reclusado, Sumito Nagasawa*

CONVERGENCE ANALYSIS OF STRESS FIELDS TO HOMOGENIZATION PREDICTIONS IN OPTIMAL PERIODIC COMPOSITE DESIGN …….. 2236
*Pedro Coelho, Rui Reis, José Guedes*

A NUMERICAL STUDY FOR THE BUCKLING CAPACITY OF WIND TURBINE BLADES: GEOMETRY, LOADING AND MATERIAL INFLUENCE …….. 2247
*Efstathios E. Theotokoglou, George A. Balokas, Evgenia K. Savvaki*

**800 MULTISCALE PROBLEMS**

**MS 801: MULTISCALE COMPUTATIONAL HOMOGENIZATION FOR BRIDGING SCALES IN THE MECHANICS AND PHYSICS OF COMPLEX MATERIALS**

A SEMI-IMPLICIT MICROPLAR DISCRETE-TO-CONTINUUM METHOD FOR GRANULAR MATERIALS …….. 2266
*Kun Wang, Waiching Sun*

MODELING OF LOW-ALLOYED TRIP-STEELS BASED ON DIRECT MICRO-MACRO SIMULATIONS …….. 2280
*Stefan Prüger, Ashutosh Gandhi, Daniel Balzani*

HOMOGENIZATION METHOD FOR DESIGNING NOVEL ARCHITECTURED CELLULAR MATERIALS …….. 2292
*Zheng-Dong Ma*

A HOMOGENIZATION TECHNIQUE FOR ELASTO-PLASTIC COMPOSITES …….. 2316
*Federica Covezzi, Stefano De Miranda, Sonia Marfia, Elio Sacco*

**MS 802: ADVANCES IN THE MODELLING OF MULTI-SCALE, MULTI-PHYSICS AND MULTI-UNCERTAINTY PROBLEMS**

THEORETICAL PREDICTION OF PLASTIC INSTABILITY FOR ASYMMETRICAL MATERIALS …….. 2323
*Shenghua Wu, Nannan Song, F. M. Andrade Pires*

MULTISCALE SIMULATION OF THREE-DIMENSIONAL UNSTEADY GAS FLOWS IN MICROCHANNELS OF TECHNICAL SYSTEMS …….. 2331
*Viktoriia Podryga, Yury Karamzin, Tatiana Kudryashova, Sergey Polyakov*

HIGHER-ORDER ASYMPTOTIC HOMOGENIZATION OF PERIODIC MATERIALS WITH LOW SCALE SEPARATION …….. 2346
*M. Mohammed Ameen, R.H.J. Peerlings, M.G.D. Geers*

**MS 805: ADVANCED MULTI-PHYSICS AND MULTI-SCALE TECHNIQUES FOR MODELING INELASTIC PROCESSES IN SOLIDS: DAMAGE, FRACTURE AND CONTACT MECHANICS**

MESO-SCALE MODELING OF HYBRID INDUSTRIAL/RECYCLED STEEL FIBER-REINFORCED CONCRETE …….. 2353
*Antonio Caggiano, Diego Said Schicchi, Guillermo Etse, Enzo Martinelli*

**MS 806: MULTISCALE MODELLING OF MATERIALS AND STRUCTURES**
THE KIRKENDALL SHIFT AND FRENKEL EFFECT DURING MULTI-COMPONENT DIFFUSION PROCESS ..... 2363
*Bartek Wierzba*

CRACK PATH FIELD AND STRAIN INJECTION TECHNIQUES IN DYNAMIC FRACTURE SIMULATIONS ..... 2372
*Oriol Lloberas-Valls, Alfredo E. Huespe, Javier Oliver, Ivo F. Dias*

HYBRID FE/XFE FINITE ELEMENT MODEL FOR SIMULATION OF BRITTLE-DUCTILE FRACTURES IN DUAL-PHASE STEEL GRADES. ..... 2390
*Konrad Perzynski, Lukasz Madej*

EFFECTIVE MESO AND MACRO PROPERTIES FOR FIBRE-REINFORCED-POLYMER CURING COUPLED TO VISCO-ELASTICITY ..... 2398
*Rolf Mahnken, Christian Dammann*

ON APPLICATION OF THE MESHLESS FINITE DIFFERENCE METHOD TO NUMERICAL HOMOGENIZATION ..... 2417
*Irena Jaworska*

A MESOSCALE STUDY OF DEFORMATION AND FAILURE OF ANGLE-PLY LAMINATES UNDER TENSILE LOADING BY MEANS OF NUMERICAL HOMOGENIZATION ..... 2426
*Marek Romanowicz*

MODELING OF PHASE CHANGES IN MICRO-DOMAIN INDUCED BY AN ULTRASHORT LASER PULSE ..... 2437
*Ewa Majchrzak, Lukasz Turchan, Jolanta Dziatkiewicz*

**MS 807: ADVANCED COMPUTATIONAL STRATEGIES FOR MODELLING, SIMULATION AND CHARACTERISATION OF MULTI-SCALE HETEROGENEOUS MATERIALS**
VALIDATION OF A DISCRETE CRACK MODEL FOR LIGHTWEIGHT AGGREGATE CONCRETE BEAMS ..... 2449
*D. Dias-Da-Costa, R. Graça-E-Costa, R.N.F. Carmo*

**MS 808: MULTISCALE AND MULTIPHYSICS MODELING OF CEMENTITIOUS MATERIALS**
REINFORCEMENT CORROSION IN CONCRETE DUE TO CARBONATION AND CHLORIDE INGRESS UP AND BEYOND INDUCTION PERIOD ..... 2460
*Karolina Hájková, Libor Jendele, Vit Šmilauer, T. Sajdlová, Jan Červenka*

**MS 809: MULTISCALE STOCHASTIC FINITE ELEMENT METHODS**
THE STOCHASTIC FINITE ELEMENT METHOD FOR NUCLEAR APPLICATIONS ..... 2471
*José David Arregui Mena, Lee Margetts, Llion Evans, D. V. Griffiths, Anton Shterenlikht, Luis Cebamanos, Paul M Mummery*

DETERMINATION OF MESOSCALE RANDOM FIELDS FOR THE APPARENT PROPERTIES OF SPATIALLY RANDOM COMPOSITES ..... 2484
*Dimitrios Savvas, George Stefanou*

**MS 811: MULTISCALE MODELING OF CONCRETE AND CONCRETE STRUCTURES**
NUMERICAL SIMULATION OF THE AUTOGENOUS SHRINKAGE OF HARDENING PORTLAND CEMENT PASTE ... 2501
*Peng Gao, Guang Ye, Jiangxiong Wei, Qijun Yu*

**900 NUMERICAL SIMULATION METHODS**

**MS 901: ISOGEOMETRIC METHODS**
ANALYSIS OF LAMINATED COMPOSITE PLATES USING ISOGEOMETRIC COLLOCATION METHOD ..... 2514
*G. S. Pavan, K. S. Nanjunda Rao*

MODELLING STIFFENED LIGHTWEIGHT STRUCTURES WITH ISOGEOMETRIC ANALYSIS VIA MORTAR METHODS ..... 2531
*Malte Woidt, Kay Sommerwerk, Matthias C. Haupt, Peter Horst*

DOMAIN DECOMPOSITION SOLUTION SCHEMES FOR LARGE-SCALE IGA PROBLEMS ..... 2541
*George Stavroulakis, Dimitris Tsapetis, Manolis Papadrakakis*

**MS 902: INNOVATIVE NUMERICAL APPROACHES FOR MULTI-PHYSICS PROBLEMS**
NUMERICAL PROPERTIES OF A DISCONTINUOUS GALERKIN FOMULATION FOR ELECTRO-THERMAL COUPLED PROBLEMS ..... 2558
*Lina Homsi, Christophe Geuzaine, Ludovic Noels*

**MS 903: ADVANCES IN FICTITIOUS DOMAIN METHODS FOR SOLID MECHANICS**
TOWARDS LATTICE-BOLTZMANN ON DYNAMICALLY ADAPTIVE GRIDS — MINIMALLY-INVASIVE GRID EXCHANGE IN ESPRESSO ..... 2566
*Michael Lahnert, Carsten Burstedde, Christian Holm, Miriam Mehl, Georg Rempfer, Florian Weik*

AUTOMATIC CONFORMAL DECOMPOSITION OF ELEMENTS CUT BY NURBS ..... 2591
*Jakob W. Steidl, Thomas-Peter Fries*

A DESIGN-THROUGH-ANALYSIS APPROACH USING THE FINITE CELL METHOD ..... 2601
*Benjamin Wassermann, Tino Bog, Stefan Kollmannsberger, Ernst Rank*

**MS 905: DISCONTINUOUS GALERKIN METHODS: NEW TRENDS AND APPLICATIONS**
DISCONTINUOUS GALERKIN COMPUTATION OF GASEOUS MIXTURE COAXIAL JETS ..... 2614
*Nicoletta Franchina, Marco Savini, Francesco Bassi*

DISCONTINUOUS GALERKIN SOLUTION OF THE REYNOLDS-AVERAGED NAVIER–STOKES AND KL-KT-LOG(W) TRANSITION MODEL EQUATIONS ..... 2632
*A. Bassi, Alessandro Colombo, Antonio Ghidoni, Marco Lorini, Gianmaria Noventa*

A STUDY OF MULTIGRID SMOOTHERS USED IN COMPRESSIBLE CFD BASED ON THE CONVECTION DIFFUSION EQUATION ..... 2648
*Philipp Birken, Jonathan Bull, Antony Jameson*

HIGH-ORDER LINEARLY IMPLICIT TWO-STEP PEER METHODS FOR THE DISCONTINUOUS GALERKIN SOLUTION OF THE INCOMPRESSIBLE RANS EQUATIONS ..... 2664
*Francesco Carlo Massa, Gianmaria Noventa, Francesco Bassi, Alessandro Colombo, Antonio Ghidoni, Marco Lorini*
OUTPUT-BASED SPACE-TIME ADAPTATION WITH NON-VARIATIONAL TIME INTEGRATION ..... 2684
*Krzysztof Fidkowski*

MATRIX-FREE MODIFIED EXTENDED BACKWARD DIFFERENTIATION FORMULAE APPLIED TO THE DISCONTINUOUS GALERKIN SOLUTION OF COMPRESSIBLE UNSTEADY VISCOUS FLOWS ..... 2697
*Alessandra Nigro, Carmine De Bartolo, Andrea Crivellini, Francesco Bassi*

**MS 906: MATHEMATICAL ADVANCES IN ISOGEOMETRIC ANALYSIS**
PRECONDITIONERS FOR ISOGEOMETRIC ANALYSIS BASED ON SOLVERS FOR SYLVESTER EQUATION ..... 2715
*Giancarlo Sangalli, Mattia Tani*

**MS 907: REGULARIZED ENRICHED APPROXIMATIONS AND QUADRATURE FOR DISCONTINUITIES, SINGULARITIES AND CONTINUOUS-DISCONTINUOUS TRANSITION**
EFFECTIVE 3D REGULARIZED XFEM FOR PULL-OUT OF STEEL BARS IN CONCRETE, BENDING AND SHEAR TESTS ON FRP-REINFORCED CONCRETE BEAMS ..... 2726
*Elena Benvenuti, Nicola Orlando*

**MS 908: VERIFICATION AND VALIDATION OF STRUCTURAL MECHANICS SIMULATION MODELS**
EXPERIMENTAL VALIDATION OF COMPOSITE STRUCTURES IN EXPLICIT IMPACT ANALYSIS ..... 2734
*Konstantinos Fotopoulos, George Lampeas*

CODE VERIFICATION IN SHELL ANALYSIS BY THE METHOD OF MANUFACTURED SOLUTIONS ..... 2745
*Michael Gfrerer, Martin Schanz*

EXPERIMENTAL VALIDATION OF A MODEL UPDATE PROCEDURE FOCUSING ON SMALL GEOMETRIC DEVIATIONS ..... 2757
*Thomas Maywald, Arnold Kühhorn, Sven Schrape*

OPTIMUM DESIGN, FINITE ELEMENT MODEL UPDATING AND DYNAMIC ANALYSIS OF A FULL LAMINATED GLASS PANORAMIC CAR ELEVATOR ..... 2774
*Dimitrios Giagopoulos, Iraklis Chatziparasidis*

INVESTIGATION IN THE MECHANICAL BEHAVIOUR OF LOW-VELOCITY IMPACTED CFRP PLATES UNDER DIFFERENT BOUNDARY CONDITIONS ..... 2786
*Süleyman Arslan, Zamaan Sadeghi, Athanasios Dafnis, Kai-Uwe Schröder*

A NEW SHELL FINITE ELEMENT WITH DRILLING DEGREES OF FREEDOM AND ITS RELATION TO EXISTING FORMULATIONS ..... 2803
*Robert Winkler, Dimitrios Plakomytis*

**MS 910: HIGH ORDER CFD METHODS: CONCLUSIONS AND OUTLOOK**
CURVED GRID GENERATION AND DG COMPUTATION FOR THE DLR-F11 HIGH LIFT CONFIGURATION ..... 2843
*Ralf Hartmann, Harlan Mcmorris, Tobias Leicht*

BS1 - DNS OF THE TAYLOR-GREEN VORTEX AT RE = 1600 ..... 2859
*Andrea Mastellone, Luigi Cutrone, Francesco Capuano*

**MS 911: NUMERICAL METHODS IN THE MECHANICS OF GENERALIZED CONTINUA**
THE STRUCTURAL SYMMETRY WITHIN THE CONTEXT OF NONLOCAL ELASTICITY ..... 2869
*Aurora Angela Pisano, Paolo Fuschi*

ISOGEOMETRIC GALERKIN METHODS FOR GRADIENT-ELASTIC BARS, BEAMS, MEMBRANES AND PLATES ..... 2876
*Jarkko Niiranen, Sergei Khakalo, Viacheslav Balobanov, Josef Kiendl, Antti H. Niemi, Bahram Hosseini, Alessandro Reali*

**MS 912: HIGH-ORDER METHODS, SENSITIVITY ANALYSIS AND ADAPTATION FOR THE NAVIER STOKES EQUATIONS**
ADAPTIVE MESH REFINEMENT TECHNIQUES FOR HIGH-ORDER FINITE-VOLUME WENO SCHEME ..... 2883
*Harshavardhana Srinivasan, Panagiotis Tsoutsanis*

IMPLEMENTATION OF A LOW-MACH NUMBER MODIFICATION FOR HIGH-ORDER FINITE-VOLUME SCHEMES FOR ARBITRARY HYBRID UNSTRUCTURED MESHES ..... 2900
*Nicholas Simmonds, Panagiotis Tsoutsanis, Adrian Gaylard*

APPROPRIATE COMBINATIONS OF CONTROLLER PARAMETERS FOR UNSTEADY FLOW SIMULATIONS WITH ADAPTIVE TIME STEP CONTROL ..... 2916
*Kathrin Kozulovic, Graham Ashcroft*

**MS 913: HIGH-ORDER METHODS FOR POLYGONAL AND POLYHEDRAL MESHES**
ADAPTIVE BEM-BASED FEM ON POLYGONAL MESHES FROM VIRTUAL ELEMENT METHODS ..... 2930
*Steffen Weißer*

VEM AND TOPOLOGY OPTIMIZATION ON POLYGONAL MESHES ..... 2941
*Paola F. Antonietti, Matteo Bruggi, Simone Scacchi, Marco Verani*

THE VIRTUAL ELEMENT METHOD FOR DISCRETE FRACTURE NETWORK FLOW AND TRANSPORT SIMULATIONS ..... 2953
*Matias Fernando Benedetto, Stefano Berrone, Andrea Borio, Sandra Pieraccini, Stefano Scialo*

**MS 916: DIRECT METHODS FOR LIMIT STATES OF MATERIALS AND STRUCTURES**
COMPOSITE FEM MODELS FOR LIMIT AND SHAKEDOWN ANALYSIS ..... 2971
*Leonardo Leonetti, Giovanni Garcea, Hung Nguyen-Xuan*

ON SHAKEDOWN OF ELASTIC-PLASTIC BODIES WITH TEMPERATURE DEPENDENT PROPERTIES ..... 2983
*Michael Peigney*

A NUMERICAL STUDY ON THE ENDURANCE LIMIT OF PARTICULATE REINFORCED METAL MATRIX COMPOSITES (PRMMCS) USING THE DIRECT METHOD AND THE STATISTICAL LEARNING ..... 2994
*Geng Chen, Bezold Alexander, Christoph Broeckmann, Dieter Weichert*

SHAKEDOWN ANALYSIS OF PLATE BENDING UNDER STOCHASTIC UNCERTAINTY BY CHANCE CONSTRAINED PROGRAMMING ..... 3007
*Ngoc Trinh Tran, Thanh Ngoc Tran, H.G. Matthies, G.E. Stavroulakis, M. Staat*

LIMIT ANALYSIS LOCUS OF HIGH STRENGTH STEEL PLATES UNDER NON-QUADRATIC YIELD CRITERIA ..... 3020
*Konstantinos Nikolaou, Christos Bisbos*

**MS 917: MESOSCOPIC METHODS FOR COMPLEX FLUIDS AND SOFT MATTER**
PARTICLE BASED SIMULATION OF FLUID FLOW IN PERIODICALLY GROOVED CHANNELS ..... 3026
*Dorothea Kasiteropoulou, Theodoros Karakasidis, Antonios Liakopoulos*

**MS 918: COMPUTER ALGEBRA SYSTEMS IN MODELLING STATIC AND DYNAMIC PROBLEMS IN MECHANICS OF SOLIDS**
HIGH COMPUTATIONAL EFFICIENCY THROUGH GENERIC ANALYTICAL FORMULATION FOR LINEAR SOIL PRESSURE DISTRIBUTION OF RIGID SPREAD RECTANGULAR FOOTINGS ..... 3045
*John Bellos, Nikolaos P. Bakas*

A REAL-SPACE MODAL ANALYSIS METHOD FOR NON-PROPORTIONAL DAMPED STRUCTURES ..... 3065
*Evgueni Stanoev*

A FORCE-BASED FORMULATION FOR THE ANALYSIS OF 3-DIMENSIONAL INELASTIC STRUCTURAL FRAMES ... 3091
*Theodoros Patsios, Konstantinos Spiliopoulos*

THE FEATURES OF ANALYTICAL SOLUTIONS OF BOUNDARY VALUE PROBLEMS OF THE ELASTICITY THEORY FOR FINITE DOMAINS WITH ANGULAR POINTS OF A BOUNDARY ..... 3107
*Alexander Kerzhaev, Mikhail Kovalenko, Irina Menshova*

**MS 919: RECENT ADVANCES IN NUMERICAL SIMULATION AND ANALYSIS OF KINETIC MODELS**
DISCRETE VELOCITY MODELS: A STUDY OF THE HYDRODYNAMIC LIMIT ..... 3113
*Hans Babovsky*

MESO - MACRO MODELS FOR A HARD SPHERE GAS ..... 3121
*Sergey Bogomolov, Natalya Esikova, Artem Kuvshinnikov*

**MS 921: RECENT ADVANCES IN BOUNDARY ELEMENT METHODS**
ADVANCES IN THE BOUNDARY ELEMENT METHOD IN GEOMECHANICS ..... 3139
*Gernot Beer*

ENERGETIC BEM FOR THE NUMERICAL SOLUTION OF DAMPED WAVE PROPAGATION EXTERIOR PROBLEMS ..... 3159
*Alessandra Aimi, Mauro Diligenti, Chiara Guardasoni*

MODELLING THE PROCESS OF SEQUENTIAL EXCAVATION WITH THE BOUNDARY ELEMENT METHOD ..... 3168
*Christian Duenser, Gernot Beer*

GREEN'S FUNCTION FOR THE EVALUATION OF ANCHOR LOSSES IN MEMS ..... 3181
*Attilio Frangi, Massimiliano Cremonesi*

EFFICIENT BOUNDARY ELEMENT FORMULATION OF THERMOELASTICITY ..... 3190
*Relindis Rott, Martin Schanz*

A CHEBYSHEV INTERPOALTION-BASED FAST MULTIPOLE METHOD FOR POROELASTICITY ..... 3201
*Barbara Knöbl, Thomas Traub, Martin Schanz*

**MS 922: HIGH-ORDER METHODS FOR ELASTIC WAVES AND THEIR APPLICATION**
ENERGY STABLE HIGH ORDER FINITE DIFFERENCE METHODS ON STAGGERED GRIDS: AN INITIAL INVESTIGATION ..... 3211
*Ossian O'Reilly, Tomas Lundquist, Jan Nordström*

**MS 923: NOVEL DISCRETIZATION METHODS – MATHEMATICAL AND MECHANICAL ASPECTS**
ALTERNATIVE TO RETURN -MAPPING ALGORITH FOR COMPUTING PLASTIC STRAIN. APPLICATION TO DILATANT MATERIALS ..... 3224
*Siegfried Maiolino*

DISCONTINUOUS GALERKIN METHOD WITH REDUCED INTEGRATION SCHEME FOR THE BOUNDARY TERMS IN ALMOST INCOMPRESSIBLE LINEAR ELASTICITY ..... 3231
*Hamid Reza Bayat, Stephan Wulfinghoff, Stefanie Reese*

**1000 OPTIMIZATION, INVERSE PROBLEMS AND CONTROL**

**MS 1001: STRUCTURAL AND MULTIDISCIPLINARY OPTIMIZATION**
A METHOD FOR BI-DIRECTIONAL COUPLING OF STRUCTURE AND SYSTEM IN THE OPTIMISATION OF MULTI-FUNCTIONAL COMPONENTS ..... 3239
*Ajit Panesar, Ian Ashcroft, Ricky Wildman, Richard Hague*

SHAPE-TOPOLOGY OPTIMIZATION FOR DESIGNING SHELL STRUCTURES ..... 3249
*Hirotaka Nakayama, Masatoshi Shimoda*

ON A NASH GAME FOR TOPOLOGY OPTIMIZATION UNDER LOAD-UNCERTAINTY – FINDING THE WORST LOAD ..... 3259
*Carl-Johan Thore*

ANALYSIS AND DESIGN OF REINFORCED CONCRETE STRUCTURES AS A TOPOLOGY OPTIMIZATION PROBLEM .. ..... 3271
*Matteo Bruggi*

ADJOINT OPTIMAL CONTROL PROBLEMS FOR FLUID-STRUCTURE INTERACTION SYSTEMS ..... 3280
*Daniele Cerroni, Roberto Da Vià, Sandro Manservisi, Filippo Menghini, Luca Zaniboni*

APPLICATION OF WHIRL FLUTTER OPTIMIZATION-BASED SOLUTION TO FULL-SPAN MODEL OF TWIN TURBOPROP AIRCRAFT ..... 3293
*Jiri Cecrdle*

ON THE CONSIDERATION OF UNCERTAINTY IN DESIGN: OPTIMIZATION – RELIABILITY – ROBUSTNESS ..... 3310
*Nicolas Lelièvre, Pierre Beaurepaire, Cécile Mattrand, Nicolas Gayton, Abdelkader Ostmane*

A NEW OPTIMIZED ANISOTROPIC MOVING LEAST SQUARES SURROGATE MODEL WITH MAXIMIZED PROGNOSIS ..... 3321
*Kevin Cremanns, Dirk Roos*

LATIN HYPERCUBE SAMPLING BASED ON ADAPTIVE ORTHOGONAL DECOMPOSITION ..... 3333
*Dirk Roos*

A BENCHMARK OF CONTEMPORARY METAMODELING ALGORITHMS ..... 3344
*Can Bogoclu, Dirk Roos*

TOPOLOGY OPTIMIZATION USING A KRIGING-ASSISTED GENETIC ALGORITHM WITH A NOVEL LEVEL SET REPRESENTATION APPROACH ..... 3361
*Mitsuo Yoshimura, Koji Shimoyama, Takashi Misaka, Shigeru Obayashi*

TOPOLOGY OPTIMIZATION OF CONTACT PROBLEMS BASED ON ALLEN CAHN APPROACH ..... 3377
*Andrzej Myslinski*

ON THE OPTIMAL DESIGN OF CABLE-STAYED BRIDGES ..... 3386
*Gerardo Carpentieri, Mariano Modano, Francesco Fabbrocino, Luciano Feo, Fernando Fraternali*

A MIXED-INTEGER LINEAR PROGRAMMING APPROACH FOR GLOBAL DISCRETE SIZE OPTIMIZATION OF FRAME STRUCTURES ..... 3395
*Roxane Van Mellaert, Kristo Mela, Teemu Tiainen, Markku Heinisuo, Geert Lombaert, Mattias Schevenels*

A SUPER-STRUCTURE BASED OPTIMISATION APPROACH FOR BUILDING SPATIAL DESIGNS ..... 3409
*Koen Van Der Blom, Sjonnie Boonstra, Hèrm Hofmeyer, Michael Emmerich*

BENDING STIFFNESS OF A MULTILAYERED PLATE ..... 3423
*Petr Tovstik, Tatiana M. Tovstik*

EFFICIENT SIZING OF STRUCTURES UNDER STRESS CONSTRAINTS ..... 3436
*Zhi Hong, Mostafa Abdalla*

TOPOLOGY OPTIMIZATION OF WAVE BARRIERS FOR RAILWAY INDUCED VIBRATIONS IN BUILDINGS ..... 3461
*Cédric Van Hoorickx, Mattias Schevenels, Geert Lombaert*

**MS 1002: EVOLUTIONARY ALGORITHMS AND METAHEURISTICS IN CIVIL ENGINEERING AND CONSTRUCTION MANAGEMENT**

MULTIOBJECTIVE OPTIMIZATION USING GENETIC ALGORITHMS IN TIME-COST CONSTRUCTION PROJECT SCHEDULING PROBLEM ..... 3473
*Jorge Magalhães-Mendes*

APPLICATION OF GENETIC ALGORITHMS TO STRUTTED SHEET PILE WALL DESIGN OPTIMIZATION ..... 3485
*Mohamed Eid, Remon Isaac*

NASH EVOLUTIONARY ALGORITHMS: TESTING PROBLEM SIZE IN RECONSTRUCTION PROBLEMS IN FRAME STRUCTURES ..... 3493
*David Greiner, Jacques Periaux, Jose M. Emperador, Blas Galván, Gabriel Winter*

OPTIMUM DESIGN OF REINFORCED CONCRETE FRAMES ACCORDING TO EC8 AND MC2010 WITH GENETIC ALGORITHMS ..... 3505
*Panagiotis Mergos*

CALIBRATION OF THE NUMERICAL MODEL OF A STAND IN DRAGÃO STADIUM BASED ON GENETIC ALGORITHMS ..... 3519
*Jorge Leite, Diogo Ribeiro, Hugo Marques, Rui Calçada*

**MS 1003: ADVANCES IN DESIGN OPTIMIZATION OF STRUCTURES AND MATERIALS**
DESIGN OF ISOTROPIC MICROSTRUCTURES VIA A TWO-SCALE APPROACH ..... 3534
*Alexis Faure, Georgios Michailidis, Rafael Estevez, Guillaume Parry, Grégoire Allaire*

SHAPE OPTIMIZATION OF GRAPHENE SHEETS FOR MAXIMUM FUNDAMENTAL FREQUENCY ..... 3546
*Jin-Xing Shi, Masatoshi Shimoda*

TOPOLOGY OPTIMIZATION WITH AN IMPLICIT FUNCTION AND PARAMETERIZED CUTTING SURFACE ..... 3553
*Peter Dunning*

TOPOLOGY OPTIMIZATION OF PERIODIC STRUCTURES FOR COUPLED ACOUSTIC-STRUCTURE SYSTEMS ..... 3565
*William Martins Vicente, Renato Picelli, Renato Pavanello, Yi Min Xie*

BESO APPROACH TO TOPOLOGY OPTIMIZATION OF GAN PHONONIC CRYSTALS ..... 3583
*Luca D'Alessandro, Bichoy Bahr, Luca Daniel, Dana Weinstein, Raffaele Ardito*

**MS 1004: AERODYNAMIC STRATEGIES FOR THE GLOBAL OPTIMIZATION OF FLYING CONFIGURATIONS IN SUPERSONIC FLOW**
LOW BOOM / LOW DRAG SMALL SIZE SUPERSONIC AIRCRAFT DESIGN ..... 3596
*Atsushi Ueno, Yasushi Watanabe, Itham Salah El Din, Richard Grenon, Gerald Carrier*

DEVELOPMENT OF PHYSICAL METHODS OF THE SUPERSONIC AIRPLANE NEAR-FIELD INVESTIGATION AIMED AT THE SONIC BOOM MINIMIZATION ..... 3612
*Sergey Chernyshev, Alexander Ivanov, Andrey Kiselev, Vladimir Mosharov, Leonid Teperin*

GLOBAL OPTIMIZATION OF THE SHAPE OF AN AEROSPACE VEHICLE, VIA ITERATIVE OPTIMUM-OPTIMORUM STRATEGY ..... 3621
*A. Nastase*

**MS 1005: MONITORING AND CONTROL OF STRUCTURES**
SHAPE IDENTIFICATION ANALYSIS OF CAVITY IN RESIN STRUCTURE BASED ON THERMAL NONDESTRUCTIVE TESTING METHOD ..... 3630
*Kotaro Maruoka, Takahiko Kurahashi, Tetsuro Iyama*

DESIGN AND INSTALLATION OF A PERMANENT MONITORING SYSTEM FOR PALAZZO LOMBARDIA IN MILANO, ITALY ..... 3640
*Marta Berardengo, Alfredo Cigada, Stefano Manzoni, Marcello Vanali*

**MS 1006: PARAMETER IDENTIFICATION IN SOLID MECHANICS**
AN INVERSE PROBLEM STRATEGY BASED ON FORWARD MODEL EVALUATIONS: GRADIENT-BASED OPTIMIZATION WITHOUT ADJOINT SOLVES ..... 3652
*Miguel Aguilo*

COHESIVE ZONE MODEL IDENTIFICATION WITH DCB TEST PARAMETER ESTIMATION SENSITIVITY ..... 3676
*Racine Ly, Julien Jumel, Martin Shanahan, Florian Lavelle*

**MS 1007: ADDITIVE MANUFACTURING AND OPTIMIZATION**
TOPOLOGY OPTIMIZATION FOR ADDITIVE MANUFACTURING WITH CONTROLLABLE SUPPORT STRUCTURE COSTS ..... 3689
*Matthijs Langelaar*

STUDY OF TOPOLOGY OPTIMIZATION PARAMETERS AND SCAFFOLD STRUCTURES IN ADDITIVE MANUFACTURING ..... 3700
*Alain Garaigordobil, Ruben Ansola, Estrella Veguería*

FINITE ELEMENT APPROXIMATION OF A TIME-DEPENDENT TOPOLOGY OPTIMIZATION PROBLEM ..... 3711
*Matteo Bruggi, Nicola Parolini, Francesco Regazzoni, Marco Verani*

DESIGNING MANUFACTURABLE VISCOELASTIC DEVICES USING A TOPOLOGY OPTIMIZATION APPROACH WITHIN A TRULY-MIXED FEM FRAMEWORK ..... 3724
*Paolo Venini, Marco Pingaro, Carlo Cinquini*

ISO-XFEM FOR TOPOLOGY OPTIMIZATION OF STRUCTURES UNDER MULTIPLE LOAD CASES AND ACCELERATION LOADING ..... 3739
*Meisam Abdi, Ian Ashcroft, Ricky Wildman*

**MS 1008: ULTRASONIC GUIDED WAVES TESTING AND MONITORING**
LAMINATE ELEMENT METHOD AND ITS APPLICATION TO THE STUDY OF GUIDED WAVE RESONANCE PHENOMENA IN LAYERED ELASTIC STRUCTURES WITH DEFECTS ..... 3753
*Evgeny Glushkov, Natalia Glushkova, Artem Eremin, Rolf Lammering*

INVESTIGATIONS OF MECHANICAL GUIDED WAVES PROPAGATION IN PIPES REPAIRED LOCALLY BY COMPOSITE PATCHES ..... 3761
*Slah Yaacoubi, Weina Ke, F. Dahmene, M. El Mountassir*

**MS 1009: ADJOINT METHODS FOR STEADY & UNSTEADY OPTIMIZATION**

TWO-STEPS SHAPE OPTIMIZATION ALGORITHM IMPROVING HYDRODYNAMICS STABILITY ..... 3770
*Takashi Nakazawa*

ON THE CORRECT APPLICATION OF AD CHECKPOINTING TO ADJOINT MPI-PARALLEL PROGRAMS ..... 3778
*Ala Taftaf, Laurent Hascoët*

SPHERICITY: MESH OPTIMIZATION FOR ARBITRARY ELEMENT TOPOLOGY ..... 3797
*Pavlos Alexias, Eugene De Villiers*

LINKING PARAMETRIC CAD WITH ADJOINT SURFACE SENSITIVITIES ..... 3812
*Ilias Vasilopoulos, Dheeraj Agarwal, Marcus Meyer, Trevor T. Robinson, Cecil G. Armstrong*

A CODE-COUPLING APPROACH TO THE IMPLEMENTATION OF DISCRETE ADJOINT SOLVERS BASED ON AUTOMATIC DIFFERENTIATION ..... 3828
*Jan Backhaus, Anna Engels-Putzka, Christian Frey*

CAD KERNEL AND GRID GENERATION ALGORITHMIC DIFFERENTIATION FOR TURBOMACHINERY ADJOINT OPTIMIZATION ..... 3843
*Ismael Sanchez Torreguitart, Tom Verstraete, Lasse Mueller*

ALTERNATIVE SOLUTION ALGORITHMS FOR PRIMAL AND ADJOINT INCOMPRESSIBLE NAVIER-STOKES ..... 3858
*Mattia Oriani, Guillaume Pierrot*

DEFROSTER NOZZLE SHAPE OPTIMIZATION USING THE CONTINUOUS ADJOINT METHOD ..... 3883
*Lefki Germanou, Evangelos Papoutsis-Kiachagias, Antoine Delacroix, Kyriakos Giannakoglou*

SENSITIVITY ANALYSIS FOR FORCED RESPONSE IN TURBOMACHINERY USING AN ADJOINT HARMONIC BALANCE METHOD ..... 3893
*Anna Engels-Putzka, Christian Frey*

A TWO-LEVEL HYBRID APPROACH FOR OPTIMAL ACTIVE FLOW CONTROL ON A THREE-ELEMENT AIRFOIL ..... 3912
*Anil Nemili, Emre Özkaya, Nicolas R. Gauger, Felix Kramer, Frank Thiele*

TRANSITION FROM 2D CONTINUOUS ADJOINT LEVEL SET TOPOLOGY TO SHAPE OPTIMIZATION ..... 3928
*J.R.L. Koch, E.M. Papoutsis-Kiachagias, K.C. Giannakoglou*

AERODYNAMIC SHAPE OPTIMIZATION USING THE TRUNCATED NEWTON METHOD AND CONTINUOUS ADJOINT ..... 3942
*Mehdi Ghavami Nejad, Evangelos M. Papoutsis-Kiachagias, Kyriakos C. Giannakoglou*

UNSTEADY ADJOINT TO THE CUT-CELL METHOD USING MESH ADAPTATION ON GPU'S ..... 3957
*Konstantinos Samouchos, Stergios Katsanoulis, Kyriakos Giannakoglou*

SHAPE OPTIMIZATION OF TURBOMACHINERY ROWS USING A PARAMETRIC BLADE MODELLER AND THE CONTINUOUS ADJOINT METHOD RUNNING ON GPUS ..... 3972
*K.T. Tsiakas, F. Gagliardi, X.S. Trompoukis, K.C. Giannakoglou*

GEOMETRIC CONTINUITY CONSTRAINTS FOR ADJACENT NURBS PATCHES IN SHAPE OPTIMISATION ..... 3985
*Xingchen Zhang, Yang Wang, Mateusz Gugala, Jens-Dominik Mueller*

TOWARDS MULTIDISCIPLINARY ADJOINT OPTIMIZATION OF TURBOMACHINERY COMPONENTS ..... 3999
*Marc Schwalbach, Tom Verstraete*

TOWARDS AN OUTPUT-BASED RE-MESHING FOR TURBOMACHINERY APPLICATIONS ..... 4011
*Mateusz Gugala, Marcus Meyer, Jens-Dominik Mueller*

OPTIMISATION OF A U-BEND USING A CAD-BASED ADJOINT METHOD WITH DIFFERENTIATED CAD KERNEL ..... 4023
*Salvatore Auriemma, Mladen Banovic, Orest Mykhaskiv, Herve Legrand, Jens-Dominik Mueller, Tom Verstraete, Andrea Walther*

A CONTINUOUS ADJOINT APPROACH FOR VEHICLE INTERIOR NOISE REDUCTION ..... 4035
*Christos Kapellos, Michael Hartmann*

A MIXED OPERATOR OVERLOADING AND SOURCE TRANSFORMATION APPROACH FOR ADJOINT CFD COMPUTATION ..... 4047
*Zahrasadat Dastouri, Sinan Gezgin, Uwe Naumann*

**MS 1010: INVERSE PROBLEMS, DESIGN AND OPTIMIZATION**

EFFICIENT CALIBRATION OF DISCRETE ELEMENT MATERIAL MODEL PARAMETERS USING LATIN HYPERCUBE SAMPLING AND KRIGING ..... 4061
*Michael Rackl, Carolin D. Görnig, Kevin J. Hanley, Willibald A. Günthner*

UPDATING PRIOR PARAMETERS BASED ON LIKELIHOOD FUNCTION-BAYESIAN METHOD FOR PARAMETER ESTIMATION AT HIGH MEASUREMENT UNCERTAINTY ..... 4073
*Sankaran Somasundharam, K Srinivas Reddy*

**MS 1011: SURROGATE-ASSISTED EVOLUTIONARY ALGORITHMS IN AERODYNAMIC DESIGN/OPTIMIZATION**

CONSTRAINED MULTI-POINT AERODYNAMIC SHAPE OPTIMIZATION OF THE VISCOUS DPW WING THROUGH EVOLUTIONARY PROGRAMMING AND SUPPORT VECTOR MACHINES ..... 4082
*Daniel González Juárez, Esther Andrés Pérez, Mario Jaime Martin Burgos*

SURROGATE-BASED GLOBAL OPTIMIZATION OF A CYLINDER BY THE USE OF EVOLUTIONARY ALGORITHMS, SUPPORT VECTOR MACHINES AND NON UNIFORM B-SPLINES ..... 4092
*Eesther Andrés-Pérez, Daniel González-Juárez, Mario J. Martin-Burgos, Leopoldo Carro-Calvo, Sancho Salcedo-Sanz*

PERFORMANCE COMPARISON OF KRIGING AND SVR SURROGATE MODELS APPLIED TO THE OBJECTIVE FUNCTION PREDICTION WITHIN AERODYNAMIC SHAPE OPTIMIZATION ..... 4104
*Daniel Viúdez-Moreiras, Esther Andrés-Pérez, Daniel González-Juárez, Mario J. Martin Burgos*

HEAT TRANSFER OPTIMIZATION OF A RIBBED SURFACE USING SURROGATE-ASSISTED GENETIC ALGORITHMS ..... 4117
*Panagiotis Tsirikoglou, Ghader Ghorbaniasl, Simon Abraham, Chris Lacor*

MULTI-FIDELITY EXTENSION TO NON-INTRUSIVE PROPER ORTHOGONAL DECOMPOSITION BASED SURROGATES ..... 4129
*Tariq Benamara, Piotr Breitkopf, I. Lepot, Caroline Sainvitu*

EVOLUTIONARY AERODYNAMIC SHAPE OPTIMIZATION THROUGH THE RBF4AERO PLATFORM ..... 4146
*Dimitrios Kapsoulis, Varvara Asouti, Kyriakos Giannakoglou, Stefano Porziani, Emiliano Costa, Corrado Groth, Ubaldo Cella, Marco Evangelos Biancolini*

THE RBF4AERO BENCHMARK TECHNOLOGY PLATFORM ..... 4156
*Massimo Bernaschi, Alessandro Sabellico, Giorgio Urso, Emiliano Costa, Stefano Porziani, Fabrizio Lagasco, Corrado Groth, Ubaldo Cella, Marco Evangelos Biancolini, Dimitrios H.Kapsoulis, Varvara G. Asouti, Kyriakos C. Giannakoglou*

**MS 1013: SOLUTION OF LARGE-SCALE INVERSE PROBLEMS**
DATA ASSIMILATION USING MRI DATA ..... 4164
*Geir Nævdal, Ove Sævareid, Rolf J. Lorentzen*

**MS 1014: DESIGN OPTIMIZATION AND INVERSE PROBLEMS FOR WAVE PROPAGATION PROBLEMS**
FULL-WAVEFORM INVERSION FOR THE RECONSTRUCTION OF ELECTROMAGNETIC PROPERTY PROFILES USING PLANE MICROWAVES ..... 4178
*Namho Joh, Jun Won Kang*

**1100 REDUCTION METHODS**

**MS 1101: REDUCED BASIS, POD AND PGD MODEL ORDER REDUCTION TECHNIQUES**
REDUCED BASIS METHOD FOR POISSON-BOLTZMANN EQUATION ..... 4187
*Cleophas Kweyu, Martin Hess, Lihong Feng, Matthias Stein, Peter Benner*

AN ITERATIVE MODEL ORDER REDUCTION SCHEME FOR A SPECIAL CLASS OF BILINEAR DESCRIPTOR SYSTEMS APPEARING IN CONSTRAINT CIRCUIT SIMULATION ..... 4196
*Pawan Goyal, Peter Benner*

ELECTRICAL FIELDS SIMULATION IN HETEROGENOUS DOMAINS USING THE PROPER GENERALIZED DECOMPOSITION ..... 4213
*Chady Ghnatios, Francisco Chinesta, Anais Barasinski*

MODEL ORDER REDUCTION IN STRUCTURAL DYNAMICS ..... 4221
*Raul Rodriguez Sanchez, Martin Buchschmid, Gerhard Müller*

TOWARDS A NON-INTRUSIVE PROPER GENERALIZED DECOMPOSITION SCHEME FOR MODEL ORDER REDUCTION ..... 4248
*Xi Zou, Pedro Díez, Michele Conti, Ferdinando Auricchio*

A PGD-BASED TIME SPACE DECOMPOSITION FOR THE UNSTEADY NAVIER-STOKES EQUATIONS APPLIED TO INCOMPRESSIBLE FLOWS ..... 4263
*Michel Visonneau*

**MS 1103: MATHEMATICAL SURROGATE MODELLING IN ELECTROMAGNETICS**

RATIONAL MODELING OF MULTIVARIATE MULTI-FIDELITY DATA ..... 4284
*Elizabeth Rita Samuel, Dirk Deschrijver, Luc Knockaert, Tom Dhaene, Annie Cuyt*

ON THE USE OF S-PARAMETER TRANSFORMATIONS TO IMPROVE SURROGATE MODEL BEHAVIOUR OF MULTIPORT NETWORKS ..... 4295
*Petrie Meyer*

RECENT ADVANCES IN SURROGATE MODELLING OF REFLECTOR ANTENNA SYSTEMS ..... 4302
*Dirk De Villiers*

A SHORT-TIME PRONY METHOD FOR THE DETECTION OF TRANSIENTS ..... 4313
*Annie Cuyt, Wen-Shin Lee, Min-Nan Tsai*

SURROGATE MODELING OF ANTENNA RADIATION CHARACTERISTICS BY GAUSSIAN PROCESSES ..... 4322
*Jan Pieter Jacobs, Dirk De Villiers*

ACCELERATION OF MESH-BASED PHYSICAL OPTICS FOR ELECTROMAGNETIC SCATTERING ANALYSIS ..... 4330
*Dao P. Xiang, Matthys M. Botha*

DESIGN OF AXIALLY SYMMETRIC POWER COMBINERS USING SURROGATE BASED OPTIMIZATION ..... 4339
*Ryno Beyers, Dirk De Villiers*

**1200 STRUCTURAL STABILITY AND DYNAMICS**

**MS 1201: COMPUTATIONAL STRUCTURAL DYNAMICS**

FLEXURAL VIBRATION ANALYSIS OF GRAPHENE NANOPLATELETS REINFORCED NANOCOMPOSITE BEAMS ..... 4351
*Jie Yang, Chuang Feng*

DYNAMIC RESPONSE OF REAL OFFSHORE WIND TURBINES ON MONOPILES IN STRATIFIED SEABED ..... 4361
*Guillermo M. Álamo, Juan J. Aznárez, Luis A. Padrón, Alejandro E. Martínez-Castro, Rafael Gallego, Orlando Maeso*

DYNAMIC ANALYSIS OF SPECIAL CARS ON UNEVEN ROADS ..... 4377
*Szymon Tengler*

STRUCTURED SPATIAL DISCRETIZATION OF DYNAMICAL SYSTEMS ..... 4398
*Marko Jokic, Andrej Jokic, Bruno Dogancic*

WAVE PROPAGATION IN THIN PRETWISTED ANISOTROPIC STRIPS ..... 4406
*Maloth Thirupathi, Mira Mitra, P J Guruprasad*

NUMERICAL ROUTINE FOR DYNAMIC ANALYSIS OF TRANSMISSION LINES GUYED TOWERS SUBMITTED TO BROKEN CABLE ..... 4416
*Thiago Brazeiro Carlos, João Kaminski Jr.*

INVARIANCE OF EIGENFREQUENCIES AND EIGENMODES UNDER GEOMETRIC TRANSFORMATION IN ELONGATED ELASTIC STRUCTURES ..... 4436
*Maryam Morvaridi, Michele Brun*

**MS 1202: ADVANCED BEAM MODELS**

A COMPUTATIONALLY EFFECTIVE FORMULATION TO FINITE ROTATIONS - SMALL STRAINS DESCRIPTION OF BEAM ELEMENTS ..... 4445
*Salvatore Lopez*

A COROTATIONAL FINITE ELEMENT TO MODEL BENDING VIBRATIONS OF METALLIC STRANDS ..... 4455
*Francesco Foti*

3D BEAM-COLUMN FINITE ELEMENT UNDER NON-UNIFORM SHEAR STRESS DISTRIBUTION DUE TO SHEAR AND TORSION ..... 4467
*Paolo Di Re, Daniela Addessi, Filip C. Filippou*

SINGULARLY PERTURBED PROBLEMS IN MECHANICS (SOME FUNDAMENTAL ASPECTS OF DESIGNING AND COMPUTING) ..... 4481
*Lyudmila K. Kuzmina*

A DEGENERATE-CONTINUUM BASED TIMOSHENKO BEAM APPROACH FOR THE AEROELASTIC ANALYSIS OF THE WIND TURBINE BLADES ..... 4488
*Anthoula Panteli, Dimitris Manolas, Konstantinos Spiliopoulos*

DYNAMICS OF WIND TURBINE BLADES USING A GEOMETRICALLY-EXACT BEAM FORMULATION ..... 4503
*Celso Faccio Junior, Alfredo Gay Neto*

EXTENDED FORMULA FOR THE CRITICAL VELOCITY OF A LOAD MOVING ON A BEAM SUPPORTED BY A FINITE DEPTH FOUNDATION ..... 4520
*Zuzana Dimitrovová*

FIRST-ORDER DISPLACEMENT-BASED ZIGZAG THEORIES FOR COMPOSITE LAMINATES AND SANDWICH STRUCTURES: A REVIEW ..... 4528
*Marco Di Sciuva*

**MS 1203: THE MODELS AND INVESTIGATIONS METHODS OF DYNAMICS OF THE SOLIDS SYSTEMS WITH DRY FRICTION**

VIRTUAL TESTBENCH FOR THE OMNI WHEEL DYNAMICS SIMULATION: NEW CONTACT TRACKING ALGORITHM ..... 4553
*Ivan Kosenko, Sergey Stepanov, Kirill Gerasimov, Mikhail Stavrovskiy*

MODEL AND INVESTIGATION OF DYNAMICS OF SOLID SYSTEM WITH TWO MASSIVE ECCENTRICS ON A ROUGH PLANE ..... 4572
*Sergey Semendyaev, Alexey Tsyganov*

NON-REGULAR VEHICLE DYNAMICS. APPLICATION TO COLLISION ..... 4584
*Ahmed Bouzar Essaidi, Bachir Menkouz, Moussa Haddad, Taha Chettibi*

TESTING, SIMULATING AND UNDERSTANDING UNDER-PLATFORM DAMPER DYNAMICS ..... 4599
*Chiara Gastaldi, Muzio M. Gola*

**MS 1204: NONLINEAR DYNAMICS OF ROTATING STRUCTURES**
A NEW DYNAMIC SUBSTRUCTURING METHOD FOR NONLINEAR AND DISSIPATIVE SYSTEMS ..... 4611
*Colas Joannin, Benjamin Chouvion, Fabrice Thouverez*

SOME REMARKS ON TIME INTEGRATION OF 3D ROTOR-STATOR ASSEMBLY ..... 4624
*Benoit Prabel*

DELAYED FEEDBACK CONTROL METHOD FOR CALCULATING SPACE-TIME PERIODIC SOLUTIONS OF VISCOELASTIC PROBLEMS ..... 4638
*Ustim Khristenko, Patrick Le Tallec*

COMBINING FINITE ELEMENT ANALYSIS AND ANALYTICAL MODELLING FOR EFFICIENT SIMULATIONS OF NON-LINEAR GEAR DYNAMICS ..... 4655
*Shadi Sweiki, Jakub Korta, Antonio Palermo, Rocco Adduci, Domenico Mundo*

DAMPING INDUCED BY DRY FRICTION: ANALYSES AND EXPERIMENTS FOR MODELING IMPROVEMENT ..... 4664
*Marc-André Douville, Béatrice Faverjon, Eric Chatelet, Georges Jacquet-Richardet*

**MS 1206: ADVANCES IN NUMERICAL METHODS FOR LINEAR AND NON-LINEAR DYNAMICS AND WAVE PROPAGATION**
RAYLEIGH METHOD APPLIED TO A 46-M-HIGH CONCRETE MAST ..... 4674
*Alexandre De M. Wahrhaftig, Reyolando M. L. R. F. Brasil*

A DISPERSION MINIMIZED MIMETIC METHOD FOR A COLD PLASMA MODEL ..... 4682
*Vrushali A. Bokil, Vitaliy Gyrya, Duncan A. Mcgregor*

A VOLUME-AGGLOMERATION MULTIRATE TIME ADVANCING APPROACH ..... 4703
*Emmanuelle Itam, Stephen Wornom, Bruno Koobus, Alain Dervieux*

THERMOMECHANICAL NUMERICAL SIMULATION OF IMPACTS ON ELASTIC-PLASTIC SOLIDS WITH THE FINITE VOLUME METHOD ..... 4717
*Thomas Heuzé*

**MS 1207: COMPUTATIONAL SIMULATION OF SMART STRUCTURES AND MATERIALS**
TRANSVERSE WAVE PROPAGATION IN A ONE-DIMENSIONAL STRUCTURE COUPLED TO ITS ELECTRICAL ANALOGUE: COMPARISON OF TRANSFER MATRIX MODELS ..... 4728
*Boris Lossouarn, Mathieu Aucejo, Jean-François Deü*

**MS 1208: BIFURCATIONS AND STABILITY**
BUCKLING OF ANNULAR PLATE JOINT WITH CIRCULAR BEAM ..... 4741
*Sergei Filippov, Maria Boyarskaya*

THE CURL AND FLUTING OF PAPER: THE EFFECT OF ELASTO-PLASTICITY ..... 4752
*Anna-Leena Erkkilä, Teemu Leppänen, Tero Tuovinen*

STABILITY OF STEEL STRUCTURES WITH CLEARANCES AND IMPERFECTION ..... 4770
*Katarzyna Rzeszut, Andrzej Garstecki*

**MS 1209: DYNAMIC ANALYSIS OF BEAMS UNDER MOVING VEHICLES: APPLICATION TO RAILWAY TRACK MODELLING, DESIGN AND REHABILITATION**
NUMERICAL ANALYSIS FOR THE DYNAMIC RESPONSE CHARACTERISTICS OF THE PRESTRESSED CONCRETE SLEEPER ..... 4780
*Tsutomu Watanabe, Kodai Matsuoka, Shintaro Minoura*

NUMERICAL DYNAMIC ANALYSIS OF BEAMS ON NONLINEAR ELASTIC FOUNDATIONS UNDER HARMONIC MOVING LOAD ..... 4794
*Diego Froio, Roberto Moioli, Egidio Rizzi*

**MS 1210: ADVANCES IN MODELING AND ANALYSIS OF FGM STRUCTURES**
ELASTOSTATIC AND MODAL AND BUCKLING ANALYSIS OF SPATIAL FGM BEAM STRUCTURES ..... 4810
*Justin Murin, Mehdi Aminbaghai, Juraj Hrabovsky, Vladimir Kutis, Juraj Paulech, Stephan Kugler*

THERMO-ELASTICITY IN SHELL STRUCTURES MADE OF FUNCTIONALLY GRADED MATERIALS ..... 4838
*Stephan Kugler, Peter Fotiu, Justin Murin*

FINITE BEAM ELEMENT WITH PIEZOELECTRIC LAYERS AND FUNCTIONALLY GRADED MATERIAL OF CORE ..... 4863
*Vladimír Kutiš, Justín Murín, Juraj Paulech, Juraj Hrabovský, Roman Gogola, Jakub Jakubec*

NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS WITH NON-CONSTANT COEFFICIENTS ..... 4878
*Juraj Hrabovský, Justín Murín, Mehdi Aminbaghai, Vladimír Kutiš, Juraj Paulech*

TWO-WAY COUPLED ELECTRO-THERMAL ANALYSIS OF FGM SYSTEM CALCULATED BY THE NEW LINK FINITE ELEMENT ..... 4888
*Juraj Paulech, Vladimír Kutiš, Juraj Hrabovský, Justín Murín, Roman Gogola, Tibor Sedlár, Gabriel Gálik*

**MS 1211: COMPUTATIONAL STRATEGIES FOR STRUCTURAL ROBUSTNESS ASSESSMENT**
PROGRESSIVE COLLAPSE FRAGILITY MODELS OF RC FRAMED BUILDINGS BASED ON PUSHDOWN ANALYSIS ..... 4900
*Emanuele Brunesi, Roberto Nascimbene, Fulvio Parisi*

MECHANICAL PROPERTIES OF S355 UNDER EXTREME COUPLED EFFECT OF HIGH TEMPERATURES AND HIGH STRAIN RATES ..... 4920
*Daniele Forni, Bernardino Chiaia, Ezio Cadoni*

IMPLEMENTATION OF PROGRESSIVE DAMAGE IN FINITE ELEMENT CODES FOR THE ASSESSMENT OF ROBUSNTESS ..... 4928
*Domenico Asprone, Bernardino Chiaia, Valerio De Biagi, Gaetano Manfredi, Fulvio Parisi*

SEISMIC RESIDUAL CAPACITY ASSESSMENT OF FRAMED STRUCTURES DAMAGED BY EXCEPTIONAL ACTIONS ..... 4942
*Antonio Formisano, Giuseppe Iazzetta, Giuseppe Marino, Francesco Fabbrocino, Raffaele Landolfo*

**MS 1212: DYNAMICS AND SEISMIC RESPONSE OF ROCKING AND SELF-CENTERING STRUCTURES**
ROCKING RESPONSE OF MASONRY BLOCK STRUCTURES USING MATHEMATICAL PROGRAMMING ..... 4959
*Francesco Portioli, Lucrezia Cascini, Raffaele Landolfo*

CONTROLLED ROCKING, DISSIPATIVE CONTROLLED ROCKING AND MULTI-HIERARCHICAL ACTIVATION: NUMERICAL ANALYSIS AND EXPERIMENTAL TESTING ..... 4969
*Royce Liu, Alessandro Palermo*

THE ROLE OF THE PRESTRESSED TENDONS ON THE SEISMIC PERFORMANCE OF HYBRID ROCKING BRIDGE BENTS ..... 4983
*Anastasios I. Giouvanidis, Elias G. Dimitrakopoulos*

MODELLING CONTACT IN ROCKING STRUCTURES WITH A NONSMOOTH DYNAMICS APPROACH ..... 4997
*Anastasios I. Giouvanidis, Elias G. Dimitrakopoulos*

A SIMPLE ANALYTICAL MODEL FOR THE ROCKING PREWEC SYSTEM ..... 5010
*Dimitrios Kalliontzis, Sri Sritharan*

AN ANALYTICAL MODEL FOR DYNAMIC RESPONSE OF AN ELASTIC SDOF SYSTEM FIXED ON TOP OF A ROCKING SINGLE-STORY FRAME STRUCTURE: EXPERIMENTAL VALIDATION ..... 5023
*Jonas A. Bachmann, Christoph Jost, Quentin Studemann, Michalis F. Vassiliou, Bozidar Stojadinovic*

SMOOTH-ROCKING OSCILLATOR UNDER NATURAL ACCELEROGRAMS ..... 5063
*Blerta Lipo, Gianmarco De Felice*

EXPERIMENTAL INVESTIGATION OF THE SEISMIC RESPONSE OF A COLUMN ROCKING AND ROLLING ON A CONCAVE BASE ..... 5074
*Jonas A. Bachmann, Patrick Blöchlinger, Matthias Wellauer, Michalis F. Vassiliou, Bozidar Stojadinovic*

SEISMIC ASSESSMENT OF MASONRY ROCKING COLUMNS AND FRAMES UNDER GROUND MOTION EXCITATIONS ..... 5098
*Ioannis Kavvadias, Lazaros Vasiliadis*

PARAMETRIC INVESTIGATION OF THE DYNAMIC RESPONSE OF RIGID BLOCKS SUBJECTED TO SYNTHETIC NEAR-SOURCE GROUND MOTION RECORDS ..... 5113
*Michalis Fragiadakis, Ioannis Psycharis, Yenan Cao, George P. Mavroeidis*

A MACRO-ELEMENT FORMULATION FOR ROCKING FLEXIBLE BODIES WITH A DEFORMABLE BASE ..... 5129
*Evangelos Avgenakis, Ioannis N. Psycharis*

**MS 1213: INNOVATIVE STRUCTURAL SYSTEMS FOR SEISMIC RESISTANT BUILDINGS**

NONLINEAR ANALYSIS AND EXPERIMENTAL BEHAVIOUR OF AN INNOVATIVE STEEL FRAME WITH REINFORCED CONCRETE INFILL WALLS ..... 5142
*Andrea Dall'Asta, Graziano Leoni, Francesco Morelli, Walter Salvatore, Alessandro Zona*

MODAL PROPERTIES AND SEISMIC RESPONSE OF EXISTING BUILDING RETROFITTED BY EXTERNAL BRACINGS WITH VISCOUS DAMPERS ..... 5156
*Laura Gioiella, Enrico Tubaldi, Fabrizio Gara, Andrea Dall'Asta*

**MS 1214: HISTORIC MASONRY STRUCTURES: MODELLING, ASSESSMENT & RETROFIT**

SIMPLE CLOSED FORM HOMOGENIZATION MODEL FOR THE NON LINEAR STATIC AND DYNAMIC ANALYSIS OF RUNNING BOND MASONRY WALLS IN AND OUT OF PLANE LOADED ..... 5172
*Elisa Bertolesi, Gabriele Milani*

CONSTRUCTION PHASES ANALYSIS OF UNREINFORCED MASONRY BUILDINGS THROUGH EQUIVALENT FRAME MODEL ..... 5190
*Francesco Pugi, Alessio Francioso, Giacomo Sevieri*

A COMBINED EXPERIMENTAL AND NUMERICAL STUDY OF THE PULL-OUT MECHANISM OF THREADED TITANIUM BARS EMBEDDED IN MARBLE BLOCKS ..... 5207
*Stavros K. Kourkoulis, Aikaterini Marinelli, Ioanna Dakanali*

FINITE-DISCRETE ELEMENT MODELLING OF MASONRY INFILL WALLS SUBJECTED TO OUT-OF-PLANE LOADS ..... 5219
*Laura Liberatore, Marta Bruno, Omar Al Shawa, Monica Pasca, Luigi Sorrentino*

PALAZZO LA SAPIENZA IN PISA: STRUCTURAL ASSESSMENT AND RETROFIT OF AN HISTORICAL MASONRY BUILDING IN ITALY ..... 5230
*Silvia Caprili, Federico Mangini, Nicola Mussini, Walter Salvatore*

THREE-DIMENSIONAL NONLINEAR BEHAVIOUR OF MASONRY WALLS MODELLED WITH DISCRETE ELEMENTS ..... 5248
*Daniele Baraldi, Antonella Cecchi*

NUMERICAL VALIDATION OF EQUIVALENT-FRAME MODELS FOR URM WALLS ..... 5262
*Rossella Siano, Guido Camata, Vincenzo Sepe, Enrico Spacone, Pere Roca, Luca Pelà*

SEISMIC ASSESSMENT OF HISTORICAL MASONRY TOWERS IN THE NORTH-EAST REGION OF ITALY ..... 5274
*Marco Valente, Gabriele Milani*

NATIVITY CHURCH IN BETHLEHEM: FULL 3D NON-LINEAR FE APPROACH FOR STRUCTURAL DAMAGE PREDICTION ..... 5287
*Gabriele Milani, Marco Valente, Claudio Alessandri*

DYNAMIC BEHAVIOUR AND EARTHQUAKE PERFORMANCE OF GREEK BASILICA CHURCHES WITH FOUNDATION DEFORMABILITY ..... 5301
*George Manos, Evaggelos Kozikopoulos, Lambros Kotoulas*

PROBLEMS RELATED TO THE USE OF FIBER REINFORCED CEMENTITIOUS MATERIALS AS STRENGTHENING OF MASONRY MEMBERS ..... 5314
*Daniela Sinicropi, Antonio Borri, Marco Corradi, Michele Paradiso*

FAST KINEMATIC LIMIT ANALYSIS OF FRP REINFORCED MASONRY VAULTS THROUGH A NEW GENETIC ALGORITHM NURBS-BASED APPROACH ..... 5326
*Andrea Chiozzi, Gabriele Milani, Antonio Tralli*

THE ROLE OF RESTORATION MORTARS IN THE EARTHQUAKE PROTECTION OF KAISARIANI MONASTERY ..... 5340
*Antonia Moropoulou, Maria Apostolopoulou, Petros Moundoulas, Eleni Aggelakopoulou, Louiza Siouta, Asterios Bakolas, Maria Douvika, Panagiotis Karakitsios, Panagiotis G. Asteris*

FIRST RESULTS OF THE VIBRATION-BASED STRUCTURAL HEALTH MONITORING OF A MASONRY DOME ..... 5359
*Nicola Cavalagli, Gabriele Comanducci, Massimiliano Gioffrè, Vittorio Gusella, Filippo Ubertini*

BEHAVIOR OF SINGLE STORY BEARING WALL MASONRY STRUCTURES IN VARIOUS CONDITIONS ..... 5373
*Abdelraouf Kassem*

SEISMIC PROTECTION OF MONUMENTS AND HISTORIC STRUCTURES – THE SEISMO RESEARCH PROJECT ..... 5382
*Constantine Spyrakos, Charilaos Maniatakis*

**MS 1215: NONLINEAR VIBRATIONS OF CONSERVATIVE AND NONCONSERVATIVE SYSTEMS: PHENOMENA AND ADVANCED NUMERICAL METHODS**

FORCED RESPONSE OF SHROUDED BLADES WITH VARIABLE OPERATING POINTS ..... 5396
*Ferhat Kaptan, Lars Panning-Von Scheidt, Jörg Wallaschek, Victor Salit*

GEOMETRY EFFECTS ON THE NONLINEAR OSCILLATIONS OF VISCOELASTIC CYLINDRICAL SHELLS ..... 5406
*Zenon Del Prado, Marco Amabili, Paulo Gonçalves, Frederico Da Silva*

REDUCED ORDER MODELS FOR DYNAMIC BEHAVIOR OF PRESTRESSED ELASTOMER DAMPING DEVICES ..... 5417
*Antoine Legay, Jean-François Deü, Benjamin Morin*

LOCALIZATION OF VIBRATORY ENERGY OF A LINEAR SYSTEM IN A CHAIN OF FOUR NONLINEAR OSCILLATORS ..... 5427
*Simon Charlemagne, Claude-Henri Lamarque, Alireza Ture Savadkoohi*

CONTINUATION OF QUASI-PERIODIC SOLUTIONS WITH TWO-FREQUENCY HARMONIC BALANCE METHOD .. 5438
*Louis Guillot, Pierre Vigué, Christophe Vergez, Bruno Cochelin*

**MS 1216: STRUCTURAL ANALYSIS AND VIBRATIONS**

ASYMPTOTIC ANALYSIS OF DEFORMATIONS OF THE SLIGHTLY ORTHOTROPIC SPHERICAL LAYER UNDER NORMAL PRESSURE ..... 5444
*Andrei L. Smirnov, Svetlana M. Bauer, Liudmila A. Venatovskaia, Eva B. Voronkova*

EXPERIMENTAL VALIDATION OF FINITE ELEMENT MODELS FOR REINFORCED CONCRETE BEAMS WITH AND WITHOUT DISCONTINUITIES ..... 5454
*Marios Filippoupolitis, Carl Hopkins, Siu-Kui Au*

**MS 1217: COMPUTATIONAL METHODS IN EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS**

SEISMIC BEHAVIOR INDICES OF OLD TYPE REINFORCED CONCRETE MEMBERS ..... 5464
*Anthos Ioannou, Stavroula J. Pantazopoulou*

EFFECT OF SOIL SPATIAL VARIABILITY ON THE DYNAMIC BEHAVIOR OF A SLOPE ..... 5483
*Michael Michael, Tamara Al-Bittar, Abdul-Hamid Soubra*

ENERGY-MOMENTUM METHOD FOR NONLINEAR DYNAMIC OF 2D COROTATIONAL BEAMS ..... 5496
*Sophy Chhang, Mohammed Hjiaj, Jean-Marc Battini, Carlo Sansour*

THE INFLUENCE OF UNCERTAINTY OF DESIGN PARAMETERS ON DYNAMIC CHARACTERISTICS OF STRUCTURE WITH DAMPERS ..... 5507
*Magdalena Lasecka-Plura, Roman Lewandowski*

FINITE-DISCRETE NUMERICAL MODELLING OF REINFORCED CONCRETE STRUCTURES ..... 5523
*Nikolina Zivaljic, Zeljana Nikolic, Hrvoje Smoljanovic, Ivan Balic*

BEAM DYNAMIC STRESSES INCREMENTS AFTER PARTIAL DECONSTRUCTION OF FOUNDATION ..... 5533
*Vladimir Gordon, Olga Pilipenko, Timur Gasimov*

RELIABILITY OF SYSTEMS EQUIPPED WITH VISCOUS DAMPERS WITH VARIABLE PROPERTIES ..... 5550
*Andrea Dall'Asta, Laura Ragni, Fabrizio Scozzese, Enrico Tubaldi*

AN INVESTIGATION ON EFFECT OF SUBGRADE REACTION MODULUS OF SOIL ON NATURAL PERIOD OF LOW RISE BUILDINGS ..... 5567
*Rafet Sisman, Abdurrahman Sahin*

A PROBABILISTIC APPROACH TOWARDS AN EVALUATION OF EXISTING CODE PROVISIONS FOR SEISMICALLY ISOLATED STRUCTURES ..... 5574
*Anastasios Tsiavos, Bozidar Stojadinovic*

EFFECT OF THE SEISMIC EXCITATION'S INCIDENCE ANGLE ON THE NONLINEAR BEHAVIOR OF BASE ISOLATED BUILDINGS CONSIDERING POUNDING TO ADJACENT MOAT WALLS ..... 5585
*Eftychia A. Mavronicola, Panayiotis C. Polycarpou, Petros Komodromos*

NUMERICAL NON-LINEAR SIMULATION OF THE IN-PLANE BEHAVIOUR OF R/C FRAMES WITH MASONRY INFILLS UNDER SEISMIC TYPE LOADING ..... 5598
*George Manos, Vasilios Soulis*

MODELING THE MECHANICAL RESPONSE OF A LEAD-CORE BEARING DEVICE: DAMAGE MECHANICS APPROACH ..... 5613
*Todor Zhelyazov, Rajesh Rupakhety, Simon Olafsson*

INVESTIGATION OF THE PERFORMANCE OF VARIOUS MODAL CORRELATION CRITERIA IN STRUCTURAL DAMAGE IDENTIFICATION ..... 5626
*Manolis Georgioudakis, Vagelis Plevris*

**MS 1218: STABILITY AND CONTROL OF FLEXIBLE STRUCTURES**

EXPLOITING MULTIPLE REFERENCE MODELS FOR ADAPTIVE CONTROL OF FLEXIBLE STRUCTURES ..... 5646
*Ilaria Venanzi, Laura Ierimonti*

PASSIVE CONTROL OF WAVE PROPAGATION IN PERIODIC ANTI-TETRACHIRAL META-MATERIALS ..... 5658
*Marco Lepidi, Andrea Bacigalupo*

FLEXIBLE COMPLEX SYSTEM OF A DOUBLE-STRING UNDER EXTREME MOVING LOADS ..... 5670
*Jarosław Rusin*

ON THE MODELING OF SELF-DAMPING IN STRANDED CABLES ..... 5680
*Francesco Foti, Luca Martinelli, Federico Perotti*

COMPUTATIONAL EFFECTIVENESS OF LMI DESIGN STRATEGIES FOR VIBRATION CONTROL OF LARGE STRUCTURES ..... 5689
*Francisco Palacios-Quiñonero, Josep Rubió-Massegú, Josep M. Rossell, Hamid Reza Karimi*

A SMART BASE RESTRAINT FOR WIND TURBINES TO MITIGATE UNDESIRED EFFECTS DUE TO STRUCTURAL VIBRATIONS ..... 5702
*Nicola Caterino, Christos T. Georgakis, Mariacristina Spizzuoco, Antonio Occhiuzzi*

**MS 1220: DYNAMICS OF STRUCTURES SUBJECT TO SEISMIC EXCITATIONS**

NONLINEAR DYNAMIC BEHAVIOUR OF BASE-ISOLATED BUILDINGS WITH THE FRICTION PENDULUM SYSTEM SUBJECTED TO NEAR-FAULT EARTHQUAKES ..... 5712
*Fabio Mazza, Sandro Sisinno*

ABSOLUTE FINITE ELEMENT COORDINATES IN THE DYNAMICS OF LARGE FLEXIBLE STRUCTURES ..... 5726
*Evtim Zahariev*

MODAL ANALYSIS OF A FRAME MODEL UNDER UNMEASURED SEISMIC INPUT ..... 5742
*Claudio Valente, Vincenzo Sepe, Marco Di Pilla, Fabrizio Iezzi, Rossella Siano, Luigia Zuccarino*

EQUIVALENT UNIAXIAL ACCELEROGRAM FOR CSS-BASED ISOLATION SYSTEMS ASSESSMENT UNDER TWO-COMPONENTS SEISMIC EVENTS ..... 5753
*Marco Furinghetti, Alberto Pavese*

**MS 1221: COMPUTATIONAL STRUCTURAL STABILITY**

MIXED SOLID MODELS IN NUMERICAL ANALYSIS OF SLENDER STRUCTURES ..... 5768
*Domenico Magisano, Leonardo Leonetti, Giovanni Garcea*

COMPUTATIONAL ANALYSIS OF THE COLLAPSE BEHAVIOUR OF THIN-WALLED POLYGONAL STEEL BEAMS .... 5789
*Raffaele Ardito*

INFLUENCE OF STRAIN DEFINITIONS ON TRUSSES CRITICAL LOADS ..... 5799
*Reyolando Brasil, Jose Balthazar*

ULTIMATE STRENGTH ANALYSIS OF STIFFENED PANELS OF SHIP STRUCTURES UNDER COMBINED LOAD ..... 5810
*Li Hong, Meng Linghua, Qin Zhongwen, Zhang Enguo, Li Li*

**MS 1222: INFLUENCE OF LIQUEFIABLE SOIL ON SINGLE AND CLOSELY CLUSTERED STRUCTURES**
NUMERICAL ANALYSES OF THE INFLUENCE OF STRUCTURAL SLENDERNESS ON THE SEISMIC RESPONSE OF SINGLE AND CLUSTERED STRUCTURES ..... 5817
*Gonzalo Barrios, Tam Larkin, Nawawi Chouw*

EFFECT OF HIGHER MODES ON STRUCTURAL RESPONSE WITH NONLINEAR SOIL-FOUNDATION-STRUCTURE INTERACTION ..... 5826
*Xiaoyang Qin, Nawawi Chouw, Tam Larkin*

**MS 1224: INNOVATIVE SOLUTIONS FOR THE SEISMIC PROTECTION OF INDUSTRIAL BUILDINGS**
SEISMIC RETROFIT OF AN INDUSTRIAL STRUCTURE THROUGH AN INNOVATIVE SELF-CENTERING HYSTERETIC DAMPER: MODELLING, ANALYSIS AND OPTIMIZATION. ..... 5834
*Francesco Morelli, Andrea Piscini, Walter Salvatore*

PERFORMANCE ASSESSMENT OF SEISMIC RETROFITTING MEASURES ON SILO STRUCTURES USING INNOVATIVE SEISMIC PROTECTION SYSTEMS ..... 5851
*Marius Pinkawa, Benno Hoffmeister, Markus Feldmann*

SEISMIC RETROFIT OF INDUSTRIAL SILOS BY MEANS OF BASE ISOLATION DEVICES ..... 5868
*Alper Kanyilmaz, Carlo Andrea Castiglioni, Julia Georgi*

EFFICIENCY OF SEISMIC ISOLATION ON INDUSTRIAL PLANTS - CASE STUDY OF A GAS TANK ..... 5896
*José Henriques, Francesco Morelli, Bram Vandoren, Walter Salvatore, Hervé Degée*

PERFORMANCE-BASED NONLINEAR RESPONSE HISTORY ANALYSIS FRAMEWORK FOR THE "PROINDUSTRY" PROJECT CASE STUDIES ..... 5912
*Marco Faggella, Raffaele Laguardia, Rosario Gigliotti, Francesco Morelli, Franco Braga, Walter Salvatore*

PERFORMANCE BASED EARTHQUAKE ASSESSMENT OF AN INDUSTRIAL SILOS STRUCTURE AND RETROFIT WITH SLIDING ISOLATORS ..... 5926
*Edoardo Rossi, Michelangelo Ventrella, Marco Faggella, Rosario Gigliotti, Franco Braga*

**MS 1225: SEISMIC PERFORMANCE ASSESSMENT OF STRUCTURES AND SEISMIC RISK MITIGATION STRATEGIES**
A CRITICAL REVIEW OF FRAGILITY CURVES FOR EXISTING RC BUILDINGS ..... 5934
*Monica Mastroberti, Marco Vona*

DYNAMIC IDENTIFICATION AND MODELLING OF A RC HOSPITAL BUILDING ..... 5942
*Giorgio Lacanna, Pauline Deguy, Maurizio Ripepe, Massimo Baglione, Marco Tanganelli, Stefania Viti, Mario De Stefano*

SEISMIC HAZARD OF URBAN AREAS: A CASE-STUDY ..... 5952
*Giorgio Lacanna, Pauline Deguy, Maurizio Ripepe, Massimo Coli, Barbara Paoletti, Sara Barducci, Marco Tanganelli, Stefania Viti, Mario De Stefano*

SPO2FRAG V1.0: SOFTWARE FOR PUSHOVER-BASED DERIVATION OF SEISMIC FRAGILITY CURVES ..... 5962
*Iunio Iervolino, Georgios Baltzopoulos, Dimitrios Vamvatsikos, Roberto Baraschino*

COMPUTER-AIDED SEISMIC RISK ASSESSMENT AT URBAN SCALE. MODEL DEFINITION AND VALIDATION ON A CASE STUDY ..... 5977
*Alberto Basaglia, Alessandra Aprile, Francesco Pilla, Enrico Spacone*

TOWARDS INTEGRATED SEISMIC RISK ASSESSMENT IN PALESTINE - APPLICATION TO THE CITY OF NABLUS ... 5987
*Ricardo Monteiro, Paola Ceresa, Vania Cerchiello, Jamal Dabeek, Antonella Di Meo, Barbara Borzi*

REASSESS V1.0: A COMPUTATIONALLY-EFFICIENT SOFTWARE FOR PROBABILISTIC SEISMIC HAZARD ANALYSIS ..... 5999
*Iunio Iervolino, Eugenio Chioccarelli, Pasquale Cito*

SIMPLIFIED PERIOD ESTIMATION OF ITALIAN RC BRIDGES FOR LARGE-SCALE SEISMIC ASSESSMENT ..... 6013
*C. Zelaschi, R. Monteiro, R. Pinho*

**1300 UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION**

**MS 1301: THE STOCHASTIC COMPUTER METHODS IN MECHANICS**
FLOW FIELD ESTIMATION IN OPEN CHANNEL BASED ON KALMAN FILTER FINITE ELEMENT METHOD ..... 6029
*Taichi Yoshiara, Takahiko Kurahashi, Yasuhide Kobayashi, Toshihiko Eto*

STOCHASTIC TECHNIQUES FOR THE NUMERICAL SOLUTION OF ENGINEERING BOUNDARY VALUE PROBLEMS ..... 6041
*Victor Maceiras, Manuel Casteleiro*

THE BOOTSTRAP APPROACH TO THE STATISTICAL SIGNIFICANCE OF PARAMETERS IN THE FIXED EFFECTS MODEL ..... 6061
*Jacek Pietraszek, Renata Dwornicka, Agnieszka Szczotok*

ON FULLY COUPLED THERMO-ELASTO-PLASTIC STOCHASTIC FINITE ELEMENT ANALYSIS OF STEEL STRUCTURES ..... 6069
*Marcin Kaminski, Michał Strakowski*

AN INTRODUCTION TO STOCHASTIC FINITE ELEMENT METHOD ANALYSIS OF HYPERELASTIC STRUCTURES ... 6078
*Marcin Kamiński, Damian Sokołowski*

**MS 1303: ANALYSIS AND DESIGN OF SAFETY CRITICAL SYSTEMS UNDER UNCERTAINTY**
IDENTIFICATION OF INTERVAL FIELDS FOR SPATIAL UNCERTAINTY REPRESENTATION IN FINITE ELEMENT MODELS ..... 6091
*Matthias Faes, Jasper Cerneels, Dirk Vandepitte, David Moens*

INTRODUCTION OF BEAD INGS INTO A CRASH TUBE USING A ROBUST OPTIMISATION APPROACH ..... 6099
*Stefan Scheiblhofer*

DYNAMICAL ANALYSIS OF PLATE MODELS WITH UNCERTAIN STRUCTURAL PROPERTIES USING THE INTERVAL FIELD METHOD ..... 6107
*Maurice Imholz, Dirk Vandepitte, David Moens*

POWER GRID ROBUSTENSS TO SEVERE FAILURES: TOPOLGICAL AND FLOW BASED METRICS COMPARISION ..... 6121
*Roberto Rocchetta, Edoardo Patelli*

**MS 1304: STOCHASTIC MODELING AND IDENTIFICATION OF UNCERTAINTIES IN COMPUTATIONAL MECHANICS**
MULTILEVEL STOCHASTIC REDUCED-ORDER MODEL IN LINEAR STRUCTURAL DYNAMICS FOR COMPLEX STRUCTURES ..... 6136
*Olivier Ezvan, Anas Batou, Christian Soize*

TO ESTIMATE NON-STATIONARY STOCHASTIC DISTRIBUTED LOAD ON A BEAM STRUCTURE FROM RESPONSE SAMPLES ..... 6148
*S.Q. Wu, J. Zhu*

A ROBUST POLYNOMIAL CHAOS KALMAN FILTER FRAMEWORK FOR CORROSION DETECTION IN REINFORCED CONCRETE STRUCTURES ..... 6160
*Wael Slika, George Saad*

ANALYSIS OF STOCHASTIC DYNAMIC SOIL-STRUCTURE INTERACTION PROBLEMS BY MEANS OF COUPLED FINITE ELEMENTS-PERFECTLY MATCHED LAYERS ..... 6172
*Manthos Papadopoulos, Stijn François, Geert Degrande, Geert Lombaert*

A MULTISCALE FRAMEWORK FOR THE STOCHASTIC ASSIMILATION AND MODELING OF UNCERTAINTY ASSOCIATED NCF COMPOSITE MATERIALS ..... 6186
*Loujaine Mehrez, Roger Ghanem, Colin Mcauliffe, William R. Rodgers, Venkat Aitharaju*

**MS 1306: ERCOFTAC SIG-45: UNCERTAINTY QUANTIFICATION IN CFD AND FLUID STRUCTURE INTERACTION**
A MULTI-FIDELITY ADAPTIVE SAMPLING METHOD FOR METAMODEL-BASED UNCERTAINTY QUANTIFICATION OF COMPUTER SIMULATIONS ..... 6196
*Riccardo Pellegrini, Cecilia Leotardi, Umberto Iemma, Emilio Campana, Matteo Diez*

A LEAST-SQUARES, ADAPTIVE UNCERTAINTY PROPAGATION APPROACH FOR A PLASMA-COUPLED COMBUSTION SYSTEM ..... 6213
*Kunkun Tang, Luca Massa, Jonathan Wang, Jonathan Freund*

UNCERTAINTY QUANTIFICATION IN NUMERICAL SIMULATIONS OF THE FLOW IN THORACIC AORTIC ANEURYSMS ..... 6226
*Alessandro Boccadifuoco, Alessandro Mariotti, Simona Celi, Nicola Martini, Maria Vittoria Salvetti*

**MS 1307: NON-INTRUSIVE SURROGATE MODELS FOR UNCERTAINTY QUANTIFICATION IN HIGH DIMENSIONS**
OPTIMIZATION OF A PHOTOACOUSTIC GAS SENSOR USING MULTIFIDELITY RBF METAMODELING ..... 6250
*Cedric Durantin, Justin Rouxel, Jean-Antoine Desideri, Alain Gliere*

BEARING CAPACITY OF STRIP FOOTINGS ON SPATIALLY RANDOM SOILS USING KRIGING AND MONTE CARLO SIMULATION ..... 6260
*Jawad Thajeel, Tamara Al-Bittar, Nour Issa, Abdul Hamid Soubra*

AN OPTIMAL SENSOR PLACEMENT METHOD FOR SHM BASED ON BAYESIAN EXPERIMENTAL DESIGN AND POLYNOMIAL CHAOS EXPANSION ..... 6272
*Giovanni Capellari, Eleni Chatzi, Stefano Mariani*

CALIBRATION OF NESTED COMPUTER MODELS ..... 6283
*Sophie Marque-Pucheu, Guillaume Perrin, Josselin Garnier*

A STATISTICAL APPROACH FOR BUILDING SPARSE POLYNOMIAL CHAOS EXPANSIONS ..... 6307
*Simon Abraham, Ghader Ghorbaniasl, Chris Lacor*

**MS 1308: MODELLING AND INVERSE METHODS IN NONLINEAR DYNAMICAL SYSTEMS**
FINITE ELEMENT METAMODELING OF UNCERTAIN STRUCTURES ..... 6316
*Vasilis Dertimanis, Dimitris Giagopoulos, Eleni Chatzi*

DYNAMICS OF MULTIBODY SYSTEMS SUBJECT TO UNILATERAL MOTION CONSTRAINTS USING A NEW NUMERICAL ALGORITHM FOR CONTACT DETECTION ..... 6328
*Antonios Pournaras, Fotios Karaoulanis, Sotirios Natsiavas*

EQUATIONS OF MOTION FOR MECHANICAL SYSTEMS SUBJECT TO ACATASTATIC CONSTRAINTS ..... 6342
*Elias Paraskevopoulos, Sotirios Natsiavas*

PARAMETER ESTIMATION OF NONLINEAR LARGE SCALE SYSTEMS THROUGH STOCHASTIC METHODS AND MEASUREMENT OF ITS DYNAMIC RESPONSE ..... 6369
*Dimitrios Giagopoulos, Alexandros Arailopoulos*

**MS 1309: SCALABLE MULTI-FIDELITY MODELING FOR DESIGN, UNCERTAINTY QUANTIFICATION, AND INVERSE PROBLEMS**
VARIABLE FIDELITY MODELLING IN MODERN AIRCRAFT DESIGN ..... 6383
*Marian Zastawny*

**MS 1310: COMPUTATIONAL METHODS FOR THE SOLUTION OF STOCHASTIC DIFFERENTIAL EQUATIONS**
FOUR-DIMENSIONAL PATH INTEGRATION METHOD FOR ESTIMATING THE STOCHASTIC ROLL RESPONSE ..... 6398
*Wei Chai, Arvid Naess, Bernt Leira*

SEISMIC RELIABILITY ASSESSMENT OF FRAME STRUCTURES WITH FLEXIBILITY-BASED STOCHASTIC BEAM-COLUMN ELEMENTS ..... 6407
*Georgios Balokas, Michalis Fragiadakis*

**1400 SPECIAL TECHNOLOGICAL SESSIONS**

**MS 1401: TOICA: THERMAL OVERALL INTEGRATED CONCEPT AIRCRAFT**
THERMAL TRADE OFF SUSTAINED BY MULTI DISCIPLINARY AND MULTI LEVEL OPTIMIZATION ..... 6420
*Olivier Tabaste, Cedric Campguilhem*

COLLABORATIVE ENGINEERING TECHNOLOGIES ENABLING MULTI-PARTNER THERMAL ANALYSIS IN EARLY DESIGN STAGES OF AIRCRAFT ..... 6436
*Erik Baalbergen, Wim Lammen, Bert De Wit, Robert Maas, Marie Moghadasi, Johan Kos, Fabio Chiacchio*

VISUAL ANALYTICS FOR EVALUATION OF VALUE IMPACT IN ENGINEERING DESIGN ..... 6461
*Timoleon Kipouros, Ola Isaksson*

**STS 3: INNOVATIVE DESIGN OPTIMIZATION TOOLS LINKED TO INDUSTRIAL AERONAUTICAL APPLICATIONS: TARGETING GREENER PERFORMANCES**
COMBINING AN RBF-BASED MORPHER WITH CONTINUOUS ADJOINT FOR LOW-SPEED AERONAUTICAL OPTIMIZATION APPLICATIONS ..... 6471
*Evangelos Papoutsis-Kiachagias, Matej Andrejasic, Stefano Porziani, Corrado Groth, David Erzen, Marco Evangelos Biancolini, Emiliano Costa, Kyriakos Giannakoglou*

**STS 5: TRANSITION LOCATION EFFECT ON SHOCK WAVE BOUNDARY LAYER INTERACTION**
LOAD CONTROL OF NATURAL-LAMINAR-FLOW WING VIA BOUNDARY LAYER CONTROL ..... 6485
*Wieńczysław Stalewski, Janusz Sznajder*

**STS 7: MORPHING TECHNOLOGIES FOR AIRCRAFT WINGS**
AN ADAPTIVE TRAILING EDGE FOR LARGE COMMERCIAL AIRCRAFT ..... 6496
*Antonio Concilio, Ignazio Dimino, Rosario Pecora*

**STS 9: ADVANCED WING HIGH-LIFT SYSTEMS**
APPLICATION OF SYNTHETIC JETS ACTUATORS IN WING-PYLON JUNCTION AREA TO IMPROVE THE HIGH LIFT PERFORMANCES ..... 6505
*Petr Vrchota*

**YOUNG INVESTIGATORS MINISYMPOSIUM**
A NANO-MACRO BOTTOM-UP APPROACH TOWARDS BRITTLE FRACTURE ..... 6515
*Sandeep Patil, Yousef Heider, Carlos Hernandez Padilla, Eduardo Cruz Chu, Bernd Markert*

**ECCOMAS OLYMPIAD MINISYMPOSIUM**
INTEGRATION OF DESIGN AND ANALYSIS THROUGH BOUNDARY INTEGRAL EQUATIONS ..... 6526
*Benjamin Marussig, Jürgen Zechner, Gernot Beer, Thomas-Peter Fries*

## Contributed Sessions

**CS 110: NUMERICAL MODELS IN BIOMECHANICS**

EVALUATION OF ANGULAR KINEMATICS OF LOWER LIMB AMPUTEES USING QUANTITATIVE FLUOROSCOPIC IMAGING ..... 6539
*Alexander Breen, Mihai Dupac*

PRELIMINARY STUDY TO INVESTIGATE THE EFFECT OF PISTON-LIKE AND ROCKING MOTIONS OF THE STAPES FOOTPLATE ON THE BASILAR MEMBRANE VIBRATION ..... 6547
*Philipp Wahl, Sebastian Ihrle, Pascal Ziegler, Peter Eberhard*

IN-HOUSE MULTIBODY HUMAN MODEL BASED ON EULER PARAMETERS FOR THE FAST IMPACT SCENARIO CALCULATION ..... 6556
*Jan Spicka, Michal Hajzman, Tomasz Bonkowski*

EXPERIMENTAL STUDY ON THE FUZZY-PID HYBRID CONTROL ALGORITHM FOR UNLOADING SYSTEM IN MECHATRONIC DEVICE FOR GAIT RE-EDUCATION ..... 6567
*Slawomir Duda, Grzegorz Gembalczyk*

FE BONE STRUCTURAL ANALYSIS WITH CT MAPPING OF INHOMOGENEOUS MATERIAL PROPERTIES ..... 6574
*Miguel Tobias Bahia, Emílio Graciliano Ferreira Mercuri, Mildred Ballin Hecke*

SOLVING THE INVERSE PROBLEM OF ESTIMATING FUZZY VISCOELASTIC CONSTITUTIVE PARAMET ERS ..... 6588
*Ruifei Peng, Haitian Yang*

DEVELOPMENT OF A CENTRIFUGAL BLOOD PUMP FOR ECMO AND VAD OPERATIONS ..... 6606
*Shinhwa Choi, Nahmkeon Hur, Mohammad Moshfeghi, Seongwon Kang, Wonjung Kim*

EFFECT OF PERIVASCULAR TISSUE ON INFLATION-EXTENSION BEHAVIOR OF ABDOMINAL AORTA ..... 6616
*Tereza Voňavková, Lukáš Horný, Jan Veselý, Tomáš Adámek, Rudolf Žitný*

EFFECT OF STERILIZATION ON MECHANICAL PROPERTIES OF BIOLOGICAL COMPOSITE ..... 6625
*Jan Vesely, Hynek Chlup, Rudolf Zitny, Tomas Grus*

FINITE-ELEMENT MODEL OF INTRAOCULAR PRESSURE MEASUREMENT BY MAKLAKOV APPLANATION TONOMETER ..... 6631
*Dmitry V. Franus*

**CS 210: NUMERICAL SIMULATION OF COMPOSITE MATERIALS**

MOMENT-CURVATURE DIAGRAMS FOR EVALUATION OF SECOND ORDER EFFECTS IN RC ELEMENTS ..... 6637
*Helena Barros, Carla Ferreira, Tatiana Marques*

INVESTIGATIONS ON AN INNOVATIVE METAL-THERMOPLASTIC COMPOSITE ASSEMBLY ..... 6652
*Simon Paroissien, Patrick Rozycki, Thierry Renault*

A FAILURE CRITERIA FOR UNIDIRECTIONAL FIBER REINFORCED COMPOSITES BASED ON MICROMECHANICS BY ASYMPTOTIC HOMOGENIZATION ..... 6664
*Rafael Quelho De Macedo, José Miranda Guedes, Rafael Thiago Luiz Ferreira, Maurício Vicente Donadon*

EVOLUTION EQUATIONS BASED APPROACH FOR MODELING FATIGUE IN AMORPHOUS GLASSY POLYMERS. ON THE INVESTIGATION OF FATIGUE DAMAGE DEVELOPMENT IN POLYCARBONATE ..... 6675
*Sami Holopainen*

**CS 211: CRACK PROPAGATION**
ULTRASONIC GUIDED WAVES INSPECTION OF PIPES FROM ONE END TO THE FIRST BENT ..... 6688
*Mihai Valentin Predoi, Marian Soare, Ovidiu Vasile, Mihail Boiangiu*

**CS 212: NUMERICAL MODELING OF DAMAGE, FAILURE AND FRACTURE**
DUCTILE FRACTURE CRITERIA IN PREDICTION OF SLANT FRACTURE ..... 6699
*Petr Kubik, Frantisek Sebek, Jindrich Petruska*

DAMAGE MODELING USING STRAIN GRADIENT BASED FINITE ELEMENT FORMULATION ..... 6711
*Filip Putar, Jurica Soric, Tomislav Lesicar, Zdenko Tonkovic*

RELIABILITY ANALYSIS IN FRACTURE MECHANICS ACCORDING TO COMBINED FAILURE CRITERIA ..... 6721
*Rudy Chocat, Paul Beaucaire, Loïc Debeugny, Jean-Pierre Lefebvre, Caroline Sainvitu, Piotr Breitkopf, Eric Wyart*

NEW BENCHMARK FOR THE LIFE ASSESSMENT OF A THIN-WALLED PIPE SUBJECTED TO STRESS ASSISTED CORROSION ..... 6734
*Yulia Pronina, Elena Sedova*

A HIGHER ORDER PHASE-FIELD APPROACH TO FRACTURE FOR FINITE-DEFORMATION CONTACT PROBLEMS ..... 6741
*Marlon Franke, Christian Hesch, Maik Dittmann*

MODELLING THERMAL BARRIER COATINGS AND THEIR INFLUENCE ON THE LIFETIME OF ROCKET ENGINE NOZZLE STRUCTURES ..... 6764
*Marek Fassin, Stephan Wulfinghoff, Stefanie Reese*

ANALYTICAL SOLUTION FOR DYNAMIC FRACTURE OF TWO COPLANAR LIMITED-PERMEABLE CRACKS IN MAGNETO-ELECTRO-ELASTIC MATERIAL ..... 6774
*Peiwei Zhang*

DAMAGE DETECTION OF A BRIDGE BY PARAMETRIC STATISTICAL MOMENT METHOD ..... 6788
*Isabella Failla, Nicola Impollonia, Giuseppe Ricciardi*

**CS 230: MODELLING OF CONCRETE AND MEASURING**
A COHESIVE ZONE MODEL FOR THE CHARACTERISATION OF THE INTERFACIAL TRANSITION ZONE (ITZ) BETWEEN CEMENT PASTE AND AGGREGATE ..... 6799
*Etienne Malachanne, Marie Salgues, Mouad Jebli, Frederic Jamin*

AN EXACT SHEAR STRAIN APPROACH FOR RC FRAME ELEMENTS WITH AXIAL-SHEAR INTERACTION ..... 6810
*Alexander Kagermanov, Paola Ceresa*

CROSS-SECTIONAL FAILURE CRITERION COMBINED WITH STRAIN-HARDENING DAMAGE MODEL FOR SIMULATION OF THIN-WALLED TEXTILE-REINFORCED CONCRETE SHELLS ..... 6823
*Ehsan Sharei, Rostislav Chudoba, Alexander Scholzen*

SEISMIC ASSESSMENT OF A 5-STOREY RETROFITTED RC BUILDING ..... 6832
*Hussein Bark, George Markou, Christos Mourlas, Manolis Papadrakakis*

**CS 310: CAD, CAM AND CAE**
VECTOR FIELD GUIDED TOOL PATHS FOR FIVE-AXIS MACHINING ..... 6847
*Stanislav Makhanov*

AN 'A PRIORI' MODEL REDUCTION FOR ISOGEOMETRIC BOUNDARY ELEMENT METHOD ..... 6882
*Shengze Li, Jon Trevelyan, Weihua Zhang, Zhuxuan Meng*

RESEARCH OF DEFORMATION QUALITIES OF POROUS MATERIALS ENHANCING COMPUTER DESIGN OF WETSUIT ..... 6892
*Irina Cherunova, Elena Sirota, Nikolai Kornev, Mathias Paschen, Sebastian Schreier, Tatyana Lesnikova, Pavel Cherunov*

MATHEMATICAL MODELLING OF CITY AERODYNAMICS ..... 6900
*Svetlana Valger, Natalya Fedorova, Alexander Fedorov*

**CS 320: GRID GENERATION AND ADAPTIVE TECHNIQUES**
A STABLE AND CONSERVATIVE TIME-DEPENDENT INTERFACE FORMULATION ON SUMMATION-BY-PARTS FORM: AN INITIAL INVESTIGATION ..... 6911
*Samira Nikkar, Jan Nordström*

A LAPLACIAN MESH DEFORMATION TECHNIQUE FOR SIMULATION-DRIVEN DESIGN OPTIMIZATION ..... 6928
*Mario J. Martin-Burgos, Daniel González-Juárez, Esther Andrés-Pérez*

AN INTEGRATED FRAMEWORK FOR WRAPPING AND MESH GENERATION OF COMPLEX GEOMETRIES ..... 6938
*David Martineau, Jeremy Gould, Jacques Papper*

**CS 410: COMPUTATIONAL FLUID MECHANICS**
THE VORTICITY CREATION PROCESS AT PHYSICAL SURFACES ..... 6955
*Horia Dumitrescu, Vladimir Cardos*

HEAT RECOVERY BY CROSS FLOW ..... 6966
*Karel Adamek, Jan Kolar, Pavel Peukert*

THE INLET AND OUTLET BOUNDARY PROBLEM WITH THE PREFERENCE OF MASS FLOW ..... 6979
*Martin Kyncl, Jaroslav Pelant*

BUDGET ANALYSIS OF TURBULENT KINETIC ENERGY IN CORNER SEPARATION : RANS VS LES ..... 6999
*Jean-François Monier, Feng Gao, Jérôme Boudet, Liang Shao, Lipeng Lu*

A SHALLOW WATER EQUATION BASED ON DISPLACEMENT AND PRESSURE AND THE ZU-CLASS SYMPLECTIC METHOD ..... 7014
*Feng Wu, Wan-Xie Zhong*

IMMERSED BOUNDARY METHODS FOR COMPRESSIBLE LAMINAR FLOWS ..... 7029
*Rakesh Ramakrishnan, Anant Girdhar, Santanu Ghosh*

IMPACT OF EXTERNAL SURROUNDINGS ON NATURAL CONVECTION IN A VERTICAL CHANNEL ASYMMETRICALLY HEATED ..... 7045
*Delphine Ramalingom, Alain Bastide*

INFLUENCE OF ADJACENT ROOMS ON THE DEVELOPMENT OF GAS EXPLOSION ..... 7056
*Iurii Polandov, Sergei Dobrikov*

DOMAIN DECOMPOSITION APPROACH FOR NEAR-WALL TURBULENCE MODELING ..... 7066
*Sergey Utyuzhnikov*

LARGE EDDY SIMULATION OF SALTATION OVER GAUSSIAN HILLS ..... 7074
*Gang Huang, Catherine Le Ribault, Serge Simoëns, Ivana Vinkovic, J.M.Vignon*

AN EXPERIMENTAL AND NUMERICAL STUDY ON THE PERFORMANCE OF AN INNOVATIVE VERTICAL-AXIS WIND TURBINE ..... 7087
*Katarzyna Kludzinska, Krzysztof Tesch, Piotr Doerffer*

NUMERICAL SIMULATION OF NONLINEAR FREE SURFACE WATER WAVES USING A COUPLED POTENTIAL FLOW-URANS/VOF APPROACH ..... 7095
*Bulent Duz, Tim Bunnik, Geert Kapsenberg*

THE OPTIMAL CONTROL OF A MULTI-MASS VIBRATION PROPULSION SYSTEM IN A VISCOUS INCOMPRESSIBLE FLUID ..... 7121
*Artem Nuriev, Olga Zakharova*

HIGH-RESOLUTION SIMULATION OF INTERNAL WAVES ATTRACTORS AND IMPACT OF INTERACTION OF HIGH AMPLITUDE INTERNAL WAVES WITH WALLS ON DYNAMICS OF WAVES ATTRACTORS ..... 7130
*Ilias Sibgatullin, Michael Kalugin*

NUMERICAL SIMULATION OF HEAT AND MASS TRANSFER PROCESSES IN LARGE-SCALE FLUIDIZED BED COMPLEX STRUCTURE APPARATUS AS AN EXAMPLE OF THE REACTOR OF ISOPARAFFINS DEHYDROGENATION ..... 7138
*Sergei Solovev, Svetlana Egorova, Alexander Lamberov, Olga Soloveva*

**CS 420: MULTI-PHASE AND CHEMICALLY REACTING FLOWS**

SHOCK CAPTURING COMPUTATIONS WITH STABILIZED POWELL-SABIN ELEMENTS ..... 7149
*Giorgio Giorgiani, Herve Guillard, Boniface Nkonga*

NUMERICAL INVESTIGATION OF METASTABLE CONDENSING FLOWS WITH AN IMPLICIT UPWIND METHOD ... 7165
*Lucia Azzini, Teus Van Der Stelt, Matteo Pini*

LEVEL SET METHOD FOR SIMULATING THE INTERFACE KINEMATICS: APPLICATION OF A DISCONTINUOUS GALERKIN METHOD ..... 7182
*Roozbeh Mousavi, Florian Kummer, Martin Oberlack, Peter F. Pelz*

LARGE EDDY SIMULATION OF TURBULENT COMPRESSIBLE FLOWS USING THE CHARACTERISTIC BASED SPLIT SCHEME AND MESH ADAPTATION ..... 7222
*Renato Linn, Armando Awruch*

NUMERICAL SIMULATION OF HYDROGEN JET INJECTION AND IGNITION IN SUPERSONIC FLOW ..... 7237
*Yulia Zakharova, Natalya Fedorova, Svetlana Valger, Marat Goldfeld, Olga Vankova*

A 3D UNSTEADY NUMERICAL SIMULATION OF THE REFRIGERANT FLOW EVAPORATION IN A PLATE HEAT EXCHANGER ..... 7247
*Mirza Popovac, Gerwin Schmid, Michael Lauermann*

RESEARCH OF UNSTEADY FLOW REGIMES IN CHANNEL OF HYPERSONIC INLET ..... 7255
*Natalya Fedorova, Marat Goldfeld*

EXPLICIT DARCY'S LAW BOUNDARY CONDITION WITH COMBINED CONTINUUM AND DISCRETE MODEL FOR PRESSURE DRIVEN MEMBRANE APPLICATIONS ..... 7266
*Tomi Naukkarinen, Teemu Turunen-Saaresti*

**CS 450: NUMERICAL METHODS AND CONVERGENCE ACCELERATION IN CFD**

A MULTIGRID FORMULATION FOR FINITE DIFFERENCE METHODS ON SUMMATION-BY-PARTS FORM: AN INITIAL INVESTIGATION ..... 7274
*Andrea Alessandro Ruggiu, Per Weinerfelt, Tomas Lundquist, Jan Nordström*

IMPROVED DUAL TIME–STEPPING USING SECOND DERIVATIVES ..... 7285
*Jan Nordström, Andrea Ruggiu*

A DIRECT SOLVER FOR THE ADVECTION-DIFFUSION EQUATION USING GREEN'S FUNCTIONS AND LOW-RANK APPROXIMATION ..... 7302
*Jonathan Bull*

AN IMPROVED DISCRETIZATION METHOD FOR BOUNDED CONVECTIVE SCHEMES ON UNSTRUCTURED CO-LOCATED GRIDS ..... 7317
*Vlado Przulj*

A LOW-MACH, LOW-REYNOLDS PRECONDITIONING SCHEME WITH PARTICULAR ATTENTION TO VISCOUS TIME-STEPPING ..... 7334
*Jens Fiedler, Graham Ashcroft*

PERFORMANCE IMPROVEMENT OF FLOW COMPUTATIONS WITH AN OVERSET-GRID METHOD INCLUDING BODY MOTIONS USING A FULL MULTIGRID METHOD ..... 7347
*Kunihide Ohashi, Hiroshi Kobayashi*

SOLVING LINEAR SYSTEMS WITH MULTIPLE RIGHT-HAND SIDES WITH GMRES : AN APPLICATION TO AIRCRAFT DESIGN ..... 7358
*Aloïs Bissuel, Grégoire Allaire, Laurent Daumas, Frédéric Chalot, Michel Mallet*

COMPARISON OF DIFFERENT SPATIAL/ANGULAR AGGLOMERATION MULTIGRID SCHEMES FOR RADIATIVE HEAT TRANSFER COMPUTATIONS ..... 7372
*Georgios Lygidakis, Ioannis Nikolos*

**CS 460: UNSTEADY FLOW COMPUTATION**
LINEARISED FREQUENCY DOMAIN GUST ANALYSIS OF LARGE CIVIL AIRCRAFT ..... 7390
*Philipp Bekemeyer, Reik Thormann, Sebastian Timme*

CONSISTENT NON-REFLECTING BOUNDARY CONDITIONS FOR BOTH STEADY AND UNSTEADY FLOW SIMULATIONS IN TURBOMACHINERY APPLICATIONS ..... 7403
*Daniel Schlüß, Christian Frey, Graham Ashcroft*

REDUCED ORDER MODELLING OF GUST ANALYSIS USING COMPUTATIONAL FLUID DYNAMICS ..... 7423
*Reik Thormann, Philipp Bekemeyer, Sebastian Timme*

ON THE REGULARISATION OF NON-REFLECTING BOUNDARY CONDITIONS NEAR ACOUSTIC RESONANCE ..... 7435
*Christian Frey, Hans-Peter Kersken*

ON GRID RESOLUTION REQUIREMENTS FOR LES OF WALL-BOUNDED FLOWS ..... 7454
*Saleh Rezaeiravesh, Mattias Liefvendahl, Christer Fureby*

A HYBRID MESH HARMONIC BALANCE SOLVER FOR THE AEROELASTIC ANALYSIS OF TURBOMACHINERY ..... 7466
*Hans-Peter Kersken, Graham Ashcroft, Christian Frey*

NUMERICAL AND EXPERIMENTAL INVESTIGATIONS OF DYNAMIC STALL PROBLEM ON A 2D HELICOPTER ROTOR BLADE SECTION ..... 7477
*Antonello Marino, Serena Russo, Giovanni Paolo Reina, Gennaro Esposito, Francesco Capizzano, Carlo De Nicola*

INVESTIGATION OF A MODERN & HYBRID TURBULENCE MODELLING APPROACH FOR TRANSIENT AUTOMOTIVE AERODYNAMICS SIMULATIONS ..... 7488
*Rene Devaradja, Petr Simanek, Jacques Papper, Pavla Policka*

FREE-SURFACE FLOW SIMULATIONS FOR MOORED AND FLOATING OFFSHORE PLATFORMS ..... 7515
*Arthur Veldman, Roel Luppes, Peter Van Der Plas, Henri Van Der Heiden, Bulent Duz, Henk Seubers, Joop Helder, Tim Bunnik, Rene Huijsmans*

MODELLING OF UNSTEADY SECONDARY VORTICES GENERATED BEHIND THE RADIAL GAP OF THE AXIAL TURBINE BLADE WHEEL ..... 7532
*Petr Straka*

PATHWAYS TO IMPROVED AERODYNAMIC DESIGN ..... 7544
*James Page, Paul Hield, Callum Mantell, Paul Tucker*

SIMULATIONS OF BLAST WAVE PROPAGATION IN OPEN AND CLOSED SPACE ..... 7557
*Natalya Fedorova, Svetlana Valger, Yulia Zakharova*

**CS 500: HIGH PERFORMANCE COMPUTING**
ON A DYNAMIC SCHEDULING ALGORITHM FOR MASSIVELY PARALLEL COMPUTATIONS OF ATOMIC ISOTOPE ..... 7572
*Elizaveta Dorofeeva, Jatin Arora, Stefan Typel, Gevorg Poghosyan, Peter Sanders, Achim Streit*

APPLICATION THE SPARSE MATRIX METHOD TO CALCULATE THE METAL ELASTIC STRESS-STRAIN STATE, USING THE FINITE ELEMENT METHOD ..... 7581
*Inna Barankova, Uliana Mikhailova, Victor Demidovich*

**CS 610: AERO-ACOUSTICS**
FLUID ANALYSIS USING FICTITIOUS DOMAIN FINITE ELEMENT METHOD ..... 7589
*Yukihiro Terakado, Tahahiko Kurahashi*

ON THE SIMULATION OF AERODYNAMIC NOISE WITH DIFFERENT TURBULENCE MODELS ..... 7599
*Xin Huang, Michael Schäfer*

NUMERICAL INVESTIGATION OF THERMOACOUSTIC ENGINE USING IMPLICIT LARGE EDDY SIMULATION ..... 7609
*Mustapha Mahdaoui, Rachid Bennacer, Smaïne Kouidri, N. Martaj*

SEMI-DISCRETE COINCIDENCE IN THE MID-FREQUENCY SOUND TRANSMISSION THROUGH RIB-STIFFENED PANELS ..... 7620
*Edwin Reynders, Cédric Van Hoorickx, Arne Dijckmans*

**CS 620: COMPUTATIONAL CONTACT MECHANICS**
SOLVING UNILATERAL CONTACT PROBLEMS IN MULTIBODY DYNAMICS USING A PRIMAL-DUAL INTERIOR POINT METHOD ..... 7628
*Dario Mangoni, Alessandro Tasora*

**CS 630: SIMULATION OF FLUID-STRUCTURE INTERACTION**
3D DRBEM MODELING FOR ROTATING INITIALLY STRESSED ANISOTROPIC FUNCTIONALLY GRADED PIEZOELECTRIC PLATES ..... 7640
*Mohamed Abdelsabour Fahmy*

DIRECT EDDY CURRENT METHOD FOR VOLUMETRIC FLAWS OF CYLINDRICAL SHAPE ..... 7659
*Valentina Koliskina, Andrei Kolyshkin, Rauno Gordon, Olev Märtens*

NUMERICAL SIMULATION OF BUOYANT PLUMES USING A FIXED POINT ITERATIVE METHOD ..... 7666
*Blanca Bermúdez, Alejandro Rangel-Huerta, W. Fermín Guerrero S., José David Alanis Urquieta*

INVESTIGATION OF THE FLOW OVER AN OSCILLATING CYLINDER WITH A VERY LARGE EDDY SIMULATION MODEL ..... 7674
*Anastasia Kondratyuk, Michael Schäfer, Awais Ali*

NUMERICAL STUDY OF TWO OPTIMIZED COUPLING INTERFACE TREATMENTS FOR STEADY CONJUGATE HEAT TRANSFER PROBLEMS ..... 7682
*Marc Paul Errera, Roch Roukoz El Khoury*

EFFECT OF DISC GEOMETRY ON THE DYNAMIC STABILITY OF DIRECT SPRING OPERATED PRESSURE RELIEF VALVES ..... 7695
*István Erdődi, Csaba Hős*

CFD SIMULATION OF THERMO- AERODYNAMIC INTERACTION IN A SYSTEM HUMAN-CLOTH-ENVIRONMENT UNDER VERY LOW TEMPERATURE AND WIND CONDITIONS ..... 7703
*Irina Cherunova, Sina Samarbaksch, Nikolai Kornev*

**CS 720: COMPUTATIONAL MATERIALS SCIENCE**
ON THE INFLUENCE OF THE INCLUSIONS' MORPHOLOGY ON THE ACCURACY OF THE PREDICTION OF REINFORCED COMPOSITES MECHANICAL BEHAVIOUR ..... 7711
*Viwanou Hounkpati, Vladimir Salnikov, Philippe Karamian-Surville, Alexandre Vivet*

NUMERICAL SIMULATION OF ADVANCED PULTRUSION PROCESSES WITH MICROWAVE HEATING ..... 7720
*Evgeny Barkanov, Pavel Akishin, Rudolf Emmerich, Matthias Graf*

EFFICIENT NUMERICAL SIMULATION OF INDUSTRIAL SHEET METAL BENDING PROCESSES ..... 7739
*Christian Zehetner, Paula Reimer, Franz Hammelmüller, Hans Irschik, Wolfgang Kunze*

NUMERICAL MODELLING AND SIMULATION OF SHEET METAL CUTTING PROCESSES ..... 7749
*Paula Reimer, Christian Zehetner, Franz Hammelmüller, Wolfgang Kunze*

SIMULATION OF PRECIPITATION IN V-CONTAINING HSLA STEEL FOR THE STRENGTHENING ENHANCEMENT ..... 7757
*Piyada Suwanpinij, Paolo Massaro, Annalisa Pola, Prasonk Srichareonchai*

**CS 750: COMPUTATIONAL MODELING OF COMPOSITES**
NUMERICAL ANALYSIS OF SANDWICH PANELS SUBJECTED TO POINT LOADS ..... 7765
*Zbigniew Pozorski, Jolanta Pozorska*

BOND OF REINFORCEMENT IN REACTIVE POWDER CONCRETE: EXPERIMENTAL STUDY ..... 7771
*Mingde Sun, Ri Gao, Aili Li, Yongjing Wang*

MODELLING SOIL-FIBRE COMPOSITE BEHAVIOUR USING A MICROMECHANICAL APPROACH ..... 7779
*Thomas Bower, Anthony Jefferson, Peter Cleall, Paul Lyons*

NUMERICAL SIMULATION OF MECHANICAL BEHAVIOR OF WOVEN COMPOSITE AT DIFFERENT STRAIN RATE BY A COLLABORATIVE ELASTO-PLASTO-DAMAGE MODEL WITH FRACTIONAL DERIVATIVES ..... 7786
*Alina Krasnobrizha, Patrick Rozycki, Laurent Gornet, Pascal Cosson*

NUMERICAL SIMULATION OF MECHANICAL BEHAVIOUR OF COMPOSITE SANDWICH PANELS WITH DEFECTS ..... 7800
*Aleksandr Anoshkin, Valerii Zuiko, Mikhail Alikin, Anna Tchugaynova*

TIMOSHENKO BEAM ELEMENT WITH ANISOTROPIC CROSS-SECTIONAL PROPERTIES ……. 7810
*Alexander R. Stäblein, Morten H. Hansen*

STABILIZATION OF THE NUMERICAL PRIMING TO STABILIZATION PHENOMENON IN THE CRUSHING OF A COMPOSITE PANEL USING LOCAL AND DIFFUSE DAMAGE LAWS …….. 7820
*Christine Espinosa, Floran Tostain, Samuel Rivallant*

STRUCTURAL ANALYSIS OF ADHESIVE BONDING FOR THICK-WALLED TUBULAR COMPOSITE PROFILES …….. 7837
*Geminiano Mancusi, Agostina Orefice, Luciano Feo, Fernando Fraternali*

**CS 751: SMART MATERIALS AND STRUCTURES**
TOWARDS REAL-TIME STRUCTURAL HEALTH MONITORING DAMAGE DETECTION WITHOUT USER INPUT ….. 7853
*Mohmmad Salmanpour, Zahra Sharif Khodaei, Ferri Aliabadi*

FIBROUS TECTONICS:A RETHINKING OF COMPOSITE PRODUCTION THROUGH INNOVATION AND EXPLORATION OF MOLDING TECHNIQUES AND METHODOLOGIES …….. 7863
*David Costanza*

**CS 830: COMPUTATIONAL NANOTECHNOLOGY**
EFFECT OF NANOSIZED ASPERITIES AT THE SURFACE OF A NANOHOLE …….. 7875
*Mikhail Grekov, Aleksandra Vakaeva*

**CS 840: MULTI-SCALE COMPUTATIONAL METHODS FOR SOLIDS AND FLUIDS**
THE GEMA FRAMEWORK – AN INNOVATIVE FRAMEWORK FOR THE DEVELOPMENT OF MULTIPHYSICS AND MULTISCALE SIMULATIONS …….. 7886
*Carlos Augusto Teixeira Mendes, Marcelo Gattass, Deane Roehl*

NONLINEAR PATCH METHOD AND APPLICATION …….. 7895
*Antoine Brunet, Pierre Sarrailh, François Rogier, Jean-François Roussel, Denis Payan*

ASSESSMENT OF A CONTINUUM MICROMECHANICS-BASED MULTISCALE MODEL FOR CONCRETE BY MEANS OF SENSITIVITY ANALYSIS AND UNCERTAINTY PROPAGATION …….. 7905
*Luise Göbel, Andrea Osburg, Tom Lahmer*

SPATIAL CLUSTERING STRATEGIES FOR HIERARCHICAL MULTI-SCALE MODELLING OF METAL PLASTICITY …… 7922
*Md Khairullah, Jerzy Gawad, Dirk Roose, Albert Van Bael*

BLACK-BOX SOLVER FOR ONE-DIMENSIONAL MULTISCALE MODELLING USING THE QTT FORMAT …….. 7938
*Ivan Oseledets, Maxim Rakhuba, Andrei Chertkov*

SCALE EFFECT IN MICROFLOWS MODELLING WITH THE MICROPOLAR FLUID THEORY …….. 7948
*Anna Kucaba-Pietal*

**CS 930: HIGH-ORDER DISCRETIZATION METHODS**

A COMPARISON OF VARIOUS NODAL DISCONTINUOUS GALERKIN METHODS FOR THE 3D EULER EQUATIONS ..... 7956
*Michael Bergmann, Svetlana Drapkina, Graham Ashcroft, Christian Frey*

A 3D ISOGEOMETRICAL BOUNDARY ELEMENT ANALYSIS FOR NON-LINEAR GRAVITY WAVE PROPAGATION ... 7967
*Jorge Maestre, Jordi Pallarés, Ildefonso Cuesta*

HYBRID OPENMP/MPI PARALLELIZATION OF A HIGH–ORDER DISCONTINUOUS GALERKIN CFD SOLVER ..... 7992
*Francesco Bassi, Alessandro Colombo, Andrea Crivellini, Matteo Franciolini*

ON THE CHOICE OF SHOCK CAPTURING SCHEMES FOR THE SOLUTION OF THE LWR TRAFFIC FLOW EQUATION USING A HIGH ORDER MODAL DISCONTINUOUS GALERKIN DISCRETIZATION ..... 8013
*Alberto Costa Nogueira Junior, João Lucas De Sousa Almeida, Cláudio Alessandro De Carvalho Silva*

COMPARISON OF FINITE VOLUME HIGH-ORDER SCHEMES FOR THE WO-DIMENSIONAL EULER EQUATIONS ..... 8033
*Jens Wellner*

NURBS-BASED ISOGEOMETRIC ANALYSIS FOR BALLISTIC EVALUATION OF TITANIUM PLATES ..... 8047
*Mattia Montanari, Petros Siegkas, Antonio Pellegrino, Nik Petrinic*

HYBRID RIEMANN SOLVERS FOR LARGE SYSTEMS OF CONSERVATION LAWS ..... 8057
*Birte Schmidtmann, Mariia Astrakhantceva, Manuel Torrilhon*

**CS 940: EXTENDED DISCRETIZATION METHODS**

THE USE OF ENRICHED BASE FUNCTIONS IN THE THREE-DIMENSIONAL SCALED BOUNDARY FINITE ELEMENT METHOD ..... 8066
*Sascha Hell, Wilfried Becker*

METHODOLOGY OF COMBINED APPLICATION OF DIRECTIONAL DERIVATIVES AND THE EXTENDED FINITE ELEMENT METHOD (X-FEM) FOR SOLVING VIBRATION EIGENVALUE PROBLEMS ..... 8085
*Daria Serbichenko, Pascal Cosson, Patrick Rozycki*

XFEM ANALYSIS – INCLUDING BUCKLING – OF COMPOSITE SHELLS CONTAINING DELAMINATION ..... 8103
*Wilhelm J.H. Rust, Saleh Yazdani, Peter Wriggers*

**CS 960: MESHLESS METHODS**

THREE-DIMENSIONAL CRACK PROPAGATION ANALYSIS USING MESHLESS POINT COLLOCATION METHOD .... 8115
*Eiji Tanaka*

A CELL-BASED SMOOTHED POINT INTERPOLATION METHOD FOR AXISYMMETRIC PROBLEMS ..... 8133
*Arash Tootoonchi, Arman Khoshghalb*

EFFICIENT SHAPE FUNCTION COSTRUCTION OF EFG MESHLESS METHOD ..... 8142
*P. Metsis, M. Papadrakakis*

**CS 980: NUMERICAL AND SYMBOLIC COMPUTATION**
MODERNIZING SOFTWARE IN SCIENCE AND ENGINEERING: FROM C/C++ APPLICATIONS TO MOBILE PLATFORMS ..... 8162
*Liliana Favre, Claudia Pereira, Liliana Martinez*

NEW EDUCATIONAL MATLAB TOOL TO EXPLAIN TWO DIMENSIONAL FINITE ELEMENT METHOD TO GRADUATE STUDENTS ..... 8177
*Rafet Sisman, Abdurrahman Sahin*

CONTRIBUTION OF NUMERICAL THERMAL SIMULATION OF MASONRY TO THE GLOBAL OBJECTIVE OF BUILDINGS ENERGY EFFICIENCY ..... 8187
*Hipólito Sousa, Rui Sousa, Luís Silva, Gonçalo Sousa*

**CS 990: PARTICLE-BASED METHODS**
TOWARDS MULTIPHYSICS SIMULATION OF DEEP PENETRATION LASER WELDING USING SMOOTHED PARTICLE HYDRODYNAMICS ..... 8196
*Haoyue Hu, Peter Eberhard, Florian Fetzer, Peter Berger*

MULTIPLE TIMESCALE MODELLING OF PARTICLE SUSPENSIONS IN METAL MELTS SUBJECTED TO EXTERNAL FORCES ..... 8207
*Anton Manoylov, Georgi Djambazov, Valdis Bojarevics, Koulis Pericleous*

STATISTICAL DISTRIBUTION FUNCTION OF CHARGED PARTICLES IN MAGNETIC FIELD ..... 8224
*B.Chetverushkin, Nicola D'Ascenzo, Valeri Saveliev*

A STABILISED TOTAL LAGRANGIAN CORRECTED SMOOTH PARTICLE HYDRODYNAMICS TECHNIQUE IN LARGE STRAIN EXPLICIT FAST SOLID DYNAMICS ..... 8231
*Giorgio Greto, Sivakumar Kulasegaram, Chun H. Lee, Antonio J. Gil, Javier Bonet*

BLOOD FLOW SIMULATION USING SMOOTHED PARTICLE HYDRODYNAMICS ..... 8241
*Mohammed Al-Saad, Sivakumar Kulasegaram, Stephane P.A. Bordas*

**CS 1010: COMPUTATIONAL INVERSE PROBLEMS AND OPTIMIZATION**
ON THE VALIDITY OF TIDAL TURBINE ARRAY CONFIGURATIONS OBTAINED FROM STEADY-STATE ADJOINT OPTIMISATION ..... 8247
*Christian T. Jacobs, Matthew D. Piggott, Stephan C. Kramer, Simon W. Funke*

DEVELOPMENT OF ACCURATE PNEUMATIC TYRE FINITE ELEMENT MODELS BASED ON AN OPTIMISATION PROCEDURE ..... 8262
*Chrysostomos-Alexandros Bekakos, George Papazafeiropoulos, Dan J. O'Boy, Jan Prins*

METEOROLOGICAL DATA ASSIMILATION USING AN ADJOINT PROGRAM GENERATED BY AUTOMATIC DIFFERENTIATION ..... 8275
*Yasuyoshi Horibata*

EVOLUTIONARY TOPOLOGY OPIMIZATION USING PARAMETERIZED B-SPLINE SURFACE ..... 8283
*Igor Pehnec, Damir Vucina, Frane Vlak*

GEARBOX DESIGN VIA MIXED-INTEGER PROGRAMMING ..... 8294
*Bastian Dörig, Thorsten Ederer, Peter F. Pelz, Marc E. Pfetsch, Jan Wolf*

A FRAMEWORK FOR THE DESIGN BY OPTIMIZATION OF HYDROFOILS UNDER CAVITATING CONDITIONS ..... 8305
*Paolo Olivucci, Stefano Gaggero*

A COMPUTATIONAL STRATEGY FOR TRAJECTORY OPTIMIZATION OF UNDERACTUATED MULTIBODY SYSTEMS WITH CONTACTS ..... 8324
*Silvia Manara, Alessio Artoni, Marco Gabiccini*

ANALYSIS OF NON-LINEAR ENVIRONMENTAL LOAD COMBINATIONS BY EXTENDED CONTOUR-LINE ALGORITHMS ..... 8337
*Bernt Leira*

ADJOINT-BASED ROBUST OPTIMIZATION USING POLYNOMIAL CHAOS EXPANSIONS ..... 8351
*Joao Miranda, Dinesh Kumar, Chris Lacor*

APPLICATION OF THE DECONVOLUTION METHODS FOR PROCESSING OF MEASUREMENT SIGNALS IN THE FAST PROCESSES ..... 8365
*Marat Goldfeld, Valery Pickalov*

INVERSION METHODS BASED ON KALMAN FILTERING FOR IDENTIFICATION IN TUNNELING PROBLEMS ..... 8378
*Tamara Nestorović, Luan T. Nguyen*

**CS 1020: EVOLUTIONARY AND DETERMINISTIC METHODS FOR DESIGN, OPTIMIZATION AND CONTROL**
EVALUATION OF DIFFERENT APPROACHES FOR THE OPTIMIZATION OF LAYOUT AND CONTROL OF BOOSTER STATIONS ..... 8391
*Philipp Pöttgen, Peter F. Pelz*

MANIFESTATION OF ASCENDANCY OF EXTINCTION-REIGNITION ON SOUNDING HYBRID ROCKET USING DESIGN INFORMATICS ..... 8403
*Kazuhisa Chiba, Masahiro Kanazaki, Toru Shimada*

COMPUTING AN OPERATING STRATEGY FOR AN ACTIVE BODY CONTROL WITH DYNAMIC PROGRAMMING IN THE AUTOMOTIVE AREA ..... 8419
*Marlene Utz, Phillipp Hedrich, Peter F. Pelz*

SEISMIC RELIABILITY-BASED DESIGN OF STRUCTUTRES ISOLATED BY FPS ..... 8431
*Paolo Castaldo, Guglielmo Amendola, Bruno Palazzo*

SEISMIC RETROFIT OF EXISTING BUILDINGS THROUGH THE DISSIPATIVE COLUMNS ..... 8445
*Paolo Castaldo, Bruno Palazzo, Francesco Perri, Ivana Marino, Marco Maria Faraco*

AN EXAMPLE OF ENERGY DISSIPATION OPTIMIZATION FOR STEEL MRFS WITH PIN-JOINTED COLUMN BASES ..... 8460
*Elide Nastri, Annabella Paciello*

ECCENTRICALLY BRACED FRAMES DESIGNED FOR THE ENERGY DISSIPATION OPTIMIZATION ..... 8476
*Elide Nastri*

AN EXAMPLE OF COLUMN DESIGN OPTIMIZATION FOR FAILURE MODE CONTROL OF REINFORCED CONCRETE FRAMES ..... 8492
*Roberta Muscati*

A COMPARISON OF MILP AND MINLP SOLVER PERFOMANCE ON THE EXAMPLE OF A DRINKING WATER SUPPLY SYSTEM DESIGN PROBLEM ..... 8509
*Lea Rausch, Philipp Leise, Thorsten Ederer, Lena C. Altherr, Peter F. Pelz*

RELIABILITY TEST DEMONSTRATION METHOD FOR EXPONENTIAL LIFE SYSTEM WITH RELIABILITY GROWTH UNDER THE CONDITION OF IN-TIME CORRECTIVE STRATEGY ..... 8528
*Yunyan Xing, Ping Jiang, Feng Yao, Zhiwei Yang, Michael Emmerich, Thomas Bäck*

**CS 1100: REDUCTION ORDER METHODS**
EVALUATION OF LINE-FITTING METHOD OF MODEL ORDER REDUCTION ..... 8539
*Valery Makhavikou, Roland Kasper, Dmitry Vlasenko*

AN EFFICIENT ORDER REDUCTION STRATEGY IN EARTHQUAKE NONLINEAR RESPONSE ANALYSIS OF STRUCTURES ..... 8552
*Franz Bamer, Abbas Kazemi Amiri, Christian Bucher*

ACTIVE SUBSPACES FOR THE PRELIMINARY FLUID DYNAMIC DESIGN OF UNCONVENTIONAL TURBOMACHINERY ..... 8572
*Juan S. Bahamonde, Matteo Pini, Piero Colonna*

EXTENSION AND APPLICATION OF A NONLINEAR REDUCED ORDER MODEL TO GUST LOAD PREDICTION IN TIME DOMAIN ..... 8587
*Christoph Strobach, Klemens Lindhorst, Matthias Haupt, Peter Horst*

**CS 1200: STRUCTURAL DYNAMICS**
RESEARCH ON HYSTERETIC BEHAVIORS OF A SAPARATED SHOCK ABSORBER APPLIED IN RAILWAY BRIDGE ..... 8597
*Aili Li, Ri Gao, Mingde Sun, Jilei Zhang*

FINITE ELEMENT ANALYSIS OF TIMBER BEAMS WITH FLAWS ..... 8606
*Janka Kovacikova, Mats Ekevad, Olga Ivankova, Sven Berg*

NONLINEAR FORCED RESPONSE OF A STATOR VANE WITH MULTIPLE FRICTION CONTACTS USING A COUPLED STATIC/DYNAMIC APPROACH ..... 8612
*Marco Lassalle, Christian Maria Firrone*

COMPUTATIONAL METHOD OF DETERMINATION OF INTERNAL EFFORTS IN LINKS OF MECHANISMS AND ROBOT MANIPULATORS WITH STATICALLY DEFINABLE STRUCTURES CONSIDERING THE DISTRIBUTED DYNAMICALLY LOADINGS ..... 8627
*Zhumadil Baigunchekov, Muratulla Utenov, Nurzhan Utenov, Saltanat Zhilkibayeva*

FE MODELLING OF SFRC BEAMS UNDER IMPACT LOADS ..... 8640
*Pegah Behinaein, Ali Abbas, Demetris Cotsovos*

AN APROXIMATE ANALYTICAL SOLUTION FOR NONLINEAR FGM SHELL STRUCTURE WITH VARIABLE IN TIME PARAMETERS ..... 8654
*Victor Gristchak, Yuliia Fatieieva*

UNCERTAINTY OF MODELS IN INTELLIGENT SYSTEMS UNDER STOCHASTIC LOADING ..... 8665
*Amalia Moutsopoulou, Georgios Stavroulakis, Tasos Pouliezos*

COMPUTATIONAL MODELLING OF COLD-FORMED STEEL SCREWED CONNECTIONS AT AMBIENT AND ELEVATED TEMPERATURES ..... 8678
*Luis Mesquita, Rui Dias, Armandino Parente, Paulo Piloto*

MONITORING AND ANALYSIS OF STRESS FIELD FOR ORTHOTROPIC STEEL DECK OF DASHENGGUAN YANGZTE RIVER BRIDGE ..... 8691
*Ying Wang, Y.S. Song*

**CS 1201: COMPUTATIONAL SOIL MECHANICS**
A MORE COMPREHENSIVE MODELING OF CONTACT FORCE DURING SHEAR TESTING USING DEM ..... 8708
*Varvara Roubtsova, Mohamed Chekired*

THREE-DIMENSIONAL BE-FE MODEL OF BUCKET FOUNDATIONS IN POROELASTIC SOILS ..... 8725
*Jacob D. R. Bordón, Juan J. Aznárez, Orlando F. Maeso*

**CS 1202: STRUCTURAL ANALYSIS AND MULTI BODY DYNAMICS**
ON THE NUMERICAL INFLUENCES OF INERTIA REPRESENTATION FOR RIGID BODY DYNAMICS ..... 8739
*Xiaoming Xu, Wanxie Zhong*

APPROACHES TO THE CREATION OF MULTIBODY MODELS OF THE VVER 1000 NUCLEAR REACTOR CONTROL ASSEMBLY ..... 8756
*Pavel Polach, Michal Hajžman*

**CS 1300: UNCERTAINTY QUANTIFICATION AND ERROR ESTIMATION**
THE UPLIFT CAPACITY OF HORIZONTAL PLATE ANCHORS IN SPATIALLY VARIABLE CLAY USING SPARSE POLYNOMIAL CHAOS EXPANSIONS ..... 8769
*Tom Charlton, Mohamed Rouainia*

FATIGUE RELIABILITY OF AGEING RAILWAY BRIDGES: FEASIBILITY OF PROBABILISTIC APPROACH ..... 8778
*Nirosha D. Adasooriya*

AN EFFICIENT AERODYNAMIC SHAPE OPTIMIZATION FRAMEWORK FOR ROBUST DESIGN OF AIRFOILS USING SURROGATE MODELS ..... 8787
*Daigo Maruyama, Dishi Liu, Stefan Görtz*

QUANTIFICATION OF SPATIAL VARIABILITY FOR TRANSVERSE ELASTIC MODULUS OF SPRUCE WOOD …….. 8801
*Alireza Farajzadeh Moshtaghin, Steffen Franke, Thomas Keller, Anastasios Vassilopoulos*

COMPARISON BETWEEN A POLYNOMIAL CHAOS SURROGATE MODEL AND MARKOV CHAIN MONTE CARLO FOR INVERSE UNCERTAINTY QUANTIFICATION BASED ON AN ELECTRIC DRIVE TEST BENCH …….. 8809
*Philipp Glaser, Michael Schick, Kosmas Petridis, Vincent Heuveline*

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EVALUATION OF ANGULAR KINEMATICS OF LOWER LIMB AMPUTEES USING QUANTITATIVE FLUOROSCOPIC IMAGING

**Alex C. Breen**[1], **Mihai Dupac**[2]

[1] Anglo-European College of Chiropractic
Bournemouth, BH5 2DF, UK
e-mail: AlexBreen@aecc.ac.uk

[2] Department of Design and Engineering, Bournemouth University
Poole BH12 5BB, UK
e-mail: mdupac@bournemouth.ac.uk

**Keywords:** Fluoroscopy, Imaging, Angular Kinematics, Lower Limb, Prosthesis.

**Abstract.** *In the design of a functional prosthetic socket for lower limb amputees it is desirable to mimic the actions of a healthy limb as close as possible. While limiting the degree of freedom to which the residual limb is able to move in the prosthetic socket during ambulation (walking) is a prerequisite for an adequate fitting, with inadequate understanding of the specific anatomy of the patient including residuum changes after amputation, an overly tight socket fit can lead to pain, pressure ulcers and infections. Furthermore, changes in the initial alignment between the residual limb and socket have been shown to affect the pressure distribution within the socket.*

*The purpose of this study was to measure and report the angular kinematic data/variables of lower limb amputees under protocols designed to infer the kinematics within the socket during walking and static weight hold on the artificial limb. The study details the angular motion kinematics of the tibia within the residuum with and without the addition of a 5kg mass used to replicate forces applied to the residual limb while undergoing the swing phase of slow and fast walking. Using quantitative fluoroscopic imaging techniques, continuous frame by frame tracking of tibia and prosthesis have been recorded and measures of residual limb/tibia angular kinematic have been performed. The understanding of the angular motion of the tibia within the residuum conveys a new insight in socket design (comfort and performance considerations) and furnishes a straightforward view into the effects of surgical amputation and prosthetic fit techniques.*

## 1 INTRODUCTION

Clinical studies showing that a majority of people with lower-limb amputation experience socket instability, discomfort, or residual limb pain [8,9,13] indicate the need of a better design for a functional and stable prosthetic to enable individuals to maximize their quality of life by better engaging in a day by day physical activities.

Recent biomechanical studies [1] have determined that in walking and running the socket absorbs the impacts [12,21] and behaves as a variable stiffness device which store and release energy. Moreover, for the below-knee amputee running, relatively shorter aerial and swing times (-35%), and longer ground contact times (15%), which increase reaction forces in the socket have been observed. These biomechanical findings suggest that the dynamics of the stump inside the socket can lead to undesirable side effects [8].

While reflective markers placed at specific locations convey a good understanding of the joints and (prosthetic foot) kinematics [17,18] it gives few indication of the stump motion inside the socket [2,3]. Since the weight distribution insight the socket is playing an important role in socket design, a better understanding of the angular kinematics between the residual tibia and the prosthetic is absolutely essential [16]. To answer the need, the Quantitative Fluoroscopy (QF) technology developed at the Institute for Musculoskeletal Research and Clinical Implementation at the Anglo-European College of Chiropractic (AECC), which can perform semi-automated continuous motion analysis of multiple images of the osseous linkages [6,7], have been used for the analysis of angular kinematics between the residual bone and the socket.

Using a new coronal plane QF imaging protocol the angular kinematics of the residual limb (residuum) within the socket have been assessed during walking and static weight hold using the participant's prosthesis. The study details the residual limb angular kinematics with and without the addition of a 5kg mass used to replicate slow and fast walking. The understanding of the angular motion of the tibia within the residuum conveys a new insight in socket design (comfort and performance considerations) and furnishes a straightforward view into the effects of surgical amputation and prosthetic fit techniques.

## 2 METHODS

A protocol to measure the angular motion of the tibia while simulating the swing phase of gait was designed based on some studies reported in literature in [3,10,11]. The protocol using QF was approved in November 2012 by the National Research Ethics Service (REC reference: 13/SW/0248) due to the need for 'Approval for research involving ionizing radiation' [15], however, the effective dose received by any individual was minimal, with a maximum dose of received by any participant across the cohort was 5x10-6 mSv (in addition to the UK average annual background radiation dose of 2.7mSv). To simulate angular kinematics under slow walking, fast walking and static weight hold on the artificial limb gait the protocol requires the participants to apply load as a weight transfer from foot to foot, that is, rocking from one foot to another. A platform with handrails (Figure 1) has been used to level the participant(s) to the fluoroscope height where the angular kinematics images of the limb-prosthesis interface are recorded (Figure 2). The image intensifier was positioned as close as possible to the participant's amputated limb ensuring that magnification/distortion effects were minimal and the base of the prosthetic socket would not leave the image field at any time.

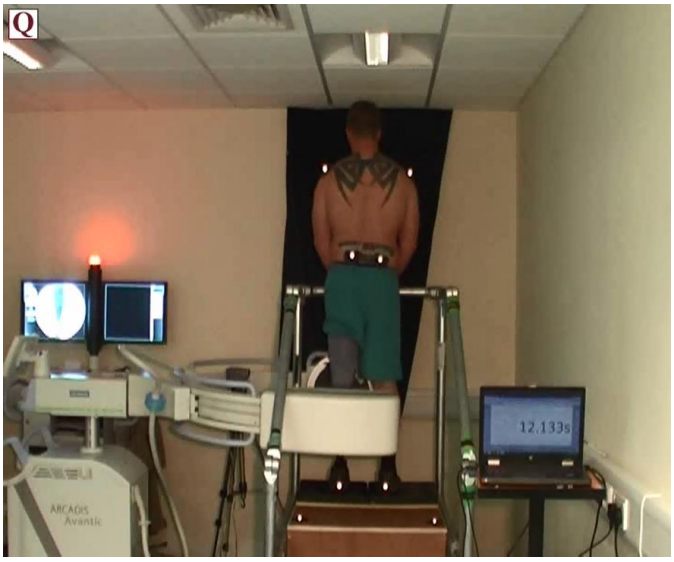



Figure 1: Depicting the stance of a unilateral trans-tibial amputee volunteer during residual limb fluoroscopy imaging procedure with and without the addition of a 5kg mass used to replicate slow and fast walking.

Residual limb image acquisition, which quantifies tibial motion within the socket while accounting for magnification effects [5], has been performed for 3 test configurations named:

i) Rocking from Foot to Foot (RFF): a gait cycle duration of 2 seconds was determined to be of optimal duration. This was designed to be as close as possible to the natural gait cycle and avoid image blur across the image sequence allowing accurate image tracking of the tibia. During this simulated gait cycle the participants were been asked to move their full weight from rock from foot to foot (using a metronome for guidance) during which 10 seconds of fluoroscopic imaging took place. The fluoroscopic recording acquired in the anterior-posterior (AP) view of the limb/prosthesis interface started once the participants confirmed their comfort with the motion. This allowed a recording of consecutive 5 gait cycles.

ii) Rocking with 5 kg mass on Prosthetic Limb (R5kgPL): The RFF procedure was repeated with the addition of a 5 kg mass to simulate the force applied during a slightly faster walking swing phase [14]. All participants mentioned that they are comfortable with the addition of the 5 kg mass which was attached to the base of the prosthetic limb.

iii) Static Weight-Bearing Hold (SWBH): carrying no additional mass weight the participants have been asked to slowly put their weight on their prosthetic limb and held it there for a short period of time (about 10 seconds).

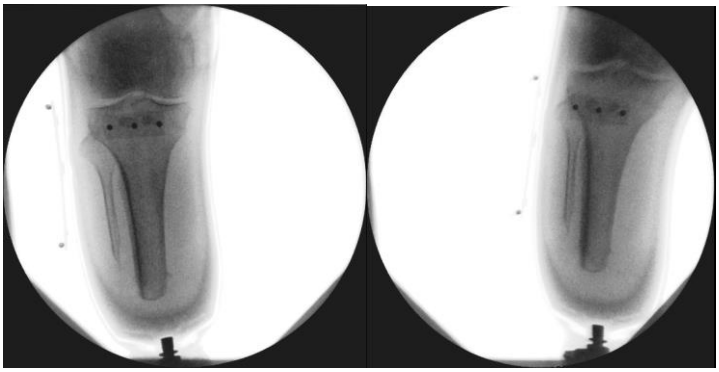


Figure 2: Example of successful tracking of tibial rotation during image acquisition

Two force plates placed on the platform were used to measure the body weight distribution during these testing protocols.

## 3 DATA ANALYSIS

In order to perform the data analysis, all the recorded images have been transferred from the fluoroscope to a dedicated computer for enhancement and analysis. Each fluoroscopic image sequence typically contained up to 150 individual DICOM images and was 350 megabytes in size. Individual image frames (Figure 3) were extracted from the sequence into '.jpg' format files using a JPEG compression performed in the Matlab environment.

Following enhancement of the images, templates have been manually placed around the tibia and metallic base of the prosthetic socket. These templates, called the 'reference templates', are used the verification process as well as a simplified representation of the tibia. The tracking algorithm makes note of the grayscale pixel information contained within each tracking template, as well as its location in the image. This information is then compared to the pixel data for the same location in the next frame. The template is then automatically moved both laterally and in rotation by small increments into locations near to its previous location and the process is repeated.

Via cross-correlation methods, each image in each position is compared to that of the previous image. This process is then repeated for each subsequent image and tracking template for that image sequence [6,7].



Figure 3: Individual image frames - examples of the variation in the residual tibia length and shape

The benefit of this process is its ability to quantify the location of the residual limb within the socket and to provide standardization of the measurements. However, there were some complications within this quantification process. While identification of the tibial outline and anatomical landmarks were found to be satisfactory, tracking of the tibia often failed as it moved though the image field at speed (i.e. shifting weight).

In the full loading and full traction positions, where the limb is momentarily motionless, the tibia was readily identifiable. However, between these two positions, when the tibia is moving at speed, motion blur reduces the likelihood of successful image tracking. To account for these difficulties, manual or discrete identification of tibia positions were necessary at extremes of motion. To achieve this, the original templates have been manually replaced in a corrected position before continuing automated tracking. This allowed the relative sizes and shapes of the reference templates to be maintained and to give consistent results.

Additional quality assurance procedures included inspection for image distortion, magnification and out of plane rotation of the prosthesis. Image distortions such as pin cushion effect were not observed, since this is automatically corrected by the image interface of the Siemens fluoroscope used (Siemens Arcadis Avantic – Siemens GMBH, Germany). To account for magnification errors in the fluoroscopic images of the prosthesis limb an imaging protocol for measuring the proximo-distal motion of the residual tibia within the soft tissues of the residual limb and estimating errors which may arise in these measurements have been considered [7].

Although the literature does not elaborate on tibial angulation in the coronal plane, previously conducted studies using sagittal or oblique imaging views have required adaptations to overcome the angular alignment change between the prosthesis and tibia as the tibia as it descends into the socket and soft tissue [3, 14,19,20]. In the course of this investigation it became apparent that such a method was needed and that it must be applicable across a range of tibia shapes and socket/suspension style combinations.

## 4 RESULTS

To standardize the results and to account for coronal plane rotation variations in the alignment of the residual limb and socket, a method for measuring the proximo-distal motion of the tibia was derived from similar radiographic techniques which imaged the residual limb in the sagittal plane [14,19]. The method considered the proximo-distal direction defined in the coronal plane passing through the center of the tibia, and represented by a line running from the center of the tibial plateau, perpendicular to it and extended to the bottom of the prosthetic socket (Figure 4). This compensate for asymmetries at the distal end of the tibia due to the surgical techniques used or to post-surgical bone growth.

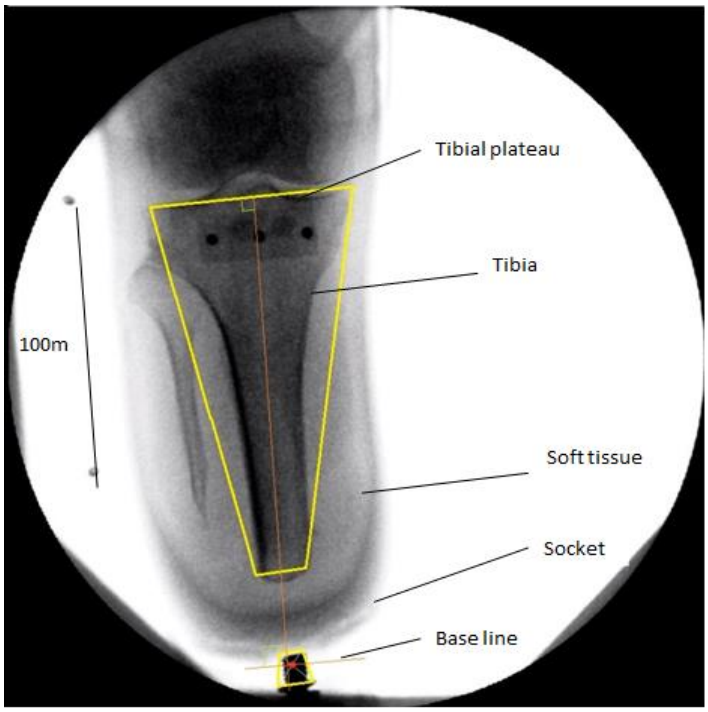



Figure 4: Tibia angular kinematics - a method to measure the angular changes

To account for angular changes, the socket base was identified as a line, perpendicular to the proximo-distal direction and passing through the center of the prosthesis base marker. The tibia angular kinematics as a function of body weight and static hold is shown in Figure 5. It can be seen that the addition of a 5 kg mass shown in Figure 5.b which is used to replicate forces applied to the residual limb while undergoing the swing phase of fast walking, affect the angular kinematics in Figure 5.a by an increase from 4.8 to 8.3 degrees. From Figure 5.c it can be also observed that the angular kinematics is minimal when a static hold is performed.

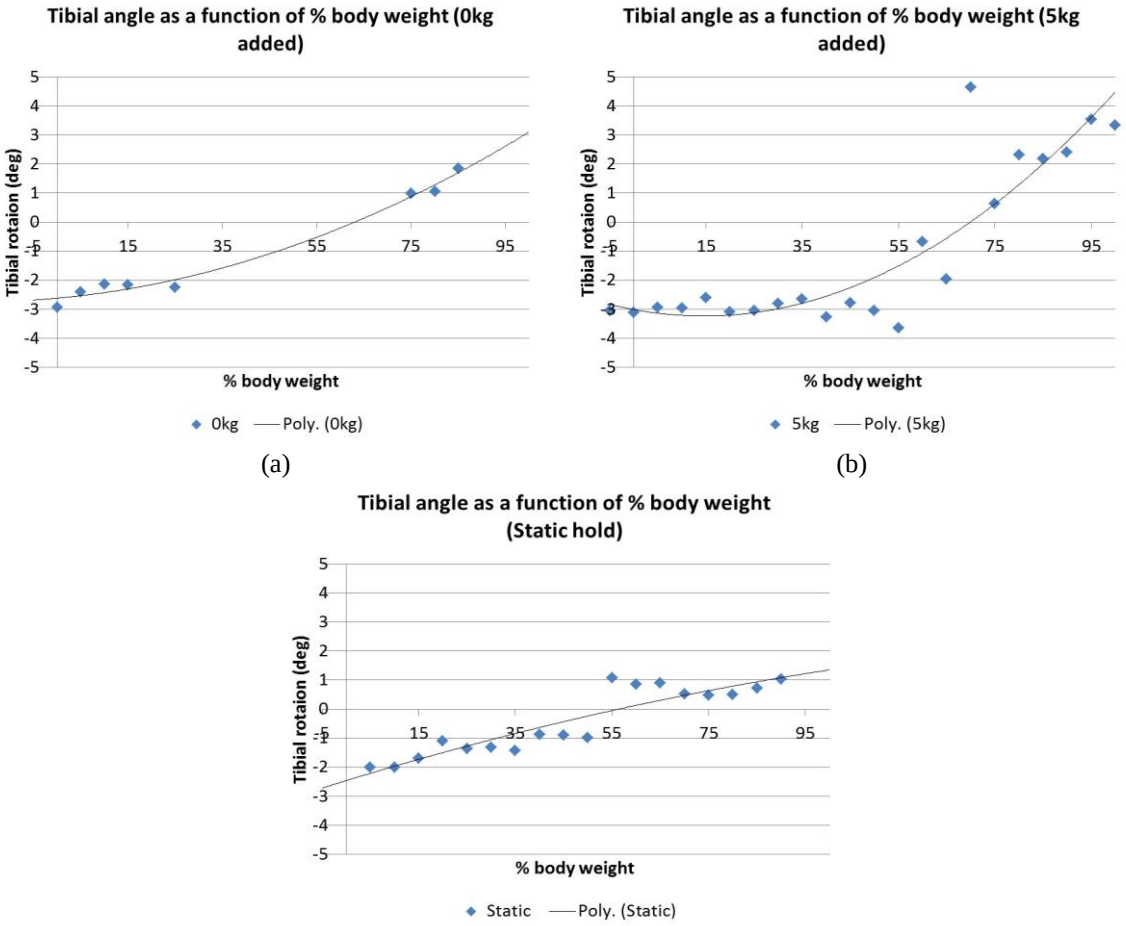



Figure 5: Angular kinematics of the prosthetic foot/tibia for (a) no weight attached, i.e., slow walking (b) 5 kg weight attached, i.e., fast walking, and (c) full body weight is slowly applied to residual limb, i.e., static hold

It was also observed (but not presented here) that the variation in the residual tibia length and shape shown in Figure 3 can affect both the initial angle of the tibia with respect to the base of the socket as well as the angular kinematics of the tibia inside the socket.

## 5 CONCLUSIONS

In this study the angular kinematics of lower limb amputees have been assessed for walking and static weight hold on the artificial limb using Quantitative Fluoroscopic imaging techniques. The study details the angular motion kinematics (maximal angular amplitude) of the tibia within the residuum with and without the addition of a 5kg mass used to replicate forces applied to the residual limb while undergoing the swing phase of slow and fast walking. The study, which provides a better understanding of the pressure distribution within the socket through its angular kinematics, conveys a new insight into socket design for comfort, performance and amputation techniques.

## REFERENCES


[1] R.W. Wirta, F.L. Golbranson, R. Mason, K. Calvo, Analysis of below knee suspension systems: Effect on gait, *Journal of Rehabilitation Research & Development*, 27, 385-396, 1990

[2] B. Soderberg, L. Ryd, B.M. Persson, Roentgen Stereophotogrammetric Analysis of Motion between the Bone and the Socket in a Transitbial Amputation Prosthesis: A Case Study, *Journal of Prosthetics and Orthotics*, **15(3)**, 95-99, 2003.

[3] C.J. Tucker, J.M. Wilken, D.J. Stinner, K.L. Kirk, A Comparison of Limb-Socket Kinematics of Bone-Bridging and Non-Bone-Briding Wartime Transtibial Amputations. *The Journal of Bone and Joint Surgery*, **94(10)**, 924-930, 2012.

[4] G. Papaioannou, C. Mitrogiannis, G. Nianios, G. Fiedler, Assessment of Internal and External Prosthesis Kinematics during Strenuous Activities Using Dynamic Roentgen Stereophotogrammetric Analysis, *Journal of Prosthetics and Orthotics*, **22(2)**, 91-105, 2010.

[5] A. Breen, M. Dupac, S. Noroozi, N. Osborne, A pre-operational study magnification measurement and error estimation of residual tibia kinematics within below knee prosthetics. *Proceedings of the 4th International Conference on Engineering Optimization*, CRC Press/Balkema, 927-930, Lisbon, Portugal, September 8-11, 2014.

[6] A. Breen, J. Muggleton, F. Mellor, An objective spinal motion imaging assessment (OSMIA): reliability, accuracy and exposure data, *BMC Musculoskeletal Disorders*, **7(1)**, 1-10, 2006.

[7] A.C. Breen, The measurement of the kinematics of the human spine using digital videofluoroscopy. Southampton: University of Southampton. *Research Report No. ME/89/32*, 1989.

[8] P.L. Ephraim, S.T. Wegener, E.J. MacKenzie, T. R. Dillingham, L.E. Pezzin, Phantom pain, residual limb pain, and back pain in amputees: results of a national survey. *Archives of Physical Medicine and Rehabilitation*, **86(10)**, 1910-1919, 2005.

[9] R. Gailey, K. Allen, J. Castles, J. Kucharik, M. Roeder, Review of secondary physical conditions associated with lower limb amputation and long-term prosthesis use. *Journal of Rehabilitation, Research and Development*, **45(1)**, 15-30, 2008.

[10] G. Gholizadeh, N.A. Abu Osman, M. Kamyab, A. Eshraghi, W.A.B. Wan Abas, M.N. Azam, Transtibial prosthetic socket pistoning: Static evaluation of Seal-In X5$^{R}$ and Dermo$^{R}$ Liner using mostion analysis system. *Clinical Biomechanics*, **27(1)**, 34-39, 2012.

[11] H. Gholizadeh, N.A.A. Osman, A.G. Luviksdottir, A. Eshraghi, M. Kamyab, W.A.B.W. Abas, A new approach for the pistoning measurement in transitbial prosthesis, *Prosthetics and Orthotics International*, **35(4)**, 360-364, 2011.

[12] S.K. Au, J. Weber, H. Herr, Powered Ankle–Foot Prosthesis Improves Walking Metabolic Economy, *IEEE Transactions on Robotics*, **25(1)**, 51-66, 2009

[13] G. Murdoch, A. Bennett-Wilson, *A Primer on Amputations and Artificial Limbs*. Charles C Thomas, 1998.

[14] H. Narita, K. Yokogushi, S. Shii, M. Kakizawa, T. Nosaka, Suspension effect and dynamic evaluaiton of the total surface bearing (TSB) trans-tibial prosthesis: a

comparison with the patellar tendon bearing (PTB) trans-tibial prosthesis. *Prosthetics and Orthotics International*, **21(3)**, 175-178, 1997.

[15] NHS, Service Development and Commissioning Directives: Chronic, Non-Malignant Pain. *Welsh Assembley Government Report*, 2008.

[16] G. Papaioannou, D. Tsiokos, G. Fiedler, C. Mitrogiannis, I. Avdeev, J. Wood, R. McKinney, Dynamic Radiography Imaging as a Tool in the Design and Validation of a Novel Intelligent Amputee Socket. *In:* Tavares, J. M. R. S., and Jorge, R. M. N., (Eds), eds. *Computational Vision and Medical Image Processing: Recent Trends.* Springer, 91-112, 2011.

[17] D. Pasparakis, , Darras, N., Normal walking principles, basic concepts, terminology 3-dimensional clinical gait analysis. *EEXOT*. **60(4)**, 183-194, 2009.

[18] L. Torburn, J. Perry, E. Ayyappa, S.L. Shanfield, Below-knee amputee gait with dynamic elastic response prosthetic feet: A pilot study, *Journal of Rehabilitation Research and Development*, **27(4)**, 369-384, 1990.

[19] Grevsten, S. and Eriksson, U., Stump-socket contact and skeletal displacement in a suction patellar-tendon bearing prosthesis. *Journal of Bone and Joint Surgery Am*, 56, 1692-1696, 1974.

[20] C.R. Bocobo, J.M. Castellote, D. MacKinnon, A. Gabrielle-Bergman, Videofluoroscopic evaluation of prosthetic fit and residual limbs following transtibial amputation. *Journal of Rehabilitation Research and Development*, **35(1)**, 6-13, 1998.

[21] T.M. Assman, Humanoid push recovery stepping in experiments and simulations, *Master's thesis, DC 2012.059*, Eindhoven University of Technology, Eindhoven, October 2012

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# PRELIMINARY STUDY TO INVESTIGATE THE EFFECT OF PISTON-LIKE AND ROCKING MOTIONS OF THE STAPES FOOTPLATE ON THE BASILAR MEMBRANE VIBRATION

**Philipp Wahl, Sebastian Ihrle, Pascal Ziegler, and Peter Eberhard**

Institute of Engineering and Computational Mechanics, University of Stuttgart,
Pfaffenwaldring 9, 70569 Stuttgart, Germany
{philipp.wahl, pascal.ziegler, peter.eberhard}@itm.uni-stuttgart.de

**Keywords:** Biomechanics, Cochlea, Basilar Membrane Vibration, Stapes Footplate, Piston-like Motion, Rocking Motion.

**Abstract.** *The inner ear or cochlea is a bone structure of spiral shape and is composed of mainly two conical chambers which are filled with fluid and separated by a soft membrane, the basilar membrane. In case of a healthy ear, the closed hydraulic system is excited through the vibration of the stapes. According to present hearing theory, this leads to pressure waves in the cochlear fluid which in turn results in the characteristic vibration behavior of the basilar membrane. Related to the sound frequency, hair cells in certain areas of the basilar membrane are stimulated and cause hearing nerve stimulation.*

*As reported in literature, the stapedial motion is mainly piston-like for low frequencies, whereas for higher frequencies rocking motions increasingly occur. Since purely rocking motions of the stapes footplate generate no net fluid displacement, several researches doubt that these motion components can lead to basilar membrane vibration and thus to hearing impression. Therefore, in this study a Finite Element model with simplified geometry of the human cochlea is developed using Eulerian-based acoustic elements to model the inner ear fluid. First the vibrations of the basilar membrane are calculated for a purely piston-like excitation mode. Then, these results are compared with the basilar membrane vibration pattern evoked by purely rocking motion around the short axis of the stapes footplate.*

## 1 INTRODUCTION

The inner ear or cochlea is the sensory organ where mechanical vibrations are transferred to nerve stimuli. It is a bone structure of spiral shape and is composed of conical chambers which are filled with fluid and separated by a soft membrane, the basilar membrane. At the apical end, the upper chamber, scala vestibuli, is connected to the lower chamber, scala tympani, through the helicotrema. At the basal end, the scala tympani is closed by the soft round window membrane, whereas the scala vestibuli is linked to the fluid-filled vestibulum. This in turn is connected to the middle ear through the stapes located at the top of the vestibulum. The cochlea can, therefore, be considered as a closed hydraulic system with fluid properties similar to water.

For a healthy ear, sound is received by the eardrum, transmitted through the middle ear ossicular chain and finally excites the hydraulic system through the vibration of the stapes footplate. The stapedial vibration leads to the propagation of pressure waves in the cochlear fluid which in turn results in the characteristic vibration of the basilar membrane. According to present hearing theory, the oscillation of the basilar membrane reaches a maximum amplitude at a characteristic location along the cochlear partition, which depends on the excitation frequency. This behavior is also called tonotopy. Particularly at this frequency-depending location, hair cells on the basilar membrane are stimulated and cause hearing nerve stimulation.

Experiments on temporal bones indicate, that the physiological motion of the stapes footplate is mainly piston-like for acoustic stimulation frequencies below $1\,\mathrm{kHz}$. For higher frequencies, the rocking motions along the long and short axis of the stapes footplate become increasingly dominant and the stapes footplate motion becomes complex [1, 2]. However, it is not clarified whether and to what extend these rocking motions are capable to evoke the characteristic basilar membrane vibration, because these rotational components lead to no net fluid displacement in the scala vestibuli [3]. Since the cochlea with its complex geometry represents a closed hydraulic system, the small vibrations of the basilar membrane and fluid pressure distributions are almost impossible to measure. Therefore, in this study a Finite Element (FE) model of the inner ear with simplified geometry is developed. First the basilar membrane vibration due to purely piston-like stapes footplate motion is investigated. Then, these results are compared with the basilar membrane vibration resulting from purely rocking motions of the stapes.

## 2 FE MODEL

In the following section, the geometry, material properties, and applied boundary conditions for the simplified FE model of the human cochlea are presented. Since the cochlear fluid is slightly compressible and represents an enclosed fluid volume with no mean flow, acoustic elements are used to describe the behavior of the inner ear fluid. The discretized equation for the pressure-formulated, linear acoustic elements as well as the formulation between acoustic and structural elements to describe the cochlear fluid-structure interactions are introduced.

### 2.1 Geometry, material properties, and boundary conditions

The FE model with its simplified geometry is shown in Figure 1. For preliminary investigations, the chambers are modeled cube-shaped and uncoiled. The fluid-filled scala vestibuli and scala tympani are divided through the basilar membrane, which is embedded in the rigid bony wall. At the apex, the chambers are connected through the helicotrema. At the base, the scala vestibuli is linked to the fluid-filled vestibulum, which interacts with the middle ear through the stapes footplate. The scala tympani, however, is closed by the round window membrane which is embedded in the lateral basal wall.

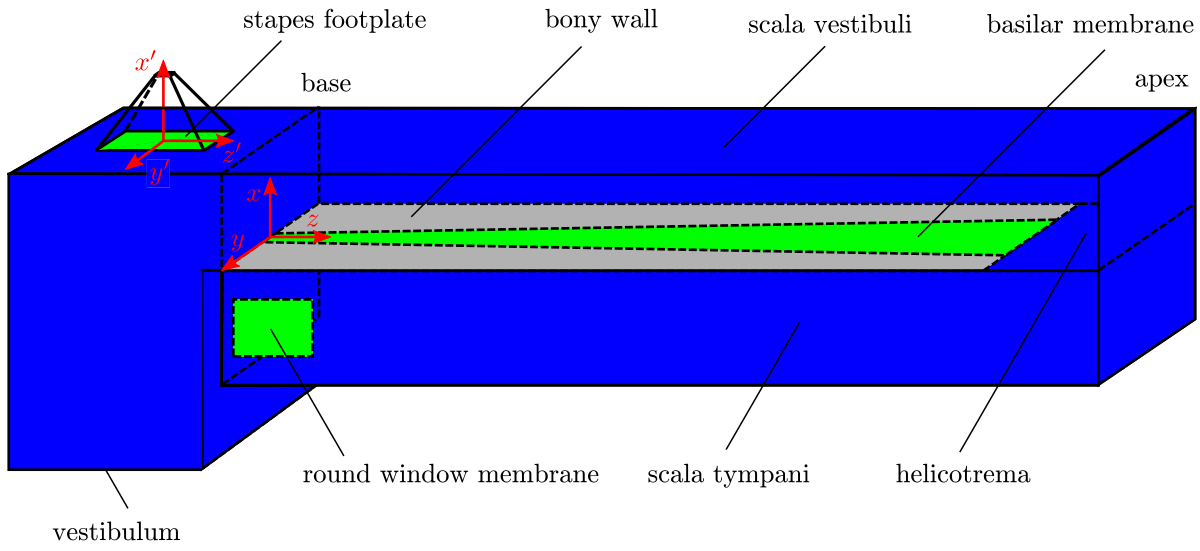



Figure 1: Model of the human cochlea with simplified geometry.

The dimensions of the cochlear fluid spaces and structural components are summarized in Table 1 and are gathered from anatomical data published in literature, from the evaluation of additional CT-data or from other cochlea FE models. The fluid chambers have a width of $1.37\,\mathrm{mm}$ and are slightly different in height, thus the fluid volume of the scala vestibuli is slightly lower compared to that of the scala tympani. The entire cochlea model has a length of $35.9\,\mathrm{mm}$. It should be noted, that the width of the basilar membrane increases from base towards the apex, in contrast its thickness decreases along the cochlear partition. Taking into account the area of the stapes footplate and round window membrane reported in literature, their shapes are assumed to be rectangular.

The material properties used for the FE model are listed in Table 2. The density, viscosity, and sound velocity of the inner ear fluid are in the same range as water [19]. As reported in literature, the basilar membrane comprises transversally orientated fibers which are embedded in an elastic ground substance. Therefore, the orthotropic material formulation described in [20] for a guinea pig is used fitting the Young's modulus to that of a human basilar membrane [13]. Further, a structural damping for the basilar membrane is applied with a modal damping of $\xi = 0.1$. For the round window membrane linear elastic material properties are assumed. The rims of the basilar and round window membrane have fixed translations. The stapes footplate and bony wall are represented as rigid bodies, the latter one is additionally fixed in space.

## 2.2 Coupling acoustic with structural elements for fluid-structure interaction

The FE model is created using ANSYS Mechanical APDL 15.0. The structural components are meshed with standard shell elements. For the fluid domain the pressure-formulated or Eulerian-based acoustic elements are used, referred to as FLUID30 in ANSYS. The acoustic wave equation for a fluid with viscosity $\eta_\mathrm{f}$, sound velocity $c_\mathrm{f}$, and density $\rho_\mathrm{f}$ is defined by

$$\nabla^2 p - \frac{1}{c_\mathrm{f}^2}\frac{\partial^2 p}{\partial t^2} + \frac{4\eta_\mathrm{f}}{3\rho_\mathrm{f} c_\mathrm{f}^2}\nabla^2\frac{\partial p}{\partial t} = 0. \tag{1}$$

The equation describes the spatial propagation of pressure waves in a fluid with no net flow rate and no source, neglecting effects of thermal conductivity. Further, a compressible fluid

Table 1: Dimensions of the cochlea model.

| | | | |
|---|---|---|---|
| *fluid spaces* | | | |
| vestibulum | volume | $50\,\mathrm{mm}^3$ | [4, 5] |
| scala vestibuli | volume | $35.3\,\mathrm{mm}^3$ | [6, 7, 8, 9, 10] |
| scala tympani | volume | $44.2\,\mathrm{mm}^3$ | [6, 7, 8, 9, 10] |
| helicotrema | area ($y$ x $z$) | $0.64\,\mathrm{mm}^2(1.37\,\mathrm{x}\,0.46\,\mathrm{mm})$ | [7] |
| *structural components* | | | |
| basilar membrane | length | $30.4\,\mathrm{mm}$ | [11] |
| | width basal | $0.1\,\mathrm{mm}$ | [12, 13, 5] |
| | width apical | $0.5\,\mathrm{mm}$ | [12, 13, 5] |
| | thickness basal | $7.5\,\mu\mathrm{m}$ | [12, 13, 14] |
| | thickness apical | $2.5\,\mu\mathrm{m}$ | [12, 13, 14] |
| bony wall | length | $30.4\,\mathrm{mm}$ | |
| | width | $1.37\,\mathrm{mm}$ | |
| | thickness basal | $7.5\,\mu\mathrm{m}$ | |
| | thickness apical | $2.5\,\mu\mathrm{m}$ | |
| round window membrane | area ($x$ x $z$) | $2.1\,\mathrm{mm}^2(1.04\,\mathrm{x}\,2\,\mathrm{mm})$ | [15] |
| | thickness | $65\,\mu\mathrm{m}$ | [16, 17] |
| stapes footplate | area ($y'$ x $z'$) | $3\,\mathrm{mm}^2(1.18\,\mathrm{x}\,2.5\,\mathrm{mm})$ | [18] |

where density changes are due to pressure variations is assumed and the equation is limited to relatively small pressures $p$, such that the changes in density are small compared to the mean density. By discretizing Eq. 1, the finite element equation for the pressure-formulated, linear acoustic element is

$$\mathbf{M}_\mathrm{f}\,\ddot{\mathbf{p}} + \mathbf{C}_\mathrm{f}\,\dot{\mathbf{p}} + \mathbf{K}_\mathrm{f}\,\mathbf{p} = \mathbf{F}_\mathrm{f}, \tag{2}$$

where $\mathbf{M}_\mathrm{f}$, $\mathbf{C}_\mathrm{f}$ and $\mathbf{K}_\mathrm{f}$ are the equivalent fluid mass, damping and stiffness matrices. Further, $\mathbf{F}_\mathrm{f}$ represents the vector of applied fluid loads and $\mathbf{p}$ the vector of the unknown nodal acoustic pressures.

To describe the interaction between the inner ear fluid and the stapes footplate, basilar membrane and round window membrane, the acoustic pressure is related to the resulting normal displacement of the structure, using a coupling matrix $\mathbf{R}$ that accounts for the effective surface area associated with each node on the interface. This leads to the equations of the two-way coupled fluid-structure interaction problem

$$\begin{aligned} \mathbf{M}_\mathrm{f}\,\ddot{\mathbf{p}} + \mathbf{C}_\mathrm{f}\,\dot{\mathbf{p}} + \mathbf{K}_\mathrm{f}\,\mathbf{p} &= \mathbf{F}_\mathrm{f} - \rho_\mathrm{f}\mathbf{R}^T\ddot{\mathbf{u}}, \\ \mathbf{M}_\mathrm{s}\,\ddot{\mathbf{u}} + \mathbf{C}_\mathrm{s}\,\dot{\mathbf{u}} + \mathbf{K}_\mathrm{s}\,\mathbf{u} &= \mathbf{F}_\mathrm{s} + \mathbf{R}\mathbf{p}. \end{aligned} \tag{3}$$

There, $\mathbf{M}_\mathrm{s}$, $\mathbf{C}_\mathrm{s}$ and $\mathbf{K}_\mathrm{s}$ are the structural fluid mass, damping and stiffness matrices, $\mathbf{F}_\mathrm{s}$ is the vector of applied structural loads and $\mathbf{u}$ is the vector of unknown nodal displacements. These equations can be formed into a matrix equation, where the assembled mass and stiffness matrices are unsymmetric. Using the linear pressure-formulated acoustic elements to model the fluid domain, the FLUID30-element in ANSYS has one pressure DOF and three additional displacement DOFs at the fluid-structure interface. However, elements inside the fluid domain have only pressure DOFs. For the FLUID30-elements located on the exterior surfaces of the scala vestibuli, scala tympani and vestibulum, the absorption coefficient is set to zero representing a rigid wall boundary condition.

Table 2: Material properties for the cochlea model.

| | | | |
|---|---|---|---|
| *cochlear fluid* | | | |
| density | $\rho_\mathrm{f}$ | $1000\,\mathrm{kg/m^3}$ | [19] |
| viscosity | $\eta_\mathrm{f}$ | $0.001\,\mathrm{Pa\,s}$ | [19] |
| sound velocity | $c_\mathrm{f}$ | $1500\,\mathrm{m/s}$ | [19] |
| *basilar membrane* | | | |
| Young's modulus | basal $E_{y,\mathrm{b}}$ | $50\,\mathrm{MPa}$ | [13] |
| | apical $E_{y,\mathrm{a}}$ | $3\,\mathrm{MPa}$ | [13] |
| | $E_x = E_z$ | $E_{y,\mathrm{a}}$ | [20] |
| shear modulus | $G_{xy} = G_{xz} = G_{yz}$ | $E_{y,\mathrm{a}}/2.99$ | [20] |
| Poisson ratio | $\nu_{xy} = \nu_{zy}$ | $0.49E_{y,\mathrm{a}}/E_y$ | [20] |
| | $\nu_{xz}$ | $1-(0.49E_{y,\mathrm{a}}/E_y)$ | [20] |
| modal damping | $\xi$ | $0.1$ | |
| *round window membrane* | | | |
| density | $\rho_\mathrm{rw}$ | $1200\,\mathrm{kg/m^3}$ | [13] |
| Young's modulus | $E_\mathrm{rw}$ | $0.35\,\mathrm{MPa}$ | [13] |
| Poisson ratio | $\nu_\mathrm{rw}$ | $0.3$ | [13] |

## 3 RESULTS

The cochlea model is excited by the motion of the stapes footplate. In this study, first the vibration of the basilar membrane due to purely piston-like motion of the stapes footplate along the $x'$ axis for $0.5, 1, 2$ and $4\,\mathrm{kHz}$ is investigated. Then, these results are compared at $1\,\mathrm{kHz}$ with the basilar membrane vibration pattern for purely rocking motion around the short axis $y'$ of the footplate. For both excitation modes, physiological amplitudes are used and applied to the rigid stapes footplate as kinematic boundary conditions, Table 3. For the following results, the transversal deflections in $x$-direction of the steady basilar membrane vibration are evaluated in the $xz$-plane.

Table 3: Displacement amplitude $x'$ (piston-like) and rotational amplitude $\beta_{y'}$ (rocking) applied to the stapes footplate to excite the cochlea model. The amplitudes are gained from physiological data [2] for an ear canal pressure of $1\,\mathrm{Pa}$.

| frequency | $0.5\,\mathrm{kHz}$ | $1\,\mathrm{kHz}$ | $2\,\mathrm{kHz}$ | $4\,\mathrm{kHz}$ |
|---|---|---|---|---|
| piston motion $x'$ [nm] | 1.80 | 38.20 | 7.96 | 1.99 |
| rocking motion $\beta_{y'}$ [$\mu$rad] | | 7.96 | | |

For an excitation frequency of $1\,\mathrm{kHz}$, the transversal deflections of the basilar membrane for three discrete time points of a cycle as well as the envelope of the membrane pattern is shown in Figure 2.

The piston-like motion of the stapes footplate leads to a vibration of the basilar membrane. Due to the fluid viscosity and the structural damping, adjacent partitions of the basilar membrane vibrate with an increasing delay in phase from base towards the apex. The spatially moving oscillation nodes result in a travelling wave of the basilar membrane. Thereby, the amplitude increases from the base towards the apex reaching a maximum amplitude at a char-

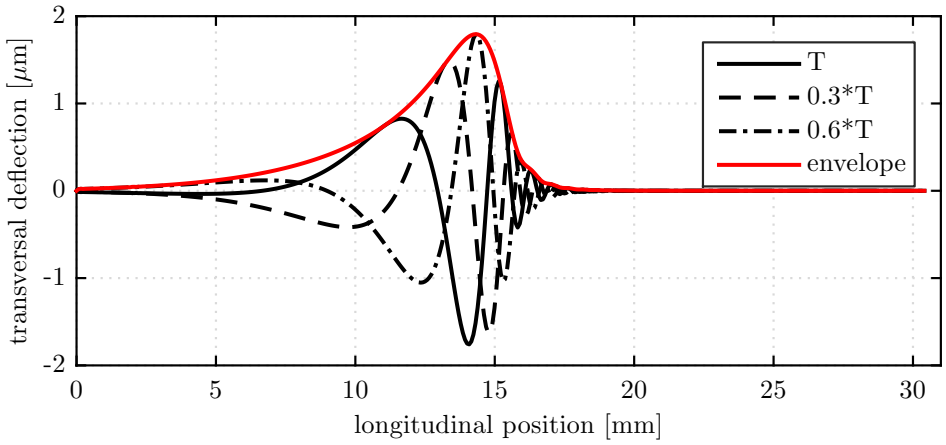



Figure 2: Vibration of the basilar membrane due to piston-like excitation for $x' = 38.2\,\mathrm{nm}$ at $1\,\mathrm{kHz}$.

acteristic point along the cochlea indicated by the envelope in Figure 2. Beyond this point, the amplitude decreases rapidly and the adjacent apical domain of the basilar membrane remains at rest. Further, the basilar membrane vibrations for the piston-like stapedial amplitudes according to Table 3 are calculated. The basilar membrane amplitudes for excitation frequencies of $0.5, 1, 2$ and $4\,\mathrm{kHz}$ are shown in Figure 3.

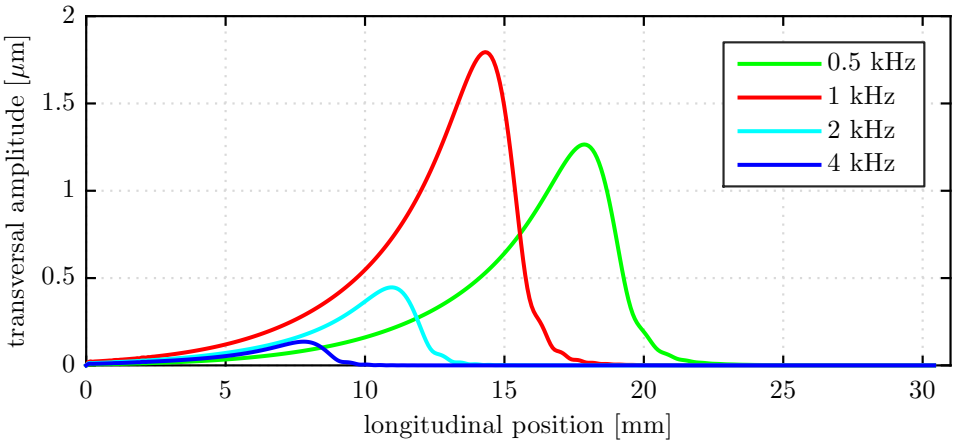



Figure 3: Envelopes of the basilar membrane vibration due to piston-like stapes footplate motion for different excitation frequencies.

By increasing the excitation frequency, the maximum amplitude is shifted towards the base of the cochlear partition. This behavior leads to a unique mapping of each excitation frequency to a distinct location along the basilar membrane. This tonotopic behavior as well as the vibration pattern of the basilar membrane are in agreement with the present hearing theory, described in Section 1. Comparing the maximum basilar membrane amplitudes for the different excitation frequencies, the global maximum amplitude occurs for an excitation frequency of $1\,\mathrm{kHz}$. Assuming, that the amplitude at the characteristic place is a measure for the perceived sound intensity, this result is consistent with the lowest hearing threshold level for human ears at around $1\,\mathrm{kHz}$ [21].

Finally, the cochlea model is excited by rocking motion around the short axis of the stapes

footplate with an amplitude of $\beta_{y'} = 7.96\,\mu\text{rad}$. The normalized envelope and the transversal deflections of the basilar membrane for one point of time of a cycle are shown in Figure 4.

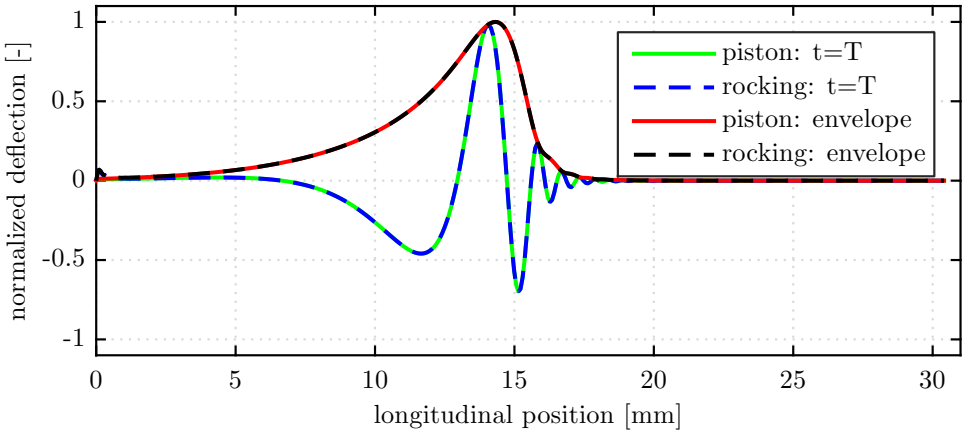



Figure 4: Vibrations of the basilar membrane at 1 kHz for purely piston-like and rocking motions of the stapes footplate.

For an excitation frequency of 1 kHz, the rocking motion of the stapes footplate leads to a vibration of the basilar membrane. The comparison with the normalized membrane deflections of Figure 2 shows, that the evoked basilar membrane pattern is the same as for the piston-like excitation mode. Obviously, the vibration pattern of the basilar membrane depends primarily on the excitation frequency, not on the mode of excitation. Except in the basal domain, slight differences in the vibration pattern are visible resulting from differences between the basal fluid pressure distributions, Figure 5.

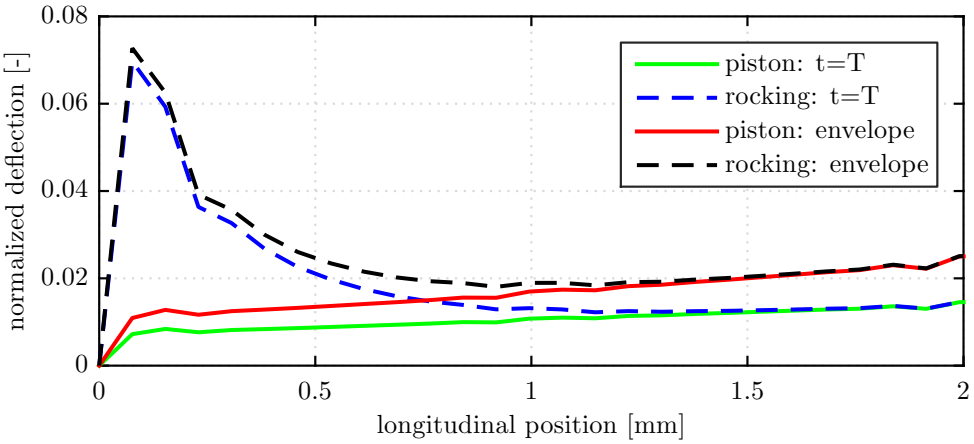



Figure 5: Vibrations of the basilar membrane in the basal domain for 1 kHz.

For the excitation frequency of 1 kHz, the maximum amplitude of the basilar membrane due to stapedial rocking motion is around 43 dB lower compared with that for the piston-like motion. Thus, for this excitation frequency the contribution of stapedial rocking motion to hearing impression is assumed to be low. However, at this frequency the amplitude of the piston-like motion of the stapes footplate is maximum and decreases considerably for higher

frequencies, Table 3. Therefore, further investigations are intended to study the vibration of the basilar membrane over the entire auditory frequency range for both modes of excitation.

## 4 CONCLUSION

In this study, an FE model of the human cochlea with a simplified geometry is developed. Using Eulerian-based acoustic elements to describe the inner ear fluid, the related equations for the two-way coupled fluid-structure interaction are described. For the piston-like motion of the stapes footplate, the FE model represents the characteristic basilar membrane vibration with a distinct maximum amplitude along the cochlear partition and the results are in agreement with the tonotopic hearing theory. It is further found that also purely rocking motion around the short axis of the stapes footplate is capable to evoke the characteristic basilar membrane vibration. Apart from slight differences in the basal domain, the resulting patterns of the basilar membrane for both excitation modes are identical. For an excitation frequency of $1\,\mathrm{kHz}$, the maximum basilar membrane amplitude due to rocking motion of the stapes footplate is around $43\,\mathrm{dB}$ lower than for the piston-like motion. However, according to physiological data the stapedial amplitude for the piston-like motion decreases for higher frequencies. Therefore, in a next step the basilar membrane vibrations for both excitation modes need to be studied for the entire auditory frequency range.

## REFERENCES

[1] Asai, M.; Roberson, J.B.; Goode, R.L.: Acoustic Effect of Malleus Head Removal and Tensor Tympani Muscle Section on Middle Ear Reconstruction. Laryngoscope, Vol. 107, pp. 1217–1222, 1997.

[2] Sim, J.H.; Chatzimichalis, M.; Lauxmann, M.; Röösli, C.; Eiber, A.; Huber, A.: Complex Stapes Motion in Human Ears. Journal of the Association for Research in Otolaryngology, Vol. 11, No. 3, pp. 329–341, 2010.

[3] Hato, N.; Stenfelt, S.; Goode, R.L.: Three-Dimensional Stapes Footplate Motion in Human Temporal Bones. Audiology & Neuro-Otology, Vol. 8, pp. 140–152, 2003.

[4] Tinz, L.: Drei-dimensionale Rekonstruktion des menschlichen Mittel- und Innenohres. Doctoral thesis, Technische Universität München, 2005.

[5] Koike, T.; Sakamoto, C.; Sakashita, T.; Hayashi, K.; Kanzaki, S.; Ogawa, K.: Effects of a Perilymphatic Fistula on the Passive Vibration Response of the Basilar Membrane. Hearing Research, Vol. 283, pp. 117–125, 2012.

[6] Braun, K.; Böhnke, F.; Stark, T.: Three-Dimensional Representation of the Human Cochlea Using Micro-Computed Tomography Data: Presenting an Anatomical Model for Further Numerical Calculations. Acta Otolaryngologica, Vol. 132, pp. 603–613, 2012.

[7] Penteado, C.: Geometric Reconstruction of a Human Cochlea With Basis on Micro-CT Images. Tech. rep., Institute of Engineering and Computational Mechanics, 2011.

[8] Wysocki, J.: Dimensions of the Human Vestibular and Tympanic Scalae. Hearing Research, Vol. 135, pp. 39–46, 1999.

[9] Thorne, M.; Salt, A.N.; DeMott, J.E.; Henson, M.M.; Henson, O.W.; Gewalt, S.L.: Cochlear Fluid Space Dimensions for Six Species Derived From Reconstructions of Resonance Images. Laryngoscope, Vol. 109, pp. 1661–1668, 1999.

[10] Igarashi, M.; Ohashi, K.; Ishii, M.: Morphometric Comparison of Endolymphatic and Perilymphatic Spaces in Human Temporal Bones. Acta Otolaryngologica, Vol. 101, pp. 161–164, 1986.

[11] Poznyakovskiy, A.A.; Zahnert, T.; Kalaidzidis, Y.; Lazurashvili, N.; Schmidt, R.; Hardtke, H.J.; Fischer, B.; Yarin, Y.M.: A Segmentation Method to Obtain a Complete Geometry Model of the Hearing Organ. Hearing Research, Vol. 282, pp. 25–34, 2011.

[12] Kwacz, M.; Marek, P.; Borkowski, P.; Mròwka, M.: A Three-Dimensional Finite Element Model of Round Window Membrane Vibration Before and After Stapedotomy Surgery. Biomechanics and Modeling in Mechanobiology, Vol. 12, pp. 1243–1261, 2013.

[13] Gan, R.Z.; Reeves, B.P.; Wang, X.: Modeling of Sound Transmission From Ear Canal to Cochlea. Annals of Biomedical Engineering, Vol. 35, pp. 2180–2195, 2007.

[14] Böhnke, F.; Arnold, W.: 3D-Finite Element Model of the Human Cochlea Including Fluid-Structure Couplings. Journal for Oto-Rhino-Laryngology, Vol. 61, pp. 305–310, 1999.

[15] Atturo, F.; Barbara, M.; Rask-Andersen, H.: Is the Human Round Window Really Round? An Anatomic Study With Surgical Implications. Otology & Neurotology, Vol. 35, pp. 1354–1360, 2014.

[16] Carpenter, A.M.; Muchow, D.; Goycoolea, M.V.: Ultrastructural Studies of the Human Round Window Membrane. Archives of Otolaryngology- Head and Neck Surgery, Vol. 115, pp. 585–590, 1989.

[17] Becvarovski, Z.: Absorption of Intratympanic Topical Antibiotics. Ear Nose Throat Journal, Vol. 83, pp. 18–19, 2004.

[18] Sim, J.H.; Röösli, C.; Chatzimichalis, M.; Eiber, A.; Huber, A.: Characterization of Stapes Anatomy: Investigation of Human and Guinea Pig. Journal of the Association for Research in Otolaryngology, Vol. 14, pp. 159–173, 2013.

[19] Baumgart, J.: The Hair Bundle: Fluid-Structure Interaction in the Inner Ear. Doctoral thesis, Technische Universität Dresden, 2010.

[20] Fleischer, M.; Schmidt, R.; Gummer, A.W.: Compliance Profiles Derived From a Three-Dimensional Finite-Element Model of the Basilar Membrane. The Journal of the Acoustical Society of America, Vol. 127, pp. 2973–2991, 2010.

[21] Lenarz, T.; Boenninghaus, H.G.: Hals-Nasen-Ohren-Heilkunde. Berlin Heidelberg: Springer, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# IN-HOUSE MULTIBODY HUMAN MODEL BASED ON EULER PARAMETERS FOR THE FAST IMPACT SCENARIO CALCULATION

**Jan Špička[1], Michal Hajžman[2], and Tomasz Bonkowski[2]**

[1]New Technologies – Research Centre
University of West Bohemia
Univerzitní 8, Pilsen, 306 14, Czech Republic
e-mail: spicka@ntc.zcu.cz

[2] European Centre of Excellence, NTIS –
– New Technologies for Information Society
Faculty of Applied Sciences
University of West Bohemia
Univerzitní 8, Pilsen, 306 14, Czech Republic
e-mail: mhajzman@kme.zcu.cz, tomasz@ntis.zcu.cz

**Keywords:** Multibody dynamics, Human body model, Euler parameters, Population diversity.

**Abstract.** *Purpose of this paper is to build an in-house spatial model of a human body based on multibody approaches, that could be utilized in numerical simulations of various impact scenarios. The model currently contains 17 rigid bodies and 16 spherical joints. The joints are modelled as range-unlimited in order to keep model robust for the further analysis. The limits of the motion will be further modelled within internal stiffness. However the approach of the spatial joints with unlimited motion suffers with so-called singular positions and thus the numerical calculation can loose its stability. To solve this problem, the concept of four Euler parameters instead of widely used three Euler angles is used here. The dimensions and masses of the particular segments are based on the real data of the Czech population measurement in 80s. The model is parametrized with the height and weight of the particular human. All the dimensions are calculated with respect to these values, which makes the model scalable to capture human population diversity. The contact problem of the human body with the external shape or self-contact between segments is solved using the continuous contact force model, where normal and tangential forces are evaluated as a function of the penetration, stiffness and coefficients of restitution. Algorithm for ellipsoid-to-plane contact is applied here. MATLAB software is used to build the in-house software for a human body model. The model will be further validated to capture the real human kinematics and to achieve human-like behaviour under different impact conditions.*

# 1 INTRODUCTION

Modelling of the human body is an expanding discipline of biomechanical research. With the advantages of the fast computer processing, the researchers have started to use computers for the biomechanical modelling more often. The goals ate to investigate numerical models of the human by different approaches for different situations. The example of the highly expanding area of human modelling is safety/automotive field, where virtual models of the human body or the parts of the body are often used to predict or eliminate fatal injuries or death. For this purpose, the detail and deformable finite element modelling method is usually applied. Such complex models of particular tissues are widely developed. With the proper description of materials and geometry, these models can serve to calculate deformations, velocities, accelerations, stresses or rupture of any human segments. On the other hand, there might be some cases, where such complex description are inefficient or inappropriate. If we are interested only in global behaviour (kinematics and dynamics) of the system (human body) the detailed description within finite elements are inefficient and time and cost consuming. Let us imagine a simple example of a man pedalling his bike. If we consider the man and the bike as one mechanical system and the goal is to describe the kinematics of the bike based on position of the legs, the knowledge of stresses and deformations of the segments or internal organs is not required. For such purpose the classical description of mechanisms can be enough to capture the motion. Moreover, the complex description can be considered with the articulated rigid body approach. If the researchers are interested in muscle forces, the musculoskeletal modelling method is an appropriate tool. Such model consists of constrained rigid bodies (articulated rigid body, multibody system) and the model of muscles. On the other hand, the requirements of the detailed tissue injury and deformations always require application of the finite element modelling.

Purpose of this work is to build an in-house software for fast simulations of the global behaviour of the scalable human body under different loading scenarios. Authors use the multibody (articulated rigid bodies) approach for the description of the human body. Such method allows to calculate the motion under various loading in a real time simulation (or close to real time). Authors replaced the well known Euler angles with the principle of Euler parameters. The Euler angles is a widely used method for description of the spatial rotational motion of the rigid body. However, this approach can suffer from the singular positions [1] , when the angle of nutation is close to $\pm i\pi, i \in < 0, 1, 2... >$. To solve this problem, the principle of Euler parameters [2, 3, 4, 5], that can be used without the singularities, was applied here.

# 2 METHODS

## 2.1 Euler parameters

Four Euler parameters are used for the parametrization of the 3 degrees-of-freedom spatial (spherical) motion of the rigid body. Therefore, these parameters are not totally independent and the constraint equation between Euler parameters can be define as

$$\sum_{k=1}^{4} (\beta_k)^2 = 1, \tag{1}$$

or written using vector of Euler parameters $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^T \boldsymbol{\beta} - 1 = 0. \tag{2}$$

When using e.g. Lagrange equations for the formulation of system dynamics, the second derivative of the constraint equations is further added to the equation of motion. Thus the Eq. 2 is twice

differentiated and one obtain

$$\boldsymbol{\beta}^T\ddot{\boldsymbol{\beta}} = -\dot{\boldsymbol{\beta}}^T\dot{\boldsymbol{\beta}}. \tag{3}$$

### 2.1.1 Euler parameters in terms of Euler angles

Euler parameters define the 3 degrees-of-freedom rotation usually described using three Euler angles (precession $\psi$, nutation $\theta$ and spin $\phi$). Since the Euler parameters can be hard to interpret in the physical meaning, it is useful to define Euler parameters as functions of well known Euler angles as

$$\beta_1 = \cos\left[\frac{1}{2}(\psi+\phi)\right]\cos\frac{\theta}{2}, \tag{4}$$

$$\beta_2 = \cos\left[\frac{1}{2}(\psi-\phi)\right]\sin\frac{\theta}{2}, \tag{5}$$

$$\beta_3 = \sin\left[\frac{1}{2}(\psi-\phi)\right]\sin\frac{\theta}{2}, \tag{6}$$

$$\beta_4 = \sin\left[\frac{1}{2}(\psi+\phi)\right]\cos\frac{\theta}{2}, \tag{7}$$

and similarly the derivatives of Euler angles as a function of Euler velocities

$$\dot{\beta}_1 = -\sin\left[\frac{1}{2}(\psi+\phi)\right]\ \frac{1}{2}(\dot{\psi}+\dot{\phi})\cos\frac{\theta}{2} - \cos\left[\frac{1}{2}(\psi+\phi)\right]\sin\frac{\theta}{2}\ \frac{\dot{\theta}}{2}, \tag{8}$$

$$\dot{\beta}_2 = -\sin\left[\frac{1}{2}(\psi-\phi)\right]\ \frac{1}{2}(\dot{\psi}-\dot{\phi})\sin\frac{\theta}{2} + \cos\left[\frac{1}{2}(\psi-\phi)\right]\cos\frac{\theta}{2}\ \frac{\dot{\theta}}{2}, \tag{9}$$

$$\dot{\beta}_3 = \cos\left[\frac{1}{2}(\psi-\phi)\right]\ \frac{1}{2}(\dot{\psi}-\dot{\phi})\sin\frac{\theta}{2} + \sin\left[\frac{1}{2}(\psi-\phi)\right]\cos\frac{\theta}{2}\ \frac{\dot{\theta}}{2}, \tag{10}$$

$$\dot{\beta}_4 = \cos\left[\frac{1}{2}(\psi+\phi)\right]\ \frac{1}{2}(\dot{\psi}+\dot{\phi})\cos\frac{\theta}{2} - \sin\left[\frac{1}{2}(\psi+\phi)\right]\sin\frac{\theta}{2}\ \frac{\dot{\theta}}{2}. \tag{11}$$

## 2.2 Equations of motion

### 2.2.1 Equations of motion of unconstrained bodies

Equations of motion of one unconstrained (free) body were defined based on the notation shown in [4]. Here, the body has 7 coordinates (three translational coordinates and four Euler parameters). The dynamic equations of motion of such unconstrained body can be define as

$$\underbrace{\begin{bmatrix}\mathbf{m}^i_{RR} & \mathbf{0}\\ \mathbf{0} & \mathbf{m}^i_{\beta\beta}\end{bmatrix}}_{\mathbf{M}^i}\underbrace{\begin{bmatrix}\ddot{\mathbf{R}}^i\\ \ddot{\boldsymbol{\beta}}^i\end{bmatrix}}_{\mathbf{q}^i} = \underbrace{\begin{bmatrix}(\mathbf{Q}^i_e)_R\\ (\mathbf{Q}^i_e)_\beta\end{bmatrix}}_{\mathbf{Q}^i_e} + \underbrace{\begin{bmatrix}\mathbf{0}\\ (\mathbf{Q}^i_\nu)_\beta\end{bmatrix}}_{\mathbf{Q}^i_\nu}, \qquad \text{i=1, 2, ...,n,} \tag{12}$$

where $\mathbf{M}^i$ is the mass matrix of $i$-th body, $\mathbf{q}^i$ is the vector of 7 generalized coordinates of the body and $(\mathbf{Q}^i_e)_R$ and $(\mathbf{Q}^i_e)_\beta$ are the vectors of generalized external forces associated with the generalized translations and orientations, respectively. Vector $\mathbf{Q}^i_\nu$ is so called quadratic velocity vector [4].

Equation (12) is the matrix equation that governs the unconstrained motion of the rigid body. This equation was simplified for the centroidal case, i.e when the origin of the local body-fixed coordinate system is rigidly attached to the centre of mass. When the sub-matrices are expressed in terms of the vector of Euler parameters, the overall equations of motion of the unconstrained body can be written in the following form

$$\underbrace{\begin{bmatrix} m^i\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}^i_{cog}\boldsymbol{G}^i \\ 0 & \boldsymbol{\beta}^{iT} \end{bmatrix}}_{\mathbf{M}^i} \underbrace{\begin{bmatrix} \ddot{\mathbf{R}}^i \\ \ddot{\boldsymbol{\beta}}^i \end{bmatrix}}_{\mathbf{q}^i} = \underbrace{\begin{bmatrix} \sum_{j=1}^{n_F} \boldsymbol{F}^i_j \\ \boldsymbol{G}^{iT}[\sum_{k=1}^{n_M} \boldsymbol{M}^i_j + \sum_{j=1}^{n_F} \boldsymbol{u}^i_k \times \boldsymbol{F}^i_k] \end{bmatrix}}_{\mathbf{Q}^i_e} + \underbrace{\begin{bmatrix} \mathbf{0} \\ -\begin{bmatrix} \boldsymbol{I}^i_{cog}\dot{\boldsymbol{G}}^i\dot{\boldsymbol{\beta}}^i - (\boldsymbol{G}^i\dot{\boldsymbol{\beta}}^i \times \boldsymbol{I}^i_{cog}\boldsymbol{G}^i\dot{\boldsymbol{\beta}}^i) \\ \dot{\boldsymbol{\beta}}^{iT}\dot{\boldsymbol{\beta}}^i \end{bmatrix} \end{bmatrix}}_{\mathbf{Q}^i_\nu}, \qquad \text{i=1, 2, ...,n,} \tag{13}$$

where

$$\boldsymbol{I}_{cog} = \boldsymbol{A}\overline{\boldsymbol{I}}_{cog}\boldsymbol{A}^T \tag{14}$$

is the inertia matrix defined in the global coordinate system, in which matrix $\boldsymbol{A}$ is the direction cosine matrix (transformation matrix between local and global coordinate systems) and is defined as

$$\boldsymbol{A} = \begin{bmatrix} 1-2\beta_3^2-2\beta_4^2 & 2(\beta_2\beta_3-\beta_1\beta_4) & 2(\beta_2\beta_4+\beta_1\beta_3) \\ 2(\beta_2\beta_3+\beta_1\beta_4) & 1-2\beta_2^2-2\beta_4^2 & 2(\beta_3\beta_4-\beta_1\beta_2) \\ 2(\beta_2\beta_4-\beta_1\beta_3) & 2(\beta_3\beta_4+\beta_1\beta_2) & 1-2\beta_2^2-2\beta_3^2 \end{bmatrix}. \tag{15}$$

Matrices $\boldsymbol{G}$ and $\overline{\boldsymbol{G}}$ are defined (in the global and the local coordinate system, respectively) as

$$\boldsymbol{G} = 2\begin{bmatrix} -\beta_2 & \beta_1 & -\beta_4 & \beta_3 \\ -\beta_3 & \beta_4 & \beta_1 & -\beta_2 \\ -\beta_4 & -\beta_3 & \beta_2 & \beta_1 \end{bmatrix} \tag{16}$$

and

$$\overline{\boldsymbol{G}} = 2\begin{bmatrix} -\beta_2 & \beta_1 & \beta_4 & -\beta_3 \\ -\beta_3 & -\beta_4 & \beta_1 & \beta_2 \\ -\beta_4 & \beta_3 & -\beta_2 & \beta_1] \end{bmatrix}. \tag{17}$$

Symbol $m^i$ in equation (13) is mass of $i$-th body, $\boldsymbol{I}$ is 3x3 identity matrix, $\boldsymbol{F}_i$ is the vector of applied external forces, $\boldsymbol{u}_k$ is the position vector of point P, where an external force acts, $n_F$ is the total number of external forces, $\boldsymbol{M}_i$ is the vector of applied external torques and $n_M$ is the total number of the torques.

### 2.2.2 Equations of motion of constrained bodies

Let us consider a multibody system containing $N$ bodies. The bodies are interconnected using kinematic joints. The kinematic constraints (equations) of the joints can be defined as

$$\boldsymbol{C}(\boldsymbol{q}, t) = \mathbf{0}. \tag{18}$$

The overall equations of motion of the kinematic system containing $N$ bodies can be written as

$$\begin{bmatrix} \mathbf{M}^1 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C}_{\mathbf{q}^1}^T \\ \mathbf{0} & \mathbf{M}^2 & \cdots & \mathbf{0} & \mathbf{C}_{\mathbf{q}^2}^T \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & 0 & \cdots & \mathbf{M}^N & \mathbf{C}_{\mathbf{q}^N}^T \\ \mathbf{C}_{\mathbf{q}^1} & \mathbf{C}_{\mathbf{q}^2} & \cdots & \mathbf{C}_{\mathbf{q}^N} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \ddot{\mathbf{q}}^1 \\ \ddot{\mathbf{q}}^2 \\ \vdots \\ \ddot{\mathbf{q}}^N \\ -\boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_e^1 + \mathbf{Q}_\nu^1 \\ \mathbf{Q}_e^2 + \mathbf{Q}_\nu^2 \\ \vdots \\ \mathbf{Q}_e^N + \mathbf{Q}_\nu^N \\ \mathbf{Q}_d \end{bmatrix} \tag{19}$$

or written in a more condensed form as

$$\begin{bmatrix} \mathbf{M} & \mathbf{C}_q^T \\ \mathbf{C}_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \ddot{\boldsymbol{q}} \\ -\boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_e + \mathbf{Q}_\nu \\ \mathbf{Q}_d \end{bmatrix}. \tag{20}$$

Vector $\boldsymbol{C_q}$ is a Jacobian matrix that is defined using derivatives of constraint equation (18) with respect to generalized coordinates $\boldsymbol{q}$ as

$$\boldsymbol{C_q} = \begin{bmatrix} \frac{\partial C_x}{\partial q_1} & \cdots & \frac{\partial C_x}{\partial q_N} \\ \frac{\partial C_y}{\partial q_1} & \cdots & \frac{\partial C_y}{\partial q_N} \\ \frac{\partial C_z}{\partial q_1} & \cdots & \frac{\partial C_z}{\partial q_N} \end{bmatrix}. \tag{21}$$

Constraint equation (18) and consequently equation (21) depend on the type of the particular kinematic joint. Vector $\boldsymbol{Q}_d$ is the vector that absorbs quadratic terms in the velocities and it is defined as

$$\mathbf{Q}_d = \begin{bmatrix} -\sum_{\alpha=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 \boldsymbol{C}_1}{\partial q_j \partial q_\alpha} \dot{q}_j \dot{q}_\alpha \\ -\sum_{\alpha=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 \boldsymbol{C}_2}{\partial q_j \partial q_\alpha} \dot{q}_j \dot{q}_\alpha \\ \vdots \\ -\sum_{\alpha=1}^{N}\sum_{j=1}^{N} \frac{\partial^2 \boldsymbol{C}_r}{\partial q_j \partial q_\alpha} \dot{q}_j \dot{q}_\alpha \end{bmatrix}, \tag{22}$$

in which $r$ is the number of constraints in the model.

### 2.3 Contact treatment

The goal of this work is to develop the human body model, or external shape of the body, respectively, that consists of spheres and ellipsoids linked in the open kinematic tree structure. This model is built on purpose simulations of the global behaviour of the human body under external loading including contact with some infrastructure or self-contact between bodies. Despite the fact that there are no limitations for the shape of the infrastructure, we can always approximate such shape with the finite number of planes. Thus, only the contact between an ellipsoid and a plane is solved. In authors' previous work [6] the algorithm for the solution of the ellipsoid-to-plane contact was presented. The main idea of this method lays in the creation of a new plane, parallel to the initial one and tangential to the ellipsoid, see Figure 1. When common point C of the new plane and of the ellipsoid is detected, the distance between this point and the initial plane can be calculated.

In case of collision the generated contact forces are calculated based on the continuous contact force model [7] as

$$\mathbf{F}_{ij}^c = -K_n \delta_n \mathbf{n}_c - C_n(\mathbf{v}_c \mathbf{n}_c)\mathbf{n}_c + \mathbf{F}_{ij}^t \quad \text{(normal force)}, \tag{23}$$

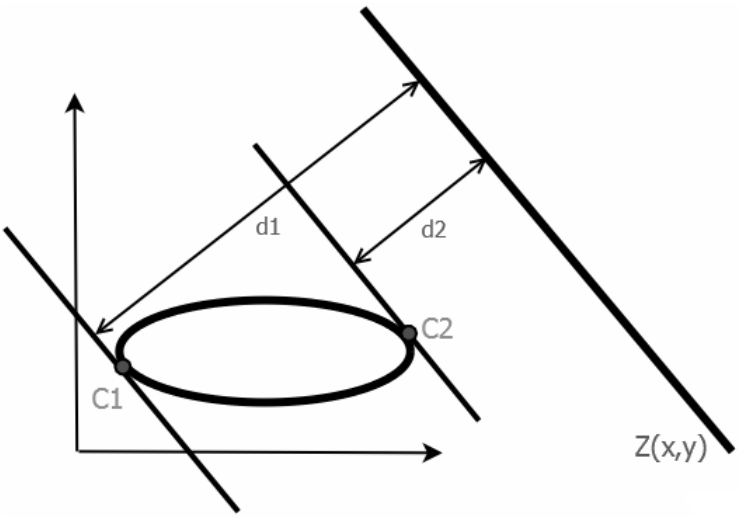



Figure 1: Ellipsoid and parallel planes.

$$\mathbf{F}_{ij}^{t} = -K_t \mathbf{v}_c^t + C_t(\mathbf{v}_c \times \mathbf{n}_c) \times \mathbf{n}_c \quad \text{(tangential force)}, \tag{24}$$

$$\mathbf{M}_{ij}^{c} = \mathbf{R}_{ij}^{C} \times \mathbf{F}_{ij}^{c} \quad \text{(torque of the contact force)}. \tag{25}$$

The contact force expressions (23), (24) and (25) are functions of distance (penetration) $\delta_n$, relative normal velocity $\mathbf{v}_c$ collinear with vector of normal $\mathbf{n}_c$ and tangential velocity vector $\mathbf{v}_c^t$ and the contact parameters; i.e. virtual normal and tangential stiffness $K_n$ and $K_t$ respectively, and virtual coefficients of restitution $C_n$ and $C_t$. Vector $\mathbf{R}_{ij}^{C}$ is the position vector of the contact point defined in the global coordinate system.

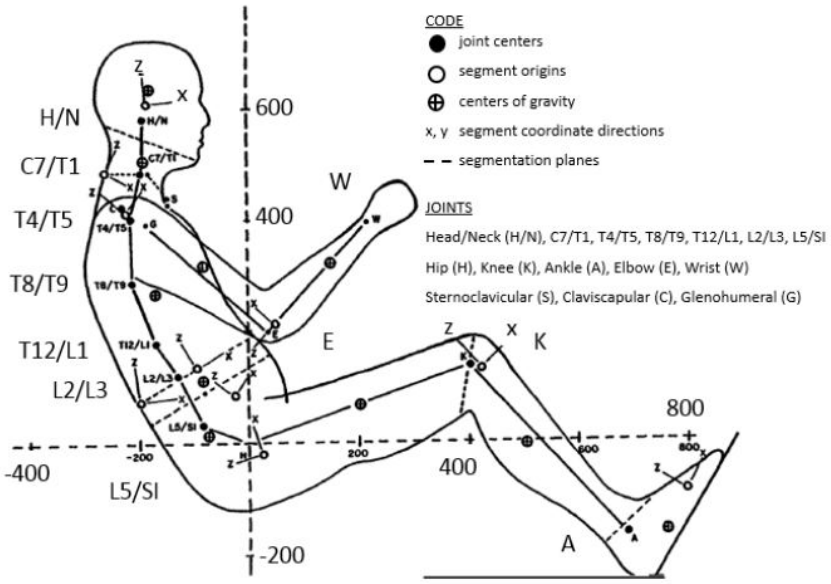



Figure 2: Segmentation of a human body [8].

## 2.4 Human body model

The basic concept of this work is in the segmentation of the human body defined by Robbins [8]. Human body is divided into the segments, whose motion respects correct kinematics and dynamics of the human body caused be external loading, see Figure 2.

Based on this segmentation, the spatial human body model was created as an open kinematic tree-structure containing 17 rigid bodies. Hynčík [8] suggested the simple 11-bodies structure of the 2D model, as is shown in Figure 3, where all bodies are modelled as ellipses or circles, respectively.

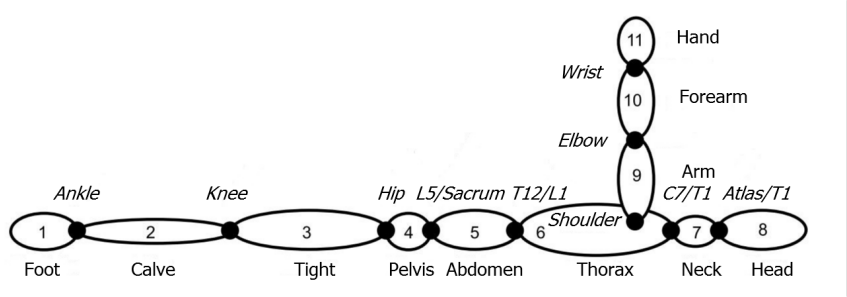



Figure 3: Segmentation of 2D human body model [8].

The spatial segmentation is expanded by the pairs of extremities and thus the model has 17 rigid bodies, based on Figure 2 and Figure 3:

- Pelvis
- Abdomen
- Thorax
- Neck
- Head
- Left and right arms
- Left and right forearms
- Left and right hands
- Left and right tights
- Left and right calves
- Left and right feet

The connections between bodies are realized by means of 16 spherical joints, representing real human's joints:

- Joint between abdomen and pelvis, i.e. vertebrae L5/Sacrum
- Joint between thorax and abdomen, i.e. vertebrae T12/L1
- Joint between neck and thorax, i.e. vertebrae C7/T1
- Joint between head and neck, i.e. vertebrae Atlas/T1
- Left and right shoulder joints
- Left and right elbow joints
- Left and right wrist joints
- Left and right hip joints
- Left and right knee joints
- Left and right ankle joints

### 2.4.1 Mass distribution, geometry and inertia properties

The proposed model is developed in the way that the total mass and the total height of the human are inputs of the in-house software. The dimensions and masses of particular segments are calculated based on these defined values. The masses of the body segments can be calculated from the total mass and height of the human [9, 10]. The method of Zaciorsky and Salujanov [10] is based on the radioisotope method of experimental measuring on 100 humans and defined coefficients for each body segments. The mass of the segments can be expressed using the simple equation as

$$m_i = \beta_{0i} + \beta_{1i} m + \beta_{2i} v, \tag{26}$$

where $m$ [kg] is the total mass of the body, $v$ [cm] is weight of body and coefficients $\beta$ are defined in the Table 1.

| **Name of the segment** | $\boldsymbol{\beta_0}$ **[kg]** | $\boldsymbol{\beta_1}$ **[-]** | $\boldsymbol{\beta_2}$ $[\boldsymbol{kg.cm^{-1}}]$ |
|---|---|---|---|
| Pelvis | -7.498 | 0.0976 | 0.04896 |
| Abdomen | 7.181 | 0.2234 | -0.0663 |
| Thorax | 8.2144 | 0.1862 | -0.0584 |
| Neck | 0.096 | 0.0031 | 0.0022 |
| Head | 1.2 | 0.014 | 0.0123 |
| Arm | 0.25 | 0.03013 | -0.0027 |
| Forearm | 0.3185 | 0.01445 | -0.00114 |
| Hand | -0.1165 | 0.0036 | 0.00175 |
| Tight | -2.649 | 0.1463 | 0.0137 |
| Calf | -1.592 | 0.03616 | 0.0121 |
| Foot | -0.829 | 0.0077 | 0.0073 |

Table 1: Table of coefficients for calculation of body segment's masses [10].

Geometry of the segments was defined based on the Robbins database [11], where values of human segments' dimensions for the large, medium and small man and woman respectively, were published. Authors use average male of height equals to 180 cm as a reference one (measured in Czech Republic in the year 2011) [12], and all the dimensions from Robbins database (medium male) are related to this value. Thus, the dimensions of each body segment are functions of the total height of the human. Such algorithm makes the proposed in-house model scalable for various humans: height – Robbins database (Figure 2) and weight – Zaciorsky method (26). The inertia properties of the body segments are calculated from mass and geometry of the particular parts.

The in-house software model is parametrized with the height and weight, which makes the model scalable in order to capture the diversity of human population [13, 14, 15].

## 3 RESULTS

The mathematical methods for the development of the scalable multibody model of the human were described in the previous chapter. The purpose of such model is the description of the global behaviour of the human under different loading scenarios. The examples of the various human body models are shown in the Figure 4. Figure 5 shows the motion of the reference human body model (180 cm, 80 kg) for the prescribed initial velocities at the trunk and the extremities.

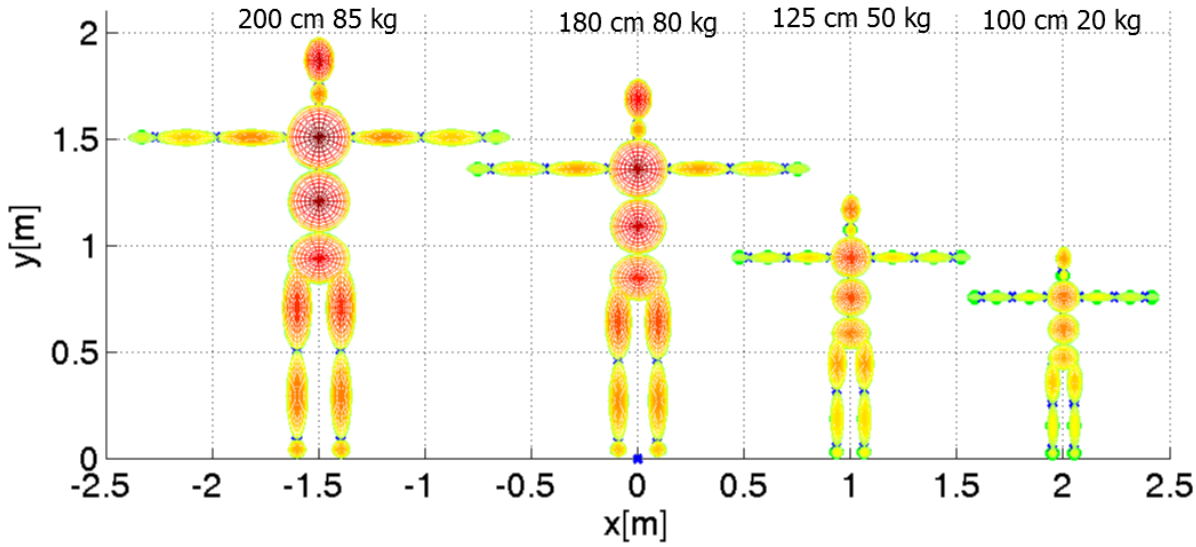



Figure 4: Human body model family.

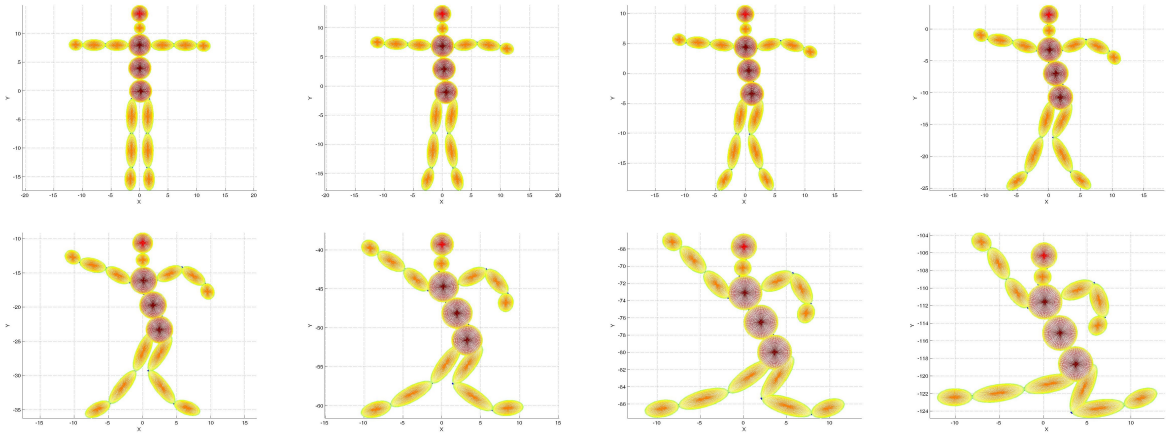


Figure 5: Illustrative motion of the human body model.

## 4 CONCLUSION

The presented work is focused on the investigation of the multibody principle in the modelling of the human body. Authors used Euler parameters set for the parametrization of the spatial rotations. This paper introduced such principle and pointed out the advantages and disadvantages of such approach. Generally, the Euler parameters can solve the problem with the singularities that occur with the Euler angles, however it brings an extra equation (constraint equation among Euler parameters) into the equation of motion for each body. Dynamics equations of motion for the case of unconstrained and constrained body, respectively, are defined here. Detection of the contact scenario is solved with the ellipsoid-to-plane contact algorithm and the consequent contact forces (normal and tangential) are calculate based on the continuous contact force model as a function of penetration. The definition of the forces consists of the virtual stiffness and virtual coefficients of restitution that will be further tuned up in order to correspond with the behaviour of the real human body. The presented model used the wide database of geometric properties of the human body, measured in 1970's in Czech Republic. The model is parametrized with the initial height and weight and the dimensions and particular masses depend on these values. The scaling process keeps this model scalable with respect to the particular human. This simple multibody model of human body was evaluated on the purpose of description of global behaviour of the human body. The goal was to evaluated the in-house software including contact scenario, scalable, easy-to-position model that can be easy modified for the further purposes, such as driving of the model or human cognitive behaviour.

### Acknowledgement

This work was supported by the internal grant project SGS-2016-059 and by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

## REFERENCES

[1] Jr. Ang, H. Marcelo, V. D. Tourassis, Singularities of Euler and roll-pitch-yaw representations. *Aerospace and Electronic Systems*, IEEE Transactions on 3: pp. 317–324, 1987.

[2] F. Amirouche, *Fundamentals of multibody dynamics: theory and applications*. Springer Science & Business Media, 2007.

[3] P. E. Nikravesh, *Computer-aided analysis of mechanical systems* Prentice-Hall. Inc., 1988.

[4] A. A. Shabana, *Computational Dynamics*. Second Edition, Wiley, Chicago, 2011.

[5] R. A. Wehage, *Quaternions and Euler parameters a brief exposition*. Computer aided analysis and optimization of mechanical system dynamics. Springer Berlin Heidelberg, pp. 147–180, 1984.

[6] J. Špička, M. Hajžman, P. Polach, A Bouncing Ball Benchmark Problem and Solution of An Ellipsoid-to Plane Contact. in *Proceedings of The 3rd Joint Conference on Multibody System Dynamics, The 7th Asian Conference on Multibody Dynamics*, pp. 29-30, Busan, Korea, 2014.

[7] J. Alves, P. Flores, N. Peixinho, Contact analysis for different viscoelastic contact force models. in *Proceedings* ECCOMAS, 2013.

[8] Hynčík, L, *Škálování hybridního modelu člověka*. Habilitační práce v oboru Mechanika, Západočeská univerzita v Plzni,(2014).

[9] J. Křen, P. Janíček, J. Rosenberg, *Biomechanika* Západočeská univerzita, Fakulta aplikovaných věd, 2001.

[10] E. Tlapáková, *Geometrie hmotností lidského těla* , [Online]. Available at http://biomech.ftvs.cuni.cz/staff/tlapakova/Geometrie hmotností lidského těla, 2011.

[11] L.W. Schneider, D.H. Robbins, M.A. Pflüg, R.G. Snyder, *Development of anthropometrically based design specifications for an advanced adult anthropomorphic dummy family*. volume 1. Final report, 1983.

[12] M. Kovařík, *Antropometrický výzkum dospělé populace a jeho aplikace v oblasti interiéru a architektury*. Vysoké učení technické v Brně. Fakulta architektury, 2011.

[13] P. Bláha, *Antropometrie československé populace od 6 do 55 let*. Ústř. štáb Čs. spartakiády 1985, 1987.

[14] Hynčík, L, Rigid body based human model for crash test purposes. *Engineering Mechanics*, vol. 8, pp. 1-6, 2001.

[15] D. H. Robbins, *Anthropometry of motor vehicle occupants*. Technical report, 1983.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EXPERIMENTAL STUDY ON THE FUZZY-PID HYBRID CONTROL ALGORITHM FOR UNLOADING SYSTEM IN MECHATRONIC DEVICE FOR GAIT RE-EDUCATION

**Slawomir Duda**[1], **Grzegorz Gembalczyk**[1]

[1] Silesian University of Technology
Akademicka 2A, 44-100 Gliwice, Poland
{Slawomir.Duda, Grzegorz.Gembalczyk}@polsl.pl

**Keywords:** Instructions, ECCOMAS Congress, Computational Methods, Engineering Sciences, Proceedings.

**Abstract.** *In the article, the control system dedicated to the unit responsible for the patient unloading, which is one of the components of the mechatronic device for gait reeducation will be discussed and experimentally verified. The control algorithm of the unloading system should guarantee an expected constant value of the unloading force, regardless of the position and direction of the patient movement.*

*The article presents a solutions in which the hybrid control algorithm based on fuzzy logic coupled with PID controller has been implemented. The optimization of parameters for regulators was conducted based on numerical simulations in which the walk of a patient through a step was analyzed. The experimental results show that the proposed control algorithm realized the primary goal (rapid reaction to the changes of the unloading force values) very well.*

## 1 INTRODUCTION

Modern mechatronic devices find an increasing number of applications aimed at aiding the activities performed by humans. Advanced mechatronic systems are used both in common household devices as well as in various industries. One of the more interesting applications of modern mechatronic technologies is found in medicine and rehabilitation. This might be exemplified by surgical robots such as Vinci or Robin Heart. These robots are required to allow for a precise positioning of surgical instruments as well as to provide a precise view of the surgical field [1,2,3].

The use of automated devices is beneficial in the process of rehabilitation - especially of the upper and lower limbs. Devices such as Lokomat, zeroG or exoskeletons allow to increase the intensity of the training and perform a larger number of repetitions, which – as reported by numerous studies – translates into a significant improvement in rehabilitation effects. Moreover, the devices also support the work of physical therapists, allowing them to concentrate on observing the patient's technique in performing the exercises [4,5].

Rehabilitation devices are also required to adjust the resistance to motion in a manner adequate to the strength of the patient's muscles and to minimize the excessive loads. These goals are reached by means of biofeedback or with the use of mechanical solutions [6].

One of the innovative projects in the field of rehabilitation equipment is the mechatronic device dedicated to the gait re-education, developed in the Institute of Theoretical and Applied Mechanics at Silesian University of Technology.

## 2 MECHATRONIC DEVICE FOR GAIT RE-EDUCATION

The structure of the device resembles a single-girder overhead travelling crane which perform a keeping up the movement behind the patient. In addition, the dedicated special purpose winch relieves the patient with a constant force in the vertical direction.

Each of the device's axes is driven by servo drives with permanent magnet excited synchronous motors. Due to the use of dedicated servo inverters, it is possible to adjust the dynamics of the device, e.g. by limiting the acceleration of individual elements. Moreover, all motors have been equipped with measuring encoders.

The designed control system has been created in MATLAB/Simulink and the transmission of control signals between the PC Computer and the device control unit is conducted by means of two RT-DAC4/PCI real time boards. A signal conditioning interface has also been developed that is used for the measurement of the rope inclination angle, the value of unloading force and the pressure of the patient's feet against the ground. The correlation between the components in the mechatronic device for gait re-education has been presented in figures 4 [7,8].

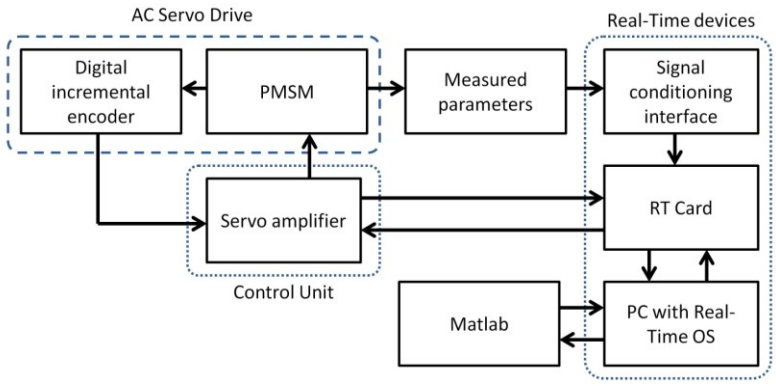



Figure 1: Block diagram of the of the components cooperation in mechatronic device for locomotor training.

The hoist assembly is of a special significance – it should rapidly react to the changes of the unloading force and simultaneously protect the patient against excessive loads. It is also important that sufficient vertical movement is provided as it is necessary e.g. for the training of climbing the stairs. To satisfy these requirements, two separately operating drive systems were installed in the unloading device. In the first of these (Z1), the drive motor drives the winding drum of the hoist by means of a planetary gear, which allows to control the vertical position of the sling. The second one has been coupled with a screw drive responsible for the movement of the travelling element which is separated from the pulley with elastic elements (Z2 drive). It is responsible for the dynamic compensation of the patient's body weight. Such an electromechanical system acts as a "Series Elastic Actuator", dedicated for rehabilitation devices or walking robots. The structure of the unloading system has been presented in fig. 2.

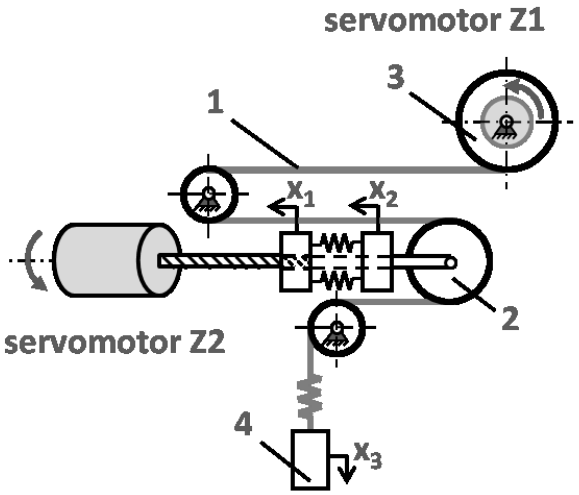



Figure 2: Scheme of unloading system.

The mathematical model of the presented device has been developed. The motion of the system has been described using the generalized displacement variables and the numerical calculations for this model were performed using the MATLAB/Simulink software [9].

Stiffness of the springs in the Z2 actuator and the stiffness of the rope were determined experimentally with the testing machine.

## 3 CONTROL ALGORITHM FOR UNLOADING SYSTEM

The development of an optimal control of the unloading system constitutes an interesting engineering problem. From the point of view of rehabilitation centres, it is desirable that the control is performed by means of a single, universal system, that would not be dependent on the type of exercises.

Experiments have shown, however, that in case of walking on a horizontal and flat surface it is best to only use the Z2 drive. The use of the PID controller allowed to maintain the unloading force with the precision of approximately 30 N. The movement of this actuator is limited by limiting switches and its range of movement is approximately 10 cm. To give consideration to the tests that have been completed, a solution was proposed in which the PID controller's operation was coupled with a fuzzy logic controller [10,11,12]. The idea of study control algorithm is shown in Fig. 3.

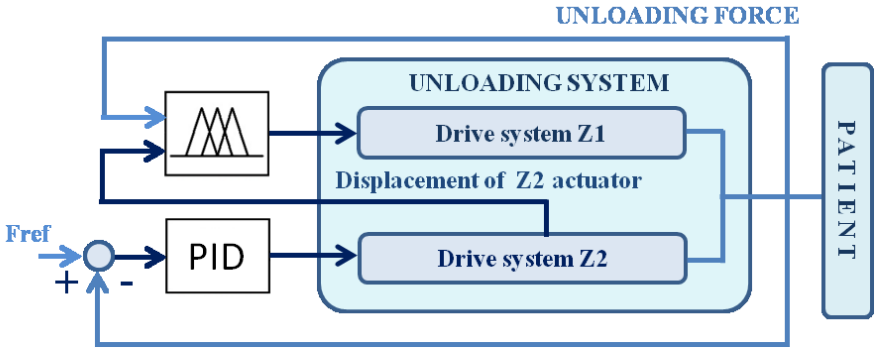



Figure 3: Block diagram of the proposed control system.

The developed set of controlling rules has been presented in table 1. The $DZ_2$ symbol denotes the displacement of the actuator in the Z2 drive, while UF relates to the unloading force misalignment value. The characters from "---" to "+++" represent the values in an increasing order. The "nom" index means that the value of the variable equals the nominal value.

| | | Displacement of actuator in Z2 drive system | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $D_{Z2}$--- | $D_{Z2}$-- | $D_{Z2}$ - | $D_{Z2nom}$ | $D_{Z2}$+ | $D_{Z2}$++ | $D_{Z2}$+++ |
| Unloading force | UF - - - | Z1++ | Z1++ | Z1+++ | Z1+++ | Z1+++ | Z1+++ | Z1+++ |
| | UF - - | Z1++ | Z1++ | Z1+++ | Z1++ | Z1++ | Z1+++ | Z1+++ |
| | UF- | Z1+ | Z1+ | Z1+ | Z1+ | Z1+ | Z1++ | Z1++ |
| | $UF_{nom}$ | Z1-- | Z1- | 0 | 0 | 0 | Z1+ | Z1++ |
| | UF + | Z1-- | Z1-- | Z1- | Z1- | Z1- | Z1- | Z1- |
| | UF ++ | Z1--- | Z1--- | Z1-- | Z1-- | Z1--- | Z1-- | Z1-- |
| | UF +++ | Z1--- | Z1--- | Z1--- | Z1--- | Z1--- | Z1-- | Z1-- |

Table 1: Applied rules as a table.

The ranges of the membership functions have been determined in the process of optimization that was conducted using the numerical model of the device. During the numerical calculations the walk of a patient through a 10 cm high step was analyzed. And the optimization was done using genetic algorithm. The settings of the PID controller have been selected independently for the case in which the patient moves on a flat surface and only the Z2 drive is active. This allowed to tune the fuzzy controller to the existing control system and largely facilitated the numerical computations by reducing the number of variables in the optimization process.

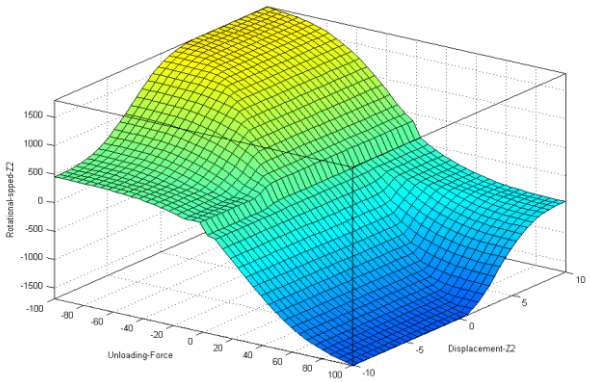



Figure 4: The output signal controlling the motor rotational speed in Z1 drive presented as a surface.

## 4 EXPERIMENTAL RESEARCH OF THE CONTROL SYSTEM

A series of experiments was conducted for the developed algorithm controlling the drives of the unloading system. In the experiments, the operation of the device was tested during walking on a flat surface and in case of climbing a stair that was approximately 10 cm high. During the study, the unloaded person moved with different speeds. Special attention was given to the operation of the motor driving the winding drum. The attached graphs exhibit the registered values of shifts of the centre of gravity of the unloaded person in vertical axis (fig. 5), the value of the unloading force (fig. 6), the displacement of the regulating unit in the Z2 drive system (fig. 7), the linear displacement of the end of the cable resulting from the operation of the winding drum (fig. 8) and the value of the control signal of the engine speed in the Z1 drive (fig. 9).

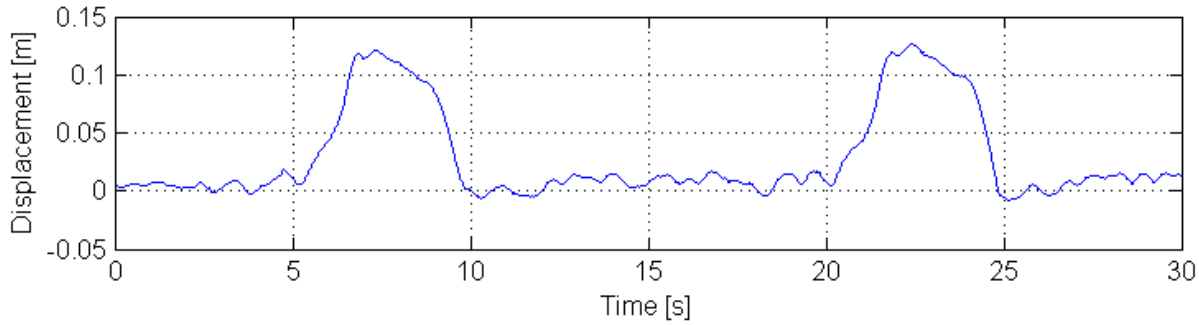



Figure 5: Registered displacement of the centre of gravity of the unloaded person in vertical axis as a function of time.

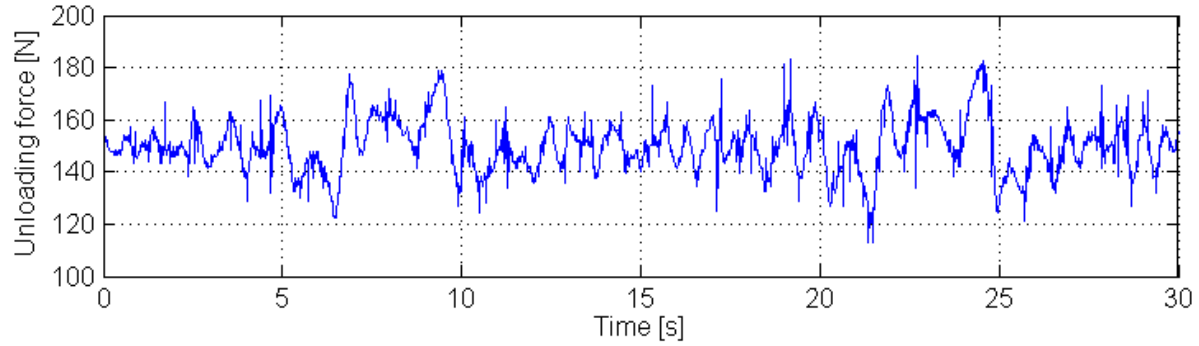



Figure 6: Registered value of the unloading force as a function of time.

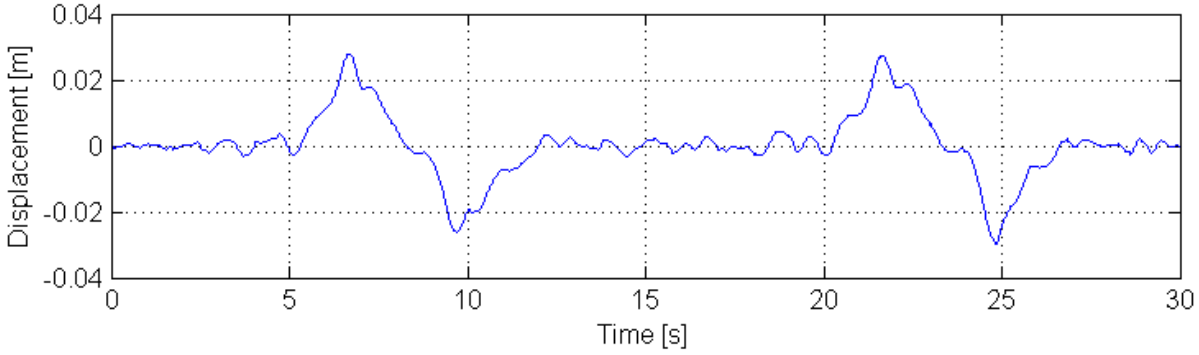



Figure 7: Registered displacement of the regulating unit in the Z2 drive system as a function of time.

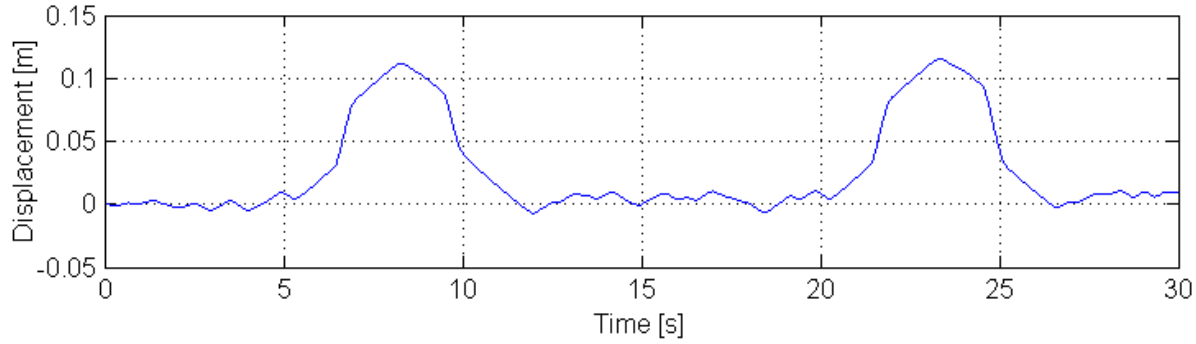



Figure 8: Registered linear displacement of the end of the cable resulting from the operation of the winding drum as a function of time.

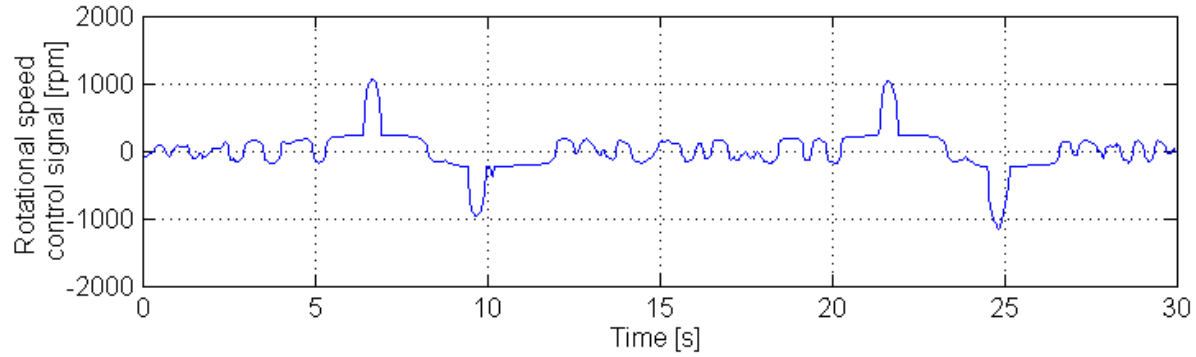



Figure 9: Registered value of the control signal of the engine rotational speed in the Z1 drive as a function of time.

## 5 CONCLUSIONS

- The work describes an innovative device used for the re-education of walk. Special attention was given to the structure and operation of the unloading system which requires the coordination of two independent drives for proper operation.
- A hybrid Fuzzy-PID control system was proposed, in which the PID controls the operation of the drive responsible for dynamic compensation of the patient's weight. The fuzzy controller controls the operation of the winding drum.
- A set of control rules has been developed for the presented fuzzy controller as well as membership functions have been defined for all input and output signals. The ranges of these functions have been selected in an optimization process with the use of the numerical model of the device.
- Experiments have confirmed the proper operation of the developed algorithm. The application of fuzzy control along a classic PID controller guaranteed the maintenance of the unloading force with the precision of 30N, notwithstanding the type of the exercises performed by the unloaded person. The experiments have also confirmed that one of the basic goals has been reached – during a walk on a flat surface the misalignment of the unloading force is regulated primarily by the dynamic compensation system. Moreover, the winding drum starts operating at high speeds only when the patient performs exercises requiring large vertical displacements.

## REFERENCES


[1] P. Abrishami, A. Boer, K. Horstman, Understanding the adoption dynamics of medical innovations: Affordances of the da Vinci robot in the Netherlands. *Social Science & Medicine*, **117,** 125-133, 2003.

[2] S.M. Yoon, W.J. Kim, M.C. Lee, Design of Bilateral Control for Force Feedback in Surgical Robot. *International Journal of Control Automation and Systems,* **13,** 916-925, 2015.

[3] Z. Nawrat, P. Kostka, Polish cardio-robot 'Robin Heart'. System description and technical evaluation. *International Journal of Medical Robotics and Computer Assisted Surgery,* **2,** 36-44, 2006.

[4] H.S. Lo, S.Q. Xie, Exoskeleton robots for upper-limb rehabilitation: State of the art and future prospects. *Medical Engineering & Physics,* **34,** 261-268, 2012.

[5] H.Y. Yu, S.N. Huang, G. Chen, Y.P. Pan, Z. Guo, Human-Robot Interaction Control of Rehabilitation Robots With Series Elastic Actuators. *IEEE Transactions on Robotics,* **31,** 1089-1100, 2015.

[6] A.L. Behrman, S.J. Harkema, Locomotor training after human spinal cord injury: a series of case studies, *Physical Therapy*, **80**, 688-700, 2000.

[7] R. Campa, R. Kelly, V. Santibanez, Windows-based real-time control of direct-drive mechanisms: platform description and experiments, *Mechatronics,* **14***,* 1021-1036, 2004.

[8] S. Duda, G. Gembalczyk, R. Lawniczek, Simulation and experimental studies on the control system for mechatronic rehabilitation device with keep-up movement. V. Fuis, *20th International Conference on Engineering Mechanics (EM),* Svratka, Czech Republic, May 12-15, 2014.

[9] R.S. Esfandiari, B. Lu, *Modeling and analysis of dynamic systems.* CRC Press, 2014.

[10] R.Y. Wan, Y. Z. Zhang, H. H. Sun, W. Zhou, Z.J. Chen, Hybrid Fuzzy PID Control for Synchrodrive System of Dual Linear Motors. *Applied Mechanics and Materials***, 130**, 804-808, 2014.

[11] A.R. Norouzzadeh Ravari, H.D. Taghirad, A novel hybrid Fuzzy-PID controller for tracking control of robot manipulators. *International Conference on Robotics and Biomimetics,* Bangkok, Thailand, February 21-26, 2009.

[12] K.Y. Tu, T.T. Lee, W.J. Wang, Design of a multi-layer fuzzy logic controller for multi-input multioutput systems. *Fuzzy sets and systems*, **111**, 199-214, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FE BONE STRUCTURAL ANALYSIS WITH CT MAPPING OF INHOMOGENEOUS MATERIAL PROPERTIES

**M. T. Bahia[1], E. G. F. Mercuri[2], and M. B. Hecke[3]**

[1] Federal Institute of Santa Catarina
Rua Pavão, 1377, Bairro Costa e Silva - CEP 89220-618, Joinville, SC, Brasil
e-mail: migueltbahia@ifsc.edu.br

[2,3] Federal University of Paraná
Centro Politécnico - Bloco Lame/Cesec, Caixa postal 19011, Bairro Jardim das Américas CEP 81531-990, Curitiba, PR, Brasil
{emiliomercuri, mildredhecke}@gmail.com

**Keywords:** Subject-specific Finite Element Modeling, Computed Tomography (CT), Image Processing, Bone Biomechanics, Young Modulus, Femur.

**Abstract.** *This paper presents a bone structural analysis of a reconstructed fractured femur assigning inhomogeneous isotropic material properties accessed from CT images into the FE model. The grayscale values accessed from CT images were converted into local Young modulus that was assigned to the associated finite element of the mesh. The FE model was obtained from CT images by a numerical reconstruction of a proximal femur which presents an atypical fracture in the sub-trochanteric region with full displacement and no continuity between the fragments. The reconstruction involved procedures such as volume reconstruction from images, segmentation by thresholding, labeling of the connected components and also the alignment of point clouds data using the Iterative Closest Point (ICP) algorithm. The resulted aligned point cloud was thus converted into a solid and finally in a finite element mesh, that was analyzed in a standard FE framework. Another FE model with constant elastic modulus for cortical and trabecular bone was also simulated to compare results with our model. The values of principal stress and strains were rather closer to tensile values but for compressive values there were obtained significant differences between the two models, indicating that inhomogeneous model captures better the mechanical behavior for compression. Results obtained are in accordance with expected bone biomechanics behavior. The assignment of inhomogeneous material properties into the FE model become the structural analyzes more realistic and this research highlight that this assignment is fundamental for reliable subject-specific models. The procedure implemented for reconstruction of fractured bones using CT images offers new insights to analyze the bone fractures since allows recovering the material state immediately before the failure.*

## 1 INTRODUCTION

The modeling of biological tissues has commonly involved the adoption of homogeneous isotropic materials, with constant Young's modulus and Poisson's ratio. Such adoption has been very useful considering aspects such as the simplicity of computational implementation and the theoretical consolidation in the scientific community. However, in the case of bone, for instance, many studies highlight that this tissue presents heterogeneous physical properties and anisotropic behavior with age and tissue density dependence [1].

On the other hand, the use of subject-specific FE models as a tool to access the state of stress and fracture risk in individual patients based on CT images is increasing among researchers [2-4]. In this perspective, the accurate knowledge of the stress and strain characteristics implies the use of reliable models that consider specific features to define the geometric and the material model.

The method used to perform CT scans is accurate; the test is non-invasive, painless and involves little radiation exposure for patients allowing acquiring images with high quality from any part of human body, even in very small structures such as, for example, brain detail.

The CT numbers or the intensity of image pixels (16 bits) is represented by Hounsfield Units (HU). The HU numbers describe quantitatively a radiological density, which corresponds to a linear transformation of measuring the linear attenuation coefficient. The radiological density of distilled water at standart temperature and pressure is set to *0* HU (zero HU) and the radiological density of air at STP is set to *-1000* HU. The HU numbers can be related to an apparent density using hydroxyapatite phantoms of known mineral density, and further performing a linear interpolation. There are several relationships between the HU values and Young modulus available at the literature, for instance in [5-6]. Thus it is possible perform a mapping into the inhomogeneus isotropic material properties and further assign the material properties into the FE model. In this work we have implemented a procedure to perform the mapping of the grayscale values accessed from CT images, conversion into bone ash densities and Young modulus and assignment of these local material properties into the FE model. Similar mapping of properties accessed from CT images have been performed by softwares like IA-FEmesh[1] and Bonemat[2] [7]. However this kind of mapping is commonly used when there is geometrical equivalence and equal reference coordinates between the images and the 3D model reconstructed.

In this work we perform a numerical reconstruction of a fractured femur preserving the grayscale values accessed from CT images. These grayscale values are correlated to bone ash densities and converted into local elastic modulus. This inhomogeneous isotropic material distribution is finally assigned to the FE model, which contribute to become the model more realistic and reliable. Besides, this numerical implementation of reconstruction of fractured bones using CT images offers new insights to analyze the bone fractures since allows recovering the material state immediately before the failure. The steps for reconstruction of the femur involved techniques of image processing such as segmentation by thresholding, contour definition and labeling of connected objects. The alignment of point clouds was implemented using an Iterative Closest Point (ICP) algorithm, commonly used for registration on CAD systems [8].

[1] Available to download at https://www.ccad.uiowa.edu/MIMX/projects/IA-FEMesh
[2] Available to download at http://www.bonemat.org/

## 2 MATERIALS AND METHODS

A fractured proximal femur was automatic 3-D reconstructed from CT images. The dataset of the fractured femur was taken *in vivo*, for diagnostic reasons. The patient was a 60 years old woman with a medical history of osteoporosis, as shown in Fig. 1. The CT scans were performed by the Center for Diagnostic Imaging at Hospital Santa Catarina (SP, Brazil). The images were kindly provided by professor Marcelo de Medeiros Pinheiro, Department of Medicine of the Federal University of São Paulo (UNIFESP). For this reconstruction we use 243 CT slices of the fractured proximal femur (diaphyseal femoral head) spaced at 1.5 mm, which means likely 3/4 of the femur. The fracture was concerned about cross producing lateral displacement and angularity between the two resulting fragments. The 3-D reconstruction, mapping and assignment of inhomogeneous isotropic material properties, generation of FE model and structural analysis were fully implemented with in-house programs written in MatLab[3]. The creation of the STL file used as an input data for mesh creation was developed using the platform MeshLab[4] and the post-processing visualization were performed in the open source Paraview 5.0[5]. In the following, the procedure employed is detailed and the methods for creating a FE model with inhomogeneous isotropic material properties are summarized.

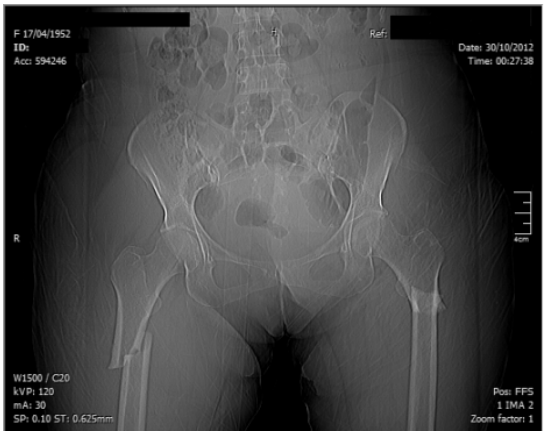



Fig. 1. CT scan of the atypical bilateral hip fracture without trauma. The donor is a female diagnosed with osteoporosis.

### 2.1 Subject-specific FE model

The CT scans were performed using a scanner with parameters: 120 kVp, 30 mA, 1.5 mm slice thickness and pixel size of 0.32 mm. In order to convert the image dataset to a 3D dataset of pixels, each image (that contains pixels information) is considered as a plane and each pixel a point in this plane. Acquiring planes at different depths so as the process is repeated, a 3D volume is acquired. Each level corresponds to the distance between each slice, as illustrated on Fig. 2.

After obtained the 3D coordinates others steps must be performed to isolate the part of 3D dataset to be analyzed. Here it was employed a segmentation by thresholding, which allows to remove all the pixels of a volume that do not satisfy a condition. In this case it was chosen an intensity threshold and only those pixels above the threshold were selected. The threshold

[3] Trademark of The MathWorks Inc., MA, USA.
[4] Available to download at http://meshlab.sourceforge.net/
[5] Available to download at http://www.paraview.org/

used in this work was 104 for a grayscale range of [0, 255]. Result of the segmentation can be viewed in Figs. 3 and 4.

Considering that in this case we are dealing with a fractured femur, it is necessary identify each fragment of the fractured femur and perform coordinates transformations in such a way that align the fragments to reconstruct the original femur before fracture. Translation of point cloud was carried out by adding the vector position of each node *i* (*i* = 1, ..., number of nodes), the desired translation vector $\vec{T} = (a, b, c)$, where *a*, *b*, and *c* are the displacements in *x*, *y*, and *z*, respectively.

$$\vec{x}_{new} = \vec{x}_{old} + \vec{T} \tag{1}$$

The rotation of the point cloud data about *x*, *y*, and *z* axes is performed by multiplying the position vector of each node by the rotation matrix, wherein:

$$R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{bmatrix}$$

$$R_y(\beta) = \begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \tag{2}$$

$$R_z(\gamma) = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Unfortunately, the rotation matrices and translations vectors necessary to this alignment are unknown. To overcome this difficulty we employ here a registration technique for alignment of point cloud data called Iterative Closest Point method (ICP) that is commonly used in CAD systems for registration of surfaces. The ICP algorithm was introduced in 1991 by [9] and independently by [10] and it was further developed by various researchers.

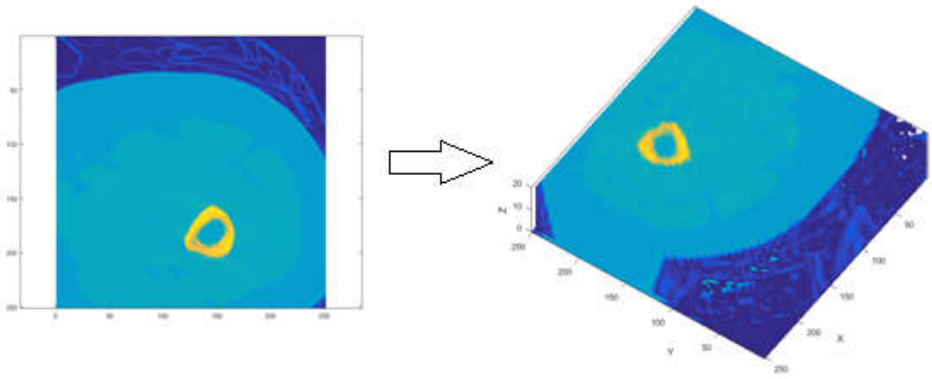


Fig. 2. Conversion of 2D CT images into a 3D volume.

The overall goal is to apply a transformation to one of the clouds and bring it as close to the other as possible. The convention is to apply the transformation to the data points in order to bring these to the best alignment with the model points. In 3-D space, such a transformation has six degrees of freedom – three rotations and three translations – and the objective function

becomes a function of six variables, see [8], [11]. Expressing a rotation in terms of the rotation matrix **R** and a translation in the vector $\vec{T}$ the problem may be seen as a minimization problem of the squared distances from points in one cloud to their nearest neighbors in the other *after* appliance of the transformation. We are interested in to find **R** and $\vec{T}$ that minimize the following objective function:

$$F = \sum_{i=1}^{N} \left\| [\mathrm{R}] p_i - \vec{T} - q_i \right\|^2 \tag{3}$$

In this work it was used as a reference point cloud a femur with no fracture. Thus, the fragments represented by point clouds were aligned using a healthy femur, to find the rotation matrix **R** and the translation vector $\vec{T}$. The resulted 3D aligned point cloud was converted into a 3D surface using the open source software Meshlab. The 3D surface was transformed in a solid finite element model discretized into 3846 nodes and 16689 linear tetrahedral elements, as illustrated on Fig. 6.

### 2.2 Mapping of pixel values in elements

After performing the above procedures to reconstruct the fractured femur preserving the pixels information and generate the FE model, the next step is assign to each finite element of the mesh the average grayscale value of the pixels that fall inside of each element. Thus it is necessary to establish a criterion for checking if the pixel fall inside or outside of the tetrahedral finite element, and further assign this value (or average of values) locally to each element.

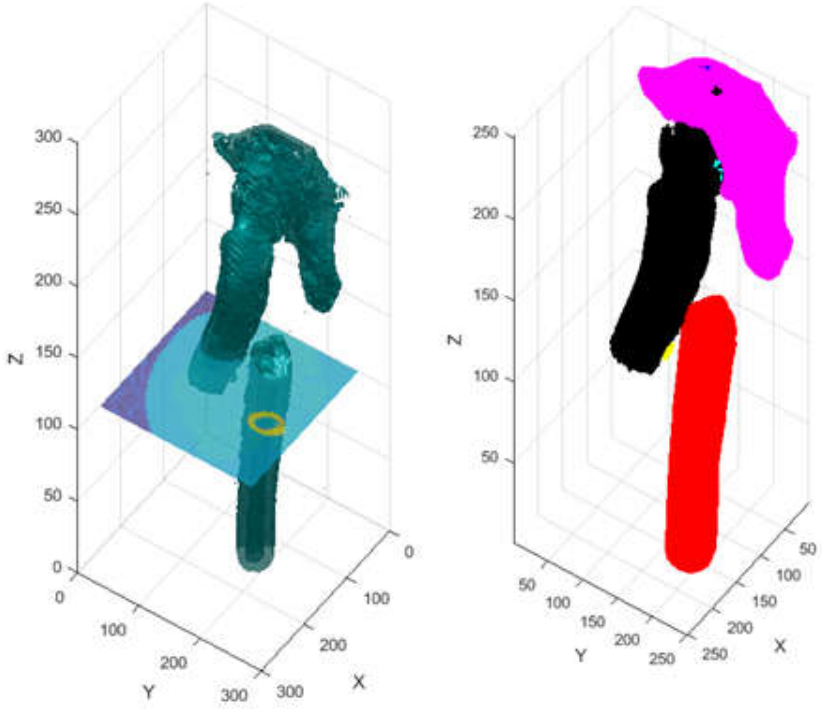



Fig. 3. Segmentation by thresholding (left) and labeling of connected components (right).

In other words, the goal here is to check whether a point defined by the pixels coordinates falls inside or outside of the convex hull defined by the vertices of each tetrahedral finite ele-

ment and assigns the average of the grayscale values of the pixels inside the convex hull to each respective element. There are several possibilities to perform this. One very interesting way adopted here was proposed by [19] and is summarized as follows.

One point of a 3D point cloud data belongs to a specified tetrahedral if the signed volume of all the tetrahedral formed by this point and each triangular face of the specified element is positive. When at least one of the volumes of the tetrahedral formed by this point and each triangular face of the element is negative the point fall outside of the element. This procedure is repeated for all the elements of the model in order to verify if the point belongs or not to a specified element.

In this work we adopt medical images of 8 bits, representing a grayscale range of [0, 255]. Fig. 5 shows a histogram of the distribution of the mapped pixels on the reconstructed femur by grayscale values provided by the 3D images. The bimodal appearance (two bells) of the graph indicates the two regions of bones: cortical and trabecular. Histograms such this one also allows to analyze the effects of CT image resolution in the material properties distribution. The result of the procedure of mapping and assignment of grayscale values into the FE mesh is illustrated at Fig. 6.

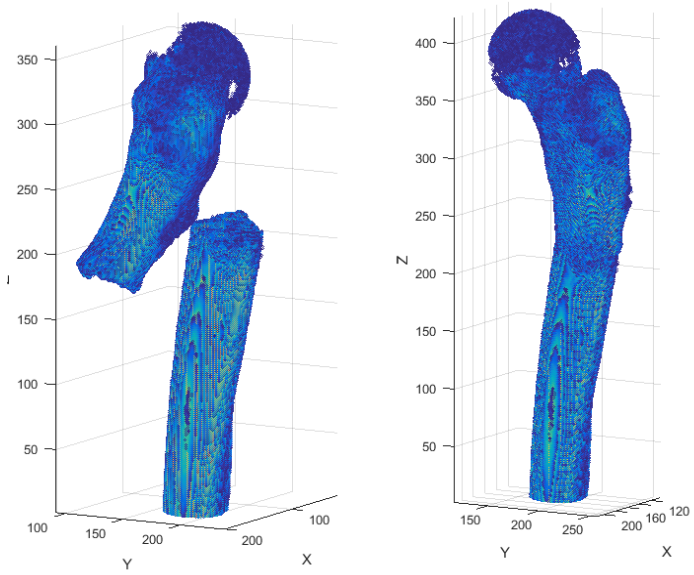



Fig. 4. Point cloud data of femora fragments (left) and aligned femora (right) using the ICP algorithm.

In order to determine the point wise Young modulus it was adopted the following procedure. For cortical bone, we assign to the whitest pixel of image (full bone with no voids) the value 1.4 $g/cm^3$ for bone ash density[6]. For water region on image (darkest pixel surrounding the bone) we attribute "zero" ash density. Other values of ash density were finding using a linear interpolation between the region with full bone with no voids ($\rho_{ash}$ =1.4 $g/cm^3$) and the water region ("zero" ash density), which produced the following equation:

$$\rho_{ash} = [p(x,y,z) - a] \quad \frac{14}{b-a} \quad [\mathrm{g/cm^3}] \tag{4}$$

where $p(x,y,z)$ is the image grayscale value at the position $(x,y,z)$, $a$ and $b$ are the grayscale values for water and full bone with no voids, respectively. In this study, the values adopted were $a$ = 66 and $b$ = 219, considering a grayscale range of [0,255]. These values should be

[6] Ash weight divided by the real volume, which neglects the effect of fat content for trabecular bone.

chosen according to the available images. After correlate the grayscale values with ash densities, it is adopted a relationship proposed by [16] and validated by [17] and [18] to determine the corresponding Young modulus (Eqs. 5 and 6):

$$E_{cort} = 10200\rho_{ash}^{20}{}^{1}\ [\text{MPa}] \tag{5}$$

$$E_{trab} = 5307\rho_{ash} + 469\ \ [\text{MPa}] \tag{6}$$

For distinguish the cortical and trabecular bone, we define $\rho_{ash} > 06g/cm^{3}$ for cortical bone and $\rho_{ash} \leq 06g/c\ m^{3}$ for trabecular bone, as proposed by [12]. Poisson's ratio was set to 0.3 for all material bins, covering the range used in other studies, such as [12] and [15]. The elastic modulus range obtained with this procedure of mapping was approximately 1.06 – 19.46 GPa.

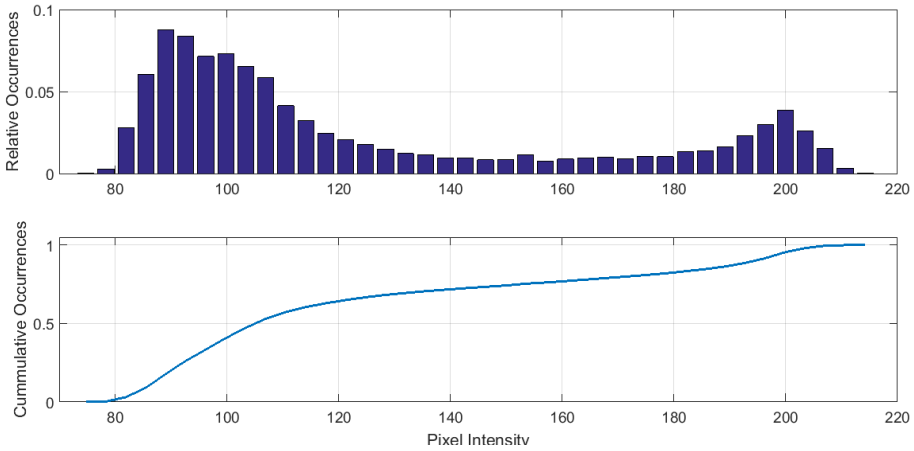



Fig. 5. Histograms of pixel intensities by relative and cumulative frequencies (above and behind, respectively).

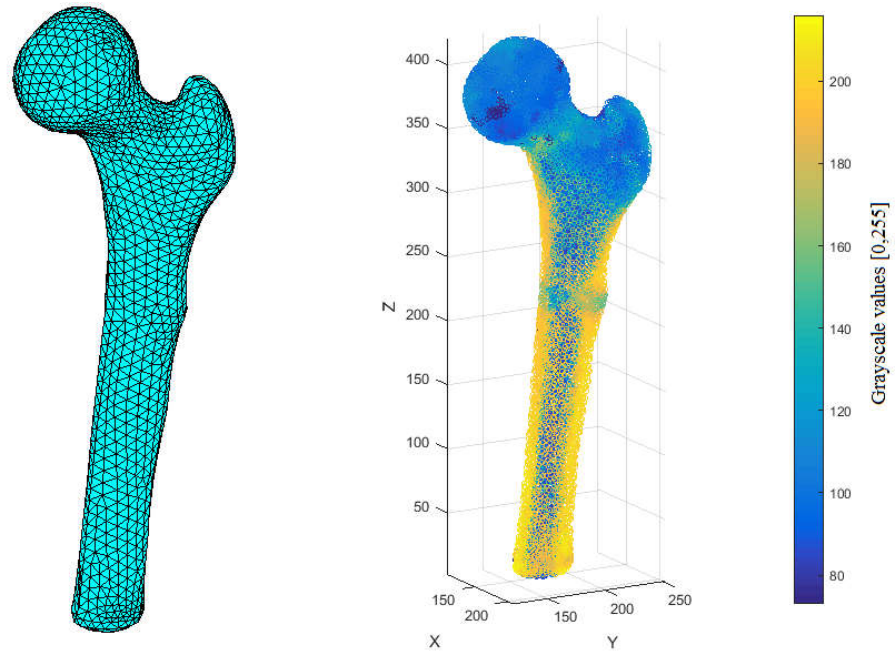



Fig. 6. Discretization of reconstructed femur in a tetrahedral finite elements mesh (left) and the assignment of the grayscale values for each finite element (right).

### 2.3 Boundary conditions of FE model

Reliable boundary conditions are fundamental for a realistic analysis of a biomechanical model. Here the models proposed by [20] and [21], were modified to define the values of adductor force on the femur head and the abductor force on the greater trochanter. For this analysis we assume an individual weighting 680 N. It is performed a balance of forces acting on the pelvis accordingly to [21], which establish the resultant adductor force upon femur head at an angle of 16° with z-axis and a resultant abductor force upon greater trochanter at an angle of 22° with z-axis. In this case, the adductor and abductor forces are 2168 N and 1632 N, respectively. These forces were proportionally distributed on the element faces of the boundaries at femur head and greater trochanter as shown in Fig. 7. The value of each face force is proportional of the angle between the normal of the element face and the resultant force. The nodes on the basis of femur were set fixed.

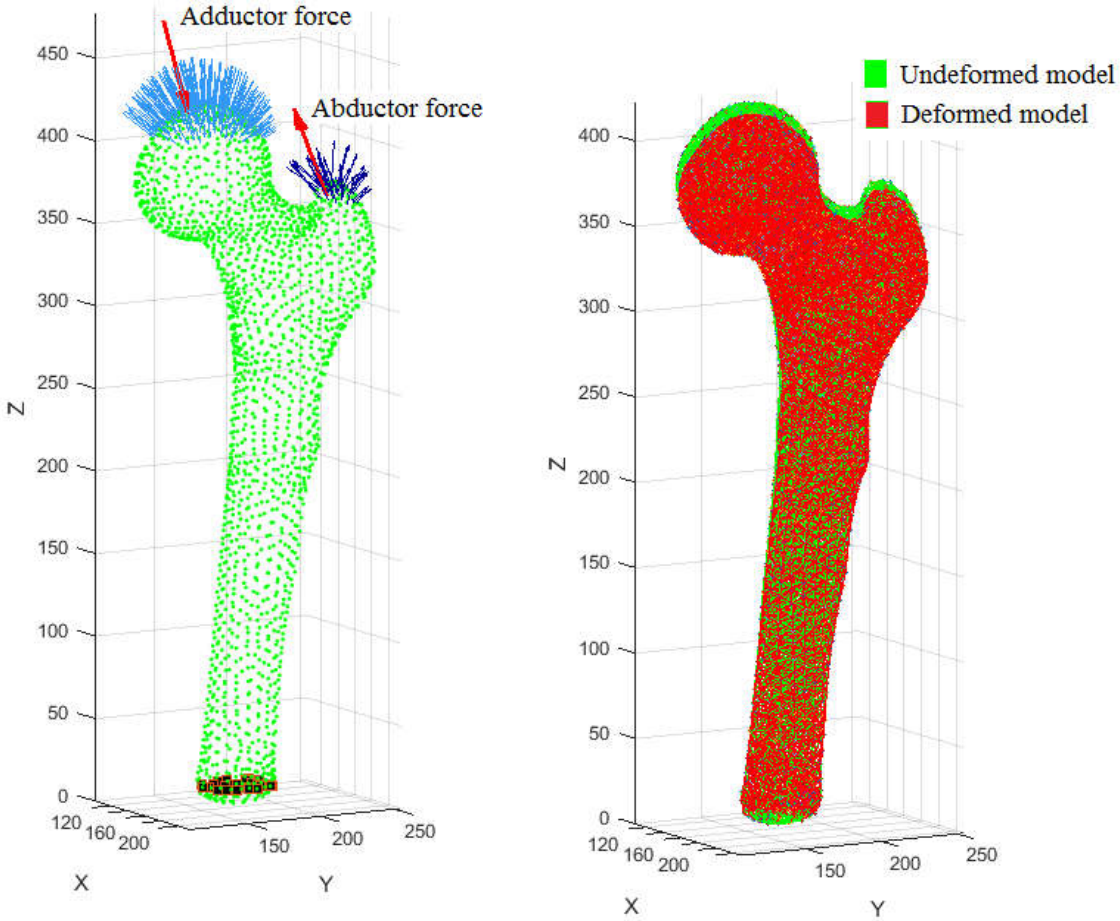



Fig. 7. Boundary conditions applied to the FE model (left) and undeformed/deformed configurations(right).

## 3 RESULTS

A fractured femur CT scanned was reconstructed in 3D solid and analyzed by a standard finite element framework developed by the authors. The ICP algorithm used to align the point clouds data seems to be very effective in this purpose. The inhomogeneous isotropic material properties accessed from CT images were assigned to the FE model. The range of elastic modulus obtained in the procedure of mapping was 1.06 – 19.46 GPa, which fits in the range

proposed by recent studies. Fig. 7 shows the applied boundary conditions and the deformed configurations obtained after simulation.

Some previous works has been adopted Von Mises equivalent stress as a local yield criterion, as pointed by [22]. Von Mises stress is suitable for materials that present a ductile behavior as metals. On the other hands, in the case of bone the tensile strength is smaller than the compressive strength, indicating that bone be classified as a brittle material, as discussed in [22] and [23]. Thus we adopted here the Drucker-Prager equivalent stress, which takes in account the hydrostatic stress contribution, following the work of [22]. Drucker-Prager yield criterion is described by the following equations:

$$F(\sigma) = \alpha J_1 + (J'_2)^{1/2} \tag{7}$$

$$J_1 = \sigma_x + \sigma_y + \sigma_z$$

$$J'_2 = \frac{1}{2}\left(\sigma'_x{}^2 + \sigma'_y{}^2 + \sigma'_z{}^2\right) + \tau_{xy}^2 + \tau_{yz}^2 + \tau_{xz}^2 \tag{8}$$

$$\sigma'_{ij} = \sigma_{ij} - \delta\sigma_m \text{ and } \sigma_m = \frac{1}{2}\sigma_{ii}$$

where $F(\sigma)$ is a yield function, $\sigma$ is the stress tensor, $\sigma_x, \sigma_y$ e $\sigma_z$ are the normal stress, $\tau_{xy}$, $\tau_{yz}$, and $\tau_{xy}$ are shear stress, the prime (') indicates deviatoric stress. $J_1$ is the first stress invariant, $J_2$ the second deviatoric stress invariant, and α is a parameter related to the proportions of the volumetric and deviatoric strains or the dilative potential of material. Here α is set to 0.07, as proposed in [24]. Fig. 8 shows the Drucker-Prager equivalent stress evaluated for this structural analysis.

| | | | Inhomogeneous<br>$E_{cortical}$ =10200$\rho_{ash}^{2.0}$ [MPa]<br>$E_{trabecular}$= 5307$\rho_{ash}$+469 | Homogeneous<br>$E_{cortical}$ = 18100 MPa<br>$E_{trabecular}$ = 15200 MPa |
|---|---|---|---|---|
| Drucker-Prager eq. stress | [MPa] | | 537.85 | 528.28 |
| Principal stress | [MPa] | Max (tensile) | 417.71 | 407.95 |
| | | Min (compressive) | -806.32 | -641.97 |
| Principal strains | [mm/mm] | Max (tensile) | 0.034099 | 0.022277 |
| | | Min (compressive) | -0.073072 | -0.035104 |

Table 1. Comparison between results obtained by inhomogeneous and homogeneous FE model

In order to compare the effectiveness of the procedure of mapping inhomogeneous isotropic material properties, we perform the same structural analysis considering a material with homogeneous distribution of elastic modulus. It was adopted here two constant values for elastic modulus: 18.1 GPa for cortical bone and 15.2 GPa for trabecular bone, as proposed by [25]. The results are summarized at Table 1. The values for Drucker-Prager equivalent stress and tensile principal stress are similar. Significant differences are observed for compressive values of principal stress and strains, what suggests that inhomogeneous model capture better the mechanical behavior on compression. The contribution of hydrostatic stress in the yield criterion on Drucker-Prager equivalent stress also reveals some remarks. As can be seen in the equations 7 and 8, $J_1$ becomes negative when hydrostatic stress is presented, what means that the yield criterion $F(\sigma)$ becomes smaller. So as discussed by [22], a larger external load is necessary to cause the yielding of an element.

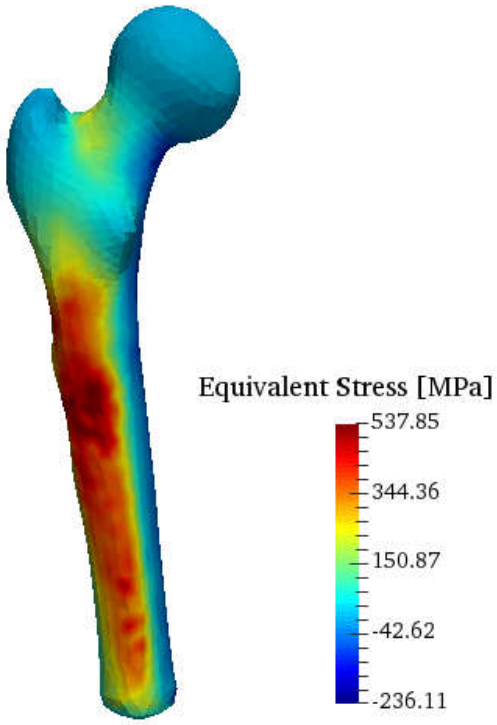



Fig. 8. Drucker-Prager equivalent stress distribution for FE model with inhomogeneous elastic modulus.

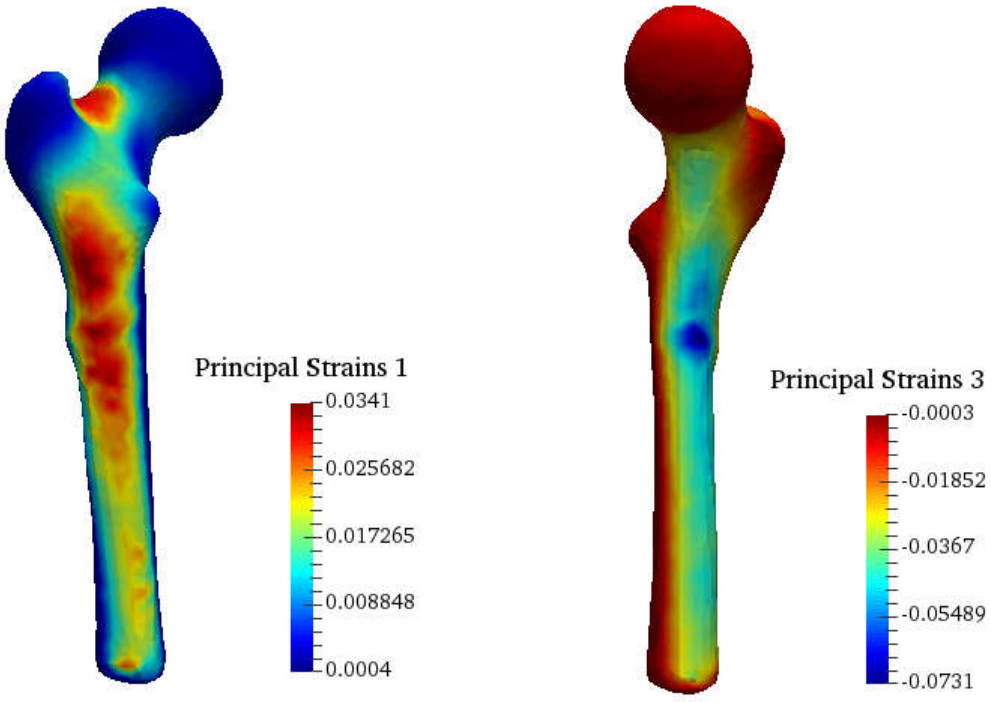



Fig. 9. Maximum (tensile) e minimum (compressive) principal strains (left and right, respectively) in FE model with inhomogeneous elastic modulus.

The maximum Drucker-Prager equivalent stress obtained in this simulation for inhomogeneous and homogeneous models were rather closer, 537.85 MPa and 528.28 MPa, respectively. These values will be used in a future work to access whether the equivalent stress exceed the element yield stress, which allows to evaluate locally the failure sites as shown in [22].

The maximum (tensile) and minimum (compressive) principal strains are shown on Figs. 9 and 10 for inhomogeneous and homogeneous elastic modulus. The results for tensile principal strains were similar but there were significant differences for compressive principal strains. The values for maximum and minimum principal strains are also are listed on Table 1. An important result can be observed in these figures. The regions which present maximum and minimum principal strains (tensile and compressive) are in a region very close to the real fracture, indicating that this model can be used to predict the failure site in a perspective of subject-specific FE models.

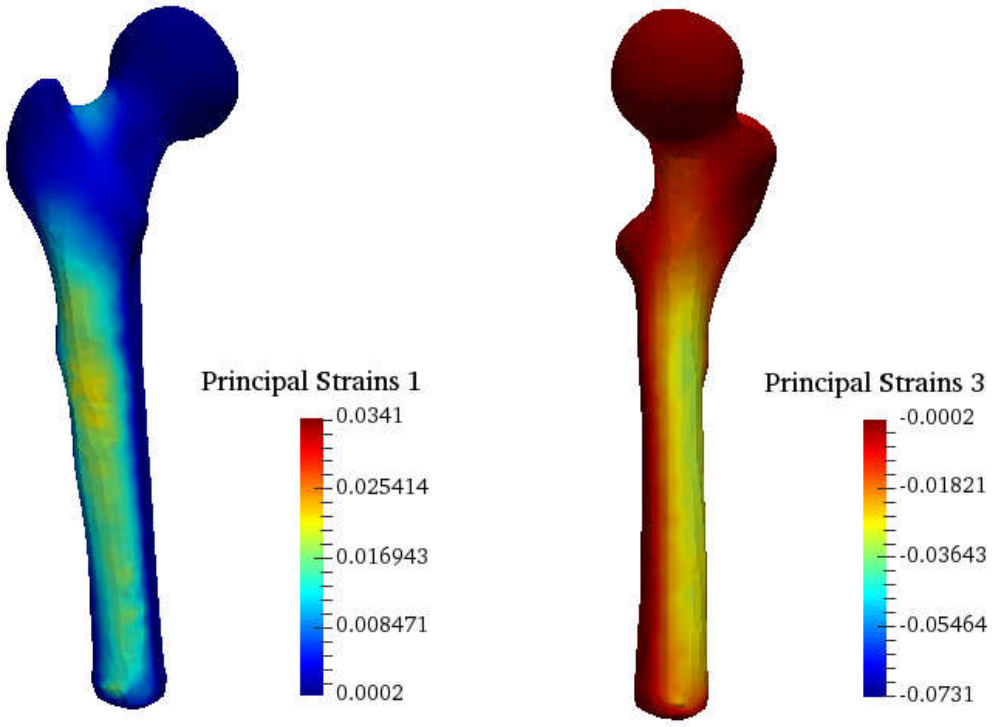



Fig. 10. Maximum (tensile) e minimum (compressive) principal strains (left and right, respectively) in FE model with homogeneous elastic modulus.

## 4 DISCUSSION

Perform a structural analysis in a reconstructed bone as show in this work presents some limitations. First of all, there is rarely specific experimental data for each patient under study to compare with the numerical simulation, which led to adopt values available in the literature. This option may seem rather strange considering that we are interested in a subject-specific model. Other important note is the impossibility to recover the real boundary conditions (muscle and ligaments forces, joint reactions) involved before fracture. So it is very difficult to achieve in the numerical simulation the same location of the real fracture without knowledge and strict control of the boundary conditions. Choose real or almost real boundary conditions constitute a major challenge in this type of numerical simulation. In order to over-

come this challenge or at least minimize its effects it is convenient analyse different scenarios of loading to cover the entire range of loading cases.

The bone modulus-density relationships available in the literature have been demonstrated good agreement with experimental data. Unfortunately, there are few works evaluating the effectiveness of these relationships when we are dealing with specific cases presenting bone diseases, such as osteoporosis and osteopenia cases.

## 5 CONCLUSIONS

- The access of material properties from CT images offers many possibilities to the biomechanics structural analysis. Reliable subject-specific FE models require information about the inhomogeneities materials properties of each particular case in study. The use of the medical image data, however, requires caution considering that obtaining reliable values of bone ash density (used in the modulus-density relationships) must consider the type of CT equipment, calibration, the type of tissue under study, and ideally should have an experimental support for ensure validation
- The procedure presented here for reconstruction of a fractured bone preserving the CT image information offers a new perspective to analyze the causes of bone fractures since allows recover the material state immediately after the failure. Find real boundary conditions remains a challenge, but the results obtained in this research offers new perspectives to investigate.

## REFERENCES

[1] M. Doblaré, J. M. García, Anisotropic bone remodelling model based on a continuum damage-repair theory. *Journal of Biomechanics* 35(1): 1-17, 2002.

[2] E. Schileo, E. Dall'ara, F. Taddei, A. Malandrino, T. Schotkamp, M. Baleani, M. Viceconti, An accurate estimation of bone density improves the accuracy of subject-specific finite element models. *Journal of Biomechanics* 41:2483–2491, 2008.

[3] E. Schileo, F. Taddei, A. Malandrino, L. Cristofolini, M. Viceconti, Subject-specific finite element models can accurately predict strain levels in long bones. *Journal of Biomechanics* 40, 2982–2989, 2007.

[4] M. Viceconti, M. Davinelli, F. Taddei, A. Cappello, Automatic generation of accurate subject-specific bone finite element models to be used in clinical studies. *Journal of Biomechanics* 37, 1597–1605, 2004

[5] J. Y. Rho, M. C. Hobatho, R. B. Ashman, Relations of mechanical properties to density and CT numbers in human bone. *Medical Engineering and Physics*, Vol. 17, No. 5, 347-355, 1995.

[6] D. R. Carter, W. C. Hayes, The compressive behavior of bone as a two-phase porous structure. *The Journal of Bone & Joint Surgery,* 59(7):954–62, 1977.

[7] F. Taddei, A. Pancanti, M. Viceconti, An improved method for the automatic mapping of computed tomography numbers onto finite element models, *Medical Engineering & Physics,* 26: 61–69, 2004.

[8] H. M. Kjer, J. Wilm, Evaluation of surface registration algorithms for PET motion correction, Bachelor thesis, Technical University of Denmark, Informatics and Mathematical Modelling, 2010

[9] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, pages 2724–2729 vol.3, 1991.

[10] P. J. Besl and H. D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[11] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987.

[12] N. Trabelsia, C. Milgrom, Z. Yosibash, Patient-specific FE analyses of metatarsal bones with inhomogeneous isotropic material properties. *Journal of Mechanical Behaviour of Medical Materials,* 20: 177-189, 2014.

[13] K. H. Kristen, , K. Berger, , C. Berger, , W. Kampla, , W. Anzbock, , S.H. Weitzel, The first metatarsal bone under loading conditions: a finite element analysis. *Foot and Ankle Clinics of North America* 10 (1), 1–14, 2005.

[14] J. M. Garcia-Aznar, J. Bayod, A. Rosas, R. Larrainzar, R. Garcia-Bogalo, M. Doblare, L.F. Llanos, Load transfer mechanism for different metatarsal geometries: a finite element study. *Journal of Biomechanical Engineering* 131 (2) 021011–1–021011–7, 2009.

[15] Y. D. Gu, X. J. Ren, J. S. Li, M. J. Lake, Q. Y. Zhang, Y. J. Zeng, Computer simulation of stress distribution in the metatarsals at different inversion landing angles using the finite element method. *International Orthopaedics* 34 (5), 669–676, 2010.

[16] T. S. Keller, Predicting the compressive mechanical behavior of bone. *Journal of Biomechanics* 27, 1159–1168, 1994.

[17] Yosibash, Z., Trabelsi, N., Milgrom, C., Reliable simulations of the human proximal femur by high-order finite element analysis validated by experimental observations. Journal of Biomechanics 40 (16), 3688–3699, 2007.

[18] N. Trabelsi, , Z. Yosibash, , C. Milgrom, Validation of subjectspecific automated p-FE analysis of the proximal femur. *Journal of Biomechanics* 42 (3), 234–241, 2009.

[19] C. Zannoni, R. Mantovani, M. Viceconti, Material properties assignment to finite element models of bone structures: a new method. *Medical Engineering & Physics*, 20: 735–740, 1998.

[20] Firoozbakhsh K, Cowin SC. An analytical model of Pauwels' functional adaptation mechanism in bone. *Journal of Biomechanical Engineering*. Nov; 103(4):246-52, 1981.

[21] J. Nedoma, J. Stehlik, *Mathematical and Computational Methods and Algorithms in Biomechanics: Human Skeletal Systems*, John Wiley & Sons, 2011.

[22] M. Bessho, I. Ohnishi, J. Matsuyama, T. Matsumoto, K. Imai K, K. Nakamura, Prediction of strength and strain of the proximal femur by a CT-based finite element method, *Journal of Biomechanics*, 40(8):1745-53 , 2007.

[23] J. Cordey, E. Gautier, Strain gauges used in the mechanical testing of bones Part I: Theoretical and technical aspects, *Injury*,30 (Suppl 1): A7 – 13, 1999.

[24] H. B. Kupfer, K. H. Gerstle, Behavior of Concrete Under Biaxial Stresses, *Journal of Engineering Mechanical Divisi*on, ASCE 99, 853 – 866, 1973.

[25] H. H. Bayraktar, E. F. Morgan, G. L. Niebur, G. E. Morris, E. W. Wong, T. M. Keaveny, Comparison of the elastic and yield properties of human femoral trabecular and cortical bone tissue, *Journal of Biomechanics*, Volume 37 , Issue 1 , 27 – 35, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SOLVING THE INVERSE PROBLEM OF ESTIMATING FUZZY VISCOELASTIC CONSTITUTIVE PARAMETERS

**Ruifei Peng[1], Haitian Yang[1]***

[1]Department of Engineering Mechanics
Dalian University of Technology
No.2 Linggong Road, Ganjingzi District, Dalian City, Liaoning Province, P.R.C., 116024
pengruifei@mail.dlut.edu.cn, haitian@dlut.edu.cn

**Keywords:** Viscoelasticity, Inverse fuzzy arithmetic, General transformation method, Scaled boundary finite element method.

**Abstract.** *This paper presents a numerical approach to solve the inverse problem of estimating fuzzy viscoelastic constitutive parameters. The direct fuzzy viscoelastic problem is solved by utilizing Scale Boundary Finite Element Method(SBFEM), a temporally adaptive algorithms and the General Transformation Method (GTM). On the basis of the solution of direct problem, the fuzzy estimation can be realized using an inverse fuzzy arithmetic and an evolutionary algorithm of Differential Evolution(DE). Two numerical examples are provide to verify the proposed approach.*

## 1 INTRODUCTION

The experiment based estimation of viscoelastic constitutive parameters is fundamental for the viscoelastic analysis, and is an important issue related with many engineering aspects. In the estimation procedure, there are various uncertainties caused by measurements, observation, and modeling etc.. Such uncertainties may result in fault results, and necessitate to take into account.

As a matter of fact, there is a big pool containing a number of literatures related with analytical and numerical analysis for the estimation of viscoelastic constitutive parameters, However, to the best of authors' knowledge, nearly all these literatures are based on the deterministic assumption.

There are mainly three mathematical means to describe the uncertainty, including probabilistic method, interval analysis and fuzzy theory that can account epistemic uncertainties caused by a lack of knowledge or imperfections in the modeling procedure[1-3],and is becoming increasingly popular for the analysis of numerical models that incorporate uncertainty in their description [4-5].

In the context of fuzzy analysis, many well documented experimental, numerical and analytical results exist in literature concerned with various aspects, but there seems no any report directly related with the inverse fuzzy viscoelastic problems of parameters estimation.

This paper aims at developing a numerical approach to solve inverse problems of estimating unknown fuzzy constitutive parameters.

## 2 VISCOELASTIC CONSTITUTIVE RELATIONSHIP

In this paper, the viscoelastic constitutive relationships is specified by a three-parameter solid viscoelasticity model (see Figure 1) in a differential form [6]

$$\alpha_1 \mathbf{D}\boldsymbol{\varepsilon}(t) + p_1 \mathbf{D}\frac{\mathrm{d}\boldsymbol{\varepsilon}(t)}{\mathrm{d}t} = (\alpha_1 + \alpha_2)\boldsymbol{\sigma}(t) + p_1 \frac{\mathrm{d}\boldsymbol{\sigma}(t)}{\mathrm{d}t} \qquad (t > 0) \tag{1}$$

$$\boldsymbol{\sigma}(t) = \mathbf{D}\boldsymbol{\varepsilon}(t) \qquad (t = 0) \tag{2}$$

where

$$\mathbf{D} = \frac{E_2}{1-\nu^2}\begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \dfrac{1-\nu}{2} \end{pmatrix} \quad \text{for the plane stress problem} \tag{3}$$

$$p_1 = \frac{\eta_1}{E_1 + E_2} \qquad \alpha_1 = \frac{E_1}{E_1 + E_2} \qquad \alpha_2 = \frac{E_2}{E_1 + E_2} \tag{4}$$

where $\nu$, $E_1$, $E_2$ and $\eta_1$ are constitutive parameters.

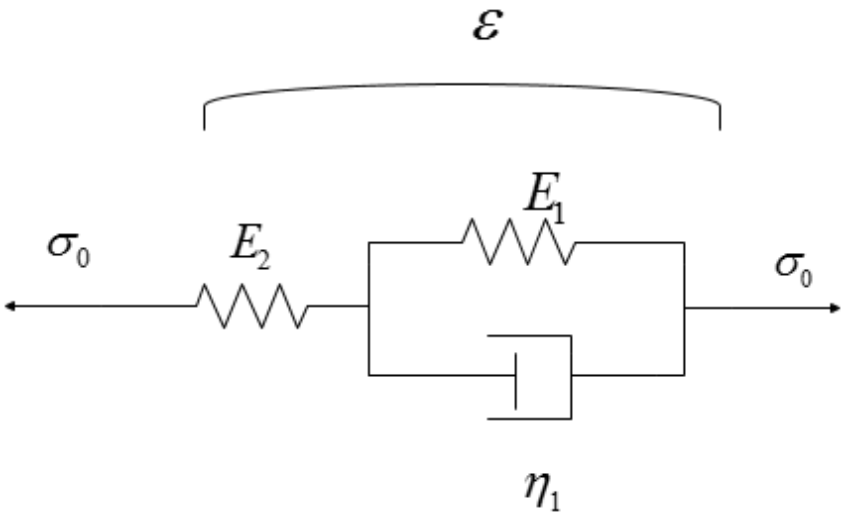



Figure 1: Three-parameter solid viscoelasticity model.

## 3 DESCRIPTION OF SOLVING DIRECT VISOELASTIC PROBLEMS WITH FUZZY UNCERTAINTY

In the direct visoelastic problem, $E_1, E_2, \eta_1$, and $\nu$, are defined by fuzzy numbers $\tilde{p}_i (i = 1,2,3,4)$, the corresponding unknown fuzzy output of displacement, $u(t)$, is defined by $\tilde{q}$.

The deterministic relationship between, $E_1, E_2, \eta_1, \nu$ and $u_s(t, \vec{x})$ can be obtained in the form [7]

$$u_s(t,\vec{x}) = u(t, \vec{x}, E_1, E_2, \eta_1, \nu..) \quad , s = 1,2,3 \tag{5}$$

where t stands for time, and $\vec{x}$ refers to a vector of coordinates.

The indeterministic relationship between $E_1, E_2, \eta_1, \nu$ and $u_s(t, \vec{x})$ is defined by

$$\tilde{q}_r = \tilde{q}_r(\tilde{p}_1, \tilde{p}_2 \tilde{p}_3 . \tag{6}$$

where $\tilde{p}_i = \left\{(x_i, \mu_{\tilde{p}_i}(x_i)) \mid \mu_{\tilde{p}_i}(x_i) \in [0,1]\right\}$, $x_i$ is a value of $\tilde{p}_i$ , $i = 1,2,...,4$

$\tilde{q}_r, r = 1,2,...,N$ is a set of N fuzzy-valued outputs at *r-th* degree of freedom (DOF) in SBFEM.

If $\mu_{\tilde{p}_i}(x_i) = 1$, $x_i$ is a member of the fuzzy set $\tilde{p}_i$.

If $\mu_{\tilde{p}_i}(x_i) = 0$, $x_i$ is not a member of the fuzzy set $\tilde{p}_i$.

If $0 < \mu_{\tilde{p}_i}(x_i) < 1$, the membership of $x_i$ is uncertain.

Using Eq.(5) and the General transformation method[8], $\tilde{q}_r$ can be obtained. The major steps to determine the membership function of $\tilde{q}_r$ using GTM include

(1) Subdividing $\mu_{\tilde{p}_i}$ into a number of $mm$ segments, equally spaced by

$$\Delta\mu_{\tilde{p}_i} = \frac{1}{mm} \quad ,i = 1,2,...,4 \tag{7}$$

where $mm$ is the number of discrete segments.

(2) Decompose $\tilde{p}_i$ into a number of intervals

$X_i^{(j)} = [a_i^{(j)}, \mathrm{b}_i^{(j)}], a_i^{(j)} \le \mathrm{b}_i^{(j)}, i = 1,2,...,4, \ j = 0,1,...,mm$ by the $\alpha$-cut at the *j-th* level ,and set $c_{l,i}^{(j)}$ within $X_i^{(j)}$, $l = 2,3,...,mm - j, \ j = 0,1,...,mm-2$ by

$$c_{l,i}^{(j)} = \begin{cases} a_i^{(j)} & l = 1, \ j = 0,1,...,mm \\ \dfrac{(\mathrm{c}_{l-1,i}^{(j+1)} + \mathrm{c}_{l,i}^{(j+1)})}{2} & l = 2,3,...,mm - j, \ j = 0,1,...,mm-2 \\ b_i^{(j)} & l = mm - j + 1, \ j = 0,1,...,mm \end{cases} \tag{8}$$

$a_i^{(j)}, b_i^{(j)}$ denote the lower and upper bounds of the interval at the *j-th* membership level for $\tilde{p}_i$.

The decomposition procedure is shown in Figure 2.

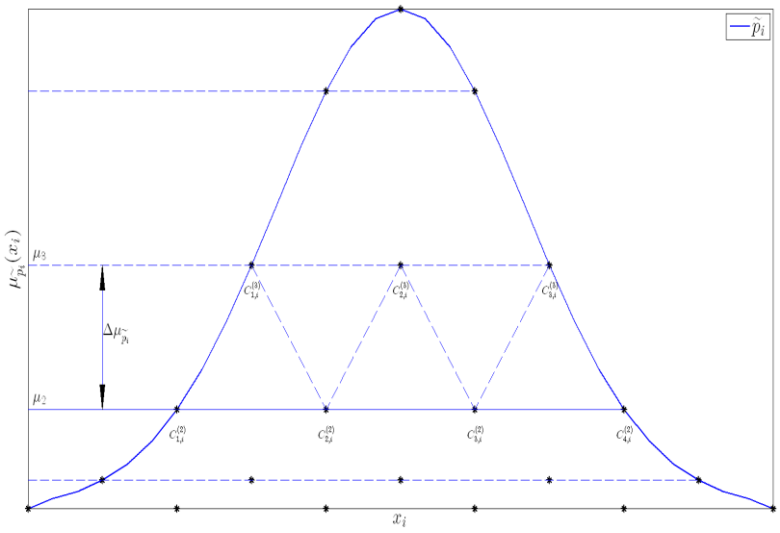



Figure 2:Decomposition of the i-*th* uncertain parameter $\tilde{p}_i$

(3) Construct a matrix $X^{(j)}$ at the $j$-*th* level

$$X^{(j)} = \begin{bmatrix} X_1^{(j)} \\ X_2^{(j)} \\ \cdots \\ X_4^{(j)} \end{bmatrix} \tag{9}$$

where $$X_i^{(j)} = \underbrace{(\gamma_{1,i}^{(j)}, \gamma_{2,i}^{(j)}, \ldots, \gamma_{(mm+1-j),i}^{(j)}), \ldots, (\gamma_{1,i}^{(j)}, \gamma_{2,i}^{(j)}, \ldots, \gamma_{(mm+1-j),i}^{(j)})}_{(mm+1-j)^{i-1} \quad (mm+1-j) \ tuples} \tag{10}$$

$$\gamma_{l,i}^{(j)} = \underbrace{(c_{l,i}^{(j)}, \ldots, c_{l,i}^{(j)})}_{(mm+1-j)^{4-i} \ elements} \tag{11}$$

$$i = 1,2,\ldots,4,\ j = 0,1,2,\ldots,mm,\ l = 1,2,3,\ldots,mm+1-j \tag{12}$$

(4) Determine ${}^k\hat{q}_r^{(j)}$ at $j$-*th* level

$${}^k\hat{q}_r^{(j)} = \tilde{q}_r({}^kX_1^{(j)}, {}^kX_2^{(j)}, \ldots, {}^kX_4^{(j)}, t),\ j = 0,1,2,\ldots,mm,\ k = 1,2,\ldots,(mm+1-j)^4 \tag{13}$$

where ${}^k\hat{q}_r^{(j)}$ denotes the $k$-*th* element of $\hat{q}_r^{(j)}$ and ${}^kX_i^{(j)}, i = 1,2,\ldots,4$ denotes the $k$-*th* column of the matrix $X^{(j)}$.

(5) Evaluate $\tilde{q}_r$ in a decomposed form

$$\tilde{q}_r^{(j)} = [a^{(j)}, \mathrm{b}^{(j)}], a^{(j)} \le \mathrm{b}^{(j)},\ j = 0,1,\ldots,mm \tag{14}$$

where

$$\begin{aligned} a^{(j)} &= \min_k(a^{(j+1)}, {}^k\hat{q}_r^{(j)}) \\ b^{(j)} &= \max_k(\mathrm{b}^{(j+1)}, {}^k\hat{q}_r^{(j)}) \end{aligned},\ j = 0,1,\ldots,mm-1 \tag{15}$$

$$a^{(mm)} = \min_k({}^k\hat{q}_r^{(mm)}) \Rightarrow \max_k({}^k\hat{q}_r^{(mm)}) = a^{(mm)} \tag{16}$$

$\tilde{q}_r^{(j)}$ is a decomposition interval for $\tilde{q}_r$ at the $j$-*th* level $\mu_j$.

(6) Evaluate $\tilde{q}_r$ in an ensemble form

By recomposing $\tilde{q}_r^{(j)}$ according to their levels of membership $\mu_j$, an approximation for the ensemble $\tilde{q}_r$ can be achieved.

## 4 DESCRIPTION OF SOLVING INVERSE VISCOELASTIC PROBLEMS WITH FUZZY UNCERTAINTY

The target of solving inverse fuzzy viscoelastic problems is to seek membership functions of unknown constitutive parameter $E_1, E_2, \eta_1, \nu$, it can be achieved by an inverse fuzzy arithmetic proposed by Hanss [9] with Eq. (5) via some partially known membership functions of $u(t)$ at some discrete time points $t_0, t_1, t_2, ..., t_{nn}$ .

The major procedure is described as following.

(1) Evaluation of peak values

Under the condition in which $\tilde{q}$ is strictly monotonic with respect to each of $\tilde{p}_i$ , $\overline{x}_1$, $\overline{x}_2$,..., $\overline{x}_4$ , the peak values of $\tilde{p}_i, i = 1,2,...,n$ can be evaluate via the inverse solutions of following equations.

$$u(t_{\mathrm{m}}) = u(t_m,\ \overline{x}_1, \overline{x}_2, .., \overline{x}_4)\ \ (m = t_0, t_1, t_2, ..., t_{nn}) \tag{17}$$

In this paper, the inverse solutions are given by DE [10].

(2) Computation of gain factors

Single-sided gain factors $\eta_{i+}^{(j)}, \eta_{i-}^{(j)}$ ,which describes the effect of the *i-th* uncertainty parameter $\tilde{p}_i$ on the uncertainty of $\tilde{q}_r$ when only positive deviations and negative deviations from the peak value $i = 1,2,...,4$ are considered, are given by

$$\eta_i^{(j)} = \frac{1}{2^3 (b_i^{(j)} - a_i^{(j)})} \times \sum_{k=1}^{2^{4-i}} \sum_{l=1}^{2^{i-1}} ({}^{k_2}\hat{q}_r^{(j)} - {}^{k_1}\hat{q}_r^{(j)}) \tag{18}$$

$$\begin{aligned} & i = 1,2,...,4,\ j = 0,1,...,mm-1 \\ & k_1 = k + (l-1)2^{5-i}, \\ & k_2 = k + (2l-1)2^{4-i}. \end{aligned} \tag{19}$$

where $a_i^{(j)}$ and $b_i^{(j)}$ are guessed lower and upper bounds of $\tilde{p}_i$ at membership $\mu_j$, ${}^{k_2}\hat{q}_r^{(j)}$ and ${}^{k_1}\hat{q}_r^{(j)}$ refer to Eq. (13).

(3) Assembly of $\tilde{p}_1, \tilde{p}_2, ..., \tilde{p}_4$

$\breve{a}_i^{(j)}$ and $\breve{b}_i^{(j)}$, which are the lower and upper bounds of the intervals of $\tilde{p}_i$ at membership $\mu_j$ can be obtained by

$$H^{(j)}\begin{bmatrix} \breve{a}_1^{(j)}-\overline{x}_1 \\ \breve{b}_1^{(j)}-\overline{x}_1 \\ \breve{a}_2^{(j)}-\overline{x}_2 \\ \breve{b}_2^{(j)}-\overline{x}_2 \\ \vdots \\ \breve{a}_4^{(j)}-\overline{x}_4 \\ \breve{b}_4^{(j)}-\overline{x}_4 \end{bmatrix}=\begin{bmatrix} c_1^{(j)}-\overline{q}_1 \\ d_1^{(j)}-\overline{q}_1 \\ c_2^{(j)}-\overline{q}_2 \\ d_2^{(j)}-\overline{q}_2 \\ \vdots \\ c_{NN}^{(j)}-\overline{q}_{NN} \\ d_{NN}^{(j)}-\overline{q}_{NN} \end{bmatrix} \tag{20}$$

where

$$H^{(j)}=\begin{bmatrix} H_{11}^{(j)} & H_{12}^{(j)} & \cdots & H_{14}^{(j)} \\ H_{21}^{(j)} & H_{22}^{(j)} & \cdots & H_{24}^{(j)} \\ \vdots & \vdots & \vdots & \vdots \\ H_{\mathrm{NN}1}^{(j)} & H_{NN2}^{(j)} & \cdots & H_{NN4}^{(j)} \end{bmatrix} \tag{21}$$

$$H_{ri}^{(j)}=\frac{1}{2}\begin{bmatrix} \eta_{i-}^{(j)}(1+\mathrm{sgn}(\eta_{i-}^{(j)})) & \eta_{i+}^{(j)}(1-\mathrm{sgn}(\eta_{i+}^{(j)})) \\ \eta_{i-}^{(j)}(1-\mathrm{sgn}(\eta_{i-}^{(j)})) & \eta_{i+}^{(j)}(1+\mathrm{sgn}(\eta_{i+}^{(j)})) \end{bmatrix} \tag{22}$$

$$i=1,2,...4.r=1,2,...NN$$

$c_i^{(j)}$ and $d_i^{(j)}$ are the known lower and upper bounds of the intervals of $\tilde{q}_r$ at membership $\mu_j$ .NN is the number of DOF with known output information.

The membership functions of the unknown $\tilde{p}_i$ can be described by $\breve{a}_i^{(j)}$ , $\breve{b}_i^{(j)}$ and $\overline{x}_k$, $i,k=1,2,...,4$.

## 5 NUMERICAL EXAMPLES

For the simplicity, all variables are assumed dimensionless. The DE parameters are set as following: maximum number of iterations is 1000, population size is 50, lower bound of scaling factor is 0.2, upper bound of scaling factor is 0.8, crossover probability is 0.2.

**Example 1** Considers the three-parameter solid viscoelasticity model(see Figure 1) under a constant loading $\sigma_0$. In this case, the model material is rock, and the constitutive relationship is defined by

$$\varepsilon(t)=\frac{\sigma_0}{E_2}+\frac{\sigma_0}{E_1}(1-e^{-\frac{E_1}{\eta_1}t}) \tag{23}$$

where $\sigma_0=1$ $E_2=9800$. $E_1,\eta_1$ refers to $\tilde{p}_1,\tilde{p}_2$, and is defined by

$$\mu_{\tilde{p}_i}(x_i)=\begin{cases}\exp\left[-\dfrac{(x_i-\bar{x}_i)^2}{2d_i^2}\right], for\left|x_i-\bar{x}_i\right|<3d_i\\ 0,\ for\left|x_i-\bar{x}_i\right|\geq 3d_i\end{cases},i=1,2 \tag{24}$$

where

$$\begin{matrix}\bar{x}_1=1960 & \bar{x}_2=52083\\ d_1=98 & d_1=2604.15\end{matrix}, \tag{25}$$

The range of time is [0,10] and the time step is set to 0.1.

Figure 3 shows the membership function of $\tilde{p}_1$, $\tilde{p}_2$ where $mm$ is set to 20.

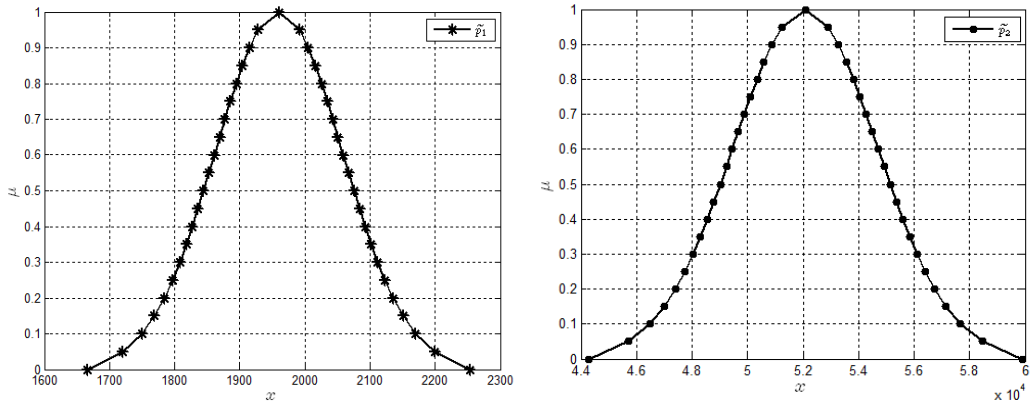
Figure 4 shows the fuzzy response of $\varepsilon$ with respect to prescribed $E_1$, $\eta_1$.

The membership functions of $\varepsilon$ at *t*=0,0.1,0.2,0.3..1.4 are employed as known information.

Figure 5 exhibits a comparison of $\tilde{p}_1$ and $\tilde{p}_2$ given by the proposed approach with prescribed ones, the maximum relative error between prescribed and estimated $\tilde{p}_1$ is 2.60%,and 1.04% between prescribed and estimated $\tilde{p}_2$.

Figure 6 shows the iteration history of DE at *0-th* membership level used in the step of inversion.

The interval boundary of prescribed and estimated $\tilde{p}_1$, $\tilde{p}_2$ at each membership is shown in Table 1 and Table 2 respectively where LB means left boundary and RB means right boundary.




(a) (b)

Figure 3:Symmetrical Gaussian fuzzy numbers $\tilde{p}_1$ and $\tilde{p}_2$ .(a) $\tilde{p}_1$ .(b) $\tilde{p}_2$

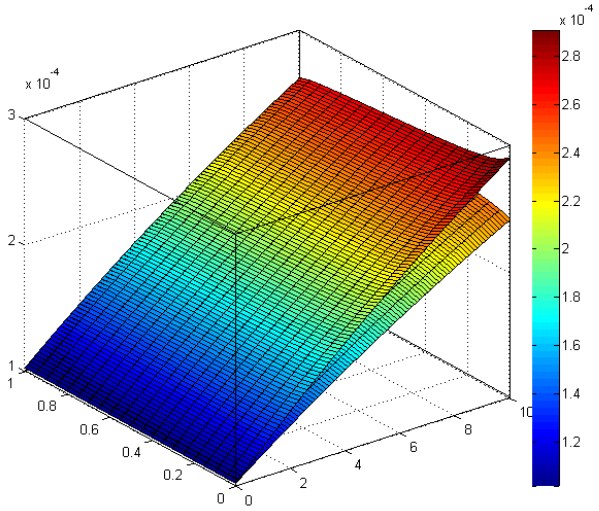
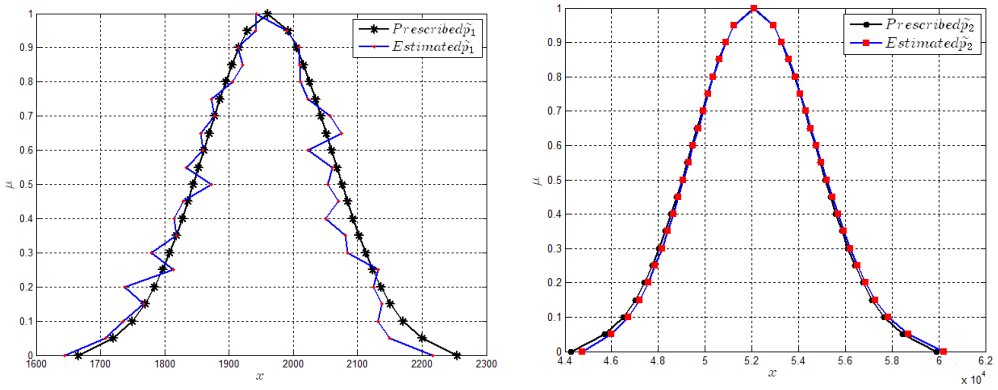



Figure 4: The fuzzy response of $\varepsilon$




(a) (b)

Figure 5 A comparison of prescribed and estimated $\tilde{p}_1$, $\tilde{p}_2$ .(a).comparison of prescribed and estimated $\tilde{p}_1$ .(b). comparison of prescribed and estimated $\tilde{p}_2$ .

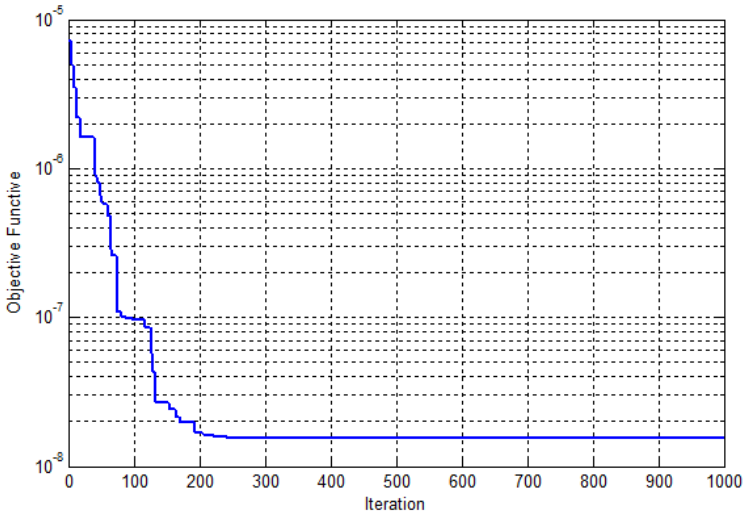



Figure 6: Iteration history of DE at *0-th* membership level.

Table 1:The interval boundary of prescribed $\tilde{p}_1$ , $\tilde{p}_2$ at each membership.

| Membership $\mu_{\tilde{p}_i}(x_i)$ | LB of $\tilde{p}_1$ | RB of $\tilde{p}_1$ | LB of $\tilde{p}_2$ | RB of $\tilde{p}_2$ |
|---|---|---|---|---|
| 1.000000 | 1960.000000 | 1960.000000 | 52083.000000 | 52083.000000 |
| 0.950000 | 1928.611441 | 1991.388558 | 51248.913118 | 52917.086881 |
| 0.900000 | 1915.013726 | 2004.986273 | 50887.581595 | 53278.418404 |
| 0.850000 | 1904.128150 | 2015.871849 | 50598.319616 | 53567.680383 |
| 0.800000 | 1894.531371 | 2025.468628 | 50343.304803 | 53822.695196 |
| 0.750000 | 1885.664293 | 2034.335706 | 50107.680307 | 54058.319692 |
| 0.700000 | 1877.229157 | 2042.770842 | 49883.533787 | 54282.466212 |
| 0.650000 | 1869.035840 | 2050.964159 | 49665.813111 | 54500.186888 |
| 0.600000 | 1860.944770 | 2059.055229 | 49450.809417 | 54715.190582 |
| 0.550000 | 1852.840058 | 2067.159941 | 49235.443242 | 54930.556757 |
| 0.500000 | 1844.613817 | 2075.386182 | 49016.847689 | 55149.152310 |
| 0.450000 | 1836.154387 | 2083.845612 | 48792.055583 | 55373.944416 |
| 0.400000 | 1827.334584 | 2092.665415 | 48557.687338 | 55608.312661 |
| 0.350000 | 1817.996537 | 2102.003462 | 48309.547786 | 55856.452213 |
| 0.300000 | 1807.927945 | 2112.072054 | 48041.995514 | 56124.004485 |
| 0.250000 | 1796.819296 | 2123.180703 | 47746.805818 | 56419.194181 |
| 0.200000 | 1784.175987 | 2135.824012 | 47410.835688 | 56755.164311 |
| 0.150000 | 1769.107672 | 2150.892327 | 47010.425974 | 57155.574025 |
| 0.100000 | 1749.695329 | 2170.304670 | 46494.582572 | 57671.417427 |
| 0.050000 | 1720.120810 | 2199.879189 | 45708.700090 | 58457.299909 |
| 0.000000 | 1666.000000 | 2254.000000 | 44270.550000 | 59895.450000 |

Table 2:The interval boundary of estimated $\tilde{p}_1$, $\tilde{p}_2$ at each membership.

| Membership $\mu_{\tilde{p}_i}(x_i)$ | LB of $\tilde{p}_1$ | RB of $\tilde{p}_1$ | LB of $\tilde{p}_2$ | RB of $\tilde{p}_2$ |
|---|---|---|---|---|
| 1.000000 | 1942.091888 | 1942.091888 | 52092.133589 | 52092.133589 |
| 0.950000 | 1941.704284 | 1989.675237 | 51241.514079 | 52924.630769 |
| 0.900000 | 1913.567391 | 2008.507551 | 50895.673356 | 53284.662664 |
| 0.850000 | 1920.806585 | 2009.527200 | 50600.240660 | 53582.141626 |
| 0.800000 | 1905.677396 | 2011.281705 | 50356.896392 | 53849.245699 |
| 0.750000 | 1872.697636 | 2022.596843 | 50135.697758 | 54084.679594 |
| 0.700000 | 1877.325231 | 2057.531698 | 49919.216384 | 54302.629658 |
| 0.650000 | 1855.714355 | 2074.421790 | 49709.882968 | 54518.237541 |
| 0.600000 | 1860.045326 | 2022.486221 | 49492.167331 | 54777.565568 |
| 0.550000 | 1833.528683 | 2060.997639 | 49299.879755 | 54974.083302 |
| 0.500000 | 1872.131222 | 2053.463510 | 49067.589422 | 55208.184748 |
| 0.450000 | 1829.143780 | 2070.269237 | 48866.174422 | 55443.235282 |
| 0.400000 | 1814.798895 | 2049.936918 | 48646.071934 | 55699.262524 |
| 0.350000 | 1817.706941 | 2080.728220 | 48403.983274 | 55946.477782 |
| 0.300000 | 1779.252500 | 2083.824668 | 48167.880228 | 56224.833575 |
| 0.250000 | 1812.915434 | 2131.030043 | 47867.654497 | 56507.772061 |
| 0.200000 | 1737.820053 | 2124.651896 | 47582.681851 | 56879.581643 |
| 0.150000 | 1767.853034 | 2137.985729 | 47188.502901 | 57297.399235 |
| 0.100000 | 1735.500517 | 2131.909586 | 46718.589420 | 57848.333607 |
| 0.050000 | 1708.101554 | 2149.767512 | 46002.731009 | 58698.303257 |
| 0.000000 | 1644.807367 | 2216.345710 | 44730.752739 | 60209.475501 |

**Example 2** Considers an estimation of $E_1$ and $\eta_1$ in Figure 1 for a square viscoelastic plate subjected to a uniform tension $p=1$ as shown in Figure 7 where $a=1, E_2=9800, \nu=0.3$.Figure 8 describes a node distribution of SBFEM with an adjacent node distance $ds=0.1$.

The membership functions of $E_1$ and $\eta_1$, refer to $\tilde{p}_1$ and $\tilde{p}_2$ defined by Eqs.(24-25), are prescribed in Figure 9 where $mm$ is set to 10.

The range of time is [0,10] and the time step is set to 0.1.

The fuzzy response of displacements are given using Eq. (5), and the membership functions of displacement of nodes 12,13,14,..20 along x direction at t=1,2,3..10 are

employed as known information.

Figure 10 shows the fuzzy response of $u_x$ at node 16 with respect to prescribed $E_1$, $\eta_1$ where $u_x$ is the displacement along x direction.

Figure 11 exhibits a comparison of $\tilde{p}_1$ and $\tilde{p}_2$ given by the proposed approach with prescribed ones, the maximum relative error between prescribed and estimated $\tilde{p}_1$ is 0.74%,and 0.59% between prescribed and estimated $\tilde{p}_2$.

Figure 12 shows the iteration history of DE at 0-th membership level used in the step of inversion.

The interval boundary of prescribed and estimated $\tilde{p}_1$, $\tilde{p}_2$ at each membership is shown in Table 3 and Table 4 respectively.

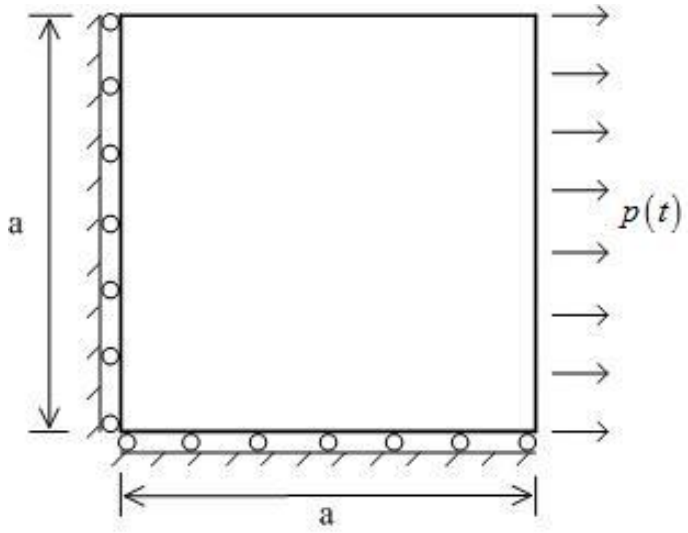



Figure 7: A square viscoelastic plate subjected to a uniform tension

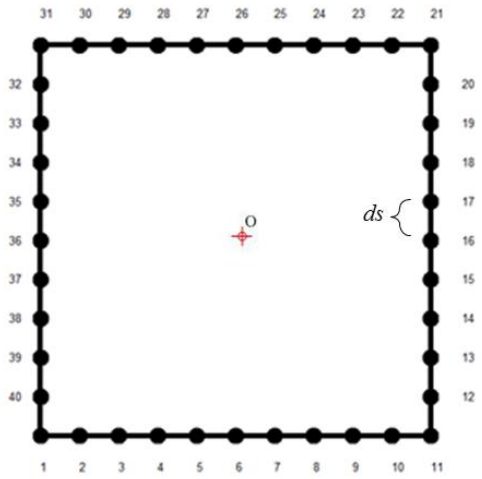
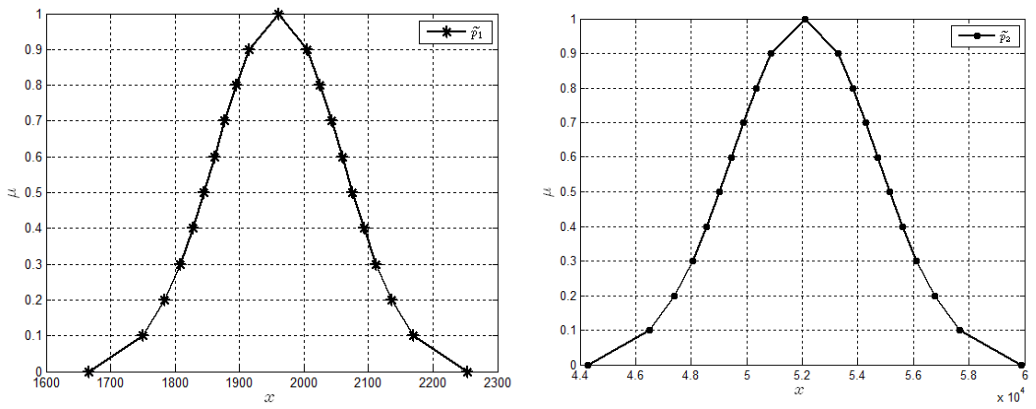



Figure 8: Node distribution of SBFEM




(a) (b)

Figure 9: Symmetrical Gaussian fuzzy numbers $\tilde{p}_1$ and $\tilde{p}_2$ .(a) $\tilde{p}_1$ .(b) $\tilde{p}_2$

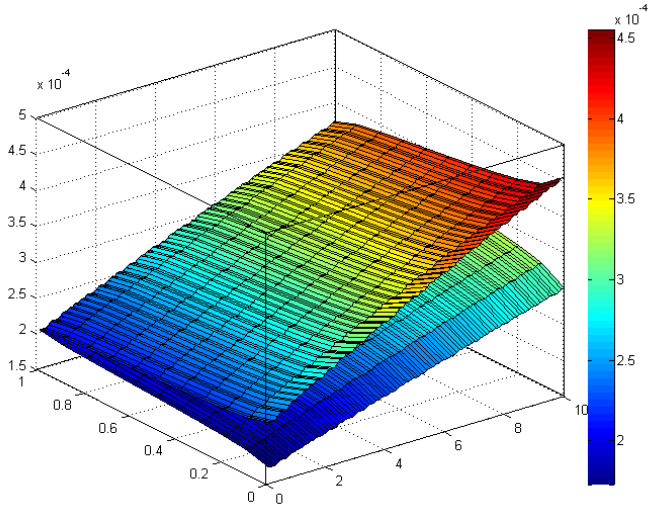
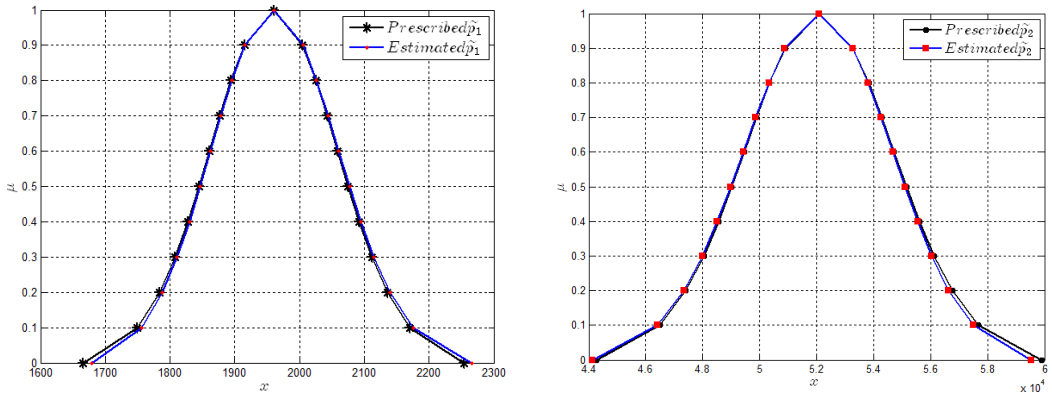



Figure 10: The fuzzy response of $u_x$ at node 16.




(a) (b)

Figure 11: A comparison of prescribed and estimated $\tilde{p}_1$ , $\tilde{p}_2$ .(a).comparison of prescribed and estimated $\tilde{p}_1$ .(b). comparison of prescribed and estimated $\tilde{p}_2$ .

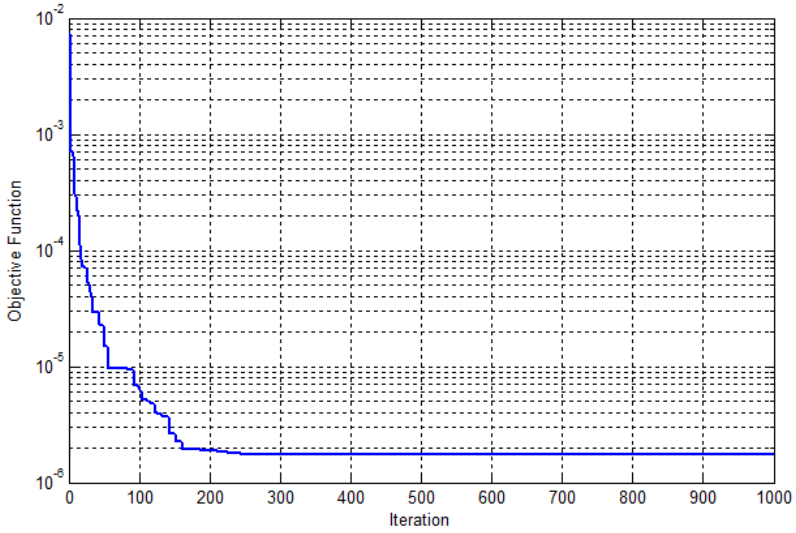



Figure 12: Iteration history of DE at 0-th membership level.

Table 3:The interval boundary of prescribed $\tilde{p}_1$, $\tilde{p}_2$ at each membership.

| Membership $\mu_{\tilde{p}_i}(x_i)$ | LB of $\tilde{p}_1$ | RB of $\tilde{p}_1$ | LB of $\tilde{p}_2$ | RB of $\tilde{p}_2$ |
|---|---|---|---|---|
| 1.000000 | 1960.000000 | 1960.000000 | 52083.000000 | 52083.000000 |
| 0.900000 | 1915.013726 | 2004.986273 | 50887.581595 | 53278.418404 |
| 0.800000 | 1894.531371 | 2025.468628 | 50343.304803 | 53822.695196 |
| 0.700000 | 1877.229157 | 2042.770842 | 49883.533787 | 54282.466212 |
| 0.600000 | 1860.944770 | 2059.055229 | 49450.809417 | 54715.190582 |
| 0.500000 | 1844.613817 | 2075.386182 | 49016.847689 | 55149.152310 |
| 0.400000 | 1827.334584 | 2092.665415 | 48557.687338 | 55608.312661 |
| 0.300000 | 1807.927945 | 2112.072054 | 48041.995514 | 56124.004485 |
| 0.200000 | 1784.175987 | 2135.824012 | 47410.835688 | 56755.164311 |
| 0.100000 | 1749.695329 | 2170.304670 | 46494.582572 | 57671.417427 |

Table 4:The interval boundary of estimated $\tilde{p}_1$, $\tilde{p}_2$ at each membership.

| Membership $\mu_{\tilde{p}_i}(x_i)$ | LB of $\tilde{p}_1$ | RB of $\tilde{p}_1$ | LB of $\tilde{p}_2$ | RB of $\tilde{p}_2$ |
|---|---|---|---|---|
| 1.000000 | 1960.054091 | 1960.054091 | 52084.882501 | 52084.882501 |
| 0.900000 | 1915.336738 | 2005.351962 | 50885.112497 | 53272.798515 |
| 0.800000 | 1895.154611 | 2026.093082 | 50332.132062 | 53809.456212 |
| 0.700000 | 1878.243287 | 2043.728483 | 49865.878574 | 54261.986362 |
| 0.600000 | 1862.373259 | 2060.422459 | 49426.076379 | 54684.788371 |
| 0.500000 | 1846.496256 | 2077.209759 | 48983.346190 | 55104.098930 |
| 0.400000 | 1829.802602 | 2095.060415 | 48514.290231 | 55546.667206 |
| 0.300000 | 1811.175145 | 2115.208505 | 47987.524316 | 56041.476602 |
| 0.200000 | 1788.518746 | 2139.975804 | 47339.448358 | 56641.253246 |
| 0.100000 | 1755.898031 | 2176.257461 | 46401.577000 | 57501.404072 |

## 6 CONCLUSIONS

The major merits of this paper include

1. The fuzzy uncertainty is taken into account for the inverse viscoelastic problem. Although there are a number of reports concerned with fuzzy uncertainty analysis in various aspects, it seems no any directly related with the inverse viscoelastic problem.

2. By integrating SBFEM, GTM, and DE, a numerical approach is presented to solve inverse problems of estimating fuzzy viscoelastic parametes.

3. Numerical verification is provided via comparisons of membership functions given by the proposed approach with prescribed ones.

## REFERENCES


[1] P.L. Woodfield , M. Monde, Estimation of uncertainty in an analytical inverse heat conduction solution. *Experimental Heat Transfer,* **22**(3), 129-143,2009.

[2] O. Hiroyasu, O. Yoshinori, System parameter identification theory and uncertainty analysis methods for multi-zone building heat transfer and infiltration .*Building and Environment* , **54**, 39-52,2012.

[3] T. Haag, S.C. Gonzalez, M. Hanss, Model validation and selection based on inverse fuzzy arithmetic. *Mechanical Systems and Signal Processing* , **32**, 116-134,2012.

[4] S. Chen, E. Nikolaidis, H. H. Cudney, R. Rosca, R. T. Haftka ,Comparison of probabilistic and fuzzy set methods for designing under uncertainty. in *Proceedings of AIAA/ASME/ASCE/AHS/ASCE Structures, Structural Dynamics, and Materials Conference and Exhibit* , American Institute of Aeronautics and Astronautics, Reston, 2860-2874,1999.

[5] B. Lallemand, G. Plessis, T. Tison, P. Level, Neumann expansion for fuzzy finite element analysis. *Engineering computations* ,**16** (5), 572-583,1999.

[6] I. H. Shames, *Elastic and inelastic stress analysis*. CRC Press, 1997.

[7] Y.Q. He, H.T. Yang, A combined approach of SBFEM and precise algorithm in time domain solving viscoelasticity problems. Submitted to *Mechanics of Time-Dependent Materials*.

[8] M. Hanss, The transformation method for the simulation and analysis of systems with uncertain parameters. Fuzzy Sets and Systems, 130(3), 277–289,2002.

[9] M. Hanss, An approach to inverse fuzzy arithmetic. In *Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American*,474-479, IEEE,2003.

[10] R. Storn, K. Price ,Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, **11**(4), 341-359, 1997.

[11] A. Klimke, R.F. Nunes , B. Wohlmuth, Fuzzy arithmetic based on dimension-adaptive sparse grids: a case study of a large-scale finite element model under uncertain parameters , *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems* ,**14**(5), 561-577, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DEVELOPMENT OF A CENTRIFUGAL BLOOD PUMP FOR ECMO AND VAD OPERATIONS

**Shinhwa Choi[1], Nahmkeon Hur[1*], Mohammad Moshfeghi[2], Seongwon Kang[1], Wonjung Kim[1]**

[1]Department of Mechanical Engineering, Sogang University
Seoul,04107 Korea
[*]e-mail: nhur@sogang.ac.kr

[2] Multi-Phenomena CFD ERC, Sogang University
Seoul,04107 Korea
e-mail: mmoshfeghi@sogang.ac.kr

**Keywords:** Centrifugal blood pump, Extra corporeal membrane oxygenation (ECMO), CFD

**Abstract.** *Using cardiopulmonary circulatory assist devices has been increased in the recent years as more models are available in the market. These devices can be employed in the situation during which both cardiac and respiratory support to a patient's heart and lungs have to be provided, either during or after surgeries, for short time or even in the case of severe disease, for a period of weeks. Hence, it is critical to know the details of the phenomena happen inside a blood pump from both mechanical performances (such as pressure head and mechanical efficiency) and biomedical factors (such as hemolysis and thrombosis) and to design an optimum pump from both aspects. This paper investigates development of centrifugal blood pump impeller, specifically with focusing on the performances during ECMO condition. The baseline model is designed by investigating existing commercial pumps and considering results of recirculation, pressure heads and mechanical efficiencies together with their biomechanical performance via Modified Indices of Hemolysis (MIH). Afterword, two more modified models are designed and simulated. Overall, a comprehensive comparison between the results of all three case demonstrate that when impeller radius and prime volume is smaller, recirculation is reduced at impeller and MIH value becomes lower. Additionally, high scalar shear stress is observed near the volute and impeller walls and inside the top cavity gap.*

## 1 INTRODUCTION

As number of patients with cardiovascular disorders has increased, using cardiopulmonary circulatory assist devices has been increased as well. These devices can be used when heart and lungs cannot perform their normal duties, such as surgeries and diseases. Cardiopulmonary circulatory assist devices are divided into two types. First, ventricular assist device (VAD) which are devices for patients of heart failure. And second, extra corporeal membrane oxygenation (ECMO) is devices for patients of cardiovascular and respiratory diseases. VAD is composed of blood pump and driving parts. Pressure drop is comparatively low in blood pump in VAD. On the other hand, an ECMO system is composed of a blood pump, a driving devices and oxygenator. ECMO has higher pressure drop because of its multi-component design [1]. In previous study investigated existing pump to find out effect of impeller diameter and rotational speed on performance of pump running in turbine model [2]. Koki et al. [3] investigated hemolysis inside centrifugal blood pump through flow visualization. Yuki et al. [4] and Yukihiko et al. [5] established the design process of centrifugal blood pump. In order to figure out hemolytic phenomenon inside of centrifugal blood pump, mathematical models are adopted with CFD simulation. The mathematical models are divided into two types [6,7]. First, Eulerian approaches consider all areas contributing to the hemolysis [6]. Second, Lagrangian approaches miss out several high scalar shear stress areas. So Lagrangian approaches can't accurately measure [7]. In this research a baseline model via designed by investigating the existing commercial blood pumps [8], by considering the results of CFD simulations, performance curves and bio-mechanical factor analyses such as hemolysis. Furthermore, the baseline model is modified according to its hemolysis performance in ECMO condition, and two more geometries are designed and simulated. The results are presented and compared in terms of flow characteristic, pump performance curve and the hemolysis index of pumps.

## 2 PHYSICS AND GOVERNING EQUATIONS

In present paper, incompressible steady flow is assumed. Therefore, the governing equations can be present as:

$$\nabla \cdot \mathbf{u} = 0 \tag{1}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \left( \frac{P}{\rho} \right) + \nu \nabla \cdot \nabla \mathbf{u} \tag{2}$$

where $\mathbf{u}$ is the velocity vector, P is the pressure and $\nu$ is kinetic viscosity.

The blood pumps rotational speed is 1200~3600 rpm which is associated with Re≈ $1.2 \times 10^5$ (based on the impeller diameter). Hence, current study has to be conducted using turbulent assumptions. Therefore, the standard k-$\varepsilon$ turbulence model has been adopted. The standard k-$\varepsilon$ turbulence model can be present as:

$$\frac{\partial (\rho k)}{\partial t} + div(\rho k \mathbf{U}) = div\left[ \frac{\mu_t}{\sigma_k} grad(k) \right] + \mu_t E_{ij} \cdot E_{ij} - \rho \varepsilon \tag{3}$$

$$\frac{\partial (\rho \varepsilon)}{\partial t} + div(\rho \varepsilon \mathbf{U}) = div\left[ \frac{\mu_t}{\sigma_\varepsilon} grad(\varepsilon) \right] + C_{1\varepsilon} \frac{\varepsilon}{k} 2\mu_t E_{ij} \cdot E_{ij} - C_{2\varepsilon} \rho \frac{\varepsilon^2}{k} \tag{4}$$

Where $C_\mu = 0.09$, $\sigma_k = 1.00$, $\sigma_\varepsilon = 1.30$, $C_{1\varepsilon} = 1.44$ and $C_{2\varepsilon} = 1.92$ is constant.

The rotational motion of the impeller has been simulated by sliding grid method. In order to consider the non-Newtonian properties of blood, the Carreau model has been applied for blood viscosity simulation [9]. The adopted Carreau model reads as:

$$\mu = \mu_\infty + (\mu_0 - \mu_\infty)(1 + (\lambda\dot{\gamma})^2)^{(n-1)/2} \tag{5}$$

Where $\lambda$=3.313 s, n=0.3568, $\mu_0$=0.056 Pa-s and $\mu_\infty$=0.00345 Pa-s are constants used for the non -Newtonian viscosity of blood.

## 3 MIH CALCULATION

Hemolysis is related to exposure of red blood cells to high shear stresses, which destroys of red blood cells [10,11]. In order to evaluate hemolysis index, ASTM F1841-97 standard [12] suggests three methods for blood damage evaluations caused by high shear stresses of a medical device: normalized index of hemolysis (NIH), normalized milligram index of hemolysis (mgNIH), modified index of hemolysis (MIH). In current study, the MIH is adopted. The numerical calculations are based on work of Wurzinger et al. [13] and Giersiepen et al. [14] who propose a power law-based model for a damage index as:

$$D = C\tau^{\alpha} t^{\beta} \tag{6}$$

where $C = 3.62\times10^{-7}$, $\alpha = 2.416$, $\beta = 0.785$ are constants defined by analysis of the experimental data, $\tau$ is shear stress and t is the time of exposure.

Moreover, Bludszuweit [15] suggests a shear stress parameter using the scalar shear stress derived from six components of the stress tensor. The scalar shear stress can be presented as:

$$\tau_{vm} = \left[\frac{1}{2}\left[\left(\sigma_{xx}-\sigma_{yy}\right)^2 + \left(\sigma_{yy}-\sigma_{zz}\right)^2 + \left(\sigma_{zz}-\sigma_{xx}\right)^2 + 6\left(\sigma_{xy}^2+\sigma_{yz}^2+\sigma_{zx}^2\right)^2\right]\right]^{\frac{1}{2}} \tag{7}$$

In order to evaluate MIH, Garon and Farinas [7] suggest numerical model, which calculates blood damage through volume integration of a damage factor. The linear damage function can be present as:

$$\bar{D}_l = D^{1/0.785} = \left(3.62\times10^{-7}\right)^{1/0.785} \tau^{2.416/0.785} t \tag{8}$$

$$D(\tau,t) = \left(\bar{D}_l\right)^{0.785} \tag{9}$$

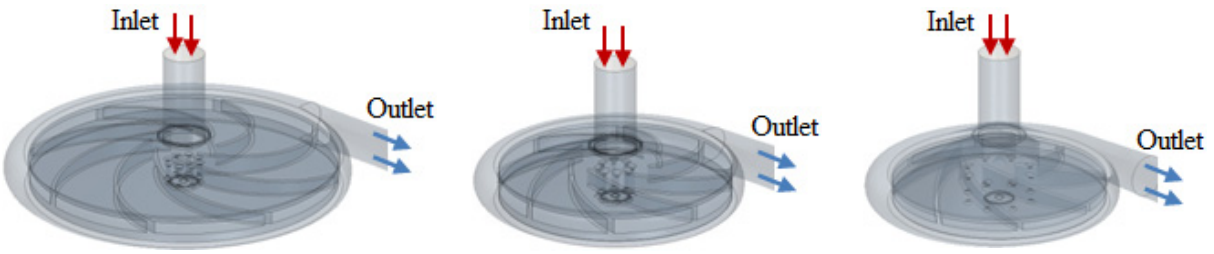
and finally the MIH can be present as:

$$MIH = D(\tau,t)\times10^6 \tag{10}$$

According to comparison presented by Chang et al. [16], the MIH values calculated based on this method in agreement with experiment.

## 4 GEOMETRY OF PUMPS

The geometry pumps is determined based on work done by Chang et al [16] and their investigation on the existing commercial blood pumps. Later, two more modified models are also designed and simulated. The 3-D geometries are shown in Figure 1. All pumps have blade-type impeller with a shroud. To design the blade, $\beta_2$ is chosen as certain value and $\beta_1$ is calculated, according to velocity triangle as shown in Figure 2.




(a) Prototype 1 (baseline) (b) Prototype 2 (c) Prototype 3

Figure 1: Isometric view of the degined models.

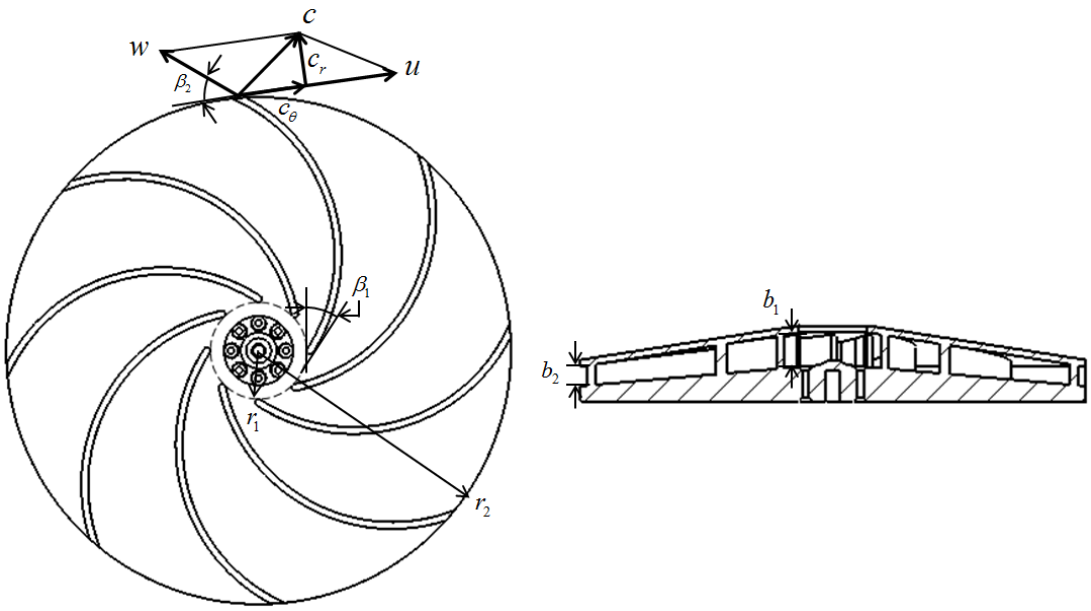



Figure 2: Convention for blade profile and velocity triangle.

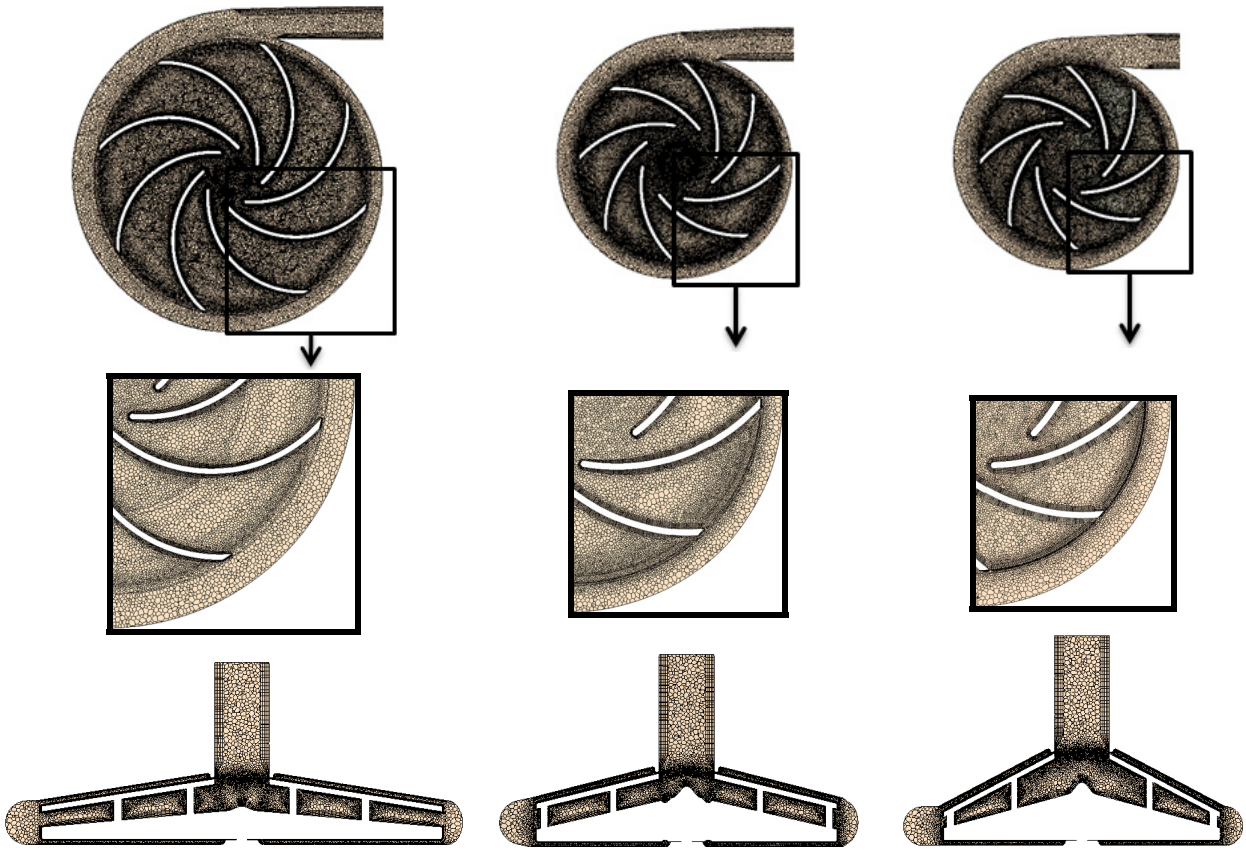
$$\beta_1 = \text{atan}\left(c_{m_1} / w_{\theta_1}\right) \tag{11}$$

where $c_{m_1} = Q_{impeller\ inlet} / A_{impeller\ inlet}$. Later, the local $\beta$ angle along to blade radius is calculated based number of blades and the impeller passage height at inlet and outlet.

The outlet pipe is connected to the casing by a volute designed based on Constant Momentum Volute (CMV) method [17]. The volume of the blood pumps, the radius of impeller and height of impeller inlet, outlet are listed in Table 1. The baseline model pump has the largest volume and longest radius of impeller. The third Prototype pump has the smallest volume and shortest impeller radius.

## 5 RESULTS AND DISCUSSIONS

The unsteady simulations are performed using STARCCM+ (v.9.06) at ECMO and VAD conditions. However, since the ECMO condition is more critical than the VAD, in the present research the flow properties of ECMO condition are presented. Convergence of each simulation is checked by pressure and velocity values at certain point in flow flied and residuals of mass and momentum conservation.



(a) Prototype 1 (baseline) (b) Prototype 2 (c) Prototype 3

Figure 3: Mesh arrangement of the designed models

| | Volume [mL] | Radius of Impeller [mm] | $\beta_1$ [degree] | $\beta_2$ [degree] |
|---|---|---|---|---|
| Prototype 1 | 95 | 37 | 33 | 30 |
| Prototype 2 | 75 | 27 | 16 | 40 |
| Prototype 3 | 66 | 25 | 14 | 43 |

Table 1: Dimensions of the designed models.

In order to investigate the flow pattern inside the pumps, Figure 4 shows the changes in the pressure head ( $\Delta P$ ) with respect to the flow rate ( $Q$ ). The performance curve is conducted by steady simulation to figure out the trends. In general, the pressure differences decrease with increasing flow rate. All cases have same trend, the largest difference is approximately 5000 [Pa] among the various flow rate. In centrifugal pump, difference of area between impeller inlet and outlet is a significant design parameter. Generally, head is controlled by difference of impeller inlet and outlet areas. The greater the difference between the areas of the impeller inlet and outlet the lower the pressure difference at the same rotational speed. For the same reason, Prototype 3 has the lowest pressure difference. So in present study Prototype 3 has the fastest rotational speed at ECMO condition to reach pressure required for ECMO condition, as shown Table 2.

Figure 5 shows velocity magnitude and streamlines for each Prototype at ECMO condition. The result of top section presents occurrence of recirculation at impeller exit area at right side of impeller and inside volute. This asymmetric recirculation is caused by non-uniform pres-

sure distribution in the volute. The side views show that effect of wash-out holes at the bottom and the wash-out cavity at top and bridge of top cavity. The washout hole diameter is 1 mm and top cavity bridge gap clearance is 0.2 mm. The top cavity flow has very fast velocity because of the narrow gap. This flow causes a noticeable swirl at impeller inlet area.

Figure 6 demonstrates pressure distribution at ECMO condition. The pressure reference point is located at pump inlet. The top view section presents non-uniform pressure distribution in the volute, and the area near the outlet pipe has the highest pressure. Especially, Prototype 3 has small pressure drop along the volute and its pressure distribution is more uniform. The side views show that effect of top cavity flow. Since the flow through top cavity has very fast velocity. The area near the impeller inlet has the lowest pressure.

Figure 7 shows the scalar shear stress distribution at ECMO condition, which is an important factor in analysis of hemolysis by Equations (8-10). The high scalar shear stresses area are located inside volute and outside of impeller walls. The side view plot shows the effect of top cavity flow, which causes the swirl flow and high scalar shear stress at the top cavity wall. The highest scalar shear stress is located in this narrow gap.

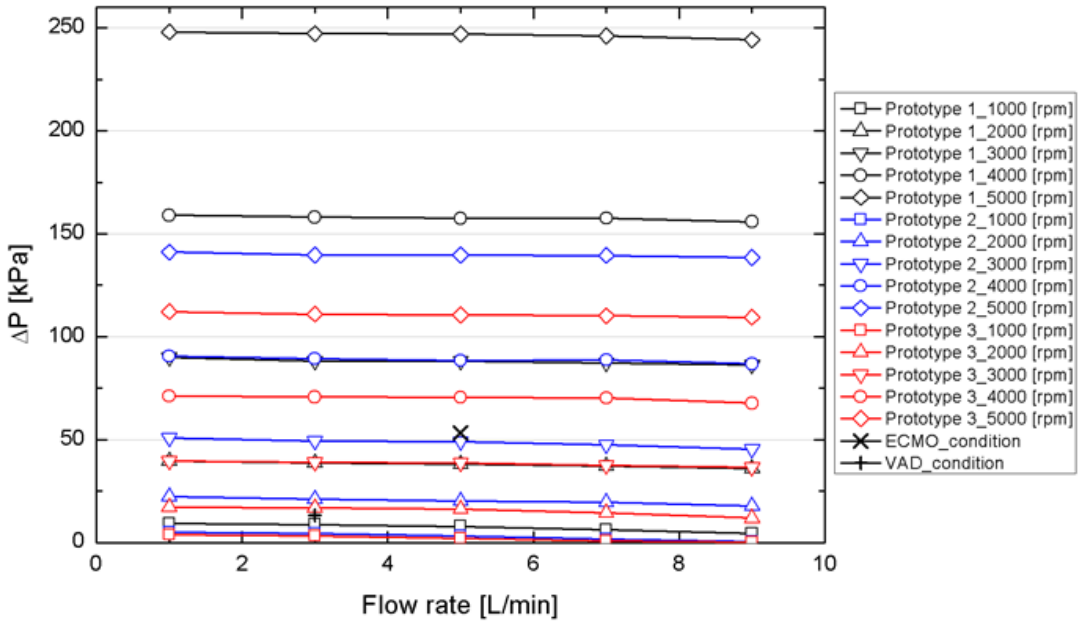
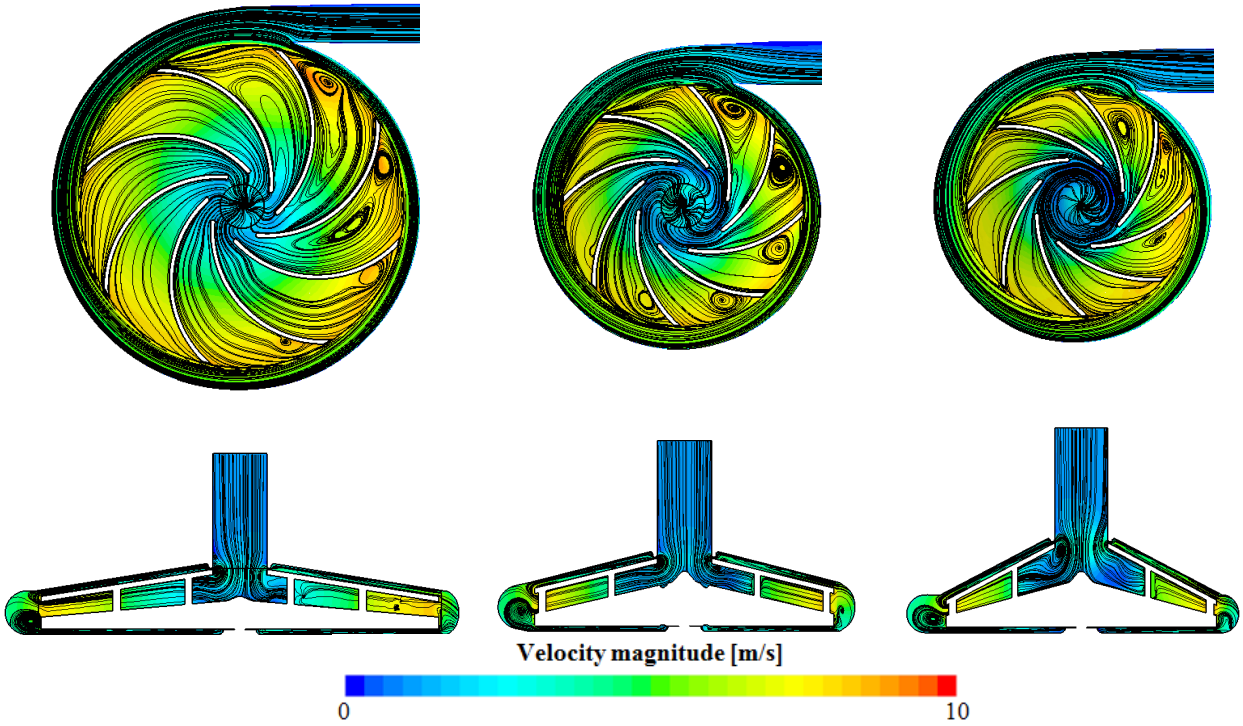



Figure 4: Pump characteristic curve

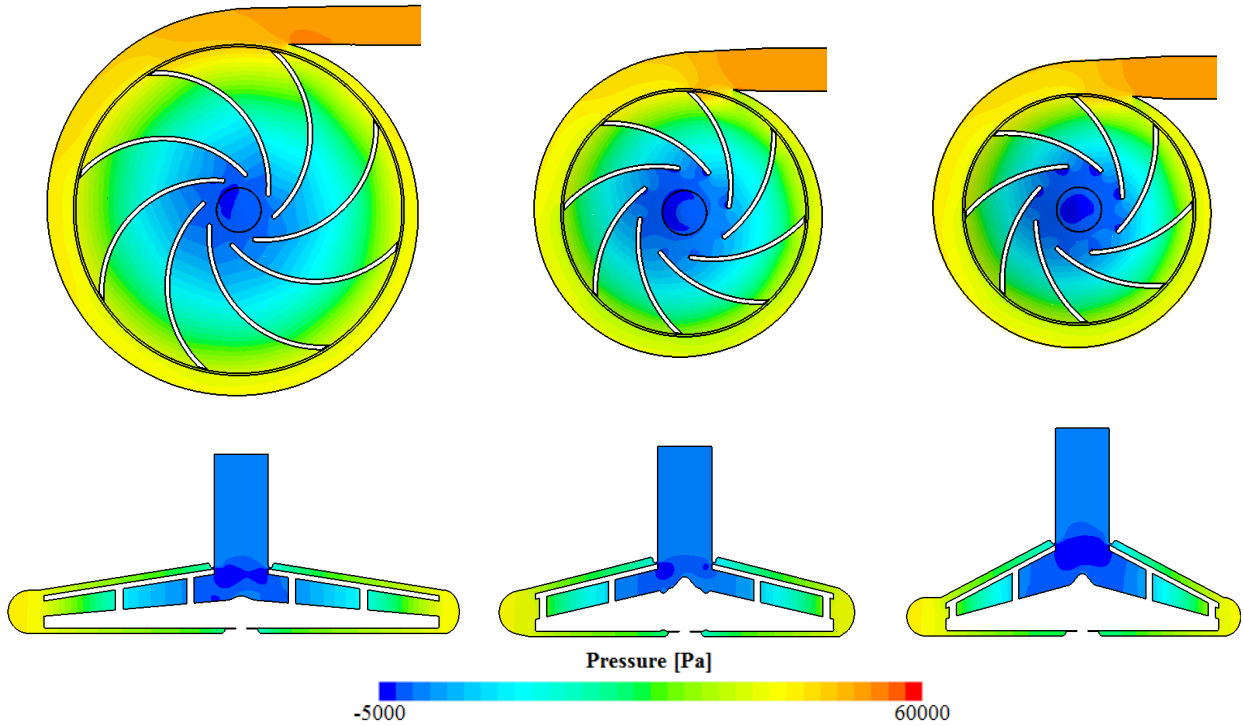
| | Pressure difference [kPa] | | Flow rate [L/min] | | Rotational speed [rpm] | |
|---|---|---|---|---|---|---|
| | ECMO | VAD | ECMO | VAD | ECMO | VAD |
| Prototype 1 (baseline) | 53 | 13 | 5 | 3 | 2368 | 1220 |
| Prototype 2 | 53 | 13 | 5 | 3 | 3120 | 1670 |
| Prototype 3 | 53 | 13 | 5 | 3 | 3619 | 1850 |

Table 2: ECMO and VAD operating condition



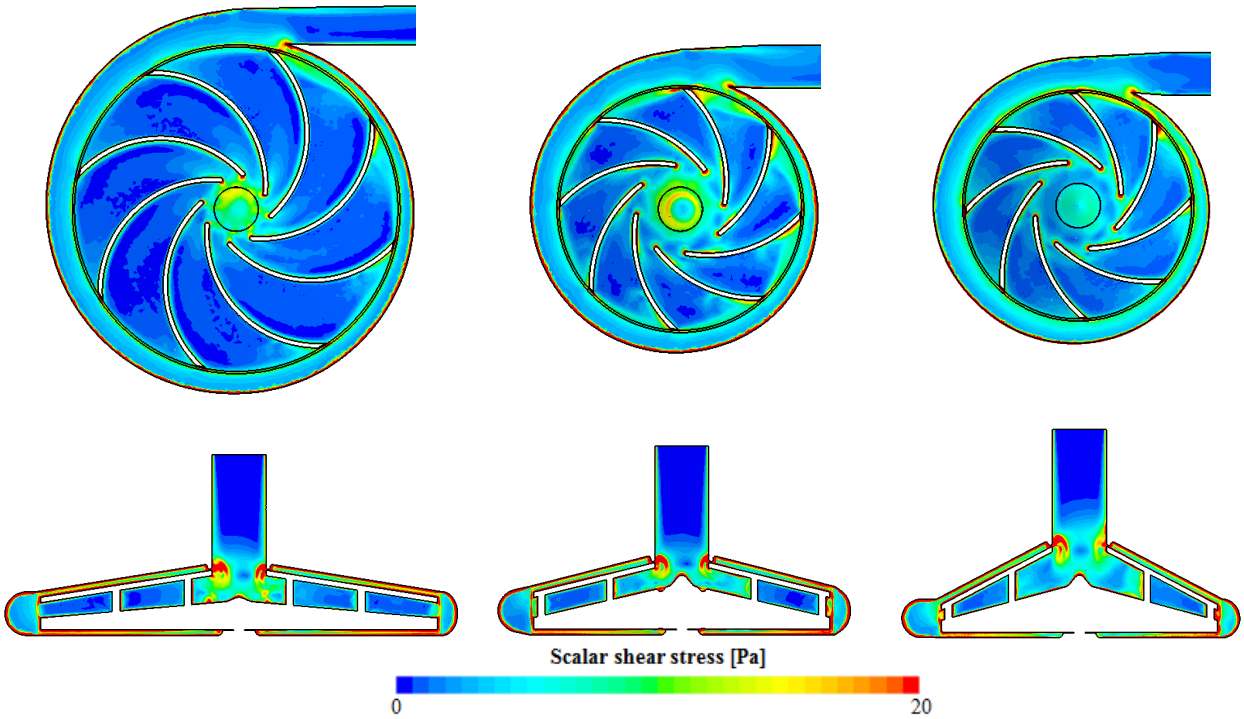

(a) Prototype 1 (baseline) (b) Prototype 2 (c) Prototype 3

Figure 5: Velocity magnitude and streamlines for designed pumps




(a) Prototype 1 (baseline) (b) Prototype 2 (c) Prototype 3

Figure 6: Pressure distribution for designed pumps




(a) Prototype 1 (baseline) (b) Prototype 2 (c) Prototype 3

Figure 7: Scalar shear stress distribution for designed pumps

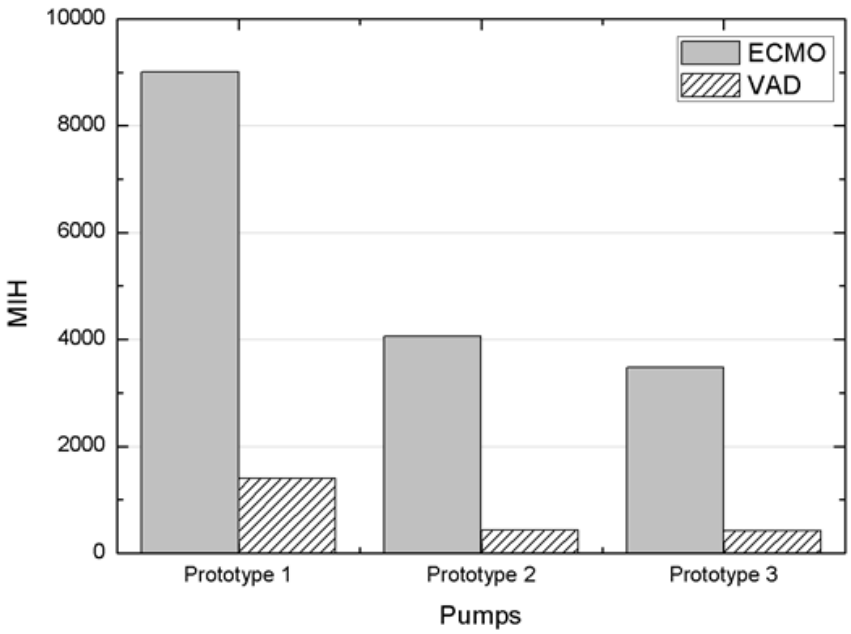



Figure 8: MIH values for designed pumps

Finally, Figure 8 displays the MIH value for three Prototypes at ECMO and VAD conditions. As it is seen, the VAD condition has smaller MIH value, because the rotational speed of the VAD condition is slower than the ECMO condition. In addition, a comparison between three models reveals that Prototype 3 has the fastest rotational speed while it has the lowest MIH values under ECMO and VAD conditions. Based on the result, the size of impeller and

priming volume are found to be more important variables than impeller rotational speed to occur hemolysis.

## 6 CONCLUDING REMARKS

In the present study, three different centrifugal blood pumps are designed and investigated through numerical simulation, using standard k-$\varepsilon$ turbulence model and sliding grid technique. In order to draw performance curve and figure out mechanical performance trend steady simulations are conducted with various flow rate (1-9 [L/min]) and rotational speed (1000-5000 [rpm]). As a result, all models have same trend. The pressure difference is reduced with increasing flow rate. Hence Prototype 3 has the lowest pressure difference because Prototype 3 has the largest difference between impeller inlet and outlet area. The results of flow streamlines show that all prototype pumps have asymmetric recirculation at the impeller exit area. The gap flow through top cavity has fast velocity. So it causes the swirl flow at impeller inlet area. Corresponding to the velocity magnitude and streamlines result, pressure is non-uniform in the volute. Because of gap flow through top cavity, the area near the impeller inlet has the lowest pressure. The results of scalar shear stress distribution, which is an important factor in analysis of hemolysis, show that high scalar shear stresses are located in volute wall, top cavity wall and impeller outside wall. At narrow gap of top cavity the highest scalar shear stress is observed. From the bio-mechanical aspect, in order to calculate MIH, an Eulerian approach is used. As result reveal, Prototype 3 has the smallest MIH value despite higher rotational speed than other models under ECMO and VAD conditions. According to results the size of impeller and total pump volume are important variables in generating the hemolysis.

## ACKNOWLEDGMENTS

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI14C0746).

## REFERENCES

[1] J.F. Fraser, K. Shekar, S. Diab, K. Dunster, S.R. Foley, C.I. McDonald, M. Passmore, G. Simonova, J.A. Roberts, D.G. Platts, D.V. Mullany, Y.L. Fung, ECMO-the clinician's view. *International Society of Blood Transfusion (ISBT) Science Series*, **7**, 82-88, 2012.

[2] S.V. Jain, A. Swarnkar, K.H. Motwani, R.N. Patel, Effects of impeller diameter and rotational speed on performance of pump running in turbine model. *Energy Conversion and Management,* **89**, 808-824, 2015.

[3] K. Takiura, T. Masuzawa, S. Endo, Y. Wakisaka, E. Tatsumi, Y. Taenaka, H. Takano, T. Yamane, M. Nishida, B. Asztalos, Y. Konishi, Y. Miyazoe, K. Ito, Development of Desogn Methods of a Centrifugal Blood Pump with In Vitro Test, Flow Visualization, and Computational Fluod Dynamics: Results in Hemolysis Tests. *Artificial Organs,* **22**, 393-398, 1998.

[4] Y. Miyazoe, T. Sawairi, K. Ito, Y. Konishi, T. Yamane, M. Nishida, B. Asztalos, T. Masuzawa, T. Tsukiya, S. Endo, Y. Taenaka, Computational Fluid Dynamics Analysis

to Establish the Design Process of a Centrifugal Blood Pump: Second Report. *Artificial Organs,* **23**, 762-768, 1999.

[5] Y. Nose, Design and Development Strategy for the Rotary Blood Pump. *Artificial Organs*, **22**, 438-446, 1998.

[6] M.E. Taksin, K.H. Fraser, T. Zhang, C. Wu, B.P. Griffith, Z.J. Wu, Evaluation of Eulerian and Lagrangian Models for Hemolysis Estimation. *ASAIO Journal,* **58**, 363-372, 2012.

[7] A. Garon, M.I. Farinas, Fast Three-dimensional Numerical Hemolysis Approximation. *Artificial Organs,* **28**, 1016-1025, 2004.

[8] M. Chang, N. Hur, M. Moshfeghi, S. Kang, W. Kim, S.H. Kang, A Numerical study on Mechanical Performance and Hemolysis for Different Type of Centrifugal Blood pumps. *International Mechanical Engineering Congress & Exposition (IMECE),* Houston, Texas, USA, November 13-19, 2015.

[9] P.J. Carreau, Rheological Equations from Molecular Network Theories. *Journal of Rheology*, **16**, 99-127, 1972.

[10] L. Goubergrits, K. Affeld, Numerical Estimation of Blood Damage in Artificial Organs. *Artificial Organs*, **28**, 499-507, 2004.

[11] S.W. Day, J.C. McDaniel, PIV Measurements of Flow in a Centrifugal Blood Pump: Steady Flow. *Journal of Biomechanical Engineering*, **127**, 244-253, 2005.

[12] ASTM Committee, *Standard Practice for Assessment of Hemolysis in Continuous Flow Blood Pumps, Vol. ⅩⅢ,* Annual Book of ASTM Standard, F1841-97, 1997.

[13] L.J. Wurzinger, R. Opitz, H. Eckstein, Mechanical Bloodtrauma. An overview. *Angeiologie*, **38**, 81-97, 1986.

[14] M. Giersiepen, L.J. Wurzinger, R. Opitz, H. Reul, Estimation of Shear Stress-related Blood Damage in Heart Valve Prosteses-In Vitro Comparison of 25 Aortic Valve. *Artificial Organs*, **13**, 300-306, 1990.

[15] C. Bludszuweit, A Theoretical Approach to the Prediction of Haemolysis in Centrifugal Blood Pumps. *Ph.D. thesis,* University of Strathclyde Glasgow, Scotland, 1994.

[16] M. Chang, N. Hur, M. Moshfeghi, S. Kang, W. Kim, S.H. Kang, Investigation on Mechanical and Bio-mechanical Performance of a Centrifugal Blood Pump. *Journal of Computational Fluids Engineering*, **20**, 88-95, 2015.

[17] Karassik, J. Igor, *Pump hand book, 2th Edition*. McGraw Hill, 1986

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFECT OF PERIVASCULAR TISSUE ON INFLATION-EXTENSION BEHAVIOR OF ABDOMINAL AORTA

**Tereza Vonavkova[1], Lukas Horny[1], Jan Vesely[1], Tomas Adamek[2,3], and Rudolf Zitny[1]**

[1] Faculty of Mechanical Engineering
Czech Technical University in Prague
Technicka 4, 166 07 Prague 6, Czech Republic
e-mail: {tereza.vonavkova, lukas.horny, jan.vesely1, 0rudolf.zitny}@fs.cvut.cz

[2] Regional Hospital Liberec
Husova 10, 460 63 Liberec
tomas.adamek@nemlib.cz

[3] Third Faculty of Medicine
Charles University in Prague
Ruska 87, 100 00 Prague
tomas.adamek@lf3.cuni.cz

**Keywords:** Abdominal Aorta, Bilayer Model, Inflation-Extension, Perivascular Tissue, Residual Stress, Thick-Walled Tube.

**Abstract.** *The research in the field of cardiovascular biomechanics is focused primarily on the heart and blood vessels, but the surrounding tissues, on the other hand, are often overlooked in the literature. This study shows that the human perivascular adipose tissue can significantly affect mechanical response of the human abdominal aorta. Analytical model of incompressible bilayer thick-walled closed cylindrical tube has been created and subsequently used to simulate the inflation-extension response of the aorta surrounded with adipose tissue. The inner layer, abdominal aorta, was assumed to exhibit residual stresses. The material of this layer was considered as anisotropic and was described by hyperelastic nonlinear constitutive model. The outer layer, fat tissue, was considered to be a hyperelastic isotropic material which does not exhibit residual stress. The outer radius of the aorta and inner radius of the external fatty tube were considered to be equal during the pressurization and axial stretch. The inflation-extension response of bilayer tube was compared with one-layer representing only the abdominal aorta. The simulations showed that the abdominal aorta is more compliant in the circumferential and axial direction in comparison with the fat tissue. The Cauchy stress across the wall thickness of bilayer model and the one-layer tube was determined and compared. The plot of the Cauchy stress versus deformed radius proved that the radial stress satisfies the boundary condition* $\sigma_{rr}({r_i}^A) = -P$, $\sigma_{rr}({r_o}^P) = 0$ *and* $\sigma_{rr}({r_o}^A) = \sigma_{rr}({r_i}^P)$. *An abrupt change in the axial stress by approx. 80 kPa occurs on the contact between the aorta and perivascular tissue. The results of this study suggest that the surrounding tissue should not be neglected in the modeling of blood vessels.*

## 1 INTRODUCTION

In the last few decades, there has been a significant growth of interest in the mechanical properties and constitutive modeling of biological soft tissue (predominantly vessels), but research on the perivascular tissue has been neglected. The adipose tissue surrounds all blood vessels and organs and influences their mechanical state. The perivascular tissue should be taken into account in solving of various types of boundary value problems describing biomechanics of abdominal aorta such as computational simulations of an aneurysm or *in vivo* constitutive modeling.

In the case of vessels, most of the works focus on modeling only one-layer tubes, but Holzapfel et al. (2000) modeled healthy carotid artery of a rabbit as bilayered structure consisting of two layers corresponding to the media and adventitia. They showed results of deformation behavior during inflation and axial torsion, bending, with and without including the residual strain and after that, they presented the distributions of the principal Cauchy stress components $\sigma_{\theta\theta}$, $\sigma_{zz}$ and $\sigma_{rr}$ along the deformed wall thickness (media and adventitia layers). Waffenschmidt et al. (2014) adopted the material, structural and geometrical parameters for a carotid artery of a rabbit from Holzapfel et al. (2000) and published extremal states of the energy of a double-layered thick-walled tube with these parameters. Bilayered model of the pressurized tube was also used by Sommer and Holzapfel (2012) who identified material parameters of invariant-based exponential elastic potential for both intact (bilayer model) and layer-dissected (one-layer model) human carotid arteries.

In the present study, the analytical model of bilayer thick-walled closed tube (the human abdominal aorta and perivascular adipose tissue as outer layer) has been created as a subject of the inflation-extension test, by using geometrical data of human abdominal aorta from Labrosse et al. (2012), material parameters of aorta from Horny et al. (2014) and material parameters of human perivascular tissue from Vonavkova et al. (2015). To the authors knowledge, this is the first study where the analytical model with individual layers (abdominal aorta and fat tissue) is studied.

## 2 MATERIAL AND METHODS

All the computations performed within this study have been conducted in Maple (Maplesoft, Waterloo, Canada). The abdominal aorta represents inner layer of analytical model and was modeled as an incompressible, hyperelastic, anisotropic residually stressed homogeneous thick-walled closed tube. The perivascular adipose tissue surrounding the aorta was considered as incompressible, hyperelastic and isotropic without residual stress.

In what follows, constitutive model and geometrical data characterizing abdominal aorta of 38 years old male donor are adopted from Horny et al. (2014). They used results, originally published by Labrosse et al. (2013), to determine material parameters of the human abdominal aorta defined in the constitutive model suggested by Gasser et al. (2006).

Vonavkova et al. (2015) published preliminary results of the uniaxial tensile tests conducted with human perivascular adipose tissue. They used hyperelastic constitutive model based on Fung-Demiray strain energy density function. This stress-strain relationship is here adopted to model mechanical behavior of the adipose tissue (male donor 29 years old).

### 2.1 Kinematics of inflation and extension of a bilayer tube

After removal from a body, the abdominal aorta is in the load-free configuration $\Omega_{\mathrm{LF}}$. However, this state is not a stress-free (reference) configuration which is denoted as $\Omega_{\mathrm{SF}}$. It is known that cutting the arterial ring in the vessel axis leads to opening of the artery due to the release of residual stress. Here, it is assumed that the open sector is the stress-free (reference)

configuration $\Omega_{\mathrm{SF}}$ (Sommer and Holzapfel, 2012), as depicted in Figure 1. Thus, the cylindrical coordinates of the tube $(\rho, \theta', z')$ in the reference configuration $\Omega_{\mathrm{SF}}$ are defined in (1).

$$\rho_i \le \rho \le \rho_o, \quad 0 \le \theta' \le (2\pi - 2\alpha), \quad 0 \le z' \le \zeta \tag{1}$$

Here $\rho_i$, $\rho_o$, $\alpha$ and $\zeta$ denote the inner and outer radius of the abdominal aorta in undeformed configuration, opening angle and the length of the stress-free tube, respectively. The opening angle for reference configuration is shown in the Figure 1.

The adipose tissue is modeled as not subjected to residual deformation, and, therefore it can be said, that its reference configuration is the load-free configuration $\Omega_{\mathrm{LF}}$ (Sommer and Holzapfel, 2012). See Figure 1.

The deformation $\chi$ (composition of the deformations $\chi_{\mathrm{L}}$ and $\chi_{\mathrm{RES}}$) which maps $\Omega_{\mathrm{SF}}$ into the deformed (loaded) configuration $\Omega_{\mathrm{L}}$ is depicted in the Figure 1. The deformation $\chi_{RES}$ maps from $\Omega_{\mathrm{SF}}$ to $\Omega_{\mathrm{LF}}$ and can be understood as a bending of curved beam corresponding to the stress-free arterial strip. The residual stress is induced by means of $\chi_{RES}$. The deformation $\chi_{\mathrm{L}}$ is associated with axial stretch and inflation and leads to the deformed configuration $\Omega_{\mathrm{L}}$ (Sommer and Holzapfel, 2012). In terms of the cylindrical coordinates $(R, \Theta, Z)$, the region of $\Omega_{\mathrm{LF}}$ is

$$R_i{}^{\mathrm{A}} \le R^{\mathrm{A}} \le R_o{}^{\mathrm{A}}, \quad 0 \le \Theta \le 2\pi, \quad 0 \le Z \le L^{\mathrm{A}}, \tag{2}$$

where $R_i{}^{\mathrm{A}}$, $R_o{}^{\mathrm{A}}$ and $L^{\mathrm{A}}$ are the inner and outer radius and the length of the abdominal aorta in the load-free configuration $\Omega_{\mathrm{LF}}$, respectively. $\Theta = \frac{2\pi}{(2\pi - 2\alpha)}\theta'$ and $Z = \delta z'$ holds in the maping $\chi_{\mathrm{RES}}$. $R_i{}^{\mathrm{P}}$, $R_o{}^{\mathrm{P}}$, $L^{\mathrm{P}}(L^{\mathrm{P}} = L^{\mathrm{A}})$, $\delta$ are the reference inner and outer radius, length and axial stretch of the perivascular tissue in the unloaded configuration.

The deformation $\chi_{\mathrm{L}}$ (Sommer and Holzapfel, 2012) takes the load-free configuration $\Omega_{\mathrm{LF}}$ into the deformed configuration $\Omega_{\mathrm{L}}$ (Figure 1). In terms of cylindrical coordinates $(r, \theta, z)$, the region of the current configuration is

$$r_i{}^{\mathrm{A}} \le r^{\mathrm{A}} \le r_o{}^{\mathrm{A}}, \quad 0 \le \theta \le 2\pi, \quad 0 \le z \le l^{\mathrm{A}}. \tag{3}$$

Here $r_i{}^{\mathrm{A}}$ , $r_o{}^{\mathrm{A}}$ and $l^{\mathrm{A}}$ denote the inner and outer radius and the length of the deformed abdominal aorta, respectively. $r_i{}^{\mathrm{P}}$, $r_o{}^{\mathrm{P}}$, $l^{\mathrm{P}}(l^{\mathrm{P}} = l^{\mathrm{A}})$ are the deformed inner and outer radius, length and axial stretch of the perivascular tissue in the loaded configuration.

The kinematics of inflation and extension is described by the deformation gradient $\mathbf{F}_{\mathrm{L}}$ in (4). The deformation gradient $\mathbf{F}_{\mathrm{RES}}$ which considered closing of opened up circular sector as shown in the equation (4). The resulting kinematics is given as (5).

$$\mathbf{F}_{\mathrm{RES}} = \begin{pmatrix} \frac{\partial R(\rho)}{\partial \rho} & 0 & 0 \\ 0 & \frac{\pi}{\pi - \alpha}\frac{R(\rho)}{\rho} & 0 \\ 0 & 0 & \delta \end{pmatrix}, \quad \mathbf{F}_{\mathrm{L}} = \begin{pmatrix} \frac{\partial r(R)}{\partial R} & 0 & 0 \\ 0 & \frac{r(\rho)}{R} & 0 \\ 0 & 0 & \lambda_Z{}^{\mathrm{A}} \end{pmatrix} \tag{4}$$

$$\mathbf{F} = \mathbf{F}_{\mathrm{L}}\mathbf{F}_{\mathrm{RES}} = \begin{pmatrix} \lambda_r(r) & 0 & 0 \\ 0 & \lambda_\theta(r) & 0 \\ 0 & 0 & \lambda_z \end{pmatrix} = \begin{pmatrix} \frac{\partial r(\rho)}{\partial \rho} & 0 & 0 \\ 0 & \frac{\pi}{\pi - \alpha}\frac{r(\rho)}{\rho} & 0 \\ 0 & 0 & \lambda_Z{}^{\mathrm{A}}\delta \end{pmatrix} \tag{5}$$

Here ${\lambda_Z}^{\mathrm{A}}$( ${\lambda_Z}^{\mathrm{A}} = {\lambda_Z}^{\mathrm{P}}$) is the axial stretch of aorta which is identical with the stretch of the adipose tissue. $\lambda_r$, $\lambda_\theta$ and $\lambda_z$ are the radial, circumferential and axial stretches, respectively.

During the deformation, the inner radius of the abdominal aorta and the inner radius of perivascular were expressed by means of the radius and the length via incompressibility condition (6) and (7), respectively.

$$\pi l^{\mathrm{A}}\left(r_o^{2\mathrm{A}} - r_i^{2\mathrm{A}}\right) = (\pi - \alpha)L^{\mathrm{A}}(\rho_o^2 - \rho_i^2), \quad {\lambda_Z}^{\mathrm{A}} = \frac{l^{\mathrm{A}}}{L^{\mathrm{A}}}. \tag{6}$$

$$\pi L^{P}\left(R_o^{2\mathrm{P}} - R_i^{2\mathrm{P}}\right) = \pi l^{\mathrm{P}}\left({r_o}^{2\mathrm{P}} - {r_i}^{2\mathrm{P}}\right), \quad {\lambda_Z}^{\mathrm{P}} = \frac{l^{\mathrm{P}}}{L^{\mathrm{P}}}. \tag{7}$$

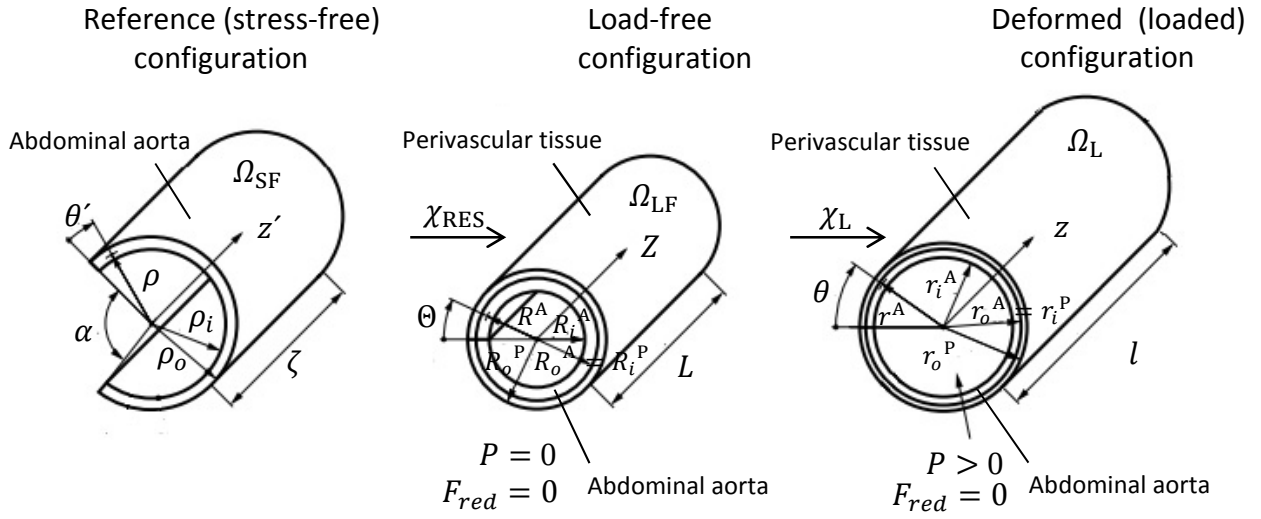



Figure 1: Kinematics of the abdominal aorta and the perivascular tissue.
The arterial ring with residual stress in the reference (stress-free) configuration $\Omega_{\mathrm{SF}}$.
The abdominal aorta with surrounding perivascular tissue in the load-free configuration $\Omega_{\mathrm{LF}}$ but without internal pressure.
The bilayer tube in the deformed (loaded) configuration $\Omega_{\mathrm{L}}$ after the application of internal pressure $P$.

## 2.2 Constitutive models

According to Ogden (1982), the constitutive equation for an incompressible hyperelastic material can be written in the form of (8).

$$\sigma = 2\mathbf{F}\frac{\partial W}{\partial \mathbf{C}}\mathbf{F}^{\mathrm{T}} - \mathbf{I}p. \tag{8}$$

Here σ denotes the Cauchy stress tensor. $\mathbf{C}$ is the right Cauchy-Green strain tensor, $\mathbf{C} = \mathbf{F}^{\mathrm{T}}\mathbf{F}$. $p$ plays the role of a Lagrangean multiplier, which represents the hydrostatic contribution to σ, not captured by $W$, due to the incompressibility constraint.

### 2.2.1. Abdominal aorta

The abdominal aorta wall was modeled as a homogenous, anisotropic, incompressible, hyperelastic material. The strain energy density function $W_{GOH}$ was used, proposed by

Gasser et al. (2006) for modeling general mechanical characteristics. The strain energy density function is expressed by equation (9).

$$W_{GOH} = \frac{\mu}{2}(I_1 - 3) + \sum_{j=4,6}\left(\frac{k_1}{2k_2}e^{k_2(K_j-1)^2} - 1\right), \tag{9}$$

$$K_j = \kappa I_1 + (1 - 3\kappa)I_j, \quad j = 4,6, \tag{10}$$

with

$$I_1 = \lambda_r^2 + \lambda_\theta^2 + \lambda_z^2, \quad I_4 = I_6 = \lambda_\theta^2 \cos^2\beta + \lambda_z^2 \sin^2\beta, \tag{11}$$

where $\mu = 15.9\,kPa$ is the infinitesimal shear modulus of the isotropic matrix. $k_1 = 78.488\,kPa$ is a stress-like material parameter and $k_2 = 4.911$ is a dimensionless parameter. The material parameters were adopted from Horny et al. (2014). $K_4$ and $K_6$ are generalised invariants related to imperfect fiber alignment. $\kappa = 0.189$ is a structural parameter which measures the degree of fiber dispersion $\left(0 \le \kappa = \frac{1}{3}\right)$. $I_1$ is the first invariant of the right Cauchy-Green strain tensor. $I_4$ and $I_6$ are additional strain invariants induced by the existence of preferred directions in a continuum. The parameter $\beta = 41.41°$ denotes the angle between the (mean) fiber direction and the circumferential direction in the individual layers, and, therefore, acts as a geometrical parameter.

### 2.2.2. Perivascular tissue

For the perivascular adipose tissue, the exponential function of strain energy density $W_D$ proposed by Demiray (1972) was used (12).

$$W_D = \frac{1}{2}\frac{\mu\left(e^{b\left(\lambda_r^2+\lambda_\theta^2+\lambda_z^2-3\right)}-1\right)}{b}, \tag{12}$$

with stress-like parameter $\mu = 18.397\,kPa$ and dimensionless parameter $b = 31.834$. These material parameters were taken from Vonavkova et al. (2015).

## 2.3 Thick-walled bilayer tube model

The equilibrium equations for the closed incompressible hyperelastic thick-walled tube in the radial and axial direction are expressed in (13) and (14) for the inner layer (the abdominal aorta) and in (15), (16) for the outer layer (the perivascular tissue). Their form is adopted from Holzapfel and Ogden (2010), Labrosse et al. (2013) and Horny et al. (2013). Detailed derivation can be found in Matsumoto and Hayashi (1996). The indices A, and P are used to distinguish between individual layers (aorta, perivascular tissue). The boundary conditions are considered in the form $\sigma_{rr}({r_i}^{\mathrm{A}}) = -P$, $\sigma_{rr}({r_o}^{\mathrm{P}}) = 0$ and $\sigma_{rr}({r_o}^{\mathrm{A}}) = \sigma_{rr}({r_i}^{\mathrm{P}})$. Further, the condition ${\lambda_Z}^{\mathrm{A}} = {\lambda_Z}^{\mathrm{P}}$ was considered.

$$P^{\mathrm{A}} = \int_{{r_i}^{\mathrm{A}}}^{{r_o}^{\mathrm{A}}} {\lambda_\theta}^{\mathrm{A}} \frac{\partial \widehat{W}^{\mathrm{A}}}{\partial {\lambda_\theta}^{\mathrm{A}}} \frac{dr^{\mathrm{A}}}{r^{\mathrm{A}}} \tag{13}$$

$${F_{red}}^{\mathrm{A}} = \pi \int_{{r_i}^{\mathrm{A}}}^{{r_o}^{\mathrm{A}}} \left(2{\lambda_z}^{\mathrm{A}} \frac{\partial \widehat{W}^{\mathrm{A}}}{\partial {\lambda_z}^{\mathrm{A}}} - {\lambda_\theta}^{\mathrm{A}} \frac{\partial \widehat{W}^{\mathrm{A}}}{\partial {\lambda_\theta}^{\mathrm{A}}}\right) r^{\mathrm{A}} dr^{\mathrm{A}} \tag{14}$$

$$P^{\mathrm{P}} = \int_{{r_i}^{\mathrm{P}}}^{{r_o}^{\mathrm{P}}} {\lambda_\theta}^{\mathrm{P}} \frac{\partial \widehat{W}^{\mathrm{P}}}{\partial {\lambda_\theta}^{\mathrm{P}}} \frac{dr^{\mathrm{P}}}{r^{\mathrm{P}}} \tag{15}$$

$${F_{red}}^{\mathrm{P}} = \pi \int_{{r_i}^{\mathrm{P}}}^{{r_o}^{\mathrm{P}}} \left(2{\lambda_z}^{\mathrm{P}} \frac{\partial \widehat{W}^{\mathrm{P}}}{\partial {\lambda_z}^{\mathrm{P}}} - {\lambda_\theta}^{\mathrm{P}} \frac{\partial \widehat{W}^{\mathrm{P}}}{\partial {\lambda_\theta}^{\mathrm{P}}}\right) r^{\mathrm{P}} dr^{\mathrm{P}} \tag{16}$$

Here $\hat{W}$ denotes the strain energy density function with eliminated explicit dependence on $\lambda_r$ by substituting $\lambda_r = \frac{1}{\lambda_\theta \lambda_z}$. $P$ indicates internal pressure and $F_{red}$ is the reduced axial (prestretching) force acting on the closed end of the tube additionally to the force generated by the pressure acting on the end (Horny et al., 2013; Horny et al., 2014; Matsumuto and Hayashi, 1996).

The distributions of the principal Cauchy stress components $\sigma_{\theta\theta}$, $\sigma_{zz}$ and $\sigma_{rr}$ across the deformed wall thickness (aorta and perivascular tissue layers) are given equations (17), (18), (19).

$$\sigma_{rr}(r) = -\int_r^{r_o} \lambda_\theta \frac{\partial \widehat{W}}{\partial \lambda_\theta} \frac{dr}{r} \tag{17}$$

$$\sigma_{\theta\theta} = \lambda_\theta \frac{\partial \widehat{W}}{\partial \lambda_\theta} + \sigma_{rr} \tag{18}$$

$$\sigma_{zz} = \lambda_z \frac{\partial \widehat{W}}{\partial \lambda_z} + \sigma_{rr} \tag{19}$$

### 2.4 Inflation-extension behavior of bilayer tube model

Horny et al. (2014) fitted experimental data from Labrosse et al. (2013) by $W_{GOH}$ hyperelastic model and estimated material parameters $(\mu,\ k_1, k_2, \beta, \kappa)$ and subsequently determined variable radii in an undeformed configuration $(\rho_i, \rho_o)$ and axial deformation which arose during closing of the aorta ring $(\delta)$.

The reference geometrical data $(R_i = 5.3\ mm, H = 1.22\ mm, \alpha = 117°)$ of human abdominal aorta (male age 38 years) from Labrosse et al. (2013), material parameters $(\mu = 15.90\ kPa,\ k_1 = 78.49\ kPa, k_2 = 4.991, \beta = 41.41, \kappa = 0.1875)$ and geometrical parameters $(\rho_i = 16.20\ mm, \rho_o = 17.42\ mm)$ from Horny et al. (2014) were chosen for purpose of this study.

The outer radius of the aorta and inner radius of the external fatty tube $({R_o}^{\mathrm{A}} = {R_i}^{\mathrm{P}},\ \ {r_o}^{\mathrm{A}} = {r_i}^{\mathrm{P}})$ were considered to be equal in the pressurization. The outer radius of adipose tissue was chosen $({R_o}^{\mathrm{P}} = 16.52\ mm)$. The value of the outer radius corresponds approximately to the amount of fat in the selected donor (male age 38 years).

The inner and outer radius of perivascular tissue $({r_i}^{\mathrm{P}},\ {r_o}^{\mathrm{P}})$ and axial stretch $\left({\lambda_Z}^{\mathrm{A}} = {\lambda_Z}^{\mathrm{B}}\right)$ have been determined by means of nonlinear least squares regression method. The objective function $Q$ (20) was minimized in Maple (Maplesoft, Waterloo, Canada) subjected to the constraint $F_{red} = 0$.

$$Q = \sum_{i=0}^{n=16} \left( \left( {P^{\exp}}_i - {P^{\mathrm{mod}}}_i \right)^2 - \left( {F_{red}^{\exp}}_i - {F_{red}^{\mathrm{mod}}}_i \right)^2 \right) \quad \text{for } i = 0,1,2,\ldots,16\ kPa \tag{20}$$

where

$${P^{\exp}}_i = {P^{\mathrm{A}}}_i + {P^{\mathrm{P}}}_i, \quad {F_{red}^{\exp}}_i = {F_{red}^{\mathrm{A}}}_i + F_{red_i}^{\mathrm{P}}. \tag{21}$$

## 3 RESULTS

The results of analytical model of bilayer thick-walled closed tube consisting of both the human abdominal aorta (index A) and surrounding perivascular tissue (index P) are summarized in the Figure 2. The bilayer model is compared with model containing only abdominal aorta. The plots of the pressure versus circumferential stretch $(\lambda_\theta)$ or axial stretch

($\lambda_z$) are depicted in the Figure 2 A, B. The radial, circumferential and axial Cauchy stress ($\sigma_{rr}, \sigma_{\theta\theta}, \sigma_{zz}$) versus deformed radius ($r$) at the pressure 16 $kPa$ are showed in the Figure 2C.

The axial stretch of the abdominal aorta and adipose tissue was equal during inflation-extension, but the analytical model shows that the tangential and axial stress differs significantly depending on the thickness of the outer tube. The radial stress satisfies the boundary condition $\sigma_{rr}(r_i^{A}) = -P$, $\sigma_{rr}(r_o^{P}) = 0$ and $\sigma_{rr}(r_o^{A}) = \sigma_{rr}(r_i^{P})$. On the contact with the aorta and perivascular tissue is an abrupt change in the axial stress by 80 $kPa$.

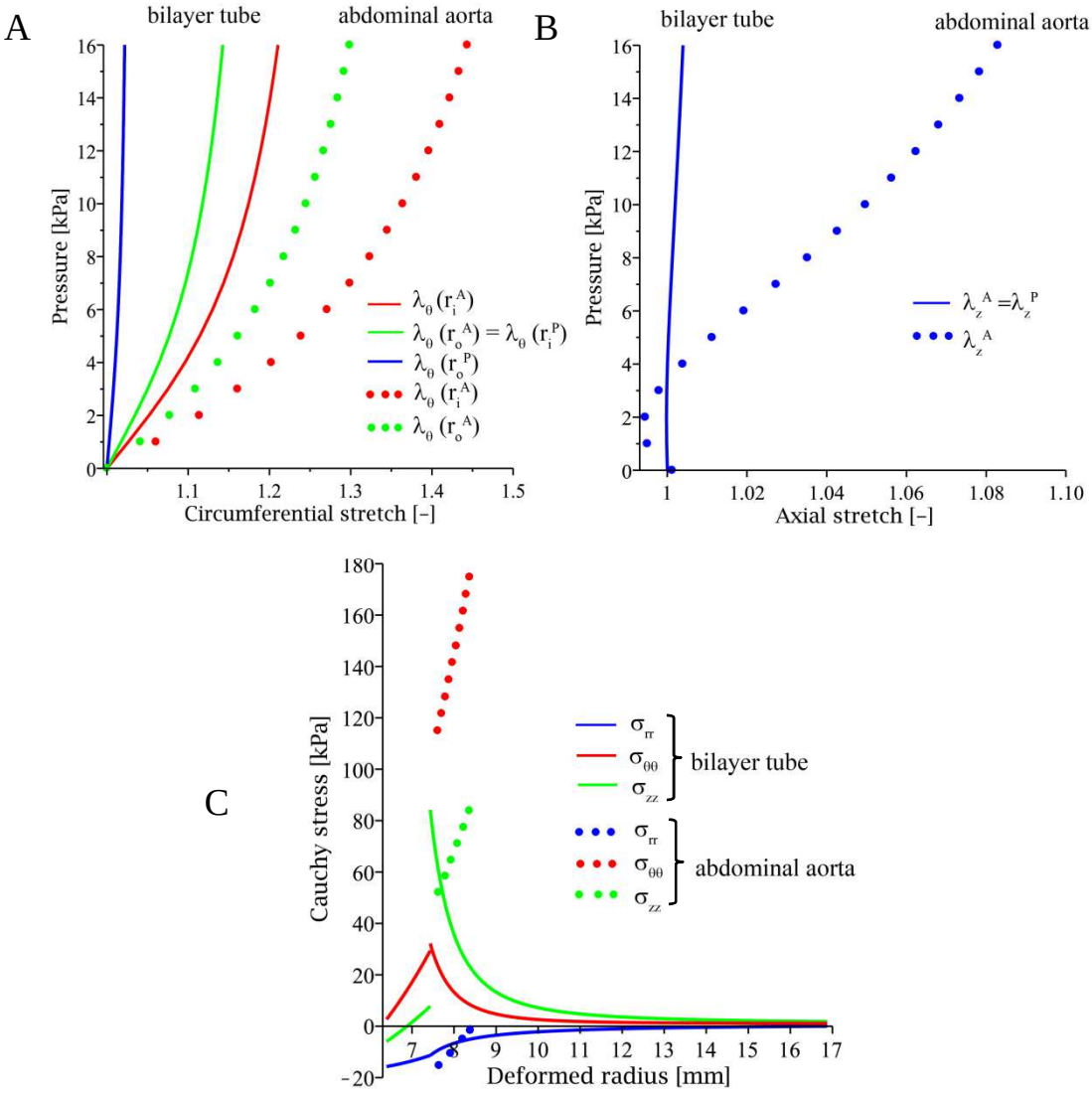



Figure 2 A The plot of the pressure versus circumferential stretch ($\lambda_\theta$) and comparison of bilayer tube (lines) with abdominal aorta (points).
B The plot of the pressure versus axial stretch ($\lambda_z$) and comparison of bilayer tube (lines) with abdominal aorta (points).
C The radial, circumferential and axial Cauchy stress ($\sigma_{rr}, \sigma_{\theta\theta}, \sigma_{zz}$) versus deformed radius for bilayer tube (lines) and abdominal aorta (points) at the pressure 16 $kPa$.

## 4 DISCUSSION

The objective of this study was to compare the mechanical behavior of the bilayer thick-walled close tube, where the inner layer was a human abdominal aorta and the outer layer was perivascular tissue, with abdominal aorta tube. The abdominal aorta and adipose tissue were

modeled as a homogeneous, anisotropic, incompressible and hyperelastic material. The strain energy density functions $W_{GOH}$, $W_D$ were used for the description of mechanical response of inner and outer layer of the model.

The simulations showed that the abdominal aorta is much more compliant in the circumferential and axial direction in comparison with the fat tissue. The stress along the wall thickness of bilayer model and only aorta tube was determined and compared.

The results of this study show that the mechanical responses are significantly different, therefore it is uncertain, whether a monolayer model without the surrounding tissue may correspond to *in vivo* reality.

**Acknowledgement**

This study has been supported by the Czech Technical University in Prague OHK2-033/16.

## REFERENCES

Demiray, H. (1972). A note on the elasticity of soft biological tissues. *Journal of Biomechanics,* 5:309–311.

Gasser, T. C., Ogden, R. W., & Holzapfel, G. A. (2006). Hyperelastic modelling of arterial layers with distributed collagen fibre orientations. *Journal of the Royal Society Interface*, 3(6): 15-35.

Holzapfel, GA, Gasser, T.C., & Ogden, R.W. (2000). A new constitutive framework for arterial wall mechanics and a comparative study of material models. *Journal of Elasticity*, 61**:** 1–48.

Holzapfel, G. A., & Ogden, R. W. (2010). Constitutive modelling of arteries. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2118), 1551-1597.

Horny, L., Netusil, M., & Vonavkova, T. (2013). Axial prestretch and circumferential distensibility in biomechanics of abdominal aorta. *Biomechanics and Modeling in Mechanobiology*, 13: 783-799.

Horny, L., Netusil, M., & Daniel, M. (2014). Limiting extensibility constitutive model with distributed fibre orientations and ageing of abdominal aorta. *Journal of the Mechanical Behavior of Biomedical Materials*, 38:39-51.

Labrosse, M.R., Gerson, E.R., Veinot, J.P, & Beller, C.J. (2013). Mechanical characterization of human aortas from pressurization testing and a paradigm shift for circumferential stress. *Journal of the Mechanical Behavior of Biomedical Materials*, 17:44-55.

Matsumoto, T., & Hayashi, K. (1996). Stress and strain distribution in hypertensive and normotensive rat aorta considering residual strain. *Journal of Biomechanical Engineering*, 118(1), 62-71.

Ogden, R.W. (1982). Elastic deformations of rubberlike solids. In Mechanics of Solids (The Rodney Hill 60th Anniversary Volume), pages 499–538. Wheaton and Co.

Ogden, R.W. (2003). Nonlinear Elasticity, Anisotropy, Material Stability and Residual stresses in Soft Tissue. *Biomechanics of Soft Tissue in Cardiovascular Systems*. 441: 65-108.

Sommer, G., & Holzapfel, G.A. (2012). 3D constitutive modeling of the biaxial mechanical response of intact and layer-dissected human carotid arteries. *Journal of the Mechanical Behavior of Biomedical Materials*, 5:116-128.

Taber, L.A. (2004). Nonlinear Theory of Elasticity: Applications in Biomechanics. ISBN:9812387358. World Scientific.

Vonavkova, T., Horny, L., Adamek, T., Kulvajtova, M., & Zitny, R. (2015). *Constitutive modelling of human perivascular adipose.* Computational Plasticity XIII, COMPLAS XIII 463-470.

Waffenschmidt, T., & Menzel, A. (2014). Extremal states of energy of a double-layered thick-walled tube – Application to residually stressed arteries. *Journal of the Mechanical Behavior of Biomedical Materials*, 29:635-654.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFECT OF STERILIZATION ON MECHANICAL PROPERTIES OF BIOLOGICAL COMPOSITE

**Jan Vesely[1], Hynek Chlup[1], Rudolf Zitny[1], and Tomas Grus[2]**

[1] Czech Technical University in Prague

Faculty of Mechanical Engineering

Technicka 4, Prague, Czech Republic

jan.vesely1@fs.cvut.cz, hynek.chlup@fs.cvut.cz, rudolf.zitny@fs.cvut.cz

[2] General University Hospital in Prague

U nemocnice 2, Prague, Czech Republic

tgrus@seznam.cz

**Keywords:** Biological composite, Sterilization, Ethylene oxide, Gamma irradiation, Collagen.

**Abstract.** *In this study, composite tubes were manufactured from biological collagenous matrix and reinforcing polyester mesh. The effect of sterilization on mechanical properties of this structure was evaluated using inflation-extension tests. Samples were exposed to two types of sterilization (ethylene oxide and irradiation). The control (non-sterilized) samples were also tested. It was found that the process of sterilization (especially irradiation) dramatically affects the final mechanical properties of the material. These findings should be taken into account when such collagenous material is assumed to be used in tissue engineering.*

## 1 INTRODUCTION

Composite materials are heterogeneous mixtures of two or more homogeneous components, which have been bonded together. In composites, properties or set of properties can be attained which could not have been obtained separately. Many in nature occurring materials can be regarded as composite e.g. bones, blood vessels, woods and others. Man-made composites are used since thousands of years, e.g. straw and natural fibers in bricks, laminated woods, etc. [1].

Over the past few decades, tissue engineering has been focused on development of biological substitutes to restore, maintain, or improve tissue functions. Collagen is the most abundant biological material used for tissue engineering. It is the basic constituent of skin, bones, ligaments and connective tissues. Collagen-based biomaterials have been studied extensively for a variety of biomedical applications, including dialysis membranes, wound dressings and artificial skin. Although native collagen possesses high tensile strength, the chemical treatment necessary for isolation makes the reconstituted collagen very poor in mechanical properties [2, 3]. A possible means to circumvent the problem is to reinforce natural polymer matrix by synthetic fibers or structures. Moreover, the properties of such composite could be modulated through composition of constituents in the material.

Routinely used sterilization process for medical products, e.g. high pressure steam (autoclaving) and dry heat cannot be considered for heat- and watersensitive biomaterials like collagen structures [4]. Currently the most widely utilized methods for collagenous materialsterilization are ethylene oxide (EO) gas infiltration and gamma irradiation. Earlier investigations focused on physical or chemical alterations after these two methods [5]. EO was claimed to alter the mechanical and physical properties of collagen slightly, but with a high risk of toxic residues [5]. Gamma irradiation, once introduced as the simplest and most effective way of sterilization without toxic substances, breaks chemical bonds, affects tensile strength and modulus, thus affecting the exposed material fundamentally [6,7].

In our study, composite tubes were manufactured from biological collagenous matrix and reinforcing polyester mesh. The effect of sterilization on mechanical properties of this structure was evaluated using inflation-extension tests.

## 2 MATERIAL AND METHODS

Tubular samples of hybrid composite were manufactured using extrusion while the polyester mesh was integrated into biological collagenous matrix (Fig 1). Two samples were sterilized using ethylene oxide, two samples by gamma irradiation and two control specimens were left unsterilized.

Samples were placed in physiological solution for 24 hours before testing. Prior to the mechanical tests, two rings were cut out from the specimen at both ends, and the mean reference dimensions of the samples (external radius, thickness) were determined by means of image analysis of digital photographs, Table 1.

### 2.1 Inflation-extension test

Each specimen was mounted in the experimental setup (Fig. 1) and marked with liquid eye-liner. The samples were pressurized until destruction using a motorized syringe (Standa Ltd, Vilnius, Lithuania). The intraluminal pressure was monitored by pressure transducer (Cressto s.r.o, Czech Republic). The experiments were performed at room temperature (22°C). The deformed geometry was recorded by a CCD camera (Dantec Dynamics, Skovlunde, Denmark). In the data post processing, the radius of the sample and length be-

tween marks during pressurization was evaluated by the edge detection algorithm in Matlab (MathWorks, MA, USA).

The longitudinal ($\lambda_z$) and circumferential ($\lambda_\theta$) stretch ratios were computed according Eq. (1). Here $L$ and $l$ is the length between marks in reference and deformed configuration, respectively. $R$ and $r$ is the unloaded and loaded middle radius of the sample, respectively.

$$\lambda_z = \frac{l}{L} \qquad \lambda_\theta = \frac{r}{R} \tag{1}$$

The influence of the sterilization was evaluated through circumferential stress, computed according thin-walled tube approximation. The circumferential stress ($\sigma_\theta$) induced by an internal pressure is computed using the Young–Laplace equation (2). In (2) $P$ is intraluminal pressure, $r$ is deformed middle radius and $h$ is deformed thickness of the sample, respectively.

$$\sigma_\theta = \frac{Pr}{h} \tag{2}$$

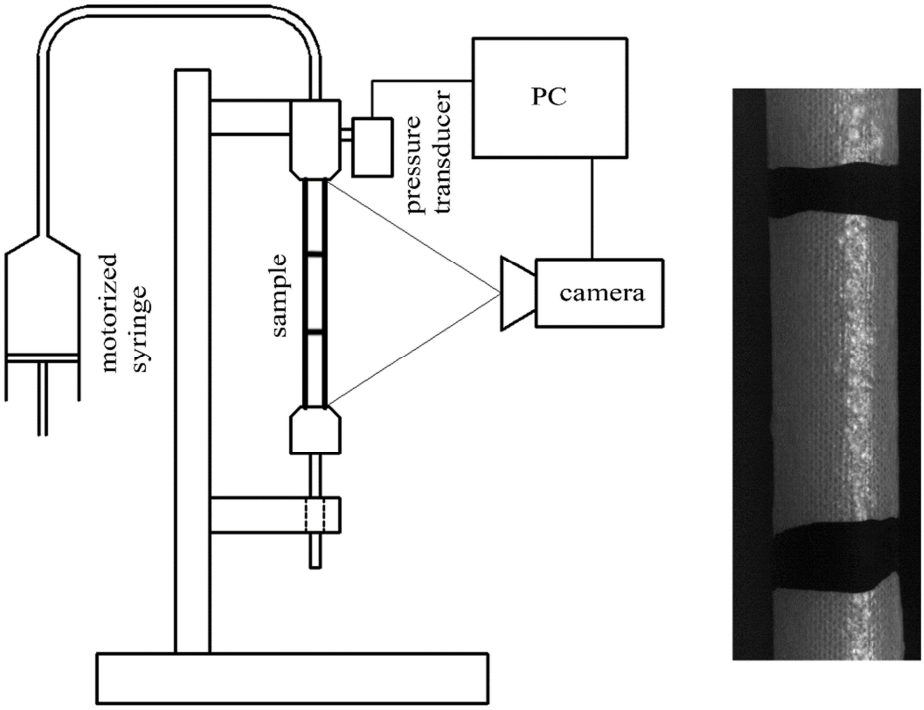



Figure 1: The experimental inflation-extension set-up (left panel) and the picture of the sample from CCD camera (right panel).

## 3 RESULTS

The reference dimensions of samples are listed in Table 1. Final circumferential stress – stretch curves are plotted in Fig. 2.

| Sample | $R_o$ [mm] (mean ± SD) | $H$ [mm] (mean |
|---|---|---|
| **Non-sterilized 1** | 3.80 ± 0.02 | 0.75 ± 0.17 |
| **Non-sterilized 2** | 4.26 ± 0.12 | 0.99 ± 0.28 |
| **Ethylene oxide 1** | 3.87 ± 0.09 | 0.62 ± 0.13 |
| **Ethylene oxide 2** | 3.65 ± 0.02 | 0.87 ± 0.25 |
| **Radiation 1** | 3.80 ± 0.06 | 0.68 ± 0.20 |
| **Radiation 2** | 3.45 ± 0.10 | 0.55 ± 0.07 |

Table 1: The reference dimensions of tubes. Here $R_o$ is the reference outer radius, $H$ is the reference overall thickness, respectively.

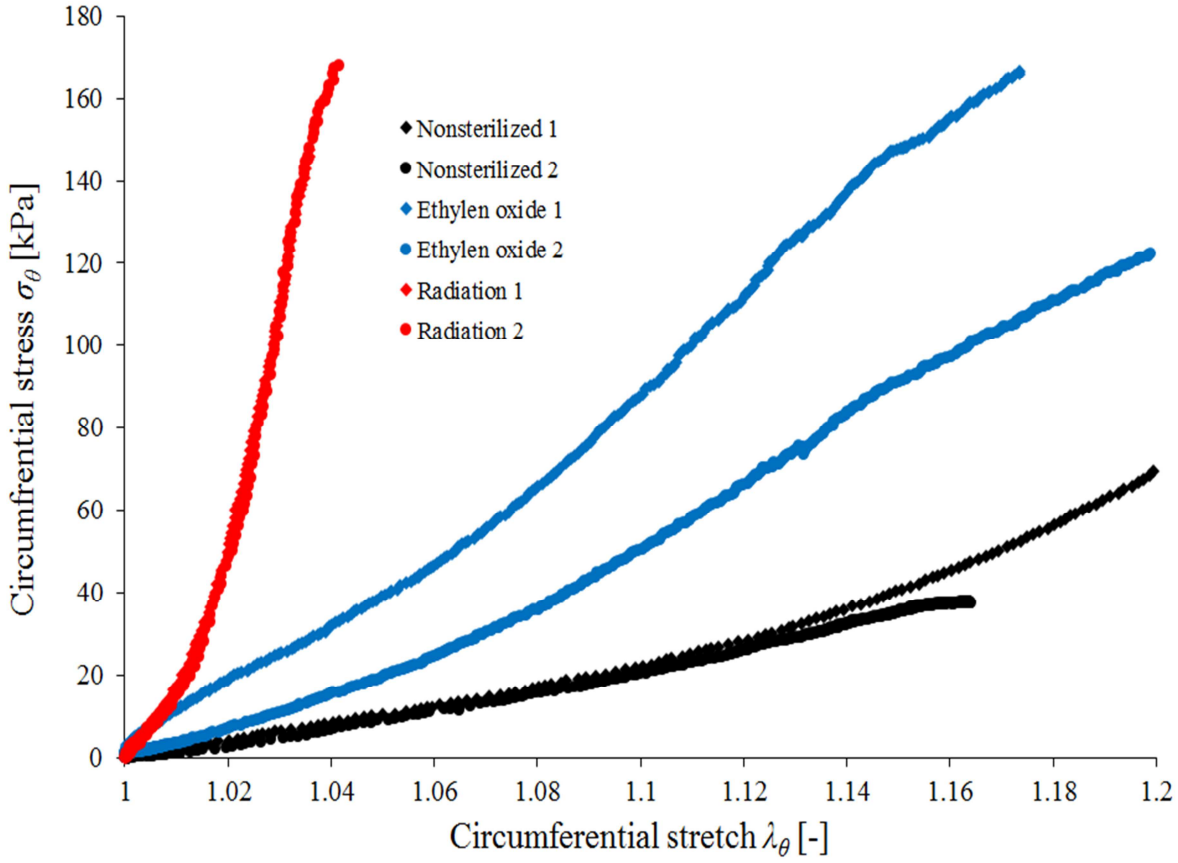



Figure 2: Stress – stretch curves for control/non-sterilized (black), ethylene oxide (blue) and irradiation (red) sterilized samples, respectively.

## 4 DISCUSSION AND CONCLUSIONS

In our study, biological composite samples were manufactured from collagenous matrix and reinforcing polyester mesh. The effect of sterilization (ethylene oxide and irradiation) on mechanical properties of this structure was examined using inflation-extension tests. This influence was evaluated through circumferential stress computed according thin-walled tube approximation. This is one of the limitations of the study, where the ratio of radius and thickness of the samples suggests using thick-walled tube model. However, authors wanted to compare the response of samples and the thin-walled model is simple and sufficient for this purpose.

Fig. 2 shows that the final mechanical properties are dramatically affected by process of the sterilization. Control samples showed the most compliant behavior and low breaking strength. Specimens sterilized using ethylene oxide are stiffer but the deformation at the destruction point is similar as for control samples, while the irradiation changed behavior of the material completely. Here the breaking strength is close to the samples sterilized by ethylene oxide, but the stiffness is very high. These observations should be interpreted through cross-linking process. During the sterilization, additional bonds are created between individual polymer (collagen) chains [8]. The more additional bonds are formed, the stiffer response is observed in the stress-stretch plot.

It was concluded, that the sterilization processes significantly change mechanical response of collagenous material. These findings should be taken into account when such biological material is assumed to be used in tissue engineering.

## ACKNOWLEDGEMENT

This study has been supported by Czech Ministry of Health through grant 15-27941A.

## REFERENCES

[1] T. Prabhuram, V. Somurajan, S. Prabhakaran., Hybrid composite materials, in proc.: *Proceedings of the International Con-ference on Frontiers in Automobile and Mechanical Engineering (FAME 2010)*, Institute of Electrical and Electronics Engineers, Chennai, India, November 25-27, 2010.

[2] R.Y. Kannana, H.J. Salacinskia, K. Salesa, P. Butlerb, A.M. Seifalian, The roles of tissue engineering and vascularization in the development of micro-vascular networks: A review. *Biomaterials*, **26**, 1857-1233, 2005.

[3] P. Giusti, L. Lazzeri, S. De Petris, M. Palla, M.G. Cascone, Collagen-based new bioartificial polymeric materials, *Biomaterials*, **15**, 1229-1233, 1994.

[4] E.M. Noah, J. Chen, X. Jiao, I. Heschel, N. Pallua, Impact of sterilization on the porous design and cell behavior in collagen sponges prepared for tissue engineering. *Biomaterials*, **23**, 2855–2861, 2002.

[5] L.H. Olde Damink, P.J. Dijkstra, M.J. Van Luyn, P.B. Van Wachem, P. Nieuwenhuis, J. Feijen, Influence of ethylene oxide gas treatment on the in vitro degradation behavior of dermal sheep collagen. *Journal of Biomedical Materials Research*, **29**, 149-155, 1995.

[6] D.T. Cheung, N. Perelman, D. Tong, M.E. Nimni, The effect of gamma-irradiation on collagen molecules, isolated alphachains, and crosslinked native fibers. *Journal of Biomedical Materials Research*, **24**, 581-589, 1990.

[7] B.C Liu, R. Harrell, R.H. Davis, M.H. Dresden, M. Spira, The effect of gamma irradiation on injectable human amnion collagen. *Journal of Biomedical Materials Research*, **23**, 833-844, 1989.

[8] V. Charulatha, A. Rajaram, Influence of different crosslinking treatments on the physical properties of collagen membranes. *Biomaterials*, **24**, 759–767, 2003.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FINITE-ELEMENT MODEL OF INTRAOCULAR PRESSURE MEASUREMENT BY MAKLAKOV APPLANATION TONOMETER

**Dmitry V. Franus**[1]

[1]Saint-Petersburg State University
28 University Ave., 198504, Stary Peterhof, Russia
e-mail: franus@mavis.ru

**Keywords:** Intraocular pressure, applanation tonometry, corneal shell.

**Abstract.** *The stress-strain state of the corneal shell of a human eye loaded with a flat base stamp of an arbitrary weight for different values of the intraocular pressure is analyzed. In the Maklakov method the intraocular pressure (IOP) is measured by deformations under loading. Similar approach, for which the stress-strain state is considered as the result of loading with a flat base stamp, is applied in [1, 2, 3].*

*This paper discusses the three dimensional finite-element model of the contact problem, when the multilayer corneal layer is loaded with a flat base stamp. This mathematical model is constructed in software package ANSYS. The cornea is modelled as an orthotropic spherical shell of variable thickness consisting of four layers: the outer layer (the epithelium), Bowmans membrane, the core layer (stroma of the cornea), and the inner layer (Descements membrane). Moreover, all layers have their specific elastic modules, that differ signicantly both in tangential direction and in transverse directions. The analysis is conducted of four flat base stamps of different weight and of different values of intraocular pressure (from 10 up to 30 mm Hg) for each stamp. The deformations of the corneal shell depends substantially on these factors. Also, the possibility of the contact detachment under loading is examined.*

*Finally, the estimation of different fundamental characteristics is made, including diameter of the contact zone, with the help of which the intraocular pressure is measured. The results of the mathematical modeling obtained in research are compared with the clinical data.*

## 1 INTRODUCTION

Determination of IOP is important in the diagnosis eye diseases and in monitoring during the postoperative period. IOP measurement is made by recording deformation of the cornea in response to mechanical impact under the load of a flat base stamp.

In the work [1] a two-dimensional model of measuring IOP using the Maklakov's tonometer is considered, in which the corneoscleral shell is modeled as two conjugated spherical segments. Meanwhile corneal shell is considered as a soft shell that has no resistance to flexural deformation and no "shrinkage" under the action of the tonometer. So deformation of the corneal shell is described by the relation $p_t = W/S$, where $p_t$ – tonometric IOP, $W$ – the weight of the tonometer, and $S$ – the contact area of the corneal shell and the tonometer (a flat base stamp). This equation is applicable only for infinitely thin and soft shells.

Based on the assumption of balancing of the weight of applied tonometer and IOP the following relation can be written as $F = W = PS$. Considering that the area of contact between the tonometer and the corneal shell has the shape of a circle, the area of the contact zone is calculated by the formula $S = \frac{\pi d^2}{4}$. Modifying the formula to express it in terms of the contact zone diameter $d$, the following is obtained:

$$d = 2\sqrt{\frac{W}{\pi P}}, \tag{1}$$

This paper presents a three-dimensional finite element modeling of the contact problem of loading the corneal shell with a flat base stamp in the mathematical software package ANSYS. It investigates the change in the stress-deformed state of the corneal shell while being loaded with stamps of different weights. In particular, existance of possible "tear" of the surface of the flat stamp from the cornea inside the contact zone is checked. Also the relationship of thickness between corneal shell and its layers is explored while loading with flat stamp and varying IOP.

## 2 FINITE-ELEMENT MODEL

Corneal shell is modeled with a spherical segment of outer radius of 7.8 mm with variable thickness, which is divided into four layers: epithelium, Bowman's membrane (BM), corneal stroma, and Descemet's membrane (DM). In this model only the stroma has variable thickness, which varies, for example, from 0,535 mm in the center up to 1,135 mm at the edge, the other layers are set with constant thicknesses. The endothelium, the fifth layer of corneal shell, is not modeled due to its insignificant thickness. All layers are modeled as homogeneous; fibration is not taken into account.

According to [4] the cornea can be considered as a transversely isotropic shell. The elastic constants satisfy to the following system of inequality [5]:

$$|\nu_i'| < \left(\frac{E_i'}{E_i}\right)^{1/2}, \; -1 < \nu_i < 1 - 2\left(\nu_i'\right)^2 \frac{E_i'}{E_i}, \; \left(E_i > 0, E_i' > 0\right), \; i = 1, ..., 4, \tag{2}$$

where $E_i$ and $E_i^{'}$ - are the moduluses of elastisity for the surface of isotropy and in the direction of normal to the surface respectively; $\nu_i$ and $\nu_i^{'}$ - the Poisson's ratios; $G_i^{'}$ is the shear modulus for any plane normal to the surface of isotropy; Shear modulus for the surface isotropy $G_i$ is given as:

$$G_i = \frac{E_i}{2\left(1 + \nu_i\right)} \tag{3}$$

Table 1: The thicknesses and elastic coefficients of the layers of the cornea

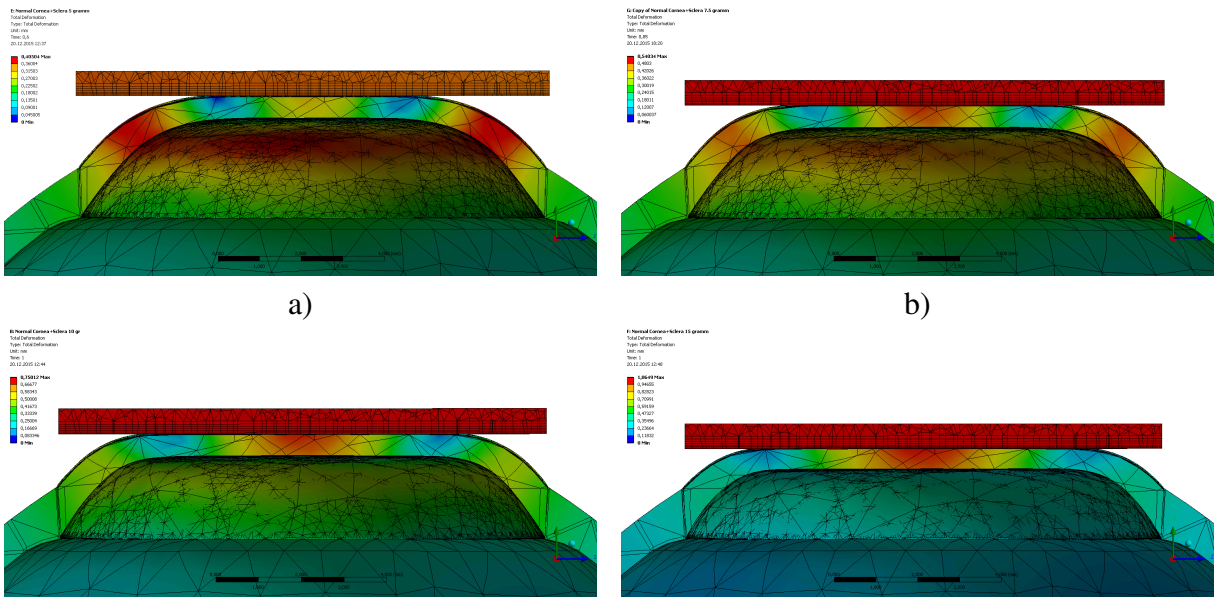
| | Units | Epithelium | BM | Stroma | DM | Sclera |
|---|---|---|---|---|---|---|
| $h_i$ | mm | 0,043 | 0,012 | 0,535-1,135 | 0,01 | 0,6-1,2 |
| $E_i$ | MPa | 0,06 | 0,6 | 0,3 | 0,9 | 5 |
| $E_i^{'}$ | MPa | 0,003 | 0,03 | 0,015 | 0,045 | 0,5 |
| $G_i$ | MPa | 0,0297 | 0,2970 | 0,1485 | 0,4455 | 2,4752 |
| $G_i^{'}$ | MPa | 0,0010 | 0,0100 | 0,0050 | 0,0150 | 0,1668 |
| $\nu_i$ | | 0,499 | 0,499 | 0,499 | 0,499 | 0,499 |
| $\nu_i^{'}$ | | 0,01 | 0,01 | 0,01 | 0,01 | 0,01 |

Table 1 specify the thickness of the layers of the corneal membrane, values of elastic moduli and Poisson's ratios in the tangential and thickness directions, which are used to solve the problem. It is assumed that the corneal tissue and scleral shell are close to incompressible, so $\nu_i$ is assumed to be 0,499. For the finite element model of the flat base stamp isotropic material with a coefficient of Young's modulus equal to $2 * 10^{11}$ Pa and Poisson's ratio $\nu = 0,3$ (steel) is used.

The IOP is set equal to 2000 Pa, which is normal, in the direction perpendicular to the inner surface. To reduce the influence of the boundary conditions on the investigated area of the corneal shell, a fixed support of the section of the sclera in the equatorial area of the eyeball is used as boundary conditions.

## 3 RESULTS

In this reseach a series of calculations is done, which correspond to the loading of the flat base stamp of different weights – 5, 7.5, 10, and 15 gram, simulating by so the process of measuring IOP by Maklakov's applanation tonometer using standard weights. Figure 1 shows



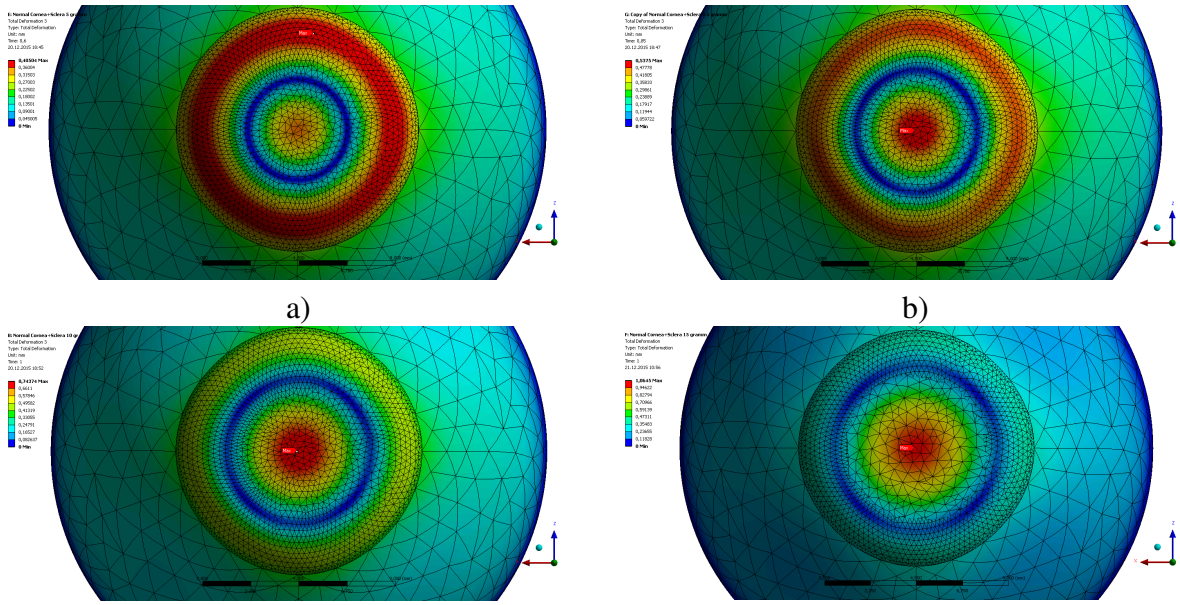
a) b)

c) d)

Figure 1: Deformation in the section $YZ$ of cornea loaded by stamp with weight a) 5 g, b) 7.5 g, c) 10 g, d) 15 g

deformations in the cross-section $YZ$, and Figure 2 shows deformation from above the model.

While loading the stamp with weight of 5 g, there is a distinct “ring”, which corresponds to the minimum deformations (close to zero by value) in the contact zone at the very edge. Moreover, these deformations become maximum (approx. 0.4 ) outside of the contact zone in the form of a ring. In case of lager weight of 7.5 g the diameter of the zone of contact is increased and another contact area of the maximum deformation appears in the center of the corneal shell. The subsequent increase in the weight of the stamp up to 10 g rises the deformations in the center and the diameter of the contact zone. This effect is enhanced by loading the heavier stamp with a mass of 15 g. Maximum values of deformations in the center of the contact zone are listed in table 3. In fact in figure 2 due to the ring of minimal deformation on the outer surface of the corneal shell, marked in blue, the diameter of the contact zone is increasing while raising the weight of the stamp.



a) b) c) d)

Figure 2: Deformation of cornea (top view) loaded by stamp with weight: a) 5 g, b) 7.5 g, c) 10 g, d) 15 g

Table 2: The diameters of the contact zone between the stamp and the epithelium of the cornea, mm

| | the weight of the stamp, grams | | | |
|---|---|---|---|---|
| IOP, mm Hg | 5 | 7,5 | 10 | 15 |
| 15 | 5,552 | 6,915 | 7,757 | 9,292 |
| 20 | 4,880 | 5,902 | 7,096 | 8,439 |
| 25 | 4,634 | 5,320 | 6,463 | 7,527 |
| 30 | 4,243 | 4,849 | 5,823 | 6,909 |

Approximating the data from table 2 we can obtain the following equation for change of the diameter of the contact zone $d$ between the corneal epithelium of the shell and loaded flat base stamp depending on $P$:

$$d_5 = -0,0002P^3 + 0,013P^2 - 0,4144P + 9,3399 \tag{4}$$

Table 3: Maximum values of deformations in the center of the contact zone

| | the weight of the stamp, grams | | | |
|---|---|---|---|---|
| IOP, mm Hg | 5 | 7,5 | 10 | 15 |
| 10 | 0,468 | 0,748 | 0,957 | - |
| 15 | 0,350 | 0,540 | 0,750 | 1,065 |
| 20 | 0,222 | 0,367 | 0,566 | 0,858 |
| 25 | 0,128 | 0,280 | 0,409 | 0,667 |
| 30 | 0,110 | 0,232 | 0,312 | 0,501 |

$$d_{7.5} = 0,0007P^3 - 0,0397P^2 + 0,6309P + 4,0951 \tag{5}$$

$$d_{10} = 0,0005P^3 - 0,0362P^2 + 0,6551P + 4,2628 \tag{6}$$

$$d_{15} = 0,0016P^3 - 0,1153P^2 + 2,5427P - 9,2137 \tag{7}$$

where equations (4) – (7) correspond to the loading of the flat base stamp with weights of 5, 7.5, 10, and 15 g respectively. In these equations $d$ – is measured in mm, and the IOP $P$ – in mm Hg. In order to calculate the value of the IOP based on the diameter of the imprint inversed relationship is obtained:

$$P_5 = 4,5884d^2 - 56,682d + 188,15 \tag{8}$$

$$P_{7.5} = 2,182d^2 - 32,942d + 138,45 \tag{9}$$

$$P_{10} = 0,0865d^2 - 9,085d + 79,635 \tag{10}$$

$$P_{15} = -0,0036d^2 - 8,1348d + 86,393 \tag{11}$$

here the equations (8) – (11) correspond to the loading of the flat base stamp with weights of 5, 7.5, 10, and 15 g respectively. The units are the same $d$ in mm, and $P$ in mm Hg. $P$ varies in the range from 15 to 30 mm Hg.

Based on received results of finite element modeling it is possible to obtain the diameter of the area of contact in dependence from the weight of loaded flat base stamp, while IOP is taken as constant. Approximating received data it is possible to obtain analytical relations between the diameter of the zone of contact $d$ and weight of the load applied to $W$:

$$d_{15} = -0,0316W^2 + 0,9174W + 1,7695 \tag{12}$$

$$d_{20} = -0,0146W^2 + 0,6538W + 1,9338 \tag{13}$$

$$d_{25} = -0,0106W^2 + 0,5106W + 2,2751 \tag{14}$$

$$d_{30} = -0,0063W^2 + 0,3998W + 2,3477 \tag{15}$$

where equation (12) – (15) correspond to loading at a constant value of IOP of 15, 20, 25, and 30 mm Hg respectively. The units of $d$ are mm and $W$ is in grams. The weight values of the flat stamp applied range from 5 to 15 g.

The corneal thickness changes in the center of the shell when loaded with a flat base stamp. As shown on figure 3 the overall thickness of the corneal shell is reduced by 18.4% under normal IOP 2000 Pa as heavier flat stamp is used. Due to its small thickness Bowman's membrane is included in the thickness of the layer of epithelium in current analysis. But in the process of loading with the 5 g stamp their combined thickness was reduced from 0.055 mm to 0,0234 mm, which is more than two times, and during loading with the 15 g stamp thickness is reduced to 0,0112 mm, this is a reduction in the thickness of the epithelium of five times. So, about 30% of the reduction in the corneal thickness falls on the epithelium.

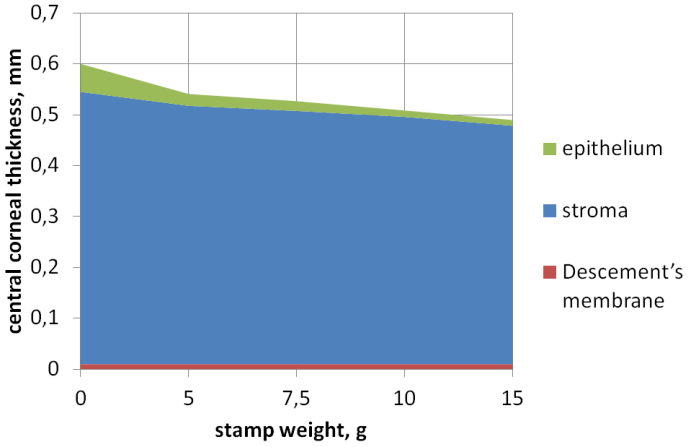



Figure 3: The change in the corneal thickness of the shell when loading by a stamp under normal IOP

## 4 CONCLUSION

There is no separation of the surfaces, in calculated range of IOP, in the contact zone of the surfaces of the stamp and of the corneal epithelium, when loading the stamp with a flat base on the corneal shell of the eye. Although the area of minimal deformations is close to zero (the ring marked in blue in figure 2), it is inside the contact zone.

As a result of loading simulation it is revealed that there is a five-times reduction in the thickness of the epithelium of the corneal shell, which is about a third of the total reduction in thickness of all the shell during loading.

Constructed finite element model of loading of the corneal shell with the flat base stamp allows to calculate the IOP depending on the weight of applied stamp and the thickness of the corneal shell of the eye in its center. Results obtained in the finite element simulation are in good agreement with clinical data.

The work was supported by the Russian Foundation for Basic Research (RFBR grant No. 15-01-06311-a).

## REFERENCES

[1] Bauer S.M., Lubimov G.A., Tovstik P.E. Mathematical Modelling of Maklakoffs Method for Measuring the Intraocular Pressure. *Fluid Dynamics*, **40(1)**, 20–33, 2005.

[2] Bauer S.M., Karamshina A.M., Kachanov A.B. Mechanical models of the measurements of intraocular pressure by Goldmann and Maklakov applanation tonometers after refractive surgery. *Russian Journal of Biomechanics*, **16(3)**, 19–24, 2012.

[3] Bauer S.M., Kachanov A.B., Semenov B.N., Sliesoraityte I. The effect of cornea thickness on measurements of intraocular pressure by applanation methods. *Ocular biomechanics*, Moscow, 69–73, 2007.

[4] Iomdina E.N. Mechanical properties of the human eye tissues *Modern problems of biomechanics*, **11**, 183–200, 2006.

[5] Radionova V.A., Titaev B.F., Chernykh K.F. *Applied theory of anisotropic plates and shells (in Russian)*. Saint-Petersburg, publisher Saint-Petersburg State University, 1996.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MOMENT-CURVATURE DIAGRAMS FOR EVALUATION OF SECOND ORDER EFFECTS IN RC ELEMENTS

**H. Barros**[1], **C. Ferreira**[1], **and T. Marques**[2]

[1] INESCC – Department of Civil Engineering, University of Coimbra

Rua Luís Reis Santos - Pólo II da Universidade, 3030-788 Coimbra, Portugal
{hbarros,Carla}@dec.uc.pt

[2] COBA, Consultores de Engenharia e Ambiente, S.A. Av. 5 de Outubro, 323, 1649-011 Lisboa

**Keywords:** Reinforced concrete, moment-curvature, stiffness, EC2, beam deflection.

**Abstract.** *The analytical (M-1/ρ) relationship is obtained with Mapple software considering the static equilibrium equations (of bending moment M and axial load N) and compatibility of deformations (between steel and concrete) at the section level [1]. The present work uses a closed form of the relation (1/ρ-M) with the Ramberg and Osgood equation [2], representing in a unique way the stages of: cracked concrete with steel in elastic range and cracked concrete with steel in the plastic range. To consider uncracked concrete in the global member behaviour the model of the European code [3] is used. The numerical results consist of moment curvature diagrams of reinforced concrete sections and load – deflection curves of reinforced concrete beams.*

## 1 INTRODUCTION

In structures composed of linear elements where one dimension is more significant than the others, such as beams and columns, a relation can be derived between the bending moment $M$ and the curvature ( $1/\rho$) at the section level, and the deformation produced by the tangential stresses disregarded. This moment ($M$) - curvature ( $1/\rho$) relation takes the form of a differential equation for the transverse displacement of the structural member in an homogeneous material.

In composite beams made of materials with different behaviour in tension and compression, such as concrete, and steel reinforcement, the ($M$) - curvature ( $1/\rho$) relation varies with the section along the structure. Several models to predict the global behaviour of these structures are used and described next.

Many ($M$ - $1/\rho$) equations are known such as the trilinear relationship, presented by Zheng et al [4], to predict flexural sectional response of beams with reinforced polymer rebars.

In the work of Torrico [5] the linearization of the moment vs curvature is used to obtain the pos-critical behaviour of high strength reinforced concrete (RC) columns, taking into account the geometrical and material nonlinearities including the confining effect of the transversal reinforcement. The values obtained are compared to experimental results of columns with different slenderness and reinforcement ratios.

The use of a moment-curvature law modelling the material softening due to cracking in reinforced concrete beams is made in Challamel et al [6]. Ponaya et al [7] uses this law to represent, in a simplified unidimensional approach, the local buckling phenomenon in steel thin-walled structures.

In Picandet et al [8] the moment-curvature relation is also used to model the geometrical softening due to the global instability in compressed columns. In this paper, a bilinear moment-curvature relation is considered with a first branch representing the elastic behaviour and a second branch the inelastic. Three alternatives in the inelastic branch are available: hardening, softening and perfect plastic. All of them are considered in the development of the analytical solution of the differential equation governing the lateral deflection of the column, clamped at the base and free at the top, where an axial load is applied. Load-deflection diagrams [8] are obtained considering simultaneously geometric and material non-linearities.

In Casandjian et al [9], the moment curvature relation at the section level is obtained using tension- deformation constitutive laws.

In the work of Caglar et al [10], the moment curvature diagrams are used to obtain the flexural stiffness of reinforced concrete columns with circular cross section, to be used in genetic programming with artificial intelligence. This technique is also applied by Cevik et al [11] and Chen et al [12].

According to European codes [3,13], the tension stiffening effect, due to the contribution of the concrete between cracks to the global behaviour, can be considered by two different ways: approaching the average strain (or curvature or stiffness) or using an effective area of concrete, which is equivalent to consider a mean modulus of elasticity. Both procedures use the concept of average constitutive laws of cracked concrete and reinforcing steel and are present in Vecchio et al [15] and Vecchio [16].

Kwak et al [17] use the average strain concept, considering the contribution of fully cracked and uncracked concrete and also a model to approach the bond-stresses between steel and surrounding concrete.

Kaklauskas et al [18] use the moment-curvature diagrams of RC beams obtained with ATENA software in a nonlinear finite element analysis, dividing the cross section into layers with perfect bond between them.

Márquez et al in [19] analyse the moment-curvature relations of concrete piles, with circular cross sections and different arrangement of reinforcement, such as the symmetric and the asymmetric, and compare them to Eurocode 2 [3] and CEB-fib Model Code 2010 [20] proposed expressions.

In the present paper the analytical ($M$- $1/\rho$) formulation takes into account concrete cracking under tension, linear elastic law in compression and steel with either elastic or plastic behaviours. Considering all these aspects the ($M$ - $1/\rho$) relation becomes nonlinear. The identified three stages are: a) uncracked concrete; b) cracked concrete and steel in the elastic domain and c) cracked concrete and steel in the plastic domain. The end of stage c) corresponds to the ultimate design, solved for multi-rectangular concrete sections in [14]. In this paper an approximation with Ramberg and Osgood equation of the inverse relation, that is the curvature ( $1/\rho$) as a function of the moment ($M$), is made. This procedure delivers a unique equation of the two concrete cracked stages within the section level, representing the behaviour of cracked sections. In order to obtain the global behaviour of the cracked member, both cracked and uncraked zones must be considered (usually named tension softening effect). The average curvature [3] is used in the present work.

The examples presented show the application of the developed model to obtain the moment-curvature of a reinforced concrete sections and the load deflection of reinforced concrete beams. The results are in good agreement with experimental and analytical results of other authors.

## 2 REINFORCED CONCRETE MEMBER BEHAVIOUR

The overall behaviour of the reinforced concrete member is ruled by the uncracked and cracked zones. The basic assumptions for singly reinforced concrete rectangular cross sections, defined by the width $b$, height $h$, concrete cover $a$ (distance between steel centroid and concrete border), steel tension area $A_s$ and steel compression area $A'_s$ (see Fig. 1), are described in the next sections.

### 2.1 Basic assumptions

The fundamental assumptions of the behaviour of flexural beams are [3]:

- the cross section remains plane after the deformation, see Fig.1;
- full adherence between concrete and steel;
- the concrete stress-strain constitutive law is linear elastic in compression and in tension, with elasticity modulus $E_c$; in tension the maximum stress in concrete is $f_{ctm}$; when the tensile concrete stress attains $f_{ctm}$ the section is considered cracked, and the global behaviour of the member depends on both cracked and uncracked sections;
- the steel stress-strain relation is linear elastic, with elasticity modulus $E_s$, up to the design yield stress $f_{yd}$. After this value a constant stress of $f_{yd}$ is considered (plastic domain).

Considering the concrete linear elastic relation in compression is an approximation allowed in EC2 [3] for service loads, for stress less than $0.4f_{cm}$ ($f_{cm}$ is the mean compressive strength). As a matter of fact the existence of cracked sections influences more the beam deflection in service, than the simplification due to the linear relation in the compressed concrete. Usually, the structures ductility is an important issue in concrete design, meaning that their collapse

occurs with steel rebars $A_s$ in the plastic domain. The possibility of the steel plastification is considered in the present model.

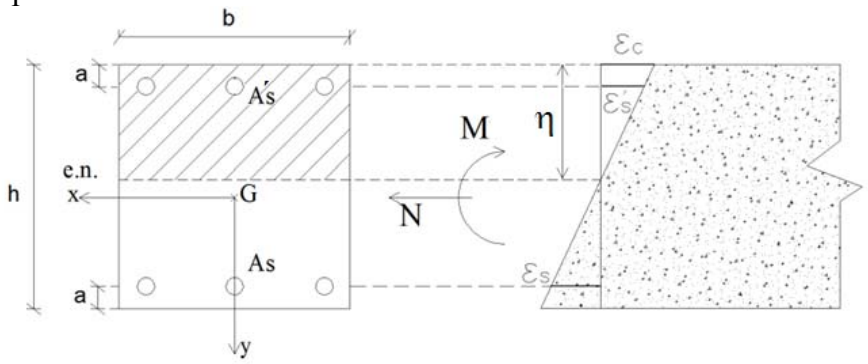



Figure 1: Geometrical definitions of the cross section and deformation.

## 2.2 Overall behaviour of the cracked member

The overall behaviour of the cracked member after the appearance of the first crack is divided into the zones: before the appearance of the first crack; formation of the cracks and the stabilized cracking, see Fig. 2a,b) [3]. In the figures state I is the ($M$ - $1/\rho$) evaluated in an uncracked section and State $II_0$ in a cracked section, both cases under pure bending. Under axial load and bending moment the curvatures are termed $1/\rho_1$ and $1/\rho_2$ in states I and II, respectively.

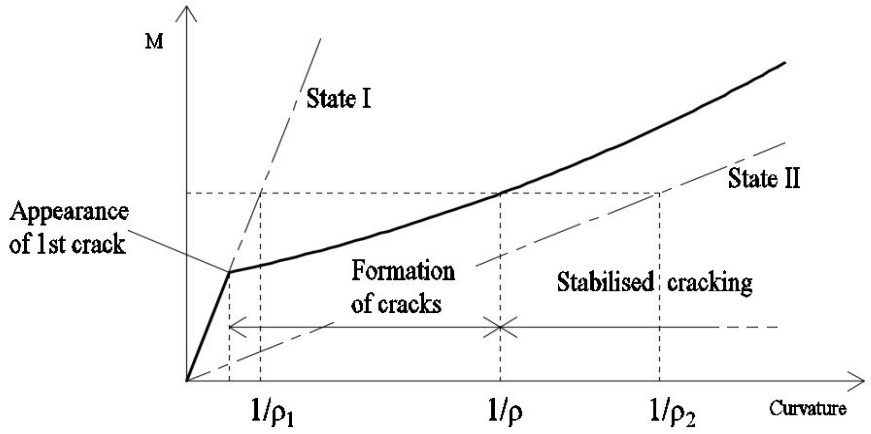



Figure 2a): Idealized moment curvature behaviour for bending moment [3].

In state I the curvature $\frac{1}{\rho_1}$ for pure bending or bending moment and axial load is considered the same [3]. The curvature for state II, bending moment and axial load, can be related to the curvature in pure bending, by $\frac{1}{\rho_2} = \frac{1}{\rho_{20}} + \frac{1}{\rho_{2N}}$. The curvature $\frac{1}{\rho_{2N}}$ is caused by the bending moment due to the axial force N acting at the centre of gravity of the total section in State I, being displaced from the centre of gravity of the cracked section.

The mean curvature $\frac{1}{\rho}$, that takes into account the contribution of the concrete between cracks, is given by, see [3]:

$$\frac{1}{\rho} = (1-\zeta)\frac{1}{\rho_1} + \zeta\frac{1}{\rho_2} \quad (1)$$

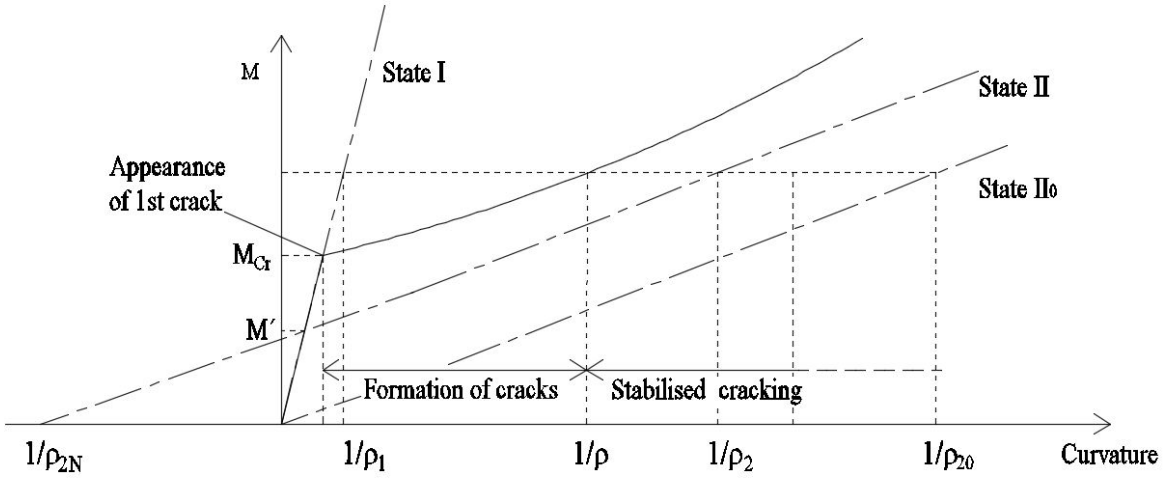



Figure 2b): Idealized moment curvature behaviour for bending moment and axial load [3].

The parameter ζ in pure bending or with bending moment and axial load is a function of: the maximum moment before cracking, $M_{cr}$; the resultant moment M at the centroid of the concrete section (middle height in the rectangular section); the adherence between steel and concrete defined by β . According to CEB Manual [13] ζ is given by:

$$\zeta = 1 - \beta \left(\frac{M_{cr}}{M}\right)^2 \text{ if } N=0 \tag{2a}$$

$$\zeta = 1 - \left(\frac{\sqrt{\beta} M_{cr} - M'}{M - M'}\right)^2 \text{ if } N \neq 0 \text{ and } \sqrt{\beta} M_{cr} \geq M' \tag{2b}$$

$$\zeta = 1 \text{ if } N \neq 0 \text{ and } \sqrt{\beta} M_{cr} < M' \tag{2c}$$

The moment $M'$ is the intersection of the moment curvature for state I and II, as represented in Fig. 2b).

### 2.3 Static equations

The static equations at the section reveal the axial load $N$ by the integration of normal stresses in concrete $\sigma_c$ and steel $\sigma_{si}$, and the bending moment $M$ by a similar procedure, that is:

$$N = \int_{A_c} \sigma_c \, dA_c + \sum_{i=1}^{n} \sigma_{si} \, A_{si} \tag{3}$$

$$M = \int_{A_c} \sigma_c \, y dA_c + \sum_{i=1}^{n} \sigma_{si} \, y_i A_{si} \tag{4}$$

where:
$A$ – area of active concrete;
$\sigma_c, \sigma_{si},$ – concrete, steel stress at $Y_i$ coordinate;
$A_{si}$– steel area;
$y_i$ and $y$ – distances to the centroid (see Fig.1);
G – centroid of the section

## 3 MOMENT CURVATURE EQUATIONS AT SECTION LEVEL

The assumptions exposed in 2.1 lead to three stages of the reinforced concrete section and corresponding equilibrium equations (3) and (4), that are:

a) uncracked concrete with the steel in the elastic range;
b) cracked concrete with the steel in the elastic range;
c) cracked concrete with the steel in the plastic domain;

In each stage the procedure is the following:

1- compute the axial load $N$;

2- the axial load value determines the neutral axis position $\eta$;

3- the bending moment $M^*$ that equilibrates the axial load is found at the neutral axis or at middle height $M$.

Section 3.1 presents the moment-curvature relation for uncracked reinforced concrete section. Section 3.2 resumes the expressions for $N$, $x$, $M$ or $M^*$ after concrete cracking. Using the methodology described in section 3.3 a moment-curvature relationship can be obtained. A closed form of the ($1/\rho$–$M$) relationship is derived by the use of Ramberg and Osgood equation and Goldberg Richard power representation, as presented in section 4.

Finally, the global behaviour of the member with the contribution of the stiffness due to the concrete between cracks is achieved.

### 3.1 Uncracked concrete

Before concrete cracking, the steel is in the elastic domain and the section is homogenized in concrete through the factor $\alpha = E_s/E_c$, multiplying the steel area [3]. The moment $M^*$ versus curvature $1/\rho$ relation is given, in function of the neutral axis position, $\eta$, by:

$$M^* = \frac{1}{\rho} E_c \left[ I_c + bh\left(\frac{h}{2} - \eta\right)^2 + A_s\alpha(h - a - \eta)^2 + A'_s\alpha(a - \eta)^2 \right] \tag{5}$$

The neutral axis in the elastic domain is located at the centroid of the section, defined by the particular position $\eta_G$, given through the equation:

$$\eta_G = \frac{\frac{bh^2}{2} + A_s\alpha(h - a) + A'_s\alpha a}{bh + (A_s + A'_s)\alpha} \tag{6}$$

The maximum moment before cracking, termed $M_{cr}$, that is the moment for which the maximum tensile stress in concrete is equal to the mean value of the tensile strength, $f_{ctm}$, and the corresponding curvature $(1/\rho)_{cr}$ are the following:

$$M_{cr} = \left[ f_{ctm} + \frac{N}{bh + (A_s + A'_s)\alpha} \right] \frac{EI}{(h - \eta_G)E_c}; \quad (1/\rho)_{cr} = \frac{M_{cr}}{EI} \tag{7}$$

With:

$$EI = E_c \left[ I_c + bh\left(\frac{h}{2} - \eta_G\right)^2 + A_s\alpha(h - a - \eta_G)^2 + A'_s\alpha(a - \eta_G)^2 \right] \tag{8}$$

and

$$I_c = \frac{bh^3}{12} \tag{9}$$

The parameter $\boldsymbol{\zeta}$ is computed by substituting the maximum moment before cracking, $\boldsymbol{M_{cr}}$, into equation (2). The moment-curvature relation in the uncracked concrete, state I, is then written as follows:

$$\left[\frac{1}{\rho}\right]_1 = \frac{M}{EI} \tag{10}$$

### 3.2 Cracked concrete

In the reinforced concrete section, after the cracking of the concrete under tension, two stages are considered: the tension steel $\boldsymbol{A_s}$ in the elastic range or in the plastic. The compression steel $\boldsymbol{A'_s}$ and the concrete are considered elastic.

#### 3.2.1 Cracked concrete and steel in the elastic range

At the initial stage, when the steel is yet in the elastic range, the static equation (3) is rewritten:

$$N = \frac{\eta^2 b E_c}{2\rho} + \frac{A_s E_s}{\rho}(a - h + \eta) + \frac{A'_s E_s}{\rho}(\eta - a) \tag{11}$$

Solving equation (11) in terms of $\eta$, the neutral axis position is given by:

$$\eta = \left[-\frac{(A_s + A'_s)\alpha}{b} + \left(\frac{\alpha^2 (A_s + A'_s)^2}{b^2} + \frac{2}{b}\left(\frac{\rho N}{E_c} + \alpha A_s (h - a) + \alpha a A'_s\right)\right)^{1/2}\right] \tag{12}$$

Equation (4) than becomes:

$$M^* = \frac{\eta^3}{3\rho} b E_c + \frac{(a - h + \eta)^2}{\rho} E_s A_s + \frac{(\eta - a)^2}{\rho} E_s A'_s \tag{13}$$

#### 3.2.2 Cracked concrete and steel in the plastic range

After the concrete cracking but with steel $A_s$ in the plastic range, that is the stress equal to $f_{yd}$, equations (11, 12 and 13) become:

$$N = \frac{1}{2\rho}\eta^2 b E_c - A_s f_{yd} + \frac{1}{\rho}(\eta - a) E_s A'_s \tag{14}$$

$$\eta = -\frac{\alpha A'_s}{b} + \left[\left(\frac{\alpha A'_s}{b}\right)^2 + \frac{2\rho}{b E_c}\left(\frac{a E_s A'_s}{\rho} + N + A_s f_{yd}\right)\right]^{1/2} \tag{15}$$

$$M^* = \frac{E_c b \eta^3}{3\rho} - A_s f_{yd}(a - h + \eta) + \frac{1}{\rho}(\eta - a)^2 E_s A'_s \tag{16}$$

The bending moment evaluated at the centroid of the concrete section (middle height) is:

$$M = M^* + N\left(\frac{h}{2} - \eta\right) \tag{17}$$

The equations (11) to (17) lead to the representation in Fig. 3 for a particular section and material properties (defined in the numerical results) and variable axial load ($N = 0; 100; 500\ 1000\ kN$), denoting that $1/\rho = \eta/(h - a)$. In Fig. 3 the elastic range both in

concrete and steel is not represented and the curves are truncated for the curvature corresponding to the ultimate compressive strain attained in the concrete. This analytical formulation, representing the two non-linear stages of the cracked concrete section, will be approximated by a single relation described in the next section.

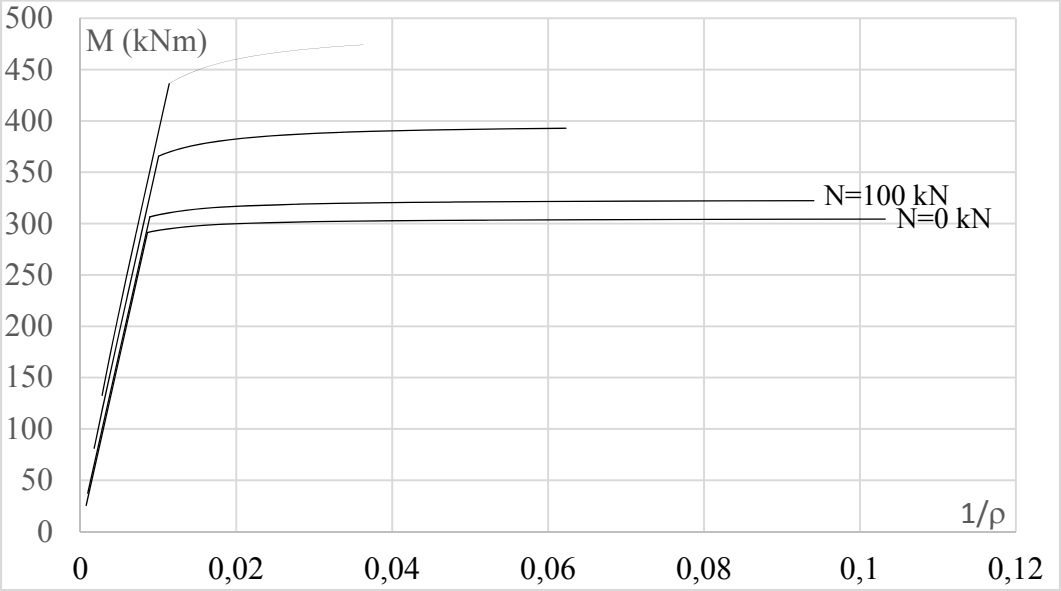



Figure 3: Analytical (*M-1/ρ*) relationship for a particular case.

### 3.3 Stiffnesses considered at elastic and plastic steel

The equivalent elastic stiffness $K_f$, point 1 in Fig. 4, is considered to represent the last stage of cracked concrete with steel in the elastic range. This means that point 1 is the end of the steel in the elastic zone. Its value is the derivative of:

$$K_f = \left.\frac{dM}{d\rho}\right|_1 \tag{18}$$

The equivalent plastic stiffness $K_p$ is computed at the point 2 with the corresponding bending moment, $M_2$, given by the derivative:

$$K_p = \left.\frac{dM}{d\rho}\right|_2 \tag{19}$$

The moment $M_2$ corresponds to the maximum bending moment obtained for the ultimate limit state (conventional rupture for bending moment with or without axial force [20]). The moment $M_0$ represented in Fig. 4, is established by:

$$M_0 = M_2 - K_p \frac{1}{\rho_2} \tag{20}$$

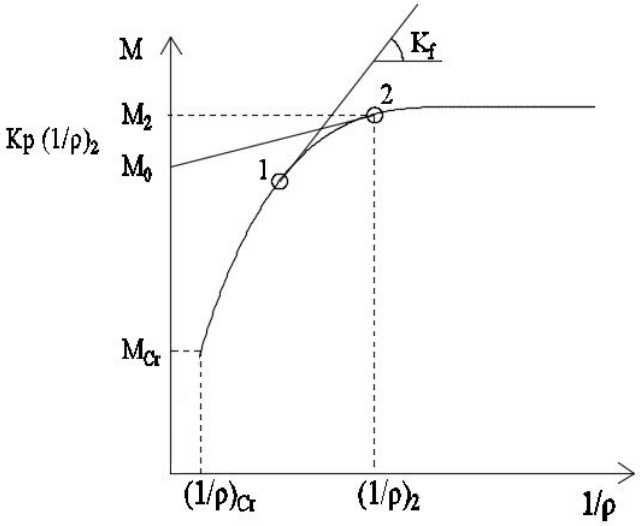



Figure 4: General moment-curvature diagram for cracked concrete.

## 4 M-$\rho$ OR $\rho$-M REPRESENTATIONS OF CRACKED SECTIONS

The representation of the moment curvature in a unique closed form can be obtained by two ways [2]: the curvature $1/\rho$ in function of the moment $M$ using the Ramberg - Osgood equation and the moment $M$ in function of curvature $1/\rho$ using the Goldberg Richard power representation. Both closed forms are described next.

### 4.1 Ramberg - Osgood equation

A unique curvature- moment ($1/\rho$–$M$) relation, after the concrete cracking, denoted by state II, can be written using the Ramberg-Osgood equation as follows [2]:

$$\left[\frac{1}{\rho}\right]_2 = \frac{M}{K_f} + \alpha \frac{M}{K_f}\left(\frac{M}{M_0}\right)^{n-1} \tag{21}$$

In the last equation n is an integer number of the Ramberg-Osgood adjustment and $\alpha$ is a parameter evaluated by:

$$\alpha = \frac{(K_f - K_p)}{K_p \left(\frac{M_2}{M_0}\right)^{n-1} n} \tag{22}$$

Fig. 5 compares the moment-curvature diagrams by Ramberg-Osgood equation with different values of the parameter n(=6, 20, 40) and the analytical curves for variable axial load applied in the particular section defined in the numerical results.

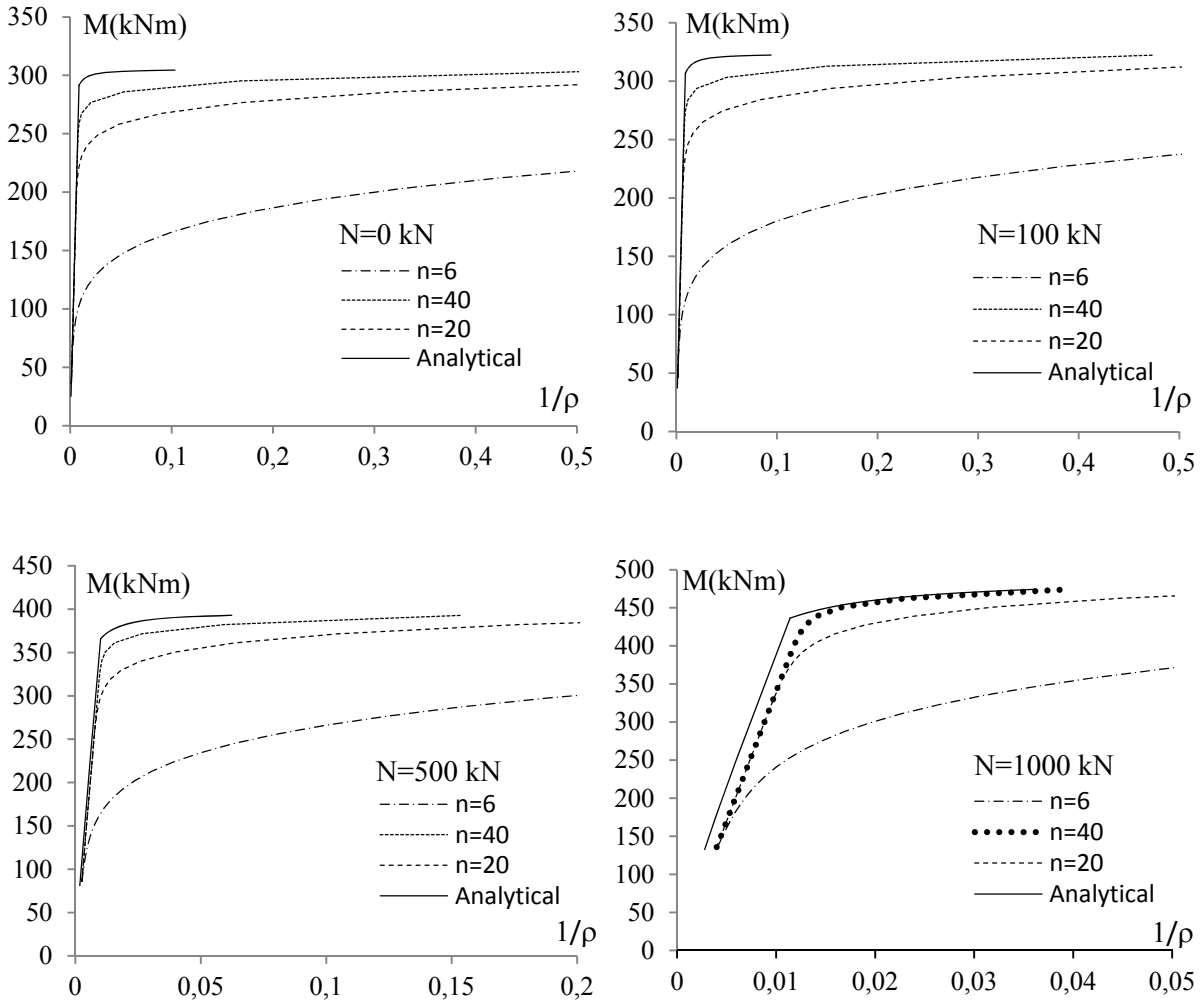



Figure 5: Moment-curvature diagram by Ramberg-Osgood equation.

## 4.2 Goldberg Richard power representation

A closed form of the ($M$-$1/\rho$) relationship is obtained using the Goldberg Richard expression [2]. This equation gives the bending moment ($M$) as a function of axial load ($N$), the initial ($K_f$) and the ultimate ($K_p$) flexural stiffnesses:

$$\boldsymbol{M} = \frac{(\boldsymbol{K_f - K_p})}{\boldsymbol{\rho}\left(\mathbf{1}+\left|\frac{(\boldsymbol{K_f-K_p})\frac{\mathbf{1}}{\boldsymbol{\rho}}}{\boldsymbol{M_0}}\right|^{\boldsymbol{m}}\right)^{\mathbf{1}/\boldsymbol{m}}} + \boldsymbol{K_p}\frac{\mathbf{1}}{\boldsymbol{\rho}} \tag{23}$$

Fig. 6 represents this closed form applied to the same particular cases of Fig. 5.

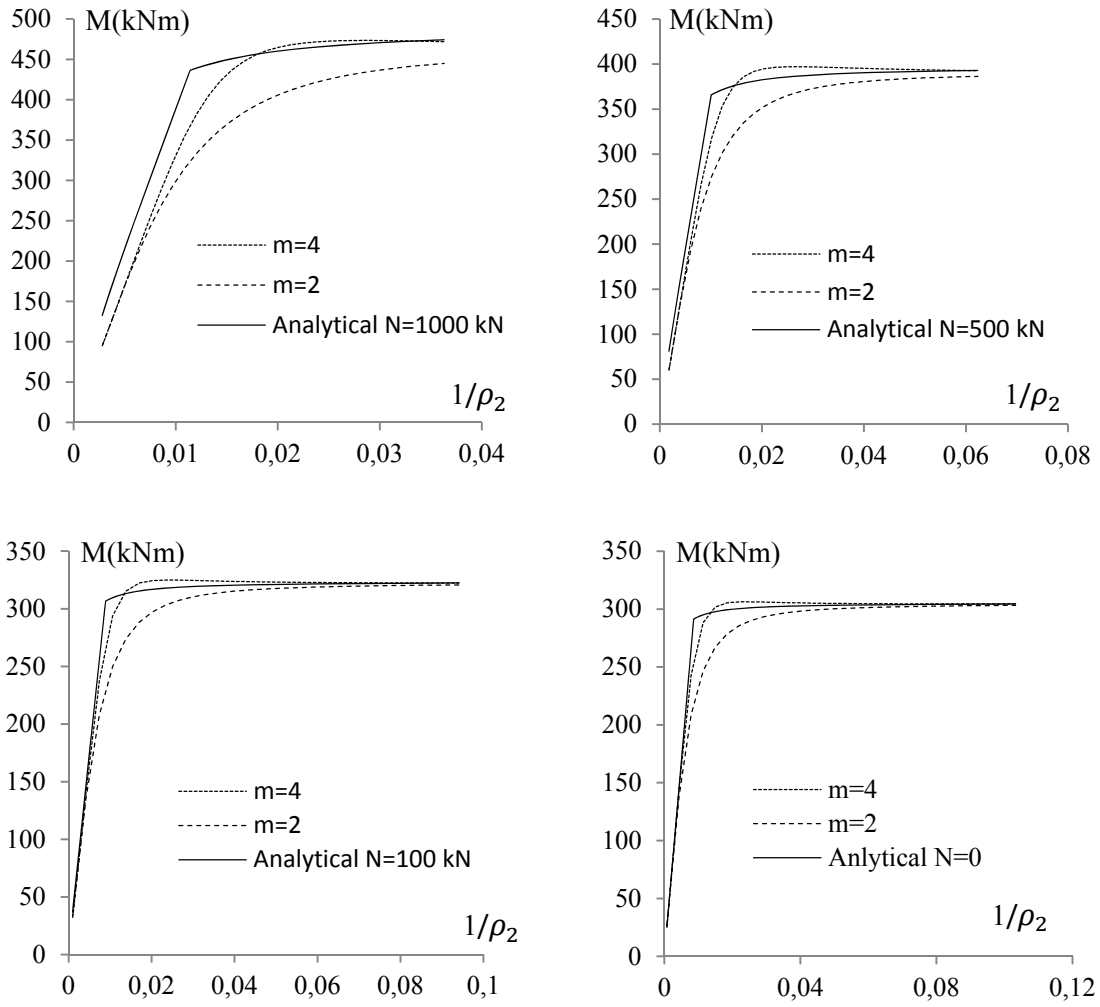



Figure 6: Goldberg Richard power representation of ($M$-$1/\rho_2$).

## 4.3 Comparison of the two representations

The previous moment-curvature approximation pointed out the need of the best parameters (n to Ramberg-Osgood, or m to Goldberg Richard equations) to employ in a particular case. Further, there is not a parameter n, for Ramberg-Osgood representation, satisfying simultaneously the moment and the corresponding curvature at the rupture for a wide range of loads in opposite to the choice of the m parameter in the Goldberg Richard equation. A comparison of the two representations is made in Fig. 7 for the same particular cases of Fig. 5, considering in the Ramberg-Osgood equation n=40 and in-the Goldberg-Richard m=4. In the figure, indicated as theoretical, is the analytical relationship composed by the equations for cracked concrete with elastic (equation (11)) and plastic steel (equation (17)). In this figure the uncracked concrete, for curvatures between *0* and $(1/\rho)_{cr}$ and moments between *0* and $M_{cr}$, is not represented, that is, only the cracked concrete is approximated by the Ramberg-Osgood and Goldberg-Richard equations.

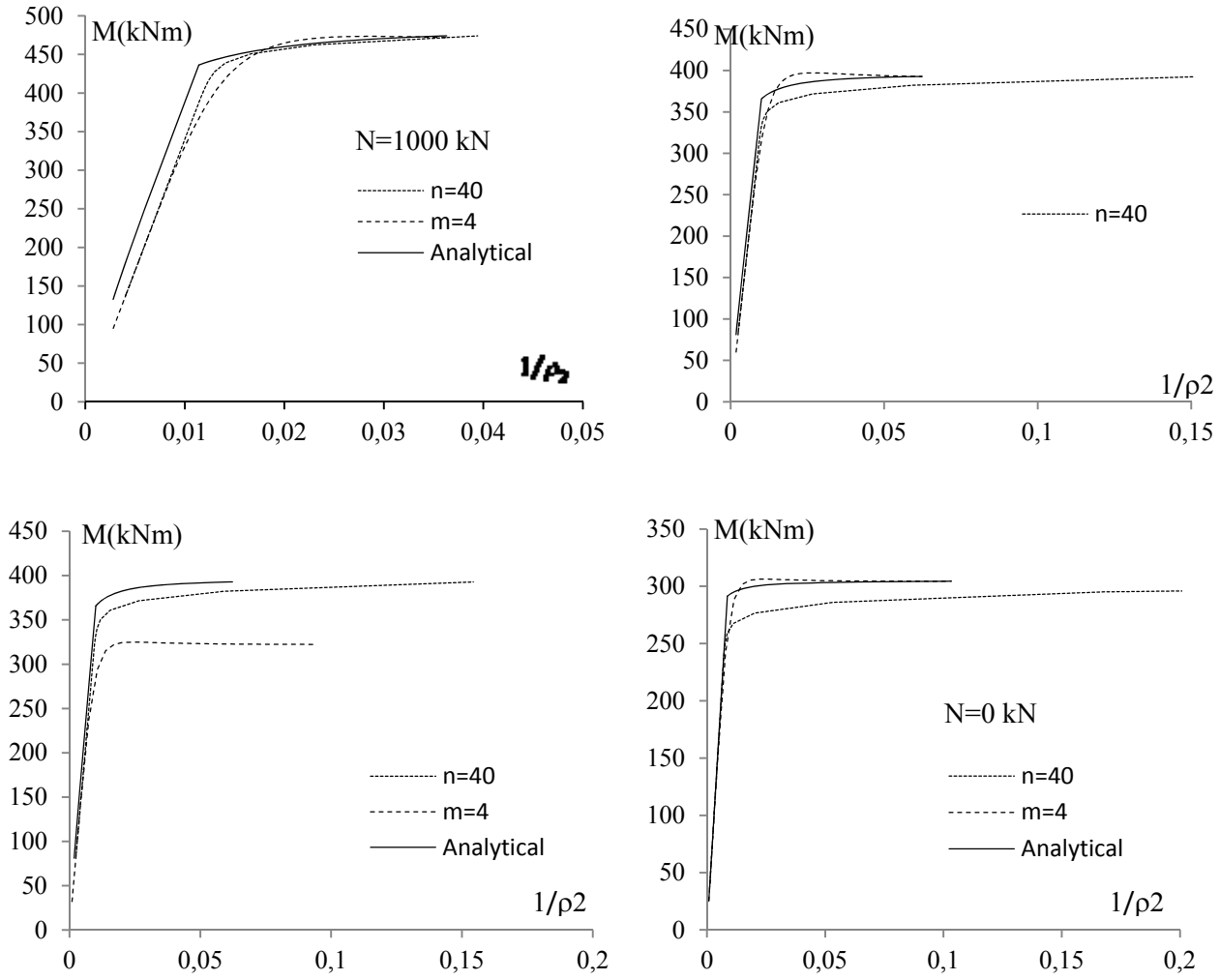



Figure 7: Comparison of the moment-curvature diagrams for cracked concrete.

The preference for the Ramberg-Osgood representation is due to the fact that the present model needs the curvature 1/ρ as a function of the bending moment, in order to apply equation (1). It can be noted that the objective is to find a unique representation to the cracked stages, $1/\rho_2$, because the interaction used in (1) considers automatically the uncracked stage $1/\rho_1$.

## 5 NUMERICAL RESULTS

### 5.1 Moment curvature diagrams: Rectangular section under variable axial load

Figure 8 shows the moment curvature diagrams obtained for a rectangular symmetric reinforced cross section with variable axial load ($N = 0, 100, 500, 1000kN$).

The cross section geometry is: b= $0{,}2m; h = 0{,}4m; a = 0{,}02m; A_s = A_s' = 0{,}0021m^2$. The material properties are : $E_s = 200GPa$; $E_c = 20GPa$; $f_{yd} = 400MPa$; $f_{ck} = 30MPa$ and $f_{ctm} = 3MPa$.

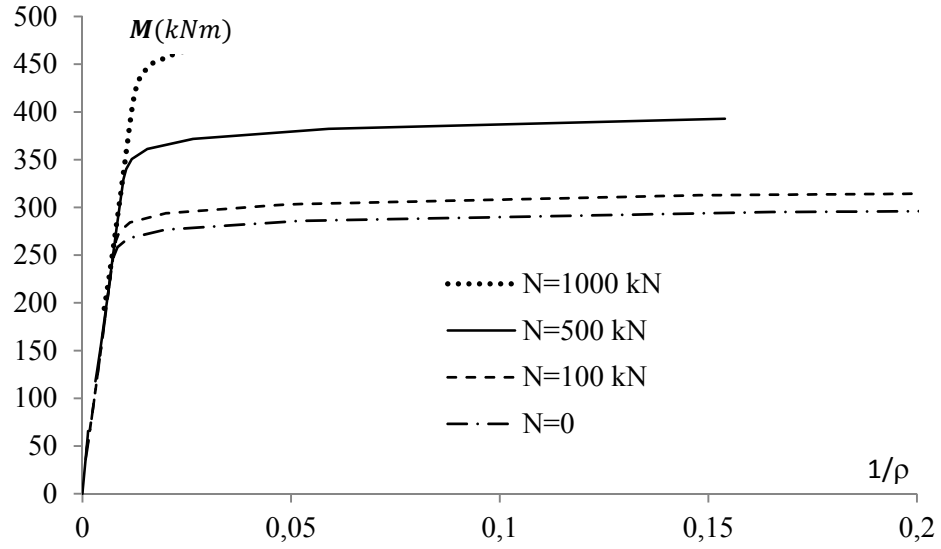



Figure 8: Moment- curvature diagrams of a RC section.

## 5.2 Moment curvature diagrams and deflection of a beam

The beam represented in Fig. 9a), is tested by Kwak et al [8]. The characteristics of the beam are the following: b=90cm; d=27.23cm; H=30.48cm; B=15.24cm; ρ´=0; ρ=0.0062; $E_s$=1.98×10$^6$kg/cm$^2$; $E_c$=2.71×10$^5$kg/cm$^2$; $f_y$=3236kg/cm$^2$; $f_c$=323kg/cm$^2$. The mean concrete tensile strength used in the present analysis is: $f_{ct}$= $f_c$/10. Figure 9b) represents the moment curvature diagram considering: the Ramberg-Osgood equation with n=80; the tension softening interaction given from equation 1 with the definition of ζ by equation 2, considering β=1.0 and 0.5; the results of Kwak et al [21] analysis made with 24 finite elements. Figure 9c) represents the load – central deflection diagram obtained with the model (for the two values of β) and the experimental and analytical results of Kwak et al [21].

It can be observed in the figures that the present analytical model approximates the experimental results.

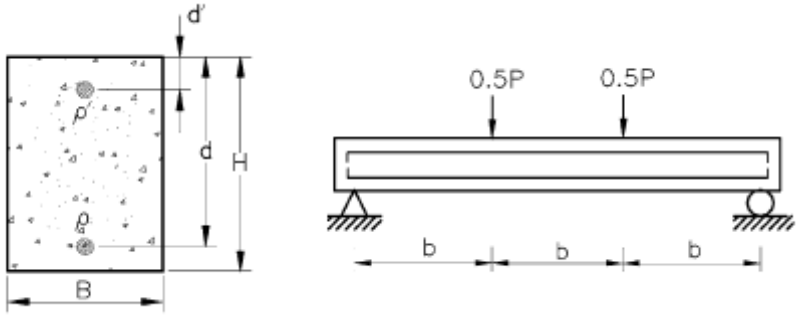



Figure 9a): Loading type and cross section of beam T1MA.

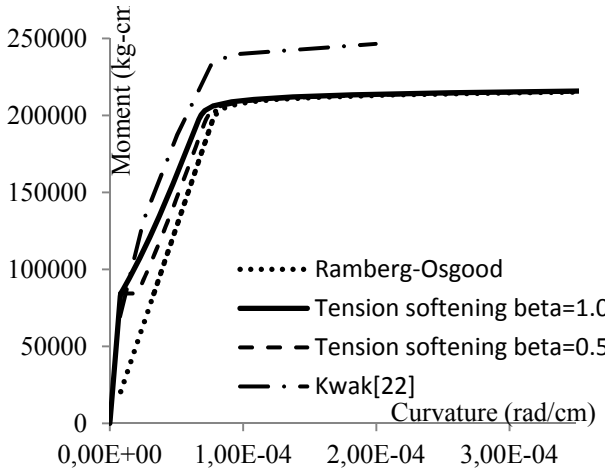



Figure 9b): Moment-curvature diagrams.

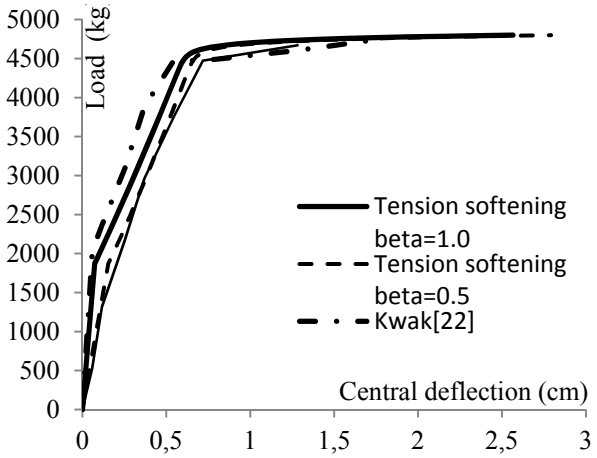



Figure 9c): Load deflection diagram.

## 6 CONCLUSIONS

The present work presents:

- the deduction of the moment curvature diagrams for singly or doubly reinforced concrete rectangular sections by MAPLE software with the three stages: *i*) uncracked concrete; *ii*) cracked concrete and steel in elastic domain; *iii*) cracked concrete and steel in plastic domain;
- the representation in a unique closed form of the curvature - moment relation using the Ramberg - Osgood equation. The moment – curvature equation with the Goldberg Richard power representation is also deduced and compared to the previous one;
- the analytical mean curvature of the structural member with the $\zeta$ parameter of EC2;
- the deflection of the cracked member by the integration of the curvatures;
- the load – deflection curve of a RC beam and the comparison to other models and experimental results.

## ACKNOWNLEGEMENTS

This work has been supported by the Fundação para a Ciência e a Tecnologia (FCT) under project grant UID/MULTI/00308/2013.

## REFERENCES

[1] Marques T., Ferreira, C. Barros H.; "Moment-curvature diagrams for reinforced concrete section design", *2nd Int. Conf. on Numerical and Symbolic Computation Developments and Applications*, SYMCOMP 2015, Faro, 2015.

[2] Faella, C.,Piluso,V., Rizzano, G; *Strutural Steel Semirigid Connections – Theory, design and software*, CRC Press, 2000.

[3] EN 1992-1-1; Eurocode 2 – *Design of concrete structures – Part 1-1: General rules and rules for buildings*, December, 2004.

[4] Zheng H., Jinping O., Bo W., "The trilinear moment vs. curvature relationship of concrete beams reinforced with fiber reinforced polymer (FRP) rebars", Composite Structures, 77, pp 30-35, 2007.

[5] Torrico, F. A., *Theoretical and experimental analysis of slender high-strength concrete columns, considering the ductility,* PhD thesis, São Paulo University, Brazil, in Portuguese, 2009.

[6] Challamel, N.; Hellesland, J.; "Buckling of softening columns in a continuum damage mechanics perspective – local versus non-local formulation", Journal of Mechanics and Solids, 39, pp 229-242, 2013.

[7] Ponaya, S.; Teeboonma, U.; Thinvongpituk, C.; "Plastic collapse analysis of thin-walled circular tubes subjected to bending", Thin Walled Structures, 47, pp 637-645, 2009.

[8] Vincent Picandet; Noel Challamel; Sovannara Hin; "Buckling and post-buckling of gradient and non-local plasticity columns experiencing softening", Int. Journal of Solids and Structures, 51, pp 4052-4067, 2014.

[9] Casandjian, C., Challamel, N, Lanos, C., Hellesland, J.; *Reinforced concrete beams, columns and frames: Mechanics and Design*, John Wiley& Sons, 2013.

[10] Naci Caglar; Aydin Demir; Hakan Ozturk; Abdulhalim Akkaya; "A simple formulation for effective flexural stiffness of circular reinforced concrete columns", Eng. Applications of Artificial Inteligence, 38, pp 79-87, 2015.

[11] Cevik, A., Arslan, M. H., Koroglu, M. A., "Genetic-programming-based modelling of RC beam torsional strength", KSCE J. Civ. Eng. 14 (3), pp 371-384, 2010.

[12] Chen, H. M., Kao, W. K., Tsai, H. C., "Genetic-programming for predicting a aseismic abilities of school buildings", Eng. Appl. Artif. Intell. 25 (6), pp 1103-1113, 2012.

[13] *CEB Design Manual on Cracking and Deformation*, CEB, 1985.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# INVESTIGATIONS ON AN INNOVATIVE METAL-THERMOPLASTIC COMPOSITE ASSEMBLY

**Simon Paroissien**[1], **Patrick Rozycki**[2], **Thierry Renault**[3]

[1] Direction de l'Expertise et de la Recherche,
IRT Jules Verne
Chemin du Chaffault, 44340 Bouguenais
E-mail: simon.paroissien@irt-jules-verne.fr

[2] Institut de Recherche en Génie Civil et Mécanique (GeM) – UMR CNRS 6183,
Ecole Centrale de Nantes
1 rue de la Noë, 44300 Nantes Cedex 3
E-mail: patrick.rozycki@ec-nantes.fr

[3] Faurecia
2 rue Hennape, 92000 Nanterre
E-mail: thierry.renault@faurecia.fr

**Keywords:** Multimaterial, FEM, CMT-Pins, Assembly, analytical

**Abstract.** *Multimaterial assembly is one of the main answers of the car industry to weight reduction issues. As their introduction in the automotive industry is still recent, many multimaterial solutions have not been tested in crash, endurance and specific loading. Pre dimensioning tools and numerical models are still very challenging. The present study aims at addressing some of these limitations. From the industrial issue of a seat back rest, an innovative multimaterial assembly based the CMT pin technology has been designed. A first model based on multilayer element is proposed to represent the global behavior of the assembly. Then, in order to investigate the multiple mechanisms (metal to composite direct adhesion, pin-composite interaction…) that rule the assembly, the present article focused on a simple unidirectional piece: a double lap shear specimen. Some studies can be found on this kind of specimen but for different materials and process. This geometry allows us to perform traction test in which both local and global data could be post treated. Several parameters have been tested to evaluate the influence of pins' arrangement on the model behavior. From this experimental campaign, a numerical model with LS Dyna software and an analytical pre dimensioning tool are then proposed and compared. These first results encourage us to continue this study with future works to find the ideal texturing answering conception specifications.*

# 1 INTRODUCTION

## 1.1 Context

During the past decades composites materials have taken more and more place in the transport industry. This is particularly true for the small series, where the price of the material have less impact on the production costs than in the large ones. One can take for example the evolution of the percentage of structural composite in the total weight of a plane in airbus production which goes from nearly 5% in the A300 (1972) to more than 50% for the more recent A350-900 XWB (2013). However, if composites are now common in this industry, their introduction in the large series of the automotive field is still very challenging. $CO_2$ emission standards required by various countries (Europe, USA, but also China) are forcing this sector to consider cutting edge solutions in weight reduction. The introduction of composite in structural applications is one of the main tracks, but this raises other issues like the way of joining the new composite parts with others that, for economic or structural reasons, will remain in metal. The joining method is a key choice so that the assembly can assure a safe load transfer while meeting the drastic specifications of the automotive industry.

## 1.2 State of the art

The two most common ways to assemble structural application in car industry are mechanical fastening and bonding. Both of them get their own advantages and limitations. Bolting benefits of a long return on experience and is a fast, cheap process but not the more efficient considering the additional weight, stress concentration and the weakening of the CFRP during the drilling process [1, 2, 5 and 13]. On the other hand bonding allowed the preservation of both substrates [7] but its efficiency decreased in extreme conditions (extremes temperature, high rate of humidity) and still need an important overlap surface. Hybrid solutions or the use of safety rivets can remedy to those limitations but, again those actions increase the weight of the structure. Apart from those industrial technologies, some other joining methods can be found in the literature. In [4] an arrow shape is laser cut and bent in a steel sheet. Those arrows are then inserted inside the composite. This study shows good results for 0° and 30° loading under various strain rates. The Surfi-Sculpt™ process, patented by The Welding Institute aims at sculpting a metal surface thanks to an electron beam. Once treated, the metal can then be assembled to a composite (Comeld™ process). Various papers explore this technology and its simulation [11 and 12]. Among the inconvenient of Comeld, the weak peeling strength, and the influence of the process on fiber orientation, and the resin distribution can be cited. The firm Alstom has also patented an assembly process between a metallic and a non-metallic part. Drops of solder formed by an arc welding process are placed, on the surface of the metallic part. Once cooled each of this droplets forms an anchoring point to the composite.

Working on the same principle, the Austrian society Fronius filled a patent (Figure 1) on a process named Cold Metal Transfer (CMT) allowing the welding of small pins on a metallic surface. This technology offers a better control of the shape of the pin formed that way than in the process mentioned above [8]. Experimental investigations have been conducted on CMT process in [10] showing promising results on thermoset CFRP – Steel assembly. The performances are then compared to those of the arrow shape in [9] showing comparable results in terms of maximum transferable joint forces and energy.

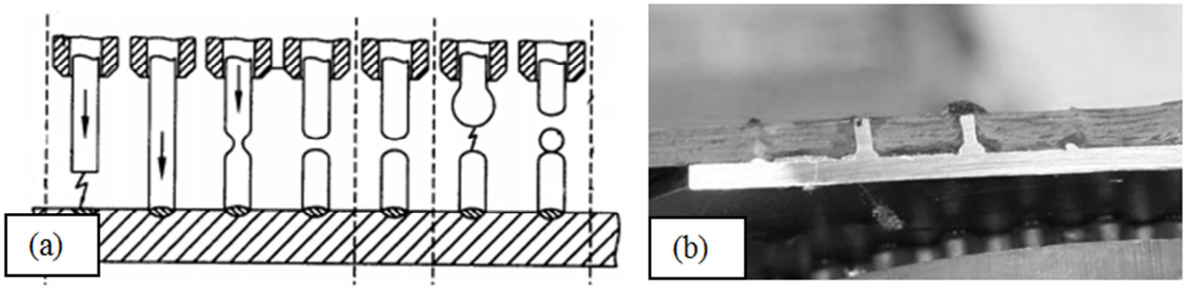



Figure 1: (a)Fronius‘ patent (b) anchoring of CMT in pins as in thermoplastic composite as developed in LIMECO project

This study focus on the evaluation and simulation of a multimaterial structure, joined with a CMT pins interface, more precisely on the case of a front seat back rest.

## 2 TRIANGLE CASE OF STUDY

### 2.1 Methodology

A thermoplastic composite triangle is assembled to the head of a steel articulation mechanism thanks to a metallic disk textured by the CMT pin technology (Figure 2). The CMT pins penetrate into the composite, locking the assembly in the transverse direction. A polyamide resin is added to avoid sliding between the two parts. Finally a second metallic triangle is welded to the articulation so that a moment could be transferred.

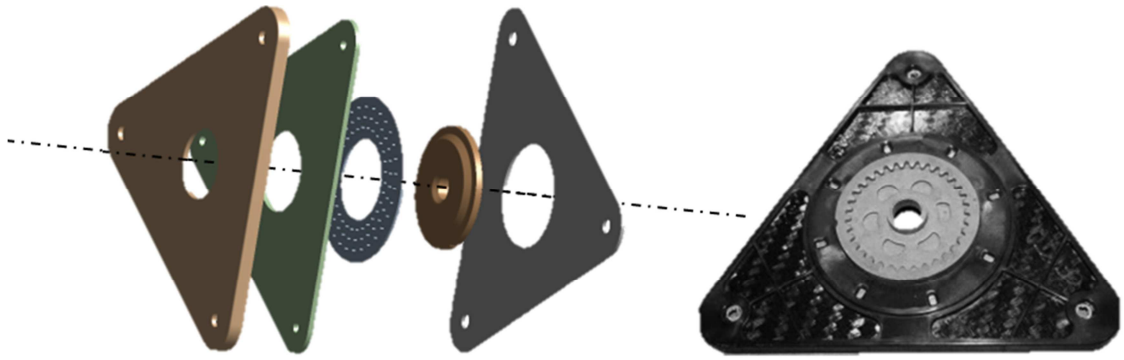


Figure 2: Assembly’s geometry

This assembly aims at transferring a torsional moment, along the rotational axis. The interface is especially complex and nonlinear, a fine representation of the local behavior of the structure would be extremely penalizing for calculation time. For this reason, it has been decided to search for the simplest possible method that conserves good global results. A first hypothesis has been to infer that the stiffness of the assembly is mainly due to interface’s texturing so the other parts of the structure can be considered perfectly bonded. The goal here is to put forward the geometrical simplifications that will reduce calculation costs. The choice has been made to represent the bonded structure with multilayer element. Those elements are an extension of the monolayer shell theory of Reissner. This type of element has the advantages to integrate different materials along its thickness (one material law per integration point), so with this element, one can represent a superposition of different parts inside a single element.

## 2.2 The multilayer model

The principle is to project on a plane surface the geometric profiles of the different pieces of the assembly. Thanks to multilayer elements, the section of each portion of the piece is then reconstructed. Here the assembly will be represented with two parts: the first part will represent the overmolding resin + composite, the other the metallic part. As the thickness of the shell is important regarding to the mesh size, several integration points per layer have been implemented.

The contribution of the interface is represented with a nonlinear torsional spring. The comportment law of this spring is composed of a torsional stiffness, a nonlinear moment-angle curve and a moment of failure. At first, these parameters are set up from experimental data; future works will propose an analytical model allowing identifying them.

To validate the model, numerical results have been compared to experimental ones from an experimental campaign. Specimens are fixed on a test bed with the following boundaries conditions: a constant low rotational speed is imposed to the steel triangle while composite triangle rotations are locked. This triangle remains free to translate along the rotation axis (Figure 3).

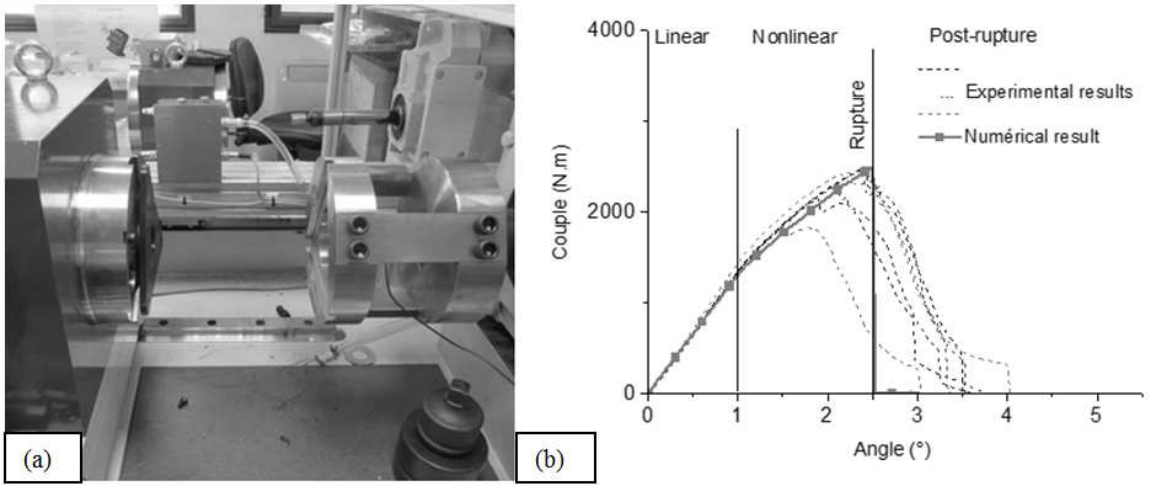



Figure 3: (a) Experimental set up, (b) Comparison between experimental and numerical results

Thanks to spring parameters set up, numerical results show good accordance with experimental ones.

## 2.3 Limitation of the model

A method giving spring behavior from the pins disposition is still necessary. To define this method, investigations have been made on another kind of geometry. Indeed, the assembly's geometry leads to complex solicitations and experimentally the metal composite interface is difficult to instrument correctly. Moreover the finite element model requires an important number of elements to model both pin scale and assembly's scale with enough precision. In order to investigate the multiple mechanisms (metal to composite direct adhesion, pin-composite interaction…) that rule the assembly, the present article focused on a simpler unidirectional piece described in the following part. Two methods of simulation, one using the Finite Element Method (FEM) and the other using an analytical algorithm, are then proposed and compared.

# 3 EXPERIMENTAL CAMPAIGN

## 3.1 The geometry tested

The geometry of the specimen should enable to isolate a single pin's behaviour, and the influence of the number of pins and their disposition on the specimen's behaviour. Moreover instrumentation should be simple and give access to the local comportment of the interface. In papers and more precisely in [13] one can find the use of a double lap shear speciemen (DLS) to characterize an interface working under a shear loading. This geometry has the advantage to limit the peeling phenomenon due to the lack of planeity in the loading. The geometry described in Figure 4, is composed of two glass fiber shells, termostamped on both sides of a metalic plate. The recovery zone is textured with the CMT pin technology. Outside of the interface area, a metal wedge is inserted between the two composites to maintain the thickness of the composite part of the specimen, as described in the figure below. Pin's height is set up at 3 mm so that the head of each pin flushes with the top of composite layer.

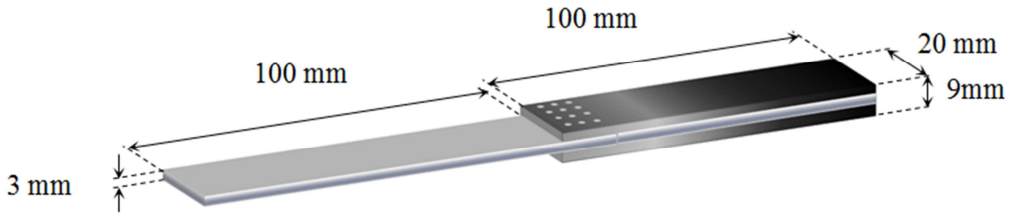



Figure 4: Specimen geometry

Three pin's dispositions have been tested: 6, 3, and 1 raw(s) of 4 pins. Only cylindrical shaped pins of 1mm diameter are welded. To ensure the repeatability five tests have been realized on each configuration.

## 3.2 Experimental set up

The tensile tests were carried out on a quasi-static testing machine INSTRON 5584 at a 1mm/min speed (Figure 5.a.). Resultant efforts were measured by a force cell. A measure of the relative axial displacement of the two parts is performed by digital image correlation (DIC). Two high-definition recording cameras filmed the test: one record the edge of the DLS (Figure 5.b.), the other one of the two faces (Figure 5.c.).

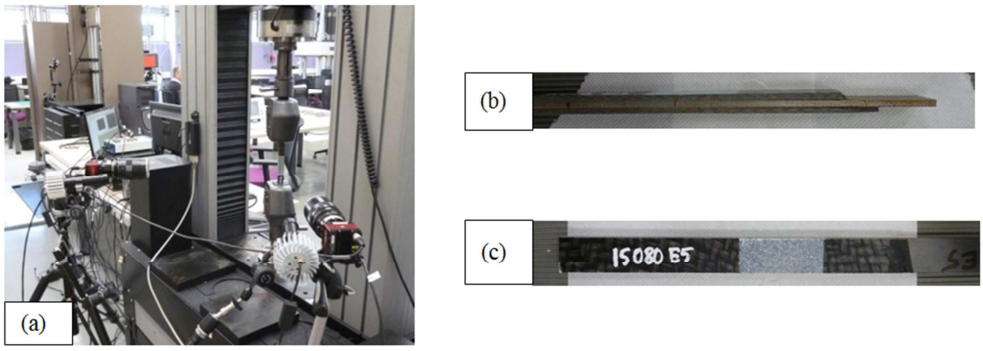



Figure 5: (a) Experimental set up, (b) DLS (edge), (c) DLS (face)

### 3.3 Results and discussion

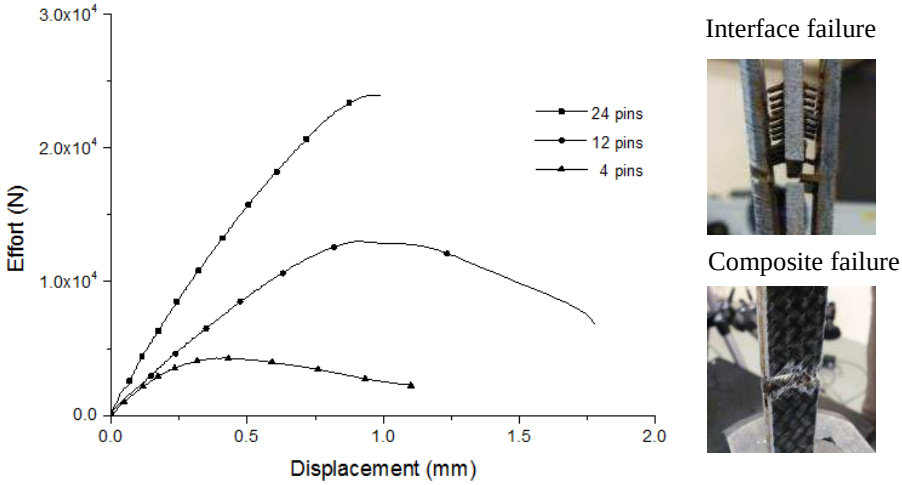



Figure 6: Effort – Displacement graph.

The experimental results are given in Figure 6, where the effort displacement law is given for the three tested configurations. For readability only one representative curve of each configuration has been displayed. Under a tensile loading, the DLS shows a linear elastic behavior, followed by a plastic softening. This softening leads to two kinds of failure, depending on fiber's orientation and effort's level. On the one hand (and in most situations) a noisy failure of the can be heard, followed by a drop of effort level. Moreover almost none normal to the interface displacement is measured by the DIC before failure. This is characteristic of a metal to composite direct adhesion [3] involved by thermo stamping. On the other hand, if fiber's orientation has been too disturbed and if the efforts are important enough, a failure of the pure composite part of the DLS can be observed.

An important dispersion can be observed between the five samples of the 24 pins configuration. This is due to resin creep during the thermo-stamping operation. In the two more flexible cases, this leads to a composite kind of failure. In the following parts of the paper, for this configuration, the simulation results will be compared to the stiffer experimental result.

## 4 NUMERICAL MODEL

### 4.1 Geometry and meshing

The geometry and meshing are realized with Ansys Workbench v15 software. The explicit solver used is LS Dyna R7.1. For this model, only the pure composite and the interface part of our DLS are modeled. The steel part is considered rigid. Two planes of symmetry allow simulating only a quarter of the specimen (Figure 7). The model is meshed with one integration point hexahedral elements. Mesh size goes from 0.1 mm on the pin to 0.3 mm on the rest of the interface.

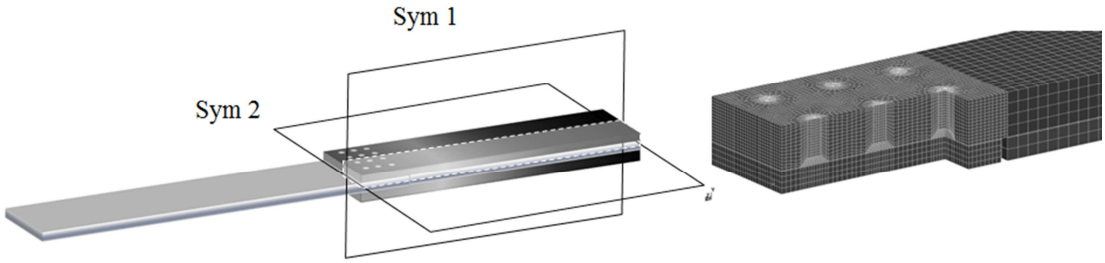



Figure 7: Geometry and meshing

To avoid initial penetrations, a small gap of 0.025mm is left between the pin and the surrounding composite. A friction coefficient of 0.05 is added. Symmetry conditions are applied to the corresponding faces. The end of composite part is considered fixed and a constant speed is imposed the other end of the DLS, on the steel section.

Concerning pins, two parts can be considered: the welding cone at pin's base and pin's stem/head. The pin, made of $G_3Si_1$, is characterized at solid state. A traction test performed on the welding wire shows important differences from the values given by the provider. It has been settled that the welding cone will be modeled according to providers' data and the wire, which didn't melt during the welding process, with the property obtained from the traction test. The material law n°261 of LS Dyna developed by Pinho & al in [6] is used. This material law allows the user to implement the in-plane shear behavior with a defined curve. At the neighborhood of each pin, the material law is replaced by an orthotropic linear law so that the local artificial constraints don't damage the composite. This hypothesis is justified by observations on the post rupture aspect of the interface. The direct adhesion is modeled with layer of cohesive elements. No data were given to implement the comportment law of those elements. So the general law n°138 has been chosen and its parameters were calibrated on experimental results.

The comparison between numerical and experimental results is given in Figure 8.

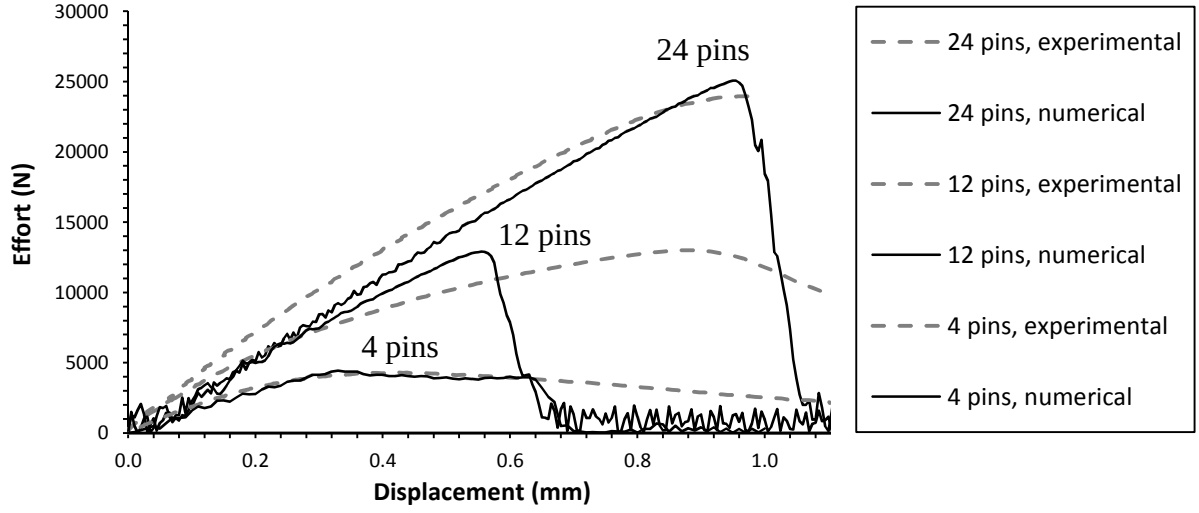



Figure 8: Comparison of the numerical and experimental results

Numerical results show a good accordance with experimental ones. Maximal efforts especially, are obtained with a 5% precision.

To link this result to the triangle issue, the local behavior of one pin in a composite environment has been studied. By refining the post processing it has been noticed, that a single force-displacement law, linking the effort at the base of the pin to its head's displacement, can

be isolate for all the pins, in all the studied configurations (Figure 10). So an assumption has been made: the behavior of a pin can be found from a numerical model at pin' scale. From this hypothesis, an analytical model giving the DLS traction law has been implemented.

### 4.2 Scale of the pin

Inside each DLS it has been noticed that a Representative Volume Element (RVE) can be isolated, (Figure 9). Thanks to geometrical and loading symmetries according to the plan (y,z), only half of the RVE has been represented.

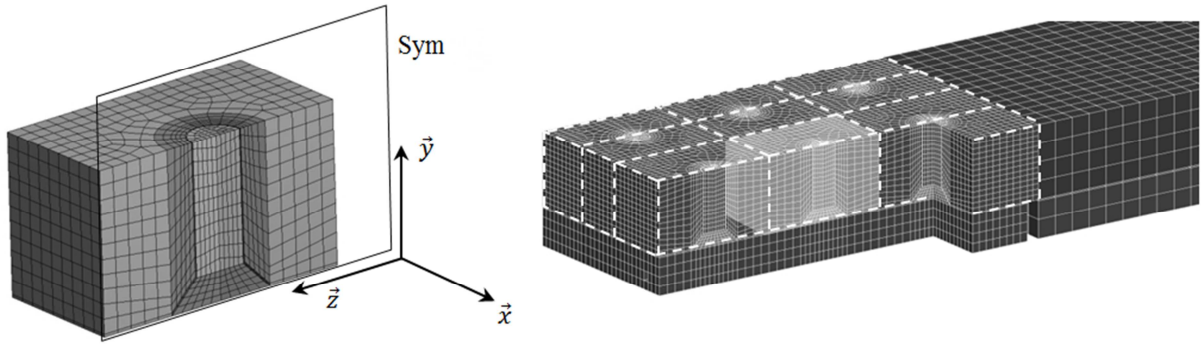



Figure 9: RVE Geometry

The same contact, material laws, and type of elements are used as in the DLS model. The bottom face of the geometry is fixed, and symmetry conditions are applied on the faces normal to x-direction. A constant speed, in the z direction is imposed on both faces normal to z. The comparison between RVE's behavior and each pins row from each configuration is given in Figure 10 (In the legend, "XX Pins RY" stands for "XX pins configuration, row number Y"):

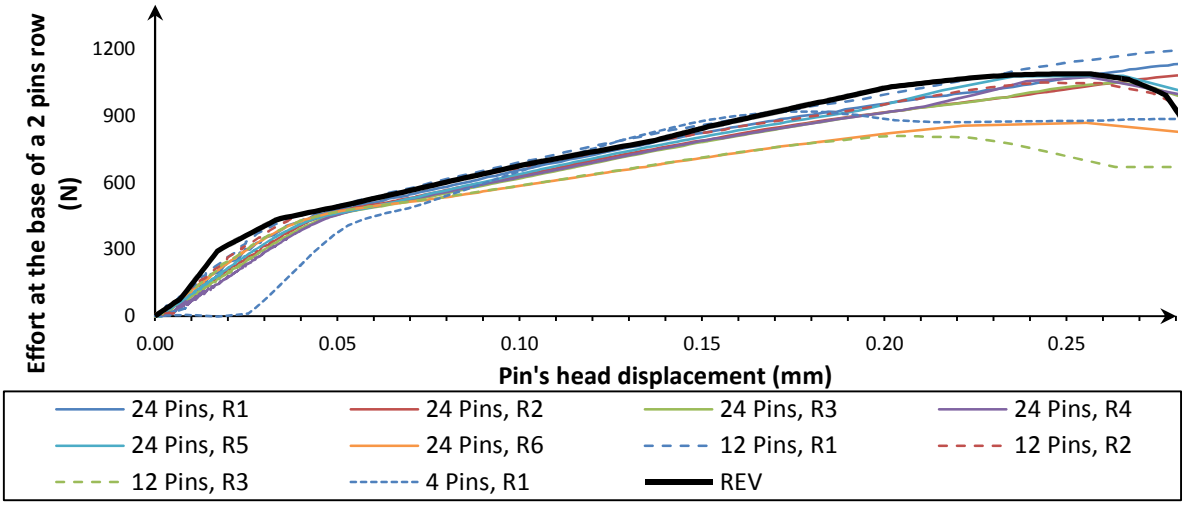



Figure 10: Pins local effort displacement law

For each pin's row, for each configuration (4, 12 and 24 pins), the Figure 10 gives the law linking pin's head displacement to the effort transferred through pin row to the metallic plate. This graph shows that the behavior of the pin inside the RVE is representative of the behavior of any pin inside the DLS. A bilinear flexion law is extracted from this graph. The rupture criterion which depends of the cohesive layer (i-e: metal to composite adhesion) is not associated to this law.

Now that the behavior of one RVE has been isolated, an analytical algorithm giving interface behavior is proposed in the next part.

# 5 ANALYTICAL MODEL

## 5.1 Presentation of the problem

For the next part, the following diagram (Figure 11: The CMT Pin link of the problem will be used:

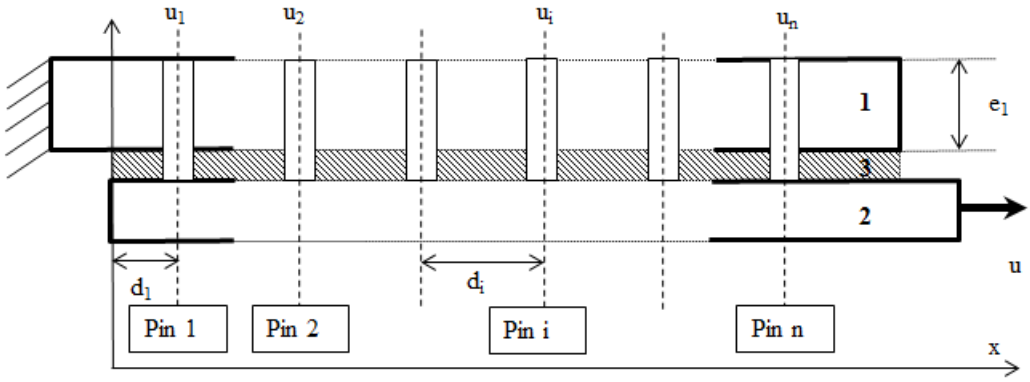



Figure 11: The CMT Pin link

## 5.2 Equivalent model

The model is made according to the following hypothesis:

- In comparison to other materials, the steel plate is considered rigid.
- The thermoplastic composite is supposed to be perfectly oriented, with the x direction as one of the orthotropic directions.
- The displacement of a section of the composite, surrounding a pin is equal to the displacement of the head of this pin, $u_i$.

Pin's behavior and cohesive layer on a RVE can be represented with a single multilinear, force displacement law $F_{RVE,i}$. The composite transmits efforts from one pin to another, with a stiffness noted $K_{dyn}$.

Once inserted in the global geometry, the system becomes:

$$
\begin{aligned}
&F = \sum_{i<n} F_{RVE,i}(u - u_i) \\
&u_1 = \frac{F}{K_{dyn,1}} \\
&u_2 = u_1 + \frac{K_{dyn,1}.u_1 - F_{RVE,1}(u - u_1)}{K_{dyn,2}} \\
&\forall i \in [2, n-1], u_{i+1} = u_i + \frac{K_{dyn,i}.(u_i - u_{i-1}) - F_{RVE,i-1}(u - u_i)}{K_{dyn,i+1}}
\end{aligned}
\tag{1}
$$

### 5.3 Different configurations

The algorithm implemented in Matlab gives the following results (Figure 12) for the different configurations experimentally tested.

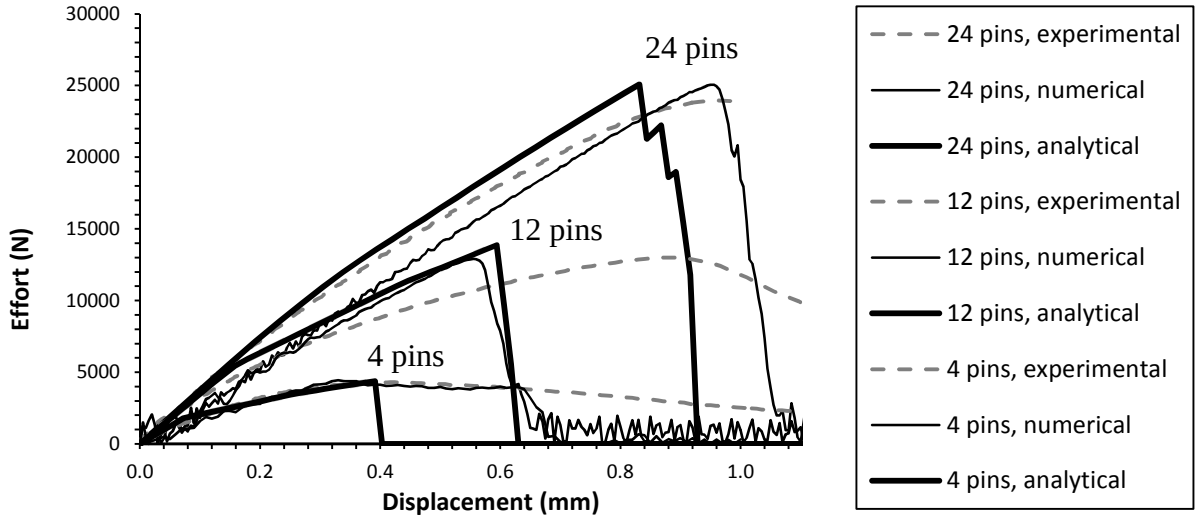



Figure 12: Comparison of analytical and experimental results for 24, 12 and 4 pins configurations

Again, the analytical results are close to experimental ones. The difference of stiffness in the 12 pins case can be due to an imperfection in fiber's orientation in the experimental test.

In this algorithm, the behavior part out of section is calculated analytically because of the simple geometry of the DLS. For a more complex case the interface can be simulated with a nonlinear spring. As pins are not modeled anymore, meshing the interface becomes much simpler and shell elements can be used to model the two parts. In the case of the 24 pin interface, the geometry used in the numerical model becomes (Figure 13):

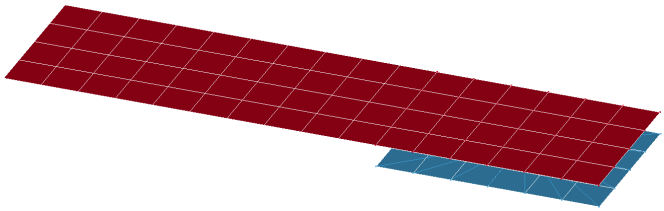


Figure 13: Simplified geometry using non linear spring as an interface

A comparison of experimental, numerical and analytical results is given in the following graph (Figure 14). In the legend, the "Ei" stands for "experimental test number i":

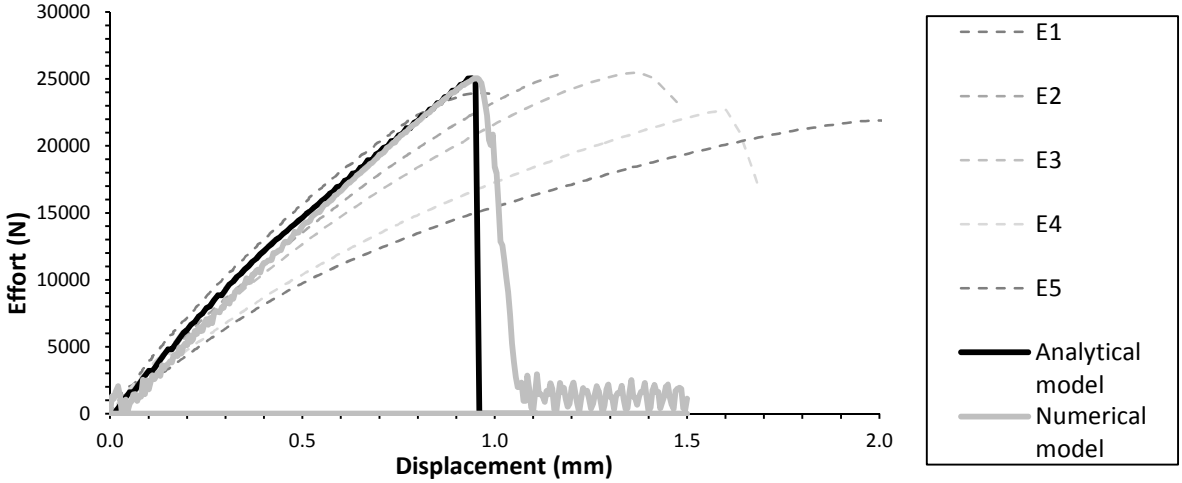



Figure 14: 24 pins configuration, experimental, numerical and analytical results

The real gain with the analytical model is on calculation time. For the numerical model the simulation was launch on a cluster with 8 nodes with a 4 x AMD Shanghaï 2380 2.5 GHz Quadcore processor and last more than 9 hours, whereas the analytical algorithm is solved quasi instantly on a single processor (i7-3630QM CPU, 2.4 GHz). This is especially convenient for the pre dimensioning or optimization phases.

## 6 CONCLUSION

The CMT pin technology is used to produce a multimaterial assembly between a steel part and a PA6, glass fiber thermoplastic composite. As a fine representation of the interface geometry would be complicate and expensive, a multilayer model combined with a nonlinear rotational spring is proposed to simulate the global behavior of the assembly. To implement spring parameters and find a method to study the local behavior of this assembly, double lap specimen are designed and tested on a quasi-static traction machine. The results show the important influence of the thermo stamping phase of the production on the behavior of the final specimen.

A numerical model is then proposed for these DLS, showing comparable results with experimental ones. Thanks to this model, the local behavior of a RVE, a pin stamped in thermoplastic composite has been extract. The flexural law obtained this way is then inserted in an analytical algorithm, which calculates the comportment law of the interface under shear loading. The whole interface can be then simulated with a simple nonlinear spring (whose parameters have been implemented from the previous algorithm) and inserted in another finite element model. The analytical algorithm works under any shear loading, and must be extended to torsional solicitations, so that it can be used to represent the triangle interface. These developments are still in progress.

This method shows very fast results for a given disposition of pin, shear loaded, and can be an important advantage for the pre dimensioning phases. The numerical model can be used if specific local post processing data are needed. This method represents a powerful tool for the comprehension and the development of the CMT pins assembly technology.

In future works an application to this method to a different texturing can be instigated. This method has also a high potential in optimization problems and can be used to find the ideal pin disposition answering interface specifications for both double lap and triangular specimen. Other kind of loading, as torsional or out of plane loading could be studied with a similar approach.

## ACKNOWLEDGEMENT

This work was performed in the framework of the LIMECO project funded by the IRT Jules Verne. The authors wish to thank all the member of this project: Faurecia for leading the project, Compose Tools for designing and realizing the mold of the triangle shaped specimen and finally the Cetim for the torsion tests realized on the triangle geometry.

## REFERENCES

[1] Camanho, P. P., & Matthews, F. L. (1997). Stress analysis and strength prediction of mechanically fastened joints in FRP: a review. *Compos Part A: Applied Science and Manufacturing*, 529-547.

[2] Dano, M. L., & al. (2000). Stress and failure analysis of mechanically fastened joints in composite laminates. *Composite Structure,* 287-296.

[3] Grujicic, M., & al. (2008). An overview of the polymer-to-metal direct-adhesion hybrid technologies for load-bearing automotive components. *Journal of materials processing technology I 97*, 363-373.

[4] Heimbs, S., & al. (2013). Failure behaviour of composite T-joints with novel metallic arrow-pin reinforcement. 16-28.

[5] Paulo, J., & al. (2004). Experimental study of drilling glass fiber reinforced plastics (gfrp) manufactured by hand lay-up. *Composites Science and Technology, Vol. 64(2)*, 289 – 297.

[6] Pinho, S. T., & al. (2006). Physically based failure models and criteria for laminated fibre-reinforced composites with emphasis on fibre kinking. Part II: FE implementation. *Composites: Part A 37*, 766 777.

[7] Sohier, L., & al. (2011). Etude expérimentale du comportement d'assemblages collés de composites sous. *Comptes Rendus des JNC 17*.

[8] Somoskői, G., & Török, I. (2013). CMT pin –Define the shape of the welded pin through welding parameters. *Production Processes and Systems, vol. 6.No. 1.*, 47-56.

[9] Ucsnik, S., & al. (2014). Composite to composite joint with lightweight metal reinforcement for enhanced damage tolerance. *ECCM16 - 16TH*.

[10] Ucsnik, S., & al. (2010). Experimental investigation of a novel hybrid metal-composite joining technology. *Composites*, 369-374.

[11] Xichang , W. (2011). Study of electron beam Surfi-sculpt during composite materials joining. *Rare metal materials and Engineering, vol. 40, suppl. 4,*, 292-296.

[12] Zhang, H. (2012). Study on the strength prediction model of Comeld composites joints. *Composites: Part B43*, 3310-3317.

[13] Zitoune, R., & Collombet, F. (2007). Numerical prediction of the thrust force responsible of delamination during the drilling of the long-fibre composite structures. *Composites Part A: Applied Science and Manufacturing, Vol. 38(3),*, 858-866.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A FAILURE CRITERIA FOR UNIDIRECTIONAL FIBER REINFORCED COMPOSITES BASED ON MICROMECHANICS BY ASYMPTOTIC HOMOGENIZATION

**Rafael Q. Macedo**[1], **José M. Guedes**[2], **Rafael T.L. Ferreira**[1*], **Maurício V. Donadon**[1]

[1]Instituto Tecnológico de Aeronáutica, DCTA
12228-900 São José dos Campos/SP Brasil
e-mail: {rquelho,rthiago,donadon}@ita.br

[2] Instituto Superior Técnico, ULisboa
1049-001 Lisbon, Portugal
e-mail: jmguedes@tecnico.ulisboa.pt

**Keywords:** Composites, Failure envelopes, Asymptotic homogenization.

**Abstract.** *This work presents a successful methodology for obtaining failure envelopes of unidirectional fiber reinforced composites based on micromechanical analysis by the asymptotic homogenization [1] method. Given a structure (in this case a composite material lamina) and external loads, plus having its material heterogeneity geometrically represented by a periodic unit cell of microstructure, the asymptotic homogenization method is able to predict its micromechanical stresses. Such stresses may be evaluated by failure criteria of the composite's constituents for several loading conditions, and this way it is possible to assess the composite's failure envelope. In the methodology developed, a periodic unit cell of the composite microstructure is isolated, consisting of a parallelepiped of polymeric matrix reinforced by cilindrical fibers oriented in one direction, and its behaviour is evaluated by an appropriate finite elements model. At first, the unit cell is tested in several directions to find strengths for the matrix and fiber and also matrix/fiber interface, thus evaluating failure characteristics of the composite constituents. The tests are carried out considering several possible orientations for the unit cell inside the related macroscopic media, which take into account the possible relative positions of the chosen unit cell inside the material heterogeneity of the composite. Then, the strengths of the constituents are used to predict the failure envelope for the fiber reinforced material, according to failure criteria devoted to the constituents. In this prediction, it is possible to say which is the constituent that fails first for each of the test loads. The results obtained are in good agreement with experimental data for carbon/epoxi and glass/epoxi composites. Moreover, the envelopes obtained are similar to the Puck & Schürmann [2] criterion, widely used to predict failure of such composites. This way, the present methodology renders good failure envelopes for fiber reinforced composites and gives information on the strengths of the constituents and material phase of failure, benefits from an incorporated micromechanical analysis.*

## 1 INTRODUCTION

Composite materials are made up of two or more materials (constituents) [3, 4] that have better properties than the constituents when alone [5]. The unidirectional fiber reinforced composite [3] is an example of such material that can achieve excellent mechanical properties having low mass. Because of that, they are heavily applied at the aeronautic industry [4]. The reinforcing phase are usually made of carbon or glass fibers and the base material is called matrix, usually made of epoxy [6].

It is very important in projects that use those materials to predict when failure is going to happen. A common procedure to investigate failure is to perform experiments. However, those are expensive, time consuming and it is not possible to investigate failure in all loading conditions. Then, with the increase of computer's capabilities, numerical methods arise as a possible manner to predict failure of composite materials.

Numerical methods can tell if failure is going to happen in a macromechanical or micromechanical analysis. The macromechanical analysis treats the composite structure as an orthotropic material and the failure criteria use the lamina strengths to determine failure [7]. The micromechanical analysis predict failure using the strengths of each constituent [8]. The micro analyses use stresses at each constituent to tell whether failure is going to happen and, because of that, different failure criteria can be considered for each constituent.

In this work, a micromechanical analysis of unidirectional fiber reinforced composites if performed using the asymptotic homogenization technique [1]. This method permits to calculate stresses at the micro level and then, failure can be predicted at each constituent. The methodology to determine failure presented here considers three regions: matrix, fiber and the interface between matrix and fiber.

## 2 ASYMPTOTIC HOMOGENIZATION

Asymptotic homogenization techniques allow to obtain micromechanics elastic properties of a composite considering its microstructure. In this technique, two levels are considered: the macro (**x**) and the micro (**y**). The micro level is considered to be periodic and it is represented by an unit cell formed by matrix and fiber. Its geometry can be changed in order to vary fiber volume fraction ($V_f$) and cross section shape. Inclusions and holes can also be inserted in it. In the present work, the hexagonal (Fig.1 a) and the square (Fig.1 b) unit cells are used to perform the homogenization.

### 2.1 Homogenized elastic properties

The homogenized elastic properties of the composite are obtained at micro scale by:

$$E^H_{ijkl} = \frac{1}{Y}\int_Y \left( E_{ijkl} - E_{ijkm}\frac{\partial \chi^{kl}_m}{\partial y_n} \right) dY. \tag{1}$$

In Eq.(1), tensorial notation is used, with the indexes vary from 1 to 3. $Y$ is the volume of the unit cell, $E_{ijkl}$ are the elastic properties of each point within the microstructure, $y_n$ are the coordinates in the unit cell, $\chi^{kl}_m$ are the auxiliar displacement fields that are obtained from the following equations,

$$\int_Y E_{ijmn}\frac{\partial \chi^{kl}_m}{\partial y_n}\frac{\partial \delta u_i}{\partial y_j} dY = \int_Y E_{ijop}\epsilon^{kl}_{op}\frac{\partial \delta u_i}{\partial y_j} dY. \tag{2}$$

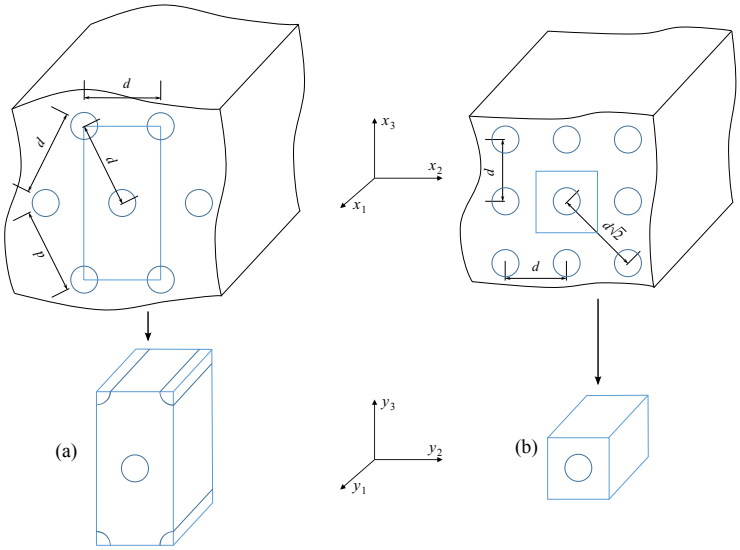



Figure 1: Periodical fiber distribution in the composite and the hexagonal unit cell.

In Eq.(2), $\epsilon_{op}^{kl}$ are unitary constant strain test fields and $\delta u_i$ are the virtual displacements. Equation (2) is here solved in the software PREMAT [1] using finite element method [9].

It can be noted in Fig.1 that the distances between each fiber in the hexagonal geometry are the same, but this not happen in the square unit cell. Because of that, transversal isotropy is expected for the perfect hexagonal unit cell [10].

### 2.2 Micro stresses

After computing the elastic properties, stresses at micro level can be calculated with the aid of the $\chi_m^{kl}$ fields. Equation (3) allows the determination of the stress tensor $\sigma_{ij}$ for each point of the unit cell:

$$\sigma_{ij} = \left( E_{ijkl} - E_{ijmn} \frac{\partial \chi_m^{kl}}{\partial y_n} \right) \frac{\partial u_k^0}{\partial x_l} = \left( E_{ijkl} - E_{ijmn} \frac{\partial \chi_m^{kl}}{\partial y_n} \right) \epsilon_{kl}^0. \tag{3}$$

In Eq.(3), $\epsilon_{kl}^0$ are the strains at macro level that are obtained from the average macroscopic displacement field $\mathbf{u}^0$ and $x_l$ are the coordinates at macro level. In our case the stress calculations are performed using the software POSTMAT [1].

## 3 FAILURE ANALYSIS

In this section, a methodology to determine failure using the asymptotic homogenization technique is presented. The failure methodology presented here is based on the work presented by [11] in which the unidirectional strengths of the composite are used to calculate the strengths of the constituents. At the composite micro level, three regions are considered in the analysis: matrix, fiber and the interface between them, and each of them has its own failure criterion.

### 3.1 Matrix failure criterion

Epoxy matrices are considered isotropic materials with different tensile and compressive strengths. There are several experiments [11] showing that the matrix failure depends on the deviatoric stress invariant $J_2$ and on the volumetric stress invariant $I_1$. A failure criterion based

on the relationship between those invariants is presented in [11]. It is known as the modified von Mises criterion for isotropic material that has different compressive and tensile strength, and is given as follows:

$$\left(\frac{\sigma_{VM}}{\sigma_{VM}^{cr}}\right)^{n_p} + \left(\frac{I_1}{I_1^{cr}}\right) = 1. \tag{4}$$

The mentioned invariants are calculated by:

$$I_1 = \sigma_{11}^m + \sigma_{22}^m + \sigma_{33}^m, \tag{5}$$

$$I_2 = -(\sigma_{11}^m\sigma_{22}^m + \sigma_{11}^m\sigma_{33}^m + \sigma_{33}^m\sigma_{22}^m) + (\sigma_{12}^m)^2 + (\sigma_{23}^m)^2 + (\sigma_{23}^m)^2, \tag{6}$$

$$J_2 = \frac{I_1^2}{3} + I_2. \tag{7}$$

In Eq.(5) and (6), $\sigma_{ij}^m$ are the components of the stress tensor of the matrix and $I_2$ is the second volumetric stress invariant. In Eq.(4), $n_p$ is a positive real number and the terms $\sigma_{VM}$, $\sigma_{VM}^{cr}$ and $I_1^{cr}$ are described below:

$$\sigma_{VM} = \sqrt{I_1^2 - 3I_2}, \tag{8}$$

$$\sigma_{VM}^{cr} = T_m\left(\frac{\alpha^{n_p} + \alpha}{\alpha + 1}\right)^{1/n_p}, \tag{9}$$

$$I_1^{cr} = T_m\left(\frac{\alpha^{n_p} + \alpha}{\alpha - 1}\right). \tag{10}$$

In Eq.(9) and Eq.(10) $T_m$ is the tensile strength of the matrix and $\alpha = C_m/T_m$, where $C_m$ is the compressive strength of the matrix.

### 3.2 Fiber failure criterion

Carbon and glass reinforcing fibers are considered to be transversally isotropic [11]. They have considerably higher elastic modulus and strength in the longitudinal direction than the matrix. In [11], the authors apply a quadratic failure criterion for fiber and show that terms regarding transverse strength can be eliminated from it. The result is a criterion that compares the normal stress in the fiber $\sigma_{11}^f$ to fiber longitudinal strengths:

$$-C_f < \sigma_{11}^f < T_f. \tag{11}$$

Where $C_f$ ($C_f > 0$) and $T_f$ are the compressive and tensile strength of the fiber in the longitudinal direction, respectively.

### 3.3 Interface failure criterion

The interface failure can be caused by debonding or detachment between fiber and matrix [11]. In this work, it is assumed that normal and shear stresses at that region are responsible for interface failure. The failure criterion is also presented in [11]:

$$\left(\frac{\langle t_n\rangle}{Y_n}\right)^2 + \left(\frac{t_s}{Y_s}\right)^2 = 1. \tag{12}$$

In Eq.(12) $Y_n$ and $Y_s$ are the normal and shear strengths of the interface, respectively, and $t_n$ and $t_s$ are the normal and shear stresses, respectively, at the interface; and the symbol $\langle \cdot \rangle$ is an operator that returns $0$ if the argument is negative. It means that only interface tractions cause failure.

### 3.4 Unit cell orientation

In this work the perfect hexagonal and square unit cells are used to sample the microstructure. It can be seen in Fig.1 that the fiber distribution using those cell within the composite has a constant pattern. However, in real unidirectional fiber reinforced composites the fiber distribution is rather random. Then, in order to account for this randomness, all possible unit cell orientations are considered as depicted in Fig.2.

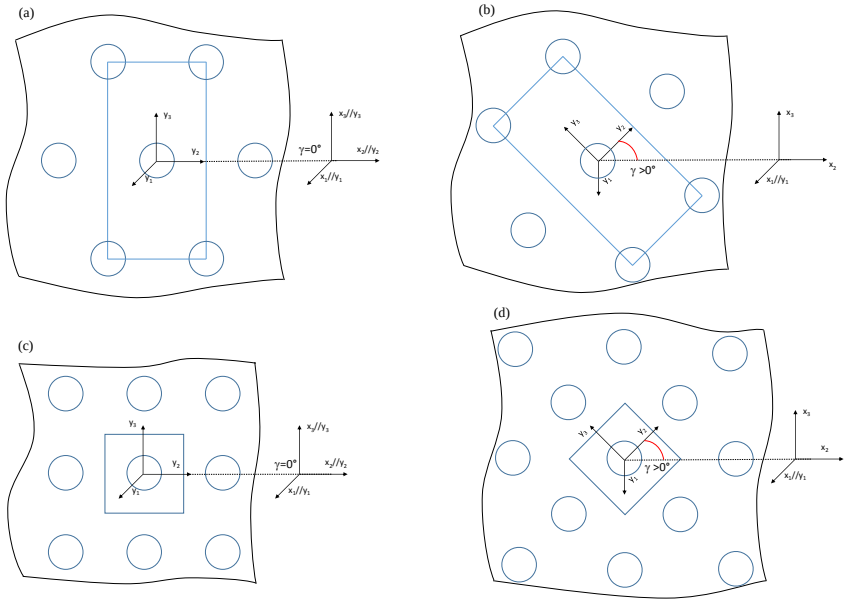



Figure 2: Possible orientations for the perfect hexagonal ((a) and (b)) and square unit cells ((c) and (d)).

Considering $\mathbf{x} = (x_1, x_2, x_3)$ the lamina coordinate system, the angle $\gamma$ accounts for the rotation of the unit cell around the axis $y_1$. All possible fiber arrangements are within the domain $\gamma \in [0, \pi]$ rad.

### 3.5 Strengths of the constituents

The unidirectional strengths of the composite are used to calculate the strengths of the constituents. The five unidirectional strengths of the composite are: $X_t$ and $X_c$ are the tensile and compressive strengths at the longitudinal direction, $Y_t$ and $Y_c$ are the tensile and compressive strengths at the transversal direction and $S_{12}$ is the in-plane shear strength. Those are considered as loads at the macro level and the micro stresses due to each of the macro loads are calculated for all possible unit cell orientations.

Three graphics can be plotted using the micro stresses, each of them representing a region of the microstructure: $\sigma_{11}^f$ is plotted in the fiber graphic, $\sigma_{VM}$ x $I_1$ in the matrix graphic and $t_s$ x $t_n$ in the interface graphic. Those are exemplified in Fig.3.

Those graphics combined with the failure criteria of Eq.(4), Eq.(11) and Eq.(12) are utilized to calculate the strengths of the constituents.

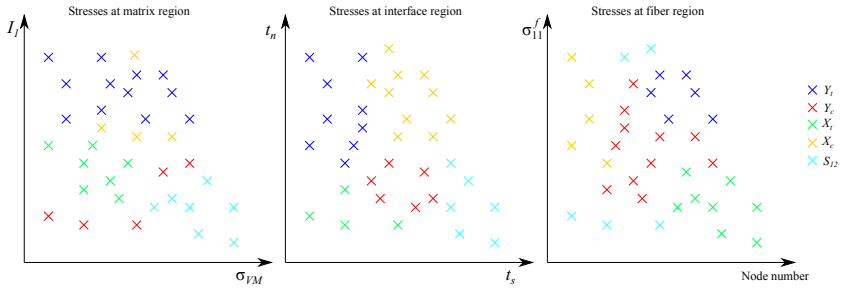



Figure 3: Stresses cloud of points of each region of the micro level.

For matrix, an optimization process is used to determine $T_m$, $C_m$ and $n_p$, where the quadratic distance between the failure criterion and the micro stresses are minimized. For the fiber and interface, the strengths $C_f$, $T_f$, $t_n$ and $t_s$ are calculated in order to guarantee that the unidirectional strengths of the composite are exactly predicted.

Finally, with the strengths of the constituents, the failure envelope can be built.

## 4 Results

In this section, the failure methodology is employed to an AS4/3501-6 carbon epoxy composite and the numerical results are compared to experimental data [7].

The elastic properties of the constituents are shown in Tab.1.

| Properties/Material | AS4 Carbon | 3501-6 epoxy |
|---|---|---|
| $E_1$ (GPa) | 225 | 4.2 |
| $E_2 = E_3$ (GPa) | 15 | 4.2 |
| $G_{12} = G_{13}$ (GPa) | 15 | 1.567 |
| $G_{23}$ (GPa) | 7 | 1.567 |
| $\nu_{12} = \nu_{13}$ | 0.2 | 0.34 |
| $\nu_{23}$ | 0.2 | 0.34 |

Table 1: Elastic properties of the constituents. Source:[12]

The elastic properties of the constituents are used as inputs to the homogenization problem. The homogenized elastic properties of the composite are calculated by Eq.(1). The fiber volume fraction used in the failure analysis was $V_f = 55\%$ due to this volume fraction yielded the least sum of error between experimental and numerical elastic properties when the perfect hexagonal unit cell was used, as can seen in Tab.2.

Now, the unidirectional strengths of the composite depicted in Tab.3 are applied as macro loads.

Applying those strengths as macro loads the micro stresses are calculated for each region of the microstructure and graphics can be built with those stresses. Those in conjunction with the failure criterion equation are used to determine the strengths of the constituents. The calculated strengths of the constituents are shown in Tab.4.

With the strengths of the constituents, the failure criterion equations of each region are defined. Figures 4 and 5 show the graphics for the perfect hexagonal and square unit cell plotted with the failure criterion equations.

| | Experimental data | Asymptotic homogenization | | | |
|---|---|---|---|---|---|
| Elastic properties | $V_f = 60\%$ | $V_f = 60\%$ | $\Delta$ (%) | $V_f = 56\%$ | $\Delta$ (%) |
| $E_1$ (GPa) | 126.000 | 136.060 | 8.0 | 127.230 | 1.0 |
| $E_2$(GPa) | 11.000 | 8.856 | 19.5 | 8.437 | 23.3 |
| $E_3$ (GPa) | - | 8.841 | - | 8.425 | - |
| $G_{12}$ (GPa) | 6.600 | 4.544 | 31.2 | 4.175 | 36.7 |
| $G_{23}$ (GPa) | - | 3.386 | - | 3.186 | - |
| $G_{13}$ (GPa) | - | 4.523 | - | 4.161 | - |
| $\nu_{23}$ | 0.400 | 0.361 | 9.8 | 0.374 | 6.5 |
| $\nu_{13}$ | - | 0.252 | - | 0.257 | - |
| $\nu_{12}$ | 0.280 | 0.251 | 10.4 | 0.257 | 8.2 |
| | | | $\overline{\Delta} = 15.8\%$ | | $\overline{\Delta} = 15.1\%$ |

Table 2: Experimental elastic properties of the AS4/3501-6 carbon epoxy composite [12] and from asymptotic homogenization using perfect hexagonal unit cell. The difference between experimental data and homogenized is $\Delta = 100\frac{|Exp-AH|}{Exp}$. The average error is $\overline{\Delta}$. The sign "-" means that experimental data were not provided.

| Strength properties | AS4/3501-6 carbon epoxy |
|---|---|
| $Y_t$ (MPa) | $60.2^{(b)}$ |
| $Y_c$ (MPa) | $273.3^{(b)}$ |
| $X_t$ (MPa) | $1950^{(a)}$ |
| $X_c$ (MPa) | $1480^{(a)}$ |
| $S_{12}$ (MPa) | $73.4^{(b)}$ |

Table 3: Unidirectional strength properties of AS4/3501-6 carbon epoxy. Sources: (a) - [12]; (b) - [7].

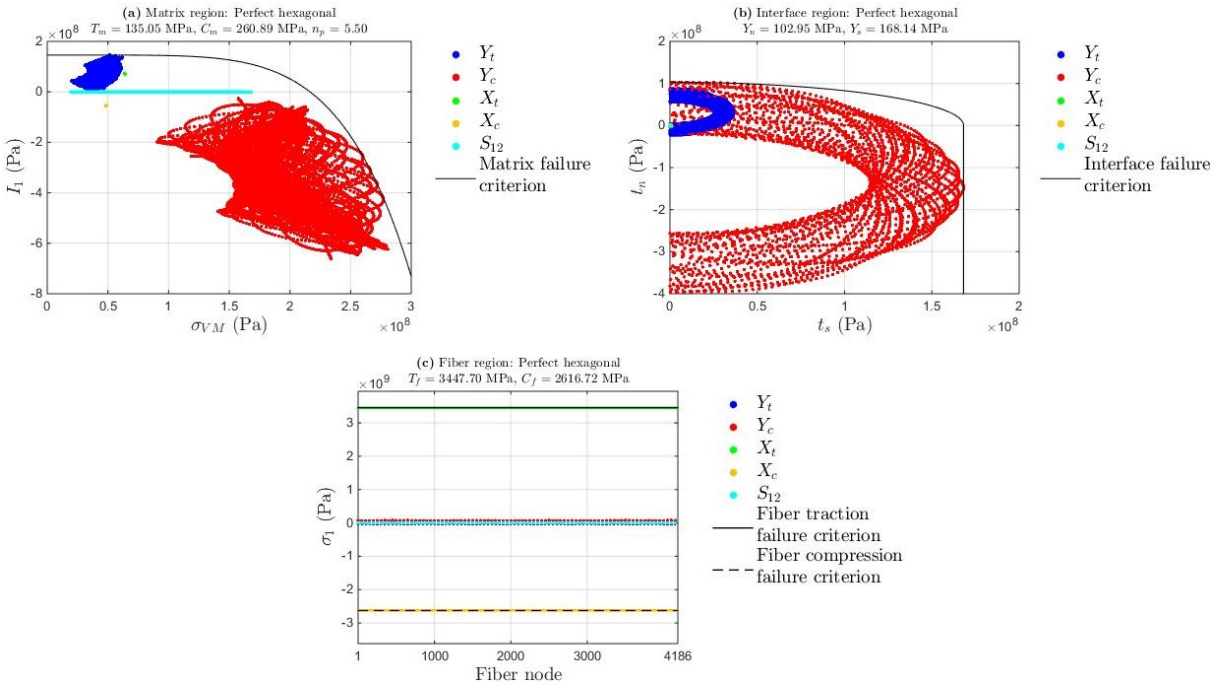



Figure 4: AS4/3501-6 carbon epoxy cloud of points and the failure criteria for each region using the perfect hexagonal unit cell.

Finally, the numerical failure envelopes can be constructed and be compared to experimental data. In addition to the asymptotic homogenization, the Puck & Schürmann [2] failure envelope is also plotted. The parameters that feed this criterion are $m_f = 1.1$ and $p_{\perp\parallel}^{(-)} = 0.3$. Figures 6

| Material | Strength parameter | Perfect hexagonal | Square |
|---|---|---|---|
| Matrix | $T_m$ (MPa) | 135.05 | 130.84 |
| | $C_m$ (MPa) | 260.89 | 270.33 |
| | $n_p$ | 5.50 | 3.66 |
| Interface | $Y_n$ (MPa) | 102.95 | 84.17 |
| | $Y_s$ (MPa) | 168.14 | 168.53 |
| Fiber | $T_f$ (MPa) | 3447.70 | 3434.07 |
| | $C_f$ (MPa) | 2616.72 | 2606.37 |

Table 4: Strengths of the constituents and parameters for the perfect hexagonal and square unit cells of the AS4/3501-6 carbon epoxy.

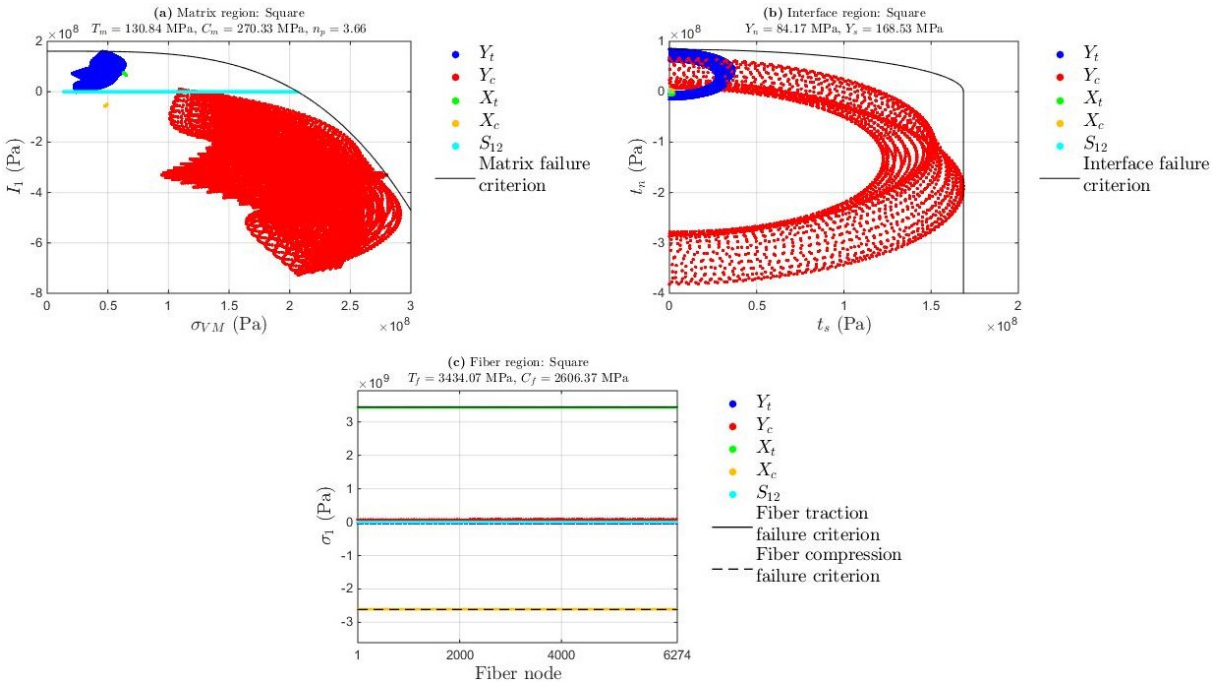



Figure 5: AS4/3501-6 carbon epoxy cloud of points and the failure criteria for each region using the square unit cell.

to 8 show the failure envelopes calculated.

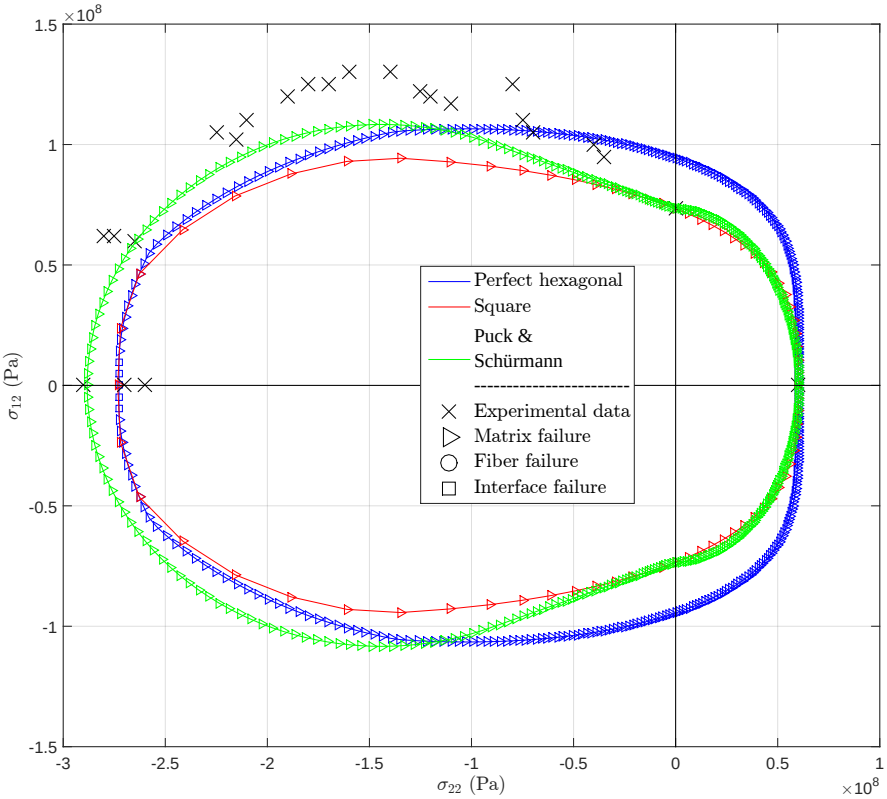



Figure 6: AS4/3501-6 carbon epoxy failure envelope at $\sigma_{22}$ x $\sigma_{12}$.

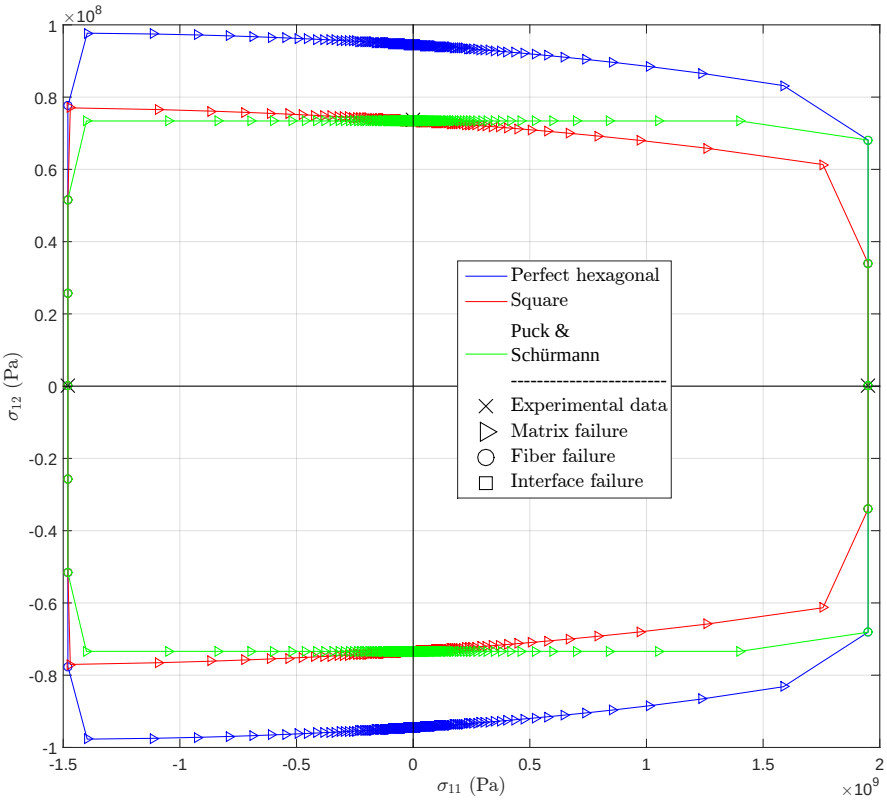



Figure 7: AS4/3501-6 carbon epoxy failure envelope at $\sigma_{11}$ x $\sigma_{12}$.

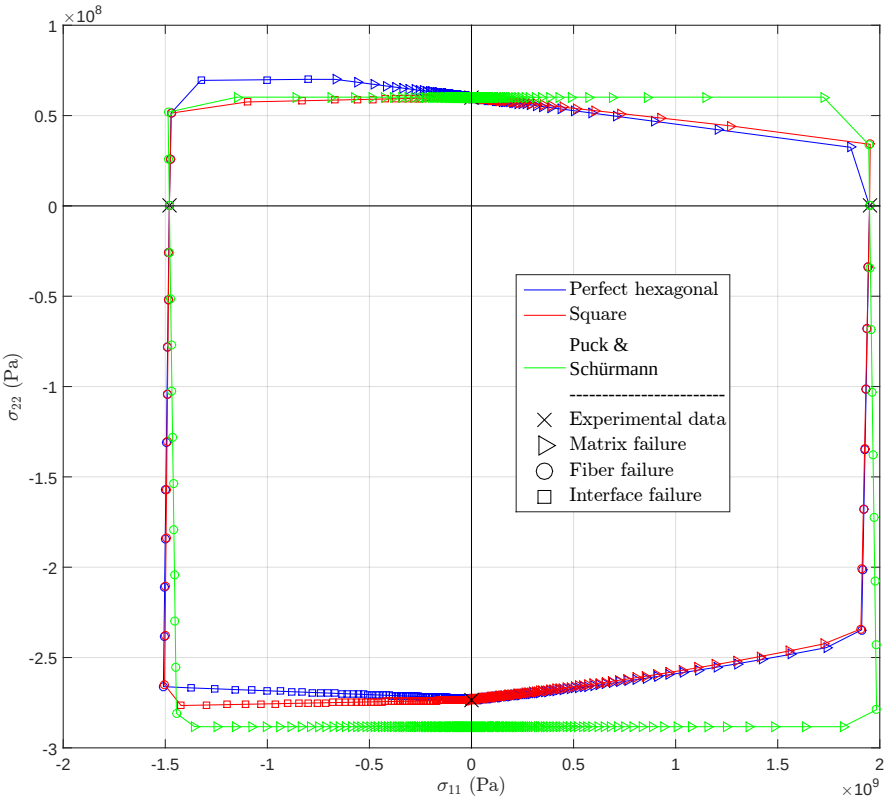



Figure 8: AS4/3501-6 carbon epoxy failure envelope at $\sigma_{11}$ x $\sigma_{22}$.

## 5 CONCLUSIONS

This work proposed a methodology to predict failure of unidirectional fiber reinforced composite using the asymptotic homogenization technique. It can be seen that the failure envelopes calculated using the methodology predict well the experimental failure points. It can also be noted that the numerical failure envelopes calculated are very similar to the Puck & Schürmann failure envelope in some regions of the graphics. The proposed methodology revealed to be a good method to predict failure of unidirectional fiber reinforced composites using the unidirectional strengths of this material.

The differences between numerical and experimental failure envelopes are caused by simplifications on the mathematical model, which are:

- The interface between fiber and matrix is considered perfect, this is: the displacements of that region are the same for matrix and fiber. In a real composite, interfacial imperfections may exist, changing the composite behaviour;
- The fiber alignment is also perfect, all fibers are parallel due to the periodicity, which do not happen in the real material;
- There is no fiber waviness, a common characteristic in real composites;
- The fibers, because of the periodicity and the unit cell model, are perfectly distributed inside the macrostructure, this is: the fiber volume fraction is constant within the microstructure. Random distributions are seen in a manufactured composite.

## ACKNOWLEDGEMENTS

The authors acknowledge the financial support received for this work from CAPES, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), process no.2014/00929-2, 2014/26377-6, and also from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), with an AVG grant.

## REFERENCES

[1] J.M. Guedes, N. Kikuchi, Preprocessing and postprocessing for materials based on the homogenization method with adaptive finite element methods. *Computer methods in applied mechanics and engineering*, **83**, 143–198, 1990.

[2] A. Puck, H. Schürmann, Failure Analysis of FRP Laminates by means of Physically based phenomenological models, *Composite Science and Technology*, **58**, 1045–1067, 1998.

[3] E.J. Barbero, *Introduction to composite materials design, 2nd Edition*. CRC Press, Boca Raton, 2010.

[4] R.M. Jones, *Mechanics of composite materials, 2nd Edition*. Taylor & Francis, Philadelphia, 1999.

[5] A. Mortensen, *Concise encyclopedia of composite materials, 2nd Edition*. Elsevier, Amsterdam, 2007.

[6] A.K. Kaw, *Mechanics of composite materials*. CRC Press, 2005.

[7] I.M. Daniel, J.J. Luo, P.M. Schubel, B.T. Werner, Interfiber/interlaminar failure of composites under multi-axial states of stress. *Composites Science and Technology*, **69**, 764–771, 2009.

[8] I.M. Daniel, Failure of composite materials. *Strain*, **43**, 4–12, 2007.

[9] J.N. Reddy, *An introduction to the finite element method, 2nd Edition*. McGraw-Hill, New York, 1993.

[10] R. de Medeiros, M.E. Moreno, F.D. Marques, V. Tita, Effective properties evaluation for smart composite materials, *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, **34**, 362–370, 2012.

[11] S.K. Ha, K.K. Jin, Y. Huang, Micro-Mechanics of Failure (MMF) for Continuous Fiber Reinforced Composites, *Journal of Composite Materials*, **42**, 1873–1895, 2008.

[12] P.D. Soden, M.J. Hinton, A.S. Kaddour, Lamina properties, Lay-Up configurations and loadings for a range of Fiber-reinforced composite laminates, *Composites Science and Technology*, **58**, 1011–1022, 1998.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EVOLUTION EQUATIONS BASED APPROACH FOR MODELING FATIGUE IN AMORPHOUS GLASSY POLYMERS. ON THE INVESTIGATION OF FATIGUE DAMAGE DEVELOPMENT IN POLYCARBONATE

**Sami Holopainen**[1]

[1]Tampere University of Technology, Department of Mechanical Engineering and Industrial Systems
P.O. Box 589, FI-33101 Tampere, Finland
e-mail: sami.holopainen@tut.fi

**Keywords:** fatigue, damage, amorphous glassy polymers, endurance surface.

**Abstract.** *While the key damage processes for focusing on the macroscopic mechanical behavior of amorphous glassy polymers are already well identified, relevant deformation mechanisms for fatigue damage are not yet established. The underlying study was born from the wish to gain some better understanding of how said mechanisms contribute to the fatigue damage in amorphous glassy matrix. In order to investigate the issue, an approach suitable for modeling fatigue in amorphous glassy polymers is proposed. The approach is calibrated to data taken from isothermal fatigue tests on dumbbell shaped PC-specimens. To investigate fatigue versus inhomogeneous deformation behavior, the approach is implemented in a finite-element program.*

## 1 INTRODUCTION

Manufacturers of polymer materials are being interested in assurance of components' lifetime, especially when the lifetime cannot be easily inspected or may lead to a catastrophe in service, [30]. Examples of mechanical components that are manufactured from polymers and may experience fatigue during their service life are found e.g. from medical industry, automotive glazing, aeronautics, and armour, [37], [24]. The design of such constructions could benefit from capable models and the strong computational capability available nowadays.

The knowledge of the ultimate behavior of amorphous glassy polymers has been developed rapidly while only a little of this work has been devoted to fatigue. Specific features of fatigue failure development are found e.g. from [40], [3], and [31]. To explore materials' microstructural characteristics with regard to fracture toughness, a strong attention has been paid on the damage mechanisms ahead of the crack tip, cf. [39], [36], [7], and [16] to mention a few. Much research has also devoted to the investigation of fatigue crack propagation in polymer composites having strong directional mechanical properties, cf. e.g. [12], [20], [19]. However, those fracture mechanics approaches neglect the crack initiation stage which may cover over 90% of the total fatigue life of amorphous polymers, [18].

Fatigue failure of amorphous polymers in their glassy state (termed amorphous glassy polymers) is generally due to a two-step process. In the first, initiation step failure is typically attributed to deficiencies or impurities affecting significant stress concentrations which exceed the strength limits of the material, [12], [23]. Under repeated loadings, those defects can nucleate and grow during the service life even at stress levels well below the nominal yield strength, [7], [27]. This part of fatigue is influenced by the localized yield-like deformation process which provides fatigue crack initiation sites controlling fatigue life (number of cycles N to failure) and thus being of a specific interest in the applied fatigue stress S (S-N curve), [23], [22]. The second, propagation step is characterized by the growth of damage through the coalescence of micro-cracks and propagation of small cracks to form large cracks which ultimately cause component failure, [21], [22]. However, the duration of the initiation step is typically orders of magnitude greater than the propagation time and thus plays most important role on fatigue behavior, [18]. Based on this observation the influence of crack propagation in the material behavior is often omitted in the fatigue models.

Fatigue failure of polymers is generally due to either mechanically or thermally dominated mechanisms. Mechanical modes that occur relatively low stresses and frequencies are characterized by the two step process described above. The macroscopic mechanical behaviour under such conditions is primarily (visco)elastic being influenced by the defects or inhomogeneities in material, [3], [17]. The fatigue life is relatively long while the ultimate fracture mechanism is rather brittle and influenced by combined mechanisms of the plastic dilatation and coalescence of inclusions or voids, [23], [28]. In thermally dominated mechanisms, at relatively high stresses and frequencies, the physical and mechanical properties of the polymer change due to hysteretic heating. This characteristic results from a high internal damping and low thermal conductivity of polymers when heat generated from mechanical fatigue cannot be dissipated to the surroundings, [41], [40], [18]. The time to thermally dominated failure is often rather short and the failure mode ductile, [4], [3], [17].

From a mesoscopic point of view, shear yielding and crazing, subsequent crack initiation and propagation are assumed to be major mechanisms for fatigue development in amorphous polymers. However, through existing literature of the field, key morphological or microstructural mechanisms that could explain their origin and subsequent progress of fatigue damage both in

homogenous and in toughened amorphous polymers are unclear at this moment.

The article continues by introducing a novel model suitable for predicting fatigue in amorphous glassy polymers. The governing constitutive model employed is an extension of the celebrated Boyce-Park-Argon (BPA) model, cf. [5], for predicting inhomogeneous plastic deformation in glassy polymers. Thereafter, Section 2.1 describes the proposed fatigue model and its numerical treatment in detail. Modeling of fatigue behavior per se is based on an appealing model introduced in [33], which model is formulated in continuum mechanics framework by using evolution equations that make the definition of damage changes per cycle redundant, i.e. cycle-counting techniques do not need to be applied. The approach is calibrated to data taken from both cold drawing experiments and isothermal fatigue tests on dumbbell shaped PC-specimens. Sections 3 and 4 are dedicated to the evaluation of the proposed approach through tangible examples where the model results are compared with experimental observations.

## 2 The model

The governing constitutive model employed here has been introduced in a previous study [13]. The model is a three dimensional extension of the BPA model, [5]. The extended model is aimed at long-term investigations of the mechanical behavior of amorphous glassy polymers under repeated loadings. Since amorphous polymers show a notable time dependent behavior under loading cycles, i.e. the polymer chains need a relaxation time to attain their equilibrium state after deformation, both viscoelastic and viscoplastic ingredients need to be included in the model. A more detailed account of the applied constitutive model and its numerical treatment is given in [13] and [14].

### 2.1 Fatigue model

When dealing with fatigue under variable complex loadings, a suitable damage rule constitutes an integral part of the analysis. The stress approach, which has been reported to be suitable for the modeling of mechanically dominated, relatively brittle and high-cycle fatigue, is considered as a basis of the proposed fatigue model. Majority of the those approaches represents fatigue-limit criteria in which the fatigue limits are described under infinite number of identical cycles, [34], [26]. For finite life predictions, however, those models are equipped with cumulative damage theories, which describe the damage increase per cycle and thus require that the loading consists of well-defined cycles, [29], [34], [26]. To define equivalent, representative cycles for load histories, cycle-counting methods need to be applied, cf. [9]. However, it is often challenging to extract equivalent cycles from complex load spectrum, which characteristic makes the cycle-counting approaches difficult for demanding applications in practice. Another way is to formulate the fatigue damage model within continuum damage mechanics (CDM) framework without need to measure damage changes per loading cycles, [32], [35], [19], [22]. However, those approaches as such are not suitable for modeling fatigue in amorphous polymers.

An appealing model suitable to describe a long-term fatigue failure behavior was proposed by Ottosen et al. [33]. According to this model, uniaxial and multiaxial stress states are treated in a unified manner for arbitrary loading histories, thus avoiding cycle-counting methods. Exploiting this evolution equation based fatigue modeling concept, a model for predicting the fatigue life of amorphous glassy polymers is proposed. The model uses only few macroscopical quantities and a single parameter set, which property makes the model simple and suitable for applications in practice.

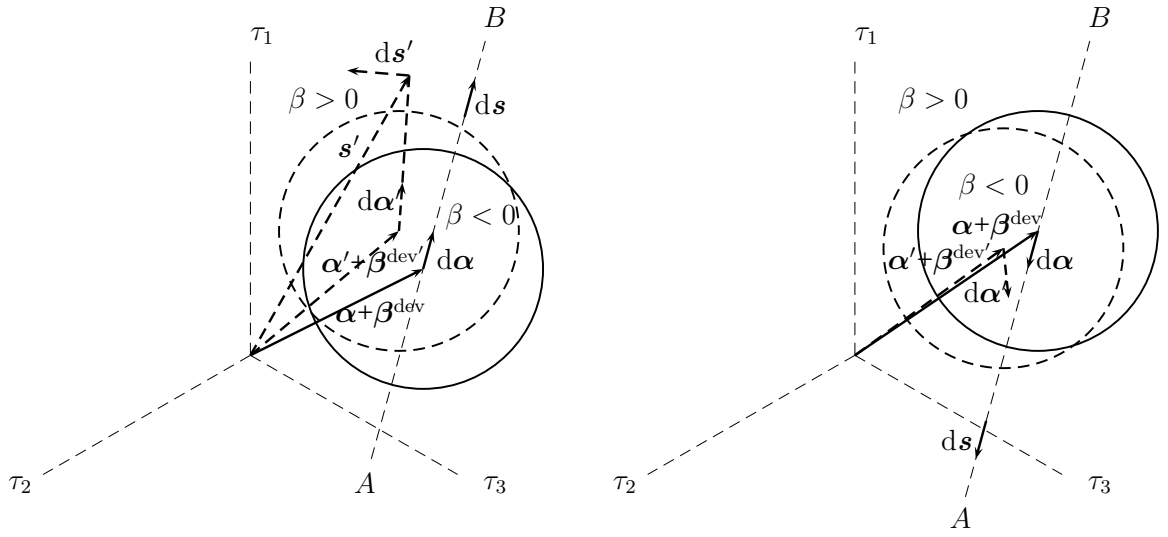



Figure 1: Alternating uniaxial stress state. The endurance surface will track the stress point and then moves between the states $A$ and $B$ (generally not fixed). Peripheries of the endurance surfaces in the initial and final state are highlighted by the dashed and solid line, respectively.

Due to crazing and the existence of voids around the chain molecules and inclusions, the microscopic yield of amorphous glassy polymers may depend on hydrostatic pressure, [38], [8]. In that context, any flow rule where the volumetric deformation is suppressed is not solely a sufficient rule for modeling fatigue. In the proposed model, the plastic deformation and fatigue damage are defined by the two evolution equations, respectively. The evolution equation for the plastic deformation does not include volumetric effects while for the fatigue damage it does. The model for fatigue is based on the concept of a moving endurance surface in the stress space and on an evolving damage variable, [33]. For many amorphous glassy polymers such an endurance surface can be identified, i.e. the cyclic lifetime increases with a decreasing accumulation of applied stress suggesting well-defined plateau when ultimate failure can finally reached at finite numbers of cycles just above the endurance limit. Wöhler curves are commonly used to illustrate those characteristics and identify polymers' endurance limits.

The endurance surface is considered as a function of the stress history and it can move in the stress space. In contrast to the plasticity theories for metals where the endurance surface may lie inside the yield surface, the fatigue damage development in amorphous glassy polymers is always induced by the propagation of plastic deformation. Also, since the polymer chains start align with the loading direction already at relatively low stresses and plastic strains, polymer materials show an anisotropic response which is in the model described by the backstress, $\boldsymbol{\beta}$. Due to that reasoning, the backstress is included to the endurance function. The expression of the backstress is defined adopting the BPA model, [13]. To include also volumetric damage effects, use is made of the endurance surface as

$$\beta = \frac{1}{S_0}(\bar{\tau} + aI_1 - S_0) = 0 \tag{1}$$

where the effective stress $\bar{\tau}$ is defined in terms of the second invariant of the reduced deviatoric stress $\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha}$ as

$$\bar{\tau} = \sqrt{3J_2(\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha})} = \sqrt{\tfrac{3}{2}(\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha}) : (\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha})}. \tag{2}$$

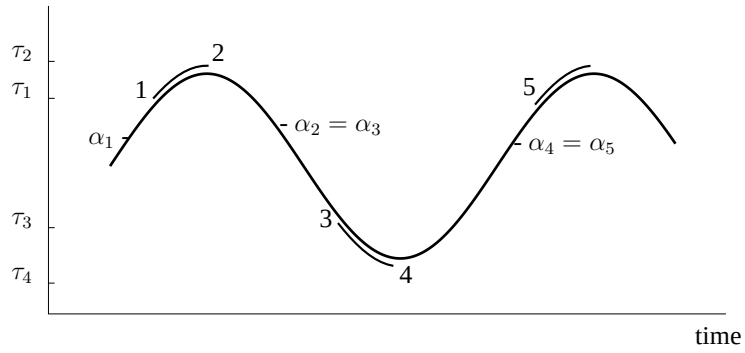



Figure 2: Alternating uniaxial loading. The damage development and movement of the endurance surface during cycling is indicated by a double curve.

In the equation (2), $\boldsymbol{s} := \boldsymbol{\tau}^{\mathrm{dev}}$ is the deviatoric part of the Kirchhoff stress, $\boldsymbol{\tau}$, of which the first stress invariant is given by $I_1 = \mathrm{tr}\,\boldsymbol{\tau}$, cf. [13]. $\boldsymbol{\alpha}$ is a fatigue backstress defined subsequently. The invariant $I_1$ reflects the effect of mean stress, i.e. the hydrostatic tension enhances the fatigue development while fatigue is suppressed under hydrostatic compression. The parameter $a$ in (1) is considered as positive and dimensionless which in uniaxial cyclic loadings determines the slope of the Haigh-diagram. The last parameter $S_0$ is the endurance limit for zero mean stress. Shape of the endurance surface in the deviatoric plane is circular as illustrated in Fig. 1.

When low-cycle fatigue is studied, the impact of the backstress $\boldsymbol{\beta}$ on fatigue is significant since its magnitude in relation to the stress is large. Then, the backstress is considered as a driving force for fatigue damage through the localized plastic deformation in the material. When reducing stresses and strains, the effect of the backstress decreases and another backstress quantity $\boldsymbol{\alpha}$ dominates the fatigue damage development. In high-cycle regime, the presence of $\boldsymbol{\alpha}$ in the fatigue model is mandatory in order to govern fatigue of virtually elastic responses (cf. e.g. steels).

The center point of the endurance surface is defined by the $\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{dev}}$ tensor as depicted in Fig. 1. Once an alternating loading is applied, the endurance surface will track the current stress point since the movement of $\boldsymbol{\alpha}$ always is in the direction of $\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha}$. It is the $\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{dev}}$ tensor which memorizes the load history and results in the movement of the endurance surface in the stress space. The evolution of $\boldsymbol{\alpha}$ is governed by a hardening rule similar to the Ziegler's kinematic hardening rule in plasticity theory, i.e.

$$\dot{\boldsymbol{\alpha}} = C(\boldsymbol{s} - \boldsymbol{\beta}^{\mathrm{dev}} - \boldsymbol{\alpha})\dot{\beta}, \tag{3}$$

where $C$ is a non-dimensional material parameter. The volumetric damage effects are included into (3) through the endurance surface (1). Since $\boldsymbol{\alpha}$ is considered an overall driving force for fatigue damage, it evolves only if the current stress state is outside the endurance surface, i.e.

$$\beta \geq 0, \quad \dot{\beta} \geq 0 \quad \Rightarrow \quad \dot{D} \geq 0, \quad \dot{\boldsymbol{\alpha}} \neq \boldsymbol{0}. \tag{4}$$

Referring to Fig. 2 for an alternating uniaxial loading, the concept is further demonstrated, see also [33] for a more detailed account. During loading from the state 1 to 2, the stress state lies outside the endurance surface and damage evolves, i.e. $\beta > 0$ and $\dot{\beta} > 0$. Between the states 2 and 3, the stress path has crossed the endurance surface and the stress state enters the space within the endurance surface. Then, $\beta > 0$ and $\dot{\beta} < 0$ when damage and the backstress do not evolve, i.e. $\dot{D} = 0$ and $\dot{\alpha} = 0$ until the stress path crosses again the current endurance surface at the state 3. It then follows that $\alpha_3 = \alpha_2$. From the state 3 to 4, damage again

evolves. In accordance with the stress path between the states 2 and 3, the damage development is inhibited until the state 5 is reached, i.e. $\alpha_5 = \alpha_4$.

### Damage evolution

Despite possible strain hardening and subsequent directional damage fields especially in large deformations, the fatigue behavior is described by a scalar valued quantity so as to keep the model simple. Assuming damage increases nonlinearly with the distance from the endurance surface, an exponential form

$$\dot{D} = K \exp(f(\beta; L_1, L_2, \vartheta))\dot{\beta}, \tag{5}$$

with the values $0 \leq D \leq 1$, is chosen for the damage evolution law. In (5), $K$, $L_1$, $L_2$, and $\vartheta$ are material parameters.

Many amorphous glassy polymers (such as PC) show only a moderate increase of the applied stress as the cyclic lifetime reduces. To capture such a behavior, a function $f$ having two linear asymptotes for positive $\beta$ is defined, i.e.

$$f(\beta; L_1, L_2, \vartheta) = L_1\beta - L_2 \left[\beta + \frac{L_2}{\vartheta}\left(\exp(-\vartheta\beta/L_2) - 1\right)\right], \tag{6}$$

which has the asymptote $L_1\beta$ when $\beta \to 0$ (HCF-regime) and $(L_1 - L_2)\beta$ when $\beta$ is large (LCF-regime). The curvature $\vartheta$ determines how rapidly the second asymptote is reached.

Since damage never decreases, it appears from (5) that $\dot{\beta} \geq 0$, i.e. damage rate increases with the distance from the endurance surface. Furthermore, damage develops only if stress states lie outside the endurance surface, i.e. $\beta > 0$ and the condition (4) is fulfilled.

## 3 Calibration of the model

The governing constitutive model was first calibrated to data obtained from cold drawing experiments on dumbbell shaped PC-specimens. Fatigue was omitted at this phase. A more detailed account for the test program involving repeated loading cycles is found from [13]. The calibrated parameters are listed in Table 1.

The parameters for modeling fatigue were determined from in situ measurements taken from [17] and [18]. Material which has been employed in the tensile fatigue test is a quenched PC (Lexan® 101R and 161R) and a specimen geometry used for fatigue studies includes a common dumbbell-shaped, injection molded tensile specimen (ASTM D638-IV), cf. [1], [2], and [10]. Isothermal test conditions for mechanically induced fatigue involve uniaxial stress submitted to load control at the room temperature of 23°C.

Table 1: Constitutive model parameters for PC. The calibration of the model is based on the the cold drawing experiments on a dumbbell shaped test specimens. The remaining viscoelastic constitutive parameters are $E = 2550$ MPa, $E_1 = 1295$ MPa, $\eta = 1.5 \cdot 10^5$ MPas, and $\nu = 0.37$.

| Parameter | $s_0$ | $s_{\mathrm{ss}}$ | $h_1$ | $\dot{\gamma}_0$ | $A$ | $C^{\mathrm{R}}$ | $N$ | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| Unit ..... | MPa | MPa | MPa | $s^{-1}$ | MPa$^{-1}K$ | MPa | | |
| Value ..... | 96 | 76 | 720 | $5.6 \cdot 10^{15}$ | 240 | 14 | 2.2 | 0.08 |

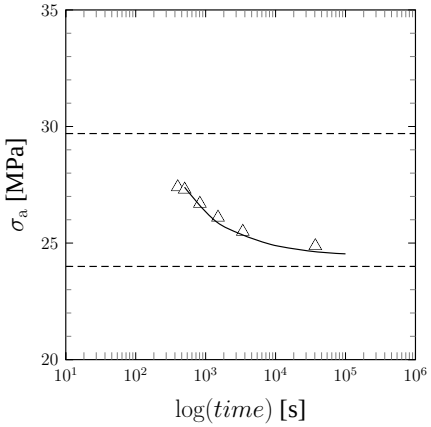



Figure 3: Fatigue strengths ($\sigma_{\mathrm{m}} = \sigma_{\mathrm{a}} + 2.2$ MPa) for PC employed (left). The solid line denotes the model result, and the marker $\triangle$ refers to data points taken from [18]. The upper and lower horizontal dashed line refer to the static tensile yield strength and an estimated endurance limit, respectively.

According to the test, the maximum stress level of a sine wave varies, while the minimum stress level is kept at 2.2 MPa. The frequency applied is 2 Hz, [18].

First, the slope $a$ of the Haigh-diagram was extracted from data given in [17] (Fig. 11a). The fatigue limit $S_0$ for zero mean stress and the remaining parameters $C$, $K$, $L_1$, $L_2$, and $\vartheta$ were calibrated to data shown in Fig. 3. The calibrated parameters are given in Table 2 and the model response is depicted in Fig. 3. The results indicate that the cyclic lifetime smoothly decreases with increasing accumulation of applied stress. Reducing the stress level, a transition in the failure mode occurs from ductile to brittle, i.e. ultimate failure can finally reached at finite numbers of cycles right above an expected endurance limit.

## 4 Results and discussion

### 4.1 Uniaxial stress state

Based on the calibrated parameters in Tables 1 and 2, the capability of the model to predict fatigue phenomena is discussed. The development of the movement of the endurance surface under a sinusoidal cyclic tension is demonstrated in Fig. 4. Right once the periodic loading is applied, the endurance surface reaches its periodic state similar to demonstrated in Fig. 1. During cycling, damage develops as the stress state is outside the endurance domain and moves away from it, i.e damage always increases when the endurance limit in Fig. 4 shows an increase and the stress is greater than the endurance limit. This situation was already illustrated in Fig. 2. Since the endurance limit in relation to the maximum stress is now low, damage increases rapidly leading to a short fatigue life as it is depicted in Fig. 5.

Table 2: Material parameters of the fatigue model for PC. The calibration is based on the cyclic tension experiments on a uniaxial tensile bar given in [17] and [18].

| Parameter | $S_0$ | $a$ | $C$ | $K \cdot 10^{-3}$ | $L_1$ | $L_2$ | $\vartheta$ |
|---|---|---|---|---|---|---|---|
| Unit ..... | MPa | | | | | | |
| Value ..... | 28.0 | 0.95 | 0.05 | 5.8 | 18.0 | 4.0 | 2.0 |

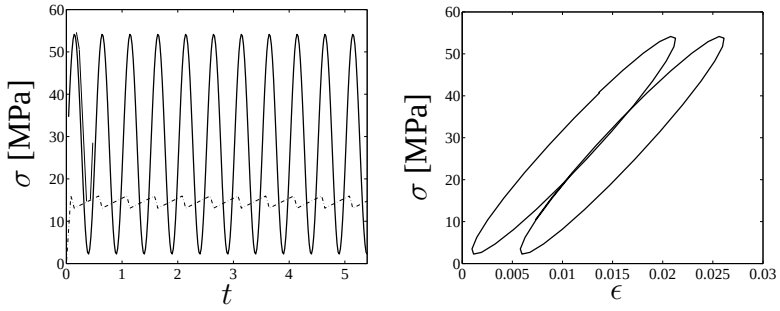



Figure 4: The cyclic true stress (solid thick line) and the periodic movement of the endurance limit (dash-and-dot line) during first few stress cycles (left). The damage development during the first cycle is indicated by the double curve. Model predictions for the second and last (prior to failure) hysteresis loops (right).

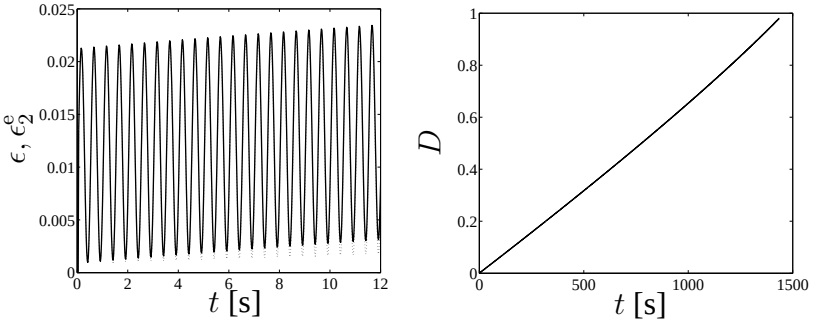



Figure 5: The periodic development of the true strain (solid line) and its viscoelastic component (dashed line) at the beginning of cycling tension. The corresponding cyclic stress is shown in Fig. 4 (left). Development of damage variable according to (5) (right).

Hysteresis loops of PC demonstrated in Fig. 4 show a constant area indicating a non-substantial energy-dissipation, i.e. hysteresis loops do not show a progressive increase in compliance and irreversible work during evolving fatigue, cf. [23], [18]. The result is a consequence of both neglected temperature effects and physical aging which is a relevant assumption under the applied low frequency loading. Fig. 5 further shows a periodically increasing development of both the true strain and its viscoelastic component at the beginning of loading. A small difference between the strains results from constantly increasing plastic strain. The growth of viscoelastic and plastic strains during the cyclic loading lead finally to a notable elongation composed of creep and plastic stretching. Considering a tensile test specimen, the elongation causes neck propagation followed by a brittle rupture as shown in [18].

Almost fully reversed uniaxial loading case (a low negative mean stress) was also studied, and the stress response as well as the corresponding hysteresis loops are shown in Fig. 6. In this situation, a reduction in the true strain is observed which feature is due to an accumulation of the viscoelastic and plastic strains. The development of the plastic strain shown in Fig. 7 refers to ratchetting, i.e. constantly accumulated plastic strain develops without bound as the cyclic loading continues, cf. [6], [21], [25]. Albeit not shown, similar phenomenon is observed in the load controlled cyclic tension, cf. Fig. 5 (left). Comparison between the first few loops and the loops prior to final failure (inset plot on the right) reveals that the area of the loops significantly increases during loading. Despite an accelerated propagation depicted in the inset plot (left),

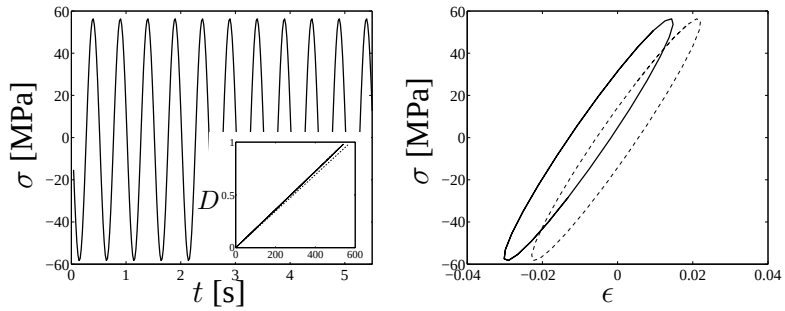



Figure 6: The cyclic true stress at the beginning of cycling as the frequency is 2 Hz and the maximum and minimum stresses are 56.4 MPa and -58.4 MPa, respectively (left). The inset plot shows damage increase without phase shift (solid line) and with phase shift $180°$ (dashed line). Model predictions for the second and last (prior to failure) hysteresis loops are highlighted by the dashed and solid line, respectively (right).

the final value prior to fatigue failure is still relatively low.

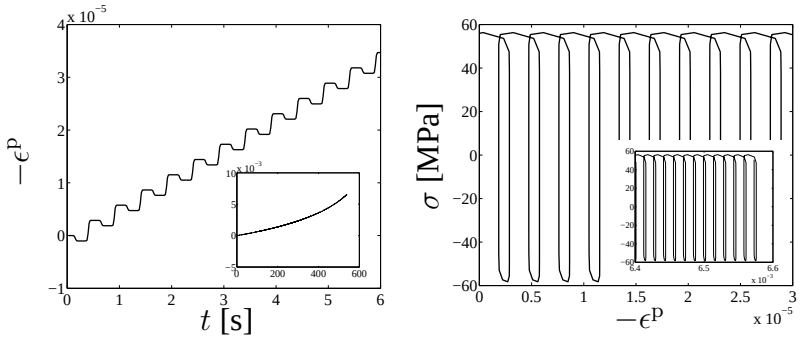



Figure 7: Periodic development of the plastic strain at the beginning of a cyclic loading (left). The inset plot shows an increase of the plastic strain up to the final failure. Stress vs plastic strain response at the beginning of loading and prior to failure (right).

## Fatigue of a dumbbell shaped PC-specimen

Due to the localization of the plastic deformation, damage grows unevenly in the material leading to a reduced fatigue life somewhere in the material. Thus, it is of interest to investigate the fatigue damage development of the entire test specimen by using a finite-element method. The specimen's geometry, loading conditions as well as details for the applied finite element mesh are found from [13].

Fig 8 shows that the damage development progresses most intensively in the gauge section of the specimen. This characteristic is because of an increasing localized plastic deformation and necking during drawing, cf. [13]. The predicted progress of fatigue damage is in line with experimentally observed macroscopic failure behavior of PC (at the same load) which shows a stable neck growth followed by a rapid rupture, cf. e.g. [18]. It is largely acknowledged that crazing is the governing micro-mechanism that triggers the fatigue damage at the sites following closely the localization of the plastic deformation in the amorphous glassy matrix, [11], [15], [7], [23].

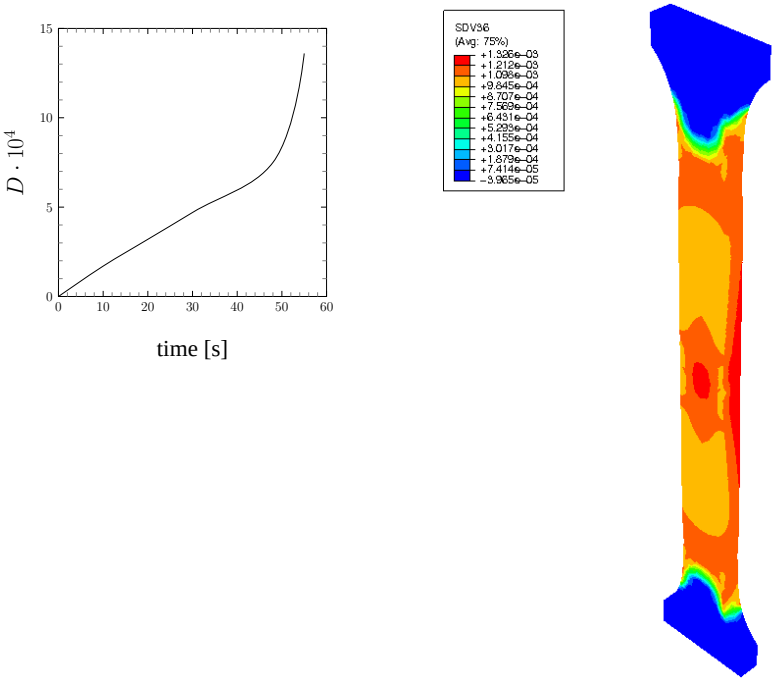



Figure 8: Damage development in a dumbbell shaped test specimen right after 100 cycles (right). The plot on the left shows the damage development in the middle of the specimen. A stable damage growth initiates after 50 seconds. The minimum and maximum pressure employed at the bottom edge are 1.1 MPa and 28.1 MPa, respectively. The corresponding average stress state in the gauge section is twofold.

A glance at Fig. 3 reveals that the fatigue life of the specimen under this loading should remain under 400 cycles. Considering the result in Fig. 8, solely the calculation of the solution for one hundred cycles took several hours. For this reason, when long term predictions are investigated, the fatigue life (e.g. for steels) is traditionally evaluated by using the location of most evolving fatigue observed at the beginning of the loading. However, due to a neck initiation, the damage response shows a nonlinear accumulation which feature makes a reliable prediction of the forthcoming damage and following fatigue life somewhat challenging. The evaluation of the fatigue life is computationally expensive since the analyzes must continue once a stable neck and damage growth are reached.

## Conclusions

The celebrated 8-chain BPA (Boyce-Parks-Argon) model was extended to cover fatigue damage behavior intrinsic to polymers of amorphous classes. When dealing with fatigue under variable complex loadings, a suitable damage rule constitutes an integral part of the analysis. A fatigue model proposed here is an extension of the appealing model given by Ottosen et al. [33], which model is formulated in continuum mechanics framework by using evolution equations that make the definition of damage changes per cycle redundant, i.e. cycle-counting techniques do not need to be applied. The model was calibrated to data taken from an accelerated uniaxial fatigue testing of PC with a high mean stress.

To investigate the propagation of fatigue damage under multiaxial cyclic loading conditions, the proposed approach was implemented in a finite element program. Finite element studies of a dumbbell-shaped test specimen were performed for analyzing the effect of plastic instabilities on fatigue damage. The proposed approach predicted the progress of fatigue damage

which was shown to be in line with experimentally observed macroscopic failure behavior of PC. The results indicated that localized yield-like deformation provides fatigue crack initiation sites which control the fatigue propagation and eventually the entire fatigue life of amorphous glassy polymers. When based solely on previous, short-term damage histories, the prediction of forthcoming damage development and the following fatigue life was shown to be challenging due to the plastic instabilities. The evaluation of the fatigue life was reliable only once a stable neck and damage growth were reached.

## References

[1] ASTM, 1985. ASTM D638, Standard for Uniaxial Tensile Testing. American Society for Testing and Materials, Philadelphia, Pa.

[2] ASTM, 1985. ASTM D671, Standard for Metric Practice. American Society for Testing and Materials, Philadelphia, Pa.

[3] Baltenneck, F., Trotignon, J.-P., Verdu, J., 1997. Kinetics of fatigue failure of polystyrene. Polym. Engngn. Sci. 37, 1740–47.

[4] Beardmore, P., Rabinowitz, S., 1975. Fatigue deformation in polymers. In: Plastic Deformation of Materials: Treatise on Materials Science and Technology; Arsenault, R. J., Ed. Academic Press: New York.

[5] Boyce, M. C., Parks, D. M., Argon, A. S., 1988. Large inelastic deformation of glassy polymers, Part I: Rate-dependent constitutive model. Mechanics of Materials 7, 15–33.

[6] Chaboche, J. L., 1991. On some modifications of kinematic hardening to improve the description of ratchetting effects. Int. J. Plasticity 7, 661–678.

[7] Drozdov, A. D., 2000. Nonlinear viscoelasticity and fatigue of glassy polymer. Mech. Res. Comm 27, 281–286.

[8] Estevez, R., Tijssens, M. G. A., Van der Giessen, E., 2000. Modeling of the competition between shear yielding and crazing in glassy polymers. J. Mech. Phys. Solids 48, 2585–2617.

[9] Fatemi, A., Yang, L., 1998. Cumulative fatigue damage and life prediction theories: a survey of the state of the art for homogeneous materials. International Journal of Fatigue 20, 9–34.

[10] Foster, A. M., May 2015. Materials testing standards for additive manufacturing of polymer materials: State of the art and standards applicability. Tech. rep., Materials and Structural Systems Division Engineering Laboratory, NISTIR 8059.

[11] Henkee, C. S., Kramer, E. J., 1984. Crazing and shear deformation in crosslinked polystyrene. Polymer 22, 721–737.

[12] Hertzberg, R. W., Manson, J. A., 1980. Fatigue of Engineering Plastics. Academic Press, New York.

[13] Holopainen, S., 2013. Modeling of the mechanical behavior of amorphous glassy polymers under variable loadings and comparison with state-of-the-art model predictions. Mechanics of Materials 66, 35–58.

[14] Holopainen, S., 2014. Influence of damage on inhomogeneous deformation behavior of amorphous glassy polymers. Modeling and algorithmic implementation in a finite element setting. Engng. Fract. Mech. 117, 28–50.

[15] Hristov, H. A., Yee, A. F., Gidley, D. W., 1994. Fatigue craze initiation in polycarbonate: study by transmission electron microscopy. Polymer 35, 3604–11.

[16] James, M. N., Christopher, C. J., Y. Lu, E. A. P., 2012. Fatigue crack growth and craze-induced crack tip shielding in polycarbonate. Polymer 53, 1558–70.

[17] Janssen, R. P. M., Govaert, L. E., Meijer, H. E. H., 2008. An analytical method to predict fatigue life of thermoplastics in uniaxial loading: Sensitivity to wave type, frequency, and stress amplitude. Macromolecules 41, 2531–40.

[18] Janssen, R. P. M., Kanter, D. K., Govaert, L. E., Meijer, H. E. H., 2008. Fatigue life predictions for glassy polymers: A constitutive approach. Macromolecules 41, 2520–30.

[19] Kawai, M., 2004. A phenomenological model for off-axis fatigue behavior of unidirectional polymer matrix composites under different stress ratios. Composites Part A: Applied Science and Manufacturing 35, 955–963.

[20] Kung, E., Mercer, C., Allameh, S., Popoola, O., Soboyejo, W., 2001. An investigation of fracture and fatigue in a metal/polymer composite. Metall Mater Trans A 32, 1997–2010.

[21] Lemaitre, J., Chaboche, J. L., 1994. Mechanics of Solid Materials. Cambridge University Press, Cambridge.

[22] Lemaitre, J., Desmorat, R., 2005. Engineering Damage Mechanics. Ductile, Creep, Fatigue and Brittle Failures. Springer-Verlag, Berlin.

[23] Lesser, A. J., 2002. Fatigue behavior of polymers. In: Encyclopedia of Polymer Science and Technology 6, 197–251.

[24] Lewis, P. R., 2009. Environmental stress cracking of polycarbonate catheter connectors. Engineering Failure Analysis 16, 1816–24.

[25] Liu, W., Gao, Z., Yue, Z., 2008. Steady ratcheting strains accumulation in varying temperature fatigue tests of PMMA. Materials Science and Engineering A 492, 102–109.

[26] Liu, Y., Mahadevan, S., 2007. A unified multiaxial fatigue damage model for isotropic and unisotropic materials. International Journal of Fatigue 29, 347–359.

[27] Lu, H., Kim, G., 2007. Characteristics of accelerated lifetime behavior of polycarbonate under athermal and high loading frequency conditions. Polymer Testing 26, 839–45.

[28] Luu, D., Maitournam, M., Nguyen, Q., April 2014. Formulation of gradient multiaxial fatigue criteria. International Journal of Fatigue 61, 170–183.

[29] Matake, T., 1977. An explanation on fatigue limit under combined stress. Bulletin of the Japan Society of Mechanical Engineers 20, 257–263.

[30] Maxwell, A. S., Broughton, W. R., Dean, G., Sims, G. D., 2005. Review of accelerated ageing methods and lifetime prediction techniques for polymeric materials. Tech. rep., National Physical Laboratory Hampton Road, Teddington, Middlesex, TW11 0LW.

[31] Melick, H. G. H. V., 2003. Deformation and failure of polymer glasses. Doctoral thesis. University Press Facilities, Eindhoven.

[32] Okui, Y., Horii, H., Akiyama, N., 1993. A continuum theory for solids containing microdefects. Int. J. Engng. Sci. 31, 735–749.

[33] Ottosen, N., Stenström, R., Ristinmaa, M., June 2008. Continuum approach to high-cycle fatigue modeling. International Journal of Fatigue 30 (6), 996–1006.

[34] Papadopoulos, I. V., 2001. Long life fatigue under multiaxial loading. International Journal of Fatigue 23 (10), 839 – 849.

[35] Peerlings, R., Brekelmans, W., de Borst, R., Geers, M., 2000. Gradient-enhanced damage modelling of high-cycle fatigue. International Journal for Numerical Methods in Engineering 49 (12), 1547–1569.

[36] Ritchie, R. O., 1999. Mechanisms of fatigue-crack propagation in ductile and brittle solids. Int. J. Fracture 100, 55–83.

[37] Shah, Q. H., Abakr, Y. A., 2008. Effect of distance from the support on the penetration mechanism of clamped circular polycarbonate armor plates. Journal of Impact Engineering 25, 1244–50.

[38] Steenbrink, A. C., Van der Giessen, E., 1998. Studies on the growth of voids in amorphous glassy polymers. Journal of Material Sciences 33, 3163–3175.

[39] Takemori, M. T., Kambour, R. P., Matsumoto, D. S., 1983. On a search for epsilon crack tip plastic zones in the fatigue of amorphous polymer. Polym. Comm 24, 297–299.

[40] Takemori, M. T., Matsumoto, D. S., 1985. Shear and craze development in the fatigue of ductile amorphous polymers. Journal of Material Sciences 20, 873–880.

[41] Tauchert, T. R., 1967. Int. J. Engng. Sci., 353–365.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ULTRASONIC GUIDED WAVES INSPECTION OF PIPES FROM ONE END TO THE FIRST BENT

**Mihai V. Predoi[1], Marian Soare[2], Ovidiu Vasile[1] and Mihail Boiangiu[1]**

[1] University Politehnica of Bucharest, Dept. of Mechanics, Splaiul Independentei 313, 060042, Bucharest, Romania
e-mail: predoi@cat.mec.pub.ro; ovidiu@icecon.ro; mboiangiu@gmail.com

[2] S.C. Nuclear NDT Research & Services S.R.L. Bucharest / ROMANIA, Berceni Road, No. 104 – Central Laboratory Building, Sector 4, Post Office No. 8, PO Box 137, Post Code 041919
soare.marian@nuclear-ndt.ro

**Keywords:** Guided waves, Ultrasonic inspection, Pipe bent.

**Abstract.** *Non-destructive inspection of long pipes can be time consuming and some localized defects might remain undetected. A new technique based on guided waves was developed in the last decade. High risk industries such as chemical or nuclear, require a reliable method covering if possible the whole length of the piping. Guided ultrasonic waves can propagate along several meters and up to tens of meters in pipes. The reflected signal, which is usually made of some of the propagating modes, brings information concerning the presence of defects and in some conditions on their location and size. There are three classes of guided modes which can propagate in a pipe: longitudinal, torsional and flexural modes. Even if many authors have investigated the guided waves in pipes, in the present research we are focusing on generating a high energy ultrasonic pulse, capable to produce detectable signals from small defects. Moreover we discuss the probability of detectability of defects near and especially on pipe bends. Using our in-house specialized software package to determine the dispersion curves and displacements fields, we determine the incident modes sent in a pipe from its free end, as in heat exchangers. These results are used for optimal numerical simulation, using the Finite Elements Method (FEM), in order to determine the defect detectability. The numerical data are compared with results obtained in laboratory experiments standard equipment for ultrasonic NDT.*

## 1 INTRODUCTION

Non-destructive inspection of long pipes, as is the case for the petro-chemical industry, can be time consuming and moreover, some localized small but dangerous defects might remain undetected. A new technique based on guided waves was developed in the last decade, as can be remarked in the works of Ditri [1], Cawley et al. [2] Lowe at al. [3], Li and Rose [4], just to mention a few of the most relevant. Guided ultrasonic waves can propagate along several meters for liquid filled or coated pipes and up to tens of meters in hollow uncoated pipes. The defect reflected signal, which is usually made of several propagating modes, brings information concerning the presence of defects and in some favorable conditions, on their location and size.

Guided waves have special propagation characteristics, as shown in classical textbooks of Auld [5], Miklowitz [6], Achenbach [7], Graff [8] or Rose [9]. There are three classes of guided modes which can propagate in a pipe: longitudinal, torsional and flexural modes as shown by Gazis theoretically in [10] and numerically in [11].

The longitudinal modes are axially symmetric and have dominant radial and axial displacements. These waves prove to be most sensitive to partly circumferential flaws. The waves are dispersive, meaning the wave phase and group velocities are strongly depending on the frequency. The torsional modes are also axially symmetric and less dispersive, especially the fundamental $SH_0$ mode which is non-dispersive at least on the straight part of a pipe. Since the radial displacements are negligible, the interaction with the surrounding various media is weak, thus reducing the wave attenuation. The interaction of these waves with dominantly axially oriented defects is more pronounced. The flexural modes which are no longer axially symmetric are in general highly dispersive and more attenuated, but these modes can also propagate in a pipe and must be taken into account.

Guided waves in pipes have been numerically investigated using several methods, for the two main aspects: dispersion curves and propagation-scattering simulation.

The dispersion equations can be solved using numerical root finding algorithms, limited to real solutions like Gazis [11] or Predoi et al. [12] or determining the complex solutions as done by the team of Lowe [13]. A better approach, providing the full spectrum of complex solutions, is to take advantage of the axial symmetry or circumferential periodicity and use the so-called Semi-Analytical Finite Elements Method (SAFEM). This method requires a finite elements mesh only along a radial segment limited by the pipe walls, as done by Predoi [14]. Another possibility is to mesh the entire cross-section of the waveguide, which can thus be used for any shape of the cross-section as shown by Hayashi at al. [15].

The propagation of acoustic ultrasonic waves can be numerically simulated by direct time marching solutions using finite differences as Gsell et al. [16], or by using the Finite Elements Method (FEM) for either the time-marching solution or the frequency domain equivalent problem. The time-marching solutions are resources consuming and thus limited to small frequency and propagation length. The frequency domain approach can be used for higher frequencies.

In the present research we are focusing on generating a high energy ultrasonic pulse, capable to produce detectable signals from small defects, at relatively high frequency (1MHz). Moreover we discuss the probability of detectability of defects near and especially on pipe bends, continuing previous researches concerning bents [17], by using FEM simulations [18] and optimal selection of modes [19].

Using our in-house specialized software packages “Tubewave” [12] for real wavenumbers and especially the SAFEM [14] for complex valued wavenumbers, we determine the dispersion curves and displacements fields, in a pipe. These results are used for optimal numerical

simulation, using the Finite Elements Method (FEM), in order to determine the defect detectability.

## 2 PROPAGATING LONGITUDINAL GUIDED WAVES

The detection of small localized defects can be done by using guided waves of short wavelengths, which implies high frequency. On the other hand, high energy flux is necessary in order to receive a detectable echo. For these reasons, in the following we focus on longitudinal axially symmetric guided waves, produced by longitudinal wave ultrasonic transducer, with central frequency of 1 MHz.

### 2.1 Dispersion curves

The geometry and displacement field notations are shown on Figure 1.

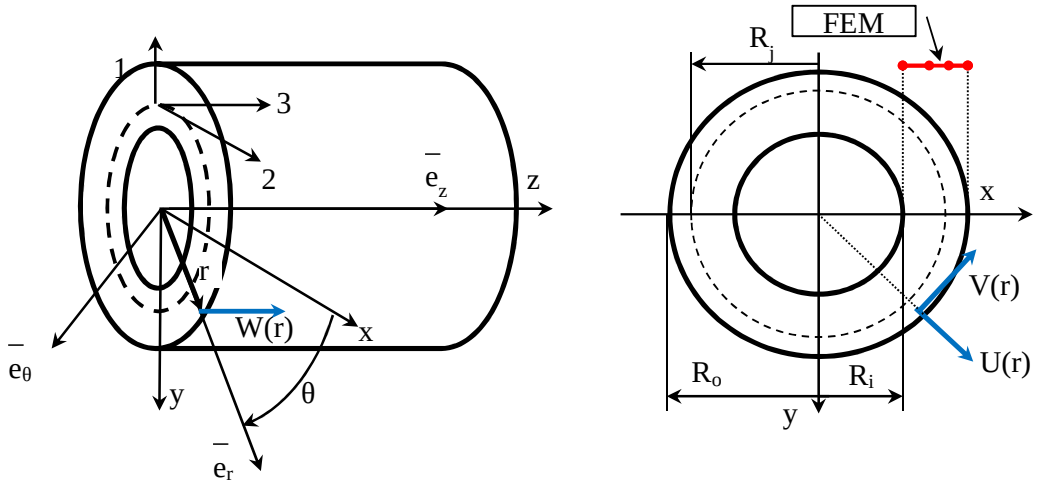



Figure 1 Pipe geometry and displacements field

Plane wave harmonic displacement field is assumed for the longitudinal waves, with *k* (1/m) the wavenumber and ω (rad/s) the angular frequency:

$$\begin{aligned} u_r(r,\bar{},z,t) &= U(r)\exp\left[i(kz-\mu t)\right] \\ u_\theta(r,\bar{},z,t) &= 0 \\ u_z(r,\bar{},z,t) &= W(r)\exp\left[i(kz-\mu t)\right] \end{aligned} \tag{1}$$

The differential equations of longitudinal plane wave propagation can be written as [14]:

$$\begin{aligned} &\frac{\partial}{\partial r}\left[rC_{11}U' + C_{12}(U + irkW)\right] - \left(C_{12}U' - irC_{66}kW'\right) - \left(\frac{C_{11}}{r}U + iC_{12}kW\right) - k^2rC_{66}U = -r\cdot\mu^2U \\ &C_{66}\frac{\partial}{\partial r}\left(rW' + irkU\right) + ikC_{12}\left(rU' + U\right) - k^2rC_{11}W = -r\cdot\mu^2W \end{aligned} \tag{2}$$

in which an isotropic homogeneous material is assumed ($C_{12} = C_{11} - 2C_{66}$), with $C_{11}$ and $C_{66}$ the Lame elasticity constants of the material, which can be also expressed in terms of Young's elasticity constant and Poisson's ratio. The pipe used in this example is a brass made pipe,

$R_o$=9mm, $R_i$=7.8 mm, mass density $\rho$=8400 kg/m$^3$; bulk longitudinal wave velocity $c_L$=4400 m/s and bulk transverse velocity $c_T$=2200 m/s, from which the Lame's constants can be easily obtained. The dispersion curves have been obtained using the algorithm detailed in [14], implemented in a FEM commercial package [20]. The variation of the wavenumbers k (labeled Re(kx)) on Figure 2 corresponds to known results, with a special remark for the L(0,5) mode which can propagate above 1650 kHz as "retro-propagative mode" up to 1800 kHz, beyond which it is propagating normally. A "retro-propagative mode" has phase and group velocities in opposite directions, which is a peculiar feature.

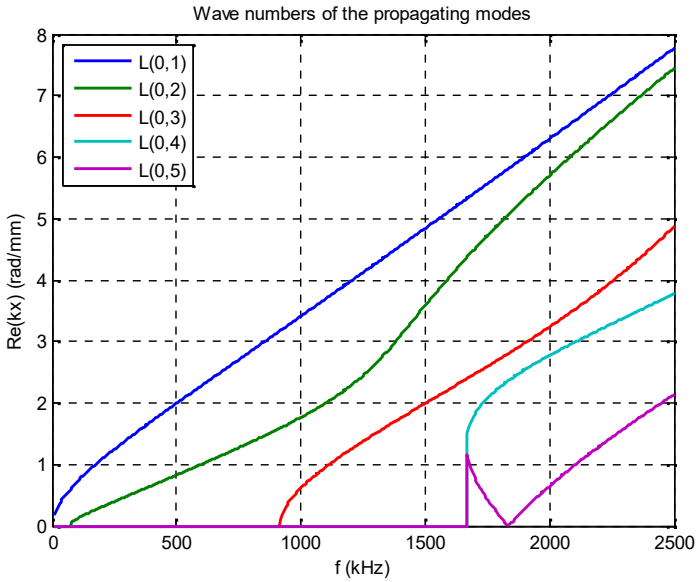



Figure 2 Dispersion curves: wavenumbers k vs. frequency for the first five propagating modes

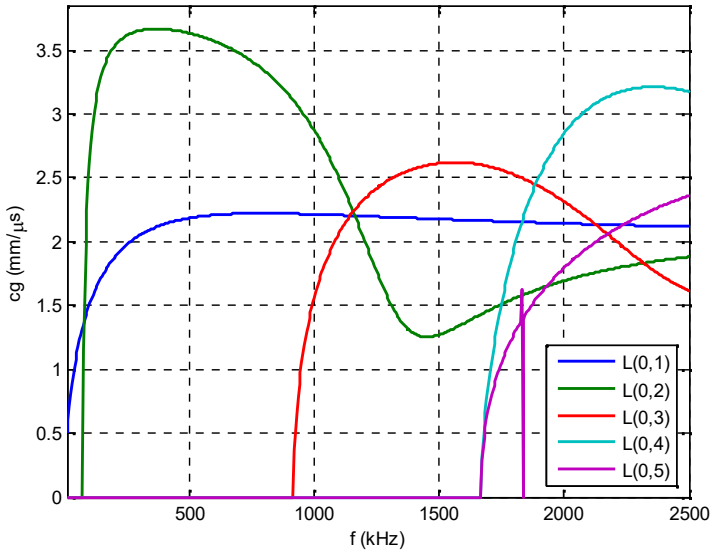



Figure 3 Dispersion curves: group velocities $c_g$ vs. frequency for the first five propagating modes

For experimental and FEM simulations purposes, it is more useful to plot the group velocities, with which travel the short acoustic pulses in the pipe (Figure 3).

## 2.2 Displacement fields

Frequency ranges for which there is a large variation of group velocity are not recommended, since the incident pulse energy will be spread into a long chirp, beginning with the fastest and ending with the slowest moving part of the mode. The effect is a long measured signal with low amplitude, which is difficult or even impossible to analyze. For this reason, most authors prefer to use lower frequencies (e.g. lower than 900 kHz in this case) for which L(0,2) mode is practically non-dispersive and the L(0,1) mode has a much higher group velocity and can be thus easily separated in the received signals. We suggest in this paper to use the frequency of 1200 kHz (or using a 1 MHz commercial transducer) because:

- higher frequency increases the defect detectability
- at this particular frequency all three propagating modes have almost the same group velocity, concentrating the acoustic energy in a single signal burst
- defect sensitivity is increased near the free inner and outer surfaces, where higher displacements amplitudes exist and higher energy flux is concentrated.

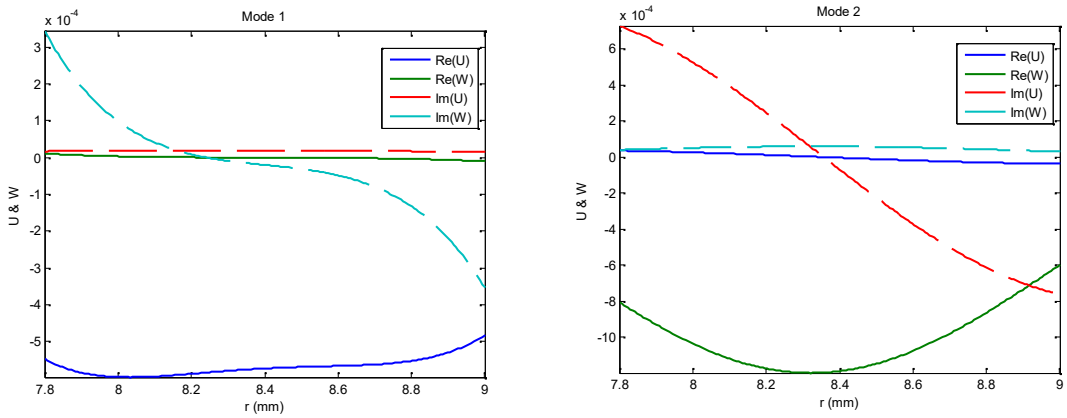



Figure 4 Displacement fields for the first two modes L(0,1) and L(0,2)

At 1MHz, the displacement fields for the first two modes L(0,1) having k=3410.8 rad/m and L(0,2) having 1764.6 rad/s are plotted on Figure 4. The existence of real and imaginary parts of the displacements is related to the phase of these displacements. The L(0,2) mode has a dominant real radial negative displacement (U) almost constant along the cross-section, with a very small imaginary part. The axial displacement (W) has an imaginary part of large values at the free surfaces and of opposite signs on these surfaces. Here, the phase lag between the radial and axial displacement is $\pi/2$, indicated by the real/imaginary significant components. The L(0,2) mode has a constant sign (negative on this plot) for the real part of the axial displacement (Re(W)), but the radial displacement changes sign along the cross-section (Im(U)) and is phase lagged by $\pi/2$. The L(0,3) mode of k=616.24 rad/s, is not shown here, being less important in the next analysis.

## 3 FEM MODEL

Since the investigation is done for multimode high frequency, the frequency domain analysis was selected. Even so, the 32 GB RAM memory (100 GB virtual memory) imposed limitations on the number of hexahedral elements 99456, which correspond to the requirement of 10 elements per shortest wavelength and a minimum number of 3 elements per pipe thickness in case of slowly radial varying displacements fields. The last condition is not correctly fulfilled since the displacements fields have rapid variations along the radius (Figure 4), but a compromise had to be made. A mesh detail is shown on Figure 5 for a 90° bent.

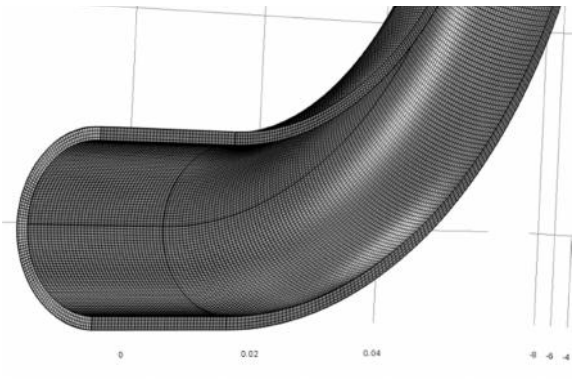


Figure 5 Mesh used for the 90° bent

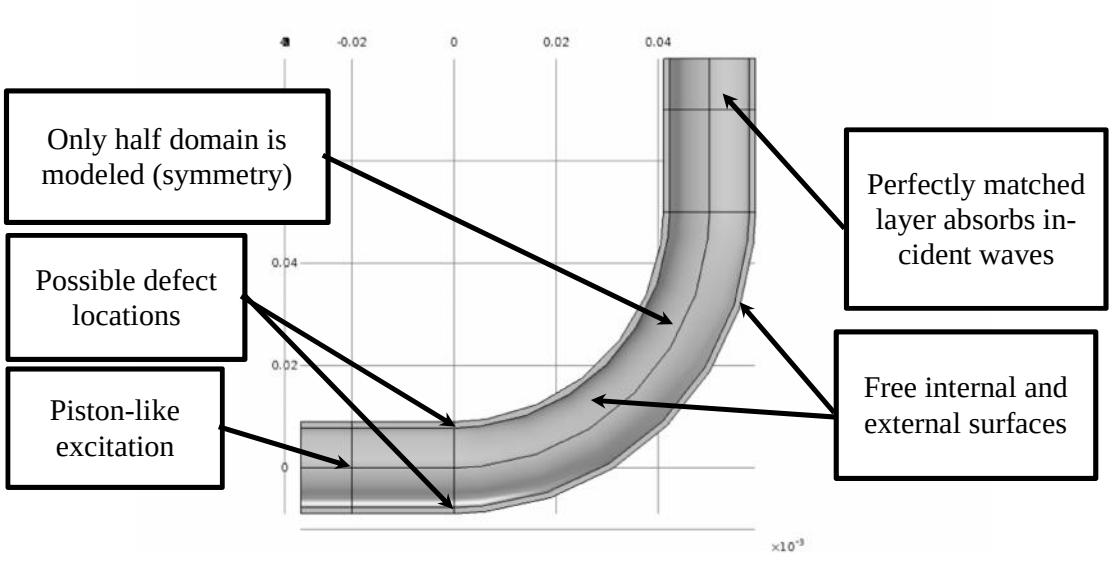



Figure 6 FEM domain and boundary conditions

The boundary conditions require special attention in frequency domain simulations. The equivalent in the time domain is a permanent harmonic excitation a single frequency. For this reason, spurious reflections from FEM domain boundaries have to be eliminated. The FEM software used in this case [20], offers two options: radiation condition and Perfectly Matched Layer (PML). The radiation condition is perfect if the incident wave is perpendicular on the

radiating surface, but this not the case in the bent pipe. Consequently, two PML domains have been attached at the two ends of the pipe FEM model. Even so, the PML is not a perfect solution since several modes can propagate in the pipe and each of them has to be absorbed in the PML. A very rapid attenuation represents a rapid variation in acoustic impedance, so reflections still occur. This is particularly important for waves of large wavelengths, for which the PML must cover more than a wavelength. Again a compromise had to be made, since these PML mean a supplementary number of finite elements, to the already large number of elements in the domain of wave propagation. The optimal parameters for the PML have been selected from a simpler model of a short straight pipe, for which the solution time is less than 2 min. The displacement field attenuation to negligible values at the exit from the PML was the selection criterion.

The excitation is done by a piston-like source placed between the first PML and the FEM physical domain, at a frequency of 1 MHz of pressure intensity 10 kPa. The rest of the surfaces are stress-free. Three simulations have been selected for the analysis: the 90° bent without defects, and with half-through flat bottomed holes on the pipe exterior, one on the inside (Figure 7a) and the other on the outside of the bent (Figure 7b), respecting the cross-section plane symmetry.

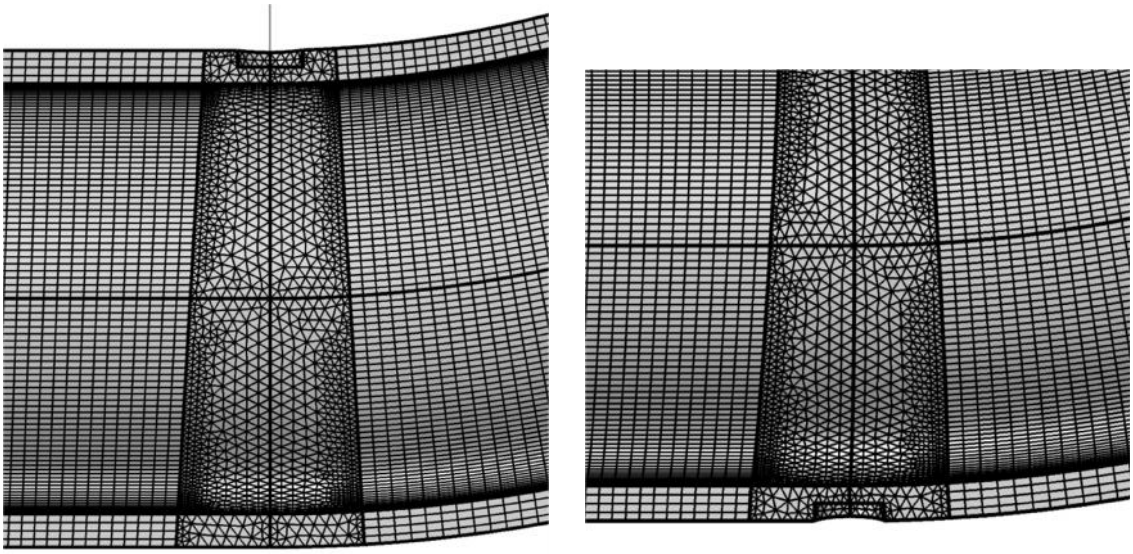


Figure 7 FEM mesh detail with flat bottom half through hole : (a) inner side of bent, (b) outer side of bent

## 4 SIMULATION RESULTS

The reference case is a 90° bent pipe without defects (Figure 8). The total displacement field indicates parallel lines of alternating amplitudes in the incident wave straight part (lower, left on the image), corresponding to plane harmonic waves. A remarkable phenomenon occurs at the passage to the bend. The waves tend to remain plane and parallel to the incident waves from the straight part and do not become perpendicular on the torus axis as some authors predict. Moreover there is a wave focalization on the outer side of the bent, near its middle. This peculiarities have been emphasized for the first time by Predoi and Petre [17], to our knowledge. A decomposition of the plane incident wave into specific toroidal plane waves seems to be not physical. For this reason we do not try to project the reflected or transmitted acoustic field on the modal basis of the torus and even on the modal basis of the straight pipe.

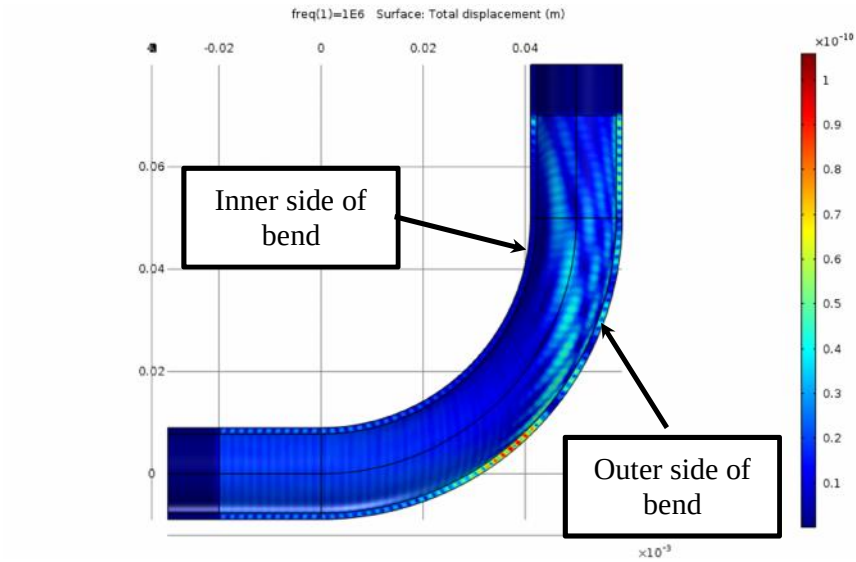



Figure 8 Total displacements for the 90° bent pipe without defects

One can easily associate the irregular pattern of maxima and minima after the middle of the bend as nonplanar partial waves which will impinge on the exit straight part a series of circumferential localized waves. A modal decomposition will be extremely complicated and mostly useless. For this reason we prefer to compare the reflected displacement field of this reference case with the fields of bents with standard defects.

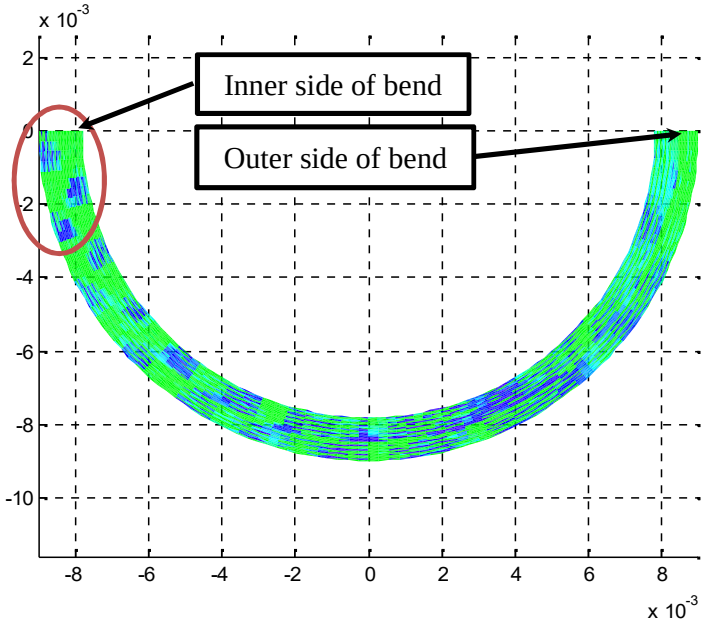



Figure 9 Relative radial displacement with a standard defect on the inner side of the bend

The radial displacement field was extracted by post-processing the FEM results data in all three cases. In order to determine the detectability of defects, the radial displacements fields of the pipes with standard notches were subtracted from the displacement fields of the reference case. All displacements are plotted on a normal cross-section on the straight part of the pipe, 10 mm from the excitation.

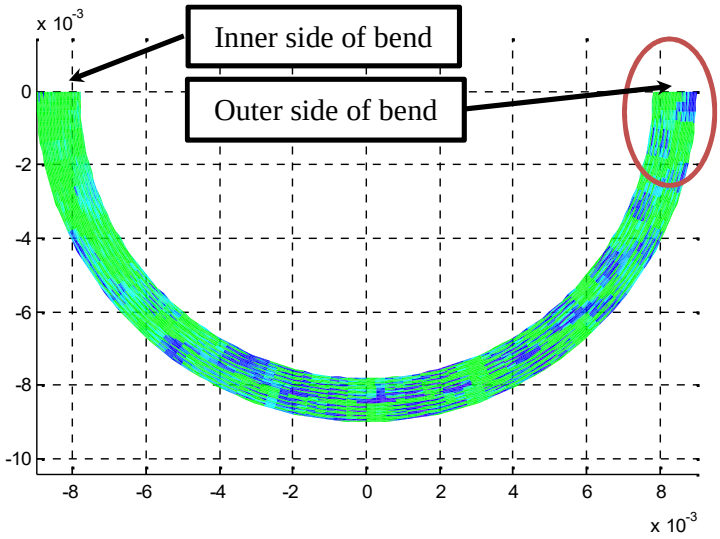



Figure 10 Relative radial displacement with a standard defect on the outer side of the bend

The radial displacements are of higher amplitude on the inner side of the bent when the standard defect is on that side and at the same time, on the outer side of the pipe there is no change in radial signal (Figure 9). If the defect is located on the outer side of the bent, a strong echo can be detected on that side of the pipe and no signal on the opposite side (Figure 10). These results prove the capability of detecting the location of the selected standard defect. The rest of the surface of the pipe is subjected to a complicated pattern of radial displacements and cannot be used for a circumferential receiver transducer.

## 5 CONCLUSIONS

- Numerical analysis using FEM in the frequency domain was used to investigate the defect presence detectability if such a defect occurs at the junction between the straight and curved part of a pipe. This is a challenge for guided waves techniques, since at this junction occurs modal scattering even in the absence of defects. Modal scattering brings spurious signals on the weak signals due to defects.
- A typical defect on either side of the junction is investigated, keeping the geometrical symmetry of the problem.
- The results are very promising, since the location of an extremely small defect (0.6 mm depth and 2.4 mm diameter) is clearly identified, even if the defect is at the junction between the straight and bent part of the pipe.
- Moreover, we have not used a difficult to obtain, single mode excitation, but a piston like one, at a common frequency of 1MHz.

- The receiver should be however another longitudinal waves transducer, placed perpendicular on the pipe, on either inner or outer generatrix, not too far from the bent.

## ACKNOWLEDGEMENT

The present work was supported by the Romanian UEFISCDI Exploratory Research Project PN-II-ID-PCE-2011-3-0512.

## REFERENCES

[1] J.J. Ditri, Utilization of guided elastic waves for the characterization of circumferential cracks in hollow cylinders , *J. Acoust. Soc. Am.*, **96**, 3769-3775, 1994.

[2] P. Cawley, M.JS. Lowe, F. Simonetti, C. Chevalier, AG. Roosenbrand, The variation of the reflection coeffcient of extensional guided waves in pipes from defects as a function of defect depth, axial extent, circumferential extent and frequency, *Proc Instn Mech Engrs*, **216**, 1131-1143, 2002.

[3] M.JS. Lowe, D.N. Alleyne, P. Cawley, Defect detection in pipes using guided waves, *Ultrasonics*, **36**, 147-154, 1998.

[4] J. Li, J.L. Rose, Excitation and propagation of non-axisymmetric guided waves in a hollow cylinder, *J. Acoust. Soc. Am.*, **109**, 457-464, 2001.

[5] BA. Auld, *Acoustic Fields and Waves in Solids*, John Wiley & Sons, 1973, vol. I.

[6] J. Miklowitz, *The Theory of Elastic Waves and Waveguides*, I ed., North Holland, 1978.

[7] J. Achenbach, *Wave Propagation in Elastic Solids*, North-Holland, 1984.

[8] K.F. Graff, *Wave Motion in Elastic Solids*, Dover Publ., 1975.

[9] J.L. Rose, *Ultrasonic Waves in Solid Media*, Cambridge University Press, 2004.

[10] D.C. Gazis, Three-Dimensional Investigation of the Propagation of Waves in Hollow Circular Cylinders. I. Analytical Foundation, *J. Acoust Soc. Am.*, **31**, 568-573, 1959.

[11] D.C. Gazis, Three-Dimensional Investigation of the Propagation of Waves in Hollow Circular Cylinders. II. Numerical Results, *J. Acoust. Soc. Am.*, **31**, 573-578, 1959.

[12] M.V. Predoi et al., Tubewave - Interactive software dedicated to guided waves in tubes, *Rom. J. Acoust. Vib.*, **XI**, 94-99, 2014.

[13] M. Lowe. (2014) Disperse software. An Interactive Program for Generating Dispersion Curves. [Online]. http://www3.imperial.ac.uk/nde/products%20and%20services/disperse

[14] M.V. Predoi, Guided waves dispersion equations for orthotropic multilayered pipes solved using standard finite elements code, *Ultrasonics*, **54**, 1825-1831, 2014.

[15] T. Hayashi, WJ. Song, J.L. Rose, Guided wave dispersion curves for a bar with an arbitrary cross-section, a rod and rail example, *Ultrasonics*, **41**, 175-183, 2003.

[16] D. Gsell, T. Leutenegger, J. Dual, Modeling three-dimensional elastic wave propagation in circular cylindrical structures using a finite-difference approach, *J. Acoust. Soc. Am.*, **116**, 3284-3293, 2003.

[17] M.V. Predoi, C.C. Petre, "Guided waves scattering by discontinuities near pipe bends," in *Proceedings of Meetings on Acoustics*, vol. 19, 2013.

[18] M.V. Predoi, M. Soare, C.C. Petre, "Pipes inspection using longitudinal guided waves based on finite elements simulation," in *28th DANUBIA-ADRIA Symposium on Advances*

*in Experimental Mechanics*, Siofok, Hungary, 2011.

[19] M.V. Predoi, C.C. Petre, M. Soare, "Selection of guided modes in pipes for their sensitivity to flaws," in *International Conference on Structural Analysis of Advanced Materials*, Sinaia, Romania, 2011.

[20] COMSOL Multiphysics User Manual. (2014) [Online]. http://www.comsol.com/comsol-multiphysics

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DUCTILE FRACTURE CRITERIA IN PREDICTION OF SLANT FRACTURE

**P. Kubík**[1], **F. Šebek**[1], **and J. Petruška**[1]

[1]Brno University of Technology, Faculty of Mechanical Engineering, Institute of Solids Mechanics, Mechatronics and Biomechanics, Technická 2896/2, 616 69 Brno, Czech Republic
e-mail: {kubik.p,sebek,petruska}@fme.vutbr.cz

**Keywords:** Ductile fracture, Coupled models, Continuum damage mechanics, Slant fracture, Localization, Explicit dynamics.

**Abstract.** *The ductile fracture of metallic materials is consequence of damage accumulation after straining accompanied by large plastic deformations. It is significantly dependent on the microstructure of particular material and micromechanical defects as vacancies or second phase particles. The nucleation, growth and coalescence of voids is the fracture mechanism which applies in high values of stress triaxiality. The shear mechanism appears in the region of negative stress triaxialities. Finally, the combination of both fracture mechanisms occurs in cases of moderate stress triaxialities. The crack or fracture surface in specimens or real components is often tilted approximately 45 degrees to applied load. This slant fracture is driven by shear mechanism and occurs in the plane of maximum shear stress. It was shown that coupled ductile fracture criteria are convenient for the prediction of slant fracture in finite element simulations. There is conducted analysis of slant fracture prediction ability of two chosen coupled and uncoupled criteria in the present paper. Those criteria were calibrated and applied to two different metallic materials. The prediction is validated on those calibration fracture tests at which the slant fracture was observed.*

## 1 INTRODUCTION

There has been a large increase of application of ductile fracture criteria to prediction of failure at various crashes and manufacturing operations as turning, cutting or cold working [1, 2, 3, 4] in the last decade due to evolution of computer technology. There are many criteria in the literature based on different approaches. Those can be simply categorized as coupled and uncoupled. Coupled criteria have mutually connected the damage with plasticity. The damage is estimated on the basis of accumulated plastic deformation and fracture strain which is dependent on the stress state. In case of uncoupled criteria, the damage is estimated by the same means but it does not influence the plasticity. The slant fracture has been still studied intensively. Huang and Xue [5] paid attention to prediction of slant ductile fracture on flat specimen, thick-walled pipe and four-point bending of a pre-cracked pipe. Morgeneyer and Besson [6] used a Gurson-like model with suggesting a shear void nucleation term based on a Lode parameter for strain rate for simulation of a fully meshed Kahn tear test sample for which a flat to slant crack transition was achieved at loads close to experimental results for aluminium. Li and Wierzbicki [7] simulated a slant fracture of a flat tensile specimens cut from advanced high strength steel sheets using partially coupled model.

The present paper deals with prediction ability of two coupled and uncoupled ductile fracture criteria on the flat specimen loaded by tension and on the smooth cylindrical specimen of Aluminium Alloy (AA) 2024 and AISI 1045 carbon steel, respectively. The most influencing parameters, mesh size and its configuration, which trigger the slant fracture were analysed. The crack initiation and propagation were realized through element deletion technique which requires the adequately small size of finite elements in principle. The numerical simulations are conducted within Abaqus/Explicit commercial finite element code with the use of user subroutine VUMAT for implementation of ductile fracture criteria into time integration scheme.

## 2 MODEL OF PLASTICITY AND FAILURE

### 2.1 Plasticity model

Isotropic hardening and von Mises yield criterion with associated flow rule were adopted. Flow curves were calibrated using tensile tests of smooth cylindrical specimens for both metals.

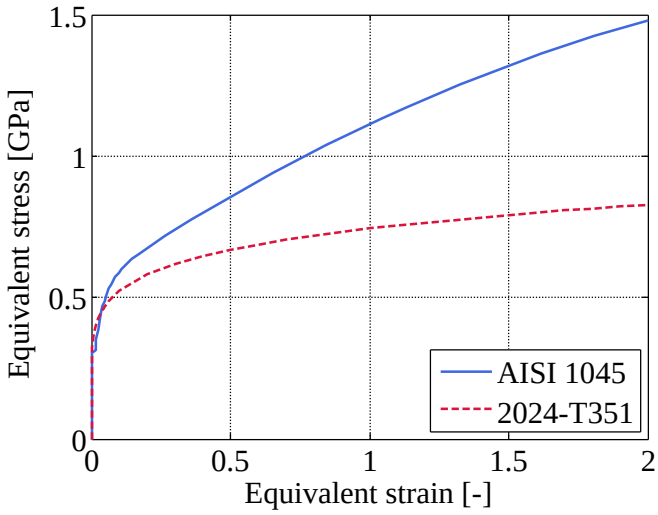



Figure 1: Flow curves of investigated metals.

Six specimens were tested having 6 mm in diameter with 30 mm gauge length of AISI 1045 steel. The flow curve until ultimate tensile strength was estimated using classical transformation formulae for true stress–strain relationship to engineering stress–strain. It is suitable to use some of correction formulae for extracting the flow curve directly from experiment beyond the plastic instability when available [8, 9, 10]. Only the elongation was measured during tensile testing in our case. Therefore, the trial and error method together with computational simulations were employed in order to change the extrapolated trend of flow curve until the satisfying match between computation and experiment was reached. Wierzbicki et al. [11] carried out a series counting fifteen experiments in a broad range of stress triaxiality $\eta$ and normalized third invariant of deviatoric stress tensor $\xi$ for AA 2024-T351. Smooth cylindrical specimen used for calibration of flow curve had 9 mm in diameter and gauge lenght of 25.4 mm. Flow curves of both materials are depicted in Fig. 1. Satisfying power law fit of those curves gave strength coefficients and strain hardening exponents summarized in Tab. 1. Hollomon hardening law, $\bar{\sigma} = K\bar{\epsilon}^n$ where $\bar{\sigma}$ is equivalent stress, $K$ is strength coefficient, $\bar{\epsilon}$ is equivalent strain and $n$ is strain hardening exponent, was used.

| Material | Strength coefficient [MPa] | Strain hardening exponent [-] |
|---|---|---|
| AISI 1045 | 1147.6 | 0.2998 |
| 2024-T351 | 744 | 0.153 |

Table 1: Strength coefficients and strain hardening exponent of investigated materials.

## 2.2 Fracture criteria

Two phenomenological models were chosen in the present study. One of them was 4-parametric Xue–Wierzbicki model [11] with material constants $F_1, \ldots, F_4$ which incorporates fracture envelope $\bar{\epsilon}^f$ symmetric with respect to plane strain condition

$$\bar{\epsilon}^f(\eta, \xi) = F_1 \mathrm{e}^{-F_2\eta} - \left(F_1 \mathrm{e}^{-F_2\eta} - F_3 \mathrm{e}^{-F_4\eta}\right)\left(1 - |\xi|^{\frac{1}{n}}\right)^n \tag{1}$$

This model captures the transition between uniaxial tension and compression, where the material exhibits the greatest ductility, very well.

Another model was Extended Mohr–Coulomb criterion which was proposed by Bai and Wierzbicki [12] with the use of plasticity developed earlier [13]. The criterion contains eight material constants but only two of them, $c_1$ and $c_2$, are related to fracture. Other constants are related to Lode and pressure dependent plasticity used in the model development. When only von Mises yield criterion and Hollomon hardening law are adopted, the fracture envelope reads

$$\bar{\epsilon}^f(\eta, \bar{\theta}) = \left[\frac{K}{c_2}\left(\sqrt{\frac{1+c_1^2}{3}}\cos\left(\frac{\pi}{6}\bar{\theta}\right) + c_1\left[\eta + \frac{1}{3}\sin\left(\frac{\pi}{6}\bar{\theta}\right)\right]\right)\right]^{-\frac{1}{n}} \tag{2}$$

Deviatoric stress state parameter is the normalized Lode angle $\bar{\theta}$ here. Extended Mohr–Coulomb criterion in the presented reduced form has lower versatility than Xue–Wierzbicki model described above. It implies that it has worse approximation ability to experimental fracture tests in the calibration stage. On the other hand, the Extended Mohr–Coulomb criterion contains cut-off region dependent on the material constant $c_1$. Limiting stress triaxiality is described as

$$\eta_c = -\frac{1}{c_1}\left[\sqrt{\frac{1+c_1^2}{3}}\cos\left(\frac{\pi}{6}\bar{\theta}\right)\right] - \frac{1}{3}\sin\left(\frac{\pi}{6}\bar{\theta}\right) \tag{3}$$

Besides tensile test of smooth cylindrical specimen, tensile tests of notched cylindrical specimens with notch radii 5, 2.5 and 1.2 mm were included in fracture criteria calibration regarding the AISI 1045 carbon steel. The diameters in the smallest cross-section were the same as in smooth specimen. Then, biaxial tests of notched tube specimens were used in the calibration as well. Various combinations of tension–torsion loading were chosen, 0, 0.5, 1, $\infty$ mm/rad [14]. Comparison of force and torque responses from experiments and computations can be found in the work of Šebek et al. [15, 16].

All fifteen tests carried out by Wierzbicki et al. [11] were used for ductile fracture calibration regarding the aluminium alloy 2024-T351. More detailed information about the specimens geometry and experiment conditions can be found in the work of Bao [17].

Material constants were identified using nonlinear least square method in MATLAB. Obtained constants for both metals are summarized in Tab. 2. Fracture envelopes of particular criteria are depicted using these constants in Figs. 2 and 3. The fracture strain of both criteria is also depicted under condition of plane stress for both investigated materials in Fig. 4.

| Material | Fracture criterion | Material constants |
|---|---|---|
| AISI 1045 | Xue–Wierzbicki model | $F_1 = 3.1288$, $F_2 = 2.1351$, $F_3 = 1.2822$, $F_4 = 1.1160$ |
| | Extended Mohr–Coulomb criterion | $c_1 = 0.2499$, $c_2 = 753.8\,\text{MPa}$ |
| 2024-T351 | Xue–Wierzbicki model | $F_1 = 0.7161$, $F_2 = 1.5768$, $F_3 = 0.3297$, $F_4 = 0.5232$ |
| | Extended Mohr–Coulomb criterion | $c_1 = 0.0621$, $c_2 = 343.8\,\text{MPa}$ |

Table 2: Material constants of both investigated materials for both criteria.

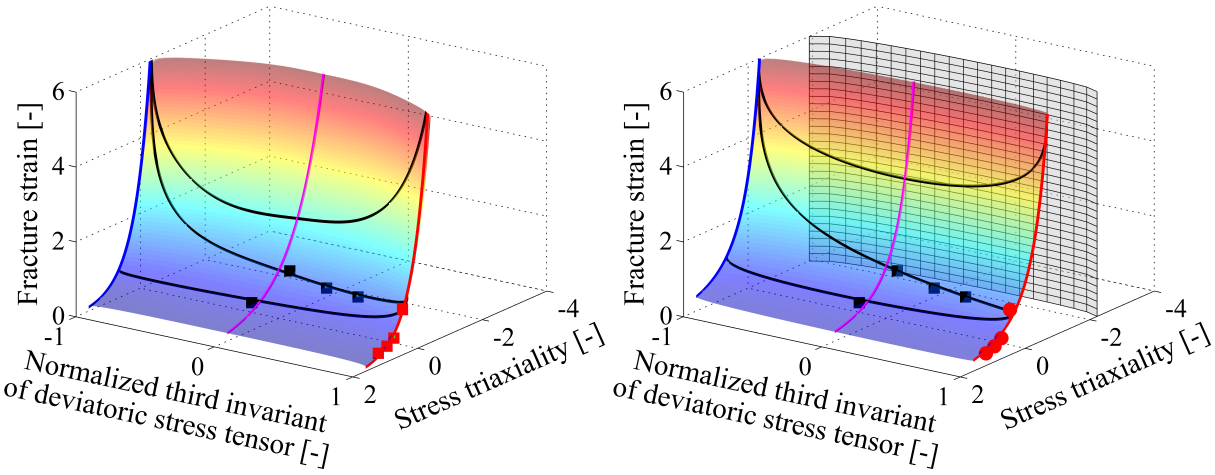



Figure 2: Fracture envelope for steel and Xue–Wierzbicki model (left), Extended Mohr–Coulomb criterion (right).

There is depicted a cut-off plane for Extended Mohr–Coulomb criterion and AISI 1045 carbon steel in Fig. 2. There is no damage accumulation below this region. There is not a cut-off plane at Extended Mohr–Coulomb criterion for aluminium alloy in Fig. 3 because it is situated in extremely low stress triaxialities, approximately $-8$, which is probably unrealistic.

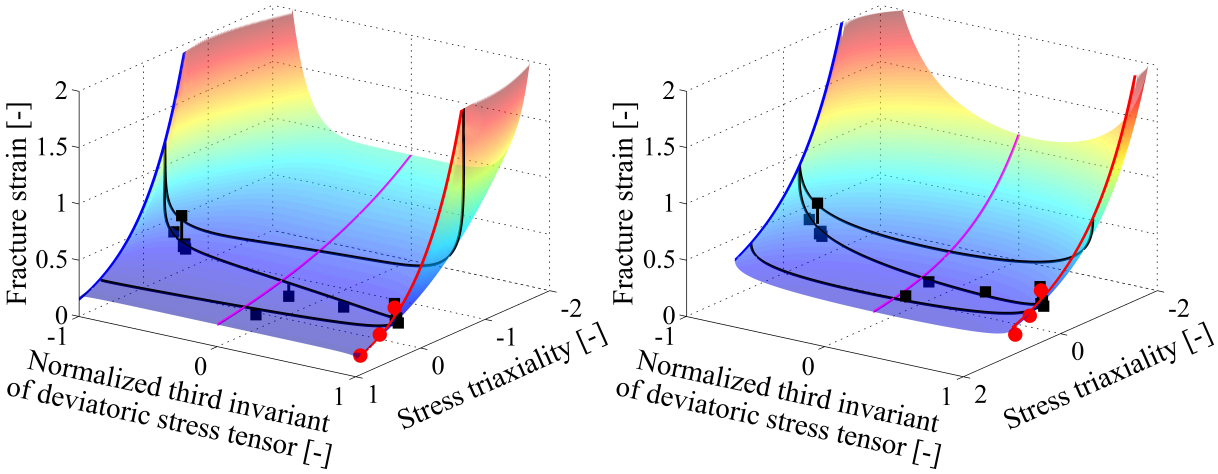



Figure 3: Fracture envelope for AA and Xue–Wierzbicki model (left), Extended Mohr–Coulomb criterion (right).

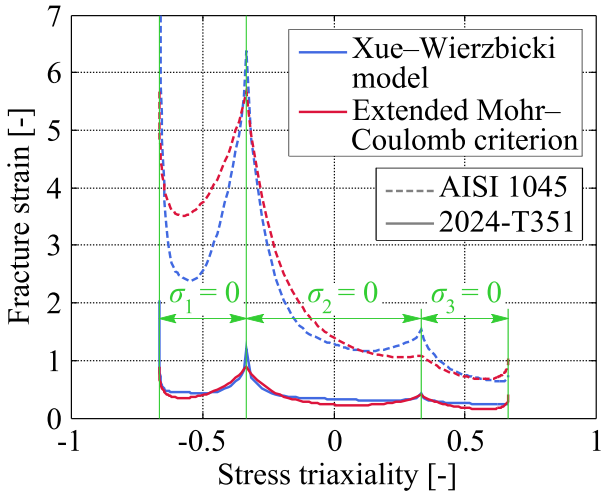



Figure 4: Fracture strain under plane stress condition for both metals.

## 2.3 Coupling failure with plasticity

The damage accumulation was governed by power law function derived by Xue [18, 19] in scope of low-cycle fatigue as

$$D = \int_0^{\hat{\epsilon}^f} m \left( \frac{\bar{\epsilon}^p}{\bar{\epsilon}^f} \right)^{m-1} \frac{\mathrm{d}\bar{\epsilon}^p}{\bar{\epsilon}^f} \tag{4}$$

Here, $m$ is the damage exponent and $\hat{\epsilon}^f$ the equivalent plastic strain for given loading path. The role of damage exponent in the finite element simulations was described in detail by Šebek et al. [20]. The coupling is ensured by weakening functiong $w$ which represents material degradation in accordance to damage accumulation through

$$w = 1 - D^{\beta} \tag{5}$$

There is additional softening parameter $\beta$ which is the weakening exponent.

The feedback of damage to plasticity is realized through multiplying the original flow stress $\bar{\sigma}_{flow}$ by weakening function as

$$\tilde{\sigma}_{flow} = w\bar{\sigma}_{flow} \tag{6}$$

In the previous equation, $\tilde{\sigma}_{flow}$ is actual flow stress of weakened material or effective stress.

Both aforementioned ductile fracture criteria were developed as uncoupled and were also calibrated with respect to that. Calibration of coupled criteria is very complicated process which requires many computational simulations. Therefore, calibrated fracture envelopes of uncoupled criteria described in previous section were used. The weakening and damage exponents, respectively, were set to have a value of 2 which is widely adopted for metallic materials [21, 22]. This is not a correct calibration procedure which will probably result in premature fracture. Nevertheless, such material constants identification is acceptable for the present paper dealing with slant fracture prediction ability.

## 3 PREDICTION OF SLANT FRACTURE

### 3.1 Experiments

Two different specimens were chosen for prediction of slant fracture. The first one is flat specimen loaded by tension which produces slant fracture through the thickness tilted approximately 45 degrees with respect to the major principal stress direction [5, 17, 24, 25]. There is post-mortem specimen of AA 2024-T351 in Fig. 5a. This geometry was overtaken from Bao [17] and used in upcoming numerical simulations. The specimen has 1.6 mm thickness and 50 mm width of the process zone with 8 mm gauge length.

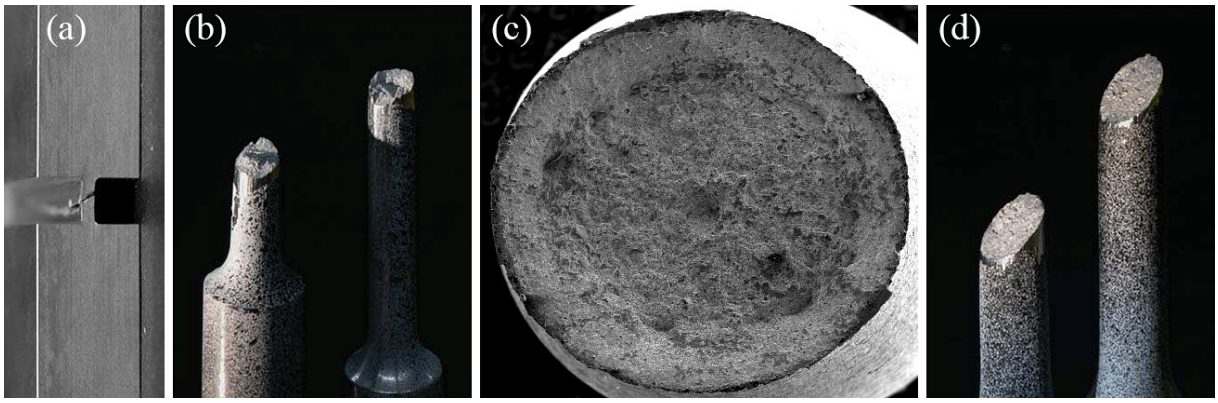



Figure 5: Post-mortem: (a) flat specimen of aluminium alloy 2024-T351 [24]; (b) smooth cylindrical specimen of aluminium alloy 2024-T351 with 6 mm in diameter [24]; (c) smooth cylindrical specimen of AISI 1045 carbon steel [25]; (d) smooth cylindrical specimen of aluminium alloy 2024-T351 with 9 mm in diameter [24].

Another investigated geometry was smooth cylindrical specimen loaded by tension. Typical cup and cone fracture prevails at most metals (Fig. 5b,c). The crack initiation occurs on the axis of symmetry due to void nucleation, growth and coalescence. The crack spreads perpendicularly to the major principal stress, that is perpendicularly to the specimen axis, forming a flat fracture. Then, the nature of mechanism changes to the slant fracture due to increasing shear stress and the crack propagates in the plane of maximum shear stress which is tilted approximately 45 degrees to the specimen axis. Unfortunately, there is often only slant fracture in case of AA 2024-T351 (Fig. 5d) which has still been waiting to be interpreted correctly. Geometry of this specimen was described earlier and used in upcoming numerical simulations.

### 3.2 Simulations

All numerical simulations were carried out in Abaqus/Explicit commercial finite element code. Fracture criteria were implemented using user subroutine Vectorized User MATerial (VUMAT). Crack initiation and propagation were realized through element deletion technique. Time increment was changed using mass scaling function in order to keep computational time reasonable while retaining negligible ratio of kinetic to internal energy for the whole model.

There was conducted analysis studying the influence of finite elements sizes which were oriented in the loading direction, hereinafter II-orientation, in case of uncoupled criteria. Elements had a size of 0.075, 0.05 and 0.025 mm. Next, there were performed simulations using uncoupled criteria with a mesh of a characteristic element length size of 0.075 mm and finite elements orientation 45 degrees to applied load, hereinafter X-orientation, thus in the direction of supposed crack propagation. Finally, there were conducted computations with elements having size of 0.075 mm with II-orientation for coupled ductile fracture criteria.

### 3.3 Flat specimen

Only planar model was created in order to save some computational time. It was discretized using 4-node rectangular plane strain elements CPE4R with reduced integration. In case of X-oriented mesh, a negligible amount of 3-node triangular plane strain elements CPE3 was used on lateral sides of the geometry. Areas lying far from the process zone were discretized by coarse mesh, again, in order to save computational time.

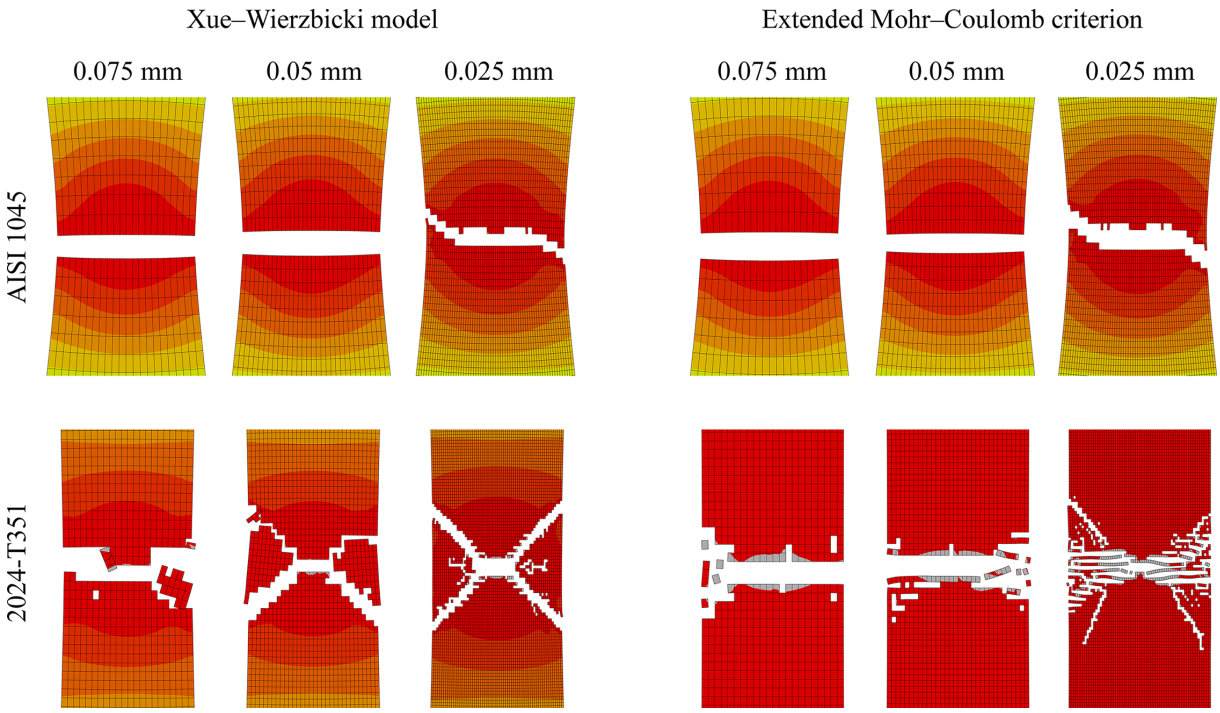



Figure 6: Influence of mesh size for uncoupled fracture criteria and flat specimen with II-orientation.

There is damage parameter in Fig. 6 immediately after the final fracture and separation of the specimen into more pieces for uncoupled criteria and flat specimen with II-orientation. The legend is not shown hereinafter but it was set from 0 to 1 and linearly divided into 20 discrete rainbow contours. The blue colour represents 0 and the red one corresponds to 1.

Both uncoupled criteria are capable of prediction the slant fracture in case of both materials for the mesh with 0.075 mm characteristic element length size.

Models behave similarly in case of AISI 1045 carbon steel, as depicted in Fig. 6. Both meshes with element sizes 0.075 and 0.05 mm, respectively, produce a flat fracture. In case of 0.025 mm, the crack propagates perpendicularly to loading and form a slant fracture in the final stage. There is no splitting into two directions. This is probably due to greater influence of imperfection because there is higher necking than in case of aluminium alloy.

The finer the mesh in case of Xue–Wierzbicki model and 2024-T351, the better the ability to produce the slant fracture. The simulation with size of 0.025 mm predicted the slant fracture very well. Extended Mohr–Coulomb criterion showed similar behaviour but there was huge amount of deleted elements in the perpendicular direction to applied load. This was probably due to fact that, among others, there was higher localization of damage parameter in case of Xue–Wierzbicki model. Crack created X-shaped fracture due to vertical symmetry of the model.

The crack initiation was in the centre of the specimen thickness, in the intersection of all planes of symmetry, for both materials and fracture criteria, respectively, which is in accordance with experimental testing. The slant fractures, which occurred, were 45 degrees to the applied load which is also in good agreement with experiments.

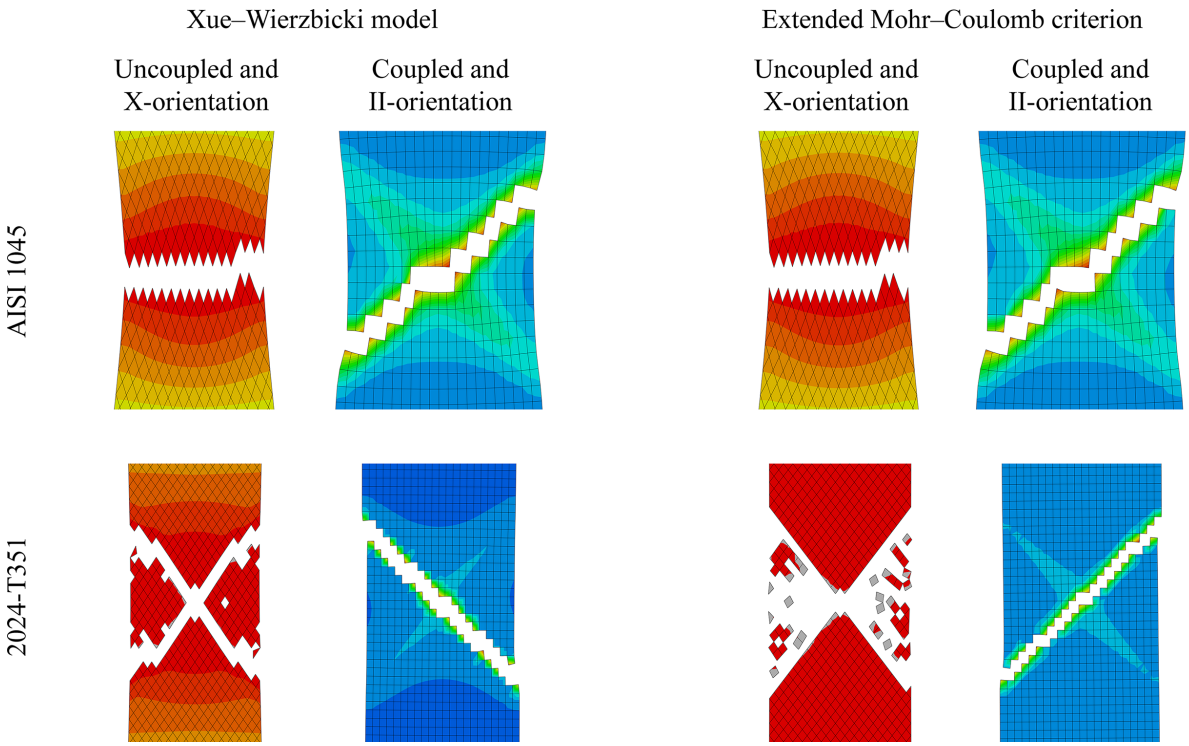



Figure 7: Prediction by uncoupled and coupled criteria for flat plate with X- and II-orientation, respectively.

It is obvious that criteria are capable to predict the slant fracture even with element size of 0.075 mm for X-orientation of the mesh in case of aluminium alloy as in Fig. 7. There is huge amount of deleted elements in case of Extended Mohr–Coulomb criterion, again. Coupled criteria predict the slant fracture without splitting into two direction and propagation of two cracks. Regarding these models, there is a rapid increase of equivalent plastic strain due to softening in the final stage just before crack initiation. This leads to localization of damage parameter and the imperfections play more significant role than in uncoupled models.

The X-orientation, the diagonal one, with characteristic element size of 0.075 mm leads to flat fracture for AISI 1045 (Fig. 7). This mechanism does not change even with using elements of 0.025 mm size. On the other hand, coupled criteria predict slant fracture similarly as in case of AA 2024-T351.

Finally, it is obvious that neither the crack initiation nor the angle of slant crack propagation were influenced by the mesh configuration or coupling the damage with plasticity.

### 3.4 Smooth cylindrical specimen

Similarly as in previous case, the model was created as axisymmetrical in order to save computational time. The geometry was discretized by 4-node rectangular axisymmetric elements CAX4R with reduced integration. In case of X-oriented mesh, there were used a small number of 3-node triangular axisymmetric elements CAX3 which were located on the axis and on the outer surface of the specimen. The regions outside the process zone were discretized by coarse mesh, again because of saving the computational time, because the deformation and fracture are concentrated into certain section.

Similarly as in case of flat specimen, both models were not capable to describe the slant fracture for both materials and element size 0.075 mm as depicted in Fig. 8.

In case of carbon steel, there was predicted flat fracture independently on the criterion or mesh size.

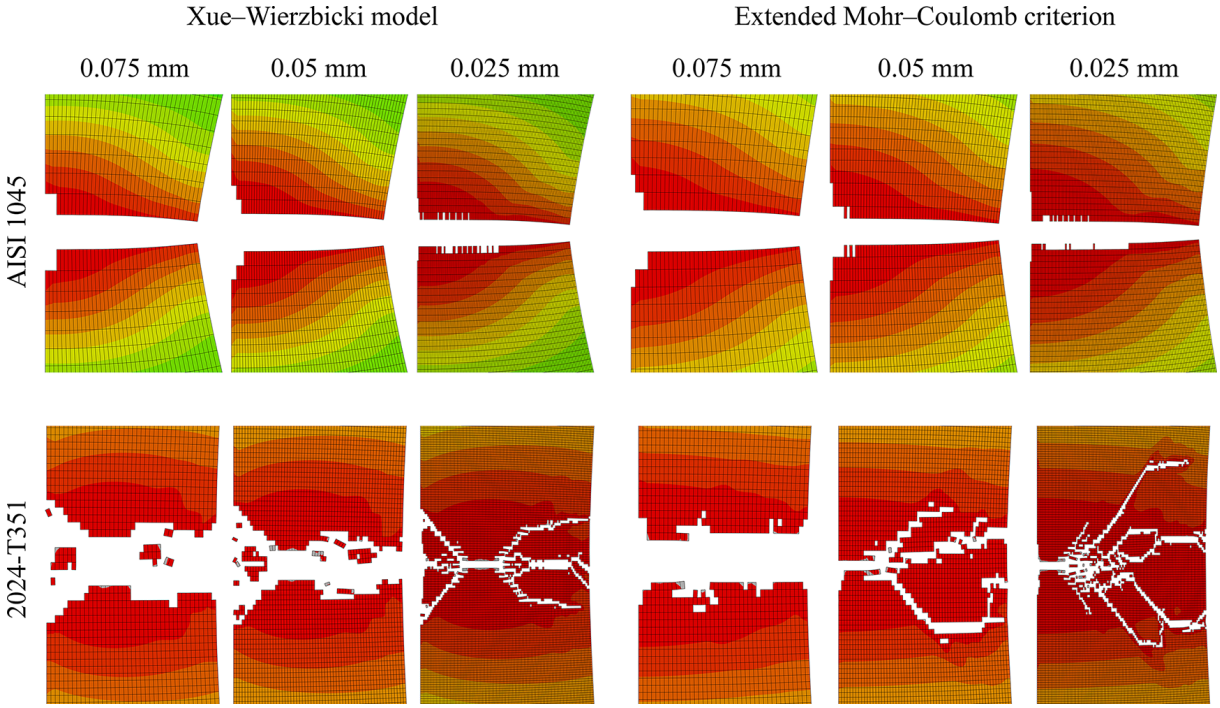



Figure 8: Influence of mesh size for uncoupled fracture criteria and smooth round specimen with II-orientation.

Xue–Wierzbicki model predicted the crack initiation not on the axis of symmetry but in a certain distance towards the outer surface for aluminium alloy which is in contradiction with experiments. Both criteria predicted the flat as well as slant fracture for 0.025 mm sized elements. The slant fracture propagated 45 degrees to applied load at first and then it changed the direction perpendicularly to specimen axis. Extended Mohr–Coulomb criterion exhibited this behaviour also in case of the finite element size of 0.05 mm.

Both uncoupled models with X-oriented mesh produced a flat fracture for carbon steel as in Fig. 9. This behaviour did not change even with decreasing the element size. Coupled Xue–Wierzbicki model produced only flat fracture while coupled Extended Mohr–Coulomb criterion predicted a slight slant fracture in the final stage close to the specimen surface without bifurcation of the crack. In case of 0.05 mm sized finite elements, both coupled models predicted slant fracture in greater distance from the surface in the final stage.

Uncoupled criteria with X-orientation and 0.075 mm sized mesh behaved for the aluminium alloy 2024-T351 the same as in case of II-orientation and 0.025 mm sized finite elements. There was better prediction of crack initiation after coupling the Xue–Wierzbicki model. The crack initiated on the axis of symmetry and there was crack bifurcation in the final stage of rupture governed by shear mechanism. Coupled Extended Mohr–Coulomb criterion predicted slight slant fracture in the final stage. Again, there was significant number of elements deleted in the plane perpendicular to the axis.

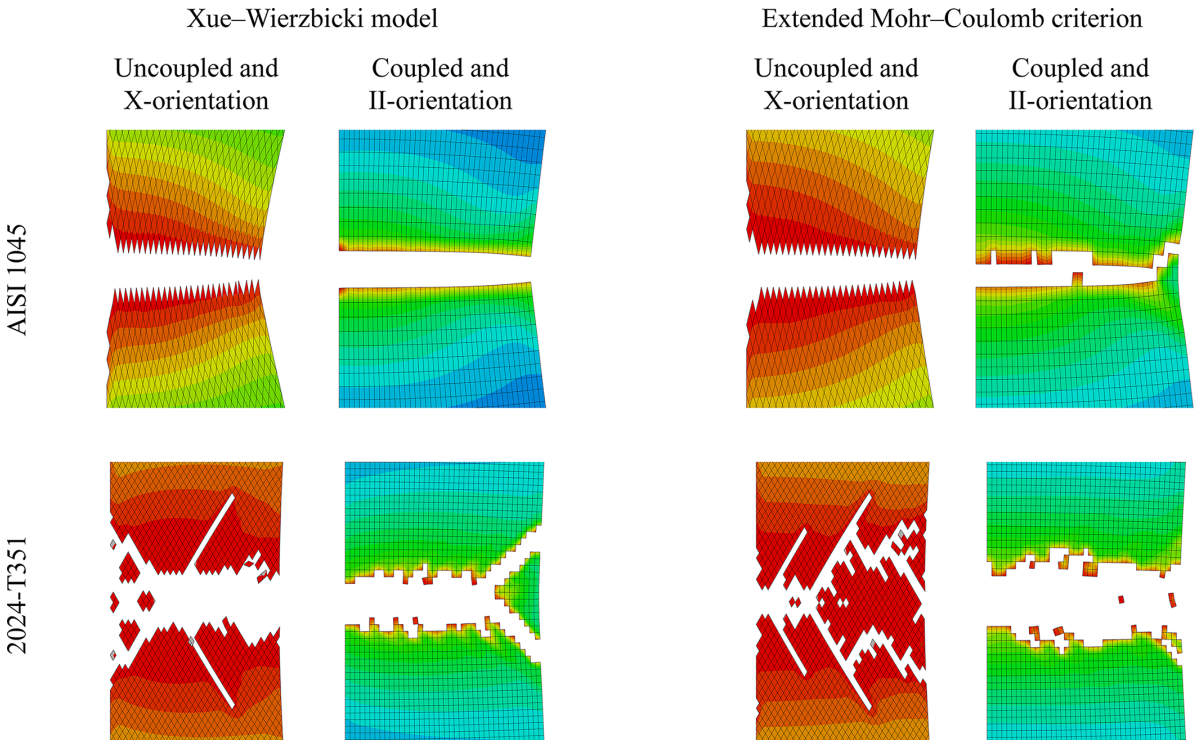



Figure 9: Results of uncoupled and coupled criteria for smooth specimen with X- and II-orientation, respectively.

## 4 CONCLUSIONS

The present paper was aimed to slant fracture prediction ability of two coupled and uncoupled ductile fracture criteria, originally phenomenological or empirical, Xue–Wierzbicki model and Extended Mohr–Coulomb criterion, for two different metals, aluminium alloy 2024-T351 and AISI 1045 carbon steel. There was conducted analysis studying the influence of the size and orientation of finite element mesh in case of uncoupled criteria. It might be generally stated that the mesh oriented in the direction of applied load exhibited similar results as a finite element mesh oriented in the supposed direction of slant fracture propagation, 45 degress to the applied load in the plane of maximum shear stress, but for much smaller element sizes.

Results also suggest that coupled models are very convenient for slant fracture prediction which is in accordance with recent literature [26, 27]. Such models describe best the real material behaviour. Nevertheless, there is a disadvantage in complex calibration procedure from the perspective of time consumption, high number of experiments needed or complicated implementing and debugging.

## 5 ACKNOWLEDGEMETNT

This work is an output of project NETME CENTRE PLUS (LO1202) created with financial support from the Ministry of Education, Youth and Sports under the "National Sustainability Programme I".

## REFERENCES

[1] L. Xue, W. Mock, T. Belytschko, Penetration of DH-36 steel plates with and without polyurea coating. *Mechanics of Materials*, **42**, 981–1003, 2010.

[2] Y. Li, M. Luo, J. Gerlach, T. Wierzbicki, Prediction of shear-induced fracture in sheet metal forming. *Journal of Materials Processing Technology*, **210**, 1858–1869, 2010.

[3] J. Petruška, P. Kubík, J. Hůlka, F. Šebek, Ductile fracture criteria in prediction of chevron cracks. *Advanced Materials Research*, **716**, 653–658, 2013.

[4] P. Kubík, F. Šebek, J. Hůlka, J. Petruška, Calibration of ductile fracture criteria at negative stress triaxiality. *International Journal of Mechanical Sciences*, **108–109**, 90–103, 2016.

[5] H. Huang, L. Xue, Prediction of slant ductile fracture using damage plasticity theory. *International Journal of Pressure Vessels and Pipping*, **86**, 319–328, 2009.

[6] T.F. Morgeneyer, J. Besson, Flat to slant ductile fracture transition: Tomography examination and simulations using shear-controlled void nucleation. *Scripta Materialia*, **65**, 1002–1005, 2011.

[7] Y. Li, T. Wierzbicki, Prediction of plane strain fracture of AHSS sheets with post-initiation softening. *International Journal of Solids and Structures*, **47**, 2316–2327, 2010.

[8] P.W. Bridgman, *Studies in large plastic flow and fracture: With special emphasis on the effects of hydrostatic pressure*. Harvard University Press, Cambridge, Massachusetts, 1964.

[9] G. La Rosa, G. Mirone, A. Risitano, Postnecking Elastoplastic Characterization: Degree of Approximation in the Bridgman Method and Properties of the Flow-Stress/True-Stress Ratio. *Metallurgical and Materials Transactions A*, **34A**, 615–624, 2003.

[10] G. Mirone, A new model for the elastoplastic characterization and the stress–strain determination on the necking section of a tensile specimen. *International Journal of Solids and Structures*, **41**, 3545–3564, 2004.

[11] T. Wierzbicki, Y. Bao, Y.-W. Lee, Y. Bai, Calibration and evaluation of seven fracture models. *International Journal of Mechanical Sciences*, **47**, 719–743, 2005.

[12] Y. Bai, T. Wierzbicki, Application of extended Mohr–Coulomb criterion to ductile fracture. *International Journal of Fracture*, **161**, 1–20, 2010.

[13] Y. Bai, T. Wierzbicki, A new model of metal plasticity and fracture with pressure and Lode dependence. *International Journal of Plasticity*, **24**, 1071–1096, 2008.

[14] P. Kubík, F. Šebek, J. Petruška, J. Hůlka, J. Růžička, M. Španiel, J. Džugan, A. Prantl, Calibration of selected ductile fracture criteria using two types of specimens. *Key Engineering Materials*, **592–593**, 258–261, 2014.

[15] F. Šebek, P. Kubík, J. Petruška, Chevron crack prediction using the extremely low stress triaxiality test. *MM Science Journal*, 617–621, 2015.

[16] F. Šebek, P. Kubík, J. Hůlka, J. Petruška, Strain hardening exponent role in phenomenological ductile fracture criteria. *European Journal of Mechanics - A/Solids*, **57**, 149–164, 2016.

[17] Y. Bao, *Prediction of ductile crack formation in uncracked bodies*. PhD thesis, Massachusetts Institute of Technology, 2003.

[18] L. Xue, Damage accumulation and fracture initiation in uncracked ductile solids subject to triaxial loading. *International Journal of Solids and Structures*, **44**, 5163–5181, 2007.

[19] L. Xue, A unified expression for low cycle fatigue and extremely low cycle fatigue and its implication for monotonic loading. *International Journal of Fatigue*, **30**, 1691–1698, 2008.

[20] F. Šebek, J. Hůlka, P. Kubík, J. Petruška, On the Proportionality of Damage Rule in Finite Element Simulations of the Ductile Failure. *Advanced Materials Research*, **980**, 189–193, 2014.

[21] B.L. Boyce, S.L.B. Kramer, H.E. Fang, T.E. Cordova, M.K. Nielsen, K. Dion, et al., The Sandia Fracture Challenge: blind round robin predictions of ductile tearing. *International Journal of Fracture*, **186**, 5–68, 2014.

[22] T.S. Cao, *Modeling ductile damage for complex loading paths*. PhD thesis, École Polytechnique Université, 2013.

[23] T.S. Cao, Numerical simulation of 3D ductile cracks formation using recent improved Lode-dependent plasticity and damage models combined with remeshing. *International Journal of Solids and Structures*, **51**, 2370–2381, 2014.

[24] L. Xue, *Ductile Fracture Modeling - Theory, Experimental Investigation and Numerical Verification*. PhD thesis, Massachusetts Institute of Technology, 2007.

[25] J. Bořkovec, *Computer simulation of material separation process*. PhD thesis, Brno University of Technology, 2008.

[26] X. Teng, Numerical prediction of slant fracture with continuum damage mechanics. *Engineering Fracture Mechanics*, **75**, 2020–2041, 2008.

[27] G. Rousselier, M. Luo, A fully coupled void damage and Mohr–Coulomb based ductile fracture model in the framework of a Reduced Texture Methodology. *International Journal of Plasticity*, **55**, 1–24, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DAMAGE MODELING USING STRAIN GRADIENT BASED FINITE ELEMENT FORMULATION

**Filip Putar, Jurica Sorić, Tomislav Lesičar and Zdenko Tonković**

Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb
Ivana Lučića 5, 10002 Zagreb, Croatia
e-mail: {filip.putar, jurica.soric, tomislav.lesicar, zdenko.tonkovic}@fsb.hr

**Keywords:** Damage, Softening, $C^1$ Continuity Finite Element, Strain Gradient Theory, Heterogeneous Material.

**Abstract.** *The present study deals with the modeling of quasi-brittle softening phenomena using the two dimensional $C^1$ continuity triangular finite element. The formulation of damage model is based on an isotropic damage law applied to the higher-order stress-strain constitutive relations originating from the full strain gradient theory. Both homogeneous and heterogeneous materials are considered by employing the second-order homogenization procedure to obtain the required constitutive matrices. For this purpose, two different heterogeneous representative volume elements are employed, where the results of the softening analyses are compared with those obtained for the corresponding homogeneous materials with equivalent internal length scales. The derived finite element formulation is implemented into the finite element program ABAQUS using user subroutines. Finally, accuracy and efficiency of the proposed higher-order gradient model are demonstrated by the modeling of a stretched plate weakened in the middle, usually used as a benchmark in strain softening analyses.*

## 1 INTRODUCTION

A large number of engineering materials, such as high-strength steels, polymers, composites, concrete and rocks can be classified as quasi-brittle, since they exhibit negligible plastic behavior before and after the damage is initiated, causing the material to soften immediately after reaching the maximum elastic deformation. It is well-known that the application of the classical continuum mechanics cannot properly resolve strain softening because it leads to the local loss of positive definiteness of the material tangent stiffness. As a consequence, the differential equations which describe the deformation process may result in the loss of ellipticity. Mathematical description of the model than becomes ill-posed and numerical solutions do not converge to a physically meaningful solution [1]. In view of the finite elements, solutions are completely dependent on the discretization, with respect to mesh refinement and mesh alignment. Energy dissipated in the fracture process tends to zero when the size of the elements involved in the softening process is reduced, and the localization zone exhibits an extreme tendency to propagate along the mesh lines [2]. Various regularization techniques have been developed in the past few decades to overcome this problem. Most of them are based on the improvement of the classical continuum model, precisely on its enrichment with the internal length scale parameters in several different ways. Micropolar [3] and viscoplastic theory [4] can preserve the ellipticity only in some specific cases, while the theories related to the nonlocal material behavior have been shown to be the most versatile. In the case of the nonlocal models, stress at a material point does not depend only on the strain and other state variables at this point, as it is the case with the classical continuum theory, but also on the strains and other state variables of the points in the surrounding area. The magnitude of this interaction is described by the aforementioned internal length scale parameter, which represents the microstructure of the material [5]. Basically, there are two different approaches when it comes to describing of non-locality in the model, integral and gradient-enhanced approach. The integral approach, introduced in [6], is based on spatial averaging of the state variables, typically strains, in the finite neighborhood of a certain point, leading to the very complicated constitutive relations made of convolution-type integrals. The gradient approach enhances the constitutive relation either by incorporation of the strain-gradients or by introduction of both strain-gradients and their stress conjugates. In case when only strain-gradients are used as an enhancement of the constitutive relation, explicit and implicit gradient formulations are usually used when dealing with softening, either in elasticity context [7], plasticity context [8] or in the analysis of the elastic wave propagation [9]. The second type of gradient approaches where both strain-gradients and their stress conjugates enter the constitutive relation has been employed less often, mainly because it is numerically more complex. In the recent developments, higher-order stress-strain theory has been employed in the context of a damage modeling of an infinitely long bar, where the authors concluded that the addition of the higher-order stress terms results in stabilizing the positive definiteness of tangent stiffness moduli when entering the strain softening regime. In such a way physically consistent solutions can be ensured and strain-softening phenomenon can be realistically reproduced [10]. Further development from one-dimensional to multi-dimensional simulation of a localized failure process has been made in [11]. In [10] and [11] element-free Galerkin (EFG) meshless method has been used for finding the approximate solutions to the corresponding boundary value problems. Another advantage of the higher-order stress-strain theory is that it can easily introduce material heterogeneity in the constitutive relations through the non-diagonal higher order material stiffness tangents [12]. The stiffness tangents can be obtained by applying the second-order homogenization technique on the representative volume element (RVE) [13].

This paper presents formulation and numerical implementation of the higher-order stress-strain damage theory into the $C^1$ continuity displacement based finite element developed in [14]. Herein, continuum damage mechanics is realized in its simplest form through isotropic damage model. Similar consideration of quasi-brittle damage response has been presented in [15]. The proposed damage algorithm is verified on the benchmark example where both homogeneous and heterogeneous materials are employed.

## 2 FORMULATION OF THE $C^1$ FINITE ELEMENT FOR SOFTENING ANALYSIS

In this section, an extension of the $C^1$ continuity finite element presented in the authors' former work [14] for the application in a softening regime is described. The developed finite element formulation is based on a small strain second gradient continuum theory for which more details can be found in [16]. Here the basic relations of the damage constitutive model are presented, and afterwards derivation of the finite element with a softening behavior is shown.

### 2.1 Damage constitutive model

The isotropic damage constitutive model describing the reduction of the elastic stiffness properties in quasi-brittle materials is based on the following stress-strain relationship [17]

$$\boldsymbol{\sigma} = \left(1 - D\right)\mathbf{C}\boldsymbol{\varepsilon}, \tag{1}$$

where $D$ is a scalar damage variable ranging from 0 (undamaged) to 1 (damaged), while $\boldsymbol{\sigma}$, $\boldsymbol{\varepsilon}$ and $\mathbf{C}$ are tensors referring to Cauchy stress, strain and elastic stiffness, respectively. In this contribution, a simple linear damage evolution law is chosen [5]

$$D = \begin{cases} \dfrac{k_u\left(\varepsilon_{eqv} - k_0\right)}{\varepsilon_{eqv}\left(k_u - k_0\right)} & k_0 \le \varepsilon_{eqv} \le k_u \\ 1 & \varepsilon_{eqv} > k_u \end{cases}, \tag{2}$$

where $k_0$ and $k_u$ are the material parameters representing the threshold strain at which damage is initiated, and the strain at which material completely loses its stiffness, respectively. $\varepsilon_{eqv}$ is the equivalent elastic strain measure which, considering the damage only due to tensile strains, can be expressed as

$$\varepsilon_{eqv} = \sqrt{\left\langle\varepsilon_1\right\rangle^2 + \left\langle\varepsilon_2\right\rangle^2} \quad \text{for} \quad \varepsilon_1, \varepsilon_2 > 0 \tag{3}$$

with $\varepsilon_1$ and $\varepsilon_2$ as principal strain components of the strain tensor $\boldsymbol{\varepsilon}$.

### 2.2 $C^1$ triangular finite element for softening analysis

The $C^1$ continuity plane strain triangular finite element proposed in [14] shown in Fig. 1 is used in the formulation derived in this contribution. The element has three nodes and 36 degrees of freedom with the displacement field approximated by the full fifth order polynomial. The nodal degrees of freedom are the two displacements and their first- and second-order derivatives with respect to the Cartesian coordinates.

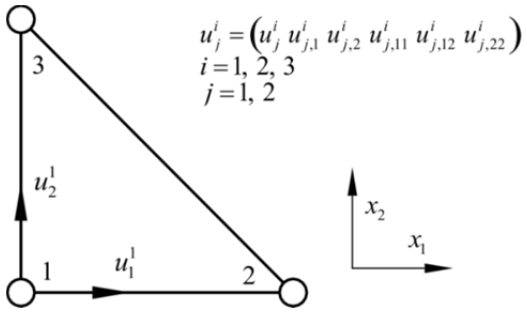



Figure 1: $C^1$ triangular finite element [14]

As usual, the element equations are derived from the principle of virtual work, which may be expressed for the strain gradient continuum as

$$\int_A \delta\boldsymbol{\varepsilon}^T\boldsymbol{\sigma}\,\mathrm{d}A + \int_A \delta\boldsymbol{\eta}^T\boldsymbol{\mu}\,\mathrm{d}A = \int_s \delta\mathbf{u}^T\mathbf{t}\,\mathrm{d}s + \int_s \delta\left(\mathrm{grad}\,\mathbf{u}^T\right)\mathbf{T}\,\mathrm{d}s\,, \tag{4}$$

where $A$ and $s$ are area and perimeter of the element, respectively. In Eq. (4), $\boldsymbol{\eta}$ represents the second-order strain tensor containing second derivatives of the displacement vector $\mathbf{u}$, while $\boldsymbol{\mu}$ is the work conjugate of the second-order strain, the so-called second-order stress or double stress tensor. $\mathbf{t}$ and $\mathbf{T}$ are the traction tensor and the double traction tensor, respectively. The stress and the second-order stress increments, $\Delta\boldsymbol{\sigma}$ and $\Delta\boldsymbol{\mu}$, are computed by the incremental constitutive relations which, for the undamaged material are defined as

$$\begin{aligned}\Delta\boldsymbol{\sigma} &= \mathbf{C}_{\sigma\varepsilon}\Delta\boldsymbol{\varepsilon} + \mathbf{C}_{\sigma\eta}\Delta\boldsymbol{\eta},\\ \Delta\boldsymbol{\mu} &= \mathbf{C}_{\mu\varepsilon}\Delta\boldsymbol{\varepsilon} + \mathbf{C}_{\mu\eta}\Delta\boldsymbol{\eta}.\end{aligned} \tag{5}$$

Herein, $\mathbf{C}_{\sigma\varepsilon}$, $\mathbf{C}_{\sigma\eta}$, $\mathbf{C}_{\mu\varepsilon}$ and $\mathbf{C}_{\mu\eta}$ are the constitutive tangent matrices which can be computed from the appropriate RVE using the second-order homogenization procedure. In case of material homogeneity, the constitutive tangent matrices $\mathbf{C}_{\sigma\eta}$ and $\mathbf{C}_{\mu\varepsilon}$ are assumed to be zero [12]. The remaining two tangent stiffness matrices can be found analytically [18], which may be written symbolically in the form

$$\begin{aligned}\mathbf{C}_{\sigma\varepsilon} &= \mathbf{C}_{\sigma\varepsilon}\left(E,\nu\right),\\ \mathbf{C}_{\mu\eta} &= \mathbf{C}_{\mu\eta}\left(E,\nu,L\right),\end{aligned} \tag{6}$$

where $L$ denotes the size of the microstructural representative volume element. It is well-known that, in the second-order computational homogenization scheme, the RVE size $L$ is linked to the internal length scale $l$ of the higher-order continuum as

$$l^2 = \frac{L^2}{12}. \tag{7}$$

After application of the isotropic damage law, Eq. (1), to the constitutive relations, Eq. (5), the constitutive damage model may be expressed in the form

$$\begin{aligned}\Delta\boldsymbol{\sigma} &= (1-D^{i-1})\left(\mathbf{C}_{\sigma\varepsilon}\Delta\boldsymbol{\varepsilon} + \mathbf{C}_{\sigma\eta}\Delta\boldsymbol{\eta}\right) - \Delta D\left(\mathbf{C}_{\sigma\varepsilon}\boldsymbol{\varepsilon}^{i-1} + \mathbf{C}_{\sigma\eta}\boldsymbol{\eta}^{i-1}\right),\\ \Delta\boldsymbol{\mu} &= (1-D^{i-1})\left(\mathbf{C}_{\mu\varepsilon}\Delta\boldsymbol{\varepsilon} + \mathbf{C}_{\mu\eta}\Delta\boldsymbol{\eta}\right) - \Delta D\left(\mathbf{C}_{\mu\varepsilon}\boldsymbol{\varepsilon}^{i-1} + \mathbf{C}_{\mu\eta}\boldsymbol{\eta}^{i-1}\right),\end{aligned} \tag{8}$$

where the stress and the second-order stress increments are computed from the values of the last converged equilibrium state ($i$–1). Here the incremental change of the damage variable is approximated by

$$\Delta D = \left( \frac{\mathrm{d}D}{\mathrm{d}\boldsymbol{\varepsilon}} \right)^{i-1} \Delta\boldsymbol{\varepsilon} \,. \tag{9}$$

Employing the relations given in [14], the strain and second-order strain increments can be expressed in terms of the nodal displacement increment vector $\Delta\mathbf{v}$ by the relations

$$\begin{aligned} \Delta\boldsymbol{\varepsilon} &= \mathbf{B}_{\varepsilon} \Delta\mathbf{v}, \\ \Delta\boldsymbol{\eta} &= \mathbf{B}_{\eta} \Delta\mathbf{v}, \end{aligned} \tag{10}$$

where $\mathbf{B}_{\varepsilon}$ and $\mathbf{B}_{\eta}$ are the matrices containing corresponding the first and second derivatives of the element interpolation functions. Substituting Eqs. (8)-(10) into Eq. (4), and after some straightforward manipulation, the following finite element equation is obtained

$$\left( \mathbf{K}_{\varepsilon\varepsilon} + \mathbf{K}_{\varepsilon\eta} + \mathbf{K}_{\eta\varepsilon} + \mathbf{K}_{\eta\eta} \right) \Delta\mathbf{v} = \mathbf{F}_{\mathrm{e}} - \mathbf{F}_{\mathrm{i}} \,, \tag{11}$$

where the particular element stiffness matrices are defined as

$$\begin{aligned} \mathbf{K}_{\varepsilon\varepsilon} &= \int_A \mathbf{B}_{\varepsilon}^T \left[ \left(1 - D^{i-1}\right) \mathbf{C}_{\sigma\varepsilon} - \mathbf{C}_{\sigma\varepsilon} \boldsymbol{\varepsilon}^{i-1} \left( \frac{\mathrm{d}D}{\mathrm{d}\boldsymbol{\varepsilon}} \right)^{i-1} - \mathbf{C}_{\sigma\eta} \boldsymbol{\eta}^{i-1} \left( \frac{\mathrm{d}D}{\mathrm{d}\boldsymbol{\varepsilon}} \right)^{i-1} \right] \mathbf{B}_{\varepsilon} \mathrm{d}A, \\ \mathbf{K}_{\varepsilon\eta} &= \int_A \mathbf{B}_{\varepsilon}^T \left(1 - D^{i-1}\right) \mathbf{C}_{\sigma\eta} \mathbf{B}_{\eta} \, \mathrm{d}A, \\ \mathbf{K}_{\eta\varepsilon} &= \int_A \mathbf{B}_{\eta}^T \left[ \left(1 - D^{i-1}\right) \mathbf{C}_{\mu\varepsilon} - \mathbf{C}_{\mu\varepsilon} \boldsymbol{\varepsilon}^{i-1} \left( \frac{\mathrm{d}D}{\mathrm{d}\boldsymbol{\varepsilon}} \right)^{i-1} - \mathbf{C}_{\mu\eta} \boldsymbol{\eta}^{i-1} \left( \frac{\mathrm{d}D}{\mathrm{d}\boldsymbol{\varepsilon}} \right)^{i-1} \right] \mathbf{B}_{\varepsilon} \mathrm{d}A, \\ \mathbf{K}_{\eta\eta} &= \int_A \mathbf{B}_{\eta}^T \left(1 - D^{i-1}\right) \mathbf{C}_{\mu\eta} \mathbf{B}_{\eta} \, \mathrm{d}A. \end{aligned} \tag{12}$$

The external and internal nodal force vectors, $\mathbf{F}_{\mathrm{e}}$ and $\mathbf{F}_{\mathrm{i}}$, in Eq. (11) are defined as

$$\begin{aligned} \mathbf{F}_{\mathrm{e}} &= \int_s \left( \mathbf{N}^T \mathbf{t} + \operatorname{grad} \mathbf{N}^T \mathbf{T} \right) \mathrm{d}s, \\ \mathbf{F}_{\mathrm{i}} &= \int_A \left( \mathbf{B}_{\varepsilon}^T \boldsymbol{\sigma}^{i-1} + \mathbf{B}_{\eta}^T \boldsymbol{\mu}^{i-1} \right) \mathrm{d}A. \end{aligned} \tag{13}$$

### 2.3 Analysis procedure

The block diagram which concisely shows the analysis procedure is presented in Fig. 2. The first step of the analysis is a preprocessing step and includes application of the second-order homogenization procedure to the appropriate RVE in order to obtain required constitutive stiffness tensors. As the linear elastic material behavior is considered in the presented damage analysis, the homogenized solutions do not depend on the macroscale deformation. Therefore, the homogenization has to be performed only once in each analysis. The homogenized stiffness tensors enter the constitutive relations, remaining constant until the end of the nonlinear damage analysis. When damage is initiated in the model, all stiffness tensors are being reduced by the term including the damage variable, as it can be seen from Eq. (8). In each finite element integration point, the incremental procedure is carried out,

where the stress and double stress tensors are calculated from the updated strain tensor, second-order strain tensor and damage variable, as well as the stiffness tensors obtained in the preprocessing step. The presented damage formulation has been implemented into the two-dimensional $C^1$ continuity triangular finite element [14] using the FE program ABAQUS and its user element subroutine UEL [19].

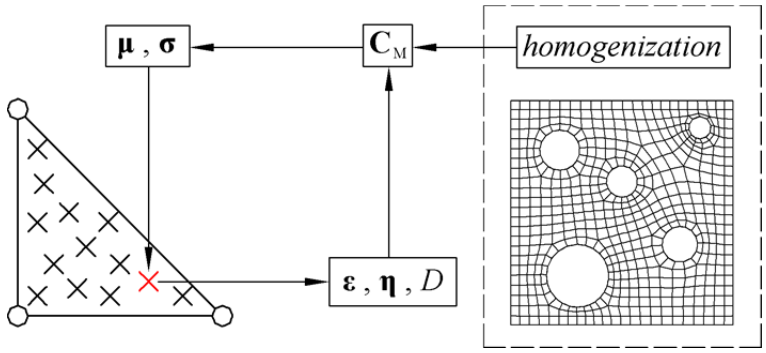



Figure 2: Scheme of the damage algorithm

## 3 NUMERICAL EXAMPLE

The algorithm presented above is verified on a benchmark problem already studied in [11] assuming homogeneous material. Here, the analysis is extended to heterogeneous materials.

As an example, a rectangular plate with an imperfect zone under tension is considered. The geometry and boundary conditions are shown in Fig. 3. The constitutive properties have been set to: Young's modulus $E = 20000$ N/mm$^2$, Poisson's ratio $\nu = 0.25$, limit elastic strain $k_0 = 0.0001$, equivalent strain corresponding to the fully damaged state $k_u = 0.0125$. The model is loaded by the horizontal displacement of 0.0325 mm at the right edge. In order to trigger localization, the Young's modulus is reduced by 10% in a 10 mm wide zone in the middle hatched area of the plate.

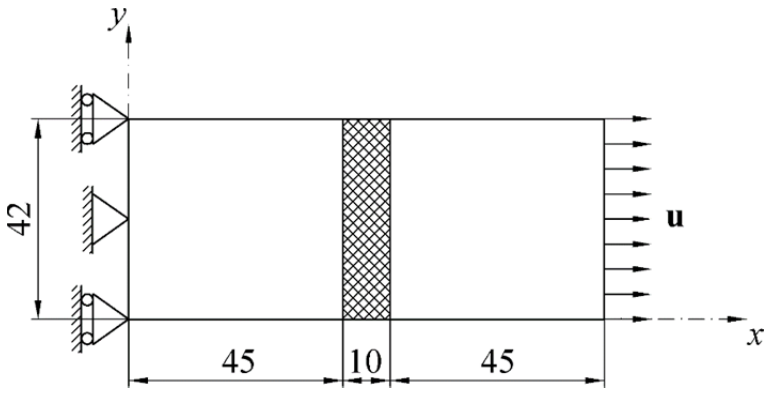



Figure 3: Geometry and boundary conditions of the plate model

The verification of the presented damage model is made using the results from [11], where the solutions are obtained in the same numerical example by means of the EFG meshless method. Therein, the constitutive tensors are derived for materials with granular microstructure, so the underlying microstructural theory differs when compared with the second-order homogenization approach. The constitutive model is restricted only to homogeneous materials, where the corresponding stiffness tensors are similar to those shown in Eq. (6). In [11] the stiffness tensors have been expressed in terms of the particle radius $r$ and they are used for the

computation by means of the proposed algorithm. The damage responses obtained for the same microstructural values and using different approaches are presented in Fig. 4.

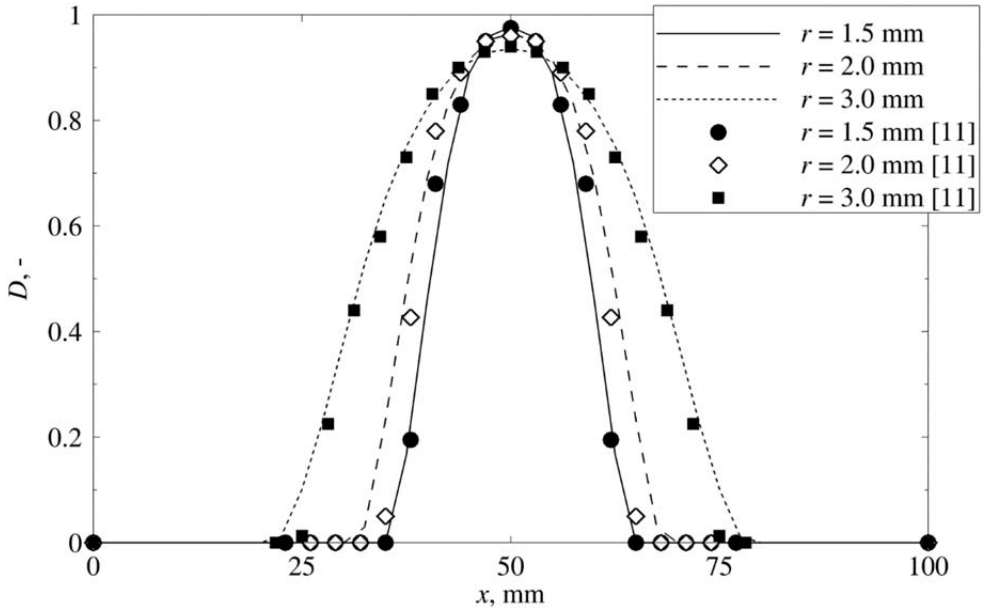



Figure 4: Comparison of damage profiles along the horizontal central axis obtained using the presented damage model with the results from the literature

As can be seen from Fig. 4, the calculated damage profiles show very good agreement with the solutions from the literature. The figure also illustrates the effect of the microstructural size on the behavior of the macroscale continuum. The increase in microstructural values leads to the expansion of the localization zone.

The evolution of the damage contour obtained with the microstructural parameter $l$ = 1.5 mm using the analytical expression of the constitutive tensors from [18], for the plate considered under imposed end displacement of $u$ = 0.0325 mm is presented in Fig. 5. For the same microstructural parameter and the displacement, the contour plot of the equivalent elastic strain $\varepsilon_{eqv}$ is displayed in Fig. 6.

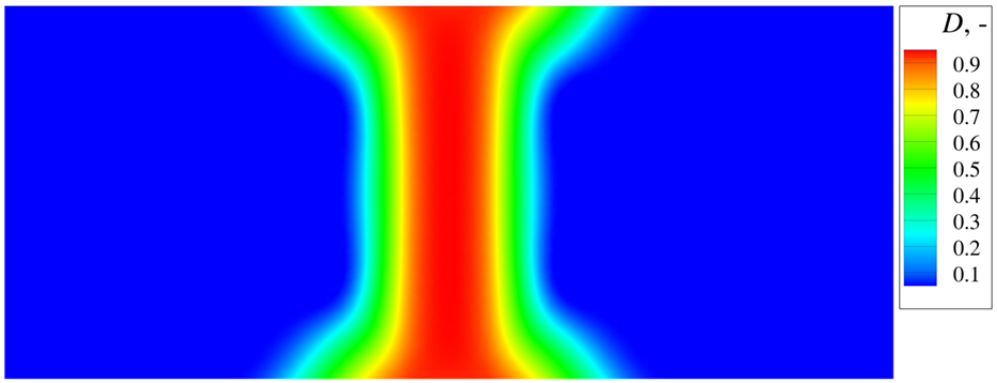



Figure 5: Distribution of damage $D$ for homogeneous material and internal length scale parameter $l$ = 1.5 mm at $u$ = 0.0325 mm

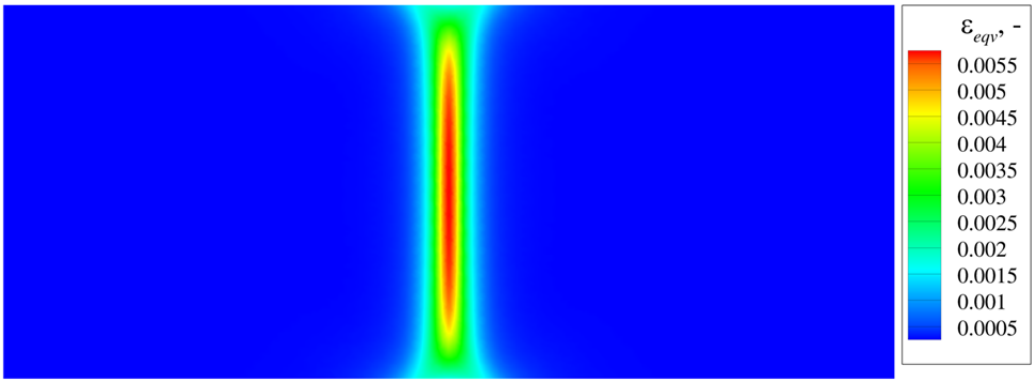



Figure 6: Distribution of equivalent elastic strain $\varepsilon_{eqv}$ for homogeneous material and internal length scale parameter $l$ = 1.5 mm at $u$ = 0.0325 mm

In the analysis of the heterogeneous material, the full form of the incremental constitutive relations represented by Eq. (8) has to be solved. Therefore, the second-order homogenization has to be applied in the preprocessing step to obtain the material stiffness tensors. Here, two cases of simple academic RVE examples are analyzed and compared to the corresponding homogeneous solutions, i.e. the solutions emerging from the corresponding homogeneous materials of the same internal length scales. The two RVEs considered have the same average hole radius $r_{ave}$ and porosity $e$, but different size $L$, as shown in Fig. 7.

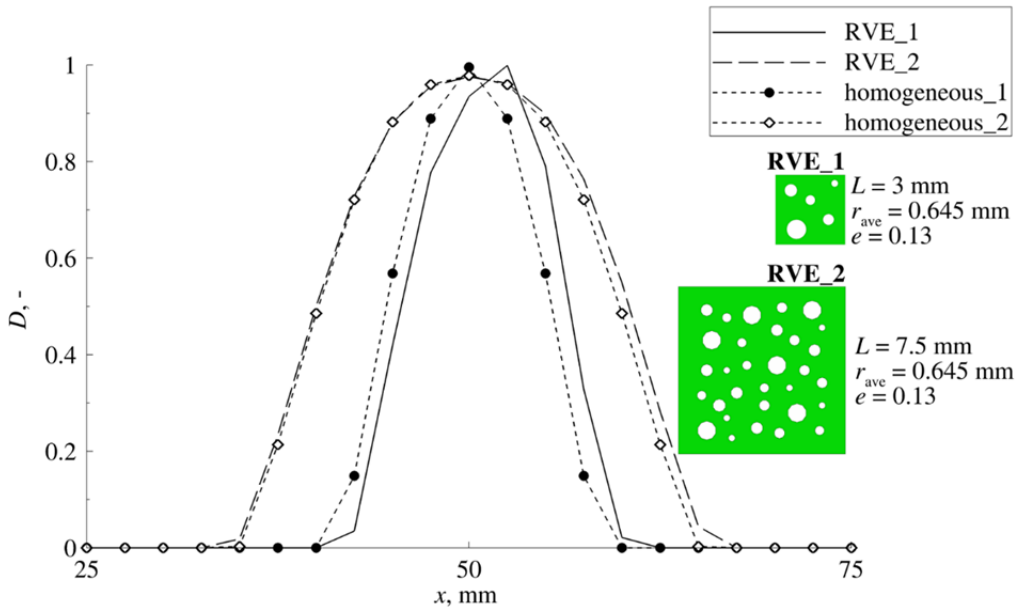



Figure 7: Comparison of damage profiles along the horizontal central axis obtained for two heterogeneous materials represented with RVEs of the same average hole radius $r_{ave}$ and porosity $e$, but different size $L$

In case of the smaller RVE, damage profile shows notable deviation from that obtained using the corresponding homogeneous material, while these differences are much smaller for the bigger RVE. This behavior can be explained with the positions of the holes as well as their distributions in the RVEs. These different damage responses can be confirmed by looking at the component values in the stiffness tensors $\mathbf{C}_{\sigma\eta}$ and $\mathbf{C}_{\mu\varepsilon}$, which are much bigger for the smaller RVE. It is to note that the influence of the RVE size and average hole radius on the

damage response of heterogeneous structures is one of the open questions and should be investigated in forthcoming research.

## 4 CONCLUSIONS

A new computational approach employing the full strain gradient theory for the modeling of quasi-brittle softening phenomena has been proposed. This model is based on the isotropic damage law so that the material tangent stiffness matrices appearing in the constitutive relations are pre-multiplied by the same term governing the damage process. The highly non-linear softening model is implemented into the triangular $C^1$ element using the FE software ABAQUS and provided UEL subroutines. The capabilities of the proposed computational strategy to simulate the strain localization has been demonstrated in a typical benchmark example consisting of a plate with an imperfect zone subjected to a tensile load. The verification of the algorithms derived has been made by the comparison with the solutions available in the literature, where the same problem has been analyzed using the EFG meshless method. Both homogeneous and heterogeneous material have been considered, where the latter is done by employing the second-order homogenization procedure to obtain the required material stiffness matrices. The homogeneous material has been analyzed employing the analytic expressions for the calculation of the stiffness matrices, and the softening state of the plate just prior to failure has been displayed by the contour plots of the damage variable and equivalent elastic strain field distributions. The heterogeneous material has been analyzed using the two different RVEs, where the damage responses have been compared to the homogeneous structures. It has been observed that the results depend on the RVE size and the number and position of the holes. The influence of heterogeneity on damage behavior requires more detailed studies which will be a topic of further research.

## ACKNOWLEDGMENT

The work has been fully supported by Croatian Science Foundation under the project "Multiscale Numerical Modeling of Material Deformation Responses from Macro- to Nanolevel" (2516).

## REFERENCES

[1] R.H.J. Peerlings, R. de Borst, W.A.M. Brekelmans, J.H.P. de Vree, Gradient enhanced damage for quasi-brittle materials, *International Journal for Numerical Methods in Engineering*, **39**, 3391-3403, 1996.

[2] R. de Borst, L.J. Sluys, H.B. Mühlhaus, J. Pamin, Fundamental issues in finite element analysis of localization of deformation, *Eng. Comput*., **10**, 99-121, 1993.

[3] C.S. Chang, L. Ma, Modeling of discrete granulates as micropolar continua, *Journal of Engineering Mechanics*, **116** (12), 2703-2721, 1990.

[4] L.J. Sluys, R. de Borst, Wave-propagation and localization in a rate dependent cracked medium model formulation and onedimensional examples, *International Journal of Solids and Structures*, **29**, 2945-2958, 1992.

[5] R.H.J. Peerlings, *Enhanced damage modeling for fracture and fatigue*, *Ph.D. Thesis*. Eindhoven University of Technology, The Netherlands, 1999.

[6] Z.P. Bažant, T. Belytschko, T.P. Chang, Continuum theory for strain-softening, *J.Eng. Mech.*, **110**, 1666–1692, 1984.

[7] R.H.J. Peerlings, R. de Borst, W. Brekelmans, M.G.D. Geers, Gradient-enhanced damage modeling of concrete fracture, *Mechanics of Cohesive-Frictional Materials*, **3**, 323–342,1998.

[8] R. de Borst, H.B. Mühlhaus, Gradient-dependent plasticity: formulation and algorithmic aspects, *Num. Int. J. Meth. Eng.*, **35**, 521–539, 1992.

[9] L.J. Sluys, R. de Borst, Wave-propagation and localization in a rate dependent cracked medium model formulation and onedimensional examples, *International Journal of Solids and Structures*, **29**, 2945-2958, 1992.

[10] C.S. Chang, H. Askes, L.J. Sluys, Higher-order strain/higher-order stress gradient models derived from a discrete microstructure, with application to fracture, *Engineering Fracture Mechanics*, **69**, 1907-1924, 2002.

[11] Y. Yang, A. Misra, Higher-Order Stress-Strain Theory for Damage Modeling Implemented in an Element-free Galerkin Formulation, *CMES: Computer Modeling in Engineering and Sciences*, **64**, 1-36, 2011.

[12] L. Kaczmarczyk, C.J. Pearce, N. Bićanić, Scale transition and enforcement of RVE boundary conditions in second-order computational homogenization, *Int J Numer Methods Eng*, **74**, 506–522, 2008.

[13] I.M. Gitman, H. Askes, L.J. Sluys, Representative volume: Existence and size determination, *Engng. Fract. Mech*., **74**, 2518–2534, 2007.

[14] T. Lesičar, Z. Tonković, J. Sorić, A second-order two-scale homogenization procedure using C-1 macrolevel discretization*, Computational Mechanics*, **54** (2), 425-41, 2014.

[15] F. Putar, T. Lesičar, J. Sorić, Z. Tonković, Damage modeling using higher-order finite element formulation. I. Kožar, N. Bićanić, G. Jelenić, M. Čanađija eds. *Proceedings of the 8th International Congress of Croatian Society of Mechanics*, CD edition. Opatija, Croatia, 2015.

[16] S. Akarapu, H.M. Zbib, Numerical analysis of plane cracks in strain-gradient elastic materials, *International Journal of Fracture*, **141** (3-4), 403-30.

[17] J. Lemaitre, *A Course on Damage Mechanics*, *2nd edition*. Springer, Berlin, 1996.

[18] V.G. Kouznetsova, M.G.D. Geers, W.A.M. Brekelmans, Size of a representative volume element in a second-order computational homogenization framework, *International Journal for Multiscale Computational Engineering*, **2** (4), 575-598, 2004.

[19] ABAQUS/Standard, in: Dassault Systemes, Providence, RI, USA, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# RELIABILITY ANALYSIS IN FRACTURE MECHANICS ACCORDING TO COMBINED FAILURE CRITERIA

**Rudy Chocat[*1,2,3], Paul Beaucaire[2], Loïc Debeugny[1], Jean-Pierre Lefebvre[2], Caroline Sainvitu[2], Piotr Breitkopf[3], and Eric Wyart[2]**

[1]Snecma, Division Moteurs Spatiaux
Forêt de Vernon BP802, 27208 Vernon FRANCE

[2]Cenaero
Rue des Frères Wright 29, 6041 Gosselies BELGIUM
e-mail: rudy.chocat@cenaero.fr

[3]Laboratoire Roberval UMR 7337 UTC-CNRS
Université de Technologie de Compiègne, 60200 Compiègne FRANCE

**Keywords:** reliability analysis, fracture mechanics, combined failure criteria.

**Abstract.** *The interest in reliability analysis related to Fracture Mechanics field grew during the past decades. This contribution proposes a reliability analysis combining two failure criteria: the crack stability and the plastic collapse by mean of the Failure Assessment Diagram (FAD). FAD and reliability method are coupled to assess the probability of failure. Some issues have been highlighted considering classic gradient-based optimization methods to determine key points for approximation method. A strategy based on surrogate-based evolutionary algorithms is proposed to solve these issues. A discussion on the most influencing parameters and their consequences in terms of reliability based design is also presented. The application of the methodology is demonstrated on a simple test case.*

# 1 INTRODUCTION

A frequent way to treat uncertainties is to consider the worse conditions and run deterministic analysis. Nevertheless, this approach can generate over-sizing and the impact on the conservatism is generally not known. Consequently, reliability analysis tools are developed to assess the probability of failure of a component considering the uncertainties to estimate the level of conservatism.

A typical Fracture Mechanics problem is subject to uncertainties related to the geometry, material properties, loads or crack location and shape. Consequently, Fracture Mechanics is one the main applications for the reliability analysis [1, 2]. To assess the probability of failure in Fracture Mechanics issues [3], two main approaches have been developed, one based on the simulation (*Monte Carlo method*) and another based on the approximation. [4] consider a failure criterion linking the Fracture toughness and Stress Intensity Factors (SIF). Nevertheless, it is necessary to consider another failure criterion to include the plastic collapse to account for the remaining strength of the structure [5]. In [6], a complete reliability study considering these two criteria is presented using *Monte Carlo method* and *First Order Reliability Methods* (*FORM*). They consider the different failure modes associated with each criterion independently and propose a performance function based on the number of cycles to failure. Another approach [7] is to combine both failure criteria using the Failure Assessment Diagram (FAD). A reliability and sensitivity analysis based on this approach is proposed in [8]. They focus the performance function on the time inspection and perform *Monte Carlo* and *FORM* analysis. The present contribution presents reliability and sensibility analysis considering a FAD margin as performance function. Some difficulties are highlighted for *FORM* computation as a classical gradient-based optimisation fails to converge, this contribution proposes to solve the optimization problem of *FORM* using efficient surrogate-based evolutionary algorithm. Our approach allows to evaluate the *FORM* approximation for black box performance function because its gradient is not necessary. In addition to reliability analysis, a complete sensitivity of the FAD margin performance function is performed on a test case.

This paper is structured as follows. The first section introduces the main methods of reliability analysis. The second section reviews the main sensitivity indices and provides recommendations for a good understanding of the reliability problem. The third section presents the FAD margin performance function and the test case. It also includes the enhancement of the *FORM* optimization algorithm and a complete sensibility analysis of the test case. Finally, the last section concludes the and discusses future work.

# 2 APPLICATION TO FRACTURE MECHANICS

## 2.1 Probability of failure

Let us consider $\mathbf{X} = (X_1, \ldots, X_d)$ as a vector of $d$ random variables $X_i$ defined by respective probability density functions $f_{X_i}$. Each random variable $X_i$ is defined by distribution law characterised by a vector of $d_\theta$ parameters $\boldsymbol{\theta}_i = \{\theta_{ik}\},\ k = 1, \ldots, d_\theta$. Realizations of random variable $X$ is denoted by $x$ and a realization vector by $\mathbf{x} = x_1, \ldots, x_d$.

A performance function $G(\mathbf{X})$ is associated to the failure modes. This function allows splitting the space into two parts: the safe space where $G(\mathbf{X}) > 0$ and the failure space defined as $G(\mathbf{X}) \leq 0$. The border is the so-called limit state function and characterised as $G(\mathbf{X}) = 0$. The aim of reliability analysis is to compute the probability that an event is in the failure space as $P_f = \text{Prob}(G(\mathbf{X}) \leq 0)$. A structural reliability problem is solved using iso-probabilistic transformation that switches the problem from the physics space $\mathbb{X}^d$ to the standard space $\mathbb{U}^d$

defined by uncorrelated standard normal random variables. $P_f$ becomes:

$$P_f = \int_{G(\mathbf{X})\leq 0} f_{\mathbf{X}}(\mathbf{x})\mathrm{d}\mathbf{x} = \int_{H(\mathbf{U})\leq 0} \phi_d(\mathbf{u})\mathrm{d}\mathbf{u}, \tag{1}$$

where $f_{\mathbf{X}}$ is the joint probability density of $X_i, i = 1, \ldots, d$, $\mathbf{U}$ are uncorrelated standard normal random variables, $\phi_d$ is the $d$-variate uncorrelated normal probability density function and $H(\mathbf{U})$ is the limit state function in the standard space.

### 2.2 Combination of failure criteria

The fundamental principle of Fracture Mechanics is that the stress field ahead of a sharp crack in a test specimen or a structural member can be characterized by a single parameter, $K$, which is the stress intensity factor (SIF). One of the main criteria in Fracture Mechanics relies on the SIF, it must not exceed the toughness $K_{1C}$ which is a material property. It is expressed using $K_r$ as:

$$K_r = \frac{K}{K_{1C}} < 1. \tag{2}$$

Another failure mode corresponds to the plastic collapse of the remaining ligament. A failure criterion $L_r$ is developed in order to measure the capacity of the component to sustain the mechanical load:

$$L_r = \frac{\sigma_n}{\sigma_{p0.2}} \leq \frac{\sigma_{p0.2} + \sigma_m}{2\sigma_{p0.2}}, \tag{3}$$

where $\sigma_n$ is the nominal stress in the remaining ligament, $\sigma_{p0.2}$ and $\sigma_m$ are respectively the yield and the ultimate strength.

The combination of both criteria is essential to assess the failure probability in Fracture Mechanics. Both criteria are combined by to the Failure Assessment Diagram (FAD) in $K_r - L_r$ space. In this diagram, the limit between safe and failure domains is defined by R6-rule [9] as a function of parameters by:

$$g(K_r, L_r) = K_r - \begin{cases} \frac{0.3+0.7\exp(-\mu L_r^6)}{\sqrt{1+0.5L_r^2}} & \text{for} \quad L_r \leq 1 \\ \frac{0.3+0.7\exp(-\mu)}{\sqrt{1.5}} L_r^{\frac{n_r-1}{2n_r}} & \text{for} \quad 1 < L_r < L_r^{\max} \\ 0 & \text{for} \quad L_r \geq L_r^{\max} \end{cases}, \quad \text{with:} \tag{4}$$

$$\mu = \min\{10^{-3}\frac{E}{\sigma_{p0.2}}, 0.6\}, \quad n_r = 0.3(1 - \frac{\sigma_{p0.2}}{\sigma_m}), \quad L_r^{\max} = \frac{\sigma_{p0.2} + \sigma_m}{2\sigma_{p0.2}},$$

where $E$ is the Young modulus. A strategy to evaluate the failure of structure in Fracture Mechanics is to determine $K_r$ and $L_r$ to define a point in the FAD diagram and to compute a function of the distance between this point and the R6-rule boundary [7]. This function is the margin, if it is negative it means a failure of the structure, and if it is positive, the structure is safe. As shown in Figure 1, the margin is function of the ratio $\frac{|\mathrm{OB}|}{|\mathrm{OA}|}$. The margin can be defined as the performance function $G(\mathbf{x})$ because it allows to combine the two failure criteria and $G(\mathbf{x})$ can be expressed as:

$$G(\mathbf{x}) = \left(\frac{|\mathrm{OB}|}{|\mathrm{OA}|} - 1\right) \begin{cases} > 0, & \text{safe} \\ \leq 0, & \text{failure} \end{cases}. \tag{5}$$

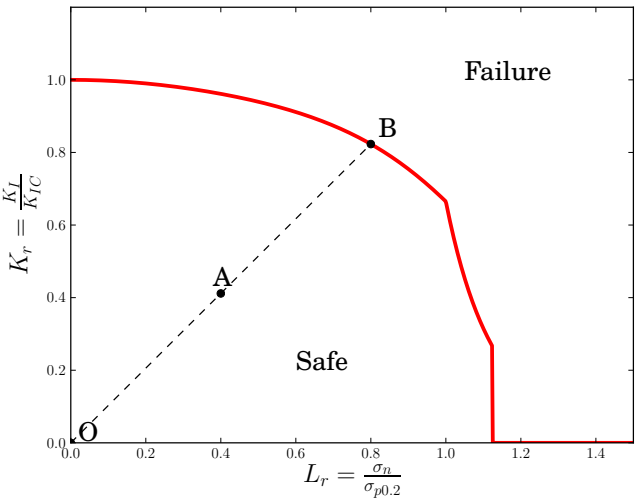



Figure 1: Failure Assessment Diagram: R6-rule and key points to evaluate the performance function.

### 2.3 Model description

A simple test case is defined. It consists in a crack in a beam (Figure 2) subjected to tensile force. The crack propagates according to the Paris law expressed by:

$$\frac{\mathrm{d}a}{\mathrm{d}N} = C(\Delta K)^m, \tag{6}$$

where $a$ is the size of the crack, $N$ is the number of cycles, $K$ is the SIF, $C$ and $m$ are crack propagation law parameters.

As shown by Virkler [10], the phenomenon of the crack propagation can be considered as a stochastic process. Indeed, considering sample of the same coupon test, experimental results show that, for a given initial crack, the speed of the crack propagation is different. A common way to integrate these uncertainties is to consider parameters of crack propagation law $C$ and $m$ as random variables. There is a strong correlation ratio between these parameters [11], which can not be independently sampled. Then, it is possible to consider $m$ as random variable and propagate the uncertainty to $C$ according to a linear relationship:

$$\ln C = \xi_1 m + \xi_0, \tag{7}$$

where $\xi_0$ and $\xi_1$ are regression coefficients identified according to experimental test results.

As presented in Table 1, eleven input parameters are set as random variables defined by classical distribution laws according to physical considerations. For the application, values of parameters are invented but representative of the reality.

### 2.4 FORM

The *Fist Order Reliability Method* (*FORM*) [12, 13] is an approximation method based on the *Hasofer Lind reliability index* [14] to compute the probability of failure. This index $\beta$ is a geometric measure of the reliability in relationship with the failure probability as $p_f = \Phi(-\beta)$. It is the shortest distance between the origin and the limit state in the standard space and is

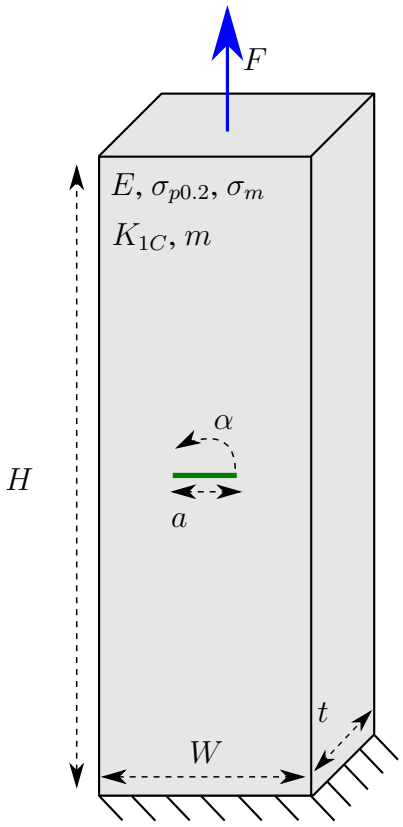



Figure 2: Test case geometry with parameters defined as random variables.

so-called the Most Probable Point (MPP) $P^*$. The coordinates $\mathbf{u}^*$ of $P^*$ are obtained by solving the following constrained optimization problem.

$$\begin{aligned} \mathbf{u}^* &= \arg\min \|\mathbf{u}\| \\ \text{s.t.} \quad & H(\mathbf{u}^*) \leq 0. \end{aligned} \tag{8}$$

This problem is commonly solved using the Rackwitz-Flessler gradient-based algorithm [15]. *FORM* implies a linearisation of the limit state in the vicinity of MPP. An approximation of the probability of failure is expressed as:

$$P_f \approx P(\underbrace{H(\mathbf{u}^*)}_{=0} + \nabla_{\mathbf{u}} H(\mathbf{u}^*)(\mathbf{u} - \mathbf{u}^*) \leq 0) = \Phi(-\beta_{\text{FORM}}), \quad \text{with: } \beta_{\text{FORM}} = \|\mathbf{u}^*\|. \tag{9}$$

Additional information about relative importance of input variables $u_i$ is provided by cosinus director $\alpha_i$ of the MPP in the $U$-space:

$$\alpha_i = -\frac{\partial \beta}{\partial u_i}|_{\{u^*\}}, \tag{10}$$

so $\alpha_i$ is the derivative of the reliability index according to variables $u$. This information is used for sensitivity analyses developed in the next section.

*FORM* is a fast and efficient method to obtain the failure probability. Moreover, it provides useful sensitivity information. Nevertheless, because it is based on an approximation, the provided value has to be checked using another reliable method such as simulations-based *Monte*

| Class | Variable | Distribution law | Description |
|---|---|---|---|
| Geometry | $L$ | $\mathcal{U}[299.9\text{ mm}; 300.1\text{ mm}]$ | Height of the beam |
| | $W$ | $\mathcal{U}[19.9\text{ mm}; 20.1\text{ mm}]$ | Width of the beam |
| | $t$ | $\mathcal{U}[9.9\text{ mm}; 10.1\text{ mm}]$ | Thickness of the beam |
| Material properties | $E$ | $\mathcal{N}(185000\text{ MPa}; 9250\text{ mm})$ | Young modulus |
| | $\sigma_{p0.2}$ | $\mathcal{N}(600\text{ MPa}; 6\text{ MPa})$ | Yield strength |
| | $\sigma_m$ | $\mathcal{N}(900\text{ MPa}; 9\text{ Mpa})$ | Shear strength |
| | $K_{1C}$ | $\mathcal{N}(200\text{ MPa}\sqrt{\text{mm}}; 2\text{MPa}\sqrt{\text{mm}})$ | Fracture toughness |
| | $m$ | $\mathcal{N}(3; 0.3)$ | Paris law parameter |
| Loads | $F$ | $\mathcal{N}(25\text{ kN}; 2.5\text{ kN})$ | Effort intensity |
| Crack | $a$ | $\mathcal{U}[0\text{ mm}; 0.5\text{ mm}]$ | Initial crack size |
| | $\alpha$ | $\mathcal{N}(0°; 1°)$ | Crack orientation |

Table 1: List of random variables with corresponding distribution laws.

*Carlo method*. The approximation can be enhanced using the Second Order Reliability Method [13] *SORM* which includes the main curvature of the limit state function.

### 2.5 Application of *FORM* with Surrogate-Based Optimization

*FORM* relies on an optimization problem to get the Most Probable Point (MPP) which is commonly solved using Rackwitz-Flessler gradient-based optimization algorithm. It has been applied on our application case but it fails to determine the MPP. Similar issues, such as slow convergence or no convergence, are also observed in [8]. Several starting points have been tested, it is observed that after few iterations, the gradient of performance function becomes close to zero and it breaks the algorithm. Null gradient underlines a saturation phenomenon of the performance function. This phenomenon reveals a structural failure before the end of the life cycle. Indeed, the crack grows and its size can become bigger than the beam size. In this case, the margin in the FAD stays the same, even if input parameters are different. Figure 3 shows the evolution of the performance function in the standard space. In order to plot it in a visual space, only two significant random variables are considered: the length of the crack $a$ and the crack propagation parameter $m$. The phenomenon of saturation can be observed in the failure space. After few iterations, the optimal point is located in the saturation domain and this stops the algorithm.

Gradient-based optimization methods are not relevant in our case. Hence, it has been proposed to use genetic algorithm to determine the MPP. Actually, genetic algorithms do not need gradient information to converge to the optimal point. Nevertheless, basic genetic algorithms need a lot of simulations to converge [16]. Then, the interest to use *FORM* in comparison with Monte Carlo can be compromised. In this contribution, a strategy based on a Surrogate-Based Optimization (SBO) to get the MPP using the in-house optimization platform Minamo [17] is built according to the following steps :

1. Firstly, a global objective function is built coupling the objective function (here the distance between the the origin and the Most Probable Point) and the inequality constraint function (here the performance function must be less or equal to zero)

2. Then, a Design Of Experiments is generated using the *Latin Hypercube Sampling* method. It allows to optimally sample the input space on the domain of definition of input parame-

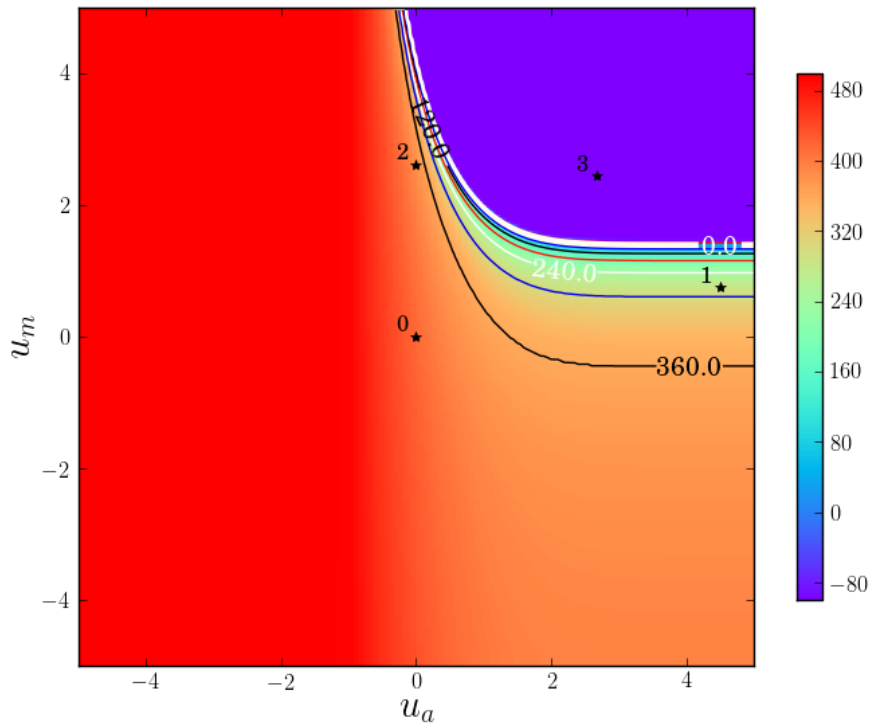



Figure 3: Evolution of the performance function in the standard space $u_a$, $u_m$. The blue part is the failure domain. Gradient in this domain is close to zero. The star points represent the iterations of FORM for the gradient-based optimization. It stops when the point is located in the saturation domain.

ters. Individuals of DOE are so-called training points. In reliability problem, the sample is generated in the standard space and each individual is mapped into the physical space by the inverse of the iso-transformation to compute the performance function at each point of the DOE.

3. A Radial Basis Function (RBF) surrogate of the global objective function in the standard space is built on these training points. The fidelity of the surrogate depends of the size of the DOE.

4. An optimal point is determined based on the surrogate model. The actual objective and constraint function are assessed on this point.

5. Optimization criteria which are typically the maximum number of iterations or the distance between subsequent approximate optimal points, are checked:

   - If criteria are satisfied, it is the end of the optimization process.
   - If they are not. The DOE is enriched by this point and then we return to the third step to build a surrogate with more training points.

This procedure is commonly named *Adaptive sampling*. At each iteration, new points enrich the DOE to refine the surrogate in interesting part of the domain. Therefore, regardless satu-

ration phenomenon, it allows to determine the MPP without performing a lot of simulations. *FORM* approximation can be applied to compute reliability information.

Table 2 presents results of the reliability analysis using *Monte Carlo method* and *FORM with SBO* comparing the probability of failure and the number of simulations. Even if there is a ratio about two between the *FORM* probability and the *Monte Carlo* one, both are in the same order of magnitude. *FORM* approximation is not exact but only about hundred simulations are necessary. A compromise has to be determined between accuracy and speed. In our case, *FORM* seems to be adapted because the scale of the probability is kept with a small number of simulations. Moreover, *FORM* provides interesting sensibility information as explained later.

| Method | $P_f$ | Simulations |
|---|---|---|
| Monte Carlo | $0.02505$ | $100000$ ($\mathrm{cv}_{\hat{P_f}} \approx 2\%$) |
| *FORM with SBO* | $0.03860$ | $115$ (LHS $40$) |

Table 2: Results of reliability methods on the test case.

## 3 Sensitivity analysis

Reliability analysis aims at getting the probability of failure. Nevertheless, it is a fuzzy information because experimental comparison are hard to establish. Therefore, sensibility of the input variables is an interesting point as it defines the significant variables that impact the probability of failure. Then, this contribution presents the main sensitivity indices useful for reliability analysis. Firstly, sensitivity analysis of the performance function is performed. Secondly, reliability sensitivities are provided.

### 3.1 Mechanical sensitivities

Mechanical sensitivities aims at characterising the impact of inputs on the performance function. Even if *screening* and *local sensitivity* are available, mechanical sensitivities are generally associated with global sensitivity providing accurate information about the influence of input variabilities on the whole domain. Based on the *ANalysis Of VAriance* (*ANOVA*) decomposition of model output, the *Sobol method* [18] is a global and model independent sensitivity analysis method that provides indices to measure the impact of the input variabilities on the output variance. First Sobol indices quantify the impact of an input variable without taking into account the interactions with other variable ans the total Sobol indices provide the global impact of a variable. Nevertheless, these indices are only focused on the variance output and can not predict a change of the mean or other statistical moments of the output. Therefore, Borgonovo [19] proposes indices based on the Probability Density Function (PDF) output. They measure the change area of the density output and allow to consider other statistical moments. To completely characterize the global impact of a variable, both indices are complementary.

Global sensitivity analysis of the performance function is achieved. Sobol first order, Sobol total and Borgonovo indices are assessed and shown in Figure 4. On one hand, the variance-based sensibility underlines three significant variables which the length of the crack $a$, the crack propagation parameter $m$ and the load $F$. It sounds in line with the definition of the test. The difference of value between the two Sobol indices can be explained by the fact that there is interaction between these three significant variables. On the other hand, density-based indices show significant indices for ten variables : The three main variables of the variance-based ap-

proach have still impact but geometric parameters and material properties have also significant indices. These last parameters impact the output density mean or skewness but they have a very small influence on the output variance. Mechanical sensitivities show that the variable $H$ has no influence on the performance. It is in line with the definition of the test case since it accounts for semi-infinite geometry.

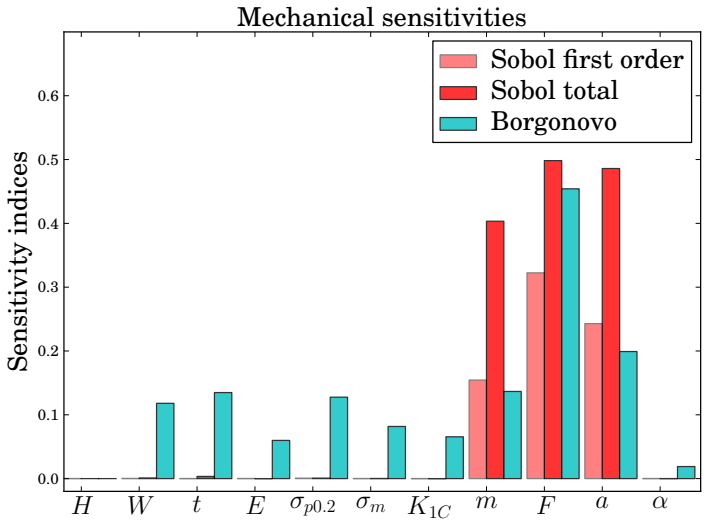



Figure 4: Mechanical sensitivities of the performance function which the margin in the FAD.

### 3.2 Reliability sensitivities

In addition to provide the probability of failure, *FORM* gives some information about the reliability sensibility. Firstly, as random variables are independent, the importance factors $\alpha_i^2$ (Figure 5) describe the weight of each variable on the reliability. As for the variance-based mechanical sensibility, there are three significant variables : $a$, $m$ and $F$. These results allow to determine, by some basic transformations[20], the reliability sensitivities of the inputs distribution parameters of random laws,

$$\frac{\partial \beta_{\text{FORM}}}{\partial \theta_{ik}} = \frac{\partial \beta_{\text{FORM}}}{\partial u_i} \frac{\partial u_i k}{\partial \theta_{ik}} = -\alpha_i \frac{\partial \mathrm{T}_{x \to u}(x_i, \boldsymbol{\theta}_i)}{\partial \theta_{ik}}, \tag{11}$$

where $\mathrm{T}_{x \to u}(x_i, \theta_f) = \Phi^{-1}(F_{X_i}(x_i, \theta_f))$ is the iso-probabilistic transformation. Considering the *FORM* relationship between $P_f$ and $\beta_{\text{FORM}}$, the sensibility of $P_f$ according to the parameters $\theta_{ik}$ is defined as:

$$\frac{\partial P_f}{\partial \theta_{ik}} \approx -\phi(-\beta_{\text{FORM}}) \frac{\partial \beta_{\text{FORM}}}{\partial \theta_{ik}}. \tag{12}$$

Reliability sensitivities of the test case (Figure 6) underline two significant variables which are $a$ and $m$. The parameters of the load distribution seem to not impact the reliability. However, reliability sensitivities quantify the impact of the variation of one unity of the input parameter on the reliability. They are not normalised, then, the comparison of the influence of variation is not obvious. Thus, it is also interesting to compare the influence of input variables quantifying

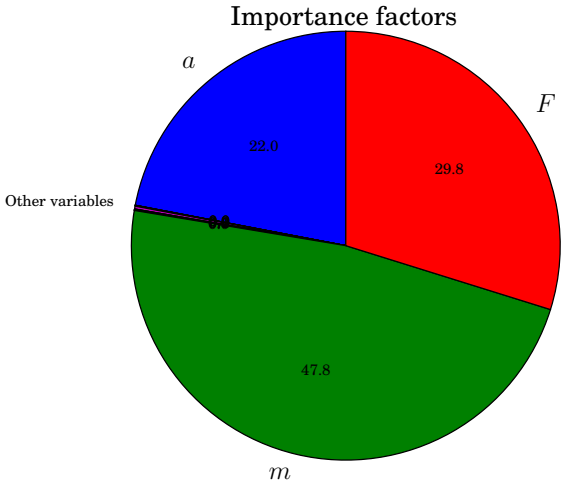



Figure 5: Influence of the parameters on the failure space.

the impact of one percent of variation. This is commonly named *elasticities* and is expressed by:

$$e_{\theta_{ik}} = \frac{\theta_{ik}}{\beta_{\text{FORM}}} \frac{\partial \beta_{\text{FORM}}}{\partial \theta_{ik}}. \tag{13}$$

According to the comparison of elasticities (Figure 7), $F$, $a$ and $m$ are the most significant variables as for the variance-based sensibility and for the importance factors. It can be due to the relatively small influence of other parameters. Indeed, results can be different if other parameters impact more the variance of the performance function. Moreover, $F$, $a$ and $m$ reliability sensitivities are positive, it implies that the reliability decreases when the mean parameters of these variables increase.

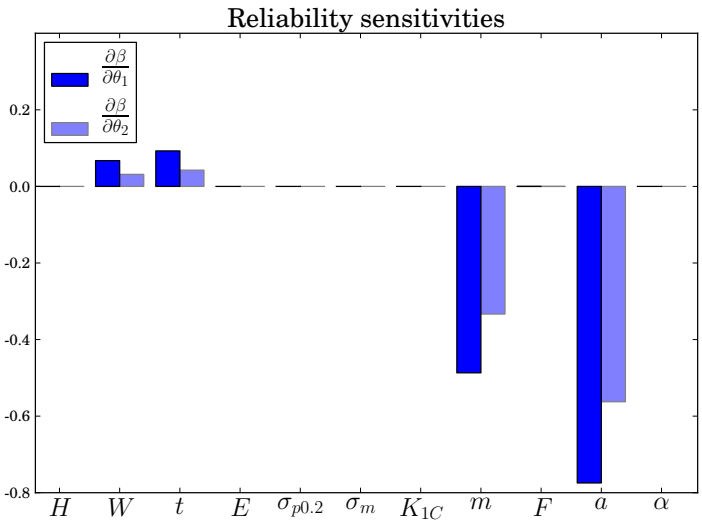



Figure 6: Reliability sensitivities of distribution parameters.

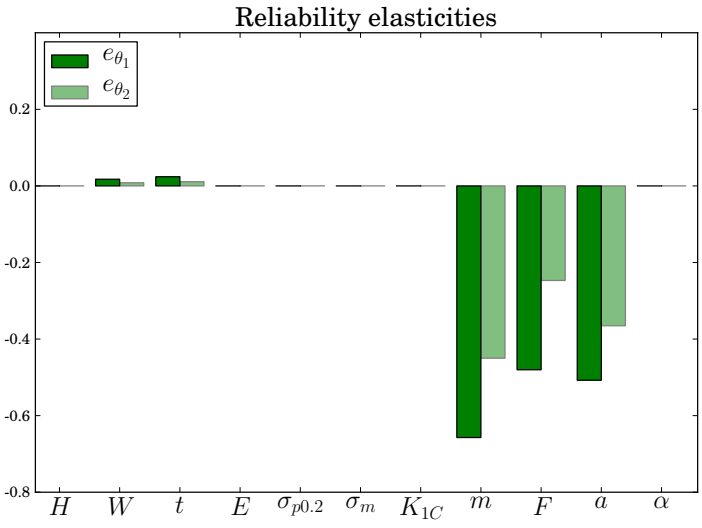



Figure 7: Reliability elasticities of distribution parameters.

To summarise, three parameters have significant influence on the variance performance function and on the reliability:

- the length of the crack $a$,
- the crack propagation parameter $m$,
- the loads $F$.

It can be explained by the fact that these parameters directly drive the crack propagation. In order to run accurate reliability analysis, specific effort has to be made to carefully fit the random laws of these parameters.

## 4 CONCLUSIONS

This contribution presents a reliability and sensibility analysis of a typical Fracture Mechanics problem which combines two failure criteria. Both are combined in the performance function using a margin function in the FAD diagram. The classical gradient-based algorithm used for the *FORM* reliability method failed to determine the Most Probable Point which is necessary to apply this method. It is due to a phenomenon of saturation of the performance function. This contribution fixes the problem using an efficient Surrogate-Based Optimization algorithm to allow to determine key information to apply the *FORM* method with a limited number of simulations. A complete sensibility analysis is performed and it underlines the significant impact other random variables as the length of the crack $a$, the Paris law crack propagation parameter $m$ and the loads $F$. Therefore, a particular consideration must be given to the modelling of these random variables.

The *FORM* method is applied here using a Surrogate-Based Optimization algorithm. Nevertheless, simulation methods can also be used using surrogate model to reduce the number of simulations. Moreover, in the surrogate enhancement, the phenomenon of saturation of the performance function should be taken into account.

## ACKNOWLEDGEMENT

The work presented in this paper is financially supported by CNES and ANRT which are gratefully acknowledged.

## REFERENCES

[1] C.R. Sundararajan, *Probabilistic structural mechanics handbook: theory and industrial applications*. Springer Science & Business Media, 2012.

[2] J.P. Lefebvre, B. Dompierre, A. Robert, M. Le Bihan, E. Wyart, C. Sainvitu, *Failure probability assessment using co-Kriging surrogate models* 6th Fatigue Design conference, Fatigue Design, 2015

[3] G.C. Sih, J.W. Provan, *Probabilistic fracture mechanics and reliability*. 2013

[4] W. K. Liu, Y. Chen, T. Belytschko Y. J. Lua, Engineering fracture mechanics *Three reliability methods for fatigue crack growth, Vol. 53, 5th Edition*, 733–752 , 1996.

[5] A.R. Dowling, C.H.A. Townley The effect of defects on structural failure: a two-criteria approach. *International Journal of Pressure Vessels and Piping*, **3** 77–107,1975

[6] A. Altamura, D. Straub, Reliability assessment of high cycle fatigue under variable amplitude loading: Review and solutions. *Engineering Fracture Mechanics*, **121**, 40–66, 2014.

[7] I. Milne, R.A. Ainsworth, A.R. Dowling, A.T. Stewart, Assessment of the integrity of structures containing defects. *International Journal of Pressure Vessels and Piping*, **32**, 3–104, 1988.

[8] G. Walz, H Riesch-Oppermann, Probabilistic fracture mechanics assessment of flaws in turbine disks including quality assurance procedures. *Structural Safety*, **28**, 273–288, 2006.

[9] British Energy Generation Limited, Assessment of the integrity of structures containing defects. *R6 Revision 4*, Amendment 7, 2003

[10] D.A. Virkler, B.M. Hillberry, P.K. Goel, The statistical nature of fatigue crack propagation. *Journal of Engineering Materials and Technology*, **101**, 148–153, 1979.

[11] M.B. Cortie, G.G. Garrett, On the correlation between the C and m in the Paris equation for fatigue crack propagation. *Engineering fracture mechanics*, **30**, 49–58, 1988.

[12] M. Lemaire, *Structural Reliability*. ISTE/Wiley, 2009.

[13] O.D. Ditlevsen, H.O. Madsen, *Structural Reliability Methods*. John Wiley Sons Inc, 1996.

[14] A.M. Hasofer, N.C. Lind, Exact and invariant second-moment code format. *Journal of the Engineering Mechanics division*, **100**, 111–121, 1974.

[15] R. Rackwitz, B. Flessler, Structural reliability under combined random load sequences. *Computers & Structures*, **9**, 489–494, 1978.

[16] R. Rackwitz, Reliability analysis - a review and some perspectives. *Structural safety*, **23** 365–395, 2001.

[17] C. Sainvitu, V. Iliopoulou, I. Lepot, Global Optimization with Expensive Functions-Sample Turbomachinery Design Application. *Recent Advances in Optimization and its Applications in Engineering*, 499–509, 2010.

[18] I.M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and computers in simulation*, **55**, 271–280, 2001.

[19] E. Borgonovo, A new uncertainty importance measure. *Reliability Engineering & System Safety*, **92**, 771–784, 2007.

[20] A. Karamchandani, C.A. Cornell, Sensitivity estimation within first and second order reliability methods. *Structural Safety*, **11**, 95–107, 1992.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NEW BENCHMARK FOR THE LIFE ASSESSMENT OF A THIN-WALLED PIPE SUBJECTED TO STRESS ASSISTED CORROSION

**Yulia G. Pronina**[1], **Elena D. Sedova**[1]

[1]Saint Petersburg State University
7/9 Universitetskaya nab., St. Petersburg, 199034 Russia
e-mail: y.pronina@spbu.ru

**Keywords:** Mechanochemical Corrosion, Thin-Walled Tube, High Pressure Vessel, Analytical Solution.

**Abstract.** *A thin-walled tube subjected to internal and external pressure under the conditions of double-sided mechanochemical corrosion is considered. The rates of corrosion are supposed to be linearly dependent on the maximum principal stress. Previous solutions for thin cylindrical shells obtained by other authors reflect only the effect of differential pressure (i.e., the difference between internal and external pressure) and do not depend on the internal and external pressure values themselves. The model proposed here allows to take into account the effect of hydrostatic pressure on the durability of the tube exposed to corrosive environment, but it is not so cumbersome as the accurate solution for the double-sided corrosion of a pressurized tube based on the solution of Lamé problem for a thick-walled cylinder. The problem is reduced to an ordinary differential equation of the first order, which is solved in a closed form. It has been observed that the computational results for the classical thin shell model and for the proposed here refined model are very close to each other when either internal or external pressure is equal to zero. But the difference between the mentioned results grows as the minimum of the pressure values increases.*

## 1 INTRODUCTION

Corrosion is a process of material destruction that occurs as a result of chemical interaction between a material and its environment. General corrosion is one of the most common types of corrosion, that damages the entire surface of the material at about the same rate, and as a consequence, results in a failure. If a structural element, as well as being in aggressive environment, is subjected to mechanical load, then the mechanochemical effect may occur when corrosion is enhanced by stress [1, 2]. In such situations for the assessment of the durability of structural elements, initial boundary value problems with evolving boundaries have to be considered. Most of them are solved by numerical methods. However, it is reasonable to have closed-form benchmarks for the verification of numerical techniques. A number of closed-form solutions have been found for the cases of uniform corrosion, e.g. by the authors of [3, 4, 5, 6, 7, 8, 9].

This paper presents an analytical solution for a thin-walled cylindrical tube subjected to internal and external pressure under the conditions of double-sided mechanochemical dissolution. The rates of corrosion are supposed to be linearly dependent on the maximum principal stress [1]. There is a lot of works devoted to thin-walled structures [3, 4, 5, 6]. However, previous solutions for thin shells obtained by other authors, do not depend on the internal and external pressures themselves, but depend only on their difference. Therefore, they do not reflect the effect of hydrostatic pressure (i.e. the minimum of the internal and external pressures) on lifetime of the vessel. At the same time, the known accurate solutions for double-sided corrosion of elastic tube under internal and external pressure based on the solution of Lamé problem are rather cumbersome [10, 11]. The model proposed here is not so complicated as these accurate solutions, but it allows to take into account the effect of hydrostatic pressure on the durability of thin shells under the mechanochemical corrosion conditions. The same questions for spherical shells were investigated in [12, 13, 14, 15]. The problem of calculation of the optimal initial thickness of a spherical vessel operating in mechanochemical corrosion conditions was solved in [16]. The results can be taken into account for the lifetime assessment of thin-walled structural elements subjected to the combined action of mechanical loads and an aggressive environment [17, 18, 19, 20, 21].

## 2 THE DESCRIPTION OF THE PROBLEM

Consider a linearly elastic long thin-walled cylindrical tube with the inner $r$ and outer $R$ radii subjected to internal $p_r$ and external $p_R$ pressure which are given constant. The article covers the following cases: an incompressible tube subjected to the plane-strain condition characterized by vanishing axial strain and a closed cylindrical vessel. The action of the ends of the cylinder is not taken into account. Cylinder is uniformly corroding from the inside and outside with the dissolution rates $v_r$ and $v_R$, respectively. Change of the tube's thickness is assumed to be quasi-static and uniform throughout the entire length: the inner radius increases while the outer radius decreases with time $t$. Let the tube's radii at the initial time $t_0$ be denoted by $r_0$ and $R_0$.

According to [1, 3], the rates of corrosion on the inner and outer surfaces are:

$$v_r = \frac{dr}{dt} = a_r + m_r\sigma_1(r), \qquad v_R = -\frac{dR}{dt} = a_R + m_R\sigma_1(R), \tag{1}$$

where $a_r,\ a_R,\ m_r$, and $m_R$ are experimentally determined constants, which, in general, are different for tension and compression [1, 7]; $\sigma_1$ is the maximum principal stress at the corresponding surface of the tube; $\operatorname{sign} m = \operatorname{sign} \sigma_1$.

Consider a cylindrical coordinate system $(\rho, \theta, z)$ with the $z$-axis coincide with the axis of

the tube. In the previous solutions for thin-walled tubes obtained by other authors (for example [4]), $\sigma_1$ was chosen equal through the thickness of the shell:

$$\sigma_1(r) = \sigma_1(R) = \sigma_{\theta\theta} = \frac{\Delta p R_c}{h}, \tag{2}$$

where $R_c = (R_0 + r_0)/2$ is the radius of the middle surface of the cylindrical tube (which is supposed to be constant during the corrosion), $h = R - r$ is the thickness of the tube, and $\Delta p = p_r - p_R$.

It's obvious that the solution derived by the use of equation (2) does not depend on the pressure values $p_r$ and $p_R$, but depends only on the difference $\Delta p$. However, it is known that hydrostatic pressure $p = \min\{p_r, p_R\}$ can affect corrosion kinetics [1].

The purpose of this paper is to derive a new (more accurate, but not more complicated than previous) solution which allow to take into account the effect of hydrostatic pressure $p$ on the assessment of the durability of thin-walled tubes subjected to mechanochemical corrosion.

## 3 THE SOLUTION OF THE PROBLEM

### 3.1 Effective stress

According to the G. Lame's formulas for a thick-walled tube under internal and external pressure, maximum principal stress at every time $t$ is determined as

$$\sigma_{\theta\theta}(\rho) = \frac{p_r r^2 - p_R R^2}{R^2 - r^2} + \frac{(p_r - p_R)}{R^2 - r^2}\frac{r^2 R^2}{\rho^2}, \qquad r \le \rho \le R. \tag{3}$$

Solution to the problem of mechanochemical corrosion obtained by the use of formula (3) is rather cumbersome [11]. Let us simplify formula (3) taking into account that for a thin-walled shell we can assume that $h/R_c << 1$. Using other notations: $r = R_c - \delta$ and $R = R_c + \delta$ (where $\delta = h/2$), we can write

$$\begin{aligned}\sigma_{\theta\theta}(r) &= \frac{p_r(2R_c^2 + 2\delta^2) - 2p_R(R_c^2 + 2R_c\delta + \delta^2)}{4R_c\delta} \\ &= \frac{p_r\,(1 + \delta^2/R_c^2) - p_R\,(1 + 2\delta/R_c + \delta^2/R_c^2)}{2\delta/R_c},\end{aligned} \tag{4}$$

and

$$\begin{aligned}\sigma_{\theta\theta}(R) &= \frac{-p_R(2R_c^2 + 2\delta^2) + 2p_r(R_c^2 - 2R_c\delta + \delta^2)}{4R_c\delta} \\ &= \frac{-p_R\,(1 + \delta^2/R_c^2) + p_r\,(1 - 2\delta/R_c + \delta^2/R_c^2)}{2\delta/R_c}.\end{aligned} \tag{5}$$

Due to the fact that $\delta/R_c << 1$, the terms $\delta^2/R_c^2$ in equations (4) and (5) can be neglected compared to 1. However, we leave the terms $\delta/R_c$ which will allow us to take into account the effect of hydrostatic pressure $p = \min\{p_r, p_R\}$. As a result we obtain

$$\sigma_{\theta\theta}(r) = \frac{\Delta p R_c}{h} - p_R, \qquad \sigma_{\theta\theta}(R) = \frac{\Delta p R_c}{h} - p_r. \tag{6}$$

### 3.2 Basic differential equation

Let us assume that during the corrosion process the radius $R_c$ remains constant, while the thickness $h$ is changing with time.

Combining equations (1) and (6) gives the ordinary differential equation of the first order

$$\frac{dh}{dt} = -(a_r + a_R - m_R p_r - m_r p_R) - (m_r + m_R)\frac{\Delta p\, R_c}{h}. \tag{7}$$

The initial condition to be satisfied at $t = t_0$ is $h|_{t_0} = R_0 - r_0$.

### 3.3 Solution of the basic equation

The integral of the differential equation (7), satisfying the above initial conditions is

$$t = t_0 + \frac{m\,\Delta p\, R_c}{a^2}\ \ln\frac{m\,\Delta p\, R_c + ah}{m\,\Delta p\, R_c + ah_0} - \frac{h - h_0}{a}, \tag{8}$$

where $a = a_r + a_R - m_R p_r - m_r p_R$ and $m = m_r + m_R$.

Maximum principal stresses at the tube's surfaces can then be calculated by the use of equations (6) for every $(t, h)$.

### 3.4 The assessment of the lifetime

In the framework of the maximum normal stress criterion, since $|\sigma_1(r)| \geq |\sigma_1(R)|$, for the assessment of the durability of the tube we should track the stress $\sigma_1(r)$. Let us denote it by $\sigma$. The lifetime of the tube can be defined as the time $t^*$ at which $\sigma$ reaches a strength limit $\sigma^* :\ \sigma(t^*) = \sigma^*$. Therefore, the lifetime can be determined by equation (8) with $h^* = \Delta p\, R_c/(\sigma^* + p_R)$ for $h$.

## 4 CALCULATION RESULTS

Let us compare calculation results obtained by the proposed model (using equations (6) for the effective stresses in (1)) with the results based on the classical thin shell theory formula (using equation (2) for the effective stress). In figure 1 the dependencies of $|\sigma|$ on $t$ are plotted for the cylindrical tube with the middle radius $R_c = 100$ $[l_c]$ and the initial thickness $h_0 = 4$ $[l_c]$; $t_0 = 0$. Here, $l_c$ is a certain unit of length. Corrosion kinetics constants are $a_r = a_R = 0.16$ $[l_c/t_c]$ and $m_r = m_R = 0.008$ $[l_c/(t_c p_c)]$). Here and below, $t_c$ is an appropriate unit of time; $p_c$ is a given unit of stress.

In order to demonstrate the effect of hydrostatic pressure $p = \min\{p_r, p_R\}$, four sets of the internal and external pressure values were used:

- $p_r = 1$ $[p_c]$, $p_R = 0$ (green dotted line);
- $p_r = 0$, $p_R = 1$ $[p_c]$ (blue dotted line);
- $p_r = 6$ $[p_c]$, $p_R = 5$ $[p_c]$ (green dashed line);
- $p_r = 5$ $[p_c]$, $p_R = 6$ $[p_c]$ (blue dashed line).

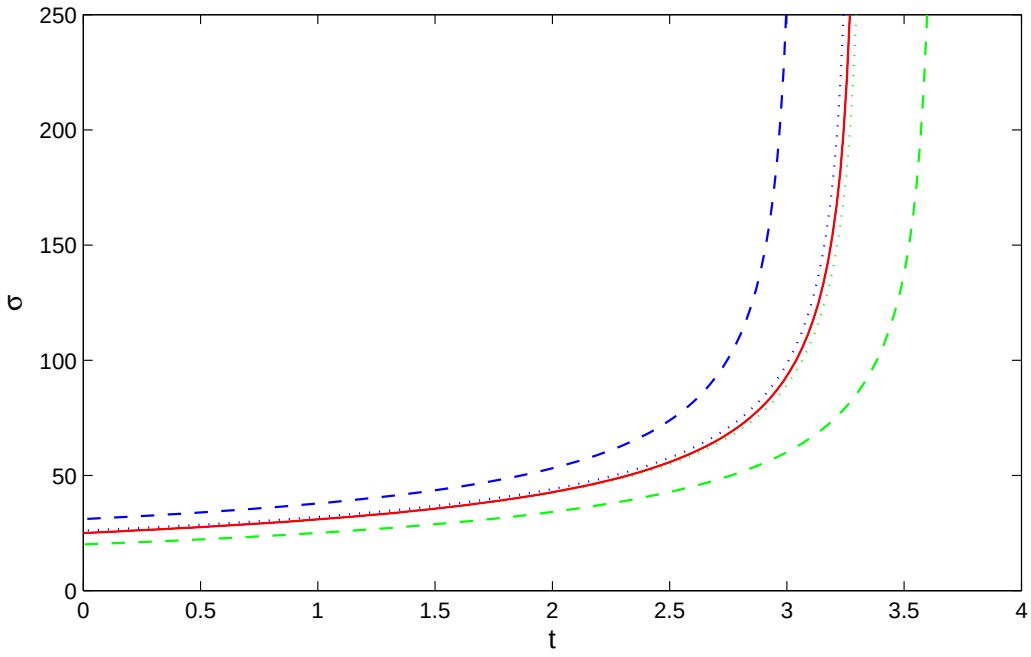



Figure 1: Evolution of $|\sigma(t)|$ for $|\Delta p| = 1$.

These sets are chosen in such a way that despite various values of the pressures $p_r$ and $p_R$, the absolute value of the pressure difference remains constant: $|\Delta p| = 1$ $[p_c]$.

It is evident that when the effective stress is given by equation (2), the dependencies of $|\sigma|$ on the time $t$ for all the above sets of pressure values are the same. In other words, for $|\Delta p| = \text{const}$, the classical thin shell theory formula provides only one curve; it is the red solid line in figure 1.

For the refined model, we have different curves for each of the considered sets of pressure values. Blue curves in figure 1 correspond to the case when $\Delta p < 0$, while green ones correspond to positive differences, $\Delta p > 0$.

It is seen from the figure that the curves for both models are close to each other at zero hydrostatic pressure, $p = \min\{p_r, p_R\} = 0$ (compare solid line and dotted lines). However, the difference between the classical thin shell model (solid line) and our refined model (dashed lines) increases with the growing $p$. Therefore, using the classical thin shell model is not reasonable for modelling the mechanochemical corrosion under high hydrostatic pressure (especially when $|\Delta p| << p$).

In the framework of our model it was revealed that the durability of the tube increases as $p$ grows if $\Delta p > 0$ (see green lines), and decreases if $\Delta p < 0$ (see blue lines).

At the same time, the computational results [11] based on the Lamé's formula and proposed here are hardy differ for any pressure values.

## 5 CONCLUSIONS

- A new mathematical model for the double-sided mechanochemical corrosion of a thin-walled cylindrical tube has been developed.
- An analytical solution of the problem has been obtained.

- The computational results for the classical thin shell model and for the proposed here refined model are very close to each other for zero hydrostatic pressure, $p = \min\{p_r, p_R\} = 0$.
- As compared to the classical thin shell model, the durability of the tube increases as $p$ grows if $p_r > p_R$, and decreases if $p_r < p_R$.
- Using the classical thin shell model is not reasonable for modelling the mechanochemical corrosion under high hydrostatic pressure.
- The solutions based on the Lamé's formula and proposed here are hardy differ for any pressure values.

## 6 ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (project N 16-08-00890).

## REFERENCES

[1] P.A. Pavlov, B.A. Kadyrbekov, V.V. Borisevich, Uniform stress corrosion and corrosion cracking of structural steels. *Soviet Materials Science*, **21** (3), 248–251, 1985.

[2] E.M. Gutman, *Mechanochemistry of solid surfaces*. World Scientific, Singapore, 1994.

[3] V.M. Dolinskii, Calculations on loaded tubes exposed to corrosion. *Chemical and Petroleum Engineering*, **3** (2), 96–97, 1967.

[4] V.G. Karpunin, S.I. Kleshchev, M.S. Kornishin, Calculation of plates and shells taking general corrosion into account. *Proceedings, 10th All-Union Conference of the Theory of Shells and Plates*, **1**, 166–174, 1975. (In Russian).

[5] I.G. Ovchinnikov, Yu.M. Pochtman, Calculation and rational design of structures subjected to corrosive wear (review). *Soviet Materials Science*, **27** (2), 105–116, 1992.

[6] E. Gutman, J. Haddad, R. Bergman, Stability of thin-walled high-pressure vessels subjected to uniform corrosion. *Thin-Walled Structures*, **38**, 43–52, 2000.

[7] A.I. Rusanov, Mechanochemistry of dissolution: Kinetic aspect. *Russian Journal of General Chemistry*, **77** (4), 491–502, 2007.

[8] I. Elishakoff, G. Ghyselinck, Y. Miglis, Durability of an elastic bar under tension with linear or nonlinear relationship between corrosion rate and stress. *Journal of Applied Mechanics, Trans. ASME,* **79** (2), 021013, 2012.

[9] O.S. Sedova, Y.G. Pronina, Initial boundary value problems for mechanochemical corrosion of a thick spherical member in terms of principal stress. *AIP Conference Proceedings*, **1648**, 260002, 2015.

[10] Y.G. Pronina, Estimation of the life of an elastic tube under the action of a longitudinal force and pressure under uniform surface corrosion conditions. *Russian Metallurgy (Metally),* **2010** (4), 361–364, 2010.

[11] Y. Pronina, Analytical solution for the general mechanochemical corrosion of an ideal elastic-plastic thick-walled tube under pressure. *International Journal of Solids and Structures*, **50** (22–23), 3626–3633, 2013.

[12] Y.G. Pronina, O.S. Sedova, S.A. Kabrits, On the applicability of thin spherical shell model for the problems of mechanochemical corrosion. *AIP Conference Proceedings*, **1648**, 300008, 2015.

[13] S.A.Kabrits, E.P. Kolpak, Numerical study of convergence of nonlinear models of the theory of shells with thickness decrease. *AIP Conference Proceedings*, **1648**, 300005, 2015.

[14] O.S. Sedova, Y.G. Pronina, Taking account of hydrostatic pressure in the modeling of corrosion of thick spherical shells. *2015 International Conference on Mechanics - Seventh Polyakhov's Reading*, 7106771, 2015.

[15] O. Sedova, Y. Pronina, A new model for the mechanochemical corrosion of a thin spherical shell. *EPJ Web of Conferences*, 108, 02040, 2016.

[16] O.S. Sedova, Y.G. Pronina, Calculation of the optimal initial thickness of a spherical vessel operating in mechanochemical corrosion conditions. *2015 International Conference on "Stability and Control Processes" in Memory of V.I. Zubov, SCP 2015 - Proceedings*, 7342165, 436–439, 2015.

[17] E.P. Kolpak, S.E. Ivanov, Mathematical and computer modeling vibration protection system with damper. *Applied Mathematical Sciences*, **9** (77–80), 3875–3885, 2015.

[18] E.P. Kolpak, S.E. Ivanov, Mathematical modeling of the system of drilling rig. *Contemporary Engineering Sciences*, **8** (13–16), 699–708, 2015.

[19] V.A. Pavlovsky, A.L. Chistov, Computation parameters of cargo reservoir fill-up process on CNG-vessels in MATLAB. *2014 International Conference on Computer Technologies in Physical and Engineering Applications, ICCTPEA 2014 - Proceedings* 6893322, 136, 2014.

[20] V.A. Pavlovsky, A.L. Chistov, K.V. Shestov, Modeling turbulent flow in a rough walls pipe. *2015 International Conference on "Stability and Control Processes" in Memory of V.I. Zubov, SCP 2015 - Proceedings*, 7342160, 424–425, 2015.

[21] N.A. Gasratova, I.A. Stareva, Reliability assessment of reinforced concrete structures. *2015 International Conference on "Stability and Control Processes" in Memory of V.I. Zubov, SCP 2015 - Proceedings*, 7342168, 378–381, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A HIGHER ORDER PHASE-FIELD APPROACH TO FRACTURE FOR FINITE-DEFORMATION CONTACT PROBLEMS

**M. Franke**[1], **C. Hesch**[1] **and M. Dittmann**[2]

[1] Karlsruhe Institute of Technology
Otto-Ammann-Platz 9, 76131 Karlsruhe, Germany
e-mail: {marlon.franke,christian.hesch}@kit.edu
[2] University of Siegen
Paul-Bonatz-Str. 9-11, 57068 Siegen, Germany
e-mail: maik.dittmann@uni-siegen.de

**Keywords:** Finite deformations, Mortar contact, fracture, phase-field, NURBS, hierarchical refinement.

**Abstract.** *The present contribution provides a comprehensive computational framework for large deformational contact and phase-fracture analysis and is based on the recently appeared publication [16]. A phase-field approach to fracture allows for the efficient numerical treatment of complex fracture patterns for three dimensional problems. Recently, the fracture phase-field approach has been extended to finite deformations (see [18] for more details). In a nutshell, the phase-field approach relies on a regularization of the sharp (fracture-) interface. Besides a second-order Allen-Cahn phase-field model, a more accurate fourth-order Cahn-Hilliard phase-field model is considered as regularization functional. For the former standard finite element analysis (FEA) is sufficient. The latter requires global $C^1$ continuity (see [3]), for which we provide a suitable isogeometric analysis (IGA) framework. Furthermore, to account for different local physical phenomena, like the contact zone, the fracture region or stress peak areas, a newly developed hierarchical refinement scheme is employed (see [19] for more details). For the numerical treatment of the contact boundaries we use the variational consistent Mortar method. The Mortar method passes the patch-test and is known to be the most accurate numerical contact method. It can be extended, in a straightforward manner, to transient phase-field fracture problems. The performance of the proposed methods will be examined in several representative numerical examples.*

## 1 INTRODUCTION

The underlying contribution deals with large deformational continuum bodies which are assumed to contact each other and are each able to fracture within considered simulation time. For the spatial discretization a modern isogeometric analysis (IGA) framework with local refinement scheme and the variational consistent Mortar contact method are employed. A structure preserving integrator is provided for dynamic problems. Note, the work is based on the recently appeared publication [16].

The phase-field method has originally been used to model phase separation. In the last two decades it furthermore has been used to regularize sharp cracks (see e.g. [26, 28, 24]). Besides the displacements the phase-field parameter $\mathfrak{s}$ is introduced as primary unknown. This phase-field parameter is driven by a suitable phase-field partial-differential equation (PDE), which is used to regularize the sharp crack interface. In the literature different PDE's are discussed (see e.g. [3, 34]). We use a second and a fourth order PDE. It is assumed, that the crack initiates or growths by attainment of a local critical energy release rate (see e.g. [9, 5, 26]). It is further possible to give a variational formulation for the crack propagation problem (see e.g. [23, 13, 12]). The phase-fracture method has been developed in the small strain regime (see e.g. [26, 28]). Recently, the phase-fracture method has been extended to the full nonlinear regime (see [18]) by using a multiplicative split of the deformation gradient into compressive and tensile parts (cf. [27]). It is important to remark, that the numerical treatment of phase-field approaches to fracture is less sophisticated than other computational crack propagation techniques for modeling sharp cracks.

In the last three decades, computational modeling of contact mechanics has been intensified (see [37, 25] for comprehensive overviews). Besides traditional nodal based contact methods, e.g. like the node-to-surface method, the variational consistent Mortar method has been well-established (see e.g. [31, 15, 30, 36]). In a recent publication (see [7]) we applied the Mortar method for thermo-mechanical frictional contact problems. Therein we used an isogeometric analysis framework, which allows for higher order approximations with arbitrary adjustable continuity. Here and in our recent publication [16] we extend the isogeometric analysis framework with the phase-field fracture approach. Note, for higher order phase-field equations standard $C^0$ Lagrangian shape functions are not sufficient anymore.

Concerning the spatial discretization both the contact zone as well as the fracture zone demand for refined meshes. Higher order spatial discretization methods are more accurate and reduce the computational demand (see [6] for a comprehensive overview). Therefore we make use of an IGA framework. To be specific we use non-uniform rational B-splines (NURBS), for which the continuity is adjustable by construction of the shape functions. For local refinement mainly T-splines and hierarchical refinement schemes have been used. The application of T-Splines have some drawbacks (see [1] for details), such that we employ an hierarchical refinement scheme (see [8, 33, 4]). Hierarchical refinement procedures replace B-spline and NURBS basis functions on the refined level by a linear combination of scaled and copied versions of themselves, maintaining the required continuity (see e.g. [22, 29]). In particular, we aim at an hierarchical refinement formulation which maintains the partition of unity and is suitable to be adapted to traditional contact mechanical formulations.

An outline of the underlying contribution is as follows. The continuum mechanical basis with application to contact and fracture mechanics and the corresponding governing equations are dealt with in Section 2. Standard FEA and IGA discretization of the weak form follows in Section 3. Furthermore, a modern Mortar contact approach will be given in Section 3. The

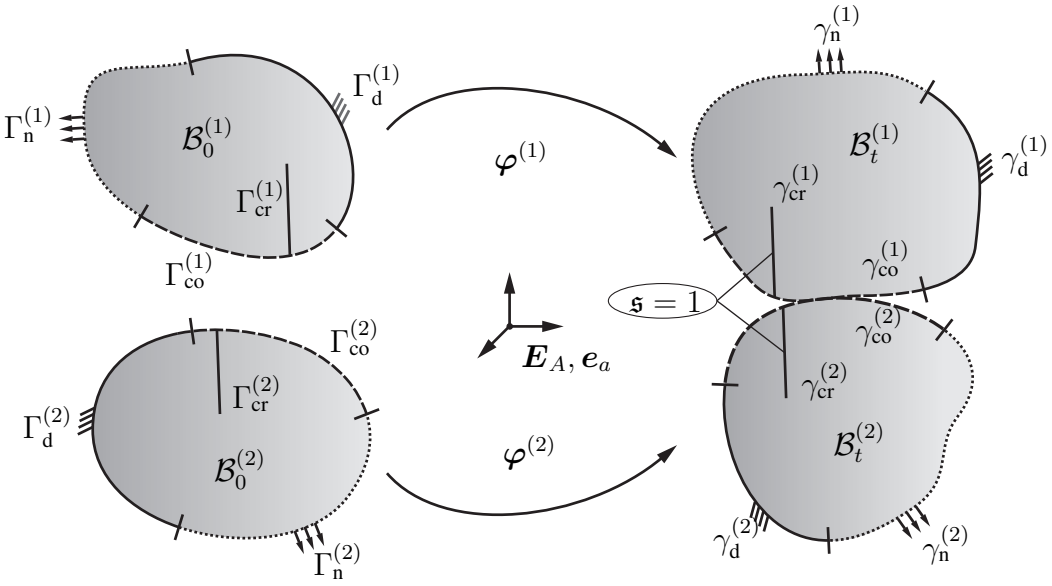



Figure 1: Typical reference (left) and current (right) configurations of bodies $\mathcal{B}^{(i)}$ with crack phase-field.

temporal discretization of the semi-discrete equations is covered in Section 4. Representative numerical examples are given in Section 5. Eventually, conclusions are drawn in Section 6.

## 2 Governing equations

The underlying contribution deals with continuum bodies $\mathcal{B}^{(i)} \in \mathbb{R}^d$, where $d \in \{1, 2, 3\}$. For ease of exposition we restrict the consideration to a two body problem, i.e. $i = 1, 2$, depicted in Figure 1. The bodies are assumed to contact each other within time $t \in \mathcal{I} = [0, T]$, where $T \in \mathbb{R}^+$. The bodies $\mathcal{B}^{(i)}$ are presented in its reference configuration $\mathcal{B}_0^{(i)}$ and its current configuration $\mathcal{B}_t^{(i)}$, accordingly finite deformations are incorporated. The reference configuration can be addressed with material coordinates $\boldsymbol{X}^{(i)} \in \mathbb{R}^d$ and corresponding material basis $\boldsymbol{E}_A$, where $A = 1, ..., d$. Moreover, the current configuration is given by the bijective mapping $\boldsymbol{\varphi}^{(i)} : \mathcal{B}_0^{(i)} \times \mathcal{I} \to \mathbb{R}^d$ and is subject to spatial basis $\boldsymbol{e}_a$, where $a = 1, ..., d$.

In order to model fracture the smooth phase-field parameter $\mathfrak{s}^{(i)} : \mathcal{B}_0^{(i)} \times \mathcal{I} \to [0, 1]$ is introduced. $\mathfrak{s}$ can be regarded as damage variable and physically represents a homogenized macroscopic crack of micro-cracks (see [28]).

The mechanical boundary $\Gamma^{(i)} := \partial\mathcal{B}^{(i)} \in \mathbb{R}^{d-1}$ (depicted in Figure 1) is decomposed into a Dirichlet boundary $\Gamma_{\mathrm{d}}^{(i)} \subset \Gamma^{(i)}$, a Neumann boundary $\Gamma_{\mathrm{n}}^{(i)} \subset \Gamma^{(i)}$ and a contact boundary $\Gamma_{\mathrm{co}}^{(i)} \subset \Gamma^{(i)}$, accordingly

$$\Gamma^{(i)} = \Gamma_{\mathrm{n}}^{(i)} \cup \Gamma_{\mathrm{d}}^{(i)} \cup \Gamma_{\mathrm{co}}^{(i)} \,. \tag{1}$$

The boundaries may not overlap each other, such that

$$\Gamma_{\mathrm{n}}^{(i)} \cap \Gamma_{\mathrm{d}}^{(i)} = \Gamma_{\mathrm{n}}^{(i)} \cap \Gamma_{\mathrm{co}}^{(i)} = \Gamma_{\mathrm{d}}^{(i)} \cap \Gamma_{\mathrm{co}}^{(i)} = \emptyset \,. \tag{2}$$

This has to be valid for the spatial counterparts of the boundaries

$$\gamma_{(\bullet)}^{(i)} = \boldsymbol{\varphi}^{(i)}(\Gamma_{(\bullet)}^{(i)}, t) \,, \tag{3}$$

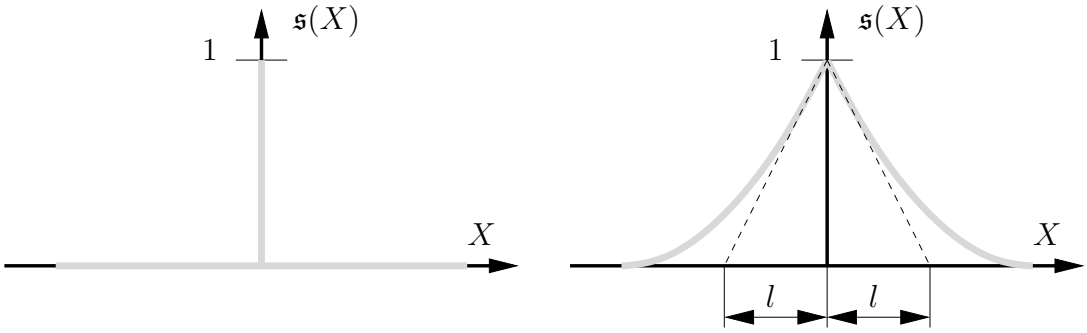



Figure 2: 1D sharp (left) and phase-field regularized (right) crack.

as well, where the abbreviation $(\bullet)$ is used to refer to the different boundaries. In addition to the mechanical boundaries a crack Dirichlet boundary $\Gamma_{\text{cr}}^{(i)}$ defines a new internal boundary of corresponding body and can be initialized within the bodies. All other phase-field boundaries can be regarded as Neumann boundaries. For the two-field problem the displacement as well as the scalar-valued phase-field parameter are the primary unknowns

$$[\boldsymbol{\varphi}, \mathfrak{s}] \in \mathbb{R}^{d+1} \,. \tag{4}$$

**Phase-field contribution** The aforementioned crack phase-field parameter $\mathfrak{s}$ has two bounds, the unbroken state with $\mathfrak{s} = 0$ and the fully broken state with $\mathfrak{s} = 1$ (see [28, 18]). Herein, we assume that crack initiates or continues only in tensile state by attainment of a critical local fracture energy density, given by $\mathcal{G}_{\text{c}}^{(i)}$, which is related to the critical Griffith-type fracture energy (see [11, 21, 28]). The sharp crack is a $d-1$-dimensional manifold. To avoid the difficult modeling of sharp cracks we regularize the crack zone with a suitable crack density functional $\gamma_{\text{cr},n}^{(i)}$, such that we are able to integrate over the $d$-dimensional domain

$$\mathcal{G}_{\text{c}}^{(i)} \int_{\Gamma^{(i)}} \mathrm{d}A \approx \mathcal{G}_{\text{c}}^{(i)} \int_{\mathcal{B}_0^{(i)}} \gamma_{\text{cr},n}^{(i)} \, \mathrm{d}V = \mathcal{G}_{\text{c}}^{(i)} \, \Gamma_{\text{cr},n}^{(i)} =: V_{\text{cr},n}^{(i)} \,.$$

Therein $n$ denotes the order of the phase-field model (cf. [3]). In [28] a second order and in [3, 34] a fourth order differential equation have been proposed[1] (see Figure 3 for the 1D analytical solution). The former is given by

$$\mathfrak{s}^{(i)} - 4\,(l^{(i)})^2\,\Delta\mathfrak{s}^{(i)} = 0\,, \tag{5}$$

with the corresponding functional

$$\mathcal{F}_{\text{cr},2}^{(i)} = \frac{1}{2} \int_{\mathcal{B}_0^{(i)}} ((\mathfrak{s}^{(i)})^2 + 4\,(l^{(i)})^2\,\nabla\mathfrak{s}^{(i)} \cdot \nabla\mathfrak{s}^{(i)}) \, \mathrm{d}V\,, \tag{6}$$

whereas the latter is given by

$$\mathfrak{s}^{(i)} - 2\,(l^{(i)})^2\,\Delta\mathfrak{s}^{(i)} + (l^{(i)})^4\,\Delta\Delta\mathfrak{s}^{(i)} = 0\,. \tag{7}$$

The corresponding functional denotes

[1] It is important to remark, that it does not exist a natural PDE to model the crack density functional (for more informations see [34])

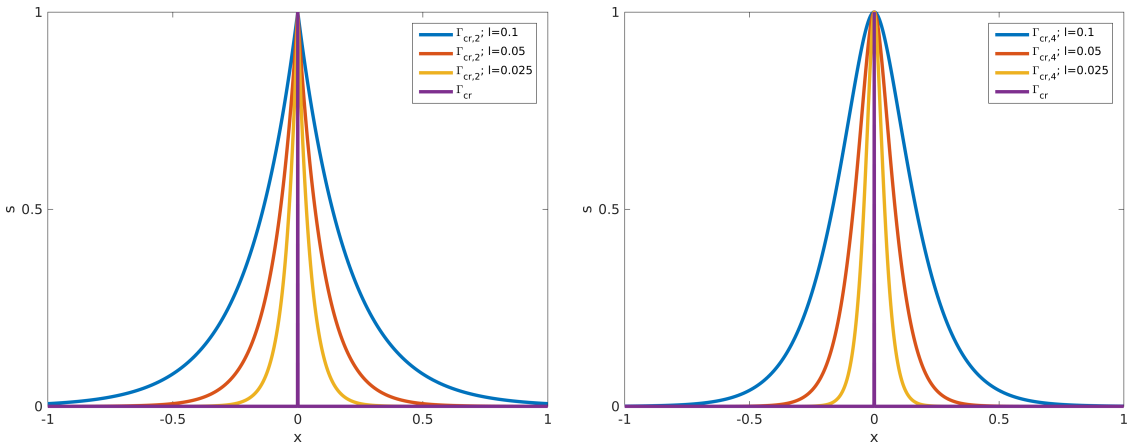



Figure 3: Analytical solution of 1D second order phase-field (5) $\mathfrak{s}(x) = e^{\frac{-|x|}{2\,l}}$ (left) and fourth order phase-field (7) $\mathfrak{s}(x) = e^{\frac{-|x|}{l}}\,(1 + \frac{|x|}{l})$ with different length scale parameters.

$$\mathcal{F}^{(i)}_{\text{cr},4} = \frac{1}{2}\int_{\mathcal{B}^{(i)}_0} \Big(\frac{1}{2}\,(\mathfrak{s}^{(i)})^2 + (l^{(i)})^2\,\nabla\mathfrak{s}^{(i)}\cdot\nabla\mathfrak{s}^{(i)} + \frac{(l^{(i)})^4}{2}\,\Delta\mathfrak{s}^{(i)}\,\Delta\mathfrak{s}^{(i)}\Big)\,\mathrm{d}V\,. \tag{8}$$

In the above equations $l^{(i)}$ denotes the length scale parameter, which determines the width of the regularization zone (see Figure 2). The length scale parameter may be treated as material parameter and should be chosen as

$$h^{(i)} < \frac{l^{(i)}}{2}\,, \tag{9}$$

where $h^{(i)}$ denotes the smallest finite element size (for more details see [28]). Note, for the fourth order approach $l^{(i)}$ can be chosen even smaller as suggested by (9) (see e.g. [16]) The regularized crack surface topologies are then constructed by $\Gamma^{(i)}_{\text{cr},2/4} := \frac{1}{l^{(i)}}\,\mathcal{F}^{(i)}_{\text{cr},2/4}$ as proposed in [28]. Accordingly, for (5) and (7), we obtain

$$\Gamma^{(i)}_{\text{cr},2} = \frac{1}{2\,l^{(i)}}\int_{\mathcal{B}^{(i)}_0} \big((\mathfrak{s}^{(i)})^2 + 4\,(l^{(i)})^2\,\nabla\mathfrak{s}^{(i)}\cdot\nabla\mathfrak{s}^{(i)}\big)\,\mathrm{d}V\,, \tag{10}$$

$$\Gamma^{(i)}_{\text{cr},4} = \frac{1}{4\,l^{(i)}}\int_{\mathcal{B}^{(i)}_0} \big((\mathfrak{s}^{(i)})^2 + 2\,(l^{(i)})^2\,\nabla\mathfrak{s}^{(i)}\cdot\nabla\mathfrak{s}^{(i)} + (l^{(i)})^4\,\Delta\mathfrak{s}^{(i)}\,\Delta\mathfrak{s}^{(i)}\big)\,\mathrm{d}V\,, \tag{11}$$

respectively. As can be observed in Figure 3, the fourth order phase-field approach has two advantages with respect to the second order phase-field approach. The fourth order phase-field approach does not contain non-differentiable areas and further by using the same length scale parameter the transition zone is smaller. Moreover, in the numerical treatment better accuracy and convergence rates of the solution have been observed (see [3]). For the second order phase-field model $C^0$-continuity, whereas for the fourth order model $C^1$-continuity is required.

**Bulk contribution** The potential of the bulk is given by

$$V^{(i)}_{\text{bulk}} = \int_{\mathcal{B}^{(i)}_0} \Psi^{(i)}\,\mathrm{d}V\,, \tag{12}$$

where $\Psi^{(i)}$ denotes the strain energy density function of body $i$. The corresponding balance of linear momentum is given by

$$\mathrm{Div}(\boldsymbol{P}^{(i)}) + \bar{\boldsymbol{B}}^{(i)} = \rho_0^{(i)}\, \ddot{\boldsymbol{\varphi}}^{(i)}\,, \tag{13}$$

where $\boldsymbol{P}^{(i)} : \mathcal{B}_0^{(i)} \times \mathcal{I} \to \mathbb{R}^{d\times d}$ denotes the first Piola-Kirchhoff stress tensor, $\bar{\boldsymbol{B}}^{(i)} \in \mathbb{R}^d$ denotes the prescribed body force density and $\rho_0^{(i)} : \mathcal{B}_0^{(i)} \times \mathcal{I} \to \mathbb{R}$ denotes the mass density. Additionally, an arbitrary hyperelastic material law can be incorporated with a suitable strain tensor. The symmetric right Cauchy-Green strain tensor is introduced by

$$\boldsymbol{C}^{(i)} = \boldsymbol{F}^{(i)}\, \boldsymbol{F}^{(i)^{\mathrm{T}}}\,, \tag{14}$$

using the deformation gradient $\boldsymbol{F}^{(i)} : \mathcal{B}_0^{(i)} \times \mathcal{I} \to \mathbb{R}^{d\times d}$. As already mentioned, we assume that only local tension rather than local compression is responsible for crack growth. Accordingly, we aim at an anisotropic description, such that we need to split the kinematic into tension and compression[2]. Therefore we employ an eigendecomposition of the deformation gradient, which yields

$$\boldsymbol{F}^{(i)} = \sum_{a=1}^{d} \lambda_a^{(i)}\, \boldsymbol{a}_a^{(i)} \otimes \boldsymbol{A}_a^{(i)}\,, \tag{15}$$

where the deformation gradient is represented in its principal stretches $\lambda_a^{(i)}$ and spatial and material directions $\boldsymbol{a}_a^{(i)}$, $\boldsymbol{A}_a^{(i)}$, respectively. As proposed by [18], an operator split of the deformation gradient is performed, such that

$$\boldsymbol{F}^{(i)} = \boldsymbol{F}^{(i),-}\, \boldsymbol{F}^{(i),+} = \sum_{a=1}^{d} \lambda_a^{(i),-}\, \lambda_a^{(i),+}\, \boldsymbol{a}_a^{(i)} \otimes \boldsymbol{A}_a^{(i)}\,. \tag{16}$$

Therein the superscripted $\pm$ denotes

$$\lambda_a^{(i),\pm} = \frac{\lambda_a^{(i)} \pm |\lambda_a^{(i)}|}{2}\,. \tag{17}$$

Accordingly, the principal strains are decomposed into tensile and compressive components. Furthermore, an anisotropic split of the principal stretches is accomplished, which yields

$$\boldsymbol{F}^{(i)} =: \boldsymbol{F}_{\mathfrak{i}}^{(i)}\, \boldsymbol{F}_{\mathfrak{e}}^{(i)}\,, \tag{18}$$

where the fracture sensitive and insensitive parts are introduced as follows

$$\boldsymbol{F}_{\mathfrak{i}}^{(i)} = \sum_{a=1}^{d} (\lambda_a^{(i),+})^{1-g(\mathfrak{s}^{(i)})}\, \boldsymbol{a}_a \otimes \boldsymbol{A}_a, \quad \boldsymbol{F}_{\mathfrak{e}}^{(i)} = \sum_{a=1}^{d} (\lambda_a^{(i),+})^{g(\mathfrak{s}^{(i)})}\, \lambda_a^{(i),-}\, \boldsymbol{a}_a \otimes \boldsymbol{A}_a\,. \tag{19}$$

Therein the degradation function $g(\mathfrak{s}^{(i)}) = 1 - \mathfrak{s}^{(i)}$ has been introduced. Note, in the linear case an additive split has been proposed by [26] and can be stated as

$$\Psi^{(i)}(\boldsymbol{\epsilon}^{(i)}) \to \tilde{\Psi}^{(i)}(\boldsymbol{\epsilon}^{i}, \mathfrak{s}^{(i)}) = (g(\mathfrak{s}^{(i)}) + k)\, \Psi^{(i),+}(\boldsymbol{\epsilon}^{(i)}) + \Psi^{i,-}(\boldsymbol{\epsilon}^{(i)}), \quad g(\mathfrak{s}^{(i)}) = (1 - \mathfrak{s}^{(i)})^2\,. \tag{20}$$

[2]Note, a less realistic but more simple choice would be an isotropic description.

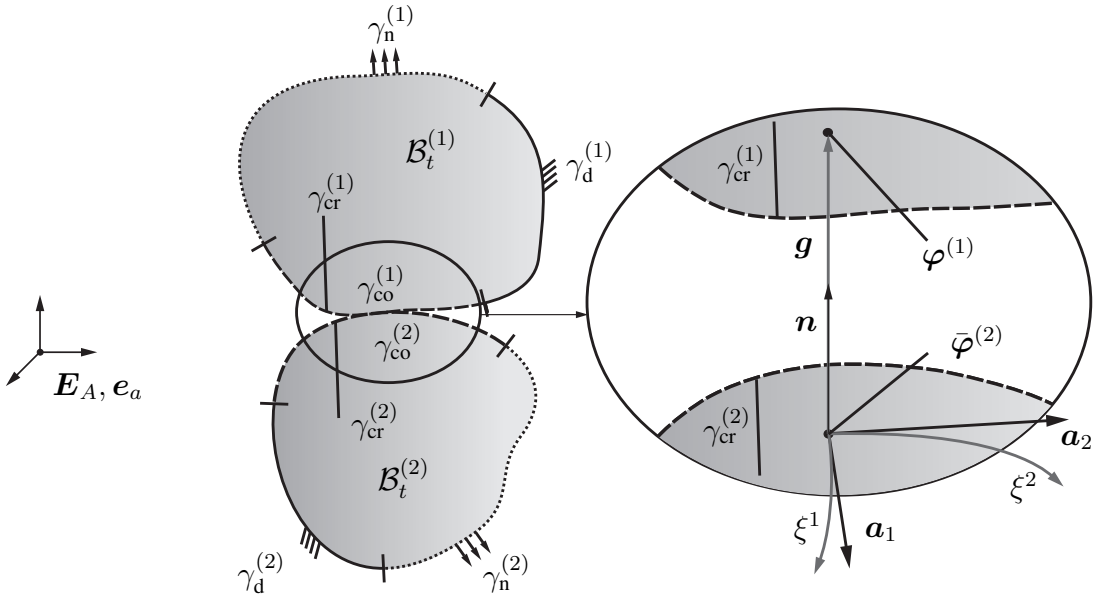



Figure 4: Parametrization of contact surface $\gamma_{\mathrm{co}}^{(2)}$.

Note further, that a more sophisticated cubic degradation function

$$g(\mathfrak{s}^{(i)}) = a^{(i)}\left((1-\mathfrak{s}^{(i)})^3 - (1-\mathfrak{s}^{(i)})^2\right) - 2\,(1-\mathfrak{s}^{(i)})^3 + 3\,(1-\mathfrak{s}^{(i)})^2, \quad a^{(i)} \geq 0\,, \tag{21}$$

is discussed in [2] and can easily be incorporated. For $a^{(i)} = 2$ the quadratic functional in equation (20) is recovered. Using the above coupled kinematic, the Helmholtz free energy depends besides the strain also on the phase-field, such that

$$\Psi^{(i)}(\boldsymbol{C}^{(i)}) \to \Psi^{(i)}(\boldsymbol{C}_{\mathfrak{e}}^{(i)}) = \Psi_{\mathfrak{e}}^{(i)}(\boldsymbol{C}^{(i)}, \mathfrak{s}^{(i)}) =: \Psi_{\mathfrak{e}}^{(i)}\,. \tag{22}$$

Accordingly, the elastic part of bulk potential (12) can be written as

$$V_{\mathrm{bulk},\mathfrak{e}}^{(i)} = \int_{\mathcal{B}_0^{(i)}} \Psi_{\mathfrak{e}}^{(i)}\,\mathrm{d}V\,. \tag{23}$$

**Contact contribution** The underlying contact description is based on the description given in [10]. Note, as mentioned earlier, the phase-field and the contact boundaries do not depend on each other. As can be observed in Figure 4 we introduce a local convective coordinate system. The closest point projection of $\boldsymbol{\varphi}^{(1)}(\boldsymbol{X}^{(1)}, t) \in \gamma_{\mathrm{co}}^{(1)}$ to opposing surface $\gamma_{\mathrm{co}}^{(2)}$ yields

$$\|\boldsymbol{\varphi}^{(1)}(\boldsymbol{X}^{(1)}, t) - \bar{\boldsymbol{\varphi}}^{(2)})\| \to \min\,. \tag{24}$$

Therein $\bar{\boldsymbol{\varphi}}^{(2)} := \boldsymbol{\varphi}^{(2)}(\bar{\boldsymbol{X}}^{(2)}(\boldsymbol{X}^{(1)}, t)$ denotes the orthogonal projected point to $\boldsymbol{\varphi}^{(1)}(\boldsymbol{X}^{(1)}, t)$. The local convective coordinate system is parametrized with convective coordinates $\xi^\alpha$, $\alpha \in \{1, 2\}$ such that

$$\bar{\boldsymbol{X}}^{(2)}(\boldsymbol{X}^{(1)}) := \boldsymbol{X}^{(2)}(\bar{\boldsymbol{\xi}}), \quad \bar{\boldsymbol{\varphi}}^{(2)} := \boldsymbol{\varphi}^{(2)}(\bar{\boldsymbol{\xi}}, t)\,. \tag{25}$$

Therein $\bar{\boldsymbol{\xi}} = [\bar{\xi^1}, \bar{\xi^2}]$ denotes the convective coordinates of the projected point and is computed via equation (24) e.g. with Newton's method. The convected basis are calculated with

$$\boldsymbol{a}_\alpha := \boldsymbol{\varphi}_{,\alpha}^{(2)}(\bar{\boldsymbol{\xi}}, t)\,. \tag{26}$$

Note that the above basis are in general not orthonormal. Accordingly, it is important to give the metric for the local coordinate system

$$m_{\alpha\beta} = \boldsymbol{a}_\alpha \cdot \boldsymbol{a}_\beta \,. \tag{27}$$

The gap vector is given by (see Figure 4)

$$\boldsymbol{g} = \boldsymbol{\varphi}^{(1)} - \bar{\boldsymbol{\varphi}}^{(2)} \,. \tag{28}$$

Furthermore, the unit outward normal to the surface $\gamma_{\text{co}}^{(2)}$ at point $\bar{\boldsymbol{\varphi}}^{(2)}$ is defined by

$$\boldsymbol{n} := \frac{\boldsymbol{a}_1 \times \boldsymbol{a}_2}{\|\boldsymbol{a}_1 \times \boldsymbol{a}_2\|} \,. \tag{29}$$

With the above the scalar-valued gap function can be computed via

$$g_{\text{N}} = \left(\boldsymbol{\varphi}^{(1)} - \bar{\boldsymbol{\varphi}}^{(2)}\right) \cdot \boldsymbol{n} \,. \tag{30}$$

Considering the balance of linear momentum at the contact boundary, we obtain

$$\boldsymbol{t}_{\text{co}}^{(1)}(\boldsymbol{X}^{(1)}, t) \, \mathrm{d}A^{(1)} = -\boldsymbol{t}_{\text{co}}^{(2)}(\bar{\boldsymbol{X}}^{(2)}(\boldsymbol{X}^{(1)}), t) \, \mathrm{d}A^{(2)} \,. \tag{31}$$

Therein, the involved contact traction can be splitted into a normal and a tangential part as follows

$$\boldsymbol{t}_{\text{co}}^{(1)}(\boldsymbol{X}^{(1)}, t) = \boldsymbol{t}_{\text{N}} + \boldsymbol{t}_{\text{T}}, \quad \boldsymbol{t}_{\text{T}} \cdot \boldsymbol{n} = 0 \,. \tag{32}$$

Moreover, the normal contact traction is given by

$$\boldsymbol{t}_{\text{N}} := \boldsymbol{t}_{\text{N}}^{(1)} = -t_{\text{N}} \, \boldsymbol{n} \,. \tag{33}$$

Frictionless contact is incorporated by using the well-known Karush-Kuhn-Tucker conditions, given by

$$g_{\text{N}} \geq 0 \,, \tag{34}$$

$$t_{\text{N}} \leq 0 \,, \tag{35}$$

$$t_{\text{N}} \, g_{\text{N}} = 0 \,. \tag{36}$$

The Karush-Kuhn-Tucker conditions are comprised of the impenetrability condition (34), the condition which only allows compressive tractions (35) and the complementarity condition (36). The Karush-Kuhn-Tucker conditions are shown in Figure 5 for both the penalty method and Lagrange multipliers. Note, the framework is readily expendable to incorporate any frictional constitutive law (e.g. Coulomb friction), but is omitted herein for convenience, i.e. $\boldsymbol{t}_{\text{T}} = \boldsymbol{0}$. Within the computational treatment of contact, we subsequently apply the active set strategy to obtain an active contact boundary $\bar{\Gamma}_{\text{co}}^{(i)}$ from the potential contact boundary $\Gamma_{\text{co}}^{(i)}$. In particular the inequalities given in (34)-(36) are implemented using the $\max$-operator (see [20] for more informations) as follows

$$\Phi_{\text{N}} = \lambda_{\text{N}} - \max(0, \lambda_{\text{N}} - c \, \Phi_{\text{N}}) = 0 \,. \tag{37}$$

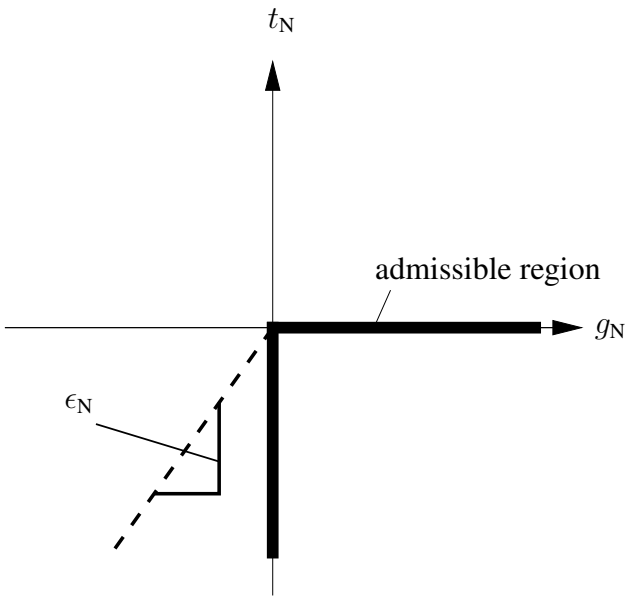



Figure 5: Karush-Kuhn-Tucker conditions for frictionless contact (solid line: exact enforcement, dotted line: penalty regularization with penalty parameter $\epsilon_{\mathrm{N}}$).

Therein $\lambda_{\mathrm{N}}$ and $\Phi_{\mathrm{N}}$ denote the Lagrange multiplier and corresponding constraint, respectively. Note $\lambda_{\mathrm{N}} := t_{\mathrm{N}}$ denotes the exact contact traction and $\Phi_{\mathrm{N}} := g_{\mathrm{N}}$ the gap function. Moreover, within the active set strategy $c \in \mathbb{R}^+$ is a constant parameter, which is suitable to influence the convergence of Newton's method but does not influence the constraint enforcement. Eventually, assuming active contact and using equation (31), we are able to introduce the contact potential employing only one integral expression, such that

$$\sum_{i=1}^{2} \bar{V}_{\mathrm{co}}^{(i)} = \int_{\bar{\Gamma}_{\mathrm{co}}^{(1)}} \boldsymbol{t}_{\mathrm{co}}^{(1)} \cdot \left(\boldsymbol{\varphi}^{(1)} - \boldsymbol{\varphi}^{(2)}\right) \, \mathrm{d}A^{(1)} := \bar{V}_{\mathrm{co}} \,. \tag{38}$$

**Initial boundary value problem** In the following the contact boundaries are assumed to be known. Collecting all introduced contributions, yields

$$V^{\mathrm{aug}} = \sum_{i} V_{\mathrm{bulk},\mathfrak{e}}^{(i)} + V_{\mathrm{cr},2/4}^{(i)} + \bar{V}_{\mathrm{co}} \,. \tag{39}$$

Regarding transient problems the kinetic energy is defined by

$$T = \sum_{i} \frac{1}{2} \int_{\mathcal{B}_0^{(i)}} \rho \, \dot{\boldsymbol{\varphi}}^{(i)} \cdot \dot{\boldsymbol{\varphi}}^{(i)} \, \mathrm{d}V \,. \tag{40}$$

With the Lagrange functional $L = T - V^{\mathrm{aug}}$ at hand, the above is used to employ Hamilton's principal, given by

$$\delta S = \int_{t_1}^{t_2} \delta L \, \mathrm{d}t = \int_{t_1}^{t_2} \left(\delta T - \delta V^{\mathrm{aug}}\right) \, \mathrm{d}t \stackrel{!}{=} 0 \,. \tag{41}$$

Equation (41) leads to the Euler-Lagrange equations. Accordingly, the initial boundary value problem (IBVP), for the coupled system with fourth order phase-field model, can be written as

$$\mathrm{Div}(\boldsymbol{F}_{\mathfrak{e}}^{(i)}\,\boldsymbol{S}_{\mathfrak{e}}^{(i)}) + \bar{\boldsymbol{B}}^{(i)} - \rho_0^{(i)}\,\ddot{\boldsymbol{\varphi}}^{(i)} = \boldsymbol{0} \quad \text{in } \mathcal{B}_0^{(i)} \forall t \in \mathcal{I}\,, \tag{42}$$

$$\mathcal{H}^{(i)} + \frac{\mathcal{G}_c^{(i)}}{4\,l^{(i)}}\,(\mathfrak{s}^{(i)} - 2\,(l^{(i)})^2\,\Delta\mathfrak{s}^{(i)} + (l^{(i)})^4\,\Delta\Delta\mathfrak{s}^{(i)}) = 0 \quad \text{in } \mathcal{B}_0^{(i)} \forall t \in \mathcal{I}\,, \tag{43}$$

$$\boldsymbol{\varphi}^{(i)} = \bar{\boldsymbol{\varphi}}^{(i)} \quad \text{on } \Gamma_{\mathrm{d}}^{(i)} \forall t \in \mathcal{I}\,, \tag{44}$$

$$\boldsymbol{P}^{(i)}\,\boldsymbol{N}^{(i)} = \bar{\boldsymbol{T}}^{(i)} \quad \text{on } \Gamma_{\mathrm{n}}^{(i)} \forall t \in \mathcal{I}\,, \tag{45}$$

$$g_{\mathrm{N}} \geq 0 \quad \text{on } \Gamma_{\mathrm{co}} \forall t \in \mathcal{I}\,, \tag{46}$$

$$t_{\mathrm{N}} \leq 0 \quad \text{on } \Gamma_{\mathrm{co}} \forall t \in \mathcal{I}\,, \tag{47}$$

$$t_{\mathrm{N}}\,g_{\mathrm{N}} = 0 \quad \text{on } \Gamma_{\mathrm{co}} \forall t \in \mathcal{I}\,, \tag{48}$$

$$\mathfrak{s}^{(i)} = 1 \quad \text{on } \Gamma_{\mathrm{cr,d}} \forall t \in \mathcal{I}\,, \tag{49}$$

$$\Delta\mathfrak{s}^{(i)} = 0 \quad \text{on } \Gamma^{(i)} \forall t \in \mathcal{I}\,, \tag{50}$$

$$\nabla((l^{(i)})^4\,\Delta\mathfrak{s}^{(i)} - 2\,(l^{(i)})^2\,\mathfrak{s}^{(i)}) \cdot \boldsymbol{N}^{(i)} = 0 \quad \text{on } \Gamma^{(i)} \forall t \in \mathcal{I}\,, \tag{51}$$

$$\boldsymbol{\varphi}^{(i)}(t=0) = \boldsymbol{\varphi}_0^{(i)} \quad \text{in } \mathcal{B}_0^{(i)}\,, \tag{52}$$

$$\dot{\boldsymbol{\varphi}}^{(i)}(t=0) = \dot{\boldsymbol{\varphi}}_0^{(i)} \quad \text{in } \mathcal{B}_0^{(i)}\,, \tag{53}$$

$$\mathfrak{s}^{(i)}(t=0) = \mathfrak{s}_0^{(i)} \quad \text{in } \mathcal{B}_0^{(i)}\,. \tag{54}$$

The above is supplemented by the constitutive equations

$$\boldsymbol{S}_{\mathfrak{e}}^{(i)} = 2\;\mathrm{D}_1\Psi_{\mathfrak{e}}^{(i)}(\boldsymbol{C}^{(i)}, \mathfrak{s}^{(i)}), \quad \mathcal{H}^{(i)} = \;\mathrm{D}_2\Psi_{\mathfrak{e}}^{(i)}(\boldsymbol{C}^{(i)}, \mathfrak{s}^{(i)})\,, \tag{55}$$

which denote the second Piola-Kirchhoff stress tensor and the driving force of the phase-field. Equations (44)-(45) are the prescribed mechanical Dirichlet and Neumann boundary conditions, whereas equations (49)-(51) denote the phase-field Dirichlet and Neumann boundary conditions. Furthermore equations (52)-(54) denote the initial conditions. Note, in order to obtain the IBVP for the second order phase-field model, equation (43) needs to be exchanged by

$$\mathcal{H}^{(i)} + \frac{\mathcal{G}_c^{(i)}}{2\,l^{(i)}}\,(\mathfrak{s}^{(i)} - 4\,(l^{(i)})^2\,\Delta\mathfrak{s}^{(i)}) = 0 \quad \text{in } \mathcal{B}_0^{(i)}\,. \tag{56}$$

Furthermore, equations (49) and (50) need to be exchanged by condition

$$\nabla\mathfrak{s}^{(i)} = 0 \quad \text{on } \Gamma^{(i)} \forall t \in \mathcal{I}\,. \tag{57}$$

Note that we do not enforce $\dot{\mathfrak{s}} \geq 0$. Accordingly, within our formulation, the phase-field may be healed except the material is fully broken, i.e. $\mathfrak{s} = 1$, which will be enforced with a Dirichlet type mechanism (see next paragraph for more informations).

**Virtual work** Analogously, we are able to collect the virtual work contributions, such that

$$G := \sum_i G_{\mathrm{dyn}}^{(i)} + G_{\mathrm{int,ext}}^{(i)} + G_{\mathrm{cr},2/4}^{(i)} + \bar{G}_{\mathrm{co}}\,. \tag{58}$$

To this end the displacement solution space is defined by

$$\mathcal{V}_s^{(i)} = \{\boldsymbol{\varphi}^{(i)} : \boldsymbol{\varphi}^{(i)}(\boldsymbol{X}^{(i)}, t) \in \mathcal{H}^1(\mathcal{B}_0^{(i)}) | \boldsymbol{\varphi}^{(i)}(\boldsymbol{X}^{(i)}, t) = \bar{\boldsymbol{\varphi}}^{(i)} \text{ on } \Gamma_{\mathrm{d}}^{(i)}\}\,, \tag{59}$$

such that the solution function $\boldsymbol{\varphi}^{(i)}$ is element of the Sobolev space $\mathcal{H}^1$ which includes the space of square-integrable functions and square-integrable first derivatives. Furthermore the solution function is required to satisfy the Dirichlet boundary condition. The space of test functions with the corresponding test function $\delta\boldsymbol{\varphi}^{(i)}$ is postulated as[3]

$$\mathcal{V}_t^{(i)} = \{\delta\boldsymbol{\varphi}^{(i)} : \delta\boldsymbol{\varphi}^{(i)}(\boldsymbol{X}^{(i)}) \in \mathcal{H}^1(\mathcal{B}_0^{(i)}) | \delta\boldsymbol{\varphi}^{(i)}(\boldsymbol{X}^{(i)}) = \boldsymbol{0} \text{ on } \Gamma_{\text{d}}^{(i)}\}\,. \tag{60}$$

Accordingly, the test-function vanishes at the Dirichlet boundary. Using standard techniques, the contributions to the weak form are obtained as follows

$$G_{\text{dyn}}^{(i)} := \int_{\mathcal{B}_0^{(i)}} \delta\boldsymbol{\varphi}^{(i)} \cdot \ddot{\boldsymbol{\varphi}}^{(i)}\, \rho_0^{(i)}\, \mathrm{d}V\,, \tag{61}$$

$$G_{\text{int,ext}}^{(i)} := \int_{\mathcal{B}_0^{(i)}} \boldsymbol{S}_{\mathfrak{e}}^{(i)} : \boldsymbol{F}_{\mathfrak{e}}^{(i)\mathrm{T}}\, \mathrm{Grad}(\delta\boldsymbol{\varphi}^{(i)})\, \mathrm{d}V - \int_{\mathcal{B}_0^{(i)}} \delta\boldsymbol{\varphi}^{(i)} \cdot \bar{\boldsymbol{B}}^{(i)}\, \mathrm{d}V - \int_{\Gamma_{\text{n}}^{(i)}} \delta\boldsymbol{\varphi}^{(i)} \cdot \bar{\boldsymbol{T}}^{(i)}\, \mathrm{d}V\,. \tag{62}$$

Using a suitable test space for the second order PDE in (5), e.g. given by

$$\mathcal{V}_{\text{cr},2}^{s,(i)} = \{\mathfrak{s}^{(i)} : \mathfrak{s}^{(i)}(\boldsymbol{X}^{(i)}, t) \in \mathcal{H}^1(\mathcal{B}_0^{(i)}) | \mathfrak{s}^{(i)} = 1 \text{ on } \Gamma_{\text{cr}}^{(i)}\}\,, \tag{63}$$

$$\mathcal{V}_{\text{cr},2}^{t,(i)} = \{\delta\mathfrak{s}^{(i)} : \delta\mathfrak{s}^{(i)}(\boldsymbol{X}^{(i)}) \in \mathcal{H}^1(\mathcal{B}_0^{(i)}) | \delta\mathfrak{s}^{(i)} = 0 \text{ on } \Gamma_{\text{cr}}^{(i)}\}\,, \tag{64}$$

where $\Gamma_{\text{cr}}^{(i)}$ denotes a sharp crack inside $\mathcal{B}_0^{(i)}$ and thus can be regarded as Dirichlet condition, yields the weak form

$$G_{\text{cr},2}^{(i)} := \frac{\mathcal{G}_c^{(i)}}{2\, l^{(i)}} \int_{\mathfrak{B}_0^{(i)}} \left( \delta\mathfrak{s}^{(i)}\, (\mathfrak{s}^{(i)} + \frac{2\, l^{(i)}}{\mathcal{G}_c^{(i)}}\, \mathcal{H}^{(i)}) + 4\, (l^{(i)})^2\, \mathrm{Grad}(\delta\mathfrak{s}^{(i)}) \cdot \mathrm{Grad}(\mathfrak{s}^{(i)}) \right)\, \mathrm{d}V\,. \tag{65}$$

For the fourth order PDE given in (7), a test space which provides higher continuity is needed, such as

$$\mathcal{V}_{\text{cr},4}^{s,(i)} = \{\mathfrak{s}^{(i)} : \mathfrak{s}^{(i)}(\boldsymbol{X}^{(i)}, t) \in \mathcal{H}^2(\mathcal{B}_0^{(i)}) | \mathfrak{s}^{(i)} = 1 \text{ on } \Gamma_{\text{cr}}^{(i)}\}\,, \tag{66}$$

$$\mathcal{V}_{\text{cr},4}^{t,(i)} = \{\delta\mathfrak{s}^{(i)} : \delta\mathfrak{s}^{(i)}(\boldsymbol{X}^{(i)}) \in \mathcal{H}^2(\mathcal{B}_0^{(i)}) | \delta\mathfrak{s}^{(i)} = 0 \text{ on } \Gamma_{\text{cr}}^{(i)}\}\,. \tag{67}$$

Accordingly, the weak form is obtained by

$$\begin{aligned} G_{\text{cr},4}^{(i)} := \frac{\mathcal{G}_c^{(i)}}{4\, l^{(i)}} \int_{\mathcal{B}_0^{(i)}} & (\delta\mathfrak{s}^{(i)}\, (\mathfrak{s}^{(i)} + \frac{4\, l^{(i)}}{\mathcal{G}_c^{(i)}}\, \mathcal{H}^{(i)}) + 2\, (l^{(i)})^2\, \mathrm{Grad}(\delta\mathfrak{s}^{(i)}) \cdot \mathrm{Grad}(\mathfrak{s}^{(i)}) \\ & + (l^{(i)})^4\, \Delta\delta\mathfrak{s}^{(i)} \cdot \Delta\mathfrak{s}^{(i)})\, \mathrm{d}V\,. \end{aligned} \tag{68}$$

Using Lagrange multipliers for the incorporation of the contact conditions, the virtual work of contact yields

$$\bar{G}_{\text{co}} = \bar{G}_{\text{co}}^{\lambda_{\text{N}}} + \bar{G}_{\text{co}}^{\boldsymbol{\varphi}}\,, \tag{69}$$

[3]Note that the test function $\delta\boldsymbol{\varphi}^{(i)}$ can also be interpreted as virtual displacement.

where $\bar{G}_{\text{co}}^{\lambda_{\text{N}}} := \delta_{\lambda_{\text{N}}} \bar{V}_{\text{co}}$ and $\bar{G}_{\text{co}}^{\boldsymbol{\varphi}} := \delta_{\boldsymbol{\varphi}} \bar{V}_{\text{co}}$ have been introduced. Eventually, the governing equations are obtained as follows[4]

$$\sum_i G_{\text{dyn}}^{(i)} + G_{\text{int,ext}}^{(i)} + \bar{G}_{\text{co}}^{\boldsymbol{\varphi}} = 0, \quad \forall \delta\boldsymbol{\varphi}^{(i)} \in \mathcal{V}_t^{(i)}\,, \tag{70}$$

$$\sum_i G_{\text{cr},2/4}^{(i)} = 0, \quad \forall \delta\mathfrak{s}^{(i)} \in \mathcal{V}_{\text{cr},2/4}^{(i)}\,, \tag{71}$$

$$\bar{G}_{\text{co}}^{\lambda_{\text{N}}} = 0, \quad \forall \delta\lambda_{\text{N}} \in \mathbb{R}\,. \tag{72}$$

Note, the abbreviation 2/4 indicates that either the virtual work of the second order phase-field equation (65) or the virtual work of the fourth order phase-field equation (68) can be taken into account.

## 3 Spatial discretization

To perform the spatial discretization for the numerical solution of the variational formulation each domain is subdivided into a finite set of non-overlapping elements

$$\mathcal{B}^{(i)} \approx \mathcal{B}^{(i),\text{h}} = \bigcup_e^{n_{\text{el}}^{(i)}} \mathcal{B}^{(i),\text{h},e}\,. \tag{73}$$

For the second order phase-field model standard FEA with $C^0$ continuous shape functions is sufficient, whereas for the fourth order phase-field model we need to apply $C^1$ continuous shape functions which we accomplish using IGA.

**Standard FEA – Allen-Cahn phase-field model** For the second order phase-field model we are able to use standard Lagrangian shape functions and isoparametric as well as Bubnov-Galerkin FE interpolations. Accordingly, solution and test functions are approximated by

$$\boldsymbol{\varphi}^{(i),\text{h}} = L_A\, \boldsymbol{q}_A^{(i)}, \quad \delta\boldsymbol{\varphi}^{(i),\text{h}} = L_A\, \delta\boldsymbol{q}_A^{(i)}, \quad \boldsymbol{X}^{(i),\text{h}} = L_A\, \boldsymbol{X}_A^{(i)}\,, \tag{74}$$

$$\mathfrak{s}^{(i),\text{h}} = L_A\, \mathfrak{s}_A^{(i)}, \quad \delta\mathfrak{s}^{(i),\text{h}} = L_A\, \delta\mathfrak{s}_A^{(i)}, \quad \forall A \in \omega = \{1, \ldots, n_{\text{node}}\}\,. \tag{75}$$

Here $L_A^{(i)}(\boldsymbol{X}^{(i)}) : \mathcal{B}_0 \to \mathbb{R}$ denote Lagrangian shape functions. Inserting these approximations into the weak forms given in (70)–(72) yields the semi-discrete virtual work

$$\sum_i \delta\boldsymbol{q}_A^{(i)} \cdot (M_{AB}^{(i)} \ddot{\boldsymbol{q}}_B^{(i)} + \boldsymbol{F}_A^{(i),\text{int}} - \boldsymbol{F}_A^{(i),\text{ext}}) + \bar{G}_{\text{co}}^{\boldsymbol{\varphi},\text{h}} = 0\,, \tag{76}$$

$$\sum_i \delta\mathfrak{s}_A^{(i)} \int_{\mathcal{B}_0^{(i)}} (L_A^{(i)}\, \mathcal{H}^{(i),\text{h}} + \frac{\mathcal{G}_c^{(i)}}{l^{(i)}} L_A^{(i)} L_B^{(i)} \mathfrak{s}_B^{(i)} + 4\, \mathcal{G}_c^{(i)}\, l^{(i)} \nabla L_A^{(i)} \cdot \nabla L_B^{(i)} \mathfrak{s}_B^{(i)})\, \mathrm{d}V = 0\,, \tag{77}$$

$$\bar{G}_{\text{co}}^{\lambda_{\text{N}},\text{h}} = 0\,. \tag{78}$$

Here, $M_{AB}^{(i)} = \int_{\mathcal{B}_0} \rho_0 L_A^{(i)}\, L_B^{(i)}\, \mathrm{d}V$ are the coefficients of the consistent mass matrix. Furthermore, $\boldsymbol{F}_A^{(i),int}$, $\boldsymbol{F}_A^{(i),ext}$ denote the nodal internal and external force vectors. Note the semi-discrete contact contributions $\bar{G}_{\text{co}}^{\boldsymbol{\varphi},\text{h}}$ and $\bar{G}_{\text{co}}^{\lambda_{\text{N}},\text{h}}$ are dealt with in the very last paragraph of this section.

[4]Strictly speaking, the underlying virtual work (38) is not an equality but an inequality equation, due to the contact constraints (34)-(36) involved. As already mentioned, here and in what follows it is assumed that the contact interface is known using e.g. an active set strategy, regularization techniques or others. Based on this assumption the virtual work can be written as an equality (see e.g. [37, 35]).

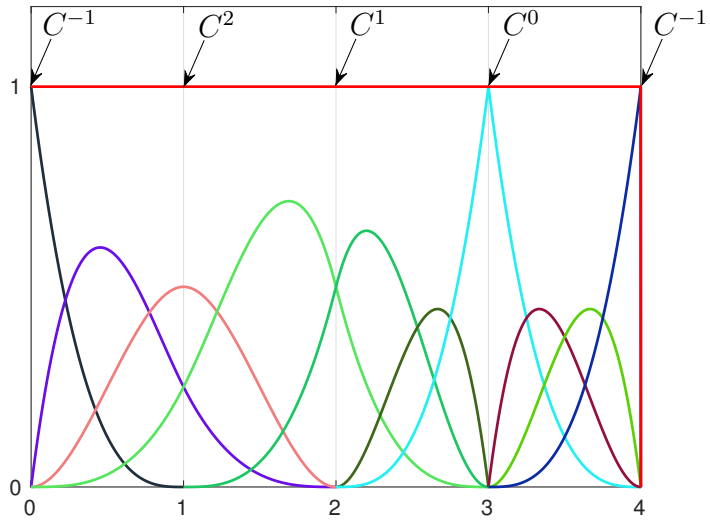



Figure 6: Continuity of cubic B-splines with knot vector $\Theta = [0\ 0\ 0\ 0\ 1\ 2\ 2\ 3\ 3\ 3\ 4\ 4\ 4\ 4]$.

**IGA – Cahn-Hilliard phase-field model** For the higher order Cahn-Hilliard phase-field model we use an IGA approach to satisfy the demand for $C^1$ continuity. For the approximation we use a linear combination of suitable splines and corresponding solution and test control variables

$$\boldsymbol{\varphi}^{(i),\mathrm{h}} = R_A^{(i)} \boldsymbol{q}_A^{(i)}, \quad \delta\boldsymbol{\varphi}^{(i),\mathrm{h}} = R_A^{(i)} \delta\boldsymbol{q}_A^{(i)}, \quad \mathfrak{s}^{(i),\mathrm{h}} = R_A^{(i)} \mathfrak{s}_A^{(i)}, \quad \delta\mathfrak{s}^{(i),\mathrm{h}} = R_A^{(i)} \delta\mathfrak{s}_A^{(i)}, \\ \forall A \in \omega = \{1, \ldots, n_{\mathrm{node}}\}. \tag{79}$$

Here $R_A^{(i)}(\boldsymbol{X}^{(i)}) : \mathcal{B}_0 \to \mathbb{R}$ denote multivariate NURBS with tensor product structure

$$R_A^{(i)} = R_{\boldsymbol{p}}^{(i),\boldsymbol{i}}(\boldsymbol{\xi}^{(i)}) = \frac{\Pi_{l=1}^{d} N_{i_l,p_l}^{(i)}(\xi^{(i),l})\, w_{i_l}^{(i)}}{\sum_{\hat{\boldsymbol{i}}} \Pi_{l=1}^{d} N_{\hat{i}_l,p_l}^{(i)}(\xi^{(i),l})\, w_{\hat{i}_l}^{(i)}}, \tag{80}$$

Therein $\boldsymbol{i} = [i_1, ..., i_d]$ denote the multi index of the B-splines $N_{i_l,p_l}^{(i)}$ with multi index $\boldsymbol{p} = [p_1, ..., p_d]$ for the respective order of each parametric direction. Moreover, $w_{\boldsymbol{i}}$ denote the NURBS weights. A comprehensive overview for IGA can be found in [6]. Note, for global $C^1$-continuity, which is required for the Cahn-Hilliard phase-field model, at least quadratic B-splines need to be provided. B-splines are recursively defined. For order $p_l = 0$ we have

$$N_{i_l,p_l=0}^{(i)}(\xi^{(i),l}) = \begin{cases} 1 & \text{if } \xi_{i_l}^{(i),l} \le \xi^{(i),l} < \xi_{i_l+1}^{(i),l} \\ 0 & \text{otherwise} \end{cases}, \tag{81}$$

whereas for order $p_l > 0$ the corresponding B-spline can be computed with

$$N_{i_l,p_l}^{(i)}(\xi^{(i),l}) = \frac{\xi^{(i),l} - \xi_{i_l}^{(i),l}}{\xi_{i_l+p_l}^{(i),l} - \xi_{i_l}^{(i),l}} N_{i_l,p-1}^{(i)}(\xi^{(i),l}) + \frac{\xi_{i_l+p_l+1}^{(i),l} - \xi^{(i),l}}{\xi_{i_l+p_l+1}^{(i),l} - \xi_{i_l+1}^{(i),l}} N_{i_l+1,p-1}^{(i)}(\xi^{(i),l}), \tag{82}$$

where problem specific knot vectors need to be provided

$$\Theta^{(i),l} = [\xi_1^{(i),l}, \ldots, \xi_{n_l+p_l+1}^{(i),l}]. \tag{83}$$

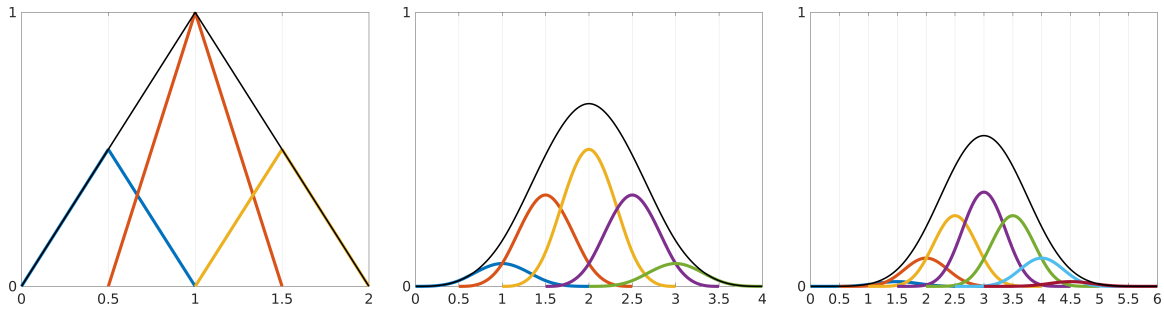



Figure 7: Hierarchical level 1 subdivision of 1D linear, cubic and quintic B-splines (black: original spline/parent, colored: 'refined' splines/children).

With the knot vectors both the finite element mesh as well as the continuity is determined (see Figure 6). B-splines and NURBS basis functions are local linear independent

$$\sum_A c_A R_A^{(i)}(\boldsymbol{\xi}^{(i)}) = 0 \Leftrightarrow c_A = 0\,. \tag{84}$$

Furthermore the partition of unity is fulfilled

$$\sum_A R_A^{(i)}(\boldsymbol{\xi}^{(i)}) = 1\,. \tag{85}$$

Another property is that B-splines as well as NURBS only have minimal support. In general B-splines and NURBS are non-interpolatory but can be constructed that they are. B-splines and NURBS basis functions are always positive, i.e.

$$R_A^{(i)}(\boldsymbol{\xi}^{(i)}) > 0\,. \tag{86}$$

As refinement techniques h-refinement e.g. achieved via knot insertion leads to expensive global refinement. For local refinement techniques there mainly exists T-splines, which are the IGA representation of hanging nodes and the concept of hierarchical refinement which we want to apply subsequently. For hierarchical B-splines the basis functions are subdivided instead of the elements. In particular the B-spline to be refined (parent) is replaced by using a linear combination of copied and scaled version of the original B-spline (children) via

$$B^{(i),A} = B_{\boldsymbol{p}}^{(i),\boldsymbol{i}}(\boldsymbol{\xi}^{(i)}) = \sum_{\boldsymbol{j}=\boldsymbol{0}}^{\boldsymbol{p+1}} \prod_{l=1}^{d} 2^{-p_l} \binom{p_l+1}{j_l} N_{i_l,p_l}^{(i)}(2\xi^{(i),l} - j_l^{(i)} h_l^{(i)} - \xi_{i_l}^{(i),l})\,. \tag{87}$$

Therein $h_l$ refers to the coarse level element length in the parameter space. The refinement procedure for one level is shown for the one-dimensional case in Figure 7 and for the two-dimensional case in Figure 8. Next we rewrite the hierarchical B-splines given in equation (87) approach by using the subdivision matrix $\boldsymbol{S}^{(i)}$, $S_{\boldsymbol{i},\boldsymbol{j}}^{(i)}$ which contains the scaling. A single B-spline on level $k$ can then be replaced by its children, which correspond to B-splines on level $k+1$

$$B_{\boldsymbol{p}}^{(i),\boldsymbol{i},k}(\boldsymbol{\xi}^{(i)}) = \sum_{\boldsymbol{j}=\boldsymbol{0}}^{\boldsymbol{p+1}} S_{\boldsymbol{i},\boldsymbol{j}}^{(i)} B_{\boldsymbol{p}}^{(i),2\boldsymbol{i}-\boldsymbol{1}+\boldsymbol{j},k+1}(\boldsymbol{\xi}^{(i)})\,. \tag{88}$$

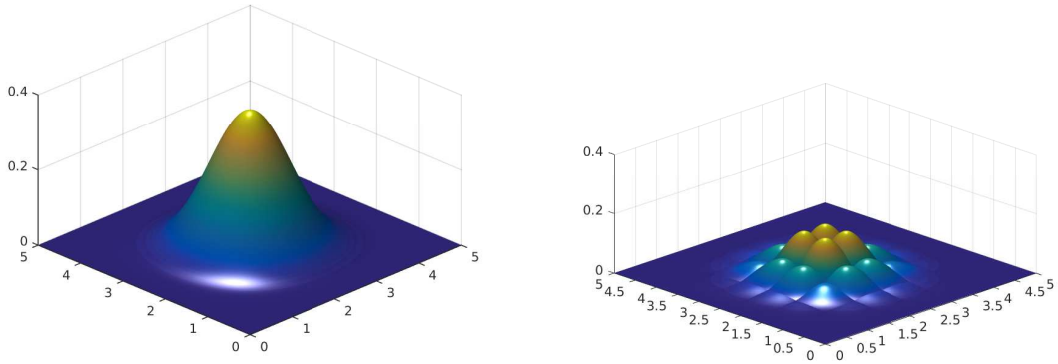


Figure 8: Subdivision of a 2D quartic B-spline (left: parent, right: children).

This is only valid for uniform knot vectors (for non-uniform knot vectors, see Sabin [32]). Afterwards, the control mesh needs to be extended by

$$\boldsymbol{q}_A^{(i),k+1} = \boldsymbol{S}_A^{(i),\mathrm{T}}\,\boldsymbol{q}_A^{(i),k}, \quad \mathfrak{s}_A^{(i),k+1} = \boldsymbol{S}_A^{(i),\mathrm{T}}\,\mathfrak{s}_A^{(i),k}\,, \tag{89}$$

where $\boldsymbol{q}_A^{(i),k+1}$ represents a set of new coordinates. The extension of B-splines to NURBS is given by

$$R_A^{(i),k} = R_{\boldsymbol{p}}^{(i),\boldsymbol{i},k}(\boldsymbol{\xi}^{(i)}) = \frac{\sum\limits_{\boldsymbol{j}=\boldsymbol{0}}^{\boldsymbol{p+1}} S_{\boldsymbol{i},\boldsymbol{j}}^{(i)} B_{\boldsymbol{p}}^{(i),\boldsymbol{2i-1+j},k+1}(\boldsymbol{\xi}^{(i)}) w_{\boldsymbol{2i-1+j}}^{(i),k+1}}{\sum\limits_{\boldsymbol{i}}\sum\limits_{\boldsymbol{j}=\boldsymbol{0}}^{\boldsymbol{p+1}} S_{\boldsymbol{i},\boldsymbol{j}} B_{\boldsymbol{p}}^{(i),\boldsymbol{2i-1+j},k+1}(\boldsymbol{\xi}^{(i)}) w_{\boldsymbol{2i-1+j}}^{(i),k+1}}\,. \tag{90}$$

The partition of unity as well as the continuity are still satisfied. For further details see [29]. Insertion of the polynomial approximations into the virtual work of the coupled phase-field system yields

$$\sum_i \delta\boldsymbol{q}_A^{(i)} \cdot (\tilde{M}_{AB}^{(i)} \ddot{\boldsymbol{q}}_B^{(i)} + \tilde{\boldsymbol{F}}_A^{(i),\mathrm{int}} - \tilde{\boldsymbol{F}}_A^{(i),\mathrm{ext}}) + G_{\mathrm{co}}^{\varphi,\mathrm{h}} = 0\,, \tag{91}$$

$$\sum_i \delta\mathfrak{s}_A^{(i)} \int\limits_{\mathcal{B}_0^{(i)}} (R_A^{(i)}\,\mathcal{H}^{(i),\mathrm{h}} + \frac{\mathcal{G}_c^{(i)}}{l^{(i)}} R_A^{(i)} R_B^{(i)} \mathfrak{s}_B^{(i)} + 4\,\mathcal{G}_c^{(i)}\,l^{(i)} \nabla R_A^{(i)} \cdot \nabla R_B^{(i)} \mathfrak{s}_B^{(i)})\;\mathrm{d}V = 0\,, \tag{92}$$

$$\bar{G}_{\mathrm{co}}^{\lambda_{\mathrm{N}},\mathrm{h}} = 0\,. \tag{93}$$

Here, $\tilde{M}_{AB}^{(i)} = \int_{\mathcal{B}_0} \rho_0 R_A^{(i)} R_B^{(i)}\;\mathrm{d}V$ are the coefficients of the consistent mass matrix. Furthermore, $\tilde{\boldsymbol{F}}_A^{(i),int}$, $\tilde{\boldsymbol{F}}_A^{(i),ext}$ denote the nodal internal and external force vectors, approximated with HNURBS. As constitutive relations the discrete second Piola-Kirchhoff stress tensor as well as the driving force for the phase-field

$$\boldsymbol{S}_{\mathfrak{e}}^{(i),\mathrm{h}} = 2\;\mathrm{D}_1\Psi_{\mathfrak{e}}^{(i)}(\boldsymbol{C}^{(i),\mathrm{h}}, \mathfrak{s}^{(i),\mathrm{h}}), \quad \mathcal{H}^{(i),\mathrm{h}} = \mathrm{D}_2\Psi_{\mathfrak{e}}^{(i)}(\boldsymbol{C}^{(i),\mathrm{h}}, \mathfrak{s}^{(i),\mathrm{h}})\,, \tag{94}$$

need to be provided.

**Mortar contact element** For the discretization of the contact interface we aim at a Mortar approach. Herein, we restrict our consideration to the IGA case for the Lagrange discretized case we refer to the Mortar method given in [15]. In order to simplify the approach without violation of the required $C^1$ continuity of the primal space we use linear approximation of the dual space as has been proposed in [17, 7]. Accordingly, the contact traction is approximated as follows

$$\boldsymbol{t}_{\text{co}}^{(1),\text{h}} = \sum_{A\in\tilde{\omega}^{(1)}} L_A \boldsymbol{\lambda}_A, \quad \delta\boldsymbol{t}_{\text{co}}^{(1),\text{h}} = \sum_{A\in\tilde{\omega}^{(1)}} L_A \delta\boldsymbol{\lambda}_A \,. \tag{95}$$

Therein $\tilde{\omega}^{(1)} = [\tilde{\boldsymbol{q}}_1, \ldots, \tilde{\boldsymbol{q}}_{n_{\text{surf}}}]$ denotes $n_{\text{surf}}$ nodes on the physical contact boundary $\gamma_{\text{co}}^{(1)}$. Moreover, $L_A : \Gamma_{\text{co}}^{(1)} \to \mathbb{R}$ are $(d-1)$-dimensional shape functions associated with nodes $A \in \tilde{\omega}^{(1)}$. It is important to remark, that the discrete Lagrange multiplier space for refined contact regions can provide possible singularities (for more informations see [16]). Eventually, we are able to state the semi-discrete contact contribution

$$\bar{G}_{\text{co}}^{\boldsymbol{\varphi},\text{h}} = \lambda_{\text{N},A}\boldsymbol{n}\cdot\left[n_{AB}\delta\boldsymbol{q}_B^{(1)} - n_{AC}\delta\boldsymbol{q}_C^{(2)}\right] + \lambda_{\text{N},A}\delta\boldsymbol{n}\cdot\left[n_{AB}\boldsymbol{q}_B^{(1)} - n_{AC}\boldsymbol{q}_C^{(2)}\right] . \tag{96}$$

Therein the nodal normal traction $\lambda_{\text{N},A}$ is employed. Furthermore, the Mortar integrals

$$n_{AB} = \int\limits_{\bar{\Gamma}_{\text{co}}^{(1),\text{h}}} L_A(\boldsymbol{\xi}^{(1)}) R_B(\boldsymbol{\xi}^{(1)}) \,\mathrm{d}A, \tag{97}$$

$$n_{AC} = \int\limits_{\bar{\Gamma}_{\text{co}}^{(1),\text{h}}} L_A(\boldsymbol{\xi}^{(1)}) R_C(\boldsymbol{\xi}^{(2)}) \,\mathrm{d}A\,, \tag{98}$$

are comprised of the above introduced linear shape functions $L_A(\boldsymbol{\xi}^{(1)})$ for the dual and quadratic shape functions $R_B(\boldsymbol{\xi}^{(1)})$, $R_C(\boldsymbol{\xi}^{(2)})$ for the primal space. In order to be able to integrate the contact contributions we provide an isoparametric transformation

$$\boldsymbol{\xi}^{(i),\text{h}}(\boldsymbol{\eta}) = \sum_{K=1}^{3} M_K(\boldsymbol{\eta})\boldsymbol{\xi}_K^{(i)}\,, \tag{99}$$

where we make use of bilinear, triangular shape functions $M_K$. The segment-wise Mortar integrals are then given by

$$n_{\kappa\beta} = \int\limits_{\triangle} L_\kappa(\boldsymbol{\xi}^{(1),\text{h}}(\boldsymbol{\eta})) R_\beta(\boldsymbol{\xi}^{(1),\text{h}}(\boldsymbol{\eta})) J_{\text{seg}} \,\mathrm{d}\eta\,, \tag{100}$$

$$n_{\kappa\zeta} = \int\limits_{\triangle} L_\kappa(\boldsymbol{\xi}^{(1),\text{h}}(\boldsymbol{\eta})) R_\zeta(\boldsymbol{\xi}^{(2),\text{h}}(\boldsymbol{\eta})) J_{\text{seg}} \,\mathrm{d}\eta\,. \tag{101}$$

Therein $J_{\text{seg}}$ denotes the segment-wise Jacobian. Using the tangential vectors in the reference configuration $\boldsymbol{A}_\alpha(\boldsymbol{\xi}) = R_{A,\alpha}(\boldsymbol{\xi})\boldsymbol{q}_A$ the segment-wise Jacobian can be computed by

$$J_{\text{seg}} = \|\boldsymbol{A}_1(\boldsymbol{\xi}^{(1),\text{h}}(\boldsymbol{\eta})) \times \boldsymbol{A}_2(\boldsymbol{\xi}^{(1),\text{h}}(\boldsymbol{\eta}))\| \det(\,\mathrm{D}\boldsymbol{\xi}(\boldsymbol{\eta}))\,. \tag{102}$$

Eventually the segment-wise contact contributions needs to be assembled into the global system. For a more detailed explanation of the Mortar method, especially the technically demanding segmentation and the assembly see [14, 15, 17, 7].

## 4 Temporal discretization

For the temporal discretization we subdivide the time interval $\mathcal{I}$ of interest into $N$ equidistant increments $\Delta t = t_{n+1} - t_{\mathrm{n}}$, as follows

$$\mathcal{I} = [0, T] = \bigcup_{n=0}^{N-1} [t_n, t_{n+1}] \,, \tag{103}$$

where in the following the variables $\boldsymbol{q}_A^n$, $\mathfrak{s}_A^n$ at time $t_n$ are assumed to be known, whereas the variables $\boldsymbol{q}_A^{n+1}$, $\mathfrak{s}_A^{n+1}$ at time $t_{n+1}$ are searched. Then we employ a second order accurate one step midpoint-type discretization. In particular the fully discrete equations

$$\begin{aligned}
\sum_i \delta \boldsymbol{q}_A^{(i)} \cdot (\tilde{M}_{AB}^{(i)} \frac{\boldsymbol{q}_B^{(i),n+1} - \boldsymbol{q}_B^{(i),n}}{\Delta t} + \int_{\mathcal{B}_0^{(i)}} \nabla R_A^{(i)}(\boldsymbol{X}^{(i)}) \cdot \boldsymbol{S}_{\mathfrak{e}}^{(i),n,n+1} \nabla R_B^{(i)}(\boldsymbol{X}^{(i)}) \, \mathrm{d}V \, \boldsymbol{q}_B^{(i),n+\frac{1}{2}} \\
- \int_{\mathcal{B}_0^{(i)}} \tilde{\boldsymbol{F}}_A^{(i),\mathrm{ext},n+\frac{1}{2}} \, \mathrm{d}V) + G_{\mathrm{co}}^{\boldsymbol{\varphi},n,n+1} = 0 \,,
\end{aligned} \tag{104}$$

$$\begin{aligned}
\sum_i \delta \mathfrak{s}_A^{(i)} \int_{\mathcal{B}_0^{(i)}} R_A^{(i)} \mathcal{H}^{(i),n+\frac{1}{2}} + \frac{\mathcal{G}_c}{2l} R_A^{(i)} R_B^{(i)} \mathfrak{s}_B^{(i),n+\frac{1}{2}} + \mathcal{G}_c \, l \nabla R_A^{(i)} \cdot \nabla R_B^{(i)} \mathfrak{s}_B^{(i),n+\frac{1}{2}} \\
+ \frac{\mathcal{G}_c \, l^3}{2} \Delta R_A^{(i)} \Delta R_B^{(i)} \mathfrak{s}_B^{(i),n+\frac{1}{2}} \, \mathrm{d}V = 0 \,,
\end{aligned} \tag{105}$$

$$G_{\mathrm{co}}^{\lambda_{\mathrm{N}},n,n+1} = 0 \,, \tag{106}$$

are obtained. In order to obtain a structure preserving time integration scheme, an algorithmic stress computation

$$\begin{aligned}
\boldsymbol{S}_{\mathfrak{e}}^{(i),n,n+1} = & \, 2 \frac{\partial \Psi^{(i),n+\frac{1}{2}}}{\partial \boldsymbol{C}^{(i),\mathrm{h}}} \\
& + 2 \frac{\Psi_{\mathfrak{e}}^{(i),n+1} - \Psi_{\mathfrak{e}}^{(i),n} + \mathcal{G}_c^{(i)} (\gamma^{(i),n+1} - \gamma^{(i),n}) - \frac{\partial \Psi_{\mathfrak{e}}^{(i),n+\frac{1}{2}}}{\partial \boldsymbol{C}^{(i),\mathrm{h}}} : \Delta \boldsymbol{C}^{(i),\mathrm{h}}}{\Delta \boldsymbol{C}^{(i),\mathrm{h}} : \Delta \boldsymbol{C}^{(i),\mathrm{h}}} \Delta \boldsymbol{C}^{(i),\mathrm{h}} \,,
\end{aligned} \tag{107}$$

using the concept of the discrete gradient is employed. Note that the concept of the discrete gradient exhibits superior stability and robustness properties even for large time steps. The fully discrete contact contributions can be calculated by

$$\begin{aligned}
G_{\mathrm{co}}^{\boldsymbol{\varphi},n,n+1} &= \lambda_A^{n,n+1} \boldsymbol{n}^{n+\frac{1}{2}} \cdot \left[ n_{AB}^n \delta \boldsymbol{q}_B^{(1)} - n_{AC}^n \delta \boldsymbol{q}_C^{(2)} \right] + \delta \lambda_A \delta \boldsymbol{n}^{n+\frac{1}{2}} \cdot \left[ n_{AB}^n \boldsymbol{q}_B^{(1)} - n_{AC}^n \boldsymbol{q}_C^{(2)} \right] \,, \\
G_{\mathrm{co}}^{\lambda_{\mathrm{N}},n,n+1} &= \delta \lambda_A \boldsymbol{n}^{n+\frac{1}{2}} \cdot \left[ n_{AB}^n \, \boldsymbol{q}_B^{(1)} - n_{AC}^n \, \boldsymbol{q}_C^{(2)} \right] \,.
\end{aligned}$$

## 5 Numerical examples

In order to show the enhanced properties of the proposed methods two representative numerical examples are examined. In particular the proposed hierarchical NURBS based discretization with the Mortar contact method and the phase-field approach to fracture are investigated in the examples, where local refinements are predefined. Note, for the simple geometries employed, all NURBS weights are chosen equal to one, such that the NURBS can be considered as B-splines.

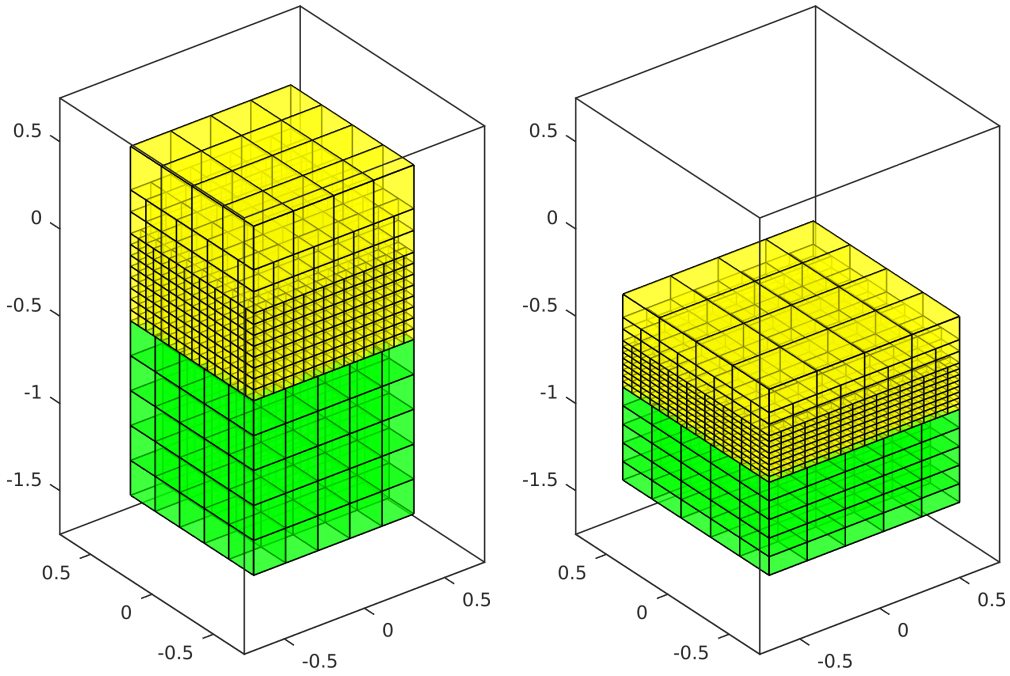



Figure 9: Patch test with two level local refinement of the upper body, where reference (left) and current configuration (right) are shown.

**Patch test** As first example we investigate the local refinement with the proposed IGA approach within a patch test (cf. [16]). The initial configuration is depicted in Figure 9. As usual for patch tests, the lower block is clamped, such that the body is able to expand in tangential directions. Furthermore, we apply a Neumann boundary to the upper surface on the upper body with constant pressure field of $\sigma = -1.3e3$.

We use quadratic B-splines and apply a two level refinement for the upper body and no local refinement for the lower body, such that we obtain non-conforming finite element meshes in contact. Large areas between coarse and fully refined elements can be observed which is due to the support of the splines. The upper body consists of $4 \times 4 \times 4$ elements on level 0. In total the upper body consists of 2192 elements with 10224 degrees of freedom (DOFs). The lower body consists of $5 \times 5 \times 5$ elements on level 0. The lower body consists of 125 elements with in total 1029 DOFs.

A compressible neo-Hookean material model is employed with Lamé parameters $\lambda^{(i)} = 1298.1$ and $\mu^{(i)} = 865.3846$, which correspond to a Young's modulus of $E^{(i)} = 2250$ and a Poisson ratio of $\nu^{(i)} = 0.3$, respectively. The corresponding strain energy density function is given by

$$\Psi^{(i)}(\boldsymbol{C}^{(i)}) = \frac{\mu^{(i)}}{2}\left(\mathrm{tr}(\boldsymbol{C}^{(i)}) - 3\right) + \frac{\lambda^{(i)}}{2}\left(\ln(J^{(i)})\right)^2 - \mu^{(i)}\,\ln(J^{(i)})\,, \tag{108}$$

where $J^{(i)} = \det(\boldsymbol{F}^{(i)})$. We are interested in a quasi-static solution, accordingly the density is neglected for both bodies.

We obtain a uniform stress distribution as can be observed in Figure 10. The corresponding displacement solution is shown in Figure 9. Accordingly, the Mortar method with higher order hierarchical NURBS approach and local refinement passes the patch test (cf. [16]).

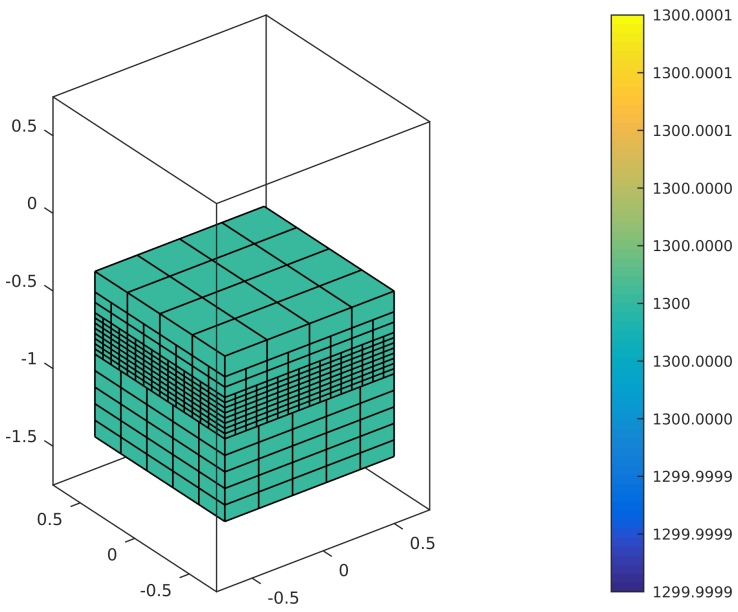



Figure 10: Von Mises stress distribution.

**Bending contact fracture problem** The next example deals with the contact of an elastic block with an elastic plate. Note, this example is taken from [16]. The initial configuration is depicted in Figure 11. As can be observed, the plate is clamped whereas the upper boundary of the block is moved downwards with a constant increment size of $\Delta \boldsymbol{u} = 5e\text{-}3\, \boldsymbol{e}_3$ for the first 200 steps and $\Delta \boldsymbol{u} = 2.5 \times 10^{-3}\, \boldsymbol{e}_3$ for the remaining steps.

The plate is of size $20 \times 30 \times 2$ , whereas the block is of size $4 \times 4 \times 4$. The center point of the plate is placed at $\begin{bmatrix} 0 & 15 & 1 \end{bmatrix}$ and the center point of the block is placed at $\begin{bmatrix} 0 & 26 & 4.5 \end{bmatrix}$. The block consists of $4 \times 4 \times 4$ elements and the plate of $13 \times 19 \times 2$ elements on level 0. Additionally, we apply a predefined local hierarchical refinement to account for the physical events (contact and fracture). Therefore the contact zone of the block is locally refined with one level. The plate is locally refined using one level refinement in the area of the expected contact boundary and two level refined at the clamping zone (see Figure 11). In total 12948 elements with overall 72912 DOFs and minimal element size of $h_{\text{min}} = 0.0769$ are employed.

Both bodies are modeled using the compressible neo-Hookean material model given in (108). The Lamé parameters of the plate are chosen as $\lambda^{(1)} = 1.1538e7$ and $\mu^{(1)} = 7.6923e6$, which

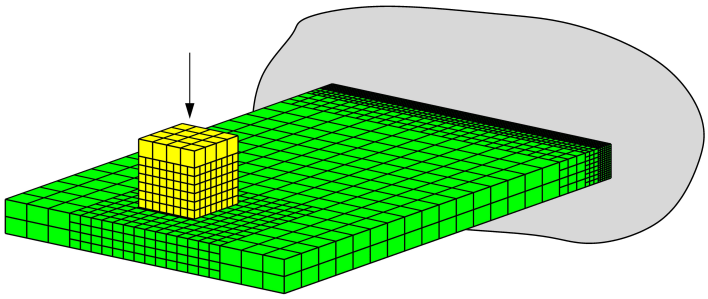


Figure 11: Reference configuration of bending contact fracture problem.

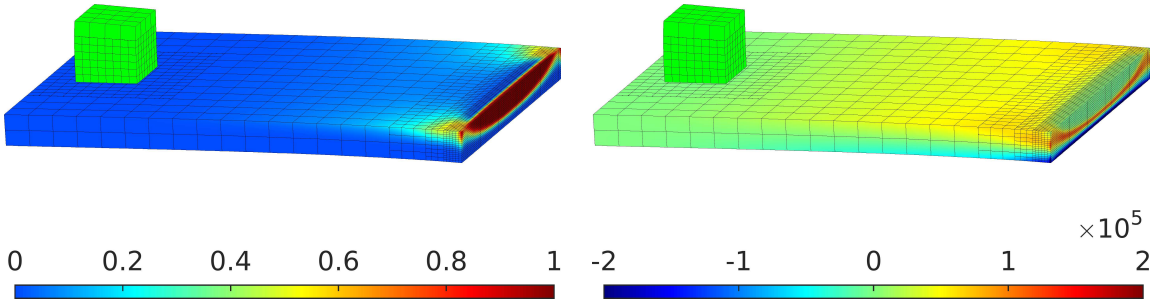



Figure 12: Final phase-field (left) and $\sigma_{22}$ stress (right) result of bending contact fracture problem.

correspond to a Young's modulus of $E^{(1)} = 2e7$ and a Poisson ratio of $\nu^{(1)} = 0.3$. The Lamé parameters of the block are chosen as $\lambda^{(2)} = 2.8846e4$ and $\mu^{(2)} = 1.9231e4$, which correspond to a Young's modulus of $E^{(2)} = 5e4$ and a Poisson ratio of $\nu^{(2)} = 0.3$. Accordingly, we expect no locking behavior and there is no need for an advanced finite element model. For the plate the phase-field parameters are chosen as $\mathcal{G}_c = 2.7 \cdot e2$ and $l = 0.1538$.

The final phase-field result as well as the final $\sigma_{22}$ stress distribution are depicted in Figure 12. As can be observed the plate ripped out of the clamping zone, such that the plate becomes statically undetermined. Accordingly, the problem can not be solved anymore with the considered quasi-static setup. Note, a more detailed explanation and examination of this example is given in [16].

## 6 CONCLUSIONS

The underlying contribution introduces a large deformation framework to describe both contact and fracture. For the former the variational consistent Mortar method, whereas for the latter the phase-field method is used. In particular an accurate and smooth fourth order phase-field model is employed which requires global $C^1$ continuity. Therefore an existing IGA framework with a sophisticated Mortar method for the contact boundaries is combined. With the proposed NURBS-IGA approach a wide range of curved geometries can be approximated exactly. Moreover, an hierarchical refinement scheme is employed which does not violate the partition of unity. This allows the local refinement of the bodies in order to account for different physical events like e.g. contact or fracture.

In total an accurate and efficient higher order method is proposed for computational modeling of large deformation Mortar contact and phase-field fracture problems.

## REFERENCES

[1] A.-V. Vuong and C. Giannelli and B. Jüttler and B. Simeon. A hierarchical approach to adaptive local refinement in isogeometric analysis. *Comput. Methods Appl. Mech. Engrg.*, 200:3554–3567, 2011.

[2] M.J. Borden. *Isogeometric Analysis of Phase-Field Models for Dynamic Brittle and Ductile Fracture*. PhD thesis, The University of Texas at Austin, 2012.

[3] M.J. Borden, T.J.R. Hughes, C.M. Landis, and C.V. Verhoosel. A higher-order phase-field model for brittle fracture: Formulation and analysis within the isogeometric analysis framework. *Comput. Methods Appl. Mech. Engrg.*, 273:100–118, 2014.

[4] P.B. Bornemann and F. Cirak. A subdivision-based implementation of the hierarchical b-spline finite element method. *Comput. Methods Appl. Mech. Engrg.*, 253:584–598, 2013.

[5] B. Bourdin, G.A. Francfort, and J.J. Marigo. The variational approach to fracture. *Journal of Elasticity*, 9:5–148, 2008.

[6] J.A. Cottrell, T.J.R. Hughes, and Y. Bazilevs. *Isogeometric Analysis: Toward Integration of CAD and FEA*. Wiley, New York, 2009.

[7] M. Dittmann, M. Franke, I. Temizer, and C. Hesch. Isogeometric Analysis and thermomechanical Mortar contact problems. *Comput. Methods Appl. Mech. Engrg.*, 274:192–212, 2014.

[8] D. Forsey and R.H. Bartels. Hierarchical B-Spline refinement. *Comput. Graph.*, 22:205–212, 1988.

[9] G.A. Francfort and J.-J. Marigo. Revisiting brittle fracture as an energy minimization problem. *Journal of the Mechanics and Physics of Solids*, 46:1319–1342, 1998.

[10] M. Franke. *Discretisation techniques for large deformation computational contact elastodynamics*. PhD thesis, Karlsruhe Institute of Technology, 2014.

[11] A.A. Griffith. The phenomena of rupture and flow in solids. *Philosophical Transactions of the Royal Society London*, 221:163–198, 1921.

[12] H. Henry. Study of the branching instability using a phase field model of inplane crack propagation. *EPL*, 83:16004, 2008.

[13] H. Henry and H. Levine. Dynamic instabilities of fracture under biaxial strain using a phase field model. *Physics Review Letters*, 93:105505, 2004.

[14] C. Hesch and P. Betsch. Transient 3D Domain Decomposition Problems: Frame-indifferent mortar constraints and conserving integration. *Int. J. Numer. Methods Eng.*, 82:329–358, 2010.

[15] C. Hesch and P. Betsch. Transient three-dimensional contact problems: mortar method. Mixed methods and conserving integration. *Computational Mechanics*, 48:461–475, 2011.

[16] C. Hesch, M. Franke, M. Dittmann, and İ. Temizer. Hierarchical NURBS and a higher-order phase-field approach to fracture for finite-deformation contact problems. *Computer Methods in Applied Mechanics and Engineering*, 301:242–258, 2016.

[17] C. Hesch and P.Betsch. Isogeometric analysis and domain decomposition methods. *Comput. Methods Appl. Mech. Engrg.*, 213:104–112, 2012.

[18] C. Hesch and K. Weinberg. Thermodynamically consistent algorithms for a finite-deformation phase-field approach to fracture. *Int. J. Numer. Meth. Engng.*, 99:906–924, 2014.

[19] Schuß S. Dittmann M. Franke M. Hesch, C. and K. Weinberg. Isogeometric analysis and hierarchical refinement for higher-order phase-field models. *Computer Methods in Applied Mechanics and Engineering*, 303:185–207, 2016.

[20] M. Hintermueller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semismooth Newton method. *SIAM J. Optim.*, 13:865–888, 2003.

[21] G.R. Irwin. Elasticity and plasticity: fracture. In S. Függe, editor, *Encyclopedia of Physics*, 1958.

[22] W. Jiang and J.E. Dolbow. Adaptive refinement of hierarchical B-spline finite elements with an efficient data transfer algorithm. *Int. J. Numer. Meth. Engng.*, 2014.

[23] A. Karma, D.A. Kessler, and H. Levine. Phase-field model of mode III dynamic fracture. *Physical Review Letter*, 81:045501, 2001.

[24] C. Kuhn and R. Müller. A continuum phase field model for fracture. *Engineering Fracture Mechanics*, 77:3625–3634, 2010.

[25] T.A. Laursen. *Computational Contact and Impact Mechanics*. Springer-Verlag, 2002.

[26] C. Miehe, M. Hofacker, and F. Welschinger. A phase field model for rate-independent crack propagation: Robust algorithmic implementation based on operator splits. *Comput. Methods Appl. Mech. Engrg.*, 199:2765–2778, 2010.

[27] C. Miehe and L.M. Schänzel. Phase field modeling of fracture in rubbery polymers. Part I: Finite elasticity coupled with brittle failure. *Journal of the Mechanics and Physics of Solids*, page in press, 2013.

[28] C. Miehe, F. Welschinger, and M. Hofacker. Thermodynamically consistent phase-field models of fracture: Variational principles and multi-field FE implementations. *Int. J. Numer. Meth. Engng.*, 83:1273–1311, 2010.

[29] R. Ortigosa, A.J. Gil, J. Bonet, and C. Hesch. A computational framework for polyconvex large strain elasticity for geometrically exact beam theory. *Computational mechanics*, 2015, DOI:10.1007/s00466-015-1231-5.

[30] A. Popp, M. Gitterle, W. Gee, and W.A. Wall. A dual mortar approach for 3D finite deformation contact with consistent linearization. *Int. J. Numer. Methods Eng.*, 83:1428–1465, 2010.

[31] M.A. Puso. A 3D mortar method for solid mechanics. *Int. J. Numer. Methods Eng.*, 59:315–336, 2004.

[32] M. Sabin, editor. *Analysis and Design of Univariate Subdivision Schemes*. Springer, Heidelberg, 2010.

[33] D. Schillinger, L. Dede, M.A. Scott, J.A. Evans, M.J. Borden, E. Rank, and T.J.R. Hughes. An isogeometric design–through–analysis methodology based on adaptive hierarchical refinement of NURBS, immersed boundary methods, and T–spline CAD surfaces. *Comput. Methods Appl. Mech. Engrg.*, 249–252:116–150, 2012.

[34] K. Weinberg and C. Hesch. A high-order finite-deformation phase-field approach to fracture. *Continuum Mechanics and Thermodynamics*, pages 1–11, 2015.

[35] K. Willner. *Kontinuums- und Kontaktmechanik*. Springer-Verlag, 2003.

[36] B.I. Wohlmuth. Variationally consistent discretization schemes and numerical algorithms for contact problems. *Acta Numerica*, 20:569–734, 2011.

[37] P. Wriggers. *Computational Contact Mechanics*. Springer-Verlag, 2nd edition, 2006.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MODELLING THERMAL BARRIER COATINGS AND THEIR INFLUENCE ON THE LIFETIME OF ROCKET ENGINE NOZZLE STRUCTURES

**Marek Fassin, Stephan Wulfinghoff, and Stefanie Reese**

Institute of Applied Mechanics
Mies-van-der-Rohe-Str.1, D-52074 Aachen (Germany)
e-mail: {marek.fassin,stephan.wulfinghoff,stefanie.reese}@rwth-aachen.de

**Keywords:** viscoplasticity, damage, thermal barrier coating, lifetime prediction

**Abstract.** *The walls of rocket engine nozzle structures undergo extreme thermomechanical loadings during the engine cycle. These loads cause an accumulation of plastic strains due to creep and relaxation which finally lead to the failure of the hot gas wall. One approach to increase the lifetime of the hot gas wall is the reduction of the temperatures in the hot gas wall since the temperature is the main driver of this failure phenomenon. Special thermal barrier coating (TBC) systems are under development in order to protect the copper alloy of the hot gas wall from high temperatures. The present paper studies the influence of such thermal barrier coatings regarding the damage behaviour of the hot gas wall made of a copper alloy. In the first step, thermal analyses are performed to define an appropriate TBC thickness and the needed coolant power such that the maximum service temperature of the TBC is not exceeded. Furthermore the temperature distributions of the hot gas wall with and without TBC are compared, which show a reduction in temperature of approximately* $200$ *K. In the second step static analyses with a recently developed viscoplastic damage model are presented. Here damage distributions as well as the deformation behaviour of the hot gas wall with and without TBC are compared. It is shown that the application of TBCs has an positive effect on the damage behaviour as well as the deformation of the hot gas wall.*

## 1 INTRODUCTION

The demand for higher payloads and the general endeavour of a performance increase of rocket engines require the development of reliable lifetime prediction models. During the operation of the rocket engine the hot gas wall undergoes extreme cyclic thermomechanical loads. These cyclic loads lead to an accumulation of plastic strains, followed by thinning and bulging of the hot gas wall. Finally, after several cycles the hot gas wall (usually made of a copper alloy) fails (see Figure 1). Since the deformed hot gas wall resembles the shape of a doghouse, the failure mode is called doghouse effect (see e.g Riccius et al. [1]). This failure phenomenon can be rather related to ductile damage than to low cycle fatigue (see e.g. Schwarz et al. [2]). One

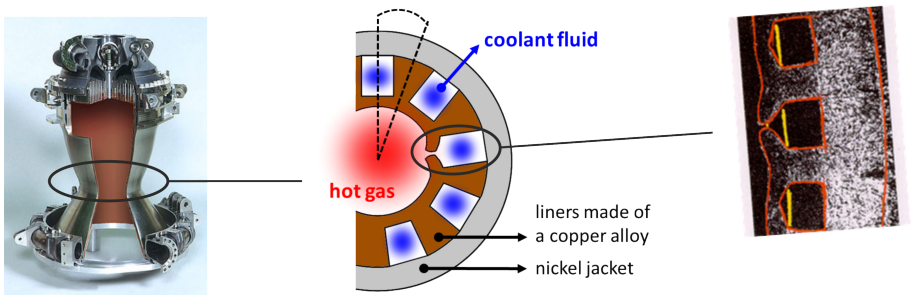



Figure 1: Left: Combustion chamber of the Vulcain rocket engine, Center: schematic cross section of the combustion chamber, Right: microscopic picture of the doghouse effect [1].

approach to increase the lifetime of the hot gas wall is the reduction of the temperatures in the hot gas wall. This seems to be a reasonable idea since the damage and failure behaviour of the copper alloy/copper substrate is highly temperature dependent. Special thermal barrier coating (TBC) systems for the application on copper alloys have been developed by the Institute for Materials at TU Braunschweig in order to protect the copper alloy from extremely high temperatures [3, 4, 5, 6, 7, 8, 9]. A thermal barrier coating typically consists of two layers: a bond coat and a top coat. In previous studies of the Institute for Materials [3, 4, 5, 6] a coating system usually used in gas turbines was investigated (bond coat: NiCrAlY, top coat: yttria-stabilized zirconia). However, this system turned out not be practical since the coatings failed at the bond coat/copper substrate interface due to (i) mismatch of the coefficient of thermal expansion (CTE) and (ii) the chemical gradient between the copper alloy substrate and the bond coat [5]. Therefore, a new bond coat NiCuCrAl was developed offering higher chemical similarity to the substrate as well as reducing the mismatch in the CTE. The current promising thermal barrier coating system is composed of a NiCuCrAl bond coat and a NiCrAlY top coat (see Fiedler et al. [8, 9]).

Butler and Pindera [10] investigated as one of the first the influence of thermal barrier coatings on the doghouse effect. They found that by the application of TBCs the temperature distribution within the hot gas wall was dramatically changed and reduced. They concluded that the strongly temperature dependent creep/relaxation of the copper alloy which is one of the main reasons for the doghouse failure can be limited by TBCs. By utilizng a viscoplastic material model they showed that the thickness of the TBC directly influences the deformation of the hot gas wall. Fiedler et al. [9] investigated with FE-simulations based on a micro model the interface stresses between the copper substrate and the bond coat. As desired, the interface stresses perpendicular to the surface are reduced strongly compared to the former bond coat made of NiCrAlY. Furthermore a temperature reduction in the copper substrate of approximately 200 K was observed.

The present study focuses on the lifetime of the copper alloy (which shall be increased by the application of a TBC), not on the lifetime of the TBC. Figure 2 left shows the representative segment of the cooling channel structure, which is highlighted in Figure 1 (center) by the dashed line, with the additional TBC on the hot gas wall (marked in red). Figure 2 (center) shows the FE model of the corresponding finite element model, where the TBC is discretized with several elements over the thickness. In reality, the TBC is a system of two layers: a top coat and a bond coat. Due to simplicity and the similarity of the thermal and mechanical properties of the two layers, the present study merges the two layers to one layer called "TBC" in the following. This approach is considered to be sufficient for the investigation of the lifetime/behavior of the subjacent copper alloy.

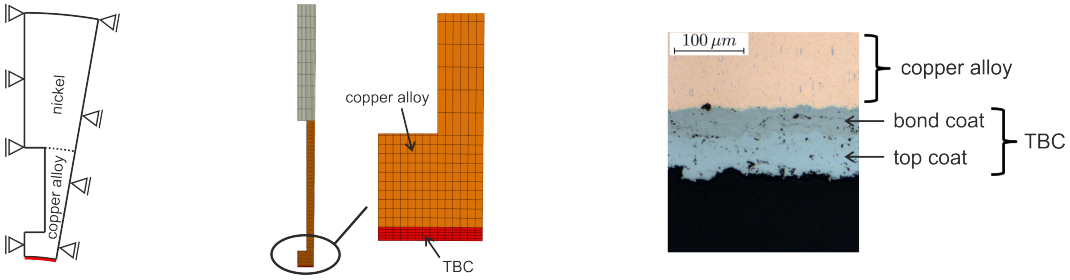



Figure 2: Left: segment of the cooling channel structure with TBC on the hot gas wall (marked in red), Center: FE model of the cooling channel segment with TBC, Right: microscopic picture of a TBC applied on a copper alloy [11].

## 2 MATERIAL MODELLING

The material model utilized for the lifetime investigations of the copper alloy has been developed and presented in previous works (see e.g. Tini et al. [12] and Kowollik et al. [13]). Therefore the material model is only briefly summarized and for more details it is referred to the latter mentioned publications. In order to take into account the relevant phenomena occuring during the thermomechanical loading of the copper alloy hot gas wall a viscoplastic material model considering isotropic damage is used. The corresponding rheological model is depicted in Figure 3. Schwarz et al. [14] compared isotropic damage models with anisotropic damage models (with and without considering micro defect closure effects) and concluded that neither damage anisotropy nor the crack closure effect strongly affect the global failure phenomenology. Due to this reason and to keep the number of parameters as low as possible the isotropic damage model is considered to be sufficient to perform first studies on the influence of the thermal barrier coatings on the lifetime of the copper alloy. An anisotropic damage model has been already developed by the authors (see e.g. Fassin et al. [15]) and was furthermore applied for lifetime simulations of rocket combustion chamber walls (see Fassin et al. [16] and Wulfinghoff et al. [17]). The intent of the authors for the future is also to use anisotropic damage models for simulations involving thermal barrier coatings.
The thermomechanical material parameters for the copper alloy are chosen according to Kowollik et al. [13] with the only difference that the coefficient of thermal expansion is adopted from Baeker et al. [18] to be now temperature dependent. The nickel material behaviour only plays a minor role and is therefore modelled as linear elastic (cf. Kowollik et al. [13]). The material behaviour of the TBC is modelled by an ideal elastoplastic material law as it was already done

in Baeker et al. [18]. The material parameters for the TBC were adopted from the bond coat material (NiCuCrAl).

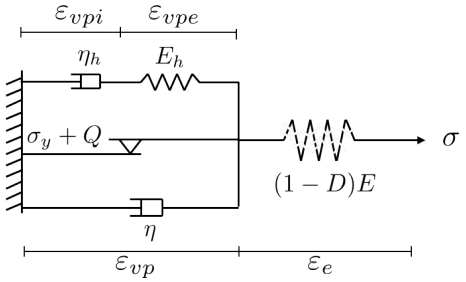



Figure 3: Rheological model of the viscoplastic material model incorporating isotropic damage.

Table 1: Summary of the constitutive equations of the viscoplastic damage model.

| | | | |
|---|---|---|---|
| total strain | $\boldsymbol{\varepsilon}$ | = | $\boldsymbol{\varepsilon}_e + \boldsymbol{\varepsilon}_{vp} + \boldsymbol{\varepsilon}_{th}$ |
| effective elasticity tensor | $\widetilde{\mathbf{C}}$ | = | $(1-D)\,\mathbf{C}$ |
| stress | $\boldsymbol{\sigma}$ | = | $\widetilde{\mathbf{C}}\,[\boldsymbol{\varepsilon}_e]$ |
| effective stress | $\widetilde{\boldsymbol{\sigma}}$ | = | $\mathbf{C}\,[\boldsymbol{\varepsilon}_e]$ |
| back stress deviator | $\boldsymbol{X}^D$ | = | $2\mu_h\left[\boldsymbol{\varepsilon}^D_{vpe}\right]$ |
| total strain | $\boldsymbol{\varepsilon}$ | = | $\boldsymbol{\varepsilon}_e + \boldsymbol{\varepsilon}_{vp} + \boldsymbol{\varepsilon}_{th}$ |
| viscoplastic strain | $\boldsymbol{\varepsilon}_{vp}$ | = | $\boldsymbol{\varepsilon}_{vpe} + \boldsymbol{\varepsilon}_{vpi}$ |
| yield function | $\Phi$ | = | $\left\|\widetilde{\boldsymbol{\sigma}}^D - \boldsymbol{X}^D\right\| - \sqrt{\frac{2}{3}}\,(\sigma_y + Q)$ |
| viscoplastic strain rate | $\dot{\boldsymbol{\varepsilon}}_{vp}$ | = | $\frac{\dot\lambda}{1-D}\,\frac{\widetilde{\boldsymbol{\sigma}}^D - \boldsymbol{X}^D}{\left\|\widetilde{\boldsymbol{\sigma}}^D - \boldsymbol{X}^D\right\|}$ |
| (local inelastic) viscoplastic strain rate | $\dot{\boldsymbol{\varepsilon}}_{vpi}$ | = | $\dot\lambda\, b\, \boldsymbol{\varepsilon}^D_{vpe}$ |
| damage | $\dot D$ | = | $\dot p\,\left(\frac{Y}{S}\right)^k\, H_{(p-p_D)}$ |
| plastic multiplier | $\dot\lambda$ | = | $\frac{\langle\overline{\Phi}\rangle^m}{\eta}$ |

## 3 THERMAL ANALYSIS

The basis of the transient thermal analysis is Fourier's heat equation

$$\Delta T = \frac{\rho c}{\zeta}\frac{\partial T}{\partial t}, \tag{1}$$

with temperature $T$ and time $t$. Here, the material parameters are $\rho$ (density), $c$ (heat capacity) and $\zeta$ (thermal conductivity). For the values of the material parameters of the copper alloy

and nickel, see Fassin et al. [16]. For the TBC the material parameters of NiCuCrAl from Baeker et al. [18] are used. The finite element model of the cooling channel segment with TBC used for the thermal as well as for the static analysis is shown in Figure 4. The thermal boundary conditions are described in the following: The interaction with the cooling fluid in the cooling channel and the hotgas on the hot gas wall is modelled via convective boundary conditions. For the cooling fluid the heat transfer coefficient $\alpha_{cf}$ and the bulk temperature $T_{cf}$ are prescribed, for the hot gas boundary condition $\alpha_{hg}$ and $T_{hg}$. One engine cycle consists of a pre cooling, a hot run, a post cooling and a relaxation phase, for which the corresponding values are listed in Table 3. A special case is present for the heat transfer coefficient $\alpha_{hg}$ on the hot gas wall surface. This value is dependent on the wall temperature $T_w$ of the hot gas wall and was adopted from Baeker et al. [18]. For the heat transfer coefficient of the cooling fluid values varying between $50\,\mathrm{kW/(m^2K)}$ and $60\,\mathrm{kW/(m^2K)}$ are applied. This variation corresponds to a varying mass flow rate of the cooling fluid. For the other free surfaces of the segment adiabatic boundary conditions are applied. In order to illustrate the influence

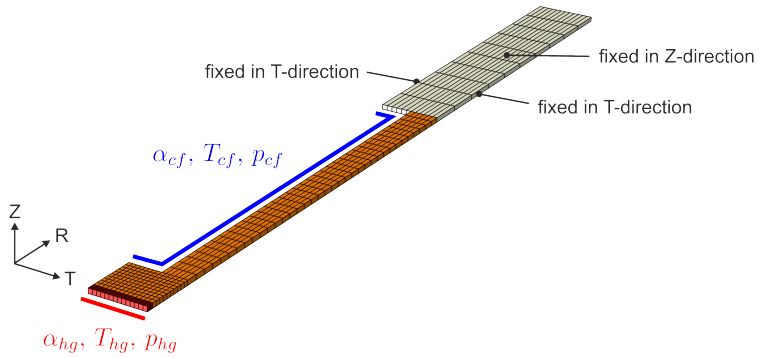



Figure 4: Finite element model of the cooling channel segment with TBC with thermal and static boundary conditions.

| **Phase** | **Time** | **Hot gas** | | **Cooling channel** | |
|---|---|---|---|---|---|
| [-] | $t$ [s] | $\alpha_{hg}\left[\frac{\mathrm{kW}}{\mathrm{m^2K}}\right]$ | $T_{hg}$[K] | $\alpha_{cf}\left[\frac{\mathrm{kW}}{\mathrm{m^2K}}\right]$ | $T_{cf}$[K] |
| Pre cooling | 0-2 | $f(T_w)$ | 293.15 | 50...60 | 50 |
| Hot run | 2 - 602 | $f(T_w)$ | 3435 | 50...60 | 50 |
| Post cooling | 602 - 604 | $f(T_w)$ | 293.15 | 50...60 | 50 |
| Relaxation | 604 - 1000 | $f(T_w)$ | 293.15 | 0.15 | 293.15 |

Table 2: Thermal boundary conditions for the hot gas and the cooling channel in one engine cycle.

of the thickness of the TBC on the temperature in the hot gas wall, several parameter studies have been performed. The conducted thermal analyses encompassed in each case only one engine cycle since the temperature distribution is approximately the same in every cycle. First of all, an appropriate TBC thickness and cooling fluid coefficient have to be specified to limit the maximum temperatures occuring in the TBC. Therefore transient thermal analyses were performed with different TBC thicknesses and cooling fluid heat transfer coefficients $\alpha_{cf}$. For the design of the TBC system the maximum occuring temperature on the surface of the TBC is an important value since it has to be under the maximum service temperature of the TBC, which is around $1500$ K for NiCuCrAl. Figure 5 shows the maximum temperature occuring on

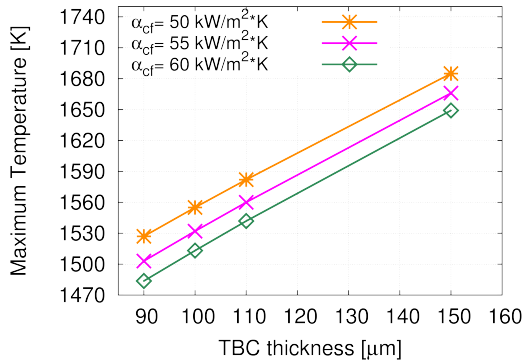



Figure 5: Maximum temperature occuring on the surface of the TBC in dependence of the thickness of the TBC for different heat transfer coefficients.

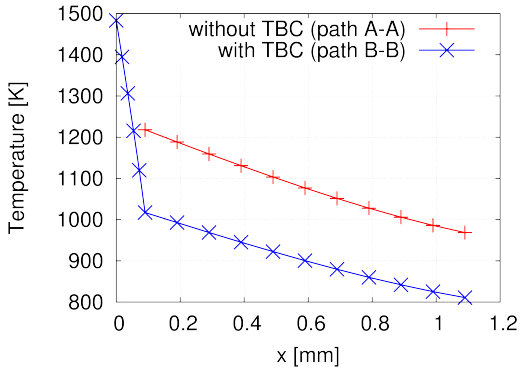



Figure 6: Comparison of the temperature distribution within the hot gas wall along the distance x for the versions "without TBC" and "with TBC".

the surface of the TBC in dependence of the thickness of the TBC for different heat transfer coefficients $\alpha_{cf}$. It can be seen that the thicker the TBC, the higher is the maximum occuring temperature. This can be explained by the insulation effect of the TBC, which has a much lower thermal conductivity than the copper alloy. Therefore the thermal resistance increases with increasing TBC thickness which leads to a lower heat flux through the hot gas wall. This finally results in a higher surface temperature. Figure 5 shows that only for a TBC thickness of $t$=90 $\mu$m and for a strong cooling with $\alpha_{cf}$=60 $\mathrm{kW/(m^2K)}$ the maximum occuring temperature lies under 1500 K. For this reason the further discussion of the results of the thermal analyses focus on this configuration. The static analyses in Section 4 will be also performed with this configuration.

In Figure 7 the temperature contour plots for the hot gas wall at the end of hot run ($t$=602 s) for the versions "without TBC" and "with TBC" are opposed to each other. Due to the application of the TBC, the temperature in the copper alloy could be drastically decreased. In contrast to that, the maximum occuring temperature on the TBC surface ($T$=1483 K) is much higher than the maximum occuring temperature on the hot gas wall surface for the version "without TBC" ($T$=1218 K). The strong temperature decrease within the TBC, which is approximately linear, is illustrated in Figure 6. Five elements over the thickness of the TBC were applied, which is shown by the six nodal temperatures plotted for path B-B for the first 0.09 mm (thickness of the TBC). The main purpose of the TBC, the reduction of the temperature within the copper alloy, is clearly depicted. The temperature reduction (difference between the red and the blue line) is round about 200 K. The maximum occuring temperature in the copper alloy is reduced from 1218 K to 1017 K. The surface temperature on the cooling channel wall is reduced from 969 K to 811 K. This temperature reduction is seen to be very promising to increase the lifetime of the copper alloy hot gas wall, which will be investigated in the next section.

## 4 STATIC ANALYSIS

The temperature field of the segment at every time point is used as input for the series of quasi static analyses with which the lifetime of the copper alloy will be investigated. The material parameters are temperature dependent as described in Section 2. Furthermore the thermal expansion of the material plays a major role in the static simulation, which is considered by the thermal strain. The static boundary conditions of the cooling channel segment are described in Figure 4. The left and the right face of the segment are fixed in tangential direction (T-direction)

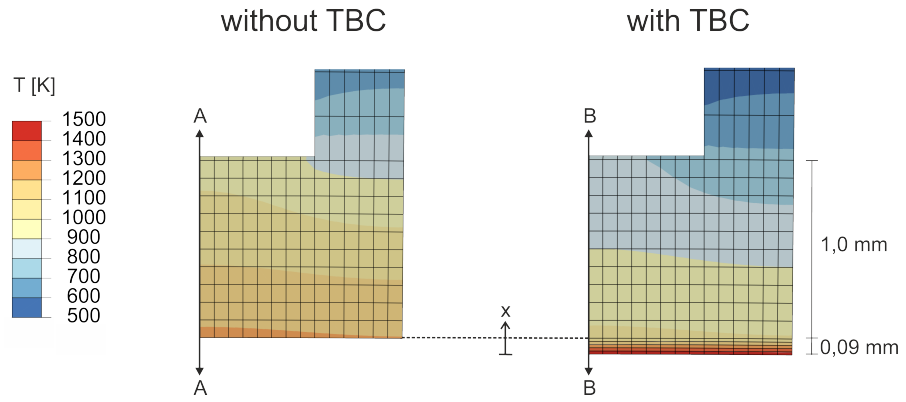



Figure 7: Temperature contour plots for the hot gas wall at the end of hot run ($t$=602 s) for the versions "without TBC" and "with TBC".

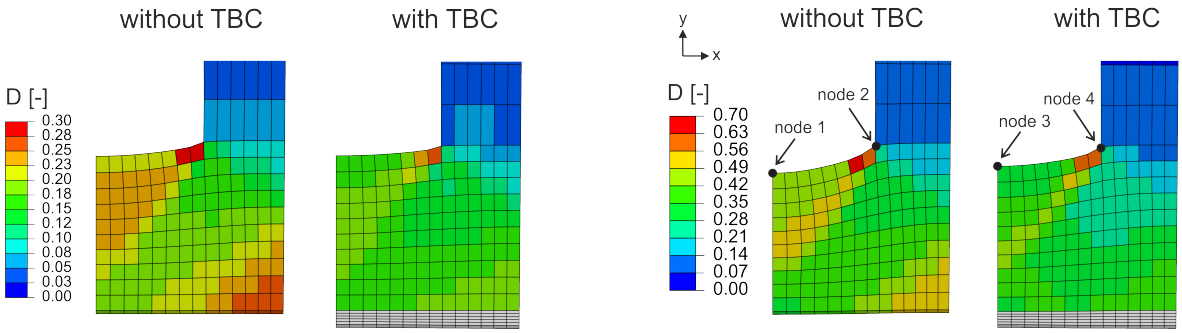
due to rotational symmetry. In thickness direction (Z-direction) a generalized plane strain condition is enforced (all elements in the R-T plane have the same strain in Z-direction) where the upper face is fixed in Z-direction and the bottom face is free. The boundary conditions of the cooling channel wall and the hot gas wall are given by means of pressures (see Table 4). For the finite element simulations 8-node bricks with reduced integration and hourglass stabilization are utilized. The time stepping is set manually. Figure 8 shows the damage distributions in the

| **Phase** [-] | **Time** $t$ [s] | **Hot gas pressure** $p_{hg}$ [MPa] | **Coolant pressure** $p_{cf}$ [MPa] |
|---|---|---|---|
| Pre cooling | 0 - 2 | 0 | 20 |
| Hot run | 2 - 602 | 10 | 20 |
| Post cooling | 602 - 604 | 0 | 20 |
| Relaxation | 604 - 1000 | 0 | 0 |

Table 3: Hot gas and coolant pressure over one engine cycle.

hot gas wall for the versions "without TBC" and "with TBC" after 10 and 20 cycles. After 10 cycles (cf. Figure 8(a)) the bulging of the hot gas wall is already clearly visible. For the version "without TBC" the highest values of the isotropic damage variable $D$ occur at the corner of the cooling channel and at the hot gas side in the middle of the fin. Furthermore damage has also accumulated in the center of the ligament. In contrast, the version "with TBC" shows mainly damage at the corner of the cooling channel and no further pronounced places of damage accumulation as for the version "without TBC". After 20 cycles (cf. Figure 8(b)) the deformation and the damage in the hot gas wall have further evolved. The differences between the damage distributions for the version "without TBC" and "with TBC" are the same as described already for the distributions after 10 cycles. In summary, it can be stated that the damage for the version "with TBC" is, as expected, smaller than for the version "without TBC". However, the differences in the maximum values of the isotropic damage varibale $D$ are smaller than expected. The influence of the TBC becomes more apparent in terms of deformation. The deformation of the hot gas wall (after 10 cycles as well as after 20 cycles) is higher for the version "without TBC". This is illustrated by the comparison of the displacement difference $\Delta u_{y,1-2}$ of node 1 and 2 with the displacement difference $\Delta u_{y,3-4}$ of node 3 and 4, which are measures for the

bulging of the hot gas wall. Figure 9 shows this deformation over the calculated 20 cycles. By using a TBC the deformation after 20 cycles is reduced by 30 %.




(a) after 10 cycles.

(b) after 20 cycles.

Figure 8: Comparison of the damage distributions in the hot gas wall for the versions “without TBC” and “with TBC” after 10 and 20 cycles.

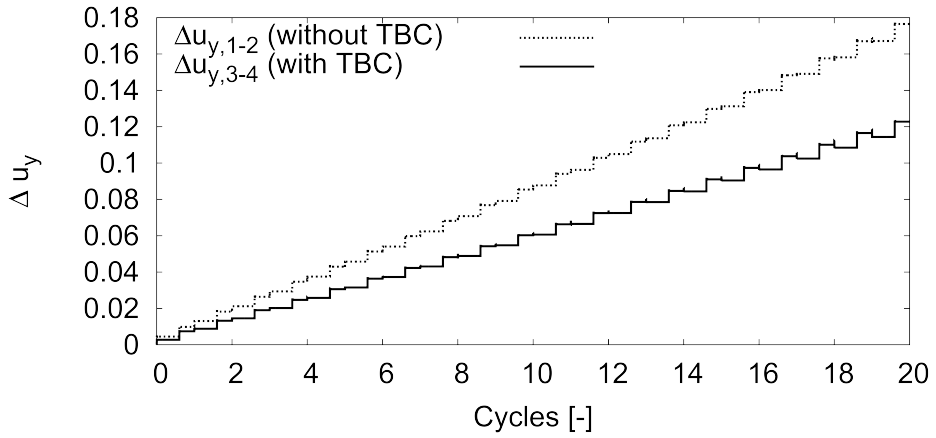



Figure 9: Comparison of the displacement differences $\Delta u_{y,1-2}$ and $\Delta u_{y,3-4}$.

## 5 CONCLUSIONS

In this work the influence of thermal barrier coatings on the lifetime and deformation behaviour of rocket engine nozzle structures made of copper alloys has been investigated. For the material modelling of the copper alloy a viscoplastic material model incorporating isotropic damage was applied. As input for the series of quasi static analyses including 20 cycles, the temperature field for one engine cycle was used repeatedly. In the first step an appropriate TBC thickness and the needed cooling fluid heat transfer coefficient were defined such that the maximum service temperature of the TBC was not exceeded. Afterwards it was shown that with the defined configuration ($t_{TBC}$=90 $\mu$m and $\alpha_{cf}$=60 kW/(m$^2$K)) a temperature reduction of approximately 200 K is obtained. The evaluation of the static simulations of 20 cycles showed that the damage distribution for a hot gas wall with TBC is slightly different from the one without TBC. The damage for the version “with TBC” was, as expected, smaller than for the version

"without TBC". However, the differences regarding the maximum values of the isotropic damage varibale $D$ were smaller as expected. But it was revealed that the deformations of the hot gas wall (bulging of the hot gas wall) induced by the thermomechanical loading were less pronounced when a TBC was applied.

**Acknowledgements**

Financial support provided by the German Research Foundation (DFG) within the projects TPD3 and TPD2 of the CRC/Transregio 40 'Fundamental Technologies for the Development of Future Space-Transport-System Components under High Thermal and Mechanical Loads' is gratefully acknowledged.

## REFERENCES

[1] Riccius JR, Haidn OJ, Zametaev EB. Influence of time dependent effects on the estimated life time of liquid rocket combustion chamber walls. *40th AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit*, 2004.

[2] Schwarz W, Schwub S, Quering K, Wiedmann D, Höppel HW, Göken M. Life prediction of thermally highly loaded components: Modelling of the damage process of a rocket combustion chamber hot wall. *CEAS Space Journal* 2011; **1**:83–97, doi:10.1007/s12567-011-0007-9.

[3] Schloesser J, Fedorova T, Bäker M, Rösler J. Thermal barrier coating development for application in rocket engines. *SFB TRR40 Annual Report 2009*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2009.

[4] Schloesser J, Kowollik D, Bäker M, Rösler J, Horst P. Failure analysis and multiscale modeling of thermal barrier coatings. *SFB TRR40 Annual Report 2010*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2010.

[5] Schloesser J, Kowollik D, Bäker M, Rösler J, Horst P. Thermal barrier coatings in rocket engines - a multiscale simulation and development approach. *SFB TRR40 Annual Report 2011*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2011.

[6] Schloesser J, Kowollik D, Bäker M, Rösler J, Horst P. Sensitivity analyses and coating development of thermal barrier coatings in rocket engines. *SFB TRR40 Annual Report 2012*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2012.

[7] Fiedler T, Fedorova T, Rösler J, Bäker M. Design of a nickel-based bond coat alloy for thermal barrier coatings on copper substrates. *SFB TRR40 Annual Report 2013*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2013.

[8] Fiedler T, Rösler J, Bäker M. Development of a cunicral bond-coat for thermal barrier coatings in rocket combustion-chambers. *SFB TRR40 Annual Report 2014*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2014.

[9] Fiedler T, Rösler J, Bäker M. Design studies on thermal-barrier-coatings for liquid-rocket combustion-chambers. *SFB TRR40 Annual Report 2014*. Lehrstuhl fr Aerodynamik und Strmungsmechanik, Technische Universitt Mnchen, 2015.

[10] Butler Jr D, Pindera MJ. Analysis of factors affecting the performance of rlv thrust cell liners. *Technical Report*, NASA Glenn Research Center 2004.

[11] T. fiedler.

[12] Tini V, Kowollik D, Reese S, Haupt M. Application of a viscoplastic damage model for the failure prediction of regeneratively cooled nozzle structures. *ECCOMAS Coupled Problems 2011 Proceedings*, 2011.

[13] Kowollik D, Tini V, Reese S, Haupt M. 3d fluid-structure interaction analysis of a typical liquid rocket engine cycle based on a novel viscoplastic damage model. *International Journal for Numerical Methods in Engineering* 2013; **94chac**:1165–1190.

[14] Schwarz W, Wiedmann D, Schwub S, Höppel H, Göken M. Assessment of different continuum damage models for life-time prediction of high thrust cryogenic combustion chambers. *Proceedings of the 4th European Conference for Aerospace Sciences (EUCASS)*, 2011.

[15] Fassin M, Wulfinghoff S, Reese S. Different numerical time integration schemes for elastoplasticity coupled to anisotropic damage. *Applied Mechanics and Materials*, vol. 784, Trans Tech Publ, 2015; 217–224.

[16] Fassin M, Wulfinghoff S, Reese S. Application of an elastoplastic material model coupled with anisotropic damage for the failure prediction of a rocket combustion chamber wall. *Proceedings of the 6th European Conference for Aerospace Sciences (EUCASS)*, 2015.

[17] Wulfinghoff S, Fassin M, Reese S. Comparison of two time-integration algorithms for an anisotropic damage model coupled with plasticity. *Applied Mechanics and Materials*, vol. 784, Trans Tech Publ, 2015; 292–299.

[18] Bäker M, Fiedler T, Rösler J. Stress evolution in thermal barrier coatings for rocket engine applications. *Mechanics of Advanced Materials and Modern Processes* 2015; **1**(1):1–10.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ANALYTICAL SOLUTION FOR DYNAMIC FRACTURE OF TWO COPLANAR LIMITED-PERMEABLE CRACKS IN MAGNETO-ELECTRO-ELASTIC MATERIAL

**Pei-Wei Zhang**[1]

[1]Jiangsu Key Laboratory of Engineering Mechanics (Southeast University), School of Civil Engineering, Southeast University, No.2, Sipailou Street, Nanjing, 210096, P.R. China.
e-mail: zhangpeiwei@seu.edu.cn

**Keywords:** analytical solution; magneto-electro-elastic material; coplanar cracks; dynamic fracture; generalized intensity factor.

**Abstract.** *Two coplanar limited-permeable rectangular cracks in magneto-electro-elastic material are modeled and solved using generalized Almansi's theorem. The mixed boundary value problem is formulated into three pairs of dual integral equations with the help of Fourier transform, in which the unknown functions are the jumps of displacements across the crack-surface. By directly expanding the unknown functions into infinite series form of Jacobi polynomials, the dual integral equations are solved, and the analytical expressions of generalized intensity factors are derived strictly. Based on the generalized intensity factors, the dynamic behaviors of two cracks are estimated under P-wave loads. To show the trends of effects of loading frequency and geometry of cracks on the generalized intensity factors, top finite terms of infinite series are numerically calculated based on the Schmidt method. Numerical results are then drawn graphically. It reveals that the trend of two cracks forming a larger crack by propagating depends strongly on crack geometry and load frequency. This work illuminate the condition inducing different crack propagation patterns, which will benifit the forecast of damage forms of transversely isotropic magneto-electro-elastic material.*

## 1 INTRODUCTIO

Magneto-electro-elastic composites, with its root in the early work of Vansuchtelen[1], have been widely applied in science and engineering. Layered magneto-electro-elastic composites, as a basic form of various electromagnetic structures, usually manufactured using tape-casting technique and bonding method, achieve a historic high electromagnetic factor. The dynamic mechanical, magnetic and electric loads are inevitable in its service, which make the layered magnetic-electric-elastic components pregnant coplanar cracks or parallel cracks from the interface or other weak position. Furthermore, the initiation of cracks will result in the reduction of electromagnetic conversion efficiency, if not fail prematurely. Therefore, researching on the dynamic fracture behaviors of magneto-electro-elastic composites is very worthwhile[2, 3]. To obtain the reliable service life time and the reliable action of related devices and to prevent failure during service, most of works concerned about the dynamic fracture problems with assumption of transversely isotropic property. Among them, Zhong et al[4, 5] investigated the anti-plane and in-plane dynamic fracture problem, and the influence of various factors on the crack propagation orientation were discussed. Feng et al[6, 7] also made lots of effort on the dynamic fracture of magneto-electro-elastic medium. They discussed four kinds of ideal crack-face assumption, and based on energy release rate, the effect of crack-face assumption on the crack extension was unfolded. As discussed in previous work[7], the boundary condition on the crack surface was complex in piezoelectric/piezomagnetic composites. The limited-permeable boundary condition first proposed in Hao's work[8] for piezoelectric materials was developed into magneto-electro-elastic composites by Zhong et al[5] and Zhou et al[9], respectively.

Although there are a lot of works focusing on the anti-plane and in-plane problems, it is insufficient to analyzing the fracture behavior in 3-Dimensions. And, it is important to take multi-cracked 3D medium into account. Layered composites were prone to generating multiple cracks in the same layer due to periodical fatigue loads or manufacturing process, so investigating the interaction between coplanar cracks is necessary. During analyzing 3D crack, model of penny, elliptical and rectangular were the most commonly used types, in which rectangular crack model is over safety in the engineering analysis.

With the lack of 3D dynamic analysis of multi-cracked magneto-electro-elastic materials in mind, in this paper, the model of transversely isotropic magneto-electro-elastic materials are established with two coplanar rectangular cracks parallel to the isotropic plane. Here the steady-state effects of the loading frequency on the limited-permeable crack initiating behavior were investigated on. The double-cracked magneto-electro-elastic material problems were formulated to three pairs of dual integral equations through Fourier transform. By solving the dual integral equations with the jumps of displacements as unknown variables, the analytical generalized intensity factors were derived strictly in the forms of infinite series which comes from Jacobi polynomial expansion of the jumps of displacements. In order to reveal some implicit links between mixed boundary conditions and generalized intensity factors, numerical results of generalized intensity factors were depicted and some useful conclusions and laws were explained, including the possibility of two crack converging and the effects of the geometry of cracks and the load frequency on the crack extension. This contribution is available for further analyzing multi-cracked 3D magneto-electro-elastic materials.

## 2 MODELING AND BASIC EQUATIONS

The simplified model of transversely isotropic homogeneous magneto-electro-elastic material is adopted in the present problem. Consider two coplanar rectangular cracks inside the isotropic plane as shown in Fig. 1. These two cracks are situated at the region of $l_1 \le |x| \le l_3$ and $|y| \le l_2$ in the plane of $z = 0$ in Cartesian coordinate system ( $x, y, z$ ). A uniform distributed normal harmonic load $\sigma_{zz}(x, y, 0) = -\sigma_0 e^{i\omega\tau}$ (here $\sigma_0$ and $\omega$ are the amplitude and the circular frequency of the incident load wave, respectively.) is directly applied on the upper and the lower crack-surfaces, to which the solution combining with a uniform field can be transformed into the solution to the double-cracked material with a remotely harmonic load according to the standard superposition technique in the fracture mechanics point of view.

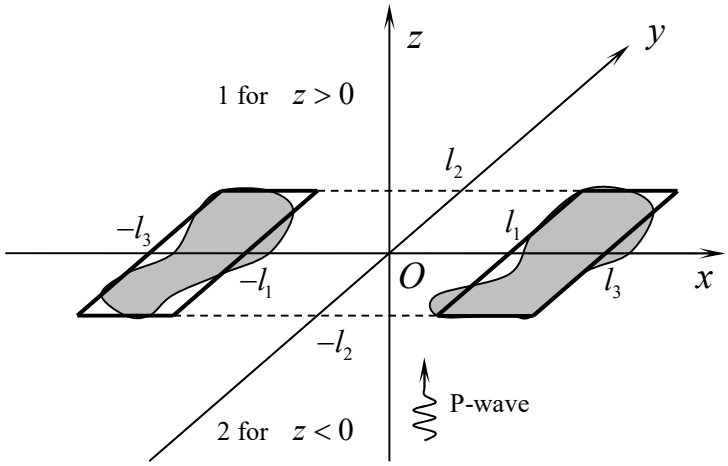



Fig.1: Geometry and coordinate system for two rectangular cracks

When the time-harmonic P-wave is applied on the magneto-electro-elastic materials and the response goes from the transient response stage into the steady state, all fields quantities can be expressed in the form as follows:

$$X_0(x, y, z, \tau) = X(x, y, z)e^{-i\omega\tau}$$

where $X$ is on behalf of the displacement $u_k$ , the electric potential $\phi$ , the magnetic potential $\psi$ , the stress $\sigma_{kl}$ , the electric displacements $D_k$ and the magnetic flux $B_k$ ( $k, l = x, y, z$ ); $\tau$ is the time variable, respectively.

In Cartesian coordinates $(x, y, z)$ of the present problem, the basic equations of linear, homogeneous, transversely isotropic magneto-electro-elastic material ignoring body force, free electric and magnetic charge are as follows:

$$\begin{cases} \sigma^{(j)}_{kl,k} = \rho_0 \ddot{u}^{(j)}_l \\ D^{(j)}_{k,k} = 0 \\ B^{(j)}_{k,k} = 0 \end{cases}, \qquad k, l = x, y, z \tag{1}$$

$$\begin{Bmatrix}\sigma_{xx}^{(j)}\\ \sigma_{yy}^{(j)}\\ \sigma_{zz}^{(j)}\\ \sigma_{xy}^{(j)}\\ \sigma_{xz}^{(j)}\\ \sigma_{yz}^{(j)}\end{Bmatrix}=\begin{bmatrix} c_{11}\partial_{,x} & c_{12}\partial_{,y} & c_{13}\partial_{,z} & e_{31}\partial_{,z} & q_{31}\partial_{,z}\\ c_{12}\partial_{,x} & c_{11}\partial_{,y} & c_{13}\partial_{,z} & e_{31}\partial_{,z} & q_{31}\partial_{,z}\\ c_{13}\partial_{,x} & c_{13}\partial_{,y} & c_{33}\partial_{,z} & e_{33}\partial_{,z} & q_{33}\partial_{,z}\\ (c_{11}-c_{12})\partial_{,y}/2 & (c_{11}-c_{12})\partial_{,x}/2 & 0 & 0 & 0\\ c_{44}\partial_{,z} & 0 & c_{44}\partial_{,x} & e_{15}\partial_{,x} & q_{15}\partial_{,x}\\ 0 & c_{44}\partial_{,z} & c_{44}\partial_{,y} & e_{15}\partial_{,y} & q_{15}\partial_{,y}\end{bmatrix}\begin{Bmatrix}u_x^{(j)}\\ u_y^{(j)}\\ u_z^{(j)}\\ \phi^{(j)}\\ \psi^{(j)}\end{Bmatrix} \tag{2}$$

$$\begin{Bmatrix}D_x^{(j)}\\ D_y^{(j)}\\ D_z^{(j)}\end{Bmatrix}=\begin{bmatrix} e_{15}\partial_{,z} & 0 & e_{15}\partial_{,x} & -\varepsilon_{11}\partial_{,x} & -d_{11}\partial_{,x}\\ 0 & e_{15}\partial_{,z} & e_{15}\partial_{,y} & -\varepsilon_{11}\partial_{,y} & -d_{11}\partial_{,y}\\ e_{31}\partial_{,x} & e_{31}\partial_{,y} & e_{33}\partial_{,z} & -\varepsilon_{33}\partial_{,z} & -d_{33}\partial_{,z}\end{bmatrix}\begin{Bmatrix}u_x^{(j)} & u_y^{(j)} & u_z^{(j)} & \phi^{(j)} & \psi^{(j)}\end{Bmatrix}^T \tag{3}$$

$$\begin{Bmatrix}B_x^{(j)}\\ B_y^{(j)}\\ B_z^{(j)}\end{Bmatrix}=\begin{bmatrix} q_{15}\partial_{,z} & 0 & q_{15}\partial_{,x} & -d_{11}\partial_{,x} & -\mu_{11}\partial_{,x}\\ 0 & q_{15}\partial_{,z} & q_{15}\partial_{,y} & -d_{11}\partial_{,y} & -\mu_{11}\partial_{,y}\\ q_{31}\partial_{,x} & q_{31}\partial_{,y} & q_{33}\partial_{,z} & -d_{33}\partial_{,z} & -\mu_{33}\partial_{,z}\end{bmatrix}\begin{Bmatrix}u_x^{(j)} & u_y^{(j)} & u_z^{(j)} & \phi^{(j)} & \psi^{(j)}\end{Bmatrix}^T \tag{4}$$

where $\rho_0$, $c_{jk}$, $e_{jk}$, $\varepsilon_{jk}$, $q_{jk}$, $d_{jk}$ and $\mu_{jk}$ ($j,k=1,2,3$) are the mass density, the elastic stiffness constants, the piezoelectric constants, the electric permittivities, the piezomagnetic constants, the electromagnetic constants and the magnetic permeabilities of magneto-electro-elastic medium, respectively. $\ddot{u}_l^{(j)}$ denotes the second-order differential of $u_l^{(j)}$ with respect to time variable and a subscript comma denotes the partial differential to the coordinates ($x,y,z$), where the superscript $j$ ($j$=1, 2) denotes the fields in the upper half space 1 ($z>0$) and the lower half space 2 ($z<0$) as shown in Fig.1. Considering the load wave propagating along $z$-axis in form of plane wave, $\ddot{u}_x^{(j)}=\ddot{u}_y^{(j)}=0$ is concluded.

Substituting Eqs.(2)-(4) into Eq.(1), the governing equations represented by the displacements, electric potentials and magnetic potentials are constructed:

$$\begin{cases}
c_{11}u_{x,xx}^{(j)}+(c_{11}-c_{12})u_{x,yy}^{(j)}/2+c_{44}u_{x,zz}^{(j)}+(c_{11}+c_{12})u_{y,xy}^{(j)}/2\\
\qquad +(c_{13}+c_{44})u_{z,xz}^{(j)}+(e_{15}+e_{31})\phi_{,xz}^{(j)}+(q_{15}+q_{31})\psi_{,xz}^{(j)}=0\\
(c_{11}-c_{12})u_{y,xx}^{(j)}/2+c_{11}u_{y,yy}^{(j)}+c_{44}u_{y,zz}^{(j)}+(c_{11}+c_{12})u_{x,xy}^{(j)}/2\\
\qquad +(c_{13}+c_{44})u_{z,yz}^{(j)}+(e_{15}+e_{31})\phi_{,yz}^{(j)}+(q_{15}+q_{31})\psi_{,yz}^{(j)}=0\\
c_{44}(u_{z,xx}^{(j)}+u_{z,yy}^{(j)})+c_{33}u_{z,zz}^{(j)}+(c_{13}+c_{44})(u_{x,xz}^{(j)}+u_{y,yz}^{(j)})\\
\qquad +e_{15}(\phi_{,xx}^{(j)}+\phi_{,yy}^{(j)})+e_{33}\phi_{,zz}^{(j)}+q_{15}(\psi_{,xx}^{(j)}+\psi_{,yy}^{(j)})+q_{33}\psi_{,zz}^{(j)}=-\omega^2\rho_0 u_z^{(j)}\\
e_{15}(u_{z,xx}^{(j)}+u_{z,yy}^{(j)})+e_{33}u_{z,zz}^{(j)}+(e_{15}+e_{31})(u_{x,xz}^{(j)}+u_{y,yz}^{(j)})\\
\qquad -\varepsilon_{11}(\phi_{,xx}^{(j)}+\phi_{,yy}^{(j)})-\varepsilon_{33}\phi_{,zz}^{(j)}-d_{11}(\psi_{,xx}^{(j)}+\psi_{,yy}^{(j)})-d_{33}\psi_{,zz}^{(j)}=0\\
q_{15}(u_{z,xx}^{(j)}+u_{z,yy}^{(j)})+q_{33}u_{z,zz}^{(j)}+(q_{15}+q_{31})(u_{x,xz}^{(j)}+u_{y,yz}^{(j)})\\
\qquad -d_{11}(\phi_{,xx}^{(j)}+\phi_{,yy}^{(j)})-d_{33}\phi_{,zz}^{(j)}-\mu_{11}(\psi_{,xx}^{(j)}+\psi_{,yy}^{(j)})-\mu_{33}\psi_{,zz}^{(j)}=0
\end{cases} \tag{5}$$

In magneto-electro-elastic material, the velocity of elastic P-wave in $z$ -axis direction $c_z = \sqrt{(c_{33} + e_{33}\frac{e_{33}\mu_{33} - q_{33}d_{33}}{\varepsilon_{33}\mu_{33} - d_{33}^2} + q_{33}\frac{q_{33}\varepsilon_{33} - e_{33}d_{33}}{\mu_{33}\varepsilon_{33} - d_{33}^2})/\rho_0}$ can be derived which only depends on the material constants.

## 3 LIMITED-PERMEABLE BOUNDARY CONDITIONS

The limited-permeable crack model, developed in Zhong's [10] and Zhou's [9, 11] works, is adopted here. The boundary conditions on crack-surfaces and at infinity for 3D problem are expressed in detail as follows:

$$\begin{cases} \sigma_{zz}^{(1)}(x,y,0^+) = \sigma_{zz}^{(2)}(x,y,0^-) = -\sigma_0 \\ \sigma_{xz}^{(1)}(x,y,0^+) = \sigma_{xz}^{(2)}(x,y,0^-) = 0, \quad \sigma_{yz}^{(1)}(x,y,0^+) = \sigma_{yz}^{(2)}(x,y,0^-) = 0 \\ D_z^{(1)}(x,y,0^+) = D_z^{(2)}(x,y,0^-), \quad B_z^{(1)}(x,y,0^+) = B_z^{(2)}(x,y,0^-) \qquad \text{for } l_1 \le |x| \le l_3, \\ D_{z0}[u_z^{(1)}(x,y,0^+) - u_z^{(2)}(x,y,0^-)] = \varepsilon_0[\phi^{(1)}(x,y,0^+) - \phi^{(2)}(x,y,0^-)] \\ B_{z0}[u_z^{(1)}(x,y,0^+) - u_z^{(2)}(x,y,0^-)] = \mu_0[\psi^{(1)}(x,y,0^+) - \psi^{(2)}(x,y,0^-)] \end{cases}$$

$$|y| \le l_2 \quad (6)$$

$$\begin{cases} \sigma_{zz}^{(1)}(x,y,0^+) = \sigma_{zz}^{(2)}(x,y,0^-), \quad \sigma_{xz}^{(1)}(x,y,0^+) = \sigma_{xz}^{(2)}(x,y,0^-) \\ \sigma_{yz}^{(1)}(x,y,0^+) = \sigma_{yz}^{(2)}(x,y,0^-), \quad u_x^{(1)}(x,y,0^+) = u_x^{(2)}(x,y,0^-) \\ u_y^{(1)}(x,y,0^+) = u_y^{(2)}(x,y,0^-), \quad u_z^{(1)}(x,y,0^+) = u_z^{(2)}(x,y,0^-) \qquad \text{for } |x| < l_1 \text{ or } |x| > l_3, \\ \phi^{(1)}(x,y,0^+) = \phi^{(2)}(x,y,0^-), \quad D_z^{(1)}(x,y,0^+) = D_z^{(2)}(x,y,0^-) \\ \psi^{(1)}(x,y,0^+) = \psi^{(2)}(x,y,0^-), \quad B_z^{(1)}(x,y,0^+) = B_z^{(2)}(x,y,0^-) \end{cases}$$

$$|y| > l_2 \quad (7)$$

$$u_x^{(j)}(x,y,z) = u_y^{(j)}(x,y,z) = u_z^{(j)}(x,y,z) = \phi^{(j)}(x,y,z) = \psi^{(j)}(x,y,z) = 0 \quad \text{for} \sqrt{x^2+y^2+z^2} \to \infty \quad (8)$$

where $D_{z0}$ and $B_{z0}$ are defined as the electric displacement and the magnetic flux inside cracks, and they are equal to $D_z^{(1)}(x,y,0^+)$ and $B_z^{(1)}(x,y,0^+)$ in magnitude, respectively. As discussed in public literatures[10, 12], the electric displacement and magnetic induction inside the opening crack are almost constants and dependent on applied magneto-electro-elastic loads, material properties, dielectric permittivity and magnetic permeability of crack interior, which would be certified briefly later. The limited-permeable boundary conditions in Eq.(6) will transform to permeable electric-magnetic boundary condition when $u_z^{(1)}(x,y,0^+) - u_z^{(2)}(x,y,0^-) = 0$ and to impermeable one when $\varepsilon_0 = \mu_0 = 0$.

## 4 SOLUTION PROCEDURE

The governing equations (5) need to be simplified by introducing displacement potential functions as discussed in reference [13]. Let

$$u_x^{(j)} = \Psi_{,y}^{(j)} - G_{,x}^{(j)}, \quad u_y^{(j)} = -\Psi_{,x}^{(j)} - G_{,y}^{(j)} \tag{9}$$

where $\Psi^{(j)}$ and $G^{(j)}$ are two potential functions. By strict derivation, the general expressions of the displacement potential function, the displacements, stresses, electric potentials, electric displacements, magnetic potentials and magnetic flux fields satisfying Eq. (8) are derived as follows:

$$\begin{cases} G^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(1)}(s,t)A_k(s,t)\cos(sx)\cos(ty)e^{-\lambda_k\sqrt{s^2+t^2}z}dsdt \\ u_x^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty [tA_0(s,t)e^{-\lambda_0\sqrt{s^2+t^2}z} + s\sum_{k=1}^{4} \chi_k^{(1)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}]\sin(sx)\cos(ty)dsdt \\ u_y^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty [-sA_0(s,t)e^{-\lambda_0\sqrt{s^2+t^2}z} + t\sum_{k=1}^{4} \chi_k^{(1)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}]\cos(sx)\sin(ty)dsdt \\ u_z^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(2)}(s,t)A_k(s,t)\cos(sx)\cos(ty)e^{-\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \phi^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(3)}(s,t)A_k(s,t)\cos(sx)\cos(ty)e^{-\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \psi^{(1)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(4)}(s,t)A_k(s,t)\cos(sx)\cos(ty)e^{-\lambda_k\sqrt{s^2+t^2}z}dsdt \end{cases} \tag{10}$$

$$\begin{cases} G^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(1^*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ u_x^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty [tB_0(s,t)e^{\lambda_0\sqrt{s^2+t^2}z} + s\sum_{k=1}^{4} \chi_k^{(1^*)}(s,t)B_k(s,t)e^{\lambda_k\sqrt{s^2+t^2}z}]\sin(sx)\cos(ty)dsdt \\ u_y^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty [-sB_0(s,t)e^{\lambda_0\sqrt{s^2+t^2}z} + t\sum_{k=1}^{4} \chi_k^{(1^*)}(s,t)B_k(s,t)e^{\lambda_k\sqrt{s^2+t^2}z}]\cos(sx)\sin(ty)dsdt \\ u_z^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(2^*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \phi^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(3^*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \psi^{(2)}(x,y,z) = \dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty \sum_{k=1}^{4} \chi_k^{(4^*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \end{cases} \tag{11}$$

where $\chi_k^{(j)}(s,t)$ and $\chi_k^{(j^*)}(s,t)$ ($k = 1,2,3,4.\quad j = 1,2,3,4,5.$) are functions of known expressions (The expressions are not shown here for brevity.).

$$\begin{cases} \sigma_{zz}^{(1)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(1)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}\cos(sx)\cos(ty)dsdt \\ \sigma_{yz}^{(1)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{5}\beta_k^{(2)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}\cos(sx)\sin(ty)dsdt \\ \sigma_{xz}^{(1)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{5}\beta_k^{(3)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}\sin(sx)\cos(ty)dsdt \\ D_z^{(1)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(4)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}\cos(sx)\cos(ty)dsdt \\ B_z^{(1)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(5)}(s,t)A_k(s,t)e^{-\lambda_k\sqrt{s^2+t^2}z}\cos(sx)\cos(ty)dsdt \end{cases} \tag{12}$$

$$\begin{cases} \sigma_{zz}^{(2)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(1*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \sigma_{yz}^{(2)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{5}\beta_k^{(2*)}(s,t)B_k(s,t)\cos(sx)\sin(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ \sigma_{xz}^{(2)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{5}\beta_k^{(3*)}(s,t)B_k(s,t)\sin(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ D_z^{(2)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(4*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \\ B_z^{(2)}(x,y,z)=\dfrac{4}{\pi^2}\int_0^\infty\int_0^\infty\sum_{k=1}^{4}\beta_k^{(5*)}(s,t)B_k(s,t)\cos(sx)\cos(ty)e^{\lambda_k\sqrt{s^2+t^2}z}dsdt \end{cases} \tag{13}$$

where $\beta_k^{(j)}(s,t)$ and $\beta_k^{(j*)}(s,t)$ ($k=0,1,2,3,4.\quad j=1,2,3,4,5.$) are functions of known expressions (The expressions are not shown here for brevity.).

Substitution of expressions (10)-(11) into limited-permeable boundary condition (6), electric displacement and magnetic flux inside crack can be generated

$$\begin{cases} D_{z0}=\varepsilon_0\dfrac{\sum_{k=1}^{4}\chi_k^{(3)}(s,t)[A_k(s,t)+B_k(s,t)]}{\sum_{k=1}^{4}\chi_k^{(2)}(s,t)[A_k(s,t)+B_k(s,t)]} \\ B_{z0}=\mu_0\dfrac{\sum_{k=1}^{4}\chi_k^{(4)}(s,t)[A_k(s,t)+B_k(s,t)]}{\sum_{k=1}^{4}\chi_k^{(2)}(s,t)[A_k(s,t)+B_k(s,t)]} \end{cases} \tag{14}$$

It is obviously that the distribution of electric displacement and magnetic flux inside crack is independent on Cartesian coordinates, means, both of the physical field inside cracks are almost constants.

To solve the fracture problem, the jumps of displacement on upper and lower crack surface are defined as follows:

$$\begin{cases} f_1(x,y) = u_x^{(1)}(x,y,0^+) - u_x^{(2)}(x,y,0^-) \\ f_2(x,y) = u_y^{(1)}(x,y,0^+) - u_y^{(2)}(x,y,0^-) \\ f_3(x,y) = u_z^{(1)}(x,y,0^+) - u_z^{(2)}(x,y,0^-) \end{cases} \tag{15}$$

It is obviously that both $f_1(x,y)$ and $f_3(x,y)$ are even functions about the variable $y$ and $f_2(x,y)$ is an odd function about the variable $y$ according to the symmetry of this problem.

Using Fourier transform, and Substituting Eqs.(10)-(13) into the crack boundary conditions Eqs.(6)-(7) and the jumps of displacements Eq.(15), the following dual integral equations can be deduced:

$$\sigma_{zz}^{(1)}(x,y,0) = \frac{2}{\pi^2}\int_0^\infty\int_0^\infty g_1(s,t)\overline{f}_3(s,t)\cos(sx)\cos(ty)dsdt = -\sigma_0, \qquad l_1 \le x \le l_3 \text{ and } 0 \le y \le l_2 \tag{16}$$

$$\sigma_{xz}^{(1)}(x,y,0) = \frac{2}{\pi^2}\int_0^\infty\int_0^\infty [g_2(s,t)\overline{f}_1(s,t) + g_3(s,t)\overline{f}_2(s,t)]\sin(sx)\cos(ty)dsdt = 0, \quad l_1 \le x \le l_3 \text{ and } 0 \le y \le l_2 \tag{17}$$

$$\sigma_{yz}^{(1)}(x,y,0) = \frac{2}{\pi^2}\int_0^\infty\int_0^\infty [g_4(s,t)\overline{f}_1(s,t) + g_5(s,t)\overline{f}_2(s,t)]\cos(sx)\sin(ty)dsdt = 0, \quad l_1 \le x \le l_3 \text{ and } 0 \le y \le l_2 \tag{18}$$

$$\begin{cases} \int_0^\infty\int_0^\infty \overline{f}_1(s,t)\sin(sx)\cos(ty)dsdt = 0 \\ \int_0^\infty\int_0^\infty \overline{f}_2(s,t)\cos(sx)\sin(ty)dsdt = 0, \quad \text{for } |x| < l_1 \text{ or } |x| > l_3 \text{ or } y > l_2 \\ \int_0^\infty\int_0^\infty \overline{f}_3(s,t)\cos(sx)\cos(ty)dsdt = 0 \end{cases} \tag{19}$$

where $\overline{f}_j(s,t)$ $(j=1,2,3)$ indicates two dimensional Fourier transform of $f_j(x,y)$. $g_k(s,t)$ $(k=1,2,3,4,5.)$ are known functions (The forms of functions are not shown here for brevity.). The foregoing three pairs of dual integral equations (16)-(19) must be solved to determine the unknown functions $f_1(x,y)$, $f_2(x,y)$ and $f_3(x,y)$.

## 5 CRACK TIP SINGULARITY FIELD AND GENERALIZED INTENSITY FACTOR

The Schmidt method [14-16] is introduced to solve the dual integral equations (16)-(19). The whole generalized fields can be solved, such as stress, electric displacement, magnetic flux and so on. But, to determine the generalized field intensity factors, the main attention must be focused on the generalized fields ahead of crack edges. In this work, $\sigma_{zz}^{(1)}$, $\sigma_{xz}^{(1)}$, $\sigma_{yz}^{(1)}$, $D_z^{(1)}$ and $B_z^{(1)}$ ahead of crack edges are obtained as follows:

$$\sigma_{zz}^{(1)}(x,y,0) = \frac{2}{\pi^2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} c_{mn}\int_0^\infty \frac{1}{t}J_{2n+1}(tl_2)\cos(ty)dt\left[\int_0^\infty \frac{G_{mn}^{(3)}(s)}{s} g_1(s,t)J_{m+1}\left(s\frac{1-l_1}{2}\right)\cos(sx)ds\right] \tag{20}$$

$$\sigma_{xz}^{(1)}(x,y,0)=\sigma_{yz}^{(1)}(x,y,0)=0 \tag{21}$$

$$D_z^{(1)}(x,y,0)=\frac{2}{\pi^2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}\int_0^{\infty}\frac{1}{t}J_{2n+1}(tl_2)\cos(ty)dt[\int_0^{\infty}\frac{G_{mn}^{(3)}(s)}{s}g_7(s,t)J_{m+1}(s\frac{1-l_1}{2})\cos(sx)ds] \tag{22}$$

$$B_z^{(1)}(x,y,0)=\frac{2}{\pi^2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}\int_0^{\infty}\frac{1}{t}J_{2n+1}(tl_2)\cos(ty)dt[\int_0^{\infty}\frac{G_{mn}^{(3)}(s)}{s}g_8(s,t)J_{m+1}(s\frac{1-l_1}{2})\cos(sx)ds] \tag{23}$$

where $g_k(s,t)$ ($k=7,8$ .) is known functions (The forms of functions are not shown here for brevity.).

The generalized intensity factors $K$ on each crack edge are respectively obtained as follows:

(I) on inner crack edge, $x=l_1$ and $0\le y<l_2$:

$$K_x^{\mathrm{I}}=\lim_{x\to l_1}\sqrt{2(l_1-x)}\cdot\sigma_{1zz}=\frac{h_0}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^{m+1}c_{mn}H_{1mn}(y) \tag{24}$$

$$K_{Dx}^{\mathrm{I}}=\lim_{x\to l_1}\sqrt{2(l_1-x)}\cdot D_{1z}=\frac{h_{70}}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^{m+1}c_{mn}H_{1mn}(y) \tag{25}$$

$$K_{Bx}^{\mathrm{I}}=\lim_{x\to l_1}\sqrt{2(l_1-x)}\cdot B_{1z}=\frac{h_{80}}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-1)^{m+1}c_{mn}H_{1mn}(y) \tag{26}$$

(II) on outer crack edge, $x=l_3$ and $0\le y<l_2$:

$$K_x^{\mathrm{E}}=\lim_{x\to l_3}\sqrt{2(x-l_3)}\cdot\sigma_{1zz}=-\frac{h_0}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}H_{1mn}(y) \tag{27}$$

$$K_{Dx}^{\mathrm{E}}=\lim_{x\to l_3}\sqrt{2(x-l_3)}\cdot D_{1z}=-\frac{h_{70}}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}H_{1mn}(y) \tag{28}$$

$$K_{Bx}^{\mathrm{E}}=\lim_{x\to l_3}\sqrt{2(x-l_3)}\cdot B_{1z}=-\frac{h_{80}}{\pi^2}\frac{\sqrt{2}}{\sqrt{l_3-l_1}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}H_{1mn}(y) \tag{29}$$

(III) on upper crack edge, $y=l_2$ and $l_1<x<l_3$:

$$K_y=\lim_{y\to l_2}\sqrt{2(y-l_2)}\cdot\sigma_{2zz}=\frac{2h_0}{\pi^2\sqrt{l_2}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}R_{2m}(x)Q_{mn} \tag{30}$$

$$K_{Dy}=\lim_{y\to l_2}\sqrt{2(y-l_2)}\cdot D_{2z}=\frac{2h_{70}}{\pi^2\sqrt{l_2}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}R_{2m}(x)Q_{mn} \tag{31}$$

$$K_{By}=\lim_{y\to l_2}\sqrt{2(y-l_2)}\cdot B_{2z}=\frac{2h_{80}}{\pi^2\sqrt{l_2}}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{mn}R_{2m}(x)Q_{mn} \tag{32}$$

where $Q_{mn} = -2\pi \dfrac{\Gamma(m+1+\frac{1}{2})}{m!} \dfrac{\Gamma(2n+1+\frac{1}{2})}{(2n)!}$.

## 6 NUMERICAL RESULTS AND DISCUSSION

Now, the analytical solution for generalized intensity factors have been derived strictly in forms of infinite polynomials, but to demonstrate the response law of cracks to kinds of factors such as loading frequency, numerical results are calculated and depicted for the generalized intensity factors. Considering the commonly used magneto-electro-elastic materials, the material constants of $BaTiO_3/CoFe_2O_4$ are adopted in all computations as shown in Table 1 to Table 3. According to the previous work[15], to achieve the available numerical solutions through the Schmidt method, only the first several terms in the infinite series should be taken into calculation. Here, the top ten terms are taken in computation, and the numerical results of the generalized intensity factors are graphically shown from Fig. 2 to Fig. 5 to indicate some significant response laws different from the static problem.

| $c_{11}$ $(N/m^2)$ | $c_{12}$ $(N/m^2)$ | $c_{13}$ $(N/m^2)$ | $c_{33}$ $(N/m^2)$ | $c_{44}$ $(N/m^2)$ | $\rho$ $(kg/m^3)$ |
|---|---|---|---|---|---|
| $22.6\times10^{10}$ | $12.5\times10^{10}$ | $12.4\times10^{10}$ | $21.6\times10^{10}$ | $4.4\times10^{10}$ | 5500.0 |

Table 1: Elastic stiffness constants and mass density of $BaTiO_3$-$CoFe_2O_4$

| $e_{31}$ $(C/m^2)$ | $e_{33}$ $(C/m^2)$ | $e_{15}$ $(C/m^2)$ | $f_{31}$ $(N/Am)$ | $f_{33}$ $(N/Am)$ | $f_{15}$ $(N/Am)$ |
|---|---|---|---|---|---|
| –2.2 | 9.3 | 5.8 | 290.2 | 350.0 | 275.0 |

Table 2: Piezoelectric constants and piezomagnetic constants of $BaTiO_3$-$CoFe_2O_4$

| $\varepsilon_{11}$ $(C^2/Nm^2)$ | $\varepsilon_{33}$ $(C^2/Nm^2)$ | $\mu_{11}$ $(Ns^2/C^2)$ | $\mu_{33}$ $(Ns^2/C^2)$ | $g_{11}$ $(Ns/VC)$ | $g_{33}$ $(Ns/VC)$ |
|---|---|---|---|---|---|
| $56.4\times10^{-10}$ | $63.5\times10^{-10}$ | $297.5\times10^{-6}$ | $83.5\times10^{-6}$ | $4.0\times10^{-9}$ | $4.7\times10^{-9}$ |

Table 3: Dielectric constants, magnetic permeablities and electromagnetic constants of BaTiO3-CoFe2O4

(I) According to Eqs.(24)-(32), the stress intensity factors, the electric intensity factors and the magnetic flux intensity factors possess of the same changing rule and only different in magnitudes coming from $h_0$, $h_{70}$ and $h_{80}$, which had been discussed in references [9, 17]. So during the following discussion, the electric displacement intensity factors and the magnetic flux intensity factors would not be elaborated particularly for the simplicity of

the article, except the generalized intensity factors show a different variation.

(II) Maintain the circular frequency $\omega$ satisfying $\omega l_2 / c_z = 0.8$ and $\omega l_2 / c_z = 1.2$, respectively, the stress intensity factors at the midpoints on the inner crack edge, the outer crack edge and the upper crack edge are depicted versus the inter space between two rectangular cracks as shown in Fig.2 and Fig.3. All stress intensity factors vibrate with the increasing of $l_1$, with the vibration amplitudes gradually tending to zero. Different from statistic problems that the stress intensity factor on the inner crack edges is always larger than ones on the outer crack edges, as shown in Fig.2(a) and Fig.3(a), the stress intensity factors at inner and outer crack edge alternately occupy the dominant position, which makes it more difficult to predicting crack converging. When $\omega l_2 / c_z = 0.8$, the cracks are more inclined to converging for $l_1 / l_2 < 0.5$. When $\omega l_2 / c_z = 1.2$, the cracks will expand outward separately for $l_1 / l_2 < 0.5$. But for $l_1 / l_2 \geq 0.5$, even though the stress intensity factor on one edge is slightly larger than that on the other edge, it is hard to say whether cracks join together or expand separately. Additionally, the stress intensity factors at the midpoint on the upper edge are smaller than that on the inner and outer edges in these two cases.

(III) Fig.4 depicts the trend of $K^*$ at midpoint on each crack edge versus the dimensionless circular frequency $\omega_0 = \omega l_1 / c_z$. Within the researched load frequency range, when the frequency is not large $\omega_0 < 1.0$, the two cracks are capable of converging. However, with larger $\omega_0$, the main behavior of two coplanar cracks is enlarging separately. Note that, as shown in Fig.4, the most dangerous load frequency prompting cracks to converge locates at $\omega_0 \approx 0.7$. Additionally, in the case of dangerous frequency, the midpoint of crack edge is not the location easy to crack according to the stress intensity factors curve as shown in Fig.5. the whole crack edges should be given adequate attention in the dynamic fracture analysis.

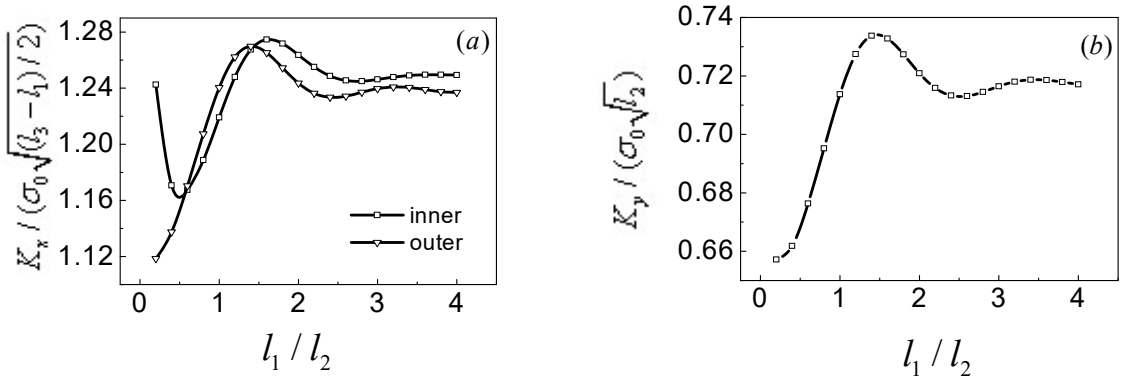



Fig.2: The stress intensity factors versus $l_1 / l_2$ for $l_2 = 1.0$, $l_3 - l_1 = 1.0$, $\omega l_2 / c_z = 0.8$, $D_0 / \varepsilon_0 = 1.0 \times 10^8$ and $B_0 / \mu_0 = 1.0 \times 10^5$.

(a) The stress intensity factors at the middle point of inner and outer crack edges

(b) The stress intensity factors at the middle point of upper crack edge

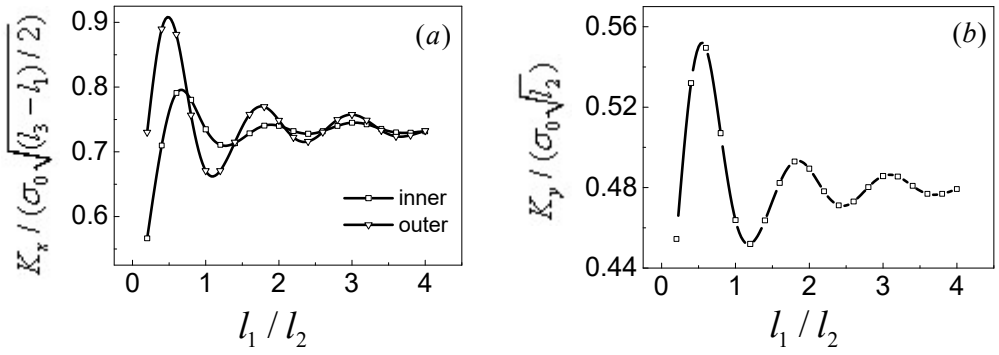



Fig.3: The stress intensity factors versus $l_1 / l_2$ for $l_2 = 1.0$, $l_3 - l_1 = 1.0$, $\omega l_2 / c_z = 1.2$, $D_0 / \varepsilon_0 = 1.0 \times 10^8$ and $B_0 / \mu_0 = 1.0 \times 10^5$.

(a) The stress intensity factors at middle points of inner and outer crack edges

(b) The stress intensity factors at the middle point of upper crack edge

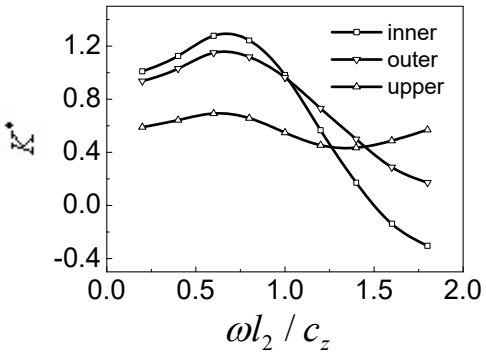



Fig.4: The stress intensity factors at middle point of inner, outer and upper crack edge versus $\omega l_2 / c_z$ for $l_1 = 0.2$, $l_2 = 1.0$, $l_3 = 1.2$, $D_0 / \varepsilon_0 = 1.0 \times 10^8$ and $B_0 / \mu_0 = 1.0 \times 10^5$.

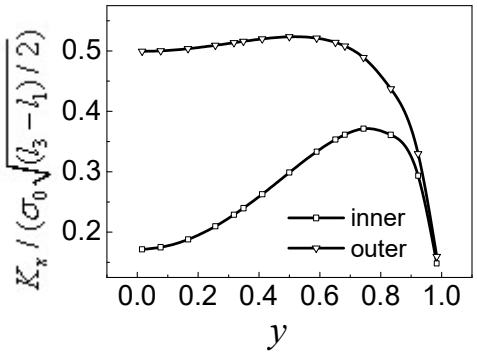



Fig.5: The stress intensity factors on inner and outer crack edges versus $y$ for $l_1 = 0.2$, $l_2 = 1.0$, $l_3 = 1.2$, $\omega l_2 / c_z = 1.4$, $D_0 / \varepsilon_0 = 1.0 \times 10^8$ and $B_0 / \mu_0 = 1.0 \times 10^5$.

## 7 CONCLUSIONS

Coupling fracture of coplanar two rectangular cracks was addressed considering piezoelectric effect, piezomagnetic effect and magneto-electric effect of materials and the load frequency. limited-permeable crack model was introduced to emulating the reality and using generalized Almansi' theorem and Fourier Transform, the analytical solution of whole generalized fields and generalized intensity factors were deduced strictly. In order to demonstrate intuitively the affection of geometry of cracks and load frequency on the initiation behavior of cracks, numerical computation was depicted with the help of Schmidt method which permitted higher accuracy with less computation. Some important factors affecting fracture behavior were pointed out to advance the material health monitory.

## ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (11202050) and Jiangsu Natural Science Foundation of China(BK2012318).

## REFERENCE

[1] J. Vansuchtelen, Product properties-new application of composite-materials. *Philips Research Reports*, **27**, 28-37, 1972

[2] J.K. Du, Y.P. Shen, B. Gao, Scattering of anti-plane shear waves by a single crack in an unbounded transversely isotropic electro-magneto-elastic medium. *Appl Math Mech-engl*, **25**, 1344-1353, 2004.

[3] G.C. Sih, Z.F. Song, Magnetic and electric poling effects associated with crack growth in BaTiO3-CoFe2O4 composite. *Theor Appl Fract Mech*, **39**, 209-227, 2003.

[4] X.C. Zhong, X.F. Li, A finite length crack propagating along the interface of two dissimilar magnetoelectroelastic materials. *Int J Eng Sci*, **44**, 1394-1407, 2006.

[5] X.C. Zhong, Analysis of a dielectric crack in a magnetoelectroelastic layer. *Int J Solids Struct*, **46**, 4221-4230, 2009.

[6] W.J. Feng, Y. Xue, Z.Z. Zou, Crack growth of an interface crack between two dissimilar magneto-electro-elastic materials under anti-plane mechanical and in-plane electric magnetic impact. *Theor Appl Fract Mech*, **43**, 376-394, 2005.

[7] W.J. Feng, Y.S. Li, Z.H. Xu, Transient response of an interfacial crack between dissimilar magnetoelectroelastic layers under magnetoelectromechanical impact loadings: Mode-I problem. *Int J Solids Struct*, **46**, 3346-3356, 2009.

[8] T.H. Hao, Z.Y. Shen, A New Electric Boundary-Condition of Electric Fracture-Mechanics and Its Applications. *Eng Fract Mech*, **47**, 793-802, 1994.

[9] Z.G. Zhou, P.W. Zhang, L.Z. Wu, Solutions to a limited-permeable crack or two limited-permeable collinear cracks in piezoelectric/piezomagnetic materials. *Arch Appl Mech*, **77**, 861-882, 2007.

[10] X.C. Zhong, X.F. Li, Magnetoelectroelastic analysis for an opening crack in a piezoelectromagnetic solid. *Eur J Mech A-solid*, **26**, 405-417, 2007.

[11] Z.G. Zhou, Z.T. Chen, Basic solution of a Mode-I limited-permeable crack in functionally graded piezoelectric/piezomagnetic materials. *Int J Solids Struct*, **45**, 2265-2296, 2008.

[12] X.F. Li, K.Y. Lee, Crack growth in a piezoelectric material with a Griffith crack perpendicular to the poling axis. *Philos Mag*, **84**, 1789-1820, 2004.

[13] W.Q. Chen, K.Y. Lee, H. Ding, General solution for transversely isotropic magneto-electro-thermo-elasticity and the potential theory method. *Int J Eng Sci*, **42**, 1361-1379, 2004.

[14] P.M. Morse, H. Feshbach, *Methods of Theoretical Physics*. McGraw-Hill, 1958.

[15] S. Itou, 3-Dimensional wave-propagation in a cracked elastic solid. *J Appl Mech-t Asme*, **45**, 807-811, 1978.

[16] P.W. Zhang, Z.G. Zhou, L.Z. Wu, Coupled field state around three parallel non-symmetric cracks in a piezoelectric/piezomagnetic material plane. *Arch Appl Mech*, **79**, 965-979, 2009.

[17] P.W. Zhang, Z.G. Zhou, B. Wang, Dynamic behavior of two collinear interface cracks between two dissimilar functionally graded piezoelectric/piezomagnetic material strips. *Appl Math Mech-engl*, **28**, 615-625, 2007.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DAMAGE DETECTION OF A BRIDGE BY PARAMETRIC STATISTICAL MOMENT METHOD

**Isabella Failla**[1], **Nicola Impollonia**[2], **and Giuseppe Ricciardi**[1]

[1] Department of Engineering, University of Messina
C.da Di Dio, Villaggio S. Agata - 98166 Messina, Italy
ifailla@unime.it, gricciardi@unime.it

[2] Department of Civil Engineering and Architecture
Viale Andrea Doria, 6 - 95123 Catania, Italy
nimpo@unict.it

**Keywords:** Damage detection, Explicit solution, Stiffness identification, Statistical moments.

**Abstract.** The statistical moment-based damage detection method with approximate parametric solutions of the stationary second-order moments of the response is proposed for damage detection of a bridge. The method is based on an approximate explicit solution recently proposed by the authors, which allows to explicitly relate second-order moments of the nodal displacements and velocity to the structural stiffness. Application on a bridge is reported to check the consistency of the method and to investigate the influence of the entity of the damage and of the application point of the force on the identification procedure.

## 1 INTRODUCTION

Moment based approaches proved to be quite effective in the identification of structural damages and positions. Statistical measures of the response under stochastic excitation have succeeded numerically and experimentally both in the non-linear and linear settings.

Nonlinearity arising from a breathing crack affects higher order statistics which can be exploited to detect and locate stiffness discontinuities with high accuracy [1]-[5].

Statistical moments are suitable damage indices also if linear behavior is preserved when the damage arises. In fact they are sensitive to local structural damage but insensitive to measurement noise so that the statistical moment-based damage detection (SMBDD) method can be conveniently resorted to [6], [7]. The SMBDD method can be more effectively employed if approximate parametric solutions for nodal stationary statistical moment are available. The benefit of having explicit solutions is conspicuous when the least squares method is applied because the residual between the simulated response and the actual statistical moments is parameterized with respect to the quantities to be identified. This offers great convenience in applying Newton's method to search for the parameters of stiffness and damping inversely when the objective function is minimized. The parametric relationship between stationary second order moments proposed in [8] are resorted to.

The objective function is defined herein as the weighted sum of squared differences between the measured data from a simulated experiment and the corresponding analytical parametric values of nodal stationary second order displacements and velocity under Gaussian input. The weights can be chosen accordingly to the sensitivity of the related measure with respect to structural parameters. The procedure has already been applied to semi-rigid connections identification [9], and its robustness against measurement noise have been assessed [10].

In this paper the method is applied to damage identification in a frame grid modeling a span bridge. The simulated experiments allow different damage entity and position, whereas a low number of measurement point is foreseen. A single frame element is given a reduced stiffness modulus so to reproduce damage. The numerical investigation always gives back the correct identification for damage entity and position when the structure is loaded by a white noise load concentrated on a node.

## 2 SECOND ORDER MOMENT DYNAMIC EQUATION

Let us consider an n-DOF structural system whose dynamic behaviour is predicted by an FE linear model, characterized by the mass, damping and stiffness $(n\times n)$ matrices $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$, respectively. Moreover, let us assume that the structure is subjected to a zero mean stationary Gaussian white noise excitation $W(t)$ with intensity $q=2\pi S_0$, $S_0$ being the constant power spectral density. The motion differential equations of the system are:

$$\mathbf{M}\ddot{\mathbf{u}}(t)+\mathbf{C}\dot{\mathbf{u}}(t)+\mathbf{K}\mathbf{u}(t)=\boldsymbol{\tau}W(t) \tag{1}$$

where $\mathbf{u}(t)$, $\dot{\mathbf{u}}(t)$ and $\ddot{\mathbf{u}}(t)$ are the n-vectors of nodal displacements, velocities and accelerations, respectively, and $\boldsymbol{\tau}$ is the influence forcing n-vector.

Let us introduce the modal coordinate transformation

$$\mathbf{u}(t)=\boldsymbol{\Phi}\mathbf{q}(t) \tag{2}$$

where $\mathbf{q}(t)$ is the *m*-vector of the modal coordinates, with $m \le n$, and $\mathbf{\Phi}$ is the reduced modal matrix, of order $n \times m$, solution of the following eigenproblem:

$$\mathbf{K\Phi} = \mathbf{\Omega M\Phi} \tag{3}$$

normalized with respect to the mass matrix $\mathbf{M}$, so that $\mathbf{\Phi}^T\mathbf{M\Phi} = \mathbf{I}_m$, whereas $\mathbf{\Phi}^T\mathbf{K\Phi} = \mathbf{\Omega} = diag[\omega_j^2]$ is the diagonal matrix listing the squares of the first *m* natural frequencies of the predicted structure. Substituting Eq.(2) into Eq.(3), the latter becomes

$$\ddot{\mathbf{q}}(t) + \mathbf{\Xi}\dot{\mathbf{q}}(t) + \mathbf{\Omega q}(t) = \mathbf{p}W(t) \tag{4}$$

where $\mathbf{p} = \mathbf{\Phi}^T\boldsymbol{\tau}$ and $\mathbf{\Xi} = \mathbf{\Phi}^T\mathbf{C\Phi}$. Eq.(4) rules the dynamic behaviour of the predicted FE model in the modal space.

By using the 2*m*-dimensional vector state approach, Eq.(4) can be written in the following matrix form:

$$\dot{\mathbf{Z}}(t) + \mathbf{D}\,\mathbf{Z}(t) = \mathbf{v}W(t) \tag{5}$$

where $\mathbf{Z}(t) = [\mathbf{q}^T(t) \quad \dot{\mathbf{q}}^T(t)]^T$ is the 2*m*-vector of the modal state variables and

$$\mathbf{D} = \begin{bmatrix} \mathbf{0} & -\mathbf{I}_m \\ \mathbf{\Omega} & \mathbf{\Xi} \end{bmatrix} \qquad \mathbf{v} = \begin{bmatrix} \mathbf{0} \\ \mathbf{p} \end{bmatrix} \tag{6}$$

being $\mathbf{I}_m$ the identity matrix of order $m$.

If the physical properties of the system, such as mass, damping and stiffness, are known, the response process is a Gaussian stationary one and its probabilistic characterization is determined once the stationary second order moments of the nodal displacements and velocities are evaluated. The latter are related to the second order moments of the modal coordinates. In fact, by introducing the *2n*-vector of the nodal state variables $\mathbf{Y}(t) = [\mathbf{u}^T(t) \quad \dot{\mathbf{u}}^T(t)]^T$, the following relation holds

$$\mathbf{m}_{\mathbf{Y},2}(t) = \mathbf{\Gamma}^{[2]}\mathbf{m}_{\mathbf{Z},2}(t) \tag{7}$$

In Eq.(7), $\mathbf{m}_{\mathbf{Y},2}(t) = \mathbf{m}_2[\mathbf{Y}(t)] = E[\mathbf{Y}(t) \otimes \mathbf{Y}(t)] = E[\mathbf{Y}^{[2]}(t)]$ is the $4n^2$-vector which collects all the second order moments of the displacements and velocities of the nodal coordinates, and $\mathbf{m}_{\mathbf{Z},2}(t) = \mathbf{m}_2[\mathbf{Z}(t)] = E[\mathbf{Z}(t) \otimes \mathbf{Z}(t)] = E[\mathbf{Z}^{[2]}(t)]$ is the $4m^2$-vector which collects the second order moments of the displacement and velocities of the *m* selected modal coordinates, where the symbol $\otimes$ means Kronecker product and the exponent in square brackets means Kronecker power [11],[12]. Furthermore,

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Phi} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi} \end{bmatrix} \tag{8}$$

is a transformation matrix of order $2n \times 2m$, so that $\mathbf{Y}(t) = \mathbf{\Gamma Z}(t)$, and $\mathbf{\Gamma}^{[2]} = \mathbf{\Gamma} \otimes \mathbf{\Gamma}$.

The differential equations describing the time evolution of the vector $\mathbf{m}_{\mathbf{Z},2}(t)$ can be written as follows [13]

$$\dot{\mathbf{m}}_{\mathbf{Z},2}(t) + \mathbf{D}_2\mathbf{m}_{\mathbf{Z},2}(t) = \mathbf{v}_2 q \tag{9}$$

where

$$\mathbf{D}_2 = \mathbf{D} \otimes \mathbf{I}_{2m} + \mathbf{I}_{2m} \otimes \mathbf{D}, \qquad \mathbf{v}_2 = \mathbf{v} \otimes \mathbf{v} = \mathbf{v}^{[2]} \tag{10}$$

Note that all the cross moments appear twice in the vector $\mathbf{m}_{\mathbf{Z},2}(t)$. Then a condensation can be employed in order to reduce the size of the problem by clearing away the repetitions. So operating the problem dimension reduces from $4m^2$ to $N = 2m^2 + m$.

The stationary response is given as solution of the algebraic equations obtained removing the time dependence from Eq.(9):

$$\mathbf{D}_2 \mathbf{m}_{\mathbf{Z},2} = \mathbf{v}_2 q \tag{11}$$

and the vector $\mathbf{m}_{\mathbf{Y},2}$ of the stationary second order moments of the nodal displacements and velocities can be evaluated by Eq.(7). From the previous equation it appears that the evaluation of the stationary second order moments of the modal coordinates and its velocities requires the inversion of the matrix $\mathbf{D}_2$.

## 3 STIFFNESS PARAMETERS VARIABILITY

Let us assume, in agreement with realistic models, the mass parameters exhibit negligible fluctuations and can be considered known. On the contrary, the stiffness of the elements commonly suffer from significant uncertainty, and their values may deviate from predicted ones. For example, this is the case of variation introduced by damage. The damage identification could be accomplished exploiting by an inverse parametric analysis aimed to detect its location and entity starting from the knowledge of statistical response measures at some points of the structure excited by a white noise, according to the SMBDD method [6].

In the following, the SMBDD method is improved according to an explicit parameterization of the stationary second order moments of the structural response, which are expressed as analytical functions of element stiffnesses. The Global stiffness matrix can deviate from the predicted value $\mathbf{K}$ and it can be written in the following form:

$$\tilde{\mathbf{K}}(\boldsymbol{\alpha}) = \mathbf{K} + \sum_{i=1}^{R_K} \alpha_i \mathbf{K}^{(i)} \tag{12}$$

where $\tilde{\mathbf{K}}(\boldsymbol{\alpha})$ is the updated global stiffness matrix, while $\mathbf{K}$ and $\mathbf{K}^{(i)}$ are known matrices determined by assembling the frame stiffness matrix and $\boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{R_K}]^T$ is the vector collecting the stiffness parameter fluctuations. In Eq.(12), $\mathbf{K}$ represents the predicted value of the stiffness matrix and the summation denotes its fluctuating components. Accordingly, the updated dynamic matrix $\tilde{\mathbf{D}}(\boldsymbol{\alpha})$ can be written as:

$$\tilde{\mathbf{D}}(\boldsymbol{\alpha}) = \mathbf{D} + \sum_{i=1}^{R_K} \alpha_i \mathbf{D}^{(i)} \tag{13}$$

where $\mathbf{D}$ is the predicted dynamic matrix and

$$\mathbf{D}^{(i)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \boldsymbol{\Omega}^{(i)} & \mathbf{0} \end{bmatrix} \tag{14}$$

with

$$\boldsymbol{\Omega}^{(i)} = \boldsymbol{\Phi}^T \mathbf{K}^{(i)} \boldsymbol{\Phi} \tag{15}$$

$\boldsymbol{\Phi}$ being the modal matrix relative to the predicted structure.

### 3.1 Stationary Stochastic Response

The stationary second order moments of the modal response are solution of the following set of algebraic equations :

$$\tilde{\mathbf{D}}_2(\alpha)\tilde{\mathbf{m}}_{\mathbf{Z},2}(\alpha) = \mathbf{v}_2 q \tag{16}$$

where

$$\tilde{\mathbf{D}}_2(\alpha) = \mathbf{D}_2 + \sum_{i=1}^{R} \alpha_i \mathbf{D}_2^{(i)} \tag{17}$$

being

$$\mathbf{D}_2^{(i)} = \mathbf{D}^{(i)} \otimes \mathbf{I}_{2m} + \mathbf{I}_{2m} \otimes \mathbf{D}^{(i)} \tag{18}$$

Eq.(16) is the parametric counterpart of Eq.(11), with unknowns given by the vector of the second order moments of the nodal response $\tilde{\mathbf{m}}_{\mathbf{Z},2}(\boldsymbol{\alpha})$.

Solving the system governed by Eq.(16) means to obtain the explicit relationships between the vector $\tilde{\mathbf{m}}_{\mathbf{Z},2}(\boldsymbol{\alpha})$ and the parameter vector $\boldsymbol{\alpha}$. This cannot be generally done by inverting $\tilde{\mathbf{D}}_2(\boldsymbol{\alpha})$, which is a parametric matrix of large dimension, so that different procedures solving this problem in an approximate way have been proposed in the literature. The approximate analytical solution presented by [8], briefly recovered in the next section, is adopted herein.

## 4 THE EXPLICIT ANALYTICAL SOLUTION

Let us consider the case in which only one fluctuating parameter $\alpha_i$ is present,so that Eq.(16) is rewritten as:

$$\tilde{\mathbf{D}}_2(\alpha_i)\tilde{\mathbf{m}}_{\mathbf{Z},2}(\alpha_i) = \mathbf{v}_2 q \tag{19}$$

The vector of the second order moments of the modal response can be written as follows:

$$\tilde{\mathbf{m}}_{\mathbf{Z},2}(\alpha_i) = \mathbf{m}_{\mathbf{Z},2} + \mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i) \tag{20}$$

where $\mathbf{m}_{\mathbf{Z},2}$ is the vector of the stationary second order moments of the predicted structure (for $\alpha_i = 0$), solution of Eq.(11). Moreover, the matrix $\tilde{\mathbf{D}}_2(\alpha_i)$ takes on the following simpler form:

$$\tilde{\mathbf{D}}_2(\alpha_i) = \mathbf{D}_2 + \alpha_i \mathbf{D}_2^{(i)} \tag{21}$$

By substituting Eq.(20) and Eq.(21) into Eq.(19), one obtains:

$$(\mathbf{D}_2 + a_i \mathbf{D}_2^{(i)})[\mathbf{m}_{\mathbf{Z},2} + \mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i)] = \mathbf{v}_2 q \tag{22}$$

Taking into account Eq.(11), the following equation for the unknown vector $\mathbf{m}_{\mathbf{Z},2}^{(i)}(\theta_i)$ is obtained:

$$[\mathbf{D}_2 + \alpha_i \mathbf{D}_2^{(i)}]\mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i) = -\alpha_i \mathbf{D}_2^{(i)} \mathbf{m}_{\mathbf{Z},2} \tag{23}$$

The solution of Eq.(23) appears not straightforward at a first glance since the inverse of the parametric matrix $(\mathbf{D}_2 + \alpha_i \mathbf{D}_2^{(i)})^{-1}$ is required. However, the expression of the vector $\mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i)$ as an explicit function of the parameter $\alpha_i$ can be readily obtained as shown in the next.

Let us consider the right and left eigenproblems related to the $(N \times N)$ matrix $\mathbf{A}_2^{(i)} = \mathbf{D}_2^{-1}\mathbf{D}_2^{(i)}$ (assuming that the condensation has been employed, i.e. $N = 2m^2 + m$):

$$\mathbf{A}_2^{(i)} \boldsymbol{\Psi}_R^{(i)} = \boldsymbol{\Psi}_R^{(i)} \boldsymbol{\Lambda}^{(i)} \qquad\qquad \boldsymbol{\Psi}_L^{(i)} \mathbf{A}_2^{(i)} = \boldsymbol{\Lambda}^{(i)} \boldsymbol{\Psi}_L^{(i)} \tag{24}$$

Since $\mathbf{A}_2^{(i)}$ is a non-defective matrix, the following ortonormalization condition can be applied:

$$\boldsymbol{\Psi}_L^{(i)} \boldsymbol{\Psi}_R^{(i)} = \mathbf{I}_N, \qquad\qquad \boldsymbol{\Psi}_L^{(i)} \mathbf{A}_2^{(i)} \boldsymbol{\Psi}_R^{(i)} = \boldsymbol{\Lambda}^{(i)} \tag{25}$$

The $p \times N$ matrix $\boldsymbol{\Psi}_L^{(i)}$ and the $N \times p$ matrix $\boldsymbol{\Psi}_R^{(i)}$ collect, respectively, the $p$ significant left and right eigenvectors, i.e. the eigenvectors related to the nonzero eigenvalues $\lambda_j^{(i)}$ (with $j = 1, 2, \ldots, p$) listed in the diagonal $p \times p$ matrix $\boldsymbol{\Lambda}^{(i)}$. By introducing the following coordinate transformation:

$$\mathbf{m}_{\mathbf{Z},2}^{(i)}(\theta_i) = \boldsymbol{\Psi}_R^{(i)} \mathbf{n}_{\mathbf{Z},2}^{(i)}(\theta_i) \tag{26}$$

into Eq.(23), pre-multiplying both sides by $\boldsymbol{\Psi}_L^{(i)}$ and taking into account Eqs.(25), Eq.(23) becomes:

$$(\mathbf{I} + \alpha_i \boldsymbol{\Lambda}^{(i)}) \mathbf{n}_{\mathbf{Z},2}^{(i)}(\alpha_i) = -\alpha\theta_i \boldsymbol{\Psi}_L^{(i)} \mathbf{A}_2^{(i)} \mathbf{m}_{\mathbf{Z},2} \tag{27}$$

From the previous equation, it is easy to obtain the exact explicit relationship between the vector $\mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i)$ and the generic parameter $\alpha_i$:

$$\mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i) = -\alpha_i \boldsymbol{\Psi}_R^{(i)} \left(\mathbf{I} + \alpha_i \boldsymbol{\Lambda}^{(i)}\right)^{-1} \boldsymbol{\Psi}_L^{(i)} \mathbf{A}_2^{(i)} \mathbf{m}_{\mathbf{Z},2} \tag{28}$$

If more parameters are present, as a first approximation the cross effects can be neglected and the superposition principle can be applied. Therefore, the solution in the modal coordinates is given by the sum of the single contributions, that is:

$$\tilde{\mathbf{m}}_{\mathbf{Z},2}(\boldsymbol{\alpha}) = \mathbf{m}_{\mathbf{Z},2} + \sum_{i=1}^{R} \mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i) \tag{29}$$

where $R$ is the number of the fluctuating parameters and the generic $\mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i)$ is given by Eq.(28). Taking into account Eq.(7), the solution in terms of nodal displacement and velocities is

$$\tilde{\mathbf{m}}_{\mathbf{Y},2}(\boldsymbol{\alpha}) = \mathbf{m}_{\mathbf{Y},2} + \sum_{i=1}^{R} \boldsymbol{\Gamma}^{[2]} \mathbf{m}_{\mathbf{Z},2}^{(i)}(\alpha_i) \tag{30}$$

Eq.(30), which gives an approximate solution if $R \geq 2$ because the joints effects of the parameter fluctuations are neglected, is sufficiently accurate for solving the inverse problem arising in structural identification. Eventually, more accurate solutions can be obtained by introducing cross terms [8].

## 5 INVERSE MEASUREMENT PROBLEM

The approximate explicit solution produced by Eq.(30) can be exploited for element stiffness identification starting from the knowledge of experimental measurements of the stationary second order moments of the response at some points of the structure. Assume that some measurements $\hat{m}_{y_k,2} = E[\hat{y}_k^2]$ are available, where $\hat{y}_k$ is the measured displacement or velocity of the k-th degree of freedom of the structure (the cap over the variable means "experimental"). The corresponding parametric expressions $\tilde{m}_{y_k,2}(\alpha)$ is evaluated as components of the vector given by Eq.(30). The inverse problem consists in retrieving the parameters $\alpha$ from given displacement or velocity measurements. Then, the identification of the $R$ parameters may be obtained using the least squares estimation approach, where the objective function

$$f(\alpha) = \sum_{k=1}^{N_R} \left[ W_k \frac{\tilde{m}_{y_k,2}(\alpha) - \hat{m}_{y_k,2}}{\hat{m}_{y_k,2}} \right]^2 \tag{31}$$

is to be minimized and $N_R$ is the number of measurements.

## 6 APPLICATION: IDENTIFICATION OF DAMAGE IN A BRIDGE

The proposed damage identification procedure has been applied to a span of a bridge, discretized by beam finite elements, simply supported on two piers at its end (see Fig.1 and 2). The simulated experiment has been performed in relation to the FE model depicted in Fig. 3, where a 18 element discretization has been used. The data of the longitudinal beams are: second moment of area $I = 0.9803\, m^4$, polar (torsional) second moment of area $J = 0.1765\, m^4$, total span $L = 34.25\, m$, mass density per unit length $\rho_{bl} = 2.55 \times 10^3\, kg/m$. The data of the transversal beams are: second moment of area $I = 0.4667\, m^4$, polar (torsional) second moment of area $J = 0.0211\, m^4$, total span $L = 8.50\, m$, mass density per unit length $\rho_{bt} = 3.50 \times 10^3\, kg/m$, mass density per unit area of the deck $\rho_d = 3.12 \times 10^3\, kg/m^2$.

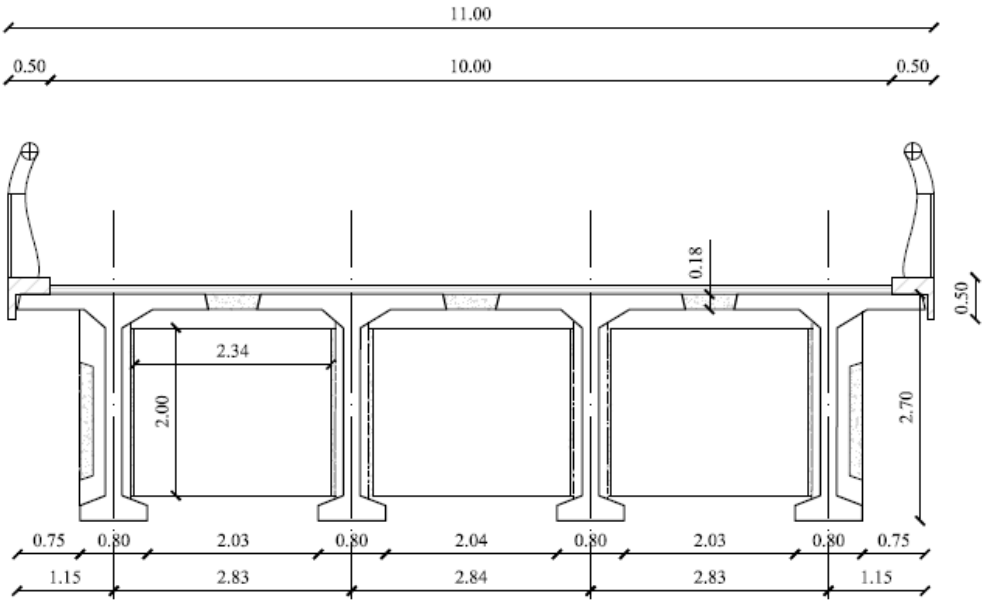



Figure 1: Cross Section of the bridge

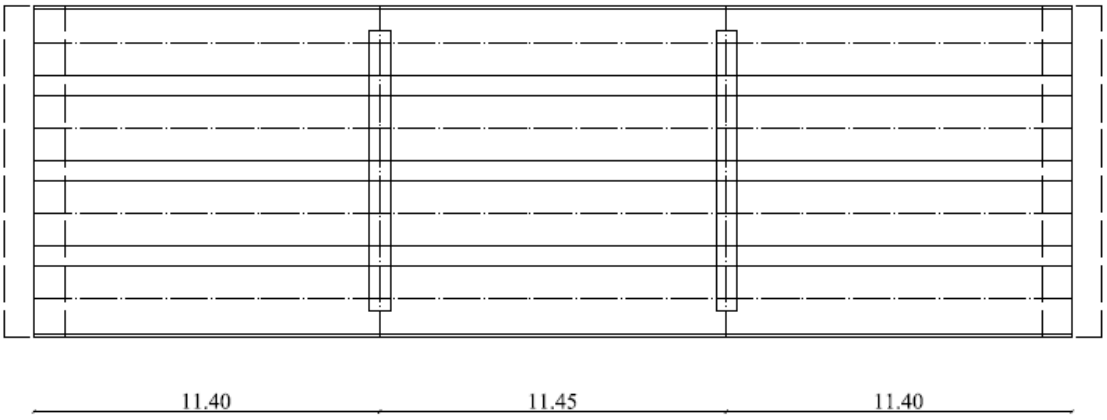



Figure 2: Schematic sketch of the bridge

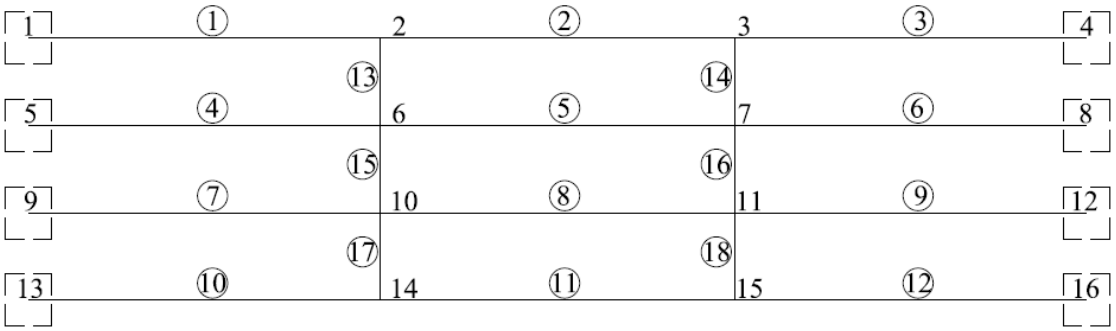



Figure 3: FE model of the bridge

Let us assume that the Young modulus of all the elements has the value $E = 3.46 \times 10^7\, kN/m^2$, except one with value $\hat{E}_i = E(1+\hat{\alpha}_i)$, where $\hat{\alpha}_i$ is a negative coefficient (with $-1 < \hat{\alpha}_i < 0$), representing the damaged zone. For simplicity, a damage due to a reduction of the Young modulus of a beam can be similarly taken into account. The bridge has been subjected to a vertical point force applied at a given node of the model, represented by a white noise. Of course more general loading could be treated. Suitable damping properties of the structure have been assumed in the cases above considered and a lumped mass matrix has been implemented.

The identification strategy relies upon the least squares estimation approach which is carried on for all possible damage scenarios by assuming a single value for the Young modulus at all the elements but the one where the damage is tentatively located. The real damage position is singled out by selecting the damage scenario with smaller residue (see Eq. (31)).

The numerical examples assume a limited number of measurements, namely $N_R = 4$, as given by second order stationary displacement at nodes 2, 3, 6 and 7.

The investigation starts from considering that the damage is located alternatively at the 10-th, 11-th and 17-th element of the simulated experimental model, with a value $\hat{\alpha}$ varying from $-0.3$ to $-0.05$.

Let us assume that the damage is present on the *j*-th element (with $j = 10, 11, 17$), with unknown value $\alpha_D^{(j)}$, and the remaining elements are supposed undamaged with value $\alpha_{ND}^{(j)}$

(with $j \neq 10, 11, 17$). Identification of the damage entity and position is determined in the element in which one has the smallest value of the residue in the minimization problem Eq.(31).

| $\hat{\alpha}$ \ FE | 10-th | Error % | 11-th | Error % | 17-th | Error % |
|---|---|---|---|---|---|---|
| –0.3 | -0.3450 | 15.00 | -0.3953 | 31.77 | -0.3114 | 3.80 |
| –0.2 | -0.2194 | 9.70 | -0.2316 | 15.83 | -0.2042 | 2.13 |
| –0.1 | -0.1046 | 4.64 | -0.1072 | 7.26 | -0.1009 | 0.93 |
| –0.05 | -0.0511 | 2.27 | -0.0517 | 3.51 | -0.0502 | 0.44 |

Table 1: Stiffness reduction identified by the parametric model for damage located at the 10-th, 11-th and 17-th element of the beam, with $\hat{\alpha} = -0.3, -0.2, -0.1$ and $-0.05$, and point force at node 3.

| $\hat{\alpha}$ \ FE | 10-th | Error % | 11-th | Error % | 17-th | Error % |
|---|---|---|---|---|---|---|
| –0.3 | -0.3462 | 15.40 | -0.3373 | 12.45 | -0.2949 | 1.68 |
| –0.2 | -0.2195 | 9.77 | -0.2161 | 8.09 | -0.1986 | 0.69 |
| –0.1 | -0.1046 | 4.66 | -0.1038 | 3.87 | -0.0998 | 0.19 |
| –0.05 | -0.0511 | 2.28 | -0.0509 | 1.89 | -0.0499 | 0.06 |

Table 2: Stiffness reduction identified by the parametric model for damage located at the 10-th, 11-th and 17-th element of the beam, with $\hat{\alpha} = -0.3, -0.2, -0.1$ and $-0.05$, and point force at node 10.

| Element | Residue | $\alpha$ |
|---|---|---|
| 1 | 6.0707E-07 | 0.00023 |
| 2 | 2.9979E-08 | -0.00731 |
| 3 | 3.3123E-07 | -0.00355 |
| 4 | 5.3909E-07 | 0.00167 |
| 5 | 5.1163E-07 | -0.00302 |
| 6 | 1.5169E-07 | -0.00555 |
| 7 | 4.7854E-07 | 0.01177 |
| 8 | 1.2607E-07 | -0.02684 |
| 9 | 3.4168E-08 | 0.01527 |
| 10 | 1.6937E-07 | 0.01194 |
| 11 | 3.2713E-07 | 0.01237 |
| 12 | 6.0940E-07 | -0.00023 |
| 13 | 4.6062E-08 | -0.05273 |
| 14 | 1.9527E-07 | 0.01106 |
| 15 | 7.6440E-08 | -0.07097 |
| 16 | 3.8792E-07 | 0.00221 |
| 17 | 8.2491E-10 | -0.20426 |
| 18 | 5.0598E-07 | 0.00117 |

Table 3: Stiffness reduction identified by the parametric model for damage located at the 17-th element of the beam, with $\hat{\alpha} = -0.2$, and point force at node 3.

| Element | Residue | $\alpha$ |
|---|---|---|
| 1 | 3.4831E-06 | 0.00969 |
| 2 | 3.4501E-07 | 0.00948 |
| 3 | 7.1014E-06 | 0.00377 |
| 4 | 7.8148E-06 | -0.00041 |
| 5 | 7.3201E-06 | 0.01713 |
| 6 | 7.8605E-06 | -0.00310 |
| 7 | 7.7307E-06 | -0.00903 |
| 8 | 3.4112E-07 | -0.03551 |
| 9 | 1.4893E-06 | -0.03172 |
| 10 | 2.3522E-06 | -0.03710 |
| 11 | 3.6456E-07 | -0.02275 |
| 12 | 2.1483E-06 | -0.04871 |
| 13 | 5.2607E-06 | 0.08237 |
| 14 | 5.9452E-06 | 0.05694 |
| 15 | 5.2748E-06 | -0.18271 |
| 16 | 6.9742E-06 | -0.05382 |
| 17 | 2.0889E-10 | -0.19862 |
| 18 | 7.0393E-06 | -0.08502 |

Table 4: Stiffness reduction identified by the parametric model for damage located at the 17-th element of the beam, with $\hat{\alpha} = -0.2$, and point force at node 10.

Table 1 and 2 show the results of the identification procedure here proposed, for different location and entity of damage $\hat{\alpha}$ for point force applied at node 3 and at node 10, revealing as the accuracy of the method increase when the value of damages decrease. The method is capable to catch also small values of the damage, common in real cases. Table 3 and 4 show the results of the identification procedure here proposed at element 17-th for a point force applied at node 3 and at node 10, and entity of damage $\hat{\alpha} = -0.2$.

## 7 CONCLUSIONS

The statistical moment-based damage detection method has been applied to identify the stiffness of a grid structure modelling the span of a bridge. The numerical applications reveal that the method is effective when damage localization should be achieved with satisfactory accuracy. The simulated experiments are related to different scenarios where a single frame is given a reduced stiffness so to reproduce a fictitious damage. In all the cases considered, the proposed method, taking advantage of accurate approximate relationship between nodal second order statistical moments and element stiffnesses, is able to spot damage position and its entity.

## REFERENCES


[1] A. Rivola, P.R. White, Bispectral analysis of the bilinear oscillator with application to the detection of fatigue cracks. *Journal of Sound and Vibration*, **216**, 889-910, 1998.

[2] P. Cacciola, N. Impollonia, G. Muscolino, Crack detection and location in a damaged beam vibrating under white noise. *Computers and Structures,* **81** 1773-1782, 2003.

[3] S. Benfratello, P. Cacciola, N. Impollonia, A. Masnata, G. Muscolino, Numerical and experimental verification of a technique for locating a fatigue crack on beams vibrating under Gaussian excitation. *Engineering Fracture Mechanics,* **74**, 2992-3001, 2007.

[4] M. Breccolotti, A.L. Materazzi, I. Venanzi, Identification of the nonlinear behaviour of a cracked RC beam through the statistical analysis of the dynamic response. *Structural Control and Health Monitoring*, **15**, 416-435, 2008.

[5] D. Boungou, F. Guillet, M. El Badaoui, P. Lyonnet, T. Rosario, Fatigue damage detection using cyclostationarity. Mechanical Systems and Signal Processing, **58/59**, 128-142, 2015.

[6] J.Zhang, Y.L. Xu, Y. Xia, J. Li, A new statistical moment-based structural damage detection method. *Structural Engineering and Mechanics*, **30**, 445-466, 2008.

[7] J. Zhang, Y.L. Xu, J. Li, Y. Xia, J.C. Li, Statistical moment-based structural damage detection method in time domain. *Journal of Earthquake Engineering and Engineering Vibration*, **12**, 13-23, 2013.

[8] N. Impollonia, G. Ricciardi, Explicit solutions in the stochastic dynamics of structural systems. *Probabilistic Engineering Mechanics*, **21**, 171-181, 2006.

[9] I. Failla, N. Impollonia, G. Ricciardi, Identification of Semi-Rigid Connections from Stationary Statistical Moments. J. Kruis, Y. Tsompanakis and B.H.V. Topping eds. *15th international conference on civil, structural and environmental engineering computing*, Prague, Czech Republic, September 1-4, 2015.

[10] N. Impollonia, I. Failla, G. Ricciardi, Parametric statistical moment method for damage detection and health monitoring. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 10.1061/AJRUA6.0000863, C4016001.

[11] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1974

[12] J.W. Brewer, Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits Sys*.;1978

[13] G. Falsone, G. Muscolino, G. and Ricciardi, Combined dynamic response of primary and multiply connected cascaded secondary subsystems. *Earthquake Eng. Struc*.; **20**, 749-67, 1991.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A COHESIVE ZONE MODEL FOR THE CHARACTERISATION OF THE INTERFACIAL TRANSITION ZONE (ITZ) BETWEEN CEMENT PASTE AND AGGREGATES

**Etienne Malachanne[1], Marie Salgues[1], Mouad Jebli[2] and Frederic Jamin[2]**

[1]Ecole des Mines d'Ales
6 Avenue de Clavieres, 30319 ALES, FRANCE
e-mail: {etienne.malachanne,marie.salgues}@mines-ales.fr

[2] LMGC UMR UM-CNRS 5508, University of Montpellier
34095 MONTPELLIER CEDEX 5, FRANCE
e-mail: {mouad.jebli,frederic.jamin}@umontpellier.fr

**Keywords:** Interfacial Transition Zone, Tensile strength, Mortar, Cohesive Zone Model

**Abstract.** *Characterization of concrete behavior needs to know mechanical properties of the two phases constituting them : mortar and aggregates. Nevertheless this bi-phasic approach reaches its limits when concrete leaves the elastic domain. At that stage and according to several studies, phenomena which occur at the interface between mortar and aggregates, or in mortar between cement paste and aggregates, must be taken into account. If the occurence of an Interfacial Transition Zone (ITZ), with weak mechanical properties in regard to the two others surrounding them is well knwon, the modeling of this third phase is not settled yet. This study focus on the characterization of the adhesion at interface between cement paste and aggregates. A mortar compounded by two limestone aggregates binded by a cement paste is considered, and tensile tests are performed on a sample of this composite. Based on these experimental results, a numerical study is developed in order to see influence of interface quality in tensile strength. For thar mortar is modeled by finite elements with a cohesive zone model at the interface, substituting ITZ. With the cohesive zone model used, coupling friction and adhesion at the interface, three parameters have to be fitted : normal and tangential stiffness, and decohesion energy. A strong correlation is found between these parameters and tensile strength, but numerical results show also low values of stiffness and decohesion energy at interface. This result could be explained by a partial adhesion between mortar and cement paste in the sample.*

# 1 INTRODUCTION

When concrete is considered as a bi-phasic material represented at a mesoscopic scale, compounded by mortar and aggregates, the characterization of its mechanical behavior depends on elastic properties of the two phases, at less until the rupture [21, 16]. Nevertheless an acurate prediction when the elastic limit is reached is often investigated. For that, a damage variable may be introduced and the behavior is described by a volumic model as developed by Pijaudier-Cabot and Mazars [20]. These models allow to describe the post-peak behavior using inverse analysis methods [3], concrete being still represented as a bi-phasic material. In a different way, some studies consider a third phase, the Interfacial Transition Zone (ITZ) between mortar or cement paste and aggregates, as an independant material with weak mechanical properties [18]. In order to take into account this third phase, some numerical studies propose a finite element model representing physically ITZ [15, 10], inducing difficulties particularly in the mesh generation because of the very thin thickness of ITZ in regard to dimensions of other phases [4]. In their experimental and numerical studies, Ke et al. [12, 13] have overcome these difficulties with the development of a micromechanical model, especially for lightweight concrete, in order to predict failure. Other works consider a mechanical damage induced by the quality of the matrix-aggregates interface [14]. In the present study we focus on the caracterization of the mechanical link between cement paste and aggregates.

A non perfect adhesion between two phases, may be modeled by introduction of Cohesive Zone Models (CZM) at the interface. These models provide good description of mechanical behavior, also out of elastic domain [6]. Among various studies using CZM in their numerical models, based on finite elements modelling [17, 7, 23] or with discrete elements [1], the one used in this paper, couples adhesion, friction and unilateral contact [22, 24]. Here, cohesive forces directed at interface are both in tangential and normal directions [2]. Cohesion at interface is represented by a damage parameter $\beta$, providing the intensity of adhesion.

The work presented here, deals with a characterization of interface quality, between cement paste and limestone aggregate. For that, the cohesive zone model quoted previously is calibrated on mechanical tests performed on a composite compounded by two cubic samples of limestone linked by a thin thickness of cement paste. For that a three-dimensionnal finite elements model of the composite is carried out using the software LMGC90 [5]. Parameters of CZM law fitted will describe the mechanical properties and the adhesion quality at the interface between cement paste aggregates.

# 2 MATERIALS AND METHODS

## 2.1 Experimental sample of concrete

All experimental results come from works of PhD thesis of Mouad Jebli, from University of Montpellier, France. Composite of mortar is realized with two cubic blocks of limestone binded by a thin cement paste, as represented in the Figure 1. Elastic properties of these two materials are obtained by uniaxial compression tests and are summarized in the Table 1.

| Material | Young's modulus (GPa) | Poisson ratio |
|---|---|---|
| Limestone | 60 | 0.3 |
| Cement paste | 12.5 | 0.3 |

Table 1: Elastic properties of aggregates and cement paste

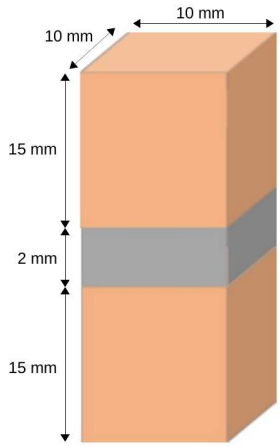



Figure 1: Sample of concrete compounded by two blocks of limestone ($10 \times 15 \times 10\,\text{mm}$) binded by a cement paste ($10 \times 2 \times 10\,\text{mm}$)

Mechanical tensile tests are carried out on the sample. A velocity of $0.01\,\text{mm}/\text{s}$ is imposed through a tensile test device, as shown by the Figure 2. Sample is loaded until rupture. The onset of cracking and rupture of the sample is recorded with use of a high-speed digital camera. For the calibration of the cohesive zone model, results on 28 days mortar composites have been taken into account.



Figure 2: Tensile test on the concrete sample

Results of uniaxial tensile test are shown on Figure 3. This curve will be used on the following to choose parameters of CZM at the interface.

## 2.2 Cohesive Zone Model

Mathematical developments of the model chosen here, using thermodynamic considerations, are detailed on several works [2, 22]. We will limit in this part to present parameters of the cohesive law, and its main constitutive equations. This model considers a non-penetration between two bodies in contact, here cement paste and aggregates. These two materials are meshed by finite elements. The unilateral contact is coupled with friction, coming from Coulomb law be-

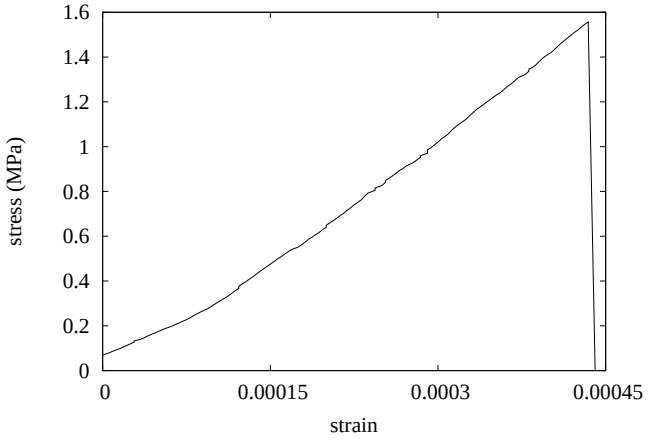



Figure 3: Stress (in MPa) vs strain for tensile test on the concrete composite

tween the two materials, and adhesion governing by a state variable which describes the contact state. This variable $\beta$, introduced by Fremond [9], giving the intensity of adhesion between the two bodies in contact, is defined with a damage parameter $D$ included between 0 (no damage) and 1 (rupture) such as :

$$\beta = 1 - D \tag{1}$$

Thus adhesion linked to damage between aggregate and cement paste is :

- $\beta = 1$ total adhesion
- $0 < \beta < 1$ partial adhesion
- $\beta = 0$ no adhesion

The thermodynamic description, based on standard material generalized theory [11], needs to define state variables associated to thermodynamical forces. For that one considers the reversible forces $R_N$ and $R_T$ associated to the dispalcements $u_N$ and $u_T$. Indexes $N$ and $T$ indicate a decomposition into normal and tengential part on the contact surface. Thus, $R_N$ and $R_T$ are considered as normal and tengential contact forces and the displacement is written :

$$[u] = [u_N]\,.n + [u_T] \quad \text{with} \quad [u_N] = [u]\,.n \tag{2}$$

In a same way, the decomposition of contact forces gives :

$$R = R_N.n + R_T \quad \text{with} \quad R_N = R.n \tag{3}$$

In this problem the state variables are $\beta$, $u_N$ and $u_T$, and we introduce a free energy potential defined such as :

$$\psi(u_N, u_T, \beta) = \frac{C_N}{2} u_N^2 \beta^2 + \frac{C_T}{2} u_T^2 \beta^2 - wb \tag{4}$$

with $C_N$ and $C_T$ the normal and tengential stiffness, $w$ the Dupre's energy and $b$ the adhesion viscosity coefficient. As explained in works of Raous et al. [22], the main constitutive laws of the contact at the interface come from the subdifferentiation of the free energy potential, and a potential of dissipation, with respect to the state variables $u_N$, $u_T$ and $\beta$. As consequence we can define the unilateral contact with adhesion :

$$\begin{aligned} R_N - C_N\,[u_N]\,\beta^2 &\geq 0 \\ [u_N] &\geq 0 \\ (R_N - C_N\,[u_N]\,\beta^2)\,[u_N] &= 0 \end{aligned} \tag{5}$$

Introducing a fricton coefficient of Coulomb $\mu$, the friction with adhesion is defined by the following equations :

$$\begin{aligned} ||R_T - C_T\,[u_T]\,\beta^2|| &\geq \mu|R_N - C_N\,[u_N]\,\beta^2| \\ ||R_T - C_T\,[u_T]\,\beta^2|| &< \mu|R_N - C_N\,[u_N]\,\beta^2| \Rightarrow [\dot{u_T}] = 0 \\ ||R_T - C_T\,[u_T]\,\beta^2|| &= \mu|R_N - C_N\,[u_N]\,\beta^2| \Rightarrow \exists\lambda > 0, [\dot{u_T}] = \lambda(R_T - C_T\,[u_T]\,\beta^2) \end{aligned} \tag{6}$$

Finally, shape of the cohesive law presented is shown by the curve of the Figure 4 giving the normal reaction with respect to normal displacement. Force-peak represents the limit of adhesion. Indeed at that stage damage occurs at interface, and energy of adhesion, considered as the area under the cruve decreases.

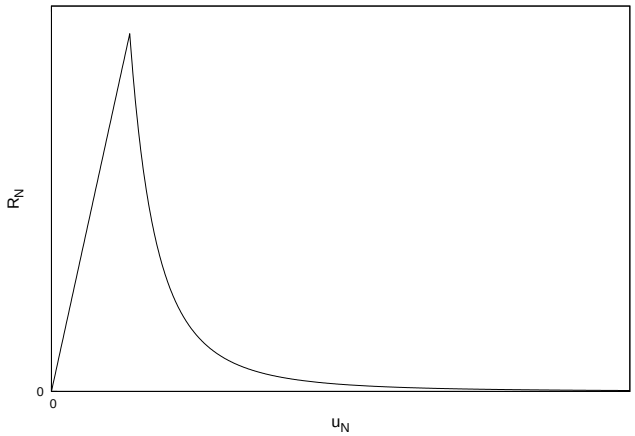



Figure 4: Shape of the cohesive law

## 2.3 Numerical computation

All the numerical computations have been proceeded with the open software LMGC90 [5], developed at the University of Montpellier, France, in which the cohesive zone model presented in past section is implemented. A tree-dimensional finite element model of the composite sample has been carried out. A symmetrical geometry has been chosen, as shown by Figure 5, in order to reduce computation time, thus conditions of symmetry will be respected. Geometry is

meshed with hexaedron finite elements H8, and the time step is taken constant and equals to $1 \times 10^{-4}$ s. A theta-method is chosen for time intregation with $\theta = 0.55$. Boundary conditions are applied on the lower and upper faces, with a velocity $v_y = 0.01\,\mathrm{mm/s}$ for the tensile load on the upper, and vertical displacements blocked on the lower due to symmetry.

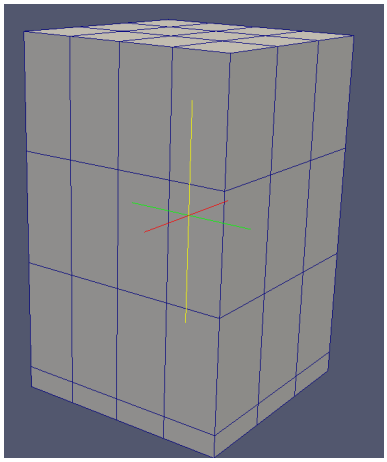


Figure 5: Meshing of a half composite sample by hexaedron finite elements

Application of the cohesive zone between cement paste and aggregate, is based on contact nodes placed in the contact surface both on cement paste and aggregates. At each time step a detection is realised and rupture occurs when contact is lost between the two bodies. To calibrate cohesive law, several parameters have to be setted : normal and tangential stiffness $C_N$ and $C_T$, friction coefficient $\mu$, adhesion energy $w$ and adhesion viscosity $b$. In this model we consider any viscosity at the interface, thus $b = 0$. Friction coefficient is determinated by experimental study, and $\mu = 0.4$. We assume that adhesion energy, is defined by the experimental curve stress-strain (Figure 3), and calculated as the aera under the curve, until the rupture. Hence :

$$w = \frac{1}{2}(1.6 \times 10^6 \times 0.015 \times 10^{-3}) = 12\,\mathrm{J/m^2} \tag{7}$$

Hence, adhesion energy calculated here is only the one for the elastic part until rupture, thus according to the Figure 4 presented in the past section we assume that the value corresponding in the cohesive law is near the double, $24\,\mathrm{J/m^2}$. We will see in the following if this assumption gives results expected.

Finally, only $C_N$ and $C_T$ have to be fitted with the experimental results. We assume that $C_N$ = $C_T$, hence their value will be calibrated with several numerical tensile test on the composite sample.

## 3 RESULTS

Numerical computation is performed until the peak-stress, and compared with the experimental stress-strain curve, the objective being to tend to this curve. On the finite element model, we observe that rupture is represented by a decohesion between the two bodies, as shown by the Figure 6.

As explained previously, in this study two parameters have to be fitted : $C_N$, $C_T$. It appears that in our case, $C_N$ and $C_T$ give the slope of the linear par and the $w$ sets the position of

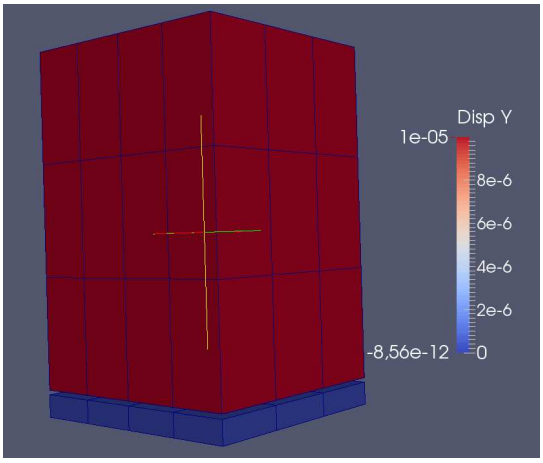



Figure 6: Displacement of the two materials after rupture

peak stress along the slope. Influence on these three parameters on the mechanical response are represented of the Figure 7. Value of decohesion energy estimated previously, has to be slightly reduced to bring numerical peak stress closer to the experimental one.

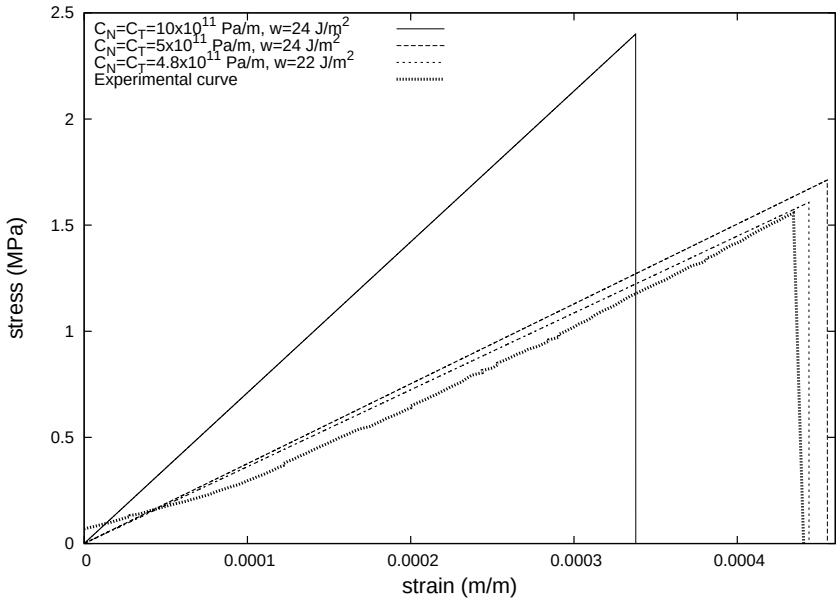



Figure 7: Influence of parameters $C_N$, $C_T$ and $w$ on the mechanical response

Thus, we can observe on Figure 7 that paramaters of cohesive zone model which allow to tend to experimental curve are $C_N = C_T = 4.8 \times 10^{11}\,\mathrm{Pa/m}$ and $w = 22\,\mathrm{J/m^2}$.

## 4 DISCUSSION

We can first notice a strong correlation between mechanical parameters defined at the interface between cement paste and aggregates, and both strain and tensile strength. Moreover numerical results show that with this cohesive zone model, it is possible to tend to experimental results. We can observe that value of decohesion calculated experimentaly, gives a good estima-

tion of the peak stress. The slighly decrease is explained by the interface considered as perfect in the numerical sample which is not the case in the experimental one. As consequence with $w$ fixed by the experiment, only 2 parameters have to be fitted.

Since decohesion occurs only in the ITZ, the latter is substituted by cohesive zones. Thus parameters fitted can be considered as characteristics of the ITZ. A comparison with other studies realised with the same cohesive law, give values of interfacial stiffness higher than the one obtained here. Indeed works of Perales et al. [19], have shown that $C_N = C_T = 2 \times 10^{18}\,\mathrm{Pa/m}$ when studies of Fouchal et al. [8] found $1.5 \times 10^{17}\,\mathrm{Pa/m}$. These values depend on material modeled, hydrided Zircaloy for the first and interface between bricks and mortar for the second. In the same way, our values of decohesion energy are also weak, particularly comparing to adhesion between bricks and mortar ($0.9\,\mathrm{J/mm^2}$). Two ways may be investigated to explain these low values of interface parameters : parameters identified with numerical computation are not representatives of the real interface properties, or an initial damage occurs at the interface, due to the composite sample fabrication. If the first hypothesis requires other experimental investigations, in particular on a mini-structure of mortar, let us focus on the second through an analysis of Young modulus.

Experimental Young modulus of the composite sample, could be calculated as being the slope of the curve of the Figure 3. Indeed, assuming that until tensile strength, in the linear part, the sample has an elastic behavior, Young modulus is obtained using the Hooke law for an uniaxial tensile test :

$$E_{sample} = \frac{\sigma}{\varepsilon} \tag{8}$$

with $\sigma$ the stress and $\varepsilon$ the strain. Here, according to the Figure 3, $E_{sample} = 3.43\,\mathrm{GPa}$. This value has to be compared with the real Young modulus of the sample, which could be calculated through a serial stiffness. Indeed if we consider that the sample is compounded by three serial elastic bodies, the equivalent stiffness is defined by :

$$k_{sample} = \frac{1}{\frac{2}{k_{agg}} + \frac{1}{k_p}} \tag{9}$$

where $k_{agg}$ and $k_p$ are aggregate and paste stiffness defined such as :

$$k_{agg,p} = \frac{S_{agg,p} \times E_{agg,p}}{L_{agg,p}} \tag{10}$$

with $S_{agg} = S_p = 100\,\mathrm{mm^2}$ the section of aggregate and cement paste, $L_{agg} = 15\,\mathrm{mm}$ the length of each aggregate and $L_p = 2\,\mathrm{mm}$ the length of the paste. Thus using these values, composite stiffness is $k_{sample} = 151515\,\mathrm{N/mm}$ leading to the theoretical equivalent Young modulus of the sample :

$$E_{th} = \frac{32 \times 151515}{100} = 48.5\,\mathrm{GPa} \tag{11}$$

A noticeable difference is observed between theoretical Young modulus and the one obtained experimentally. This result may be explained by a damage at the interface, due to the interface

quality between aggregate and cement paste, which lead to a fall on the equivalent Young modulus. According to the definition of the damage variable given by Mazars et al. [20], theoretical and experimental Young modulus are binded by the relationship :

$$E_{sample} = (1 - D)E_{th} \tag{12}$$

where the damage variable $D$ is different to zero, inducing an intensity of adhesion $\beta$ less than $1$ at the begining of the computation, as explained by equation (2.2). This observation could indicate a partial adhesion between aggregate and cement paste and explain low values of stiffness and decohesion energy at the interface. Thus the strong correlation between stiffness and adhesion energy at interface, and mechanical behavior of the composite, leads to confirm that tensile rupture is in this case governing by the adhesion quality.

## 5 CONCLUSION

Cohesive zone model applied to the numerical computation studied here, allows to calibrate decohesion between cement paste and limestone aggregate, using experimental tensile test on a composite sample compounded by both materials. It is possible to reach experimental stress-strain curve, with values of stiffness at the interface and decohesion energy according to literature. Moreover Young modulus obtained by experimental works and numerical computation, calculating the slope of the stress strain curve, is very small against the theoretical one. Other experimental studies have to be performed to confirm the existence of initial damage at the interface.

## REFERENCES

[1] R. de Borst, J. J.C. Remmers, and A. Needleman. Mesh-independent discrete numerical representations of cohesive-zone models. *Engineering Fracture Mechanics*, 73(2):160–177, Jan 2006.

[2] A.-S. Bretelle, M. Cocou, and Y. Monerie. Unilateral contact with adhesion and friction between two hyperelastic bodies. *International Journal of Engineering Science*, 39(18):2015–2032, Dec 2001.

[3] I. Comby-Peyrot, P.-O. Bouchard, F. Bay, and F. Bernard. Numerical aspects of a problem with damage to simulate mechanical behavior of a quasi-brittle material. *International Journal of Computational Materials Science and Surface Engineering*, 40(3):327–340, 2007.

[4] S. Diamond and J. Huang. The itz in concrete a different view based on image analysis and sem observations. *Cement and Concrete Composites*, 23(2-3):179–188, Apr 2001.

[5] F. Dubois, M. Jean, M. Renouf, R. Mozul, A. Martin, and M. Bagneris. Lmgc90. In *Colloque National en Calcul des Structures*. CSMA, 2011.

[6] M. Elices, G.V. Guinea, J. Gmez, and J. Planas. The cohesive zone model: advantages, limitations and challenges. *Engineering Fracture Mechanics*, 69(2):137–163, Jan 2002.

[7] M. Elices, J. Planas, and G.V. Guinea. Fracture mechanics applied to concrete. *European Structural Integrity Society*, pages 183–210, 2000.

[8] F. Fouchal, F. Lebon, and I. Titeux. Contribution to the modelling of interfaces in masonry construction. *Construction and Building Materials*, 23(6):2428–2441, Jun 2009.

[9] M. Fremond. Equilibre des structures qui adherent a leur support. *Comptes Rendus de l'Academie des Sciences, Paris*, 295 (Serie II):913–916, 1982.

[10] F. Grondin and M. Matallah. How to consider the interfacial transition zones in the finite element modelling of concrete? *Cement and Concrete Research*, 58:67–75, Apr 2014.

[11] B. Halphen and Q. S. Nguyen. Plastic and visco-plastic materials with generalized potential. *Mechanics Research Communications*, 1(1):43–47, Jan 1974.

[12] Y. Ke, S. Ortola, A.L. Beaucour, and H. Dumontet. Identification of microstructural characteristics in lightweight aggregate concretes by micromechanical modelling including the interfacial transition zone (itz). *Cement and Concrete Research*, 40(11):1590–1600, Nov 2010.

[13] Y. Ke, S. Ortola, A.L. Beaucour, and H. Dumontet. Micro-stress analysis and identification of lightweight aggregates failure strength by micromechanical modeling. *Mechanics of Materials*, 68:176–192, Jan 2014.

[14] D. Keinde, S. Kamali-Bernard, F. Bernard, and I. Cisse. Effect of the interfacial transition zone and the nature of the matrix-aggregate interface on the overall elastic and inelastic behavior of concrete under compression : a 3d numerical study. *European Journal of Environmental and Civil Engineering*, 18(10):1167–1176, 2014.

[15] K.M. Lee and J.H. Park. A numerical model for elastic modulus of concrete considering interfacial transition zone. *Cement and Concrete Research*, 38(3):396–402, Mar 2008.

[16] E. Malachanne, R. Sassine, and E. Garcia-Diaz. Numerical model for mechanical behavior of lightweight concrete and for the prediction of local stress concentration. *Construction and Building Materials*, 59:180–187, 2014.

[17] F. Moroni and A. Pirondi. Cohesive zone model simulation of fatigue debonding along interfaces. *Procedia Engineering*, 10:1829–1834, 2011.

[18] C.M. Neubauer, H.M. Jennings, and E.J. Garboczi. A three-phase model of the elastic and shrinkage properties of mortar. *Advanced Cement Based Materials*, 4(1):6–20, Jul 1996.

[19] F Perales, S Bourgeois, A Chrysochoos, and Y Monerie. Two field multibody method for periodic homogenization in fracture mechanics of nonlinear heterogeneous materials. *Engineering Fracture Mechanics*, 75(11):3378–3398, 2008.

[20] G. Pijaudier-Cabot and J. Mazars. Damage models for concrete. *Handbook of Materials Behavior Models : Failures of Materials*, 2:500–512, 2001.

[21] R. R. Prasada, J.O. Olekb, and S. Diamondb. An investigation into the influence of inter-aggregate spacing and the extent of the itz on properties of portland cement concretes. *Cement and Concrete Research*, 40:1601–1608, 2010.

[22] M. Raous, L. Cangemi, and M. Cocu. A consistent model coupling adhesion, friction, and unilateral contact. *Computer Methods in Applied Mechanics and Engineering*, 177(3-4):383–399, 1999.

[23] J. Roesler, Glaucio H. Paulino, and C. Park, K.and Gaedicke. Concrete fracture prediction using bilinear softening. *Cement and Concrete Composites*, 29(4):300–312, Apr 2007.

[24] C. Talon and A. Curnier. A model of adhesion coupled to contact and friction. *European Journal of Mechanics - A/Solids*, 22(4):545–565, Aug 2003.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EXACT SHEAR STRAIN APPROACH FOR RC FRAME ELEMENTS WITH AXIAL-SHEAR INTERACTION

**Alexander Kagermanov**[1], **Paola Ceresa**[2]

[1] UME School, Institute for Advanced Studies (IUSS)
Piazza della Vittoria, 15, 27100, Pavia, Italy
alexander.kagermanov@umeschool.it

[2] Institute for Advanced Studies (IUSS)
Piazza della Vittoria, 15, 27100, Pavia, Italy
paola.ceresa@iusspavia.it

**Keywords:** reinforced concrete, finite element analysis, frame elements, axial-shear interaction.

**Abstract.** *A fiber-section frame element for nonlinear analysis of two-dimensional, shear-critical reinforced concrete structures is developed. The element incorporates a new sectional algorithm for the computation of an exact shear strain profile and corresponding shear stress distribution. Longitudinal axial strains are evaluated based on the plane section assumption, while transverse strains are determined from equilibrium in the vertical direction, following classical beam assumptions. Axial-shear interaction at the fiber level is based on a smeared-crack, orthotropic constitutive model, which uses equivalent uniaxial material models for concrete and steel in the fixed-crack and reinforcement directions, respectively. Shear strain components at the crack, arising from deviations between principal and crack directions, are related to shear stresses by means of a shear stiffness term that fully satisfies compatibility and equilibrium conditions. The procedure was implemented into a force-based fiber element and applied to monotonic and cyclic analysis of shear-critical reinforced concrete structures. Validation examples include a number of experiments on lightly reinforced beams, shear-critical frames and columns.*

## 1 INTRODUCTION

Extensive experimental and analytical effort has been devoted to investigate the response of RC members under combined axial and shear stresses in the last 50 years. This allowed developing refined finite element models capable of predicting the global force-displacement response up to collapse, as well as local response parameters such as strains, crack widths and crack patterns. However, their general application to framed structures results cumbersome and time consuming, hence substantial effort is being addressed towards improving existing frame elements, usually based on the plane section assumption and uniaxial constitutive laws, in order to account for multi-axial loading conditions while keeping an efficient and simple formulation.

It can be said that a promising, fiber-based frame element for axial-shear interaction requires (i) an accurate multi-axial constitutive model and (ii) plausible section and beam kinematic assumptions. Extensive literature can be found concerning each of these topics. For instance, [1] introduced axial-shear interaction at the fiber level based on a dual-section analysis, whereby shear stresses over the section were calculated from the difference between axial stresses at two consecutive sections separated h/6 (h: depth of the member). [2] suggested a local procedure such that shear stresses were obtained from longitudinal gradients of deformations at a given section. Both authors used two-dimensional constitutive models, namely the Modified Compression Field Theory (MCFT), assuming that plane sections remain plane.

Several authors used Bernoulli and Timoshenko beam theories in both displacement- and force-based frame elements with axial-shear interaction [3]. The Bernoulli beam theory does not directly account for shear deformations, hence shear effects were usually introduced with an equivalent truss model, resulting only in partial coupling with axial forces and bending moments [4,5,6]. The Timoshenko beam theory provides a kinematic link between shear strains and transverse displacements of the beam axis. Hence, multi-axial constitutive models can be used at each individual fiber resulting in full axial-shear coupling [7, 8, 9, 10, 11, 12, 13]. However, the theory predicts a constant shear strain profile over the section, thus violating inter-fiber equilibrium. More recently, several higher-order theories were proposed dealing with the latter shortcomings [14, 15, 16, 17, 18]. Their application to nonlinear analysis of RC members is still in development. Moreover, their practical implementation for inelastic dynamic analysis needs further research due to complexities associated with their formulation and high computational demands.

The present work explores the possibility of an exact shear strain profile at the section level, while maintaining the plane section assumption for longitudinal strains. The rationale behind this approach is that [2], among others, demonstrated that axial-shear interaction in beams and columns can be relatively well captured assuming plane sections remain plane, whereas refined kinematic beam theories have not yet been fully verified. The exact shear strain profile is obtained from an averaged form of inter-fiber equilibrium over the section, using linear interpolation functions. The formulation was implemented into a force-based, fiber-element with smeared-crack, fixed-crack cyclic constitutive models. Verification examples on shear critical beams, frames and columns are provided at the end.

## 2 BEAM KINEMATICS

The Timoshenko beam theory is adopted to define the element kinematics. The total displacement of a given section point with coordinates (x, y) is given as:

$$u(x,y) = u_o(x) - \theta_z(x)y \tag{1}$$

$$v(x, y) = v_o(x) \tag{2}$$

where $u_o$ is the horizontal displacement, $\theta_z$ the section rotation and $v_o$ the vertical displacement of the beam axis, x the longitudinal coordinate along the beam axis, and y the vertical coordinate. Assuming small strain-displacement compatibility, fiber strains are defined from the linear Green-Lagrange strain tensor as:

$$\varepsilon_x = \frac{\partial u}{\partial x} = \frac{\partial u_o(x)}{\partial x} - \frac{\partial \theta_z(x)}{\partial x} y = \varepsilon_o - \chi_z y \tag{3}$$

$$\varepsilon_y = \frac{\partial v}{\partial y} = 0 \tag{4}$$

$$\gamma_{xy} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = \frac{\partial v_o(x)}{\partial x} - \theta_z(x) = \gamma_{oy} \tag{5}$$

where $\varepsilon_x$ is the axial strain, $\varepsilon_y$ the transverse strain and $\gamma_{xy}$ the shear strain acting on a section layer with coordinate y. $\varepsilon_o$ is the section axial strain, $\chi_z$ the section curvature and $\gamma_{oy}$ section shear deformation. The relationship between strains and section deformations can be expressed as:

$$\begin{pmatrix} \varepsilon_x \\ \gamma_{xy} \end{pmatrix} = \begin{bmatrix} 1 & -y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \varepsilon_o \\ \chi_z \\ \gamma_{oy} \end{pmatrix} \tag{6}$$

$$\boldsymbol{\varepsilon} = \mathbf{B}\mathbf{e} \tag{7}$$

Section forces are obtained from direct integration of stress distributions over the section:

$$\mathbf{S} = \int_A \mathbf{B}^T \boldsymbol{\sigma} dA \tag{8}$$

where $\mathbf{S}=[N,M,V]^T$ and $\boldsymbol{\sigma}=[\sigma_x, \tau_{xy}]^T$ are the vector of section forces and stresses, respectively. The tangent stiffness matrix of the section is obtained taking variation of section forces with respect to section deformations:

$$\mathbf{K}_s = \frac{\partial \mathbf{S}}{\partial \mathbf{e}} = \frac{\partial}{\partial \mathbf{e}} \int_A \mathbf{B}^T \boldsymbol{\sigma} dA = \int_A \mathbf{B}^T \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\varepsilon}} \frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{e}} dA = \int_A \mathbf{B}^T \frac{\partial \boldsymbol{\sigma}}{\partial \boldsymbol{\varepsilon}} \mathbf{B} dA = \int_A \mathbf{B}^T \bar{D} \mathbf{B} dA \tag{9}$$

where $\bar{D}$ is the condensed, 2×2 material stiffness matrix, given as:

$$\begin{pmatrix} \partial\sigma_x \\ \partial\tau_{xy} \end{pmatrix} = \begin{bmatrix} \bar{D}_{11} & \bar{D}_{12} \\ \bar{D}_{21} & \bar{D}_{22} \end{bmatrix} \begin{pmatrix} \partial\varepsilon_x \\ \partial\gamma_{xy} \end{pmatrix} \tag{10}$$

Condensation is carried out upon neglecting the effect of clamping forces, that is, assuming that $\sigma_y=0$, which is a reasonable approximation in regions away from local disturbance, such as supports and points of load application. In this case, the increment of transverse strains is given as:

$$\partial\varepsilon_y = \frac{(-D_{21}\partial\varepsilon_x - D_{23}\partial\gamma_{xy})}{D_{22}} \tag{11}$$

where $D_{21}$, $D_{22}$ and $D_{23}$ are coefficients of the 3×3 material stiffness matrix relating ($\partial\sigma_x$, $\partial\sigma_y$, $\partial\tau_{xy}$) and ($\partial\varepsilon_x$, $\partial\varepsilon_y$, $\partial\gamma_{xy}$), obtained from the two-dimensional constitutive model presented in the next sections.

## 3 SHEAR STRAIN PROFILE

The section model is based on the evaluation of an exact shear strain profile using the inter-fiber equilibrium equation. This can be derived from internal beam equilibrium in the longitudinal direction:

$$-\sigma_x dy - \tau_{xy} dx + \sigma_x dy + \frac{\partial \sigma_x}{\partial x} dx dy + \tau_{xy} dx + \frac{\partial \tau_{xy}}{\partial y} dy dx = 0 \tag{12}$$

$$\frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} = 0 \tag{13}$$

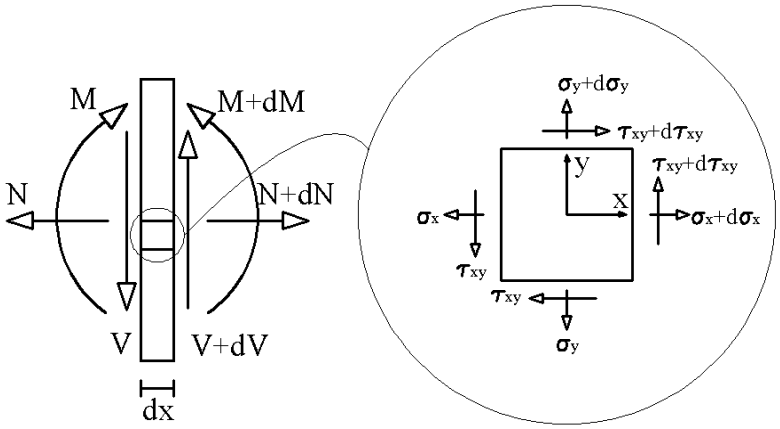



Figure 1: Internal equilibrium in a cross-section plane stress element.

Using the constitutive relationship in equation (10) in conjunction with equation (13) yields:

$$\overline{D}_{11}\frac{\partial \varepsilon_x}{\partial x} + \overline{D}_{12}\frac{\partial \gamma_{xy}}{\partial x} + \overline{D}_{21}\frac{\partial \varepsilon_x}{\partial y} + \overline{D}_{22}\frac{\partial \gamma_{xy}}{\partial y} = 0 \tag{14}$$

The first two terms in (14) contain strain derivatives in the longitudinal direction. These can be obtained from the derivative of section forces and compatibility between fiber and section strains:

$$\frac{d\mathbf{e}}{dx} = \mathbf{K}_s^{-1}\frac{d\mathbf{S}}{dx} \tag{15}$$

$$\frac{\partial \boldsymbol{\varepsilon}}{\partial x} = \mathbf{B}\frac{d\mathbf{e}}{dx} \tag{16}$$

where $d\mathbf{S}/dx=[0,-V,0]^T$ for an element under constant axial load and shear, and linear variation of bending moments. The term $\partial\varepsilon_x/\partial y$ in (14) is equal to $-\chi_z$ from (3), and $\partial\gamma_{xy}/\partial y$ is obtained from interpolation of $\gamma_{xy}$ using linear shape functions such that:

$$\gamma_{xy} = \mathbf{N}\boldsymbol{\gamma} = N_i\gamma_i + N_j\gamma_j \tag{17}$$

$$\mathbf{N} = \begin{pmatrix} N_i & N_j \end{pmatrix} = \left( \frac{y_j - y}{y_j - y_i} \quad \frac{y - y_i}{y_j - y_i} \right) \tag{18}$$

where $\gamma_i, \gamma_j$ are nodal values of shear strain at the ends of the $l^{th}$ layer, and $y_i$ and $y_j$ the corresponding coordinates (Figure 2).

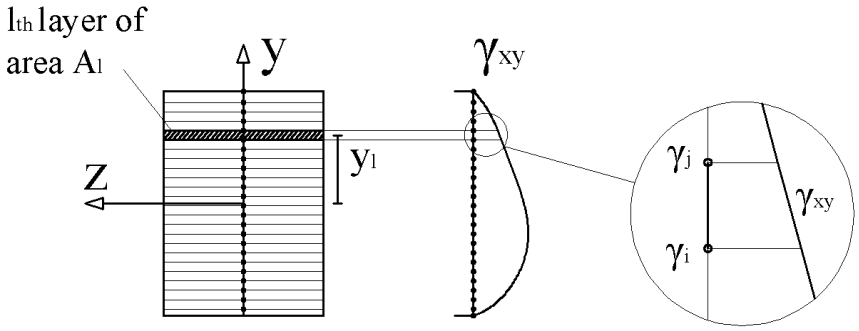



Figure 2: Interpolation of the shear strain profile.

Inter-fiber equilibrium in equation (13) is invoked in a weighted average form over the cross section, using as a weighting function the derivative of the shear strain, $\partial\gamma_{xy}/\partial y$. Upon accounting for the contribution of each individual layer, this is expressed in the following functional form:

$$\mathop{\mathbf{A}}_{l=1}^{nl} \int_{Al} \delta\gamma'^{T}_{xy} \left( \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} \right) dA = 0 \tag{19}$$

where A is the assemblage operator over the section layers, nl number of layers, $A_l$ the area of the $l^{th}$ layer, and $\delta\gamma_{xy}'$ the virtual derivative of the shear strain. Substituting equations (14) and (17) into (19) yields:

$$\mathop{\mathbf{A}}_{l=1}^{nl} \int_{Al} \delta\boldsymbol{\gamma}^{T} \left( \mathbf{N}'^{T} C + \mathbf{N}'^{T} \overline{D}_{22} \mathbf{N}' \boldsymbol{\gamma} \right) dA = 0 \tag{20}$$

$$C = \overline{D}_{11} \frac{\partial \varepsilon_x}{\partial x} + \overline{D}_{12} \frac{\partial \gamma_{xy}}{\partial x} + \overline{D}_{21} \frac{\partial \varepsilon_x}{\partial y} \tag{21}$$

Since $\delta\boldsymbol{\gamma}$, containing the virtual nodal shear strains, is non-zero, the vector of nodal shear strains $\boldsymbol{\gamma}$ can be obtained from solving:

$$\boldsymbol{\gamma} = \mathbf{k}^{-1}\mathbf{f} \tag{22}$$

where:

$$\mathbf{k} = \mathop{\mathbf{A}}_{l=1}^{nl} \int_{Al} \mathbf{N}'^{T} \overline{D}_{22} \mathbf{N}' dA \tag{23}$$

$$\mathbf{f} = -\mathop{\mathbf{A}}_{l=1}^{nl} \int_{Al} \mathbf{N}'^{T} C dA \tag{24}$$

# 4 FINITE ELEMENT IMPLEMENTATION

## 4.1 Force-based formulation

The proposed section model was implemented into a two-dimensional, force-based fiber element according to [19]. In this formulation, once the global displacements are known, an iterative state determination procedure is invoked at the element level in order to find the corresponding section strains at each integration point. Because section forces are known at the start of the state determination procedure, from the exact interpolation of element forces, and section deformations can be estimated using the current section flexibility matrix, all the ingredients for the calculation of the "exact" shear strain profile are available. Thus, the three total strain components ($\varepsilon_x$, $\varepsilon_y$, $\gamma_{xy}$) at each individual fiber are known. The corresponding stresses, ($\sigma_x$, $\sigma_y$, $\tau_{xy}$), are obtained from the constitutive model and integrated to yield the section forces. In subsequent element iterations, the initial shear strain profile is corrected based on the unbalanced vector of section forces until satisfying section equilibrium. Further details on the numerical implementation are given elsewhere [20].

## 4.2 Constitutive model

The constitutive model is based on a smeared-crack, single fixed-crack membrane model [21]. For concrete, equivalent uniaxial constitutive relationships are applied in the fixed-crack directions (figure 3), while shear components arising from deviation between principle directions and crack directions are related through the secant shear stiffness term given as:

$$G = \frac{(\sigma_1 - \sigma_2)(\sin\Delta\theta_\sigma \cos\Delta\theta_\sigma)}{2(\varepsilon_1 - \varepsilon_2)(\sin\Delta\theta_\varepsilon \cos\Delta\theta_\varepsilon)} \tag{25}$$

where $\sigma_1$, $\sigma_2$, and $\varepsilon_1$, $\varepsilon_2$ are the principal stresses and strains, respectively, and $\Delta\theta_\sigma$, $\Delta\theta_\varepsilon$ the corresponding angles between principle and crack directions.

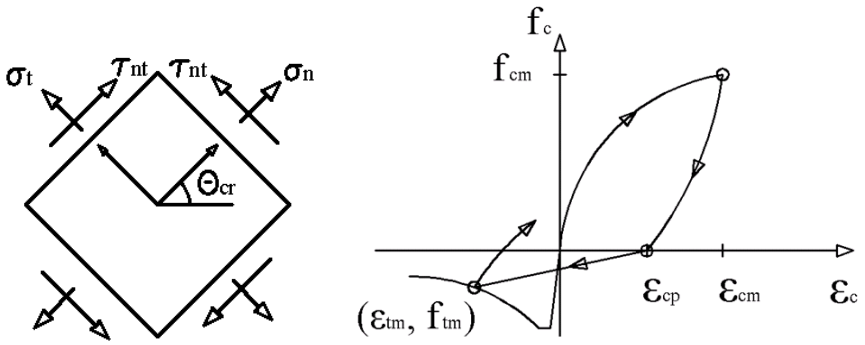



Figure 3: Fixed-crack reference system and cyclic constitutive model.

The cyclic response in tension is obtained from a physically-based approach, whereby tension stiffening, crack-opening and crack-closing branches are computed from average equilibrium on a cracked element of reinforced concrete (figure 4). Further details can be found in [21].

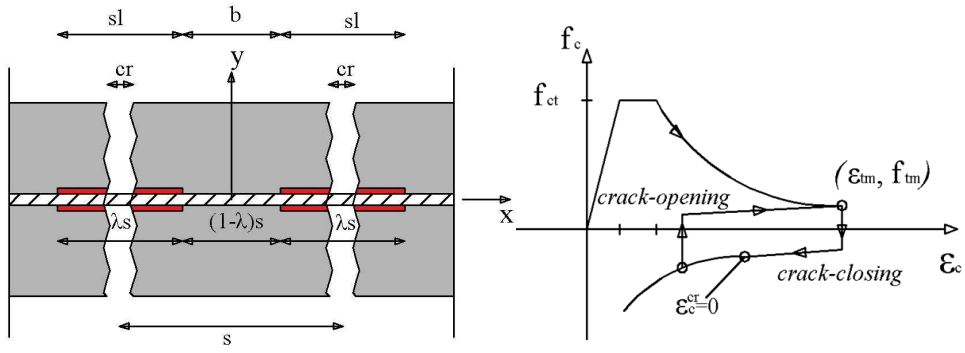



Figure 4: Cracked element of reinforced concrete and cyclic response in tension.

## 5 NUMERICAL VERIFICATION

### 5.1 Shear critical beams.

Beams OA1 and A1, tested by Bresler and Scordelis 1963 [22], were two simply supported beams with 305mm x 552mm rectangular cross sections and 3660mm clear span. Beam A1 presented 6.4mm diameter stirrups spaced at 210mm, whereas beam OA1 was unreinforced in shear and only presented bottom reinforcement (Table 2). Beam OA1 failed in diagonal-tension along the shear-critical crack with accompanying longitudinal splitting in the bottom flexural reinforcement. Beam A1 failed in shear-flexure, involving the formation of diagonal shear cracks and concrete crushing in the compression zone next to the loading plate [25].

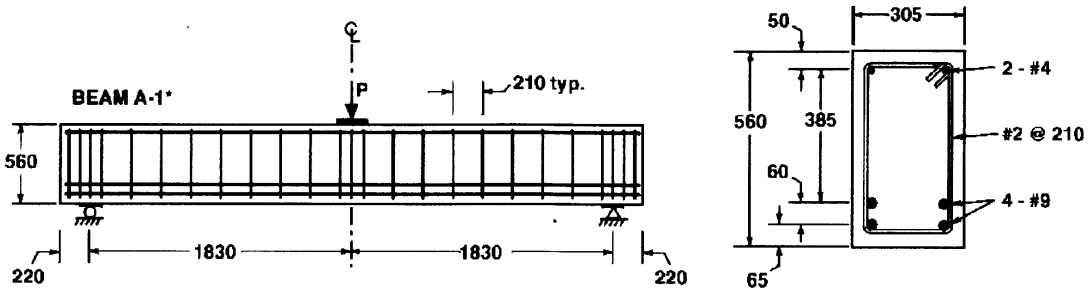



Figure 5: Beam test set up and cross section [22].

| Beam | OA1 | A1 | Steel | Diameter(mm) | $f_y$(MPa) | $f_u$(MPa) | $E_s$(MPa) |
|---|---|---|---|---|---|---|---|
| Compressive strength(MPa) | 22.6 | 24.1 | Nº2 | 6.4 | 325 | 3660 | 190000 |
| Peak Strain | 0.0018 | 0.0018 | Nº4 | 12.7 | 127 | 3660 | 201000 |
| | | | Nº9 | 28.7 | 645 | 3660 | 218000 |

Table 1: Beam material properties.

| Beam | Width(mm) | Height (mm) | Effective depth(mm) | Length(mm) | Span(mm) | Bottom Steel | Top Steel | Stirrups |
|---|---|---|---|---|---|---|---|---|
| OA1 | 310 | 556 | 461 | 4100 | 3660 | 4Nº9 | - | - |
| A1 | 307 | 561 | 466 | 4100 | 3660 | 4Nº9 | 2Nº4 | Nº2@210 |

Table 2: Beam geometric properties.

The beams were modeled with one single frame element with 5 integration points (IPs) (Gauss-Legendre integration rule), representing half of the beam due to symmetry conditions. The cross section was discretized into 21 layers of equal thickness. The influence of longitudinal reinforcement was smeared over the layers located within a distance of $7.5d_b$ ($d_b$: bar diameter) at both sides of the reinforcement. The analysis was performed under displacement control with 1mm displacement increments. Figure 6 compares the analytical and experimental responses. Also, results from FE analysis reported in [25], performed with 15×46 constant-strain, membrane finite elements based on a rotating-crack approach, are included.

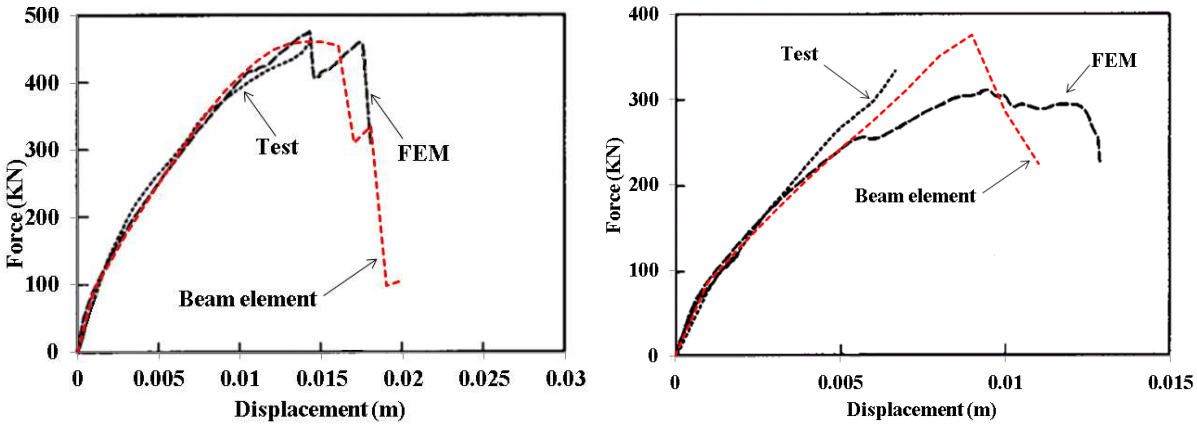



Figure 6: Force-displacement response of beams A1 (left) and OA1 (right) [25]

Figure 7 and figure 8 show calculated shear strains, shear stresses and normal concrete stresses over the depth of the section at the closest to the applied load IP, at three different displacement levels. It can be observed that the parabolic shear strain assumption approximately holds at early load stages, before significant crack-induced anisotropic behavior develops. In later stages, shear strains in the compression zone tend to decrease and concentrate in the lower portion of the web. Note that for beam A1 concrete crushing in the top compression zone occurs, suggesting that the beam ultimately failed in combined shear-compression mode, whereas for beam OA1 normal concrete stresses remain below the compressive strength. These results are consistent with the experimental observations, as previously described.

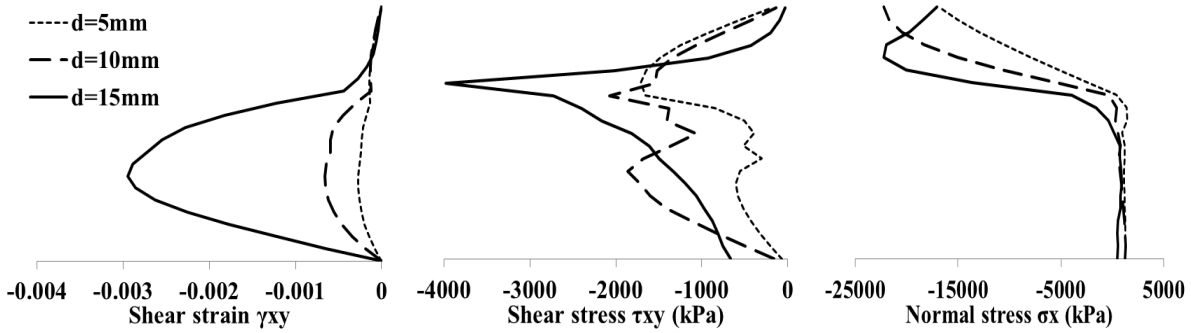



Figure 7: Shear strain, shear stress and normal concrete stress profiles for beam A1.

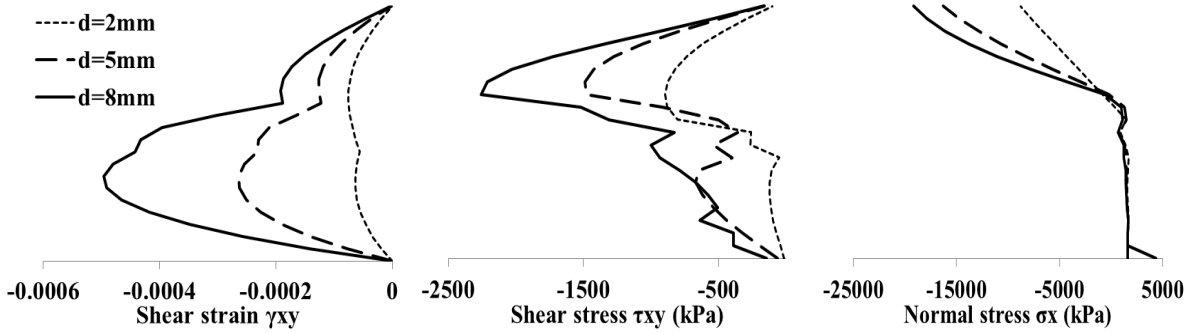



Figure 8: Shear strain, shear stress and normal concrete stress profiles for beam OA1.

## 5.2 Shear critical frame

The RC frame tested by Duong et al. 2007 [23] failed in brittle mode due to shear failure of the beams. It was a one-bay, two-story frame with an applied vertical load of 420 kN at the top of each column, subjected to an increasing lateral force at the second level (figure 9). Material properties for steel and concrete are summarized in table 3. The frame was pushed in one direction, until substantial diagonal cracking of the beams occurred, and then in the opposite direction until completing one full cycle.

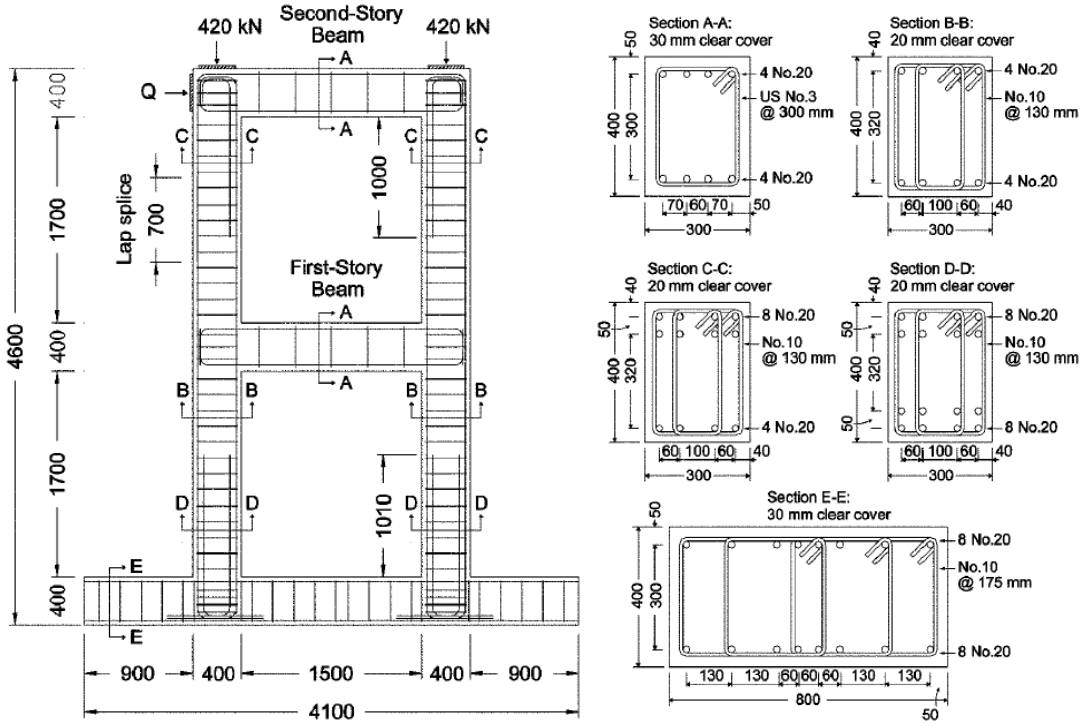



Figure 9: Geometric properties of the frame [23]

| Concrete | | |
|---|---|---|
| Compressive strength(MPa) | Peak Strain | Young Modulus(MPa) |
| 42.9 | 0.0023 | 30058 |

| Steel | | | | | | |
|---|---|---|---|---|---|---|
| Bar Size | Diameter(mm) | Area($mm^2$) | $f_y$(MPa) | $f_u$(MPa) | $E_s$(MPa) | $\varepsilon_u$‰ |
| φ20 | 19.5 | 300 | 447 | 603 | 198400 | 130.8 |
| φ 10 | 11.3 | 100 | 455 | 583 | 192400 | 129.9 |
| US #3 | 11.3 | 71 | 506 | 615 | 210000 | 134.6 |

Table 3: Frame material properties.

Each column was modeled with two frame elements per story in order to consider the increase in longitudinal reinforcement ratios at the ends of the columns. One single element was used for each beam, being the number of integration points 5/element, for both beam and column elements. The cross section was discretized into 11 layers of equal thickness. Centerline dimensions were used to define the geometry of the frame, which was assumed rigidly fixed at the foundation level. No particular modeling of the joint flexibility was undertaken. Comparison between analytical and experimental lateral force-top displacement responses is shown in figure 10, for both monotonic and cyclic load cases. The monotonic analysis predicts an ultimate displacement of about 47mm, very close to the unloading point (45mm) at which substantial shear cracking of beams was observed during the test.

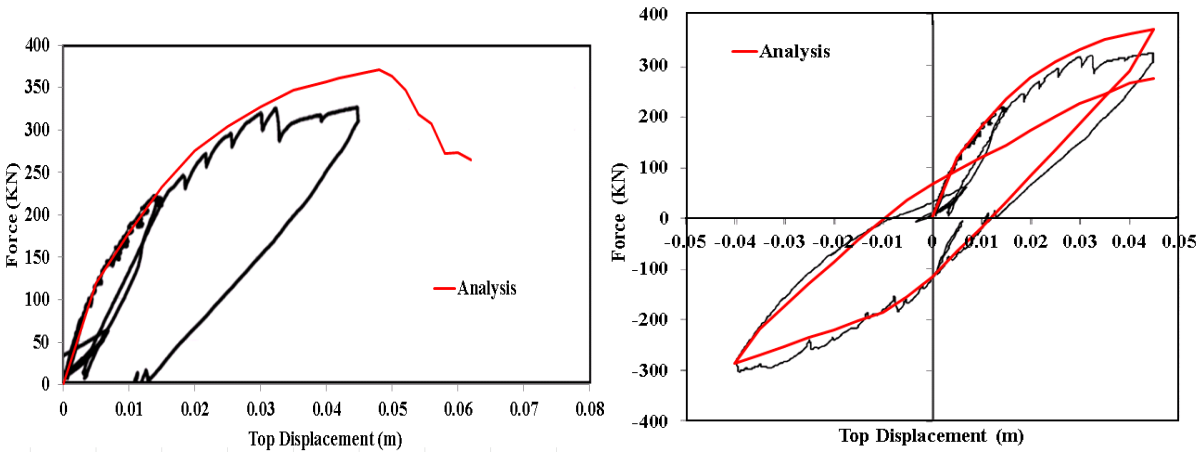



Figure 10: Force-displacement response for the monotonic (left) and cyclic (right) case.

The calculated shear strains, shear stresses and normal concrete stresses over the section are depicted in figure 11 for the first story beam, at the IP closest to the right column. Note again the reduction in shear strains in the compressive zone.

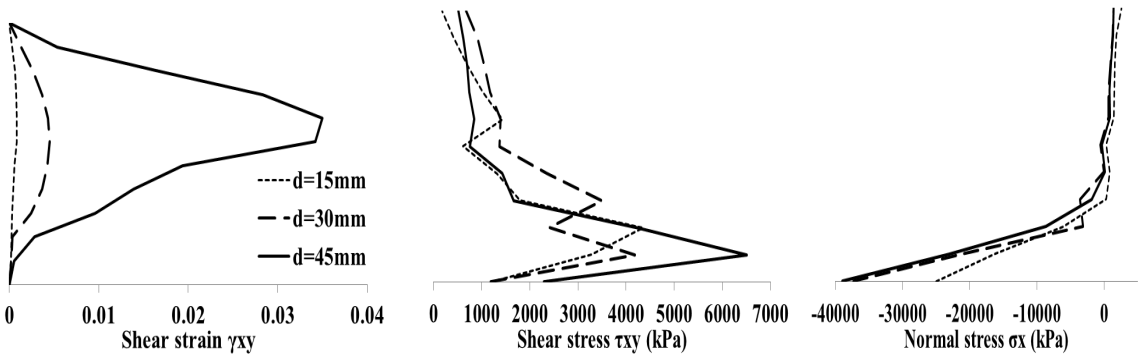



Figure 11: Shear strain, shear stress and normal concrete stress profiles for the 1st story beam.

### 5.3 RC bridge pier

The RC bridge pier R1 tested by Xiao et al. 1994 [24] had an aspect ratio of 4, 603mm x 406mm rectangular cross-section and 2398mm clear height (Figure 12). The pier was tested in double curvature, with the top and bottom rotations fixed, resulting in a shear span of 2. Longitudinal reinforcement consisted of 22, 20mm bars distributed around the perimeter of the section. Shear reinforcement consisted of 6.35mm rectangular hoops spaced at 127mm. A permanent axial load of 501kN was applied at the top of pier. The pier was subjected to increasing displacement increments at different levels of ductility.

| Concrete | | Bar Size | Diameter(mm) | $f_y$(MPa) | $f_u$(MPa) | $\varepsilon_u$‰ |
|---|---|---|---|---|---|---|
| Compressive strength(MPa) | 38 | #6 | 19.5 | 317 | 453 | 110 |
| Young Modulus(MPa) | 30820 | #2 | 6.35 | 361 | 483 | 110 |

Table 4: Bridge pier material properties.

The pier was modeled with one single element (5IPs) 2438mm long, fully fixed at the bottom and with the top node restrained against rotation. The section was discretized into 11 layers of equal thickness and the reinforcement distributed according to the layout. The top horizontal degree of freedom was subjected to increasing cyclic displacement. Experimental and analytical results are compared in Figure 12. An overestimation of the initial stiffness can be observed due to the fact that the specimen was pre-cracked under initial small-displacement excursions during the test. Good agreement is found in terms of strength, loading/unloading stiffness and permanent displacements. Results from monotonic analysis indicated imminent failure of the specimen at approximately 40mm top displacement.

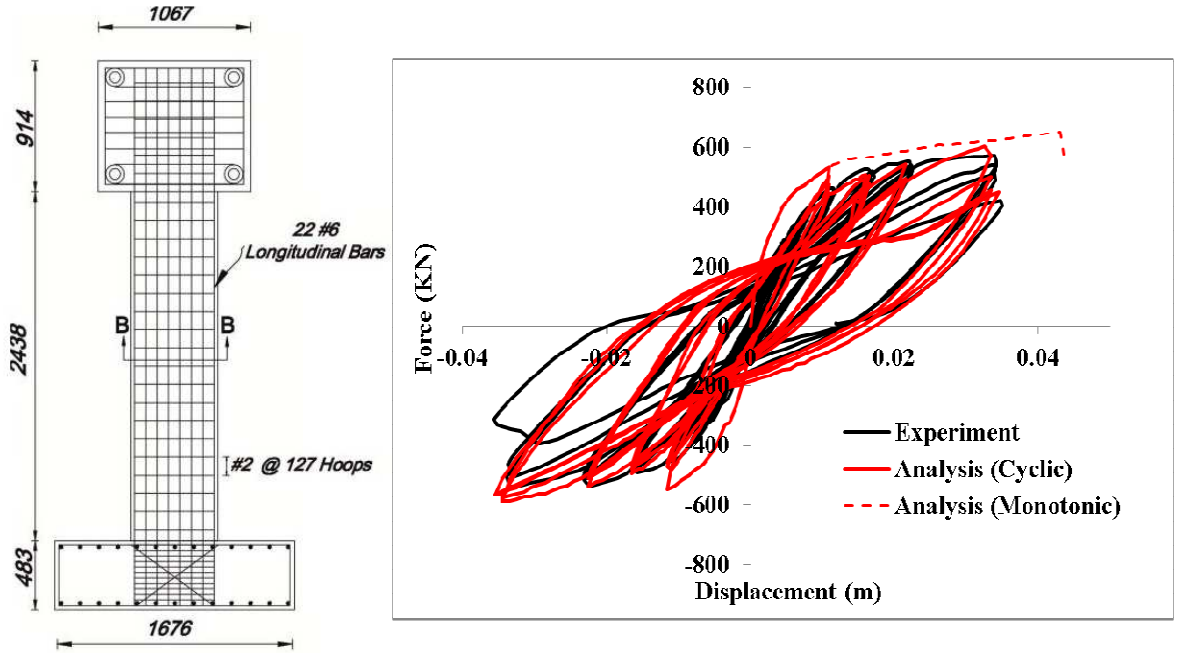



Figure 12: Bridge pier geometry and horizontal force-displacement response [24].

## 6 CONCLUSIONS

The proposed section model for the evaluation of an exact shear strain profile was satisfactorily implemented within a force-based, fiber-based frame element for monotonic and cyclic analysis of shear-critical RC structures.

Conceptually speaking, the element is superior to elements with fixed shear strain profiles, such as those with constant or parabolic patterns. However, the element kinematics still relies on the Timoshenko beam theory, i.e. plane section assumption and no section distorsion or warping, resulting in kinematic incompatibility between fiber deformations and element displacements.

A number of verification examples on shear-critical RC beams, frame and pier showed relatively good agreement in terms of post-cracked stiffness, shear strength and ultimate displacement capacity. Realistic stress and strain distributions over the section were obtained. However, further verification and comparison with experimental data is required.

## REFERENCES

[1] F.J. Vecchio, M. P. Collins, Predicting the response of RC members subjected to shear using the Modified Compression Field Theory. *ACI Structural Journal,* **85**(3), 258-268, May-June 1988.

[2] E.C. Bentz, Section analysis of reinforced concrete members. *PhD Thesis,* Department of Civil Engineering, University of Toronto, 2000.

[3] P. Ceresa, L. Petrini, R. Pinho, Flexure-Shear Fiber Beam-Column Elements for Modeling Frame Structures Under Seismic Loading — State of the Art, *Journal of Earthquake Engineering,* **11**(1), 46–88, 2007.

[4] J. Guedes, P. Pegon, A. Pinto, A Fibre/Timoshenko Beam Element in CASTEM 2000, *Special Publication Nr. I.94.31* Applied Mechanics Unit, Safety Technology Institute, Commission of the European Communities, Joint Research Centre, Ispra Establishment, Italy, 1994.

[5] L. Martinelli, Modellazione di pile di ponti in C. A. a travata soggetti ad eccitazione sismica, *PhD Thesis*, Dipartimento di Ingegneria Strutturale, Politecnico di Milano, Milano, Italy, 1998.

[6] G. Ranzo, M. Petrangeli, A fibre finite beam element with section shear modelling for seismic analysis of RC structures, *Journal of Earthquake Engineering*, **2**, 443–473, 1998.

[7] M. Petrangeli, P. E. Pinto, V. Ciampi, Fibre element for cyclic bending and shear of RC structures. I: theory, *Journal of Engineering Mechanics*, **125**(9), 994-1001, 1999.

[8] M. Petrangeli, P. E. Pinto, V. Ciampi, Fibre element for cyclic bending and shear of RC structures. II: verification, *Journal of Engineering Mechanics*, **125**(9), 1002-1009, 1999.

[9] P. Ceresa, L. Petrini, R. Pinho, R. Sousa, A fibre flexure–shear model for seismic analysis of RC-framed structures, *Earthquake Engineering and Structural Dynamics*, **38**, 565-586, 2009.

[10] J.N. Gregori, P.M. Sosa, M.A.F. Prada, F.C. Filippou, 3D frame element for the analysis of reinforced and prestressed concrete structures subjected to shear and torsion loads, *Engineering Structures*, **29**(12), 3404-3419, 2007.

[11] S. Gühner, Performance Assessment of Shear-Critical RC Frames. *PhD Thesis,* Department of Civil Engineering, University of Toronto, 2008.

[12] T. R. S. Mullapudi, A. Ayoub, Analysis of Reinforced Concrete Columns Subjected to Combined Axial, Flexure, Shear, and Torsional Loads, *Journal of Structural Engineering*, **139**(4), April 1, 2013.

[13] J. Mazars, P. Kotronis, F. Ragueneau, G. Casaux, Using multifiber beams to account for shear and torsion Applications to concrete structural elements*, Computers Methods in Applied Mechanics and Engineering*, **195**(52), 7264–7281, 2006.

[14] J.M. Bairán, A.R. Marí, Shear-Bending-Torsion Interaction in Structural Concrete Members: A Nonlinear Coupled Sectional Approach. *Computational Methods in Engineering*, **14**,249-278, 2007.

[15] S. Mohr, J.M. Bairan, A.R. Mari, A frame element model for the analysis of reinforced concrete structures under shear and bending, *Engineering Structures*, **32**,3936-3954, 2010.

[16] J. Pacheco Almeida, Flexure-Shear Element for 3D Inelastic Analysis of Frames, *PhD Thesis*, ROSE School, Università degli Studi di Pavia, 2009.

[17] V. Le Corvec, Nonlinear 3d frame element with multi-axial coupling under consideration of local effects, *PhD Thesis*, Department of Civil and Environmental Engineering, UC Berkeley, 2012.

[18] A. Papachristidis, M. Fragiadakis, M. Papadrakakis, A 3D fibre beam-column element with shear modeling for the inelastic analysis of steel structures. Computational Mechanics 45, 553–572, 2010.

[19] E. Spacone, V. Ciampi, F.C. Filippou, Mixed formulation of nonlinear beam finite element. *Computer Structures* 58(1), 71–83, 1996.

[20] A. Kagermanov. An exact shear strain evaluation for two-dimensional RC frame elements. *PhD Thesis,* UME School, Institute for Advanced Studies, Pavia, Italy, 2016.

[21] A. Kagermanov. Physically-based cyclic tensile model for RC membrane elements. *Individual Study,* UME School, Institute for Advanced Studies, Pavia, Italy, 2015.

[22] B. Bresler, A.C. Scordelis, Shear strength of RC beams, ACI Journal Proceedings, **60**(1), 51-74, 1963.

[23] K.V. Duong, S.A Sheikh, F.J. Vecchio, Seismic Behaviour of Shear-Critical Reinforced Concrete Frame: Experimental Investigation. *ACI Structural Journal*, 104(3), May-June, 304-313, 2007.

[24] Y. Xiao, M. J. N. Priestley, F. Seible, Steel Jacket Retrofit for Enhancing Shear Strength of Rectangular Short Reinforced Concrete Columns. *Structural Systems Research Project*, Report No. SSRP 92/07, University of California, San Diego, June, 1993.

[25] F.J. Vecchio, W. Shim, Experimental and Analytical Reexamination of Classic Concrete Beam Tests, *Journal of Structural Engineering*, 130(3), March 1, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# CROSS-SECTIONAL FAILURE CRITERION COMBINED WITH STRAIN-HARDENING DAMAGE MODEL FOR SIMULATION OF THIN-WALLED TEXTILE-REINFORCED CONCRETE SHELLS

**Sharei E.**[1], **Chudoba R.**[1], **and Scholzen A.**[1]

[1]Institute of Structural Concrete - RWTH Aachen University
Mies-van-der-Rohe-Str. 1
52074 Aachen, Germany
e-mail: esharei@imb.rwth-aachen.de

**Keywords:** Textile reinforced concrete, Cementitious composites, Damage mechanics, Shell structures, Failure criterion, Finite element analysis

**Abstract.** *Application of textile reinforced concrete (TRC) can significantly facilitate the manufacturing of thin-walled shell structures. Thanks to the high tensile strength and longtime durability of textile reinforcement, construction of much thinner cross sections compared to ordinary reinforced concrete has become possible. Furthermore the form-flexibility of textile reinforcement allows for much more freedom in design of shells with more complex geometries. However, these advantages make the prediction of the structural behavior under imposed loads much more difficult. Due to the complex stress redistribution in the plane of shell and also the anisotropic behavior of the composite material, different modes of failure can be observed in TRC shell structures. In this paper we focus on the failure criterion of thin-walled TRC shell elements. In particular, we propose a criterion reflecting the interaction of normal force and bending moment in a shell cross section. The failure criterion has been implemented in combination with an anisotropic damaged-based material model in order to provide realistic prediction of the structural load-carrying behavior. The accuracy of the model in prediction of different failure modes has been validated using three different types of tests setups.*

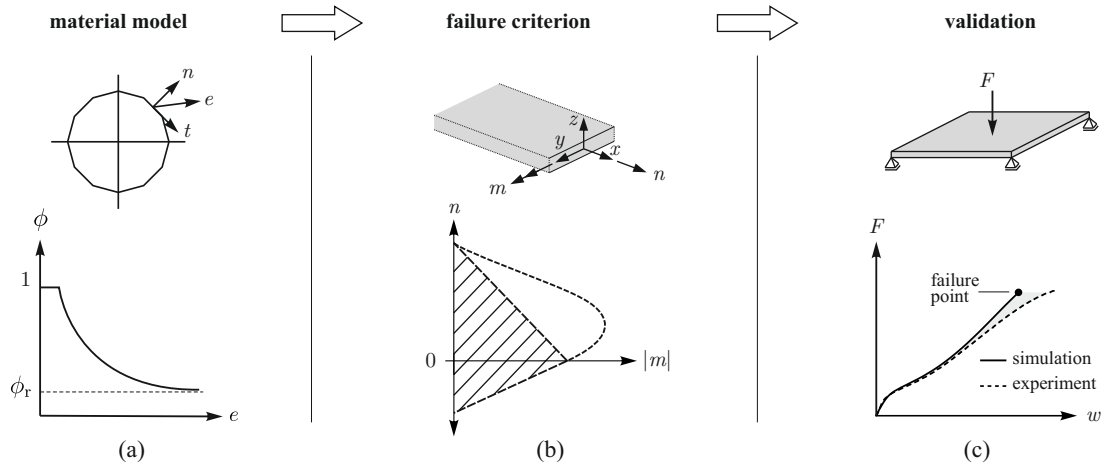



Figure 1: (a) material model formulation; (b) failure criterion of the TRC cross section based on $n$-$m$ interaction diagram; (c) validation of the failure criterion based on laboratory tests

## 1 INTRODUCTION

Textile reinforced concrete (TRC) is a composite material consisting of high performance fine aggregate matrix reinforced with textile reinforcement made of carbon or alkali-resistant glass rovings in form of fabrics. Two important characteristics of textile reinforcement are non-corrosiveness and high tensile strength. Application of this type of reinforcement increases the durability of the composite material compared to steel reinforced concrete and extends the range of its applicability in construction. Recently, TRC has been used for construction of large scale, thin-walled shells [1, 2], sandwich wall or slab panels [3], facade elements [4] and pedestrian bridges [5]. Focus of this paper is restricted to thin-walled TRC shell structures uniformly reinforced with textile fabrics.

TRC shell elements exhibit a complex behavior in terms of nonlinearity and anisotropy imposed to *in-plane* and *out-of-plane* loading conditions. Interaction of textile reinforcement with concrete matrix leads to multiple cracking under tension. Therefore, nonlinear behavior of strain-hardening type is observed in the composite material. Anisotropic behavior of TRC is due to the formation of cracks perpendicular to the principle direction of stresses [6]. At a macroscopic level, the process of crack pattern development can be viewed and modeled using *anisotropic damage* model.

In this paper behavior of the TRC shells at the material level and structural level has been investigated with the goal to formulate both realistic and pragmatic failure criterion. The overview of the modeling framework is shown in Fig. 1. In Section 2 the modeling approach with a brief description of the material model is presented. Consequently, after a brief description of calibration and validation of the model in Section 3 based on the previous studies, we investigate the failure criterion for TRC elements exposed to combined tensile force and bending moment in Section 4. The feasibility of the cross-sectional failure criterion has been demonstrated by calibrating the material model and the failure criterion using *tensile test*, *compressive test*, *bending test*. The validation for combined tensile and bending loads was performed using a *slab test*.

## 2 MODELING APPROACH OF TRC SHELLS

The crack pattern in the studied type of composite structures is fine in relation to the dimensions of the structural elements. This fine nature of cracking justifies a smeared representation

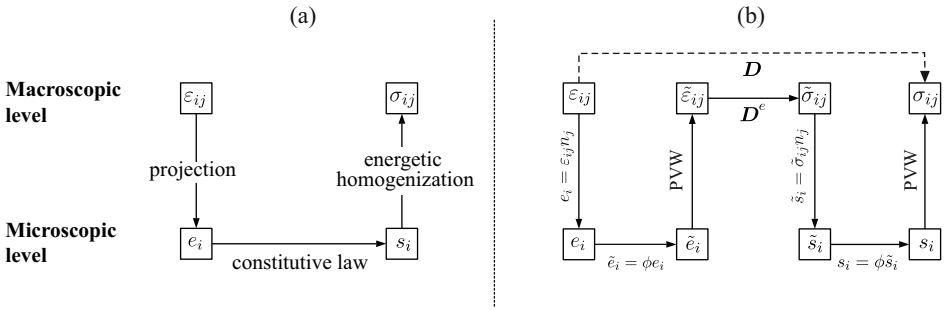



Figure 2: (a) basic principle of microplane model; (b) constitutive stress-strain relation in microplane damage model

of cracking using a homogenized strain-hardening model. In order to capture the nonlinearity and anisotropy of TRC, a damage-based model of microplane type has been be used. Material models with microplane approach are generally characterized as a constitutive relation between the strain and stress on plane of various orientations, so called *microplanes* [7]. Evolution of damage on a microplane is governed by a damage function. The shape of damage function can be tailored for either quasi-brittle materials with strain-softening behavior, e.g. plain concrete [8], or it can also reproduce the strain hardening behavior of quasi-ductile materials such as TRC [9].

Basic structure of microplane model approach is shown in Fig. 2a. The constitutive mapping of strain to stress is done in three steps: (i) geometric projection of macroscopic strain tensor $\varepsilon_{ij}$ into microplane directions providing strain vectors $e_i$, (ii) application of constitutive law between strain vector $e_i$ and stress vector $s_i$, (iii) energetic homogenization of microplane stress vectors based on the principle of virtual work (PVW) to obtain the global stress tensor $\sigma_{ij}$. The damage-based form of microplane model (microplane damage model) has been presented by Jirásek [8]. Constitutive formulation of this model is shown in Fig. 2b. In this formulation, the mapping of apparent strain/stress vectors ($e_i$ and $s_i$) to the effective strain/stress vectors ($\tilde{e}_i$ and $\tilde{s}_i$) is done on a microplane level using a damage function $\phi$. Inverse transformation of the variables from the microplane level to the macroscopic level is performed by principle of virtual work (PVW), setting the summation of virtual works of the forces applied to the system (in microscopic and macroscopic level) equal to zero. Although there is an explicit relation between effective strain and stress tensors ($\tilde{\varepsilon}_i$ and $\tilde{\sigma}_i$) in terms of elastic stiffness tensor $D^{\mathrm{e}}$, the global strain to stress mapping is only possible by consecutive steps of projection and energetic homogenization between macroscale and microplane levels to generate the stiffness tensor in damaged state $D$.

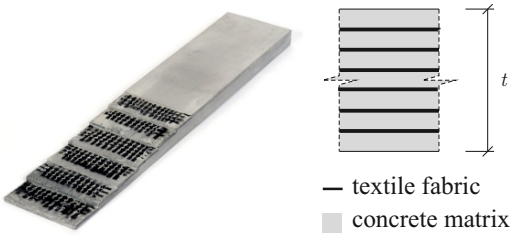
For the purpose of modeling of TRC shell structures, a two-dimensional formulation of the microplane damage model with strain-hardening for TRC has been presented by the authors in [10]. By means of dimensional reduction of the model, microplane damage model is used only for strain and stress components in the in-plane directions of the TRC shell and material behavior along the thickness of the shell is assumed to be linear elastic.

## 3 CALIBRATION AND VALIDATION

In order to determine the shape of the damage function of a certain TRC cross section, the material model needs to be calibrated. A direct incremental calibration procedure by means of a tensile test on a TRC specimen has been presented by Scholzen et al. [9]. The capability of the calibrated model in prediction of structural behavior of a large scale TRC shell has been

Table 1: Properties of TRC cross section

| description | symbol | value | unit | cross section layout |
|---|---|---|---|---|
| width | $b$ | 100 | mm | |
| thickness | $t$ | 20 | mm | |
| total cross section | $A$ | 2000 | $\mathrm{mm}^2$ | |
| number of textile layers | $n_{\mathrm{tex}}$ | 6 | – | |
| textile cross section | $A_{\mathrm{tex}}$ | 32.1 | $\mathrm{mm}^2$ | |
| reinforcement ratio | $\rho_{\mathrm{tex}}$ | 1.61 | % | |




presented by Sharei et al. [11].

The calibration is exemplified for TRC tensile test specimens reinforced with 6 equidistant layers non-impregnated carbon fabrics. Cross section properties and reinforcement layout within the cross section is given in Table. 1. The test setup with the specimen of the length $l = 1000\ \mathrm{mm}$ is shown in Fig. 3a. Strain has been measured within the length of $l_{\mathrm{m}} = 250\ \mathrm{mm}$. Stress-strain curve of the tensile test is given in Fig. 3b. The maximum tensile strain achieved in the test was $\varepsilon_{\mathrm{f}} = 7.05\ ‰$. Calibration of the material for this cross section based on [9] leads to the damage function in Fig. 3c. Failure assessment of laboratory tests has been performed by implementing the material model and the failure criterion in the finite element code ABAQUS[12].

During tensile test, damage evolves in form of a finely spaced cracks oriented perpendicularly to the direction of the tensile force. The failure occurs in the weakest crack once the ultimate strength of the textile reinforcement is achieved. The calibrated damage function represents the failure as a drop of the integrity value to zero. Thus, the tensile test serves for the calibration of the nonlinear behavior during the matrix cracking and, at the same time, for the identification of the ultimate failure.

## 4 CROSS-SECTIONAL FAILURE CRITERION

It is well known, that the bending tensile strength (also called as modulus of rupture) of quasi-brittle materials is larger than the tensile strength measured in tensile test [13]. In case of TRC, two major sources of the difference between tensile strength in pure tension and in bending can be specified distinguished:

- *Statistical size-effect:* In the tensile tests many crack bridges along the tested zone develop with weakest one defining the ultimate failure. In the three-point bending test, only the cross section at mid-span is actually tested, which may not be the weakest one.

- *Tensile load transfer:* Even for very smooth transfer of tensile load into the tensile specimen, local stress concentrations in the vicinity of clamps cannot be completely avoided and cause a premature specimen failure.

In order to properly reflect the macroscopically observed strength of a shell cross section, we propose a strength criterion defined at the level of the stress resultants of a cross section. In principle, the same approach is used in usual dimensioning approaches. In our case, we built the criterion on top of the material model implementation. Thus, except of the standard mapping

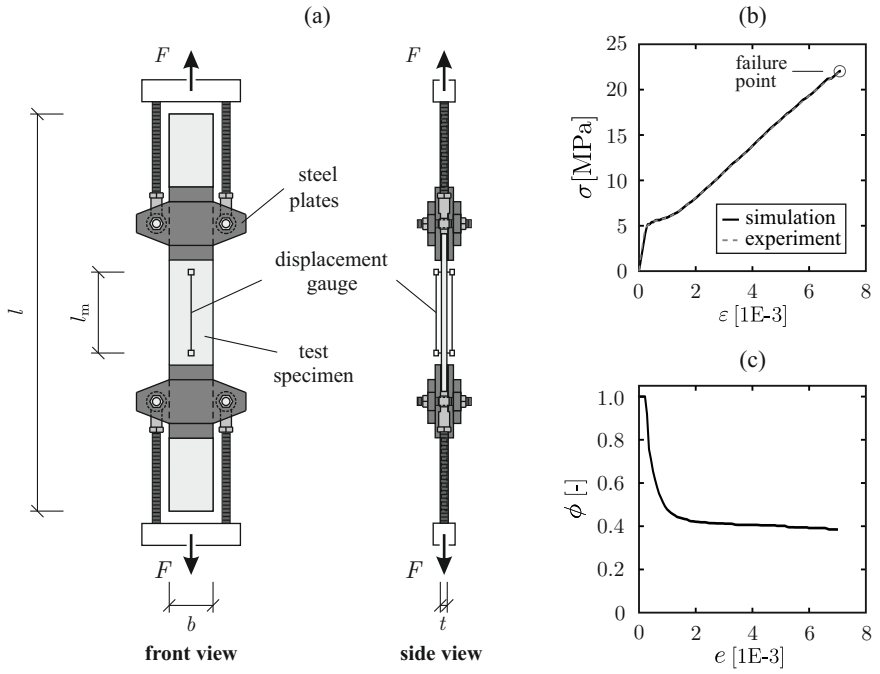



Figure 3: (a) test setup of pure tensile test; (b) stress-strain curve of the tensile test; (c) calibrated damage function

between the strain to stress tensors within an iteration loop, the material model also evaluates the stress resultants for a shell cross section and checks the level of its utilization.

In order to implement the cross-sectional criterion, the first step is to extend the strain capacity in the damage function beyond the limit measured in the tensile test given in Fig. 3c. Furthermore, the cross-sectional stress resultants, i.e. the vectors of normal forces $n$ and bending moments $m$ are calculated for each integration point of a shell. For a cross section depicted in Fig. 1b, normal force $n$ and bending moment $m$ are be determined during the analysis by integrating the stresses in the TRC cross section as:

$$n = \int_t \sigma_x dz \quad m = \int_t z\sigma_x dz \tag{1}$$

The cross-sectional resistance in presence of both normal force and bending moment is defined as a failure envelope shown in Fig. 4a. The real and the simplified $n$-$m$ envelope of a TRC cross section includes the values of strength in pure tension ($n_{\mathrm{t,Rm}}$), compression ($n_{\mathrm{c,Rm}}$) and bending ($m_{\mathrm{Rm}}$) [2]. Simplification of the failure envelope using the linear interpolation has been justified using a tests series on TRC specimens loaded with combined tensile force and bending moment [14]. The normalized form of the envelope is shown in Fig. 4b defining the utilization parameters as:

$$\eta_{\mathrm{nt}} = \frac{n}{n_{\mathrm{t,Rm}}} \quad \eta_{\mathrm{nc}} = \frac{-n}{n_{\mathrm{c,Rm}}} \quad \eta_{\mathrm{m}} = \frac{|m|}{m_{\mathrm{Rm}}}. \tag{2}$$

Utilization ratio $\eta_{\mathrm{nm}}$ of the cross section can be evaluated as:

$$\eta_{\mathrm{nm}} = \max\left\{\eta_{\mathrm{nc}}, \eta_{\mathrm{nt}}\right\} + \eta_{\mathrm{m}}. \tag{3}$$

Failure in a TRC cross section occurs once the interaction of normal force $n$ and bending moment $m$ exceeds the admissible range of $n$-$m$ envelope in Fig. 4, or in the normalized form:

$$\eta_{\mathrm{nm}} > 1. \tag{4}$$

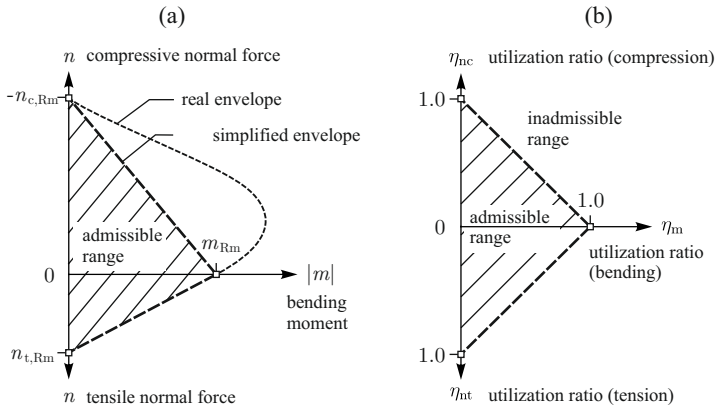



Figure 4: Failure envelope of TRC cross section based on (a) $n$-$m$ interaction diagram and (b) $\eta_{\mathrm{n}}$-$\eta_{\mathrm{m}}$ interaction diagram

Strength characteristics of the TRC cross section in Table. 1 have been determined as: tensile strength $n_{\mathrm{t,Rm}} = 476\ \mathrm{N/mm}$, compressive strength $n_{\mathrm{c,Rm}} = 1666.7\ \mathrm{N/mm}$ and bending strength $m_{\mathrm{Rm}} = 3330\ \mathrm{Nmm/mm}$.

The force-deflection curves measured and simulated for three-point bending test on a TRC specimen with the span length of $l = 460$ mm and the cross section given in Table. 1 is shown in Fig. 5a. Test has been performed with three repetitions. As indicated in Fig. 5a, the tensile strength determined in tensile test is attained at $\approx 70\ \%$ of the ultimate force in three-point bending test. Stress and strain cross-sectional profiles corresponding to the failure point are given in Fig. 5b with failure tensile strain of $1.66\ ‰$ is beyond the ultimate tensile strain in the tensile test.

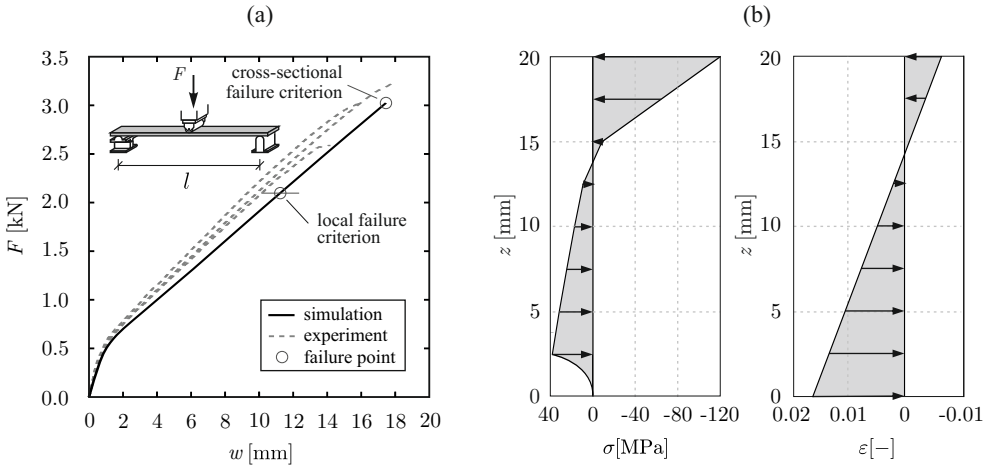



Figure 5: (a) Comparison of load-diplacement curves and ultimate failure of TRC bending tests for test and simulation applying two failure criteria; (b) stress and strain profiles in the cross section at failure

The validation of the cross-sectional failure criterion for prediction of the structural behavior including stress redistribution and complex damage pattern has been performed using TRC slab test. Test specimen had dimensions of $800 \times 800$ mm with the cross section parameters given in Table 1. The pin supported specimen was loaded at the center using displacement control as shown in Fig. 6. Simulation of the slab test was performed with and without consideration of

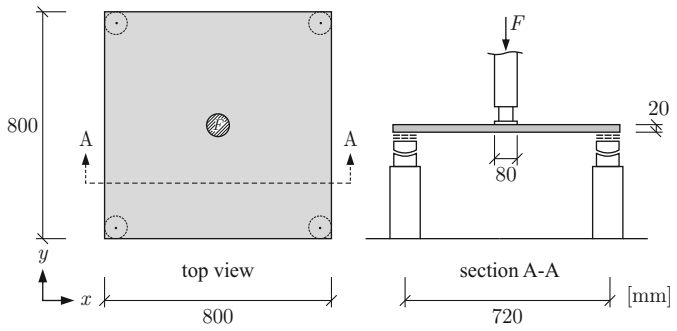



Figure 6: Test setup of the slab test

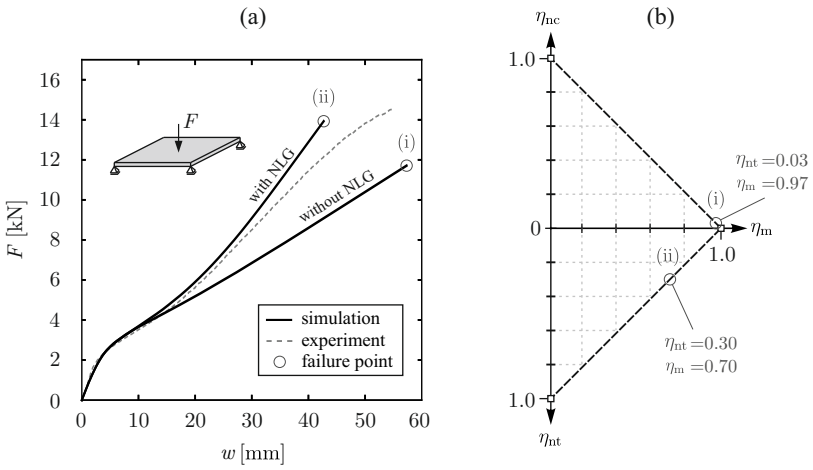



Figure 7: (a) Comparison of force-deflection curves and ultimate failure of TRC slab test for test and simulation applying cross-sectional failure criterion; (b) corresponding utilization ratios in the $\eta_n$-$\eta_m$ interaction diagram with and without consideration of geometrical nonlinearity

geometrical nonlinearity. Since the deflection of the slab was large, membrane effect could be activated during the later stages of loading. In this way, the ultimate failure was achieved for combined action of tensile force and bending moment. The geometrically linear model without the ability to reflect the membrane effect resulted in unrealistic prediction of the true behavior (force-deflection curve labeled *without NLG* in Fig. 7a). The corresponding utilization ratios are $\eta_n = 0.03$ and $\eta_m = 0.97$ for tensile force and bending moment respectively as shown in Fig. 7b, indicating a pure bending failure without activation of membrane forces in the cross section.

The force-deflection curve with the considered effect of geometrical nonlinearity, (labeled *with NLG* in Fig. 7a), has revealed a membrane effect leading to a failure with utilization ratios for tensile force and bending moment of $\eta_n = 0.30$ and $\eta_m = 0.70$, respectively as depicted in Fig. 7b, indicating a considerable tensile force in the cross section at failure. The corresponding $n$-$m$ pairs for both cases with and without geometrical nonlinearity are given in Table 2.

## 5 CONCLUSIONS

In this study failure of thin-walled TRC shell structures imposed to in-plane and out-of-plane loading conditions was investigated. Failure was predicted using two approaches: 1) local failure criterion (failure of material point) and 2) cross-sectional failure criterion (failure through $n$-$m$ interaction). Based on the simulation results we can conclude that the cross-

Table 2: Assessment of the *n-m* failure criterion for the slab tests with and without consideration of geometrical nonlinearity (NLG)

| NLG | $n$ | $m$ | $\eta_{\mathrm{nc}}$ | $\eta_{\mathrm{nt}}$ | $\eta_{\mathrm{m}}$ |
|---|---|---|---|---|---|
| | $[\mathrm{N/mm}]$ | $[\mathrm{Nmm/mm}]$ | $[-]$ | $[-]$ | $[-]$ |
| without | -46 | -3246 | 0.03 | - | 0.97 |
| with | 143 | -2350 | - | 0.30 | 0.70 |

sectional failure criterion can effectively represent the stress redistribution effects within a TRC cross section and provides a realistic estimate of the ultimate structural load.

## ACKNOWLEDGEMENTS

The financial support by *Deutsche Forschungsgemeinschaft (DFG)* (project No. 276/2-2) is gratefully acknowledged.

## REFERENCES

[1] Scholzen A., Chudoba R., and Hegger J. Thin-walled shell structure made of Textile Reinforced Concrete; Part I: structural design and construction. *Structural Concrete*, 16:106–114, 2015.

[2] Scholzen A., Chudoba R., and Hegger J. Thin-walled shell structure made of Textile Reinforced Concrete; Part II: experimental characterization, ultimate limit state assessment and numerical simulation. *Structural Concrete*, 16:115–124, 2015.

[3] Shams A., Horstmann M., and Hegger J. Experimental investigations on Textile-Reinforced Concrete (TRC) sandwich sections. *Composite Structures*, 118:643–653, December 2014.

[4] Hegger J., Kulas C., and Horstmann M. Spatial textile reinforcement structures for ventilated and sandwich facade elements. *Advances in Structural Engineering*, 15(4):665–676, April 2012.

[5] Hegger J., Goralski C., and Kulas C. A slender pedestrian bridge made of textile reinforced concrete. *International Association for Bridge and Structural Engineering (IABSE), Venice, Italy*, September 2010.

[6] Hughes B. P. and Ash J. E. Anisotropy and failure criteria for concrete. *Matriaux et Construction*, 3(6):371–374, November 1970.

[7] Bažant Z. P. and Prat P. C. Microplane model for brittle-plastic material: I. Theory. *Journal of Engineering Mechanics*, 114(10):1672–1688, 1988.

[8] Jirásek M. Comments on microplane theory. *Mechanics of Quasibrittle Materials and Structures, Hermes Science Publications*, pages 55–77, 1999.

[9] Scholzen A., Chudoba R., and Hegger J. Calibration and validation of a microplane damage model for cement-based composites applied to Textile Reinforced Concrete. In Barros H., Faria C., and Ferreira C., editors, *International Conference on Recent Advances*

*in Nonlinear Models - Structural Concrete Applications (CoRAN 2011)*, pages 417–427. ECCOMAS Thematic Conference, November 2011.

[10] Sharei E., Scholzen A., Chudoba R., and Hegger J. Anisotropic damage model for numerical simulation of Textile Reinforced Concrete shell structures. In Topping B.H.V. and Iványi P., editors, *Twelfth International Conference on Computational Structures Technology (CST 2014)*. Civil-Comp Press, September 2014.

[11] Sharei E., Scholzen A., and Chudoba R. Nonlinear analysis of textile reinforced concrete shells using an anisotropic, damaged-based material model. In Brameshuber W., editor, *11th International Symposium on Ferrocement (FERRO-11) and 3rd International Conference on Textile Reinforced Concrete (ICTRC-3)*, pages 133–139. RILEM Publications S.A.R.L., June 2015.

[12] Dassault Systèmes Simulia Corp., Providence, RI, USA. *ABAQUS User Subroutines Reference Manual, Version 6.11*, 2011.

[13] Troxell G. E., Harmer E. D., and Joe W. K. *Composition and Properties of Concrete*. McGraw Hill, second edition, 1968.

[14] Scholzen A., Chudoba R., and Hegger J. Ultimate limit state assessment of TRC structures with combined normal and bending loading. In Brameshuber W., editor, *11th International Symposium on Ferrocement (FERRO-11) and 3rd International Conference on Textile Reinforced Concrete (ICTRC-3)*, pages 159–166. RILEM Publications S.A.R.L., June 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SEISMIC ASSESSMENT OF A 5-STOREY RETROFITTED RC BUILDING

**Hussein Bark**[1], **George Markou**[1], **Christos Mourlas**[2] **and Manolis Papadrakakis**[2]

[1] ALHOSN University, Department of Civil Engineering
P.O.Box 38772, Abu Dhabi, UAE
habark01@students.alhosnu.ae, g.markou@alhosnu.ae

[2] Institute of Structural Analysis & Seismic Research, National Technical University of Athens
9 Iroon Polytechniou Str.,Zografou Campus,GR-15780 Athens,Greece
mourlasch@central.ntua.gr, mpapadra@central.ntua.gr

**Keywords:** Retrofitted Structure, Reinforced Concrete Modeling, Finite Element Method, Seismic Assessment.

**Abstract.** *Due to severe damages on structures caused by earthquakes different approaches were introduced to enhance their carrying capacity. The most common one is retrofitting, which is a technique used to improve the structural capacity of a building that is found to be inadequate against earthquake loads. The problems that arise when retrofitting reinforced concrete structures, is the use of objective and sound methods that will provide the engineer with the tools to assess the final design. This work aims to alleviate this restriction by performing a complete study of a 5-storey reinforced concrete structure that was retrofitted with reinforced concrete shear walls through the use of 3D solid finite elements. Cracking is treated with the smeared crack approach and the steel reinforcement is simulated by embedded rod elements. This detailed modeling approach is applied for the first time so as to perform a full pushover analysis of a 5-storey RC retrofitted building. The numerical results illustrate that the proposed assessment approach is computationally accurate. Furthermore, the use of parallel processing is required so as to decrease the computational demand when dealing with full scale models.*

## 1 INTRODUCTION

Retrofitting is required under several circumstances according to [1], especially when not taking into account the additional seismic loads while designing. Retrofitting is applied on existing structures that were designed according to old seismic codes, thus their lateral strength is not sufficient to carry the expected seismic forces. Moreover, building with poor construction quality can be enhanced as well.

There are mainly two types of retrofitting, global and local. The global retrofitting is concerned with the seismic resistance of a structure by placing shear walls, infill walls, steel bracings, and base isolation [2]. On the other hand, the techniques that increase the carrying capacity of inadequate members locally are considered as local strengthening methods. This includes reinforced concrete (RC), steel or fiber reinforced polymer (FRP) jacketing of the existing members such as columns, beams, joints, walls and foundations [3]. Selecting the appropriate retrofitting strategy is based on economy, performance and constructability. Sometimes it is more preferable to use more than one technique in a single structure

When it comes to design the retrofitted structural member, the available methods are rather limited given that semi-empirical formulas available in design codes account for the behavior of the section but do not integrade the effect on the overall mechanical behavior of the structure and its derived capacity after the retrofitting has been implemented locally. As it will be shown in this research work, the prementioned restrictions can be alleviated through the use of state-of-the-art 3D detailed finite element modeling.

## 2 NUMERICAL METHOD USED

The numerical method that was used in this research work was the Finite Element Method (FEM) [4]. In order to determine the mechanical behavior of the under study structure, two software have been used: Femap and ReConAn. Below follows a short description of the two finite element analysis software.

### 2.1 Femap

Femap is an advanced engineering simulation program owned by Siemens PLM Software. It is used as a “pre-processing” tool to construct finite element models and as a “post-processing” tool to display results.

### 2.2 ReConAn FEA

ReConAn is a research finite element analysis software created at the NTUA [5]. It uses two post-processing programs, Femap and ReConAn Eye. ReConAn Eye is encapsulated inside ReConAn main code structure to illustrate the crack openings that derive from the analysis as smeared cracks.

## 3 5-STOREY RC BUILDING

The under study building used in this research work was a seismically strengthened 5-storey RC building adopted from Orakdogen et al. [6], where the building was analyzed and studied for the effect of soil structure interaction (SSI) for different foundation types (see also [7]). This RC building is used herein so as to develop the numerical models that will be presented in the next sections.

### 3.1 Geometry

As it can be seen in Figure 1, the initial framing system of the building consisted of 12 columns (with rectangular sections of 25x40cm) and drop-beams (15x50cm) supporting the slabs (10cm of thickness). The initial framing system was strengthened by constructing four main shear walls (two shear walls parallel to the x axis and two parallel to the y axis) around the existing columns (Figure 1). The total length and width of the building's plan are 10.2m and 6.8m, respectively, while the total building height is 13.9m (the ground floor height is 2.15m, 3 typical floors have a 2.75m height each and the top floor has a height of 2.5m) [6].

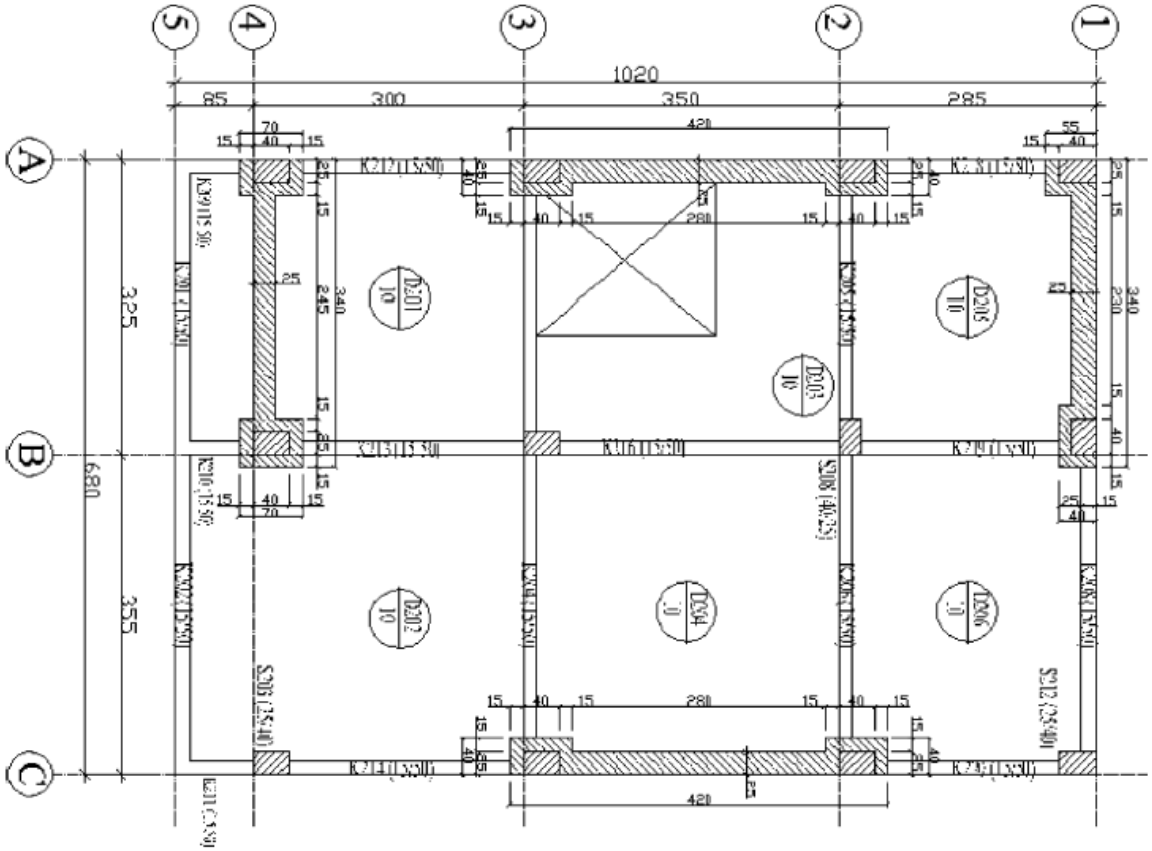



Figure 1: RC 5-storey building typical plan view [6]

### 3.2 Material properties

The material properties of the structure are divided into: (a) existing RC material, where the uniaxial compressive strength of concrete was reported to be equal to 16MPa and the steel reinforcement that had a yielding stress of 220 MPa, (b) the RC material used to strengthen the building (infill walls) had 20MPa and 420MPa, uniaxial concrete compressive strength and steel yielding stress, respectively. These were also the values incorporated in the numerical models presented below.

## 4 NUMERICAL MODELS

Two models were developed so as to study the overall effect of retrofitting. Model A foresees the discretization of the initial frame and Model B the discretization of the retrofitted frame. 20-nodded hexahedral elements where used in order to discretize the concrete meshes while rod finite elements were used for the embedded rebars. Different colors where used so

as to distinguish between different rebar diameters. Also steel plates (2cm thick) were used for the application of the external horizontal load in order to avoid any local failure. It must be noted here that the slabs were modeled through the use of a von Mises material model in an attempt to decrease the computational demand of the models. The concrete uses a 3D constitutive matrix as presented by Kotsovos and Pavlovic [8] and later modified by Markou and Papadrakakis [9].

### 4.1 Initial model (Model A)

Table 1 shows the reinforcement details for each structural member of the building.

| Element | Properties |
|---|---|
| **Beams** | Rebar: 4Ø12<br>Stirrups: Ø8/15cm |
| **Columns** | Rebar: 6Ø14<br>Stirrups: Ø12/15cm |
| **Footings** | Top & Bottom X direction: Ø12/15cm<br>Top & Bottom Y direction: Ø12/15cm |

Table 1: Reinforcement details of the initial mode

As stated above, Femap was used in order to create both numerical models (initial and retrofitted frame). The procedure followed for the model construction is summarized in four steps: (a) constructing the concrete mesh, (b) create the embedded rebars inside the concrete mesh, (c) apply the constrains, and (d) apply the loads. The initial model is shown in Figure 2a, where the concrete mesh, the applied loads and the fix support can be seen. Figure 2b shows the embedded rebars mesh, while a zoom view that illustrates the embedded rebars within a RC joint is also shown.

Three types of loads were applied in both models. (a) Dead load: the self-weight of the building which applied automatically by the software, (b) Live load: 3kN/m$^2$ distributed load were assigned on each slab. (c) Earthquake load: horizontal load was applied along the Y-direction at each floor as shown in Table 2. The pushover analysis foresaw 10 load increments and a convergence tolerance of $10^{-4}$.

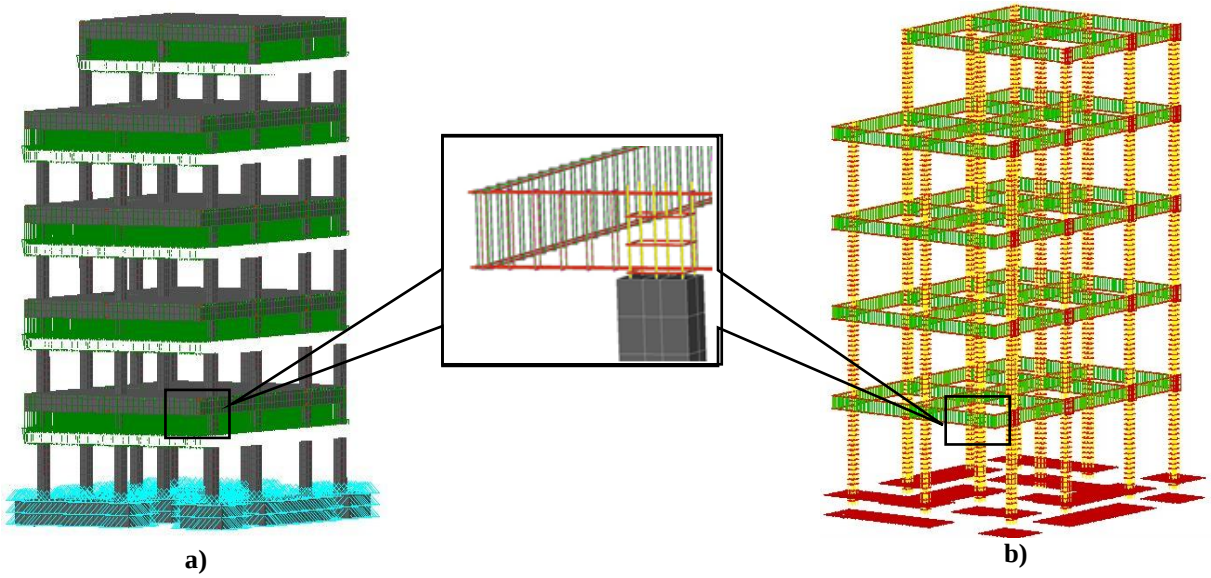



Figure 2: Model A. a) Hexa mesh. b) Embedded rebars mesh

| Storey number | Earthquake loads (kN) |
|---|---|
| **First** | 66.67 |
| **Second** | 133.33 |
| **Third** | 200 |
| **Forth** | 266.67 |
| **Roof** | 333.33 |
| **Total** | 1000 |

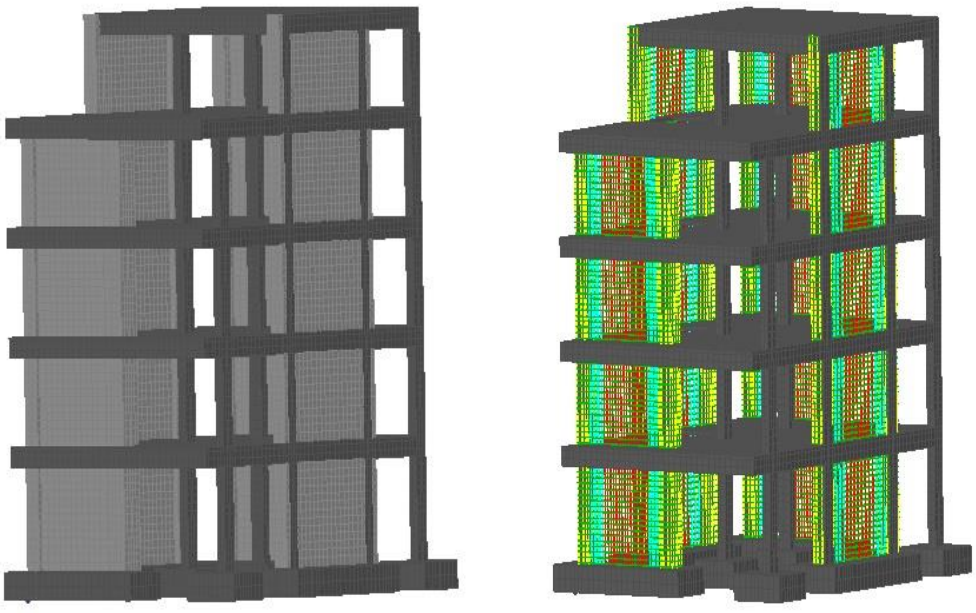
Table 2: Earthquake loads distribution

### 4.2 Retrofitted model (Model B)

As previously mentioned, RC shear walls were used in both directions to improve the overall stiffness and strength of the structure (see Figure 1). The shear walls and RC jackets were reinforced as shown in Table 3. New finite elements were added to Model A without any modifications performed on the initial mesh. Figure 3 and Figure 4 present the mesh of the second model (Model B).

| Element | Properties |
|---|---|
| **RC Jackets** | Dimensions: 15 cm around columns<br>Rebar: Ø14 at each 15 cm<br>Stirrups: Ø8/15 cm |
| **Shear Wall** | Web Rebar: 16ø12<br>Confinement Rebar: 6Ø20<br>Stirrups: Ø8/15 cm |

Table 3: Rebars used for retrofitting



a) b)

Figure 3: Model B. a) Concrete mesh of the retrofitted structure. b) rebar mesh embedded in the shear walls

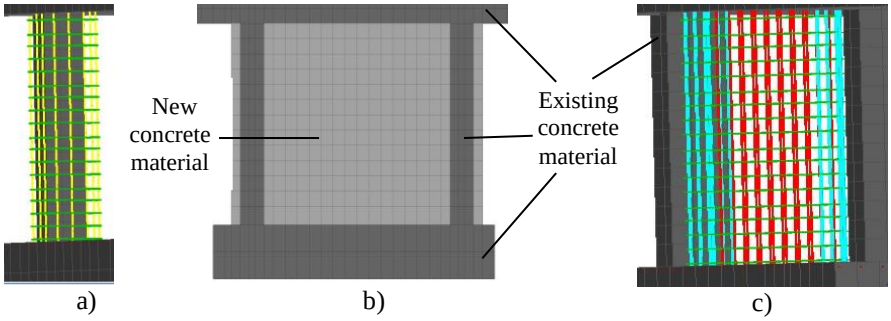



Figure 4: Model B. a) Retrofitting rebars around the column. b) Shear wall concrete mesh with initial columns. c) Shear wall reinforcement mesh

## 5 NUMERICAL RESULTS AND DISCUSSION

### 5.1 Analysis of Model A

After constructing the two models, the input data files were exported for analysis through the use of ReConAn FEA. The solution of the entire full-scale model required 29 days, which highlights the excessive computational demand of the problem. So as to assemble and store the stiffness matrix of Model A (Skyline storage), 30Gb of RAM were required. This numerical finding underlines the need of using parallel processing (a research work that is under process), while demonstrates the numerical robustness of the developed algorithm in handling this large-scale numerical problem.

Figure 5 shows the predicted P-δ curve. As it resulted from the numerical analysis, the structure managed to carry a total of 540kN prior complete failure. The failure was attributed to the failure of longitudinal reinforcement at the ground floor columns which initiated an excessive internal forces redistribution leading the frame to failure. The computed capacity does not satisfy the codes demand which was 693-954kN [6] according to the soil conditions.

As it can be seen in Figure 6a, the predicted crack initiation at load increment 1 foresees that the main crack openings are located at the joints of the structure. As the horizontal load increased, the cracks increased in size and number until complete failure (see Figure 6b).

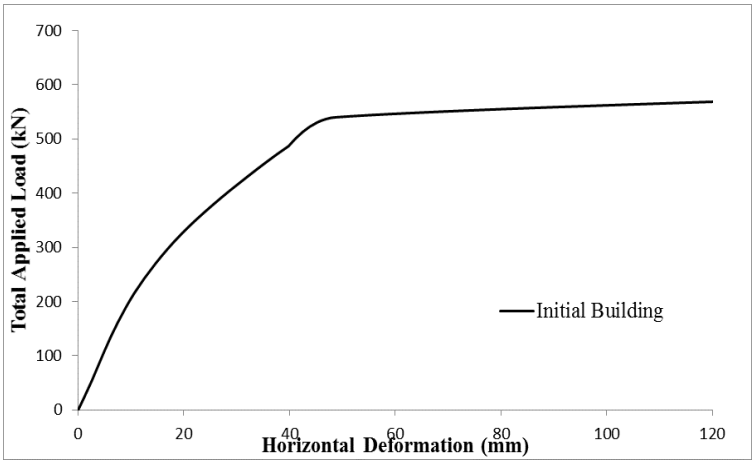



Figure 5: P-δ curve of initial building

### 5.2 Analysis of Model B

On the other hand, the solution of Model B required 66 days. So as to assemble and store the stiffness matrix of Model B, 37Gb of RAM were required due to the additional FE elements used in

order to discretize the retrofitting infill walls. The analysis details are summarized in Table 4 for both models.

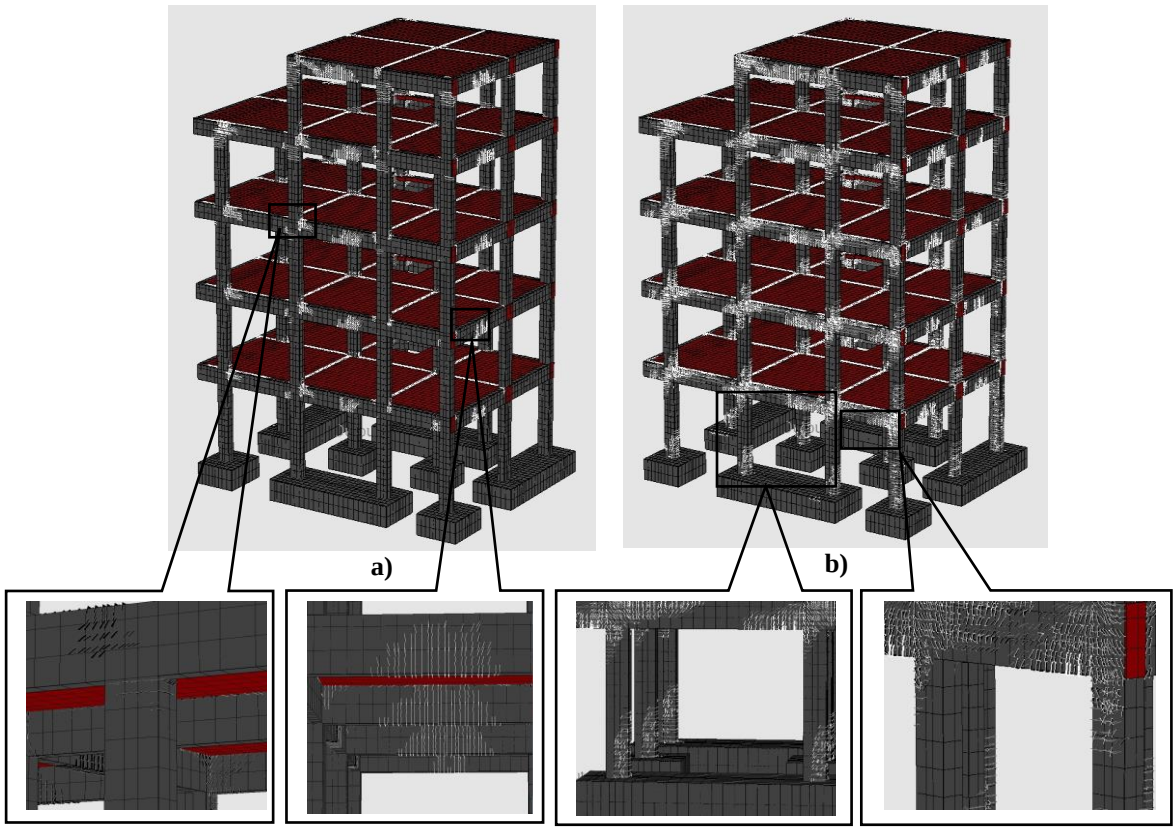


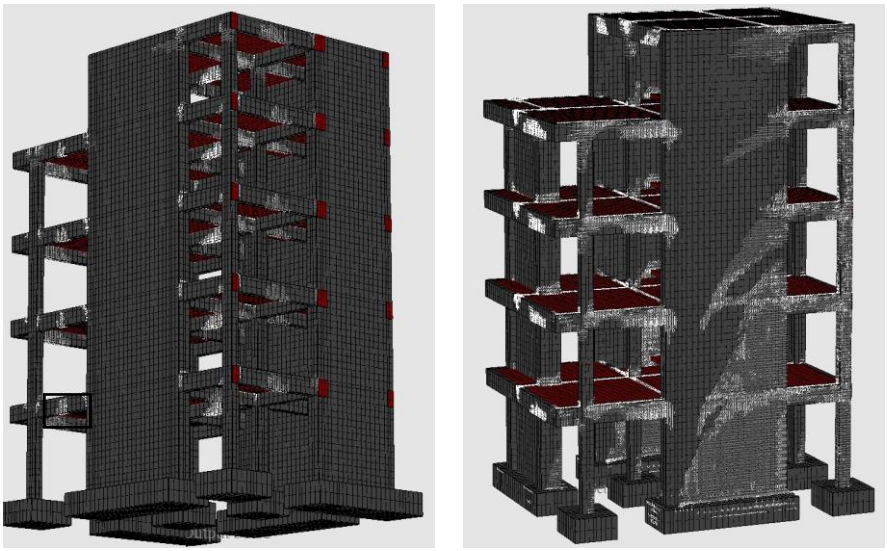

Figure 6: Model A. a) Initial cracks at load increment 1. b) Cracks prior to failure

| | **Model A** | **Model B** |
|---|---|---|
| **Mesh Data** | | |
| Total no. nodes | 159,567 | 240,060 |
| No. hexa elements | 19,044 | 29,431 |
| No. embedded rebars | 27,399 | 58,072 |
| Total no. elements | 46,443 | 87,503 |
| **Analysis Data** | | |
| No. of stiffness matrix elem. | 3,761,459,688 | 4,688,727,648 |
| Required RAM for the stiffness matrix | 30Gb | 37Gb |
| Required RAM for the analysis | 36Gb | 43Gb |
| Embedded mesh generation time | 10min 14sec | 34 min 6 sec |
| Time for solving nonlinear analysis | 29-days | 66-days |
| No. of load incr. solved successfully | 6 load incr. | 9 load incr. |
| Time for writing the output data | 25min 48sec | 1hr 28min |
| Total computational time | 29-days | 66-days |

Table 4: General numerical details that derived after the nonlinear analysis of the FE models.

As it can be seen in Figure 7a, the results from the pushover analysis of Model B, predicted the initiation of cracks occurred at the beams' end joints due to the moments that were generated from the applied horizontal load. As the horizontal load increases, the cracks increase in

size and number until complete failure (see Figure 7b). The shear walls develop horizontal and diagonal cracks as the horizontal load increases (see Figure 7b). The retrofitted model failed due to the rapture of the longitudinal reinforcement of the beams that were connected to the shear wall leading the solution to stop.



**a)** **b)**

Figure 7 Model B. a) Initial cracks at load increment 1. b) Cracks prior to failure

Figure 8 shows the comparison between the predicted P-δ curves for both models. As it resulted from the numerical analysis, the strengthened structure managed to carry a total of 1800kN prior complete failure. As mentioned above, the failure was attributed to concentration of strains at the beam-shear wall joints which led the frame to failure. It is also shown that with the increase in the stiffness due to the retrofitting shear walls, the resulted horizontal displacement decreased significantly.

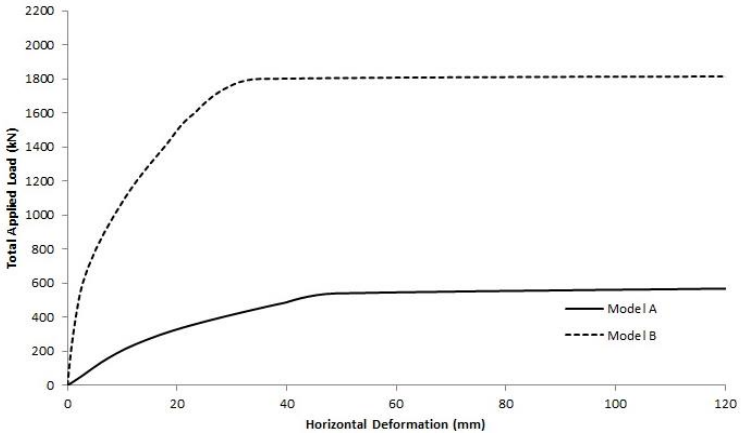



Figure 8: Comparison of P-δ curves for both models (A & B)

| | **Model A** | **Model B** |
|---|---|---|
| No. of load increments until failure | 6 | 9 |
| Numerically predicted Base Shear (kN) | 540 | 1800 |
| Max Horizontal Deformation (cm) | 5.02 | 3.94 |

Table 5: General outputs that derived from the FEA of both models.

Table 5 illustrates the comparison between the numerical results for both models. While Figure 9 shows the von Mises strain contours for both framing systems prior to failure.

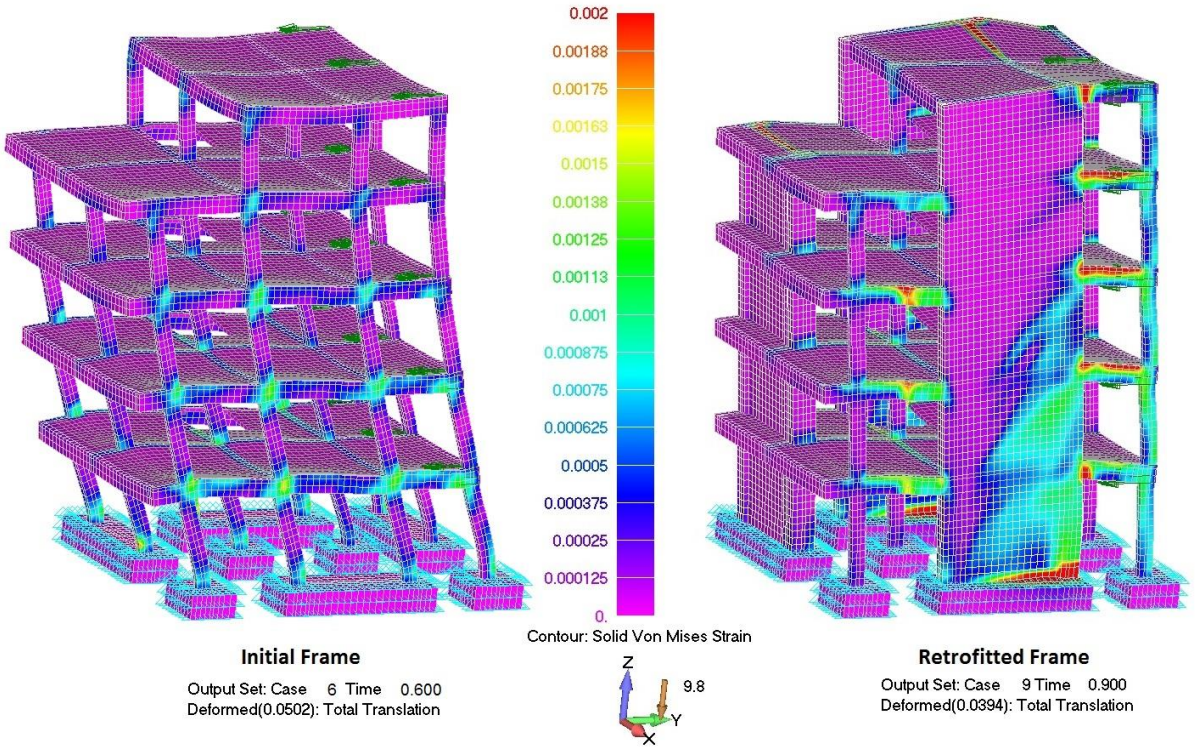



Figure 9: Comparison of von Mises contours for both models prior to failure

### 5.3 Analysis Results of a Single RC Infill Shear Wall

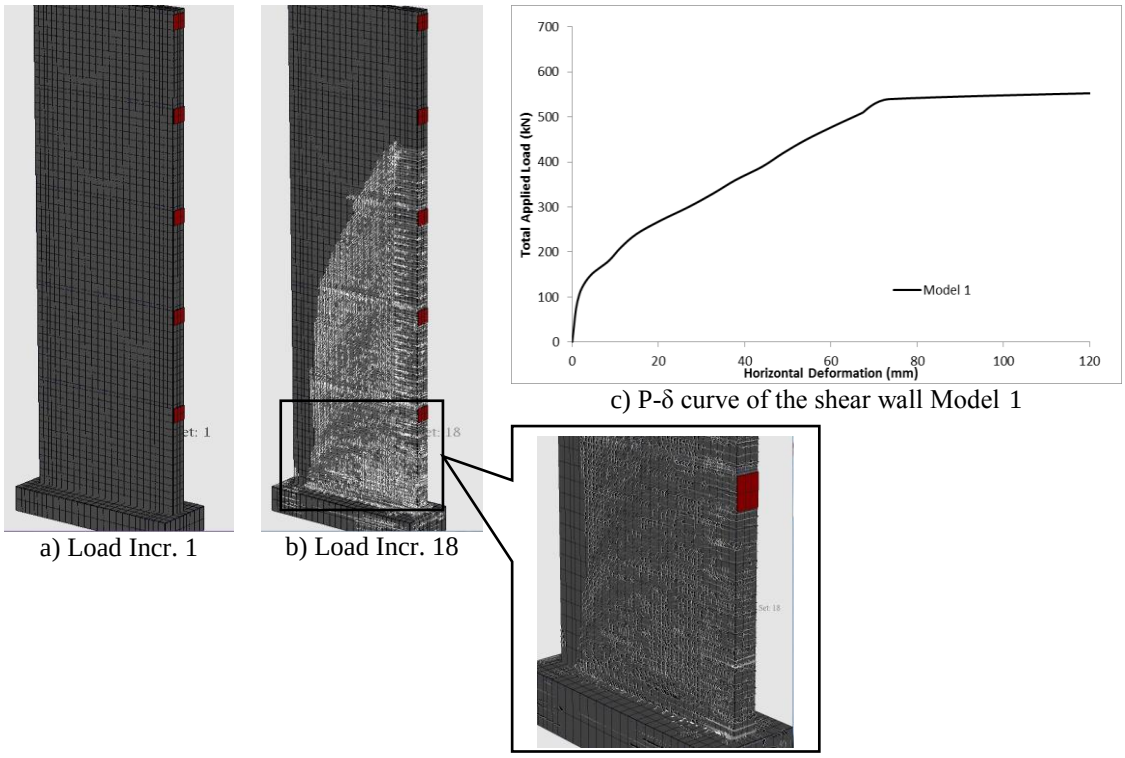
In order to investigate the overall capacity of the derived shear wall (13.9m height) after the retrofitting, a shear wall model was developed and analyzed until complete failure (see Figure 10). The retrofitting rebars that were used to construct the shear wall were assumed well anchored within the existing framing system, so as to avoid any local failure and capacitate the full strength of the shear wall. Eurocode 2 [10] was used so as to compare the numerically obtained failure load. According to Equations (1) and (2) of EC2, the resulted resistance of the shear wall section was 687kN in shear, and 5880kNm in bending moment. It must be noted here that only the characteristic values were used so as to compute the strength in shear and moment.

$$M_{Rd}= b_w \cdot l_c \cdot \rho_{vc} \cdot f_s \cdot d_1 + b_w \cdot x_u \cdot \rho_{vc} \cdot f_s \cdot d_2 + b_w \cdot (l_c + c - x_u) \cdot \rho_{vc} \cdot f_s \cdot d_3 + b_w \cdot (l_w - 2h_c) \cdot \rho_{vw} \cdot f_s \cdot d_4 + b_w \cdot x_u \cdot 0.8 f_c \cdot d_2 \quad (1)$$

$$V_{Rd,c}=[C_{Rd,c} \cdot k(100 \cdot \rho_l \cdot f_{ck})^{1/3} + k_1 \sigma_{cp}] b_{wd} \quad (2)$$

Figure 10a represent the crack pattern of the shear wall (Model l) at load increment 1 where a 5% of the horizontal load (30kN) was applied. No cracks were formed until load increment 3 where the model entered the inelastic zone (the total horizontal load was applied incrementally through 20 load increments). The shear wall failed at load increment 19 (95% of the applied horizontal load) as shown in Figure 10. The P-δ curve in Figure 10c shows the overall response of the retrofitted part of the structure as it was obtained from the analysis where it can be seen that the predicted failure load was 540kN, smaller than the load predicted

by EC2. This is attributed to the code's inability to account for the eccentricities that derive additional stresses and strains during the loading procedure, leading the shear wall to a premature failure. When the vertical static loads are applied, the shear wall develops an out of plane initial deformation thus the horizontal loads force the wall to deform in a bending and torsional manner (due to the eccentricity).




a) Load Incr. 1 b) Load Incr. 18 c) P-δ curve of the shear wall Model 1

Figure 10: Shear wall Model 1 crack patterns and P-δ curve

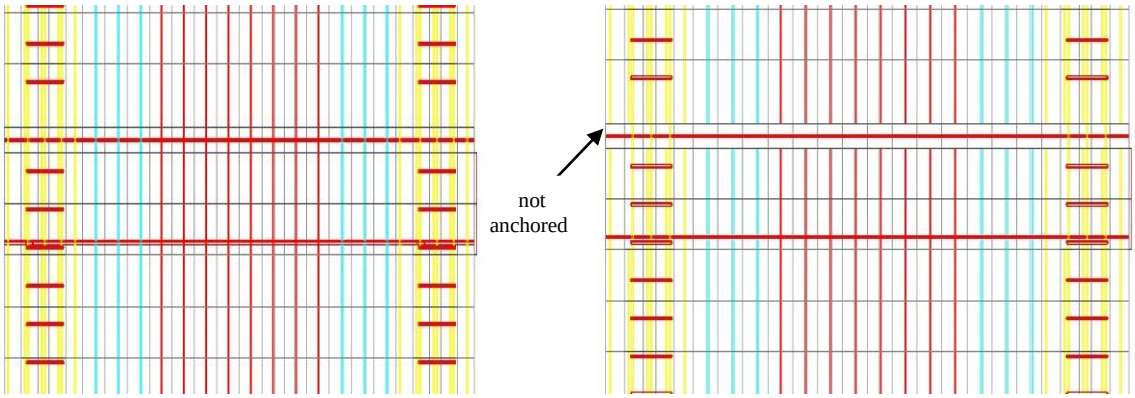
So as to further investigate the resulted carrying capacity of the shear wall, 5-additional models were developed that foresaw different material properties. The effect of the proper anchorage of the retrofitting rebars was also investigated through this study. Table 6 presents the details of each model that was developed, while Figure 8 illustrates the well anchored reinforcement (Figure 11a) and not anchored reinforcement (Figure 11b).

| Model | Retrofitting concrete (MPa) | Retrofitting Steel (MPa) | Embedded rebars anchored properly (Yes or No) | Numerically predicted Base Shear (kN) |
|---|---|---|---|---|
| **1** | 20 | 420 | Yes | 540 |
| **2** | 20 | 420 | No | 240 |
| **3** | 30 | 500 | No | 240 |
| **4** | 30 | 420 | Yes | 640 |
| **5** | 20 | 500 | Yes | 560 |
| **6** | 30 | 500 | Yes | 660 |

Table 6: Mesh and material details of the shear wall models

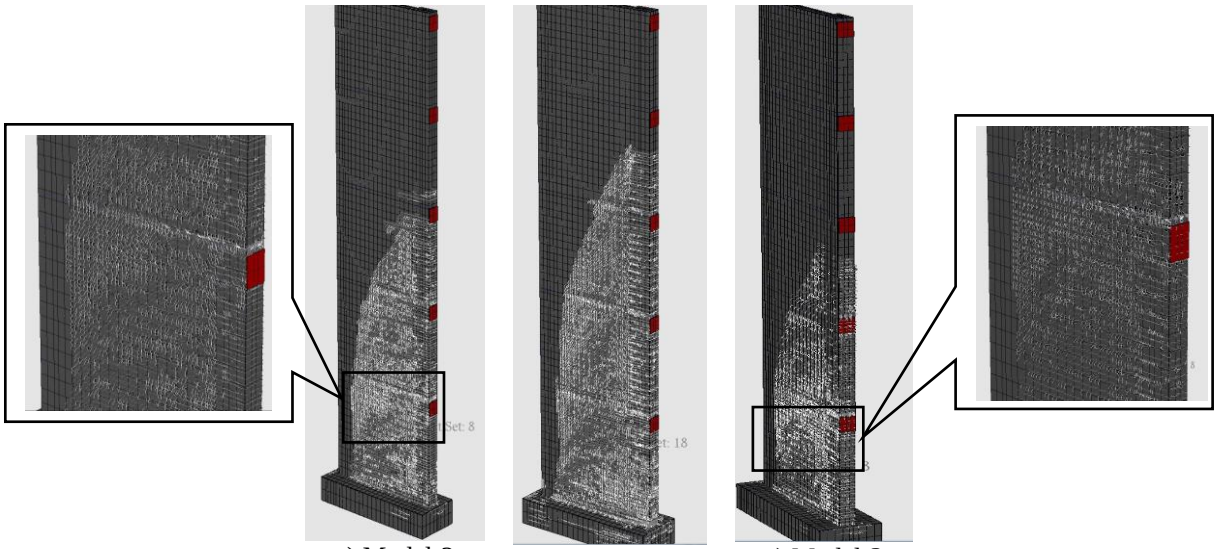
Figure 12 shows the crack patterns of the shear wall models prior to failure, where Figure 12b represents the crack pattern that was obtained from Model 1 (see also Figure 10), while Figure 12a and Figure 12c illustrate the crack patterns at load increment 8 (240kN failure load) of Models 2 and 3, respectively. As it resulted from the parametric investigation Models 2 and 3, exhibit a significantly decreased overall strength (see Table 6) due to the fact that the longi-

tudinal reinforcements were not anchored properly (throughout the height of the shear wall). Even in the case of Model 3 where the material properties were increased for both concrete and steel, the shear wall failed prematurely due to local strain concentration (see Figure 12c) that led to significant horizontal cracking (at the area where the longitudinal retrofitting reinforcement were not anchored properly).




a) Fully anchored longitudinal retrofitting rebars b) Not anchored longitudinal retrofitting rebars

Figure 11 Anchoring of retrofitted rebars



a) Model 2 b) Model 1 c) Model 3

Figure 12: Crack patterns of shear wall models prior to failure (Models 1,2 &3)

Moreover, model 4 that foresaw full anchorage, failed at load increment 16 (640kN), where the crack pattern is shown in Figure 13a. Model 5 (see Figure 13b) managed to carry 560kN, which was attributed to the increased steel yielding strength assumed in comparison to Model 1 (540kN). Figure 13c shows Model 6 that failed for a load of 660kN. This last model resulted the highest capacity given that it foresaw both increased retrofitting material properties and full anchorage of the retrofitting rebars. Figure 14 present the comparison of all numerically obtained P-δ curves.

In order to further investigate the performance of the retrofitted structural member (shear wall), a cyclic analysis was performed by using the shear wall model (Model 1). A history displacement was assigned at the top of the shear wall that foresaw the application of 5 com-

plete cycles (see Figure 15). A displacement of 75 mm was assigned at the top section of the shear wall which was then multiplied by the displacement factor shown in Figure 15.

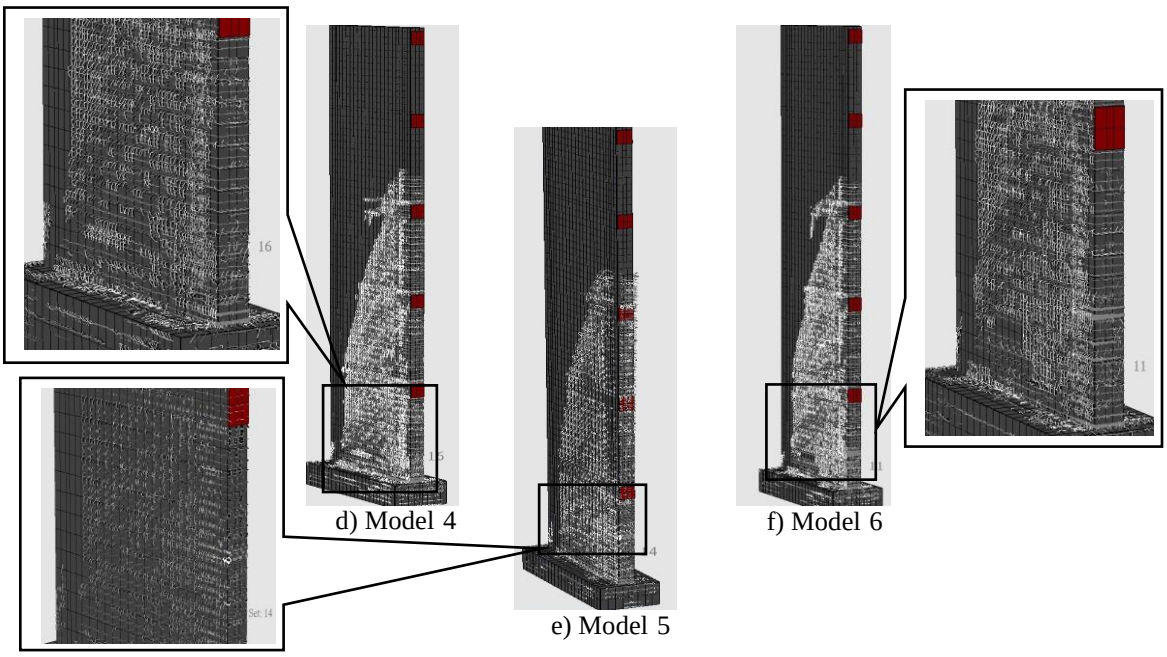



Figure 13: Crack patterns of shear wall models prior to failure (Models 4, 5 &6)

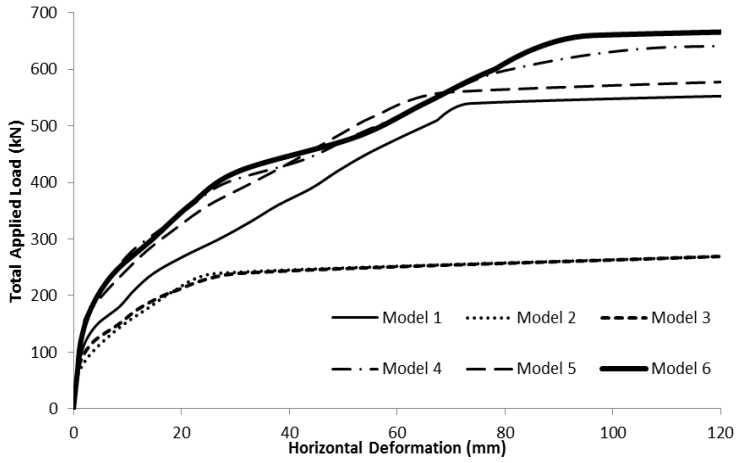



Figure 14: Comparison of shear wall models P-δ curves

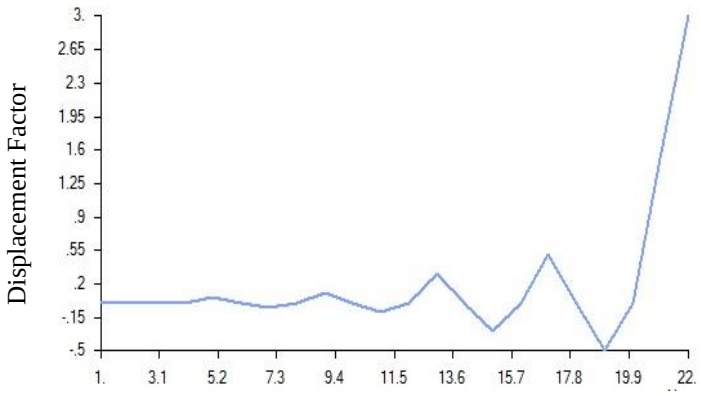
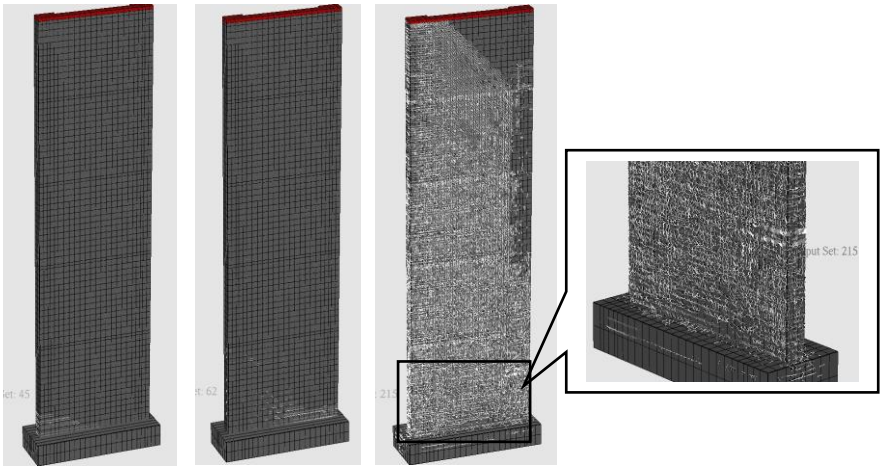



Figure 15: Displacement history curve

Figure 16 illustrates the crack patterns as they resulted from the numerical analysis. As it can be seen, opening (Figure 16a) and closing (Figure 16b) of cracks occurs according to the imposed displacements along the positive and negative x-axis. The overall hysteretic behavior is shown in Figure 17 where it is compared with the pushover analysis curve. As it derives the cyclic material model used to simulate concrete behaves in a softer manner compared to the monotonic, while managed to predict the same failure load (540kN). The opening and closing of cracks [11] cause the material to develop further deterioration thus derive a softer behavior. The full-scale 5-storey RC shear wall model required 24 hours to be analyzed for 216 load increments with an average of 10 internal iterations per load increment. This underlines the robustness of the developed algorithm [11] while the solution procedure managed to successfully converge for all load increments (the convergence criterion was set to $10^{-4}$).



a) Cracks opening b) Cracks closure c) Failure

Figure 16: Crack patterns at different load step as they resulted from the cyclic analysis

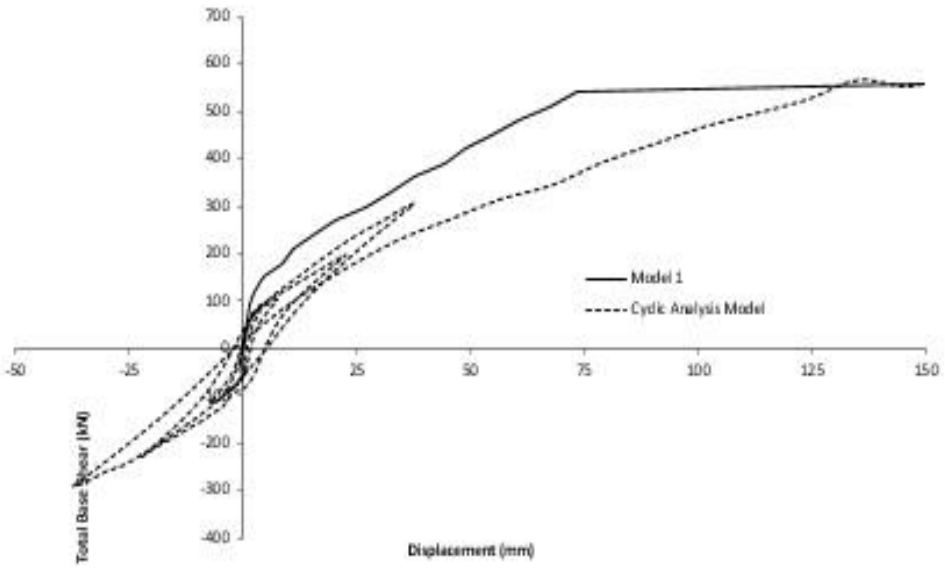



Figure 17: P-δ curves. Cyclic and monotonic analyses

## 6 CONCLUSIONS

The research work presented in this paper foresaw the numerical investigation of the mechanical behavior of retrofitted RC structures through the use of 3D detailed analyses. The full-scale model of a 5-storey RC building was developed prior and after retrofitting, while individual shear wall elements that were constructed to strengthen the structure (infill RC walls), were numerically investigated in an attempt to study the mechanical behavior and their overall effect on the retrofitted structure's response under seismic loads.

According to the numerical findings, the proposed strengthening design (addition of shear walls) significantly increases the seismic capacity of the structure. A single shear wall is capable in carrying the entire demand for seismic loads, while the overall increase of the seismic resistance of the structure due to retrofitting was found to be equal to 333%. The parametric investigation of the shear wall that derived after retrofitting the existing framing system, revealed the importance of proper implementation of the strengthening elements and connecting them rigidly to the existing frame. It was found that when the longitudinal rebars are not anchored properly within the existing framing system, the shear wall's overall capacity can be decreased more than 50%, leading to non-cost effective implementation.

An additional finding that derived from this research work was the overestimation of the carrying capacity of the shear wall (20%) that was computed when using the Eurocode formulae. The complexity of the derived strengthened shear wall's section that combines different types of materials and the irregular shape of the section itself, causes the development of a complicated deformation shape (bending and torsion in 3D) even for the case of symmetric horizontal loading. Evidently the proposed modeling approach will pave the way towards the development of improved guidelines and cost effective design.

From the analysis of the complete model with the initial framing geometry, it was found that it was not capable in carrying the seismic load demand while the computational robustness of the developed algorithm in handling an approximately half a million degrees of freedom numerical problem was illustrated. The computationally efficiency was found to require the implementation of a parallel solver which is a research work under process, while the effect of the interface between the existing and new concrete materials will be investigated in future projects (the detachment of the retrofitting concrete will decrease the overall strength of the structural members especially when excessive cracking will initiate). Finally, a future objective will be to perform a full-scale 5-storey RC building cyclic analysis.

## REFERENCES

[1] N. Gupta, P. Dhiman and A. Dhiman, "Design and Detailing of RC Jacketing for Concrete Columns," *IOSR Journal of Mechanical and Civil Engineering,* pp. 54-58.

[2] Handbook On Sesimic Retrofit of Buildings, Central Power Works Department and Indian Building Congress in association with IIT Madras, April, 2007.

[3] H. Kaplan and S. Yılmaz, Earthquake-Resistant Structures – Design, Assessment and Rehabilitation, A. Moustafa, Ed., Rijeka: InTech, 2012.

[4] R. Zienkiewicz, "The Finite Element Method: Its Basis and Fundamentals", 6th ed., Amsterdam: Elsevier Inc, 2005.

[5] G. Markou, "ReConAn v1.00 Finite Element Analysis Software User's Manual", Athens: ISAAR, 2011.

[6] E. Orakdogen, K. Girgin, M. Baguroglu, B. Buyksisli and T. Gokce, "Performance Evaluation of a Strengthened Building Considering the Soil-Structure Interaction," *Journal of Earthquake Engineering,* 12(S2), pp. 222-233, 2008.

[7] G. Markou, R. Sabouni, F. Suleiman and R. El-Chouli, "Full-Scale Modeling of the Soil-Structure Interaction Problem Through the use of Hybrid Models," *International Journal of Current Engineering and Technology,* 5, pp. 2277-4106, 2015.

[8] M. Kotsovos. and M. Pavlovic, Structural Concrete: Finite-element analysis for limit-state design, London: Thomas Telford, 1995.

[9] G. Markou, M. Papadrakakis, "Accurate and computationally efficient 3D finite element modeling of RC structures", *Computers as Concrete*, 12(4), pp. 443-498, 2013.

[10] EN1992. Eurocode 2: Design of concrete structures, 2004.

[11] Ch. Mourlas, M. Papadrakakis and G. Markou, "Accurate and Efficient Modeling for the Cyclic Behavior of RC Structural Members", ECCOMAS Congress, Crete Island, Greece, 5–10 June, 2016.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# VECTOR FIELD GUIDED TOOL PATHS FOR FIVE-AXIS MACHINING

**Stanislav S. Makhanov**
School of Information and Computer Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand
e-mail: makhanov@siit.tu.ac.th

**Keywords:** kinematics of the milling machines, error minimization, toolpath planning

**Abstract.** *We propose a new framework for toolpath generation for 5 axis machining of part surfaces represented by the StereoLithography (STL) format. The framework is based on flattening the STL part and generation of curvilinear toolpaths adapted to the vector field of optimal directions. The experiments show that there is no universal sequence of steps applicable to every surface. However, a correct choice of the tools available within the proposed ACT-framework always leads to a substantial improvement of the toolpath, in terms of its length and the machining time.*

## 1 INTRODUCTION

A milling machine is a programmable mechanism, guided by a controller fed with a CNC program (G-code) . The G-code is derived from a toolpath $\Omega = \{\Omega_0, \Omega_1, ..., \Omega_m\}$, consisting of the pairs $\Omega_p = \{M_p, I_p\}$, where $M_p = (x_p, y_p, z_p)$ is the Cartesian coordinates of the cutter location (CL) points in the machine coordinate system, and $I_p = \{I_{x,p}, I_{y,p}, I_{z,p}\}$ are the tool orientation vectors. The rotation angles, $\Re_p = (a_p, b_p)$ to control the rotary and the tilt table of the machine, are functions of the tool orientation vectors. Therefore, the toolpath can also be defined by $\Pi_M = \{\Pi_{M,0}, \Pi_{M,1}, ..., \Pi_{M,m}\}$, where $\Pi_{M,p} = \{M_p, \Re_p\}$. The Cartesian coordinates in the machine coordinates are usually derived from the corresponding toolpath in the workpiece coordinates, given by $\Pi_W = \{\Pi_{W,0}, \Pi_{W,1}, ..., \Pi_{W,m}\}$, $\Pi_{W,p} = \{W_p, \Re_p\}$, $W_p = (X_p, Y_p, Z_p) \in S(u,v)$, where $S(u,v)$ denotes the part surface and *u*,*v* the parametric coordinates. The configuration of the 5 axis milling machine is characterized by the relative positions of the rotary and linear axes, with regard to the tool and the workpiece. On a very basic level the machines can be categorized by the positions of the rotational joints as follows:

The 2-0 machine. Two rotary axes on the table.
The 1-1 machine. One rotary axis on the table and one on the tool.
The 0-2 machine. Two rotary axes on the tool.

Figures 1 and 2 display the configuration 2-0 used in this study.

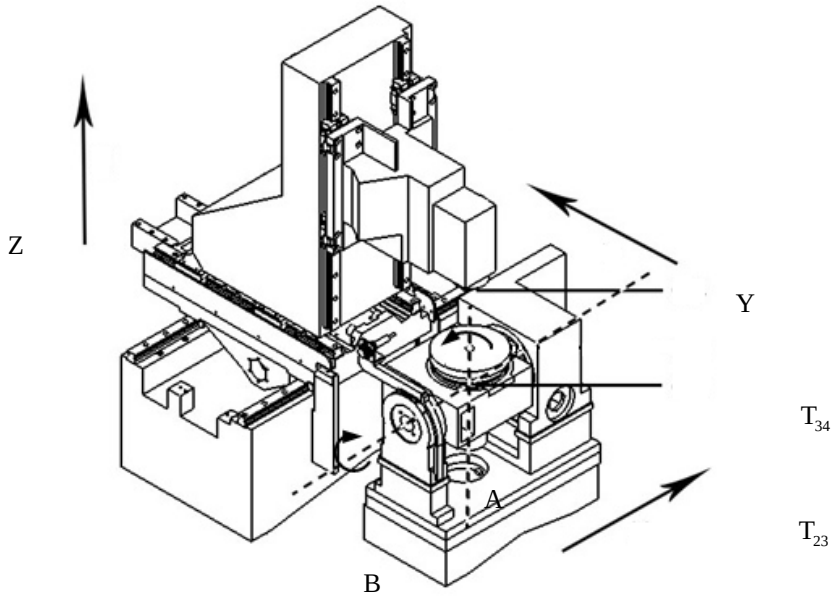



Figure 1. Five axis milling machine with rotary axes on the table



Figure 2. HAAS VF2TR

The toolpath optimization usually aims to reduce the total length of the toolpath or the total machining time, maintaining the prescribed accuracy. Alternatively, the user may wish to improve the accuracy while keeping or even reducing the machining time [1,2].

Furthermore, we differentiate the toolpath in the machine and the workpiece coordinates. In particular, our techniques account for the fact that minimization of the toolpath in the workpiece coordinates does not necessarily minimize it in the machine coordinates and does not necessarily minimize the machining time. However, optimization in the workpiece coordinates is machine independent and computationally easy. That is why the strategy to minimize the toolpath is often based on a machine independent criterion such as maximization of the machining strip [3], minimization of the scallops, etc. However, the reduction of the kinematic error and the machining time are the ultimate goals of 5–axis optimization [4-8].

Let $C$ be the quality criterion, $\varepsilon$ the difference between the actual and the desired tool trajectory, $h$ the scallop height, and $\varepsilon'$ the prescribed accuracy. The basic quality criterion is the length of the toolpath in the workpiece/machine coordinates and the machining time. Our toolpath generation employs a direct estimate of the basic criterion. It may also use another measure, which includes that criterion implicitly. For instance, instead of minimization of the length $C = L$ one can consider $C' = -R_M$ where $R_M$ is the material removal rate or $C' = \varepsilon$, where $\varepsilon$ is the kinematic error.

We also consider a quality criterion measure represented by a vector-function. At every point $W_p = (X_p, Y_p, Z_p)$, we evaluate an optimal direction (vector), using one of the above scalar criteria. The collection of these vectors constitutes the target vector field denoted by $V$.

The optimization problem is then given by

$$\underset{\Pi_W}{\text{minimize}}\, \text{dist}(V, V_W)$$

subject to

$$\varepsilon_p \le \varepsilon',$$
$$h_p \le \varepsilon',$$

where dist denotes an appropriate distance, and $V_W = \bigcup_{p=0}^{m-1} W_p \overrightarrow{W}_{p+1}$ is the desired vector field.

Although $V$ and $C'$ are not related explicitly, we will show that the solution of the above optimization problem minimizes or at least reduces $C'$ and consequently $C$.

Originally, the toolpath relative to a certain vector field was proposed in [9,10] and iso-scallop toolpaths in [1]. In [3], the part surface is covered by potential machining patches characterized by the optimal directions, corresponding to the maximum machining strip. The vector field of the optimal tool directions, to capture the "skeletal" information of toolpath (or a family of a toolpaths), was combined with the geometric constraints, and the kinematic performance of the machine and other constraints were evaluated such as the cutting force limits in [11,12]. The part surface was partitioned into clusters so that the streamlines of the vector field are close to the conventional isoparametric (ISO) or spiral in [13,14]. Optimal or nearly optimal directions can be combined with rear gouging, global gouging, and machine limits constraints [15]. Recently a machine-dependent vector field has been proposed in [43]. However, all the above algorithms construct the toolpath by offsetting an initial path and propagating the offsets inside the region. When the offset deviates from the optimal directions the algorithm requires a new initial track. Such strategies are usually sensitive to small variations of the optimization criterion. Moreover, finding the initial toolpath is a computationally expensive, NP hard problem. Additional efforts must be made to ensure that the resulting toolpath is structured and not self-intersecting.

In [16], we construct a curvilinear toolpath aligned with the optimal vector field using the elliptic grid generation. The optimal directions are evaluated using the maximum material removal rate. The proposed vector field aligned path (ACT) is a compromise between the flexibility of methods based on propagating of the initial track and the simplicity of the ISO topology. The method extends scalar based grid generation for five-axis machining originally proposed in [17,14,18-21] and independently in [22].

The flexibility of grid generation methods makes it possible to use a variety of cost functions and boundary adapted grids and domain decomposition methods. Although these options are well known in computational hydro and aerodynamics, they have not been fully exploited in the framework of five-axis machining. Furthermore, the toolpath can be further improved by applying the space filling curves (SFC) [23,24] and the biased space filling curves (BSFC) [16].
An important practical application of the toolpath generation methods is machining surfaces represented by the STL format. In order to apply the ACT to the STL surfaces, we have to either adapt the algorithms to the 3D case or flatten the STL surface and generate the required curvilinear grid in the resulting parametric domain. Noting that meshing algorithms

in 2D are more efficient than in 3D [25] , we incorporate a “radial plane” flattening approach into the proposed framework (see the forthcoming section Flattening Algorithm).
Our methodology includes the following components:

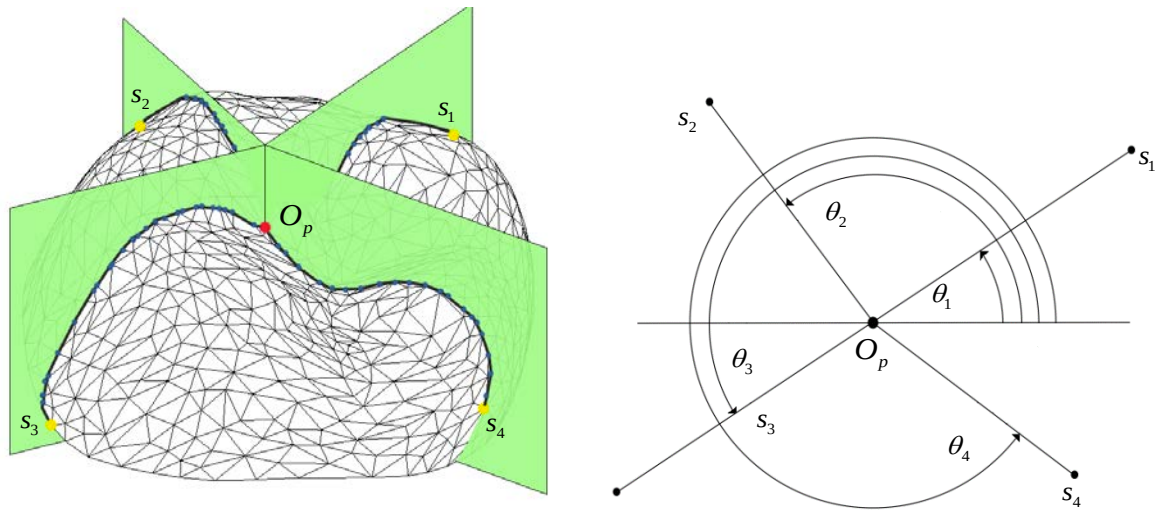
1. Flattening.
2. Evaluation of the surface characteristics in the parametric domain.
3. Grid generation in the parametric domain.
   3.1.1 Boundary fitted grids.
   3.1.2 Different types of grids, such as the C-grid, O-grid, H-grid [26].
   3.1.3 Grids adapted to control functions: scalar functions, vector fields.
4. Space filling curves
   4.1 Adaptive space filling curves.
   4.2 Biased space filling curves.

The methodology has been tested on a variety of STL surfaces, including surfaces with multiple ridges, human masks, as well as dental crowns or even tooth implants represented by the STL files. In particular, we demonstrate the efficiency of the algorithms by machining all types of human teeth: molars, premolars, canines, and incisors. The proposed tools have been tested against the standard ISO path, MasterCam generated path, and advanced toolpath generation methods of NX9. The tools available within the ACT-framework are capable of substantially improving the quality of the toolpath, in terms of its length, the machining time and the kinematic error.

## 2 FLATTENING ALGORITHM

The majority of toolpath generation techniques have been demonstrated on a single parametric surface, such as the Bezier patch, Gregory patch, B-spline, etc. However, the explicit parametric representation is rarely used in the industry. The most popular formats are STL [27], IGES [28], STEP [29], SET [30], CATIA [31], and PHIGS [32]. The STL format, which approximates the surface of a solid model by triangles, is becoming the standard for the rapid prototyping industry and is increasingly popular in multi-axis manufacturing. The main reason is its simplicity and applicability to an arbitrary surface. The algorithms for flattening triangulated or tessellated surfaces have been used in many research areas, such as generation of clothing patterns, 3D reconstruction, texture mapping reconstruction, multiresolutional analysis, formation of ship hulls, metal forming, etc. By flattening, we understand a one-to-one mapping of the STL surface to a planar domain in the parametric space. Geometrically, the flattening algorithm unfolds the surface onto a plane without self-intersections, that is, each point of the original surface is associated with a unique point $(u,v)$ in the parametric space. A variety of algorithms have been proposed for parameterization of the triangulated surfaces[33]-[39]. In this paper we propose radial plane flattening and demonstrate that it works for a particular practical case of STL surfaces representing human teeth. The method was also successfully tested on parts which can be roughly approximated by surfaces having radial symmetry. Note that this approximation could be very far from ideal, however, radial flattening still works better than some general commercial software. Besides, the proposed algorithm preserves distances along the trajectories in the radial planes and reduces angular and metric distortions in the proximity of the focus point [40].
We define a polar coordinate system centered at a point of reference $O_p$. Given the surface’s tangential plane at the reference point, a perpendicular plane through the reference point, and an arbitrary STL point $S'$, we evaluate the trajectory connecting $O_p$ and $S'$ on

the surface by finding intersections of the radial plane and the corresponding STL triangles (Figure 3).

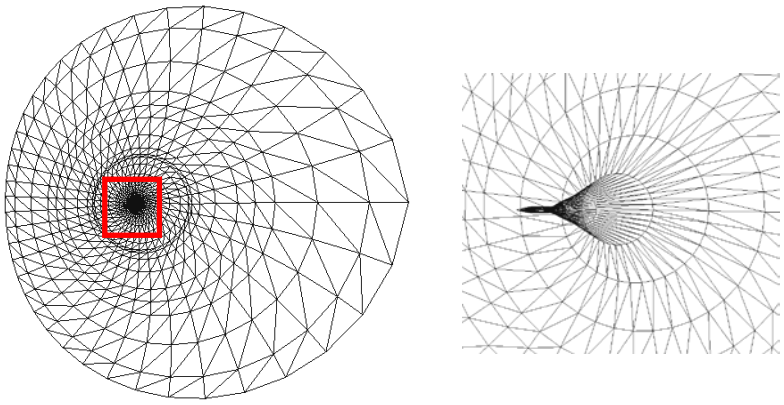



(a) (b)

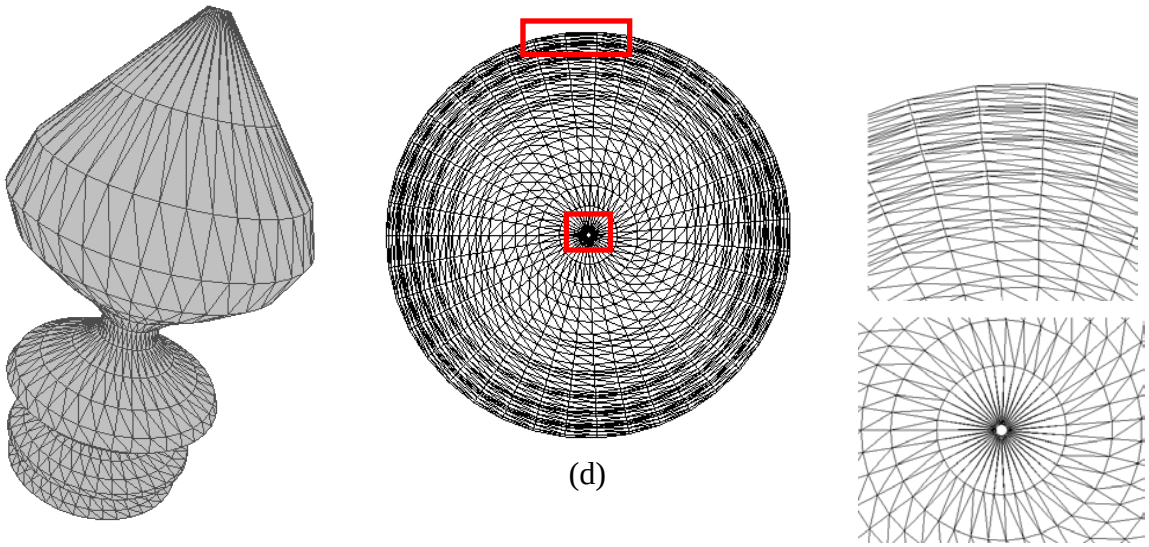
Figure 3. Radial plane flattening algorithm

The technique requires solving a large series of linear (possibly singular) systems, 2x2 and 3x3, and a special tracing procedure to connect the intersection points.

The new coordinates $(s,\theta)$ are the length of the trajectory $O_p\,S'$ and the angle between the radial plane and the reference plane, respectively (Figure 3). Note that the radial flattening does not necessarily mean that the curvilinear grid must be polar. As a matter of fact, we map the resulting triangles onto the Cartesian plane. However, using a polar parametrization and the spiral toolpath is a feasible option which can be implemented by the ACT as well. Figure 4 compares the proposed algorithm with MeshFlatten [41].



(b) (c)



(a) (d) (e)

Figure 4. Flattening (a) 3D Model (b)-(c) MeshFlatten, (d)-(e) radial plane flattening

Clearly, MeshFlatten fails to return an acceptable parametrization, whereas a correct choice of the reference point $O_p$ allows for construction of the required flattening using the proposed approach.

## 3 MATERIAL REMOVAL RATE

Let $W_1$ be an arbitrary cutter contact (CC) point on the surface (see [7] regarding the relationship between the CC and CL points). Consider a disk defined by $\Omega_{W_1} = \{W : \mathrm{dist}(W_1, W) = l_1\}$, where dist denotes an appropriate distance on the surface, and $l_1$ is an arbitrary step (see Figure 5). The corresponding disk in the machine coordinates is denoted by $\Omega_{M_1}$. The distance between the center of the disk $M_1$ and the points belonging to $\Omega_{W_1}$ is given by $l_{1,M} \equiv l_{1,M}(W)$.

The machining strip along the feed direction $\vec{W_1, W}$ is denoted by $w_1 \equiv w_1(W)$. It is evaluated by locating the intersections of the effective cutting shape and the part surface [24].

Our target vector at $W_1$ is defined as $\vec{W_1, W_2}$, where $W_2 = \arg\max_{W \in \Omega_{W_1}} C_1'(W)$, and where $C_1'(W) = w_1(W)$, $C_1'(W) = l_{1,M}(W)$ or $C_1'(W) = R_{1,M}(W)$, where $R_{1,M}(W)$ denotes the material removal rate (MRR) . In order to define the MRR, assume that only the rotation axes can exceed the machine speed limit and that the workpiece is machined with the maximum linear speed allowed for the particular material. These assumptions are usually correct for relatively small parts, exemplified in our forthcoming section “Numerical Examples and Cutting Experiments”, where the largest blank workpiece is 100x100x100mm. In this case the required machining time is evaluated by

$t = \max(t_a, t_b)$,

where $t_a = \frac{a - a_1}{v_{a,\max}}$, $t_b = \frac{b - b_1}{v_{b,\max}}$ and, where $v_{a,\max}, v_{b,\max}$ are the maximal speeds of the rotational axes $a$ and $b$, respectively.
The actual rotation speed is given by

$$v_a = \frac{a - a_1}{t},\ v_b = \frac{b - b_1}{t}.$$

Note that if $a = a_1$ and $b \neq b_1$, then $v_a = 0$. If $b = b_1$ and $a \neq a_1$, then $v_b = 0$.
If $a = a_1$ and $b = b_1$, then the maximum linear speed is allowed in each of the linear axis. Consequently, the feed rate $F = \frac{l_1}{t}$, whereas the effective MRR is evaluated by $R_M = Fw_1(W)$, where $w_1(W)$ is the corresponding machining strip.

We call the direction $W_1, \vec{W}_2$ optimal, if $W_2$ maximizes the cost function $C_1'$. Evaluating vectors $W_1, \vec{W}_2$ for each surface point and transferring them into the parametric domain $(u, v)$ generates a discrete version of the optimal vector field, $V(u,v) \equiv (v_x(u,v), v_y(u,v))$. Following $V(u,v)$ at every CC point generates an optimal tool path. However, constructing such a tool path is not always possible. Therefore, the next section presents a vector field aligned path which is nearly optimal on a set of paths topologically similar to a Cartesian zigzag.

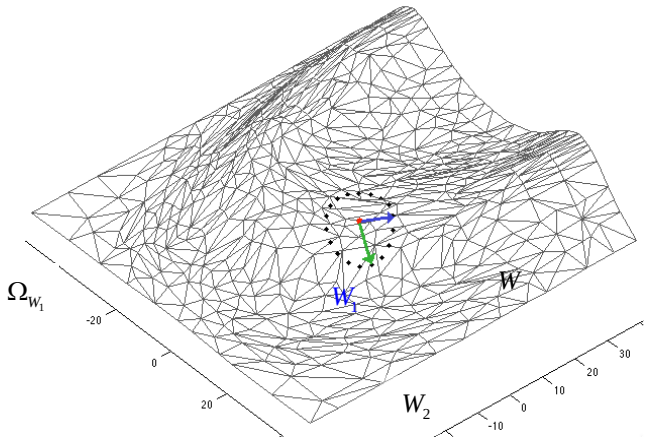



Figure 5. Material removal rate: possible directions in the workpiece coordinates

## 4 CURVILINEAR TOOL PATH

Recall that $S(u,v) \equiv (x(u,v), y(u,v), z(u,v))$ is the part surface, where $u$ and $v$ are the parametric variables. Let us consider the CC points $\{(u_{i,j}, v_{i,j}),\ 0 \le i \le N_\xi\ \ 0 \le j \le N_\eta\}$ arranged as a curvilinear grid. In other words, the CC points are a discrete analogy of a mapping $\{u(\xi,\eta), v(\xi,\eta)\}$ from the computational region $\Delta = \{0 \le \xi \le N_\xi,\ 0 \le \eta \le N_\eta\}$ onto a physical region $K$ in the parametric coordinates $u$, $v$. In other words, the rectangular grid $\{(i, j), 0 \le i \le N_\xi\ \ 0 \le j \le N_\eta\}$ being fed to $\{u(\xi,\eta), v(\xi,\eta)\}$ becomes $\{u_{i,j}, v_{i,j}\}$ (see Figure 6).

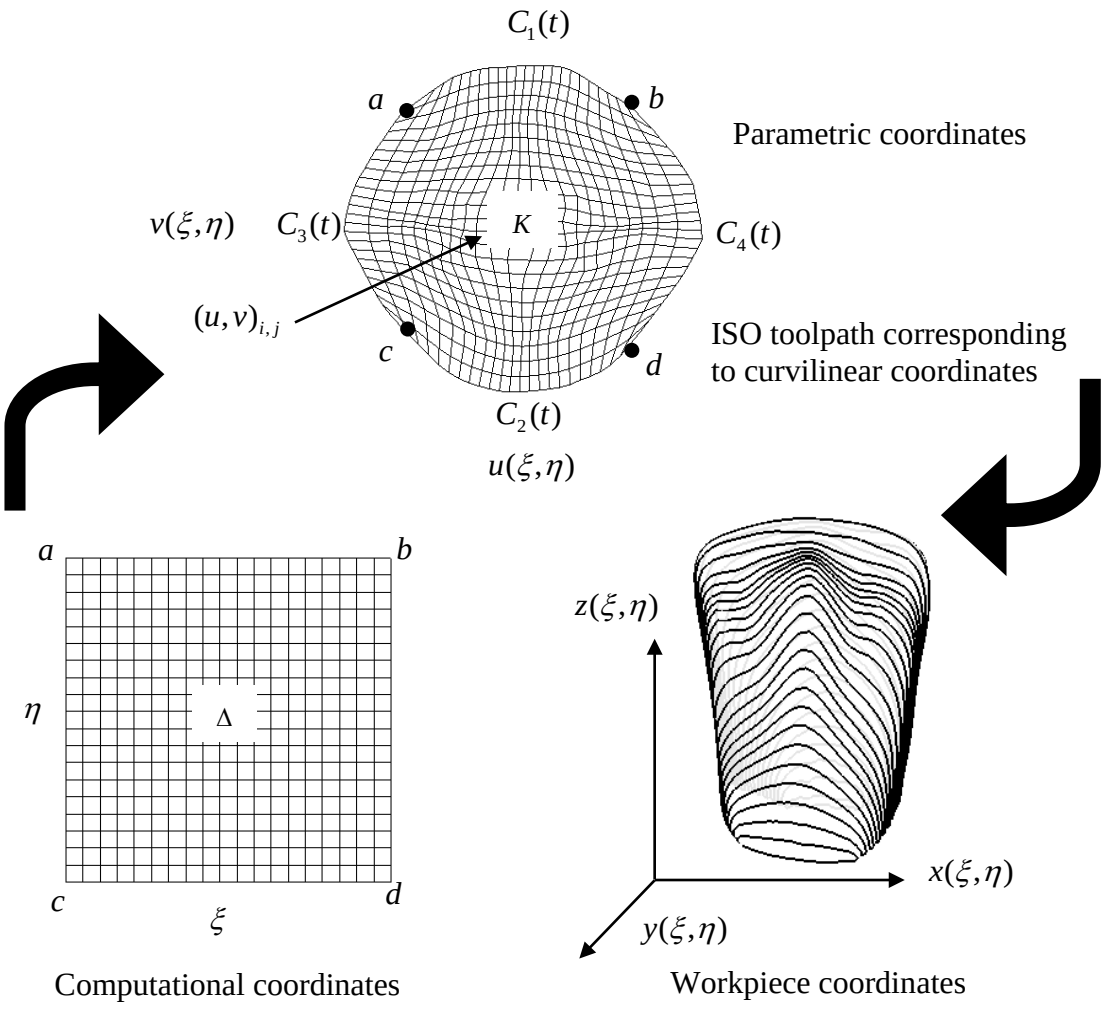



Figure 6. Coordinate transformations and the curvilinear grids, $\Delta$ denotes the computational domain, $K$ the parametric domain

We adapt the curvilinear grid to the boundary of the parametric domain obtained by flattening. Furthermore, we consider adaptation to the vector field $V(u,v)$ using a modification of classical grid generation methods [20] as follows:

The smoothness of the grid is controlled by a variational functional given by

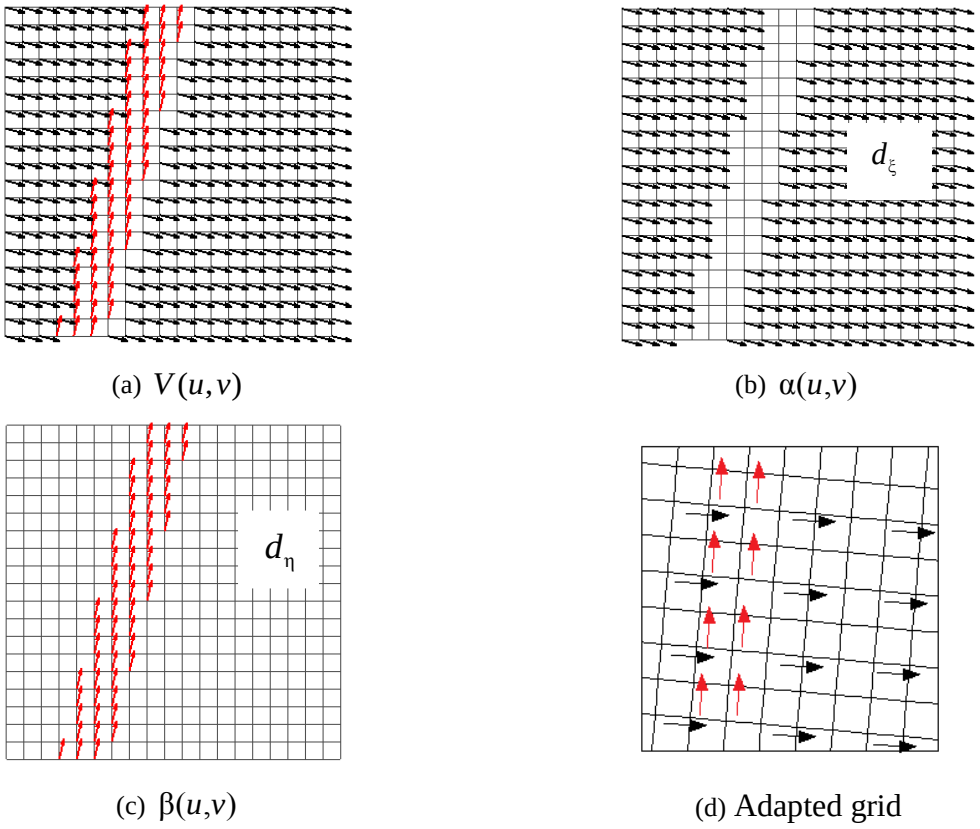
$$F_S = \int\int u_\xi^2 + u_\eta^2 + v_\xi^2 + v_\eta^2 \mathrm{d}\xi\mathrm{d}\eta \ , \tag{1}$$

where subscripts denote the partial derivatives.

Furthermore, the idea of a curvilinear grid adapted to the vector field is borrowed from computational fluid dynamics where such grids were exploited to improve the accuracy of the numerical solution of partial differential equations [44-46]. Consider a vector field $V(u,v)$ derived from a cost function $C'$ (machining strip, length of the toolpath, material removal rate, see the previous section). Furthermore, we partition $V(u,v)$ into a dual vector field $V_D(u,v) = (\alpha(u,v),\beta(u,v))$, corresponding to the $\xi$ and $\eta$ directions as follows:

$$\alpha(u,v) = \begin{cases} V(u,v) \in \Omega_\xi, \\ 0, \text{otherwise}, \end{cases} \quad \beta(u,v) = \begin{cases} V(u,v) \in \Omega_\eta, \\ 0, \text{otherwise}, \end{cases} \tag{2}$$

where $\Omega_\xi, \Omega_\eta$ are prescribed subsets of $V(u,v)$ selected according to a certain criterion. For instance, if the vector field has two major directions $d_\xi$ and $d_\eta$, the partition is performed as follows (Figure 7).




(a) $V(u,v)$ (b) $\alpha(u,v)$

(c) $\beta(u,v)$ (d) Adapted grid

Figure 7. Partition of the vector field $V(u,v)$ into $(\alpha(u,v),\beta(u,v))$

$$\alpha(u,v)=\begin{cases}\measuredangle(V,d_\xi)\approx 0 \text{ or } \measuredangle(V,d_\xi)\approx\pi,\\ 0, \quad \text{otherwise,}\end{cases}, \quad \beta(u,v)=\begin{cases}\measuredangle(V,d_\eta)\approx 0 \text{ or } \measuredangle(V,d_\eta)\approx\pi\\ 0, \quad \text{otherwise.}\end{cases}$$

The vectors from $V(u,v)$ are included into the $V_D(u,v)$ if they are almost parallel or almost anti-parallel to $d_\xi$ or $d_\eta$. In other words, $\left|\frac{(V,d)}{\|V\|\,\|d\|}\right|-1\le\varepsilon_V$, where $\varepsilon_V$ is the prescribed threshold.

The curvilinear grid is aligned with $V_D(u,v)$, using a modification of the grid generation techniques [44-46]. Let us align the family of the grid lines $\eta$=const with the vector field $\alpha(\xi,\eta)\equiv(\alpha_1(\xi,\eta),\alpha_2(\xi,\eta))$, whereas the grid lines $\xi$=const are aligned with $\beta(\xi,\eta)\equiv(\beta_1(\xi,\eta),\beta_2(\xi,\eta))$. The alignment is performed by minimizing a functional given by

$$F_V=\int\int(s_\xi\alpha')^2+(s_\eta\beta')^2\,d\xi d\eta, \tag{3}$$

where $\alpha'(\xi,\eta) \equiv (\alpha_1', \alpha_2') = (-\alpha_2, \alpha_1)$ is the vector field perpendicular to $\alpha(\xi,\eta)$, $\beta'(\xi,\eta) \equiv (\beta_1', \beta_2') = (-\beta_2, \beta_1)$ is the vector field perpendicular to $\beta(\xi,\eta)$, $s_\xi = (u_\xi, v_\xi)$, and $s_\eta = (u_\eta, v_\eta)$

Following [44], we combine the functionals $F_S$ and $F_V$ linearly as follows $\Phi = F_S + \lambda F_V$, where $\lambda$ is the weighting coefficient.

The Euler equations are

$$\Phi_u - \Phi_{\xi,u_\xi} - \Phi_{\eta,u_\eta} = 0,$$
$$\Phi_v - \Phi_{\xi,v_\xi} - \Phi_{\eta,v_\eta} = 0.$$

Substitution $s_\xi \alpha' = u_\xi \alpha_1' + v_\xi \alpha_2'$, $s_\eta \beta' = u_\eta \beta_1' + v_\eta \beta_2'$ and differentiation yields

$$\begin{aligned} u_{\xi\xi} + u_{\eta\eta} + 2\lambda\{[\alpha_1'(\alpha_1' u_\xi + \alpha_2' v_\xi)]_\xi + [\beta_1'(\beta_1' u_\eta + \beta_2' v_\eta)]_\eta = 0, \\ v_{\xi\xi} + v_{\eta\eta} + 2\lambda\{[\alpha_2'(\alpha_1' u_\xi + \alpha_2' v_\xi)]_\xi + [\beta_2'(\beta_1' u_\eta + \beta_2' v_\eta)]_\eta = 0. \end{aligned} \quad (4)$$

Furthermore, the problem can be simplified by following the optimal directions only where the alignment with the vector field leads to a significant reduction of the cost function. Therefore, the desired vector field can be replaced by

$$V_{new}(\xi,\eta) = \begin{cases} V(\xi,\eta) \text{ if } (\xi,\eta) \in \text{important points}, \\ (0,0), \text{ otherwise.} \end{cases}$$

For instance, the "important points" can be defined as points characterized by a high curvature or large rotation angles. At the unimportant points $F_V = 0$, which simplifies the numerical solution and reduces the number of iterations. The numerical approximation is based on the discrete Laplacian and the central differences for the first derivatives.

## 5 KINEMATIC ERROR

Let $W^D(s_p, s_{p+1}, t) \in S(u,v)$ be a desired trajectory between two tool positions $W_p$ and $W_{p+1}$, where $t \in [s_p, s_{p+1}]$ is a parametric coordinate along the curve. The kinematic error is a distance between $W^D(s_p, s_{p+1}, t)$, and the actual trajectories $W(s_p, s_{p+1}, t)$ are defined as follows:

$$\varepsilon = \sum_p \text{dist}(W^D(s_p, s_{p+1}, t), W(s_p, s_{p+1}, t)). \quad (5)$$

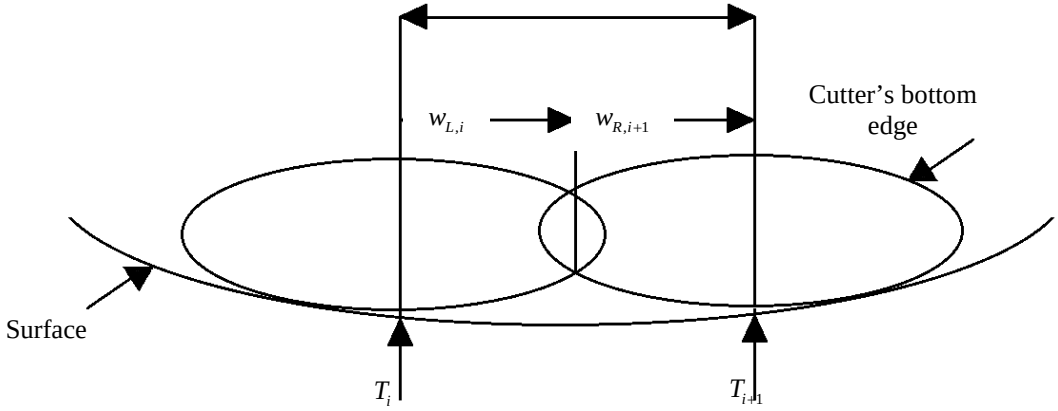
The actual trajectories are generated by the inverse kinematic transformations (see, for instance, [47], pages 40-42). Furthermore, we use a parameterization-invariant Hausdorff distance, given by

$\text{dist}_H(W, W^D) = \max\{\max_{a\in W(t)} \min_{b\in W^D(t)} |a-b|_E, \max_{b\in W^D(t)} \min_{a\in W(t)} |a-b|_E\}$, where the subscript $E$ denotes the Euclidean distance.

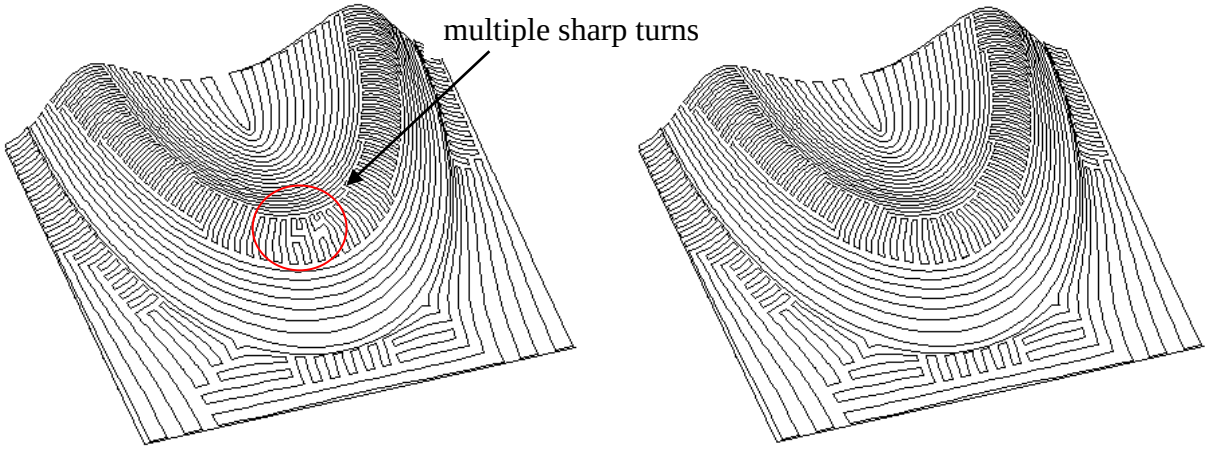
## 6. BIASED SPACE FILLING CURVE

First, the curvilinear grid $\{u_{i,j}, v_{i,j}\}$ is converted into a continuous mapping $\{u(\xi,\eta), v(\xi,\eta)\}$ using the bilinear interpolation. Next, the discrete grid is reconstructed using the condition $\mathrm{dist}_H(T_i, T_{i+1}) < w_{R,i} + w_{L,i+1}$, where $T_i$ $T_{i+1}$ are the adjacent tool tracks, and $w_{L,i}, w_{R,i+1}$ are the left and the right maximum allowed strip width (Figure 8 (a)).

The cells are connected by the biased adaptive space filling curve (BSFC) [16] so that the toolpath follows the optimal direction, where necessary. The BSFC uses the concept of important and non-important points as a compromise between reducing the machining time and the kinematic error. The idea is similar to that used for minimization of the alignment functional. If the grid is well aligned with the optimal direction, then the BSFC follows it. However, the BSFC works on the chaotic vector fields as well. If the grid is not well aligned with the important point, the BSFC defines a local optimal direction based on an estimate of the cost function around this particular point. If the point is unimportant, the BSFC connects it with a random neighboring point. Furthermore, the algorithm detects segments characterized by the multiple consecutive turns. The BSFC along this segment is re-adjusted to follow the main direction in a window around the segment as shown in Figures 8 (b) and (c).




(a) Calculating the offset between the tool tracks




(b) Before correction (c) After correction

Figure 8. BSFC correction in the case of frequent turns

## 7. NUMERICAL EXAMPLES AND CUTTING EXPERIMENTS

In this section, the ACT is applied to STL toolpath generation. The numerical technology has been tested against the standard ISO, MasterCam, advanced toolpath algorithms Helical/Spiral (HS), and Follow Periphery (FP) of NX9 (formerly UG). All surfaces have been machined using a flat end cutter. The experiments have been performed for 5 axis machines MAHO-600E and HAAS VF2TR.

The conventional methods and the ACT have been applied with an appropriate setup optimization [4]. The accuracy of the machining has been evaluated in terms of the kinematic error, roughness, and waviness. The improvement is defined as $\frac{before-after}{before}100$. The grid alignment is based on cost functions representing the material removal rate and the toolpath length. We will denote these options by subscripts *R* and *L*, respectively, for instance, $VFAG_R$ and $VFAG_L$.

### Example 1. Surface decomposition and ACT for a synthetic dental surface

The surface in Figure 9 (a) is characterized by a ridge nearby the boundary. We adopted this shape from the dental milling, where the elevated part represents the ridges of a dental crown or implant [48]. The size of the workpiece is 100x100. The surface was designed using MasterCam, then converted into the STL format (Figure 9 (a)), and flattened using the proposed radial plane approach (Figure 9 (b)).

The VFAG for such surfaces can be efficiently generated in subregions obtained automatically or manually. Clearly, the surface is composed of a flat region and a region characterized by a high curvature. The high curvature narrows the machining strip, whereas inside the flat area, the tool can cut the material using the maximum machining strip. The optimal directions corresponding to the material removal rate and machining strip width, respectively, in Figures 9 (c) and (d), show that it is possible to construct a single grid such that one family of the grid lines is adapted to $V_R$ and another to $V_L$.

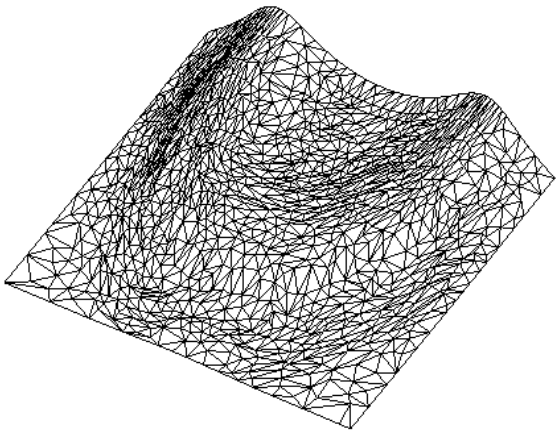
Consider the case when the user requires a minimal tool path for the rough cut to reduce the tool wear and the minimal time for the fine cut. The generated VFAG in Figure 9 (e) accomplishes both tasks. The rough machining is performed along the direction of the highest curvature, in order to maximize the machining strip ($ACT_L$). The finishing strategy employs $ACT_R$ to maximize the material removal rate and to reduce the machining time. In this case, most of the time the tool moves along the direction of the minimal curvature. Note that a boundary fitted grid in Figure 9 (f) was used as the initial approximation for VFAGs. The toolpaths, constructed by BSFC, are shown in Figures 9 (g) and (h). Clearly, the two cost functions generate very different toolpaths. The virtual and real machining results are illustrated in Figures 9 (i)-(l).

The performance of ACT, in terms of the length of the toolpath and the machining time, is summarized in Table 1. The table clearly demonstrates the advantages of the $ACT_L$ and $ACT_R$ for varying scallop height. The benchmark methods are the regular ISO, MasterCam, and advanced algorithms, HS and FP of NX9.
Compared to ISO, the length of the $ACT_L$-path is shorter by 7-20%. For instance, when the maximum allowable scallop height $h$=0.01, the length of the tool path is about 2,500 mm shorter while the entire ISO is approximately 12000mm. The machining time for the fine cut

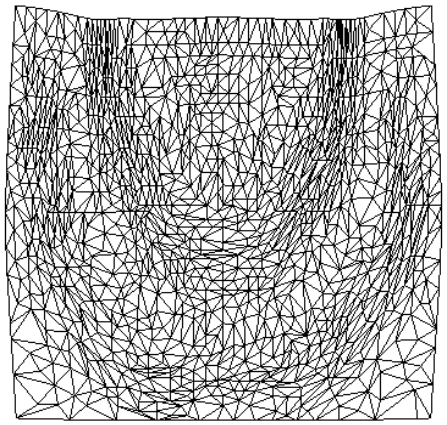
using $ACT_R$ has been reduced by 76%. In other words, the proposed method reduces the machining time by (approximately) a factor of 2. There is a clear advantage relative to NX9. For instance, on MAHO 600E for $h$=0.01, the $ACT_R$ requires 19 min versus 46 min using FP, and 9h52min using HS (see the entire evaluation in Table 1).

Table 2 compares the quality of the proposed method vs. ISO in terms of the average kinematic error $\bar{\varepsilon}$, the corresponding standard deviation $\sigma_\varepsilon$ and the number of the required CC points $N_{CC}$. Clearly, the error stays within the prescribed limits, whereas the number of the CC points has been reduced. For instance, for HAAS VF-2TR $N_{CC,ISO}=18603$, $N_{CC,R}=9625$, and $N_{CC,L}=17437$.

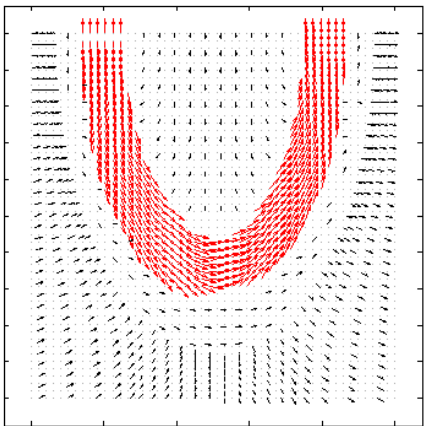
Table 3 shows the quality of the produced surface for the rough and finishing cuts evaluated by stylus profiling, [49]. The surface profile obtained by a high-resolution probe is post processed by the Gaussian filter [50]. Twenty sample profiles were measured with the standard cutoff of 0.8 mm [50]. Roughness of the machined surface is within the acceptable range for surface milling operations [51]. The rough cut by the $ACT_L$ has a better quality compared to ISO, that is, roughness: 7.8 vs. 8.1 μm and waviness 17.8 vs. 31.9. We hypothesize that the major improvement in waviness is due to a reduced variation of the rotation angles, which are implicitly included in the evaluation of the material removal rate, since maximization of the removal rate leads to the reduction in the angular distance between the CC points.

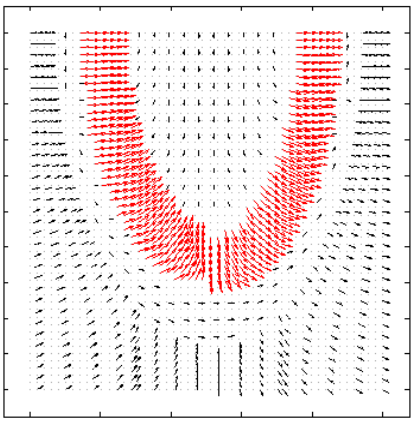


(a) Synthetic dental surface



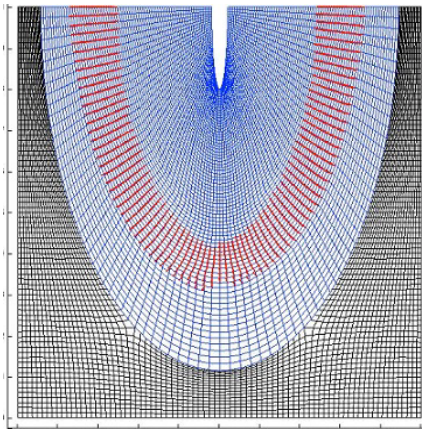
(b) Flattened surface

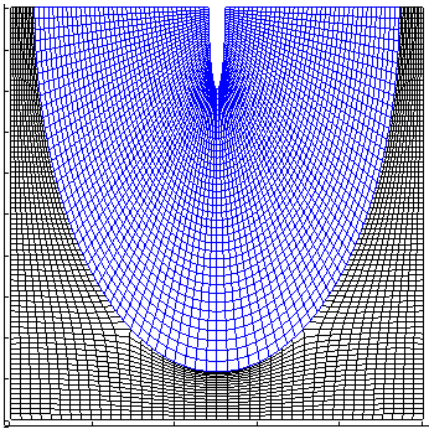




(c) $V_R$

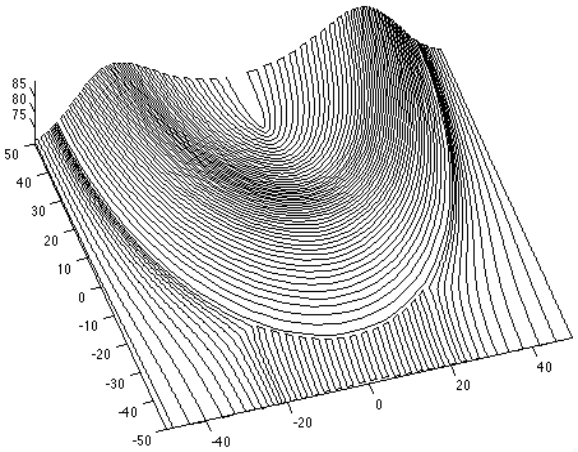
(d) Vector field $V_L$

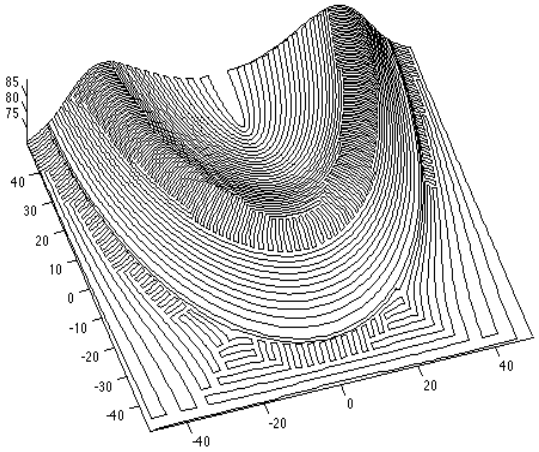




(e) VFAG

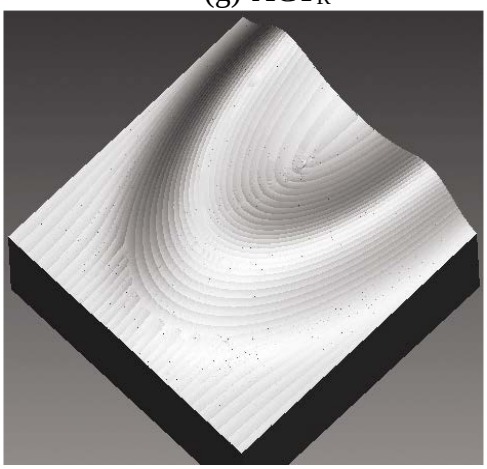
(f) Boundary fitted grid

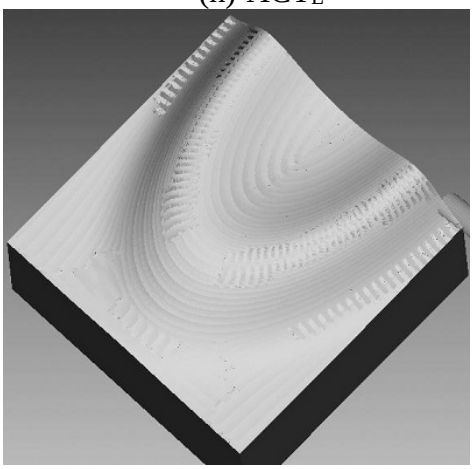



(g) $ACT_R$

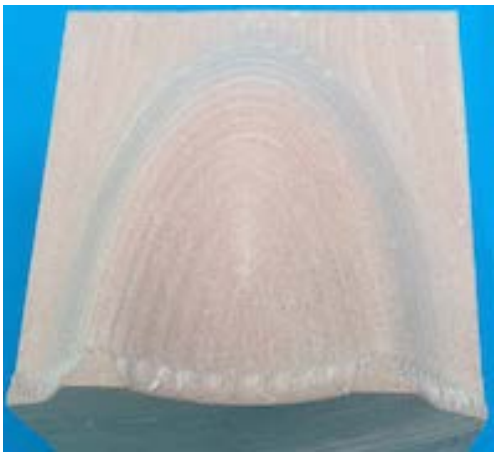
(h) $ACT_L$



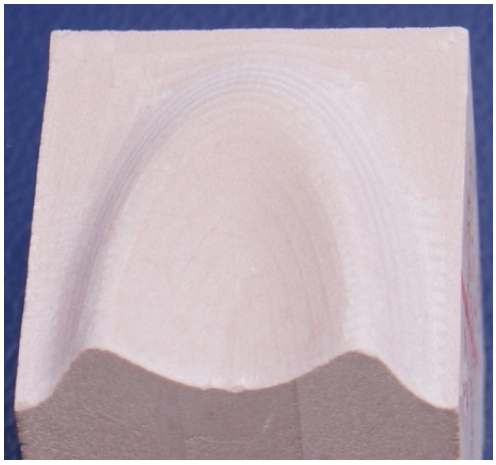


(i) Virtual $ACT_R$

(j) Virtual $ACT_L$



(k) Machining by $ACT_R$



(l) Machining by $ACT_L$

Figure 9. Synthetic (ridge-cusp) dental surface

Table 1 Synthetic dental surface: ACT vs. the reference methods

| | Toolpath Generation Methods | | | Scallop Height h (mm) | | |
|---|---|---|---|---|---|---|
| | | | | **0.25** | **0.10** | **0.05** |
| **Performance** | **Toolpath Length (mm)** | | **Master CAM** | 35553.0 | 72215.0 | 128879.0 |
| | | | **HS** | 25991.0 | 62825.0 | 125070.0 |
| | | | **FP** | 5480.4 | 8115.6 | 11175.0 |
| | | | **ISO** | 5755.2 | 8666.1 | 12034.0 |
| | | | **$ACT_R$** | 5489.1 | 7601.2 | 10075.0 |
| | | | **$ACT_L$** | 5336.9 | 7092.7 | 9545.7 |
| | **Machining Time** | **HAAS VF2TR** | **HS** | 3:26:00.6 | 8:13:54.1 | 16:22:50.4 |
| | | | **FP** | 0:45:04.7 | 1:07:20.4 | 1:33:35.8 |
| | | | **ISO** | 0:56:13.9 | 1:26:37.5 | 2:01:03.5 |
| | | | **$ACT_R$** | 0:17:46.1 | 0:24:56.1 | 0:33:36.1 |
| | | | **$ACT_L$** | 0:55:09.7 | 1:24:21.2 | 1:55:52.1 |
| | | **MAHO 600E** | **HS** | 2:03:53.1 | 4:57:21.7 | 9:52:27.2 |
| | | | **FP** | 0:22:15.4 | 0:33:09.6 | 0:46:22.0 |
| | | | **ISO** | 0:28:22.1 | 0:43:39.8 | 1:01:03.8 |
| | | | **$ACT_R$** | 0:10:25.4 | 0:14:21.7 | 0:19:22.9 |
| | | | **$ACT_L$** | 0:27:14.5 | 0:41:20.4 | 0:56:19.7 |
| **Advantage rel. to ISO** | **Toolpath Length (mm)** | | **$ACT_R$** | 4.6% | 12.3% | 16.3% |
| | | | **$ACT_L$** | 7.3% | 18.2% | **20.7%** |
| | **Machining Time** | **HAAS VF2TR** | **$ACT_R$** | 68.4% | 71.2% | **72.2%** |
| | | | **$ACT_L$** | 1.9% | 2.6% | 4.3% |
| | | **MAHO 600E** | **$ACT_R$** | 63.3% | 67.1% | **68.3%** |
| | | | **$ACT_L$** | 4.0% | 5.3% | 7.8% |

Table 2 Synthetic dental surface: CC points, kinematic error, ACT vs. ISO

| Toolpath Generation techniques | HAAS VF-2TR | | | | | | | | | MAHO 600-E | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | h=0.25 mm | | | h=0.10 mm | | | h=0.05 mm | | | h=0.25 mm | | | h=0.10 mm | | | h=0.05 mm | | |
| | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ |
| **ISO** | 5882 | 0.082 | 0.064 | 10299 | 0.040 | 0.024 | 18603 | 0.023 | 0.012 | 5047 | 0.064 | 0.063 | 9334 | 0.038 | 0.025 | 16652 | 0.022 | 0.012 |
| **$ACT_R$** | 5378 | 0.013 | 0.010 | 7157 | 0.013 | 0.008 | 9625 | 0.011 | 0.004 | 5374 | 0.009 | 0.004 | 7151 | 0.009 | 0.001 | 9331 | 0.009 | 0.001 |
| **$ACT_L$** | 4321 | 0.065 | 0.006 | 9418 | 0.025 | 0.026 | 17437 | 0.013 | 0.013 | 4059 | 0.054 | 0.002 | 9059 | 0.022 | 0.025 | 16773 | 0.012 | 0.013 |

Table 3 Synthetic dental surface: roughness and waviness, ACT vs. ISO

| Roughness $R_a$ ($\mu$m) | | | Standard Deviation | | | Waviness $R_a$ | | | Standard Deviation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ISO** | **$ACT_R$** | **$ACT_L$** | **ISO** | **$ACT_R$** | **$ACT_L$** | **ISO** | **$ACT_R$** | **$ACT_L$** | **ISO** | **$ACT_R$** | **$ACT_L$** |
| 8.1 | 6.8 | 7.8 | 10.5 | 3.1 | 8.1 | 31.9 | 12.2 | 17.8 | 7.5 | 1.5 | 6.2 |

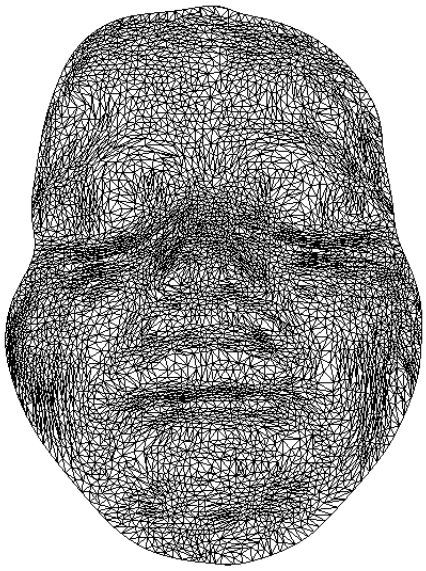
## Example 2. Dual vector field on a single Cartesian grid. STL model of the human face mask.

The ACT applied to the surface from Example 1 employs decomposition techniques to achieve the best alignment with the vector field. However, certain surfaces allow for a decomposition of the vector field $V(u,v)$ into a dual vector field $V_D(u,v) = (\alpha(u,v),\beta(u,v))$ and adaptation of the two families of the coordinate curves to $\alpha(u,v)$ and $\beta(u,v)$, respectively, within a single parametric region (see Section 4).

We apply these techniques to an STL model of the human face mask in Figure 10 (a), (b). The two families of vector fields shown in Figures 10 (c) and (d) are obtained using $V_R$ while Figure 10 (e) and (f) show the vector field obtained using $V_L$. The VFAGs are presented in Figures 10 (g) and (h). The particular cost function affects BSFCs displayed in Figures 10 (i) and (j). Virtual and real machining using the regular ISO and $ACT_R$ are presented in Figures 10 (k) - (m) and Figures 10 (l) - (n), respectively.

Table 4 presenting the performance of the ACT against the benchmark toolpath generation methods shows the advantages of the proposed framework. The $ACT_L$ is shorter by 17% and the $ATC_R$ is faster by 13%, compared to ISO. Note that the real industrial parts require thousands of hours of costly machine operations and qualified technicians. Therefore, the advantage of 13-17% is financially and technically sound. Besides, the example shows that even for a very complicated surface and seemingly chaotic vector field, the ACT based on a single VFAG can lead to important improvements (our next example shows that those improvements can be even more tangible). Table 5 shows that the efficiency of the ACT measured by the number of the required CC increases, whereas the average kinematic error practically does not change. Finally,

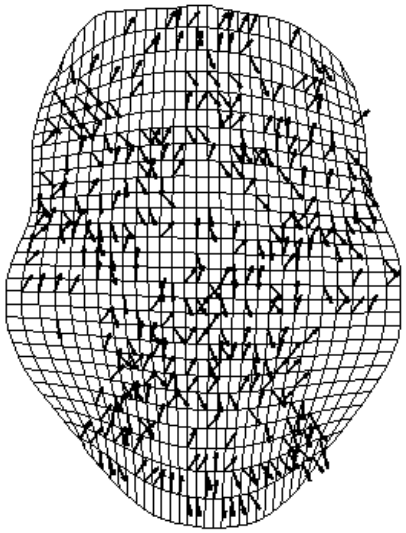
Table **6** displaying the roughness and waviness of the machined surface measured by stylus profiling, shows the advantages of the ACT.



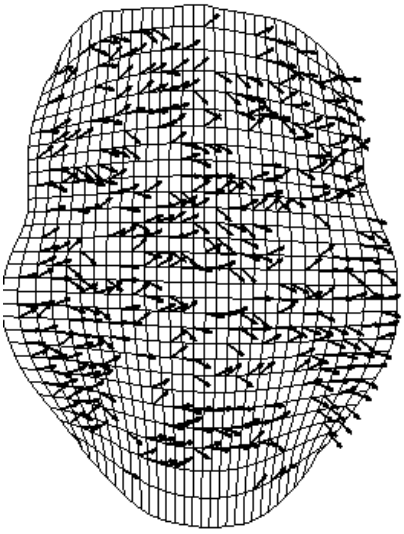
(a) Human mask



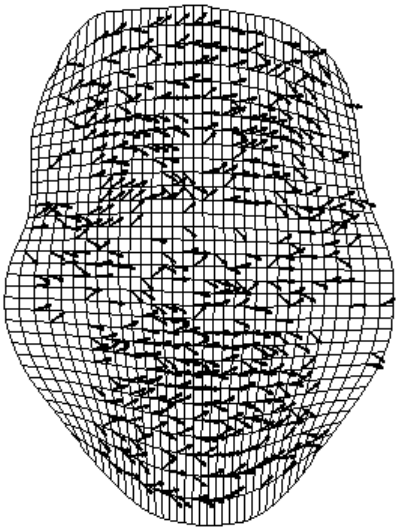
(b) Flattened human mask



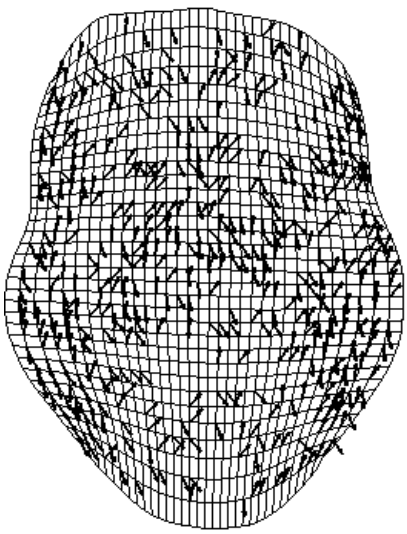
(c) $\alpha_R(u,v)$



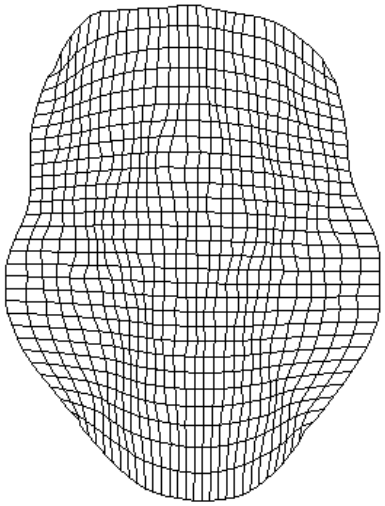
(d) $\beta_R(u,v)$



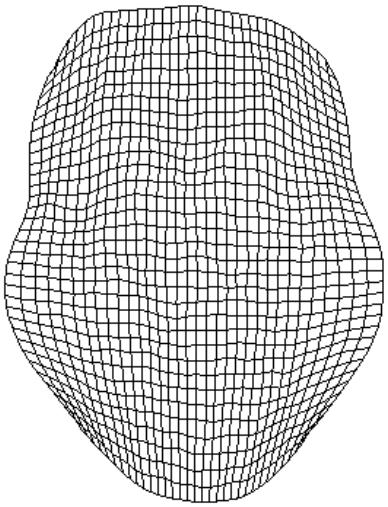
(e) $\alpha_L(u,v)$



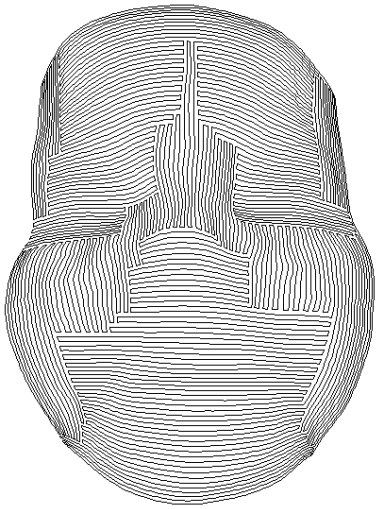
(f) $\beta_L(u,v)$



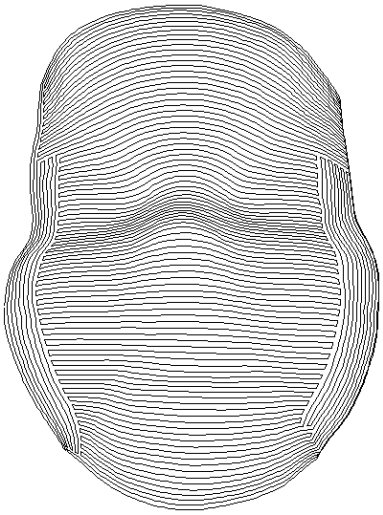
(g) $VFAG_R$



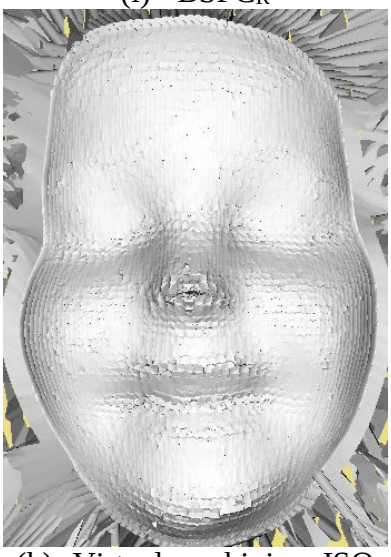
(h) $VFAG_L$



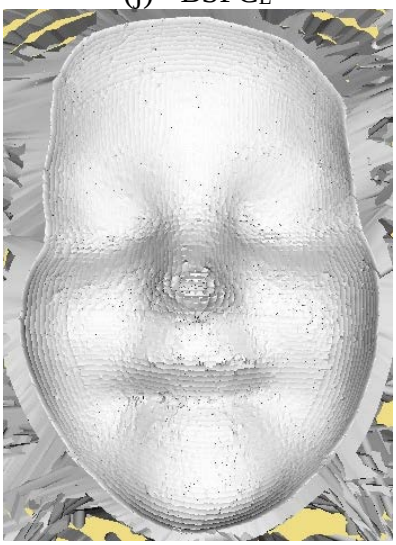
(i) $BSFC_R$



(j) $BSFC_L$



(k) Virtual machining, ISO



(l) Virtual machining, $ACT_R$



(m) Real machining, ISO



(n) Real machining, $ACT_R$

Figure 10. Machining the human face mask

Table 4 Human face surface: ACT vs. the reference methods

| | Toolpath Generation Methods | | | Scallop Height (mm) | | |
|---|---|---|---|---|---|---|
| | | | | **0.25** | **0.10** | **0.05** |
| **Performance** | **Toolpath Length (mm)** | | **Master CAM** | 17372.6 | 37669.5 | 71054.5 |
| | | | **HS** | 2968.6 | 7072.1 | 13988.8 |
| | | | **FP** | 3224.6 | 3241.2 | 4426.2 |
| | | | **ISO** | 3115.6 | 3203.0 | 4072.0 |
| | | | **$ACT_R$** | 2944.9 | 2961.8 | 3530.1 |
| | | | **$ACT_L$** | 2681.3 | 2671.1 | 3364.0 |
| | **Machining Time** | **HAAS VF2TR** | **HS** | 0:24:42.0 | 0:57:53.8 | 1:53:36.8 |
| | | | **FP** | 0:48:24.1 | 0:48:48.2 | 1:07:05.6 |
| | | | **ISO** | 0:54:27.4 | 0:55:59.7 | 1:11:29.1 |
| | | | **$ACT_R$** | 0:51:57.2 | 0:51:43.1 | 1:02:10.3 |
| | | | **$ACT_L$** | 0:53:30.3 | 0:54:41.2 | 1:09:30.9 |
| | | **MAHO 600E** | **HS** | 0:18:18.6 | 0:42:44.2 | 1:23:53.4 |
| | | | **FP** | 0:24:57.8 | 0:25:17.8 | 0:34:30.2 |
| | | | **ISO** | 0:17:52.6 | 0:18:24.9 | 0:23:27.5 |
| | | | **$ACT_R$** | 0:17:11.1 | 0:17:04.4 | 0:20:50.5 |
| | | | **$ACT_L$** | 0:17:39.2 | 0:18:07.3 | 0:22:57.1 |
| **Advantage rel. to ISO** | **Toolpath Length (mm)** | | **$ACT_R$** | 5.5% | 7.5% | 13.3% |
| | | | **$ACT_L$** | 13.9% | 16.6% | **17.4%** |
| | **Machining Time** | **HAAS VF2TR** | **$ACT_R$** | 4.6% | 7.6% | **13.0%** |
| | | | **$ACT_L$** | 1.7% | 2.3% | 2.8% |
| | | **MAHO 600E** | **$ACT_R$** | 3.9% | 7.3% | 11.2% |
| | | | **$ACT_L$** | 1.2% | 1.6% | 2.2% |

Table 5 Human face surface: CC points, kinematic error, ACT vs. ISO

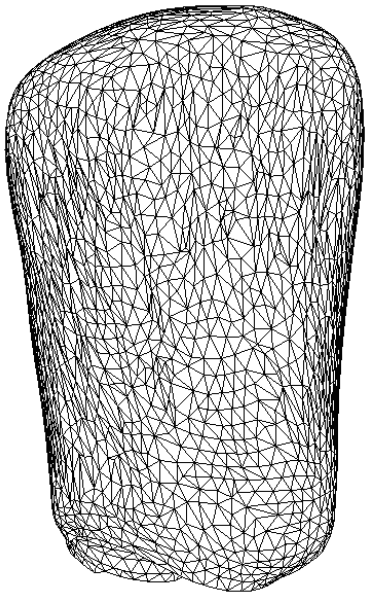
| Toolpath Generation techniques | HAAS VF-2TR | | | | | | | | | MAHO 600-E | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | h=0.25 mm | | | h=0.10 mm | | | h=0.05 mm | | | h=0.25 mm | | | h=0.10 mm | | | h=0.05 mm | | |
| | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\overline{\varepsilon}$ | $\sigma_\varepsilon$ |
| **ISO** | 1509 | 0.081 | 0.616 | 3936 | 0.056 | 0.183 | 6917 | 0.050 | 0.031 | 1260 | 0.050 | 0.154 | 1221 | 0.028 | 0.064 | 3132 | 0.022 | 0.060 |
| **$ACT_R$** | 1446 | 0.060 | 0.179 | 3896 | 0.035 | 0.104 | 6832 | 0.028 | 0.018 | 1171 | 0.034 | 0.076 | 1206 | 0.019 | 0.049 | 2704 | 0.009 | 0.030 |
| **$ACT_L$** | 1023 | 0.061 | 0.077 | 2822 | 0.035 | 0.036 | 4911 | 0.025 | 0.029 | 275 | 0.026 | 0.050 | 962 | 0.018 | 0.027 | 1941 | 0.013 | 0.020 |

Table 6 Human face surface, roughness and waviness, ACT vs. ISO

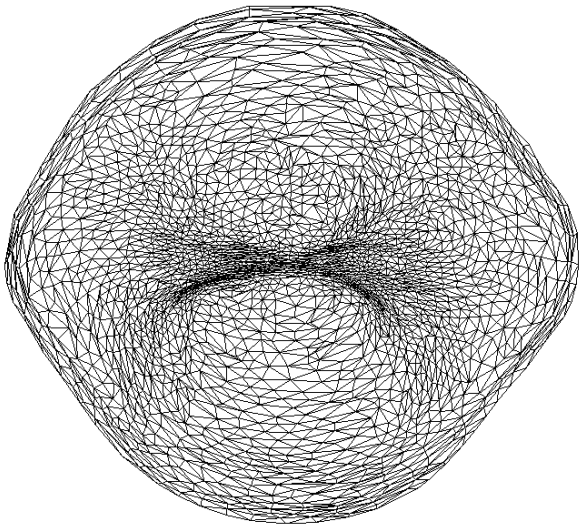
| Roughness $R_a$ ($\mu$m) | | Standard Deviation | | Waviness $R_a$ | | Standard Deviation | |
|---|---|---|---|---|---|---|---|
| **ISO** | **$ACT_R$** | **ISO** | **$ACT_R$** | **ISO** | **$ACT_R$** | **ISO** | **$ACT_R$** |
| 8.3 | 8.1 | 7.7 | 7.7 | 43.6 | 40.8 | 23.8 | 28.6 |

## Example 3. Dual vector field on a single Cartesian grid. An STL model of the incisor tooth.

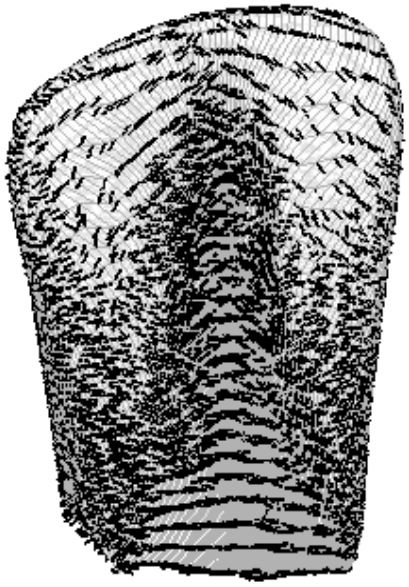
We verify our methodology on the STL models of four types of human teeth: incisor, premolar, molar, and canine. The STL model of the incisor and the corresponding flattened STL are presented in Figure 11. The incisor represents the case when the dual vector field on a single Cartesian grid leads to the significant improvement, in terms of the length of the tool path and the machining time. The components of the dual vector fields $V_{D,R}(u,v) = (\alpha_R(u,v), \beta_R(u,v))$ and $V_{D,L}(u,v) = (\alpha_L(u,v), \beta_L(u,v))_L$ are shown on the surface and in the parametric domain in Figures 11 (c) - (h). VFAGs are presented in Figures 11 (i) - (l). Finally, machined surfaces are shown in Figures 11 (m) - (n). Table 7 shows the benefits. The toolpath is 43% shorter and 46% faster against the standard ISO. The advantage with regard to NX9, in terms of the length of the tool path, is also very significant. For instance, when h=0.01, the $ACT_L$ generates the tool path of 570.9 whereas the best NX9 result is 1049.3. The advantage in the machining time on HAAS is 17%. It should be noted that the neither MasterCam nor NX9 can produce the 5 axis tool path for the STL surface directly**.** In order to compare their performance, we flattened the STL surface, generated a boundary fitted curvilinear grid, and approximated the surface on that grid. Therefore, our MasterCam and NX9 toolpath generation is partly based on the proposed techniques. Finally, Table 9 shows that the efficiency of the ACT, measured by the number of the required CC points, increases, whereas the average kinematic error practically does not change.



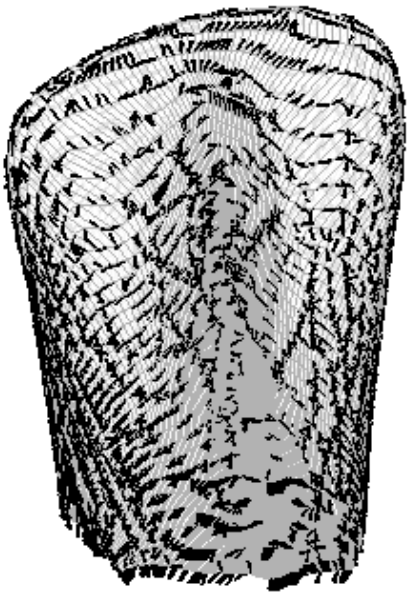
(a) STL model of the incisor



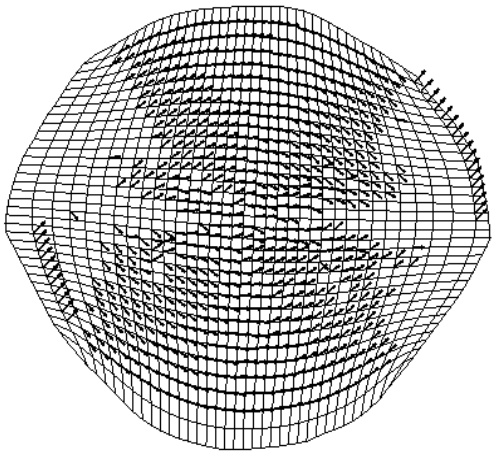
(b) flattened incisor



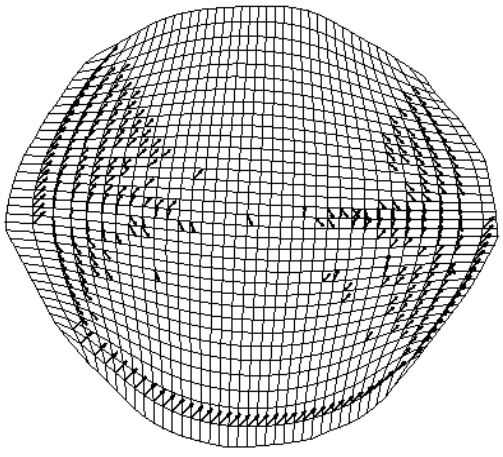
(c) $V_R$



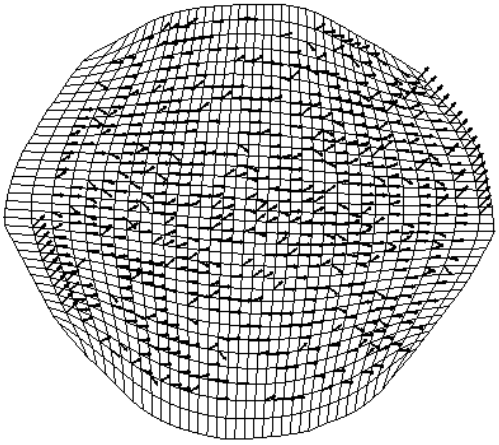
(d) $V_L$



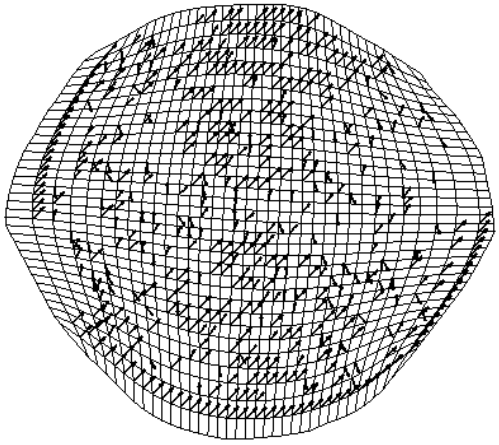
(e) $\alpha_R(u,v)$



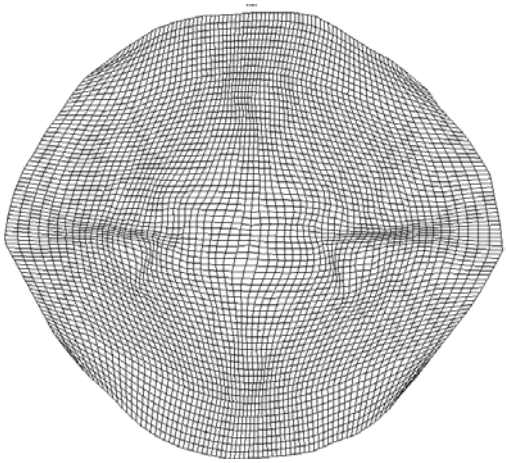
(f) $\beta_R(u,v)$



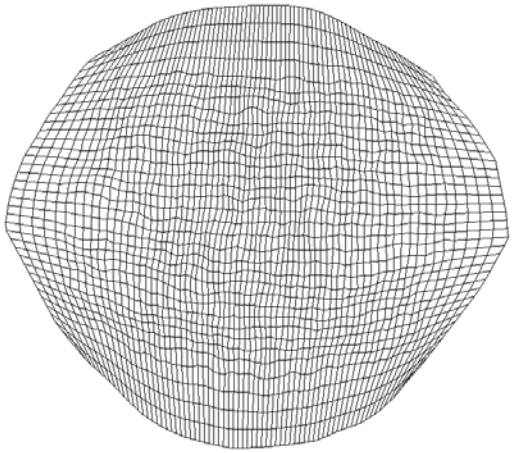
(g) $\alpha_L(u,v)$



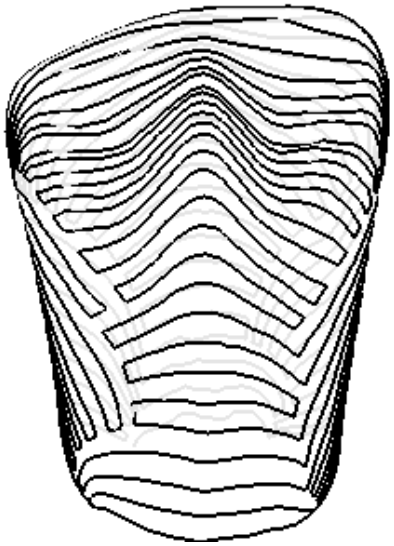
(h) $\beta_L(u,v)$



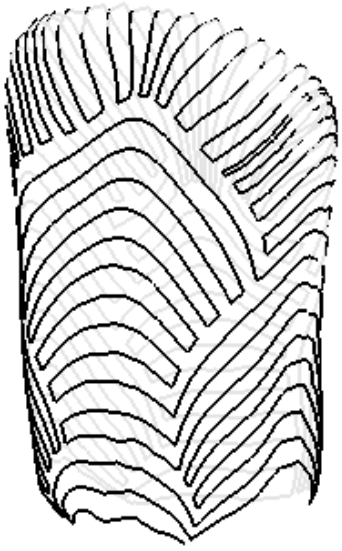
(i) $VFAG_R$



(j) $VFAG_L$



(k) $BSFC_R$



(l) $BSFC_L$



(m)Virtual milling, $ACT_R$



(n) Virtual milling, $ACT_L$

Figure 11. Machining the incisor tooth

Table 7 Incisor tooth: ACT vs. the reference methods

| | Toolpath Generation Methods | | | Scallop Height (mm) | | |
|---|---|---|---|---|---|---|
| | | | | **0.065** | **0.025** | **0.0125** |
| **Performance** | **Toolpath Length (mm)** | | **Master CAM** | 6834.8 | 14373.3 | 27614.1 |
| | | | **HS** | 655.9 | 1477.7 | 2781.5 |
| | | | **FP** | 800.1 | 940.4 | 1049.3 |
| | | | **ISO** | 756.4 | 837.9 | 1005.5 |
| | | | **ACT R** | 469.1 | 515.6 | 582.9 |
| | | | **ACT L** | 456.0 | 498.5 | 570.9 |
| | **Machining Time** | **HAAS VF2TR** | **HS** | 0:18:22.0 | 0:40:40.2 | 1:19:15.1 |
| | | | **FP** | 0:21:13.4 | 0:28:27.7 | 0:30:50.6 |
| | | | **ISO** | 0:34:19.2 | 0:38:35.2 | 0:46:14.4 |
| | | | **ACT R** | 0:19:47.5 | 0:21:34.2 | 0:25:02.3 |
| | | | **ACT L** | 0:24:50.6 | 0:27:11.2 | 0:31:25.2 |
| | | **MAHO 600E** | **HS** | 0:07:34.5 | 0:16:48.4 | 0:32:04.6 |
| | | | **FP** | 0:07:30.5 | 0:08:53.7 | 0:09:55.8 |
| | | | **ISO** | 0:11:10.7 | 0:12:49.0 | 0:15:38.9 |
| | | | **ACT R** | 0:07:13.9 | 0:07:56.6 | 0:09:11.0 |
| | | | **ACT L** | 0:07:13.1 | 0:08:09.3 | 0:09:33.1 |
| **Advantage ref. to ISO** | **Toolpath Length (mm)** | | **ACT R** | 38.00% | 38.5% | 42.0% |
| | | | **ACT L** | 39.70% | 40.5% | 43.2% |
| | **Machining Time** | **HAAS VF2TR** | **ACT R** | 42.30% | 44.1% | 45.9% |
| | | | **ACT L** | 27.60% | 29.5% | 32.0% |
| | | **MAHO 600E** | **ACT R** | 35.30% | 38.0% | 41.3% |
| | | | **ACT L** | 35.40% | 36.4% | 39.0% |

Table 8 Incisor tooth : CC points, kinematic error, ACT vs. ISO

| | HAAS VF-2TR | | | | | | | | | MAHO 600-E | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | h=0.065 mm | | | h=0.025 mm | | | h=0.0125 mm | | | h=0.065 mm | | | h=0.025 mm | | | h=0.0125 mm | | |
| | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ | $N_{CC}$ | $\bar{\varepsilon}$ | $\sigma_\varepsilon$ |
| **ISO** | 576 | 0.048 | 0.173 | 1709 | 0.030 | 0.136 | 4880 | 0.021 | 0.122 | 36 | 0.013 | 0.029 | 246 | 0.011 | 0.026 | 1282 | 0.008 | 0.024 |
| **$ACT_R$** | 222 | 0.044 | 0.069 | 923 | 0.026 | 0.056 | 2545 | 0.016 | 0.072 | 7 | 0.010 | 0.023 | 99 | 0.009 | 0.017 | 611 | 0.006 | 0.013 |
| **$ACT_L$** | 574 | 0.051 | 0.178 | 1443 | 0.031 | 0.134 | 3425 | 0.027 | 0.109 | 60 | 0.014 | 0.032 | 350 | 0.009 | 0.024 | 1137 | 0.007 | 0.024 |

## Example 4. C–grids applied to the STL models of canine, premolar, and molar teeth

The following three cases of the STL surfaces, corresponding to premolar, molar, and canine teeth, are characterized by the unstructured (chaotic) vector field. Consequently, VFAGs do not produce any improvement, even though the grids can be aligned with some small clusters of the vectors. However, we show that the tools provided by the ACT still can be used to improve the quality of the toolpath generation. We show that, boundary fitted curvilinear tool path, combined with a manual selection of the type of the grid and BSFC, lead to a substantial improvement.

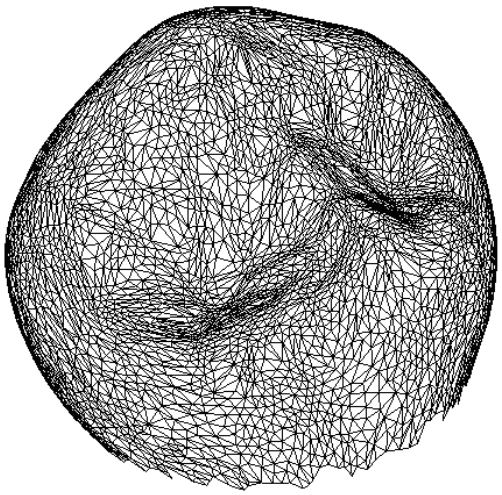
In the context of the grid generation technologies, the basic Cartesian grids include H-grids, O-grids, and C- grids[20,52]. We will show the premolar, molar, and canine STL models can be produced using the O-grid for a shorter time, using a shorter tool path.

The STL-premolar and its flattened version are presented in Figures 12 (a) - (d). The boundary fitted initial grids in Figures 12 (e) and (f) are the H-grid and the O-grid respectively. In order to evaluate the suitability of the grid, we compute the alignment functional (3) as follows:

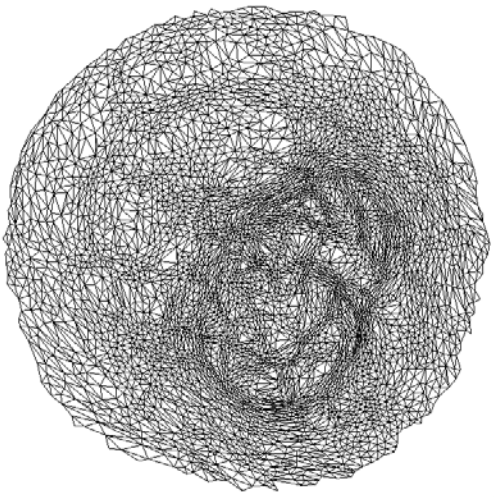
$F_{V,H,canine}=16.0$ , $F_{V,O,canine}=14.0$ , $F_{V,H,premolar}=16.7$ , $F_{V,O,premolar}=14.0$ , $F_{V,H,molar}=19.3$ , $F_{V,O,molar}=17.6$.

Clearly, the O-grid is more suitable for the three dental models. The corresponding toolpaths and virtually machined surfaces are shown in Figures 12 (g) – (l). The results for the molar and canine teeth are displayed in Figure 13 and Figure 14. Finally, Table 9 presents the benefits of the ACT, compared to ISO and to NX9 and MasterCAM. The proposed framework provides advantages in both the toolpath length and machining time, with regard to the benchmark methods. For instance, the premolar tooth can be machined by the ACT, with the advantage of 26% in length and 43% in time relative to ISO.

Table 10 provides the number of CC points versus the scallop height and the kinematic error.



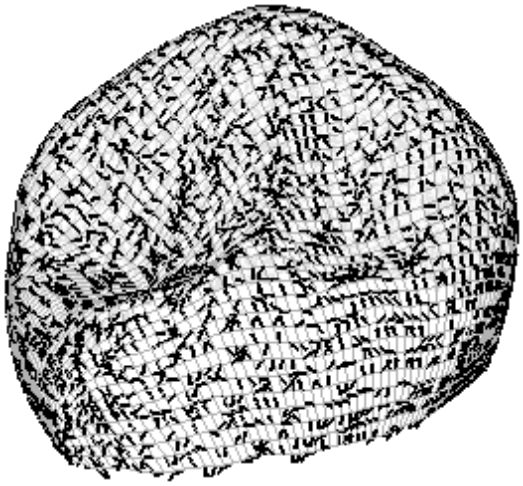
(a) STL model of the premolar tooth



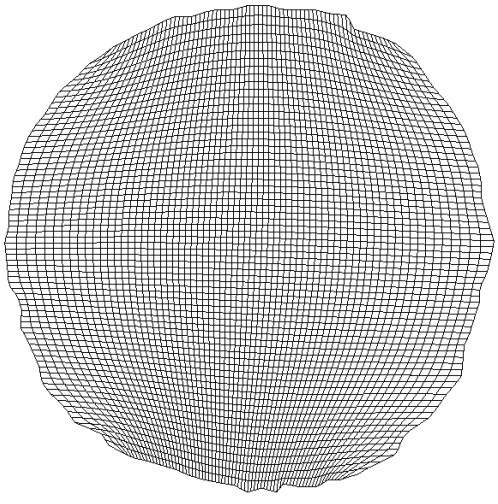
(b) flattened premolar tooth



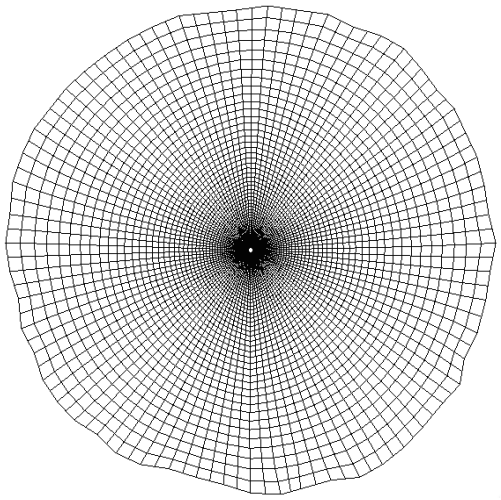
(c) $V_R$



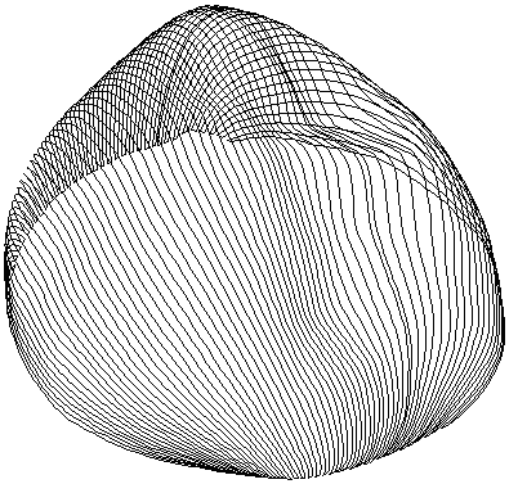
(d) $V_L$



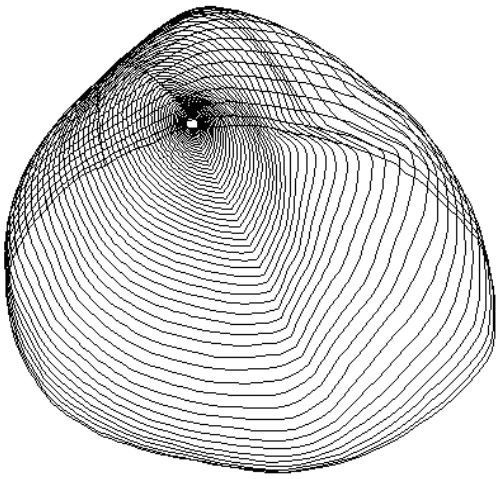
(e) H-grid



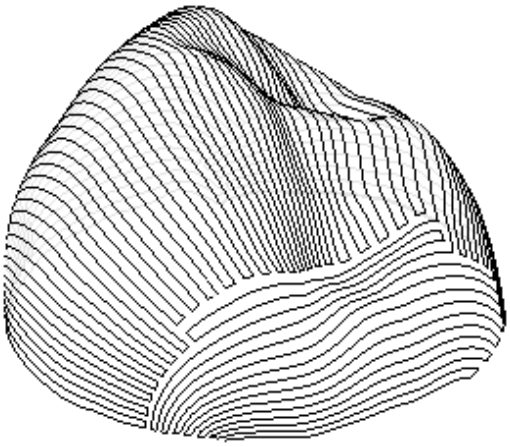
(f) O-grid



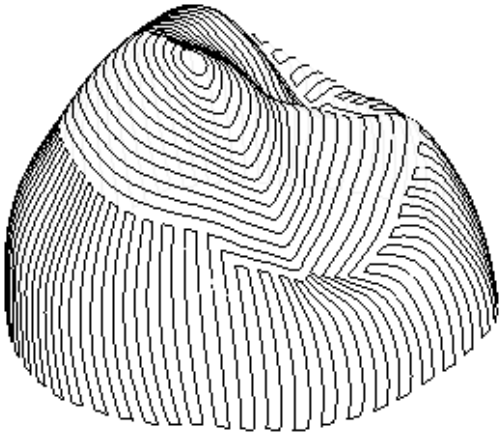
(g) $ISO_H$



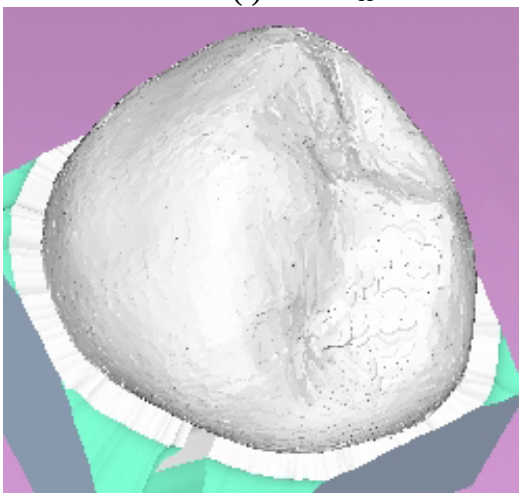
(h) $ISO_O$



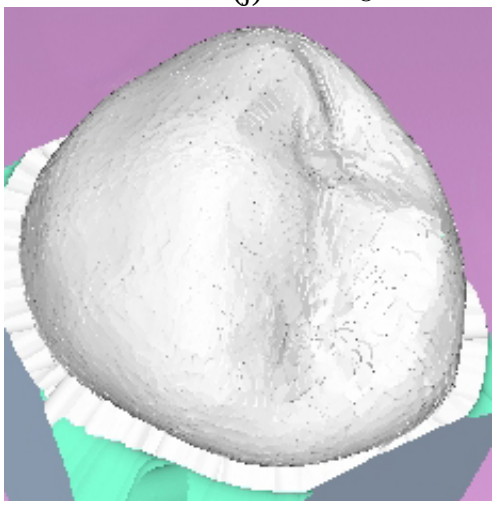
(i) $BSFC_H$



(j) $BSFC_O$

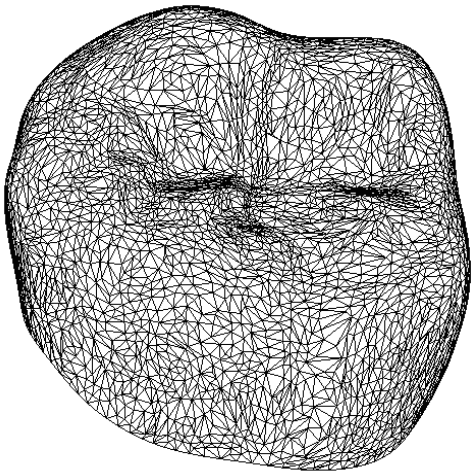


(k) Virtual milling $ISO_H$

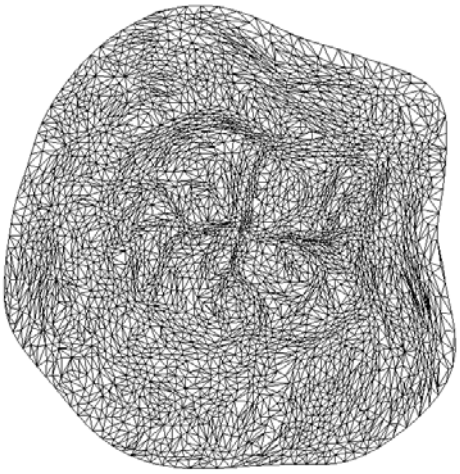


(l) Virtual milling $BSFC_O$

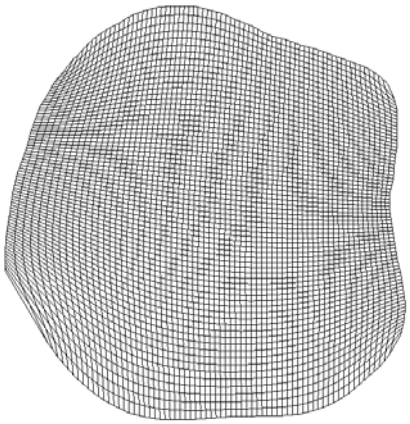
Figure 12. Machining the premolar tooth



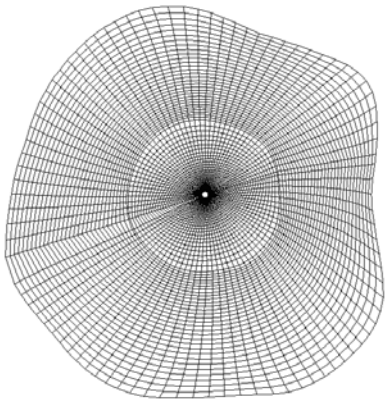
(a) STL model of the molar tooth



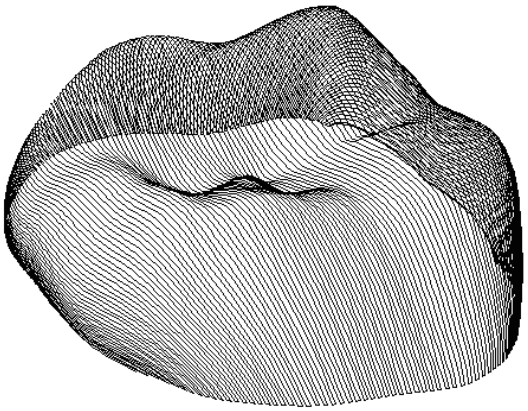
(b) flattened molar tooth



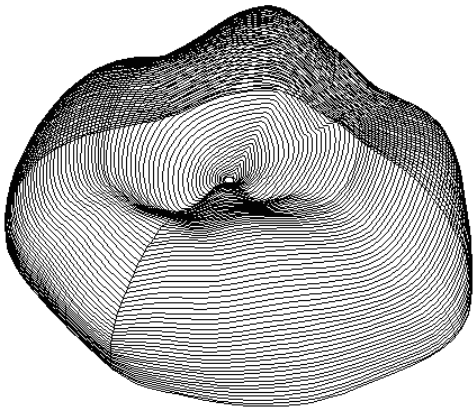
(c) H-grid



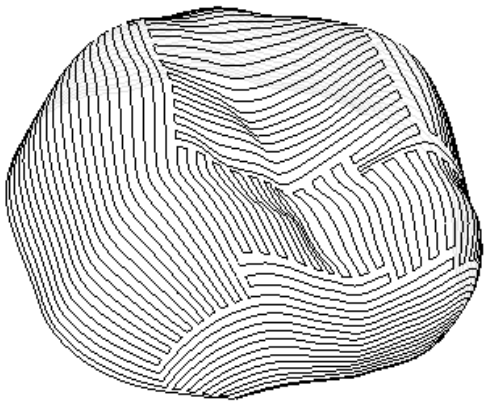
(d) O-grid



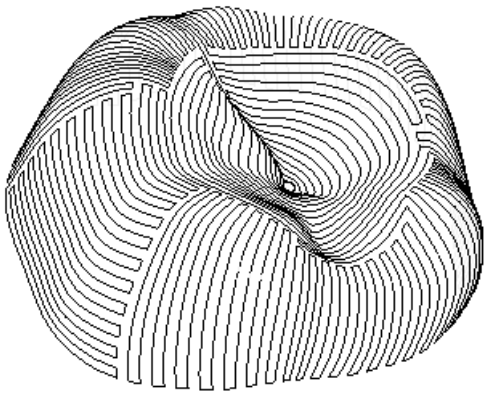
(e) $ISO_H$



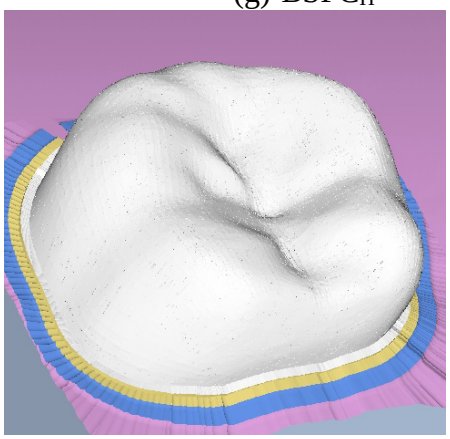
(f) $ISO_O$



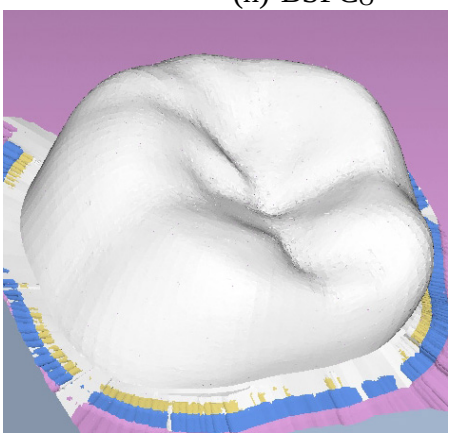
(g) $BSFC_H$



(h) $BSFC_O$

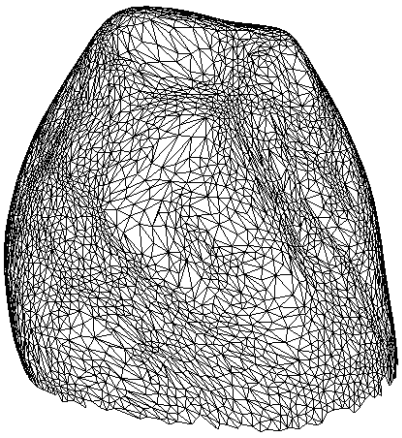


(i) Virtual milling, $ISO_H$

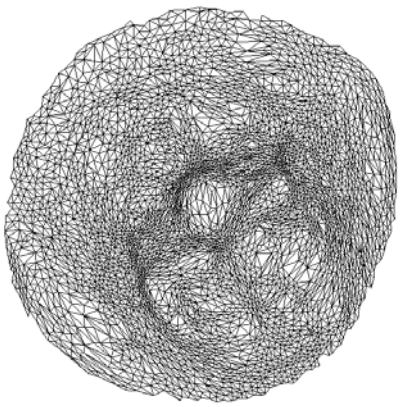


(j) Virtual milling, $BSFC_O$

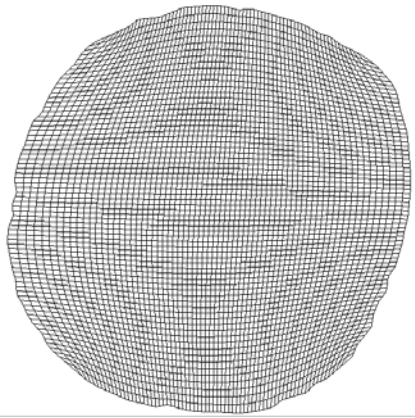
Figure 13. Machining the molar tooth



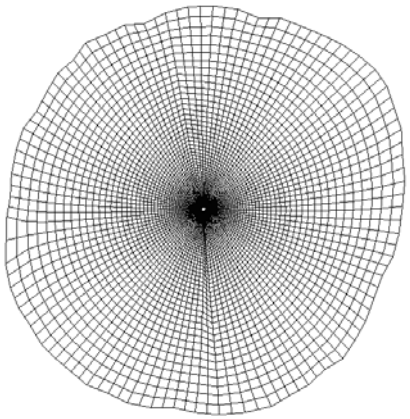
(a) STL model of a canine tooth



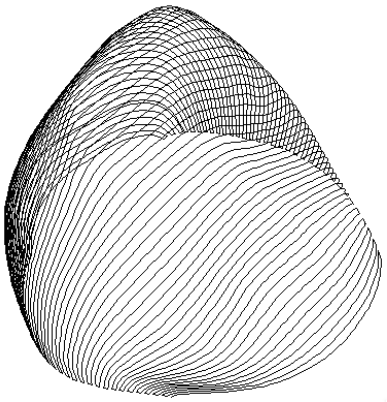
(b) Flattened canine tooth



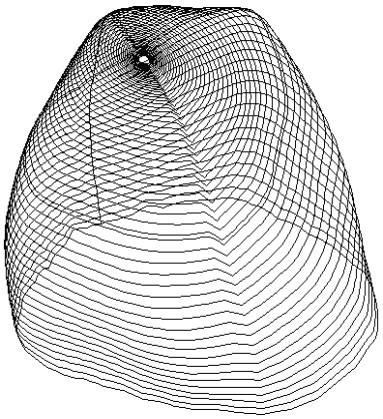
(c) H-grid



(d) O-grid



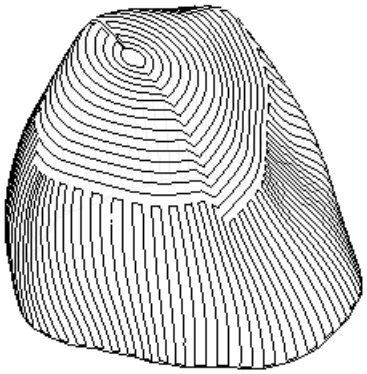
(e) $ISO_H$



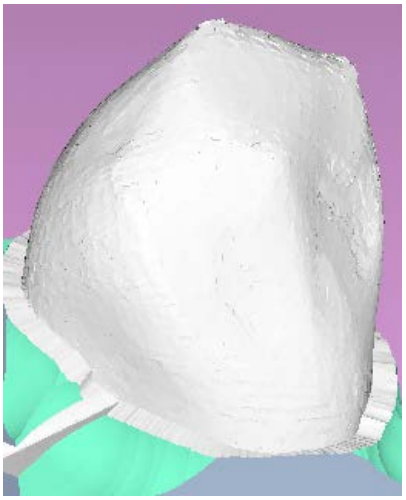
(f) $ISO_O$



(g) $BSFC_H$



(h) $BSFC_O$



(i) Virtual Milling $ISO_H$



(j) Virtual Milling $BSFC_O$

Figure 14. Machining the canine tooth

Table 9 ACT vs. the reference methods, molar, premolar, and canine tooth.

| | Toolpath Generation Techniques | | Canine | | | Premolar | | | Molar | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Scallop Height (mm) | | | Scallop Height (mm) | | | Scallop Height (mm) | | |
| | | | 0.065 | 0.025 | 0.0125 | 0.065 | 0.025 | 0.0125 | 0.065 | 0.025 | 0.0125 |
| Performance | Toolpath Length (mm) | Master CAM | 4261.3 | 9615.0 | 18802.0 | 5275.9 | 12560.3 | 25009.4 | 6483.1 | 15102.6 | 30074.0 |
| | | HS | 1013.0 | 2330.4 | 4599.0 | 1023.5 | 2360.9 | 4614.8 | 1284.6 | 2992.0 | 5912.0 |
| | | FP | 1120.5 | 1181.6 | 1230.2 | 1129.6 | 1156.9 | 1217.7 | 1099.9 | 1141.9 | 1184.5 |
| | | ISO | 766.7 | 778.4 | 801.5 | 761.7 | 773.2 | 783.5 | 847.6 | 854.4 | 855.4 |
| | | O-Grid | 664.6 | 664.5 | 664.4 | 616.9 | 622.6 | 620.5 | 826.9 | 827.7 | 828.2 |
| | | $BSFC_H$ | 610.7 | 620.3 | 634.3 | 616.1 | 617.5 | 623.5 | 782.5 | 784.8 | 785.4 |
| | | $BSFC_O$ | 617.8 | 627.4 | 641.6 | 573.2 | 574.6 | 580.1 | 742.9 | 745.1 | 745.7 |
| | HAAS VF2TR | HS | 0:20:36.5 | 0:47:34.7 | 1:33:34.8 | 0:24:28.4 | 0:56:43.5 | 1:50:48.3 | 0:37:12.3 | 1:27:20.7 | 2:51:51.9 |
| | | FP | 0:24:54.2 | 0:26:15.4 | 0:27:18.8 | 0:34:15.9 | 0:35:04.8 | 0:36:43.3 | 0:37:03.1 | 0:38:46.2 | 0:39:52.8 |
| | | ISO | 0:24:21.8 | 0:24:48.3 | 0:25:19.3 | 0:36:41.9 | 0:38:14.5 | 0:42:14.5 | 0:41:27.5 | 0:41:55.5 | 0:42:34.1 |
| | | O-Grid | 0:22:15.8 | 0:22:33.5 | 0:22:52.9 | 0:28:09.1 | 0:28:24.6 | 0:30:41.5 | 0:35:19.5 | 0:35:35.1 | 0:35:40.0 |
| | | $BSFC_H$ | 0:20:55.9 | 0:21:05.1 | 0:21:07.9 | 0:28:44.5 | 0:28:50.6 | 0:30:30.3 | 0:32:47.4 | 0:32:55.0 | 0:32:58.7 |
| | | $BSFC_O$ | 0:20:01.9 | 0:20:10.7 | 0:20:13.3 | 0:27:57.2 | 0:28:03.1 | 0:29:40.2 | 0:33:17.6 | 0:33:25.3 | 0:33:29.1 |
| | MAHO 600E | HS | 0:13:48.8 | 0:31:42.8 | 1:02:24.8 | 0:17:35.1 | 0:40:19.1 | 1:18:37.1 | 0:28:52.8 | 1:07:35.2 | 2:13:07.7 |
| | | FP | 0:10:45.1 | 0:11:20.2 | 0:11:48.2 | 0:13:48.4 | 0:14:06.2 | 0:14:42.3 | 0:19:19.1 | 0:20:13.7 | 0:20:41.2 |
| | | ISO | 0:12:20.3 | 0:12:55.3 | 0:13:18.7 | 0:19:20.1 | 0:19:30.1 | 0:19:46.5 | 0:21:19.7 | 0:21:34.5 | 0:22:25.1 |
| | | O-Grid | 0:10:13.7 | 0:10:38.5 | 0:10:55.9 | 0:12:06.4 | 0:12:11.7 | 0:12:15.6 | 0:17:48.3 | 0:17:58.1 | 0:18:02.0 |
| | | $BSFC_H$ | 0:10:15.3 | 0:10:38.3 | 0:10:39.7 | 0:13:40.9 | 0:13:42.7 | 0:13:48.7 | 0:18:32.3 | 0:18:40.6 | 0:18:47.1 |
| | | $BSFC_O$ | 0:09:18.3 | 0:09:39.1 | 0:09:40.4 | 0:11:00.3 | 0:11:01.7 | 0:11:06.6 | 0:18:45.0 | 0:18:53.3 | 0:18:59.9 |
| Advantage rel. to ISO | Toolpath Length (mm) | O-Grid | 13.3% | 14.6% | 17.1% | 19.0% | 19.5% | 20.8% | 2.4% | 3.1% | 3.2% |
| | | $BSFC_H$ | 20.3% | 20.3% | 20.9% | 19.1% | 20.1% | 20.4% | 7.7% | 8.1% | 8.2% |
| | | $BSFC_O$ | 19.4% | 19.4% | 19.9% | 24.7% | 25.7% | 26.0% | 12.3% | 12.8% | 12.8% |
| | HAAS VF2TR | O-Grid | 8.6% | 9.1% | 9.6% | 23.3% | 25.7% | 27.3% | 14.8% | 15.1% | 16.2% |
| | | $BSFC_H$ | 14.1% | 15.0% | 16.6% | 21.7% | 24.6% | 27.8% | 20.9% | 21.5% | 22.5% |
| | | $BSFC_O$ | 17.8% | 18.7% | 20.1% | 23.8% | 26.6% | 29.8% | 19.7% | 20.3% | 21.3% |
| | MAHO 600E | O-Grid | 17.1% | 17.6% | 17.9% | 37.4% | 37.5% | 38.0% | 16.5% | 16.7% | 19.6% |
| | | $BSFC_H$ | 16.9% | 17.7% | 19.9% | 29.2% | 29.7% | 30.2% | 13.1% | 13.4% | 16.2% |
| | | $BSFC_O$ | 24.6% | 25.3% | 27.3% | 43.1% | 43.4% | 43.8% | 12.1% | 12.5% | 15.3% |

Table 10 Molar, premolar and canine: CC points, kinematic error, ACT vs. ISO

| | | | HAAS VF2TR | | | MAHO 600E | | |
|---|---|---|---|---|---|---|---|---|
| **Scallop height** | **Method** | **Type** | **Molar** | **Premolar** | **Canine** | **Molar** | **Premolar** | **Canine** |
| **0.065** | **ISO** | $N_{CC}$ | 1645 | 788 | 342 | 379 | 203 | 58 |
| | | $\overline{\varepsilon}$ | 0.078 | 0.059 | 0.035 | 0.039 | 0.019 | 0.012 |
| | | $\sigma_{\varepsilon}$ | 0.170 | 0.173 | 0.065 | 0.048 | 0.074 | 0.033 |
| | **O-Grid** | $N_{CC}$ | 1637 | 708 | 302 | 577 | 163 | 59 |
| | | $\overline{\varepsilon}$ | 0.073 | 0.064 | 0.041 | 0.038 | 0.024 | 0.012 |
| | | $\sigma_{\varepsilon}$ | 0.174 | 0.280 | 0.062 | 0.058 | 0.051 | 0.031 |
| | **$BSFC_H$** | $N_{CC}$ | 1575 | 748 | 306 | 478 | 174 | 59 |
| | | $\overline{\varepsilon}$ | 0.069 | 0.060 | 0.035 | 0.038 | 0.021 | 0.012 |
| | | $\sigma_{\varepsilon}$ | 0.167 | 0.220 | 0.058 | 0.049 | 0.062 | 0.030 |
| | **$BSFC_O$** | $N_{CC}$ | 1619 | 741 | 304 | 468 | 175 | 56 |
| | | $\overline{\varepsilon}$ | 0.068 | 0.057 | 0.033 | 0.036 | 0.021 | 0.011 |
| | | $\sigma_{\varepsilon}$ | 0.167 | 0.209 | 0.056 | 0.051 | 0.060 | 0.030 |
| **0.025** | **ISO** | $N_{CC}$ | 3080 | 1654 | 984 | 776 | 445 | 171 |
| | | $\overline{\varepsilon}$ | 0.048 | 0.037 | 0.026 | 0.026 | 0.017 | 0.011 |
| | | $\sigma_{\varepsilon}$ | 0.143 | 0.163 | 0.059 | 0.046 | 0.065 | 0.018 |
| | **O-Grid** | $N_{CC}$ | 3099 | 1655 | 920 | 1171 | 435 | 164 |
| | | $\overline{\varepsilon}$ | 0.042 | 0.037 | 0.028 | 0.026 | 0.017 | 0.010 |
| | | $\sigma_{\varepsilon}$ | 0.142 | 0.253 | 0.051 | 0.032 | 0.035 | 0.018 |
| | **$BSFC_H$** | $N_{CC}$ | 2966 | 1638 | 933 | 974 | 427 | 159 |
| | | $\overline{\varepsilon}$ | 0.044 | 0.036 | 0.026 | 0.024 | 0.016 | 0.010 |
| | | $\sigma_{\varepsilon}$ | 0.128 | 0.202 | 0.053 | 0.035 | 0.048 | 0.016 |
| | **$BSFC_O$** | $N_{CC}$ | 2957 | 1632 | 936 | 974 | 418 | 165 |
| | | $\overline{\varepsilon}$ | 0.040 | 0.036 | 0.026 | 0.024 | 0.017 | 0.009 |
| | | $\sigma_{\varepsilon}$ | 0.138 | 0.196 | 0.052 | 0.036 | 0.048 | 0.016 |
| **0.0125** | **ISO** | $N_{CC}$ | 6359 | 3692 | 2115 | 1878 | 1206 | 520 |
| | | $\overline{\varepsilon}$ | 0.021 | 0.025 | 0.018 | 0.014 | 0.012 | 0.009 |
| | | $\sigma_{\varepsilon}$ | 0.130 | 0.183 | 0.048 | 0.042 | 0.057 | 0.016 |
| | **O-Grid** | $N_{CC}$ | 6309 | 3698 | 2086 | 2897 | 1134 | 490 |
| | | $\overline{\varepsilon}$ | 0.021 | 0.024 | 0.018 | 0.015 | 0.012 | 0.008 |
| | | $\sigma_{\varepsilon}$ | 0.101 | 0.236 | 0.046 | 0.023 | 0.025 | 0.016 |
| | **$BSFC_H$** | $N_{CC}$ | 6207 | 3621 | 2079 | 2292 | 1170 | 480 |
| | | $\overline{\varepsilon}$ | 0.019 | 0.025 | 0.017 | 0.014 | 0.012 | 0.008 |
| | | $\sigma_{\varepsilon}$ | 0.104 | 0.207 | 0.043 | 0.030 | 0.040 | 0.016 |
| | **$BSFC_O$** | $N_{CC}$ | 6292 | 3597 | 2093 | 2261 | 1158 | 492 |
| | | $\overline{\varepsilon}$ | 0.019 | 0.024 | 0.017 | 0.013 | 0.011 | 0.008 |
| | | $\sigma_{\varepsilon}$ | 0.105 | 0.194 | 0.044 | 0.029 | 0.038 | 0.015 |

## CONCLUSIONS

A new framework for generation of adaptive vector field aligned toolpaths for five-axis machining has been presented and analyzed. The method provides up to 70% decrease of the machining time with regard to the ISO toolpath, and up to 64% with regard to FP of NX9 (the best reference method). The advantages with regard to HS of NX9 and MasterCAM are overwhelming: the machining time is 10-20 times shorter. The tests against the benchmark toolpath generation show that the kinematic error and the roughness of the workpiece remain practically unchanged while the machining time, toolpath length, and waviness have been considerably reduced.

## ACKNOWLEGEMENT

This research is supported by the Center of Excellency in Biomedical Engineering of Thammasat University and the National Research Council of Thailand (Grant No. 168974).

## REFERENCES

1. C.C.Lo, Efficient cutter-path planning for five-axis surface machining with a flat-end cutter. *Computer-Aided Design.* **31**, 557-566, 1999.
2. A. Rao, R. Sarma, On local gouging in five-axis sculptured surface machining using flat-end tools. *Computer-Aided Design.* **32**, 409-420, 2000.
3. C.-J. Chiou, C-J, Y.-S. Lee, A machining potential field approach to tool path generation for multi-axis sculptured surface machining. *Computer-Aided Design*, **34,** 357-371, 2002.
4. W. Anotaipaiboon, S.S. Makhanov, E.L.J. Bohez, Optimal setup for five-axis machining. *International Journal of Machine Tools and Manufacture,* **46**, 964-977, 2006.
5. J. Gao, X. Chen, D. Zheng, O.Yilmaz, N. Gindy.Adaptive restoration of complex geometry parts through reverse engineering application. *Advances in Engineering Software.* **37**,592-600, 2006.
6. Y.-S. Lee, Admissible tool orientation control of gouging avoidance for 5-axis complex surface machining. *Computer-Aided Design*, **29**, 507-521, 1997.
7. W. Anotaipaiboon, S.S. Makhanov, Minimization of the kinematics error for five-axis machining. *Computer-Aided Design,* **43**, 1740-1757, 2011.
8. J.-H. Yoon, Tool tip gouging avoidance and optimal tool positioning for 5-axis sculptured surface machining. *International Journal of Production Research,* **41**, 2125-2142, 2003.
9. Y.-S. Lee, Non-isoparametric tool path planning by machining strip evaluation for 5-axis sculptured surface machining. *Computer-Aided Design*, **30**, 559-570. 1998.
10. Y.-S. Lee, Mathematical modelling using different endmills and tool placement problems for 4- and 5-axis NC complex surface machining. *International Journal of Production Research,* **36**, 785-814, 1998.
11. J. Dong, P.M. Ferreira, J.A. Stori, Feed-rate optimization with jerk constraints for generating minimum-time trajectories. *International Journal of Machine Tools and Manufacture,* **47**, 1941-1955, 2007.
12. T. Kim, S.E. Sarma, Toolpath generation along directions of maximum kinematic performance: a first cut at machine-optimal paths. *Computer-Aided Design,* **34**, 453-468, 2002.
13. C. My, E.L.J. Bohez, S.S. Makhanov, Critical point analysis of 3D vector field for 5-axis toolpath optimization. *Forth Asian Conference on Industrial Automation and Robotics, Bangkok, Thailand, 11-13 May, 2005.*

14. S.S. Makhanov, Optimization and correction of the tool path of the five-axis milling machine: Part 1. Spatial optimization. *Mathematics and Computers in Simulation*, **75**, 210-230, 2007.
15. L.L. Li, Y.F. Zhang, H.Y.Li, L. Geng, Generating tool-path with smooth posture change for five-axis sculptured surface machining based on cutter's accessibility map. *The International Journal of Advanced Manufacturing Technology*, **53**, 699-709, 2011.
16. S. Moodleah, S.S. Makhanov, 5-axis machining using a curvilinear tool path aligned with the direction of the maximum removal rate. *The International Journal of Advanced Manufacturing Technology*, **80**, 65-90, 2015.
17. E. Bohez, S.S. Makhanov, K. Sonthipermpoon, Adaptive nonlinear tool path optimization for five-axis machining. *International Journal of Production Research*, **38**, 4329-4343, 2000.
18. S.S. Makhanov, An application of variational grid generation techniques to the tool-path optimization of industrial milling robots, *Computational Mathematics and Mathematical Physics*, **39**, 1524–1535, 1999.
19. S. S. Makhanov, S.A.Ivanenko, Grid generation as applied to optimize cutting operations of the five-axis milling machine. *Applied Numerical Mathematics*, **46**, 331-351, 2003.
20. J.F. Thompson, B.K. Soni, N.P Weatherill, *Handbook of grid generation*. CRC Press, 1998.
21. S.A. Ivanenko, Generation of non-degenerate meshes. *Computational Mathematics and Mathematical Physics*, **28**, 141-146, 1998.
22. M.B Bieterman, D.R. Sandstrom, A curvilinear tool-path nethod for pocket machining. *Journal of Manufacturing Science and Engineering*, **125**, 709-715, 2003.
23. W. Anotaipaiboon, S.S. Makhanov. Toolpath generation for five-axis NC machining using space-filling curves, *The Third Asian Conference on Industrial Automation and Robotics*, Bangkok, Thailand, May 8-9, 2003.
24. W. Anotaipaiboon, S.S. Makhanov, Tool path generation for five-axis NC machining using adaptive space-filling curves. *International Journal of Production Research*, **43**, 1643-1665, 2005.
25. M.W. Bern, D. Eppstein, Quadrilateral meshing by circle packing. *International Journal Computational Geometry and Applications* **10**, 347-360, 2000.
26. P.J. Frey, P.-L. George, *Mesh Generation: Application to Finite Elements*, ISTE 2007.
27. 3D Systems Inc, Stereolithography Interface Specification. P/N 50065-S01-00, 1989.
28. The Initial Graphics Exchange Specification (IGES) Version 5.0. National Institute of Standards and Technology, Gaithersburg, MD, 1990.
29. ISO 10303-21:2002 Industrial automation systems and integration -- Product data representation and exchange -- Part 21: Implementation methods: Clear text encoding of the exchange structure, 2002.
30. Automatisation industrielle. Representation externe des donnees de definition de produits. Specification du standard d'echange et de transferts (SET), Version 85-08, Z68-300, Association Francaise de Normalisation (AFNOR) 85181, Paris. 1985.
31. http://en.wikipedia.org/wiki/CATIA, June 1, 2015.
32. http://en.wikipedia.org/wiki/PHIGS, June 1, 2015.
33. D.L. Marcum, J.A. Gaiter, Unstructured surface grid generation using global mapping and physical space approximation, *8th International Meshing Roundtable, CA, October 10-13, 1999.*
34. M. Eck, T. DeRose, T.Duchamp, H.Hoppe, M. Lounsbery, W. Stuetzle, Multiresolution analysis on arbitrary meshes. *The 22nd Annual Conference on Computer Graphics and Interactive Techniques,* Los Angeles,USA,August 06 - 11, 1995

35. M.S. Floater,Parametrization and smooth approximation of surface triangulations. *Computer Aided Geometric Design*, **14**, 231-250 (1997).
36. A. Sheffer, E.de Sturler, Parameterization of faceted surfaces for meshing using angle-based flattening, *Engineering with Computers*. **17**, 326-337, 2001.
37. J. McCartney, B.K. Hinds, B.L Seow, The flattening of triangulated surfaces incorporating darts and gussets, *Computer-Aided Design*, **31**, 249-260, 1999.
38. Y. Zhong, B. Xu, A physically based method for triangulated surface flattening. *Computer-Aided Design*, **38**, 1062-1073, 2006.
39. J. Xu, C. Jin, Boundary-conformed machining for trimmed free-form surfaces based on mesh mapping. *International Journal of Computer Integrated Manufacturing*, **26**, 720-730, 2013.
40. L. Saroul, Surface extraction and flattening for anatomical visualization. PhD. Thesis Université de Saint-Etienne, France, 2006.
41. http://www.resurf3d.com/MeshFlatten_standalone.htm (June 1, 2015).
42. S. Rusinkiewicz, Estimating curvatures and their derivatives on triangle meshes. *3D Data Processing, Visualization, and Transmission, 2nd International Symposium*, Tessaloniki, Greece, September 6-9, 2004.
43. P. Hu, K. Tang, Five-axis tool path generation based on machine-dependent potential field. *International Journal of Computer Integrated Manufacturing* (in print) DOI:101080/0951192X20151068451.
44. J.U. Brackbill, J.S. Saltzman, Adaptive zoning for singular problems in two dimensions. *Journal of Computational Physics*. **46**, 342-368, 1982.
45. J.E.Castillo, J.S. Otto, Numerical techniques for the transformation to an orthogonal coordinate system aligned with a vector field. *Computers & Mathematics with Applications*. **40**, 523-535, 2000.
46. A.E. Giannakopoulos, A.J. Engel, Directional control in grid generation. *Journal of Computational Physics*, **74**, 422-439, 1988.
47. S.S. Makhanov, W. Anotaipaiboon, *Advanced numerical methods to optimize cutting operations of five axis milling machines*. Springer, 2007.
48. M. Gaspar, F.Weichert, Integrated construction and simulation of tool paths for milling dental crowns and bridges. *Computer-Aided Design*, **45**, 1170-1181, 2013.
49. T.V. Vorburger, J. Raja, *Surface Finish Metrology Tutorial*. NISTIR 89-4088: National Institute of Standards and Technology, Gaithersburg, MD, June 1990.
50. A, Tabenkin, Function: the key to surface finish requirements. *4th Biennial International Manufacturing Technology Conference*, McLean, VA, 1988.
51. American National Standard. ASME/ANSI B46.1-1985. Surface Texture. American Society of Mechanical Engineers, 1985.
52. M.F. Siladic, Numerical grid generation and potential airfoil analysis and design. M.Sc. Thesis, The University of Texas of Austin, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN 'A PRIORI' MODEL REDUCTION FOR ISOGEOMETRIC BOUNDARY ELEMENT METHOD

**Shengze Li[1], Jon Trevelyan[2], Weihua Zhang[1] and Zhuxuan Meng[1]**

[1]College of Aerospace Science and Engineering, National University of Defense Technology
109 Deya Road, Changsha, China
e-mail: lishengze12@gmail.com, zwhkjs@163.com

[2] School of Engineering and Computing Sciences, Durham University
South Road, Durham, UK
e-mail: jon.trevelyan@durham.ac.uk

**Keywords:** Isogeometric Analysis, Boundary Element Method, Model Reduction, Karhunen-Loève Decomposition, Krylov's Subspace.

**Abstract.** *The aim of this work is to provide a promising way to accelerate the structural design procedure and overcome the burden of meshing. The concept of Isogeometric Analysis (IGA) has appeared in recent years and has become a powerful means to eliminate gaps between Computer Aided Design (CAD) and Computer Aided Engineering (CAE), giving a higher fidelity geometric description and better convergence properties of the solution. The Boundary Element Method (BEM) with IGA offers a better, more seamless integration since it uses a boundary representation for the analysis. However, the computational efficiency of IGABEM may be compromised by the dense and unsymmetrical matrix appearing in the calculation, and this motivates the present work. This study introduces an 'a priori' model reduction method in IGABEM analysis aiming to enhance efficiency. The problem is treated as a state evolution process. It proceeds by approximating the problem solution using the most appropriate set of approximation functions, which depend on Karhunen-Loève decomposition. Secondly, the model is re-analysed using a reduced basis, using the Krylov subspaces generated by the governing equation residual for enriching the approximation basis. Finally, the IGABEM calculation combined the model reduction strategy is proposed, which provides accurate and fast re-solution of IGABEM problems compared with the traditional BEM solution. Moreover, the CPU time is drastically reduced. A simple numerical example illustrates the potential of this numerical technique.*

## 1 INTRODUCTION

The Boundary Element Method (BEM) [1] is a domain discretisation technique for solving partial differential equations, which takes advantage of boundary integral equations to decrease the dimensionality of a problem by one, *i.e.*, only line integrals for 2D and surface integrals for 3D problems, and thus a smaller system of equations will be generated. The cost of mesh generation is reduced as a surface mesh generation is much more easier than domain mesh.

In a traditional structural design process, mesh generation is an essential task to bridge from CAD geometries to numerical analysis tools, and this can take a considerable amount of time. To alleviate this burden, the idea of Isogeometric Analysis (IGA) [2] has received much attention in recent years as it offers precise and efficient geometric modelling, refinement without re-meshing the CAD model and control over the smoothness of the basis functions. This provides a uniform representation for the design and analysis models, significantly reducing the overall analysis time including the creation and refinement of the analysis model.

The fact that both CAD and BEM require only a boundary representation suggest that there is much more scope for further development in linking BEM to CAD geometries than the Finite Element Method (FEM). This natural connection motivated many researchers to consider some early attempts to include CAD representations in a BEM framework. Arnold [3, 4] proposed a spline collocation method which is used to develop convergence estimates for BEM in two and three dimensions. A BEM formulation based on cubic splines[5, 6] was discovered to solve groundwater flow problems and the Laplace equation. Turco [7] presented an approach by which the elasticity problem could be solved by B-spline elements. Rivas [8] firstly put forward the combination of BEM with rational non-uniform B-splines (NURBS) in the context of the method of moments.

Recently, much of the literature on isogeometric analysis focuses on the combination with FEM. For BEM, the existing body of literature is more limited. Simpson [9] implemented the isogeometric boundary element method (IGABEM) for elastostatic analysis in two dimensions, and further Marussig [10] extended the IGABEM to three dimensions by using a hierarchical matrix which reduced the computational complexity for large scale analysis. It should be noticed that the drawback of IGABEM and traditional BEM remains, that it will generate a fully populated matrix. 3D problems present challenges in computational complexity if they have a large number of degrees of freedom (DOF). Hence, in order to mitigate this drawback, model reduction techniques will be a possible way. These provide some more efficient approximation bases which can be used for the further calculation.

One use of model reduction techniques is derived from some problems of random data processing [11], which provide a small number of shape functions in order to represent the spatially distributed state of a system. The reduced order models (ROMs) are interesting to reduce the cost of parametric studies on the state evolution of the system. This method is further used in image processing [12], and so far, it also has been successfully applied in some finite element frameworks [13, 14]. As the key of model reduction is to find the inner correlation of an evolution process, it is also suitable for the BEM analysis process as well if it follows some rules such like continuously varying parameters in time or space.

In this article, the combination of IGABEM and model reduction will be presented, and further, this framework will be applied in a reanalysis process which is tested by a quarter cylinder model under inner pressure to accelerate greatly the process of undertaking a parametric study.

## 2 B-SPLINES AND NURBS

A B-spline is a group of piecewise polynomials which are defined by a knot vector

$$\Xi = \{\xi_1, \xi_2, \ldots, \xi_{n+p+1}\} \quad \xi_A \in \mathbb{R} \tag{1}$$

which is a set of non-decreasing real numbers in the parametric space. Here, $A$ denotes the knot index, $p$ the curve order, and $n$ the number of basis functions or control points. Each real number $\xi_A$ is called a knot. The number of knots in a valid knot vector is always $n + p + 1$. The half open interval $[\xi_i, \xi_{i+1}]$ is called a knot span. The basis function $N_{A,p}$ could be defined using the Cox-de Boor recursion formula [15, 16]

$$N_{A,0}(\xi) = \begin{cases} 1 & \text{if } \xi_A \leqslant \xi < \xi_{A+1} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$N_{A,p}(\xi) = \frac{\xi - \xi_A}{\xi_{A+p} - \xi_A} N_{A,p-1}(\xi) + \frac{\xi_{A+p+1} - \xi}{\xi_{A+p+1} - \xi_{A+1}} N_{A+1,p-1}(\xi) \tag{3}$$

The B-spline geometry is found through a mapping from parametric space to physical space through a linear combination of B-spline functions, which are defined in parametric space, and the corresponding coefficients may be considered to be the geometric coordinates of a set of control points scattered in physical space. The B-spline curve can be expressed as

$$\boldsymbol{x}(\xi) = \sum_{A=1}^{n} N_{A,p}(\xi) \mathbf{P}_A \tag{4}$$

where $\boldsymbol{x}(\xi)$ is the location of the physical curve, $\xi$ the spatial coordinate in parameter space, $\mathbf{P}_A$ the control point and $N_{A,p}$ the B-spline basis function of order $p$.

Multivariate basis functions are defined by tensor products of univariate basis functions of parametric direction, given by

$$N_A(\boldsymbol{\xi}|\boldsymbol{\Xi}_A) \equiv \prod_{i=1}^{d_p} N_A^i(\xi_A^i|\Xi_A^i) \tag{5}$$

where $i$ denotes the direction index and $d_p$ is the number of dimensions.

Non-uniform rational B-splines (NURBS) are developed from B-splines but can offer significant advantages due to their ability to represent a wide variety of geometric entities. Importantly for engineering applications, they have the ability to describe circular arcs exactly, whereas the traditional piecewise polynomial FEM description cannot. The expression defining a NURBS curve is similar to that with B-splines,

$$\boldsymbol{x}(\xi) = \sum_{A=1}^{n} R_{A,p}(\xi) \mathbf{P}_A \tag{6}$$

here, $\mathbf{P}_A$ is the set of control point coordinates, $R_{A,p}$ is the NURBS basis function which is given by

$$R_{A,p}(\xi) = \frac{N_{A,p}(\xi) w_A}{W(\xi)} \tag{7}$$

with

$$W(\xi) = \sum_{A=1}^{n} w_A N_{A,p}(\xi) \tag{8}$$

where $w_A$ denotes a weight associated to each basis function or control point. It can influence the distance between the associated control point and the NURBS geometry; with higher values, the curve will be closer to that point. If all of the weights are equal to 1, the NURBS curve degenerates into a B-spline curve.

## 3 ISOGEOMETRIC BEM

### 3.1 Boundary integral equation

For a linear elastic problem, the structure occupies a continuous domain $\Gamma$, with the boundary conditions

$$\begin{aligned} u &= \overline{u} \quad \text{on } \Gamma_u \\ t &= \bar{t} \quad \text{on } \Gamma_t \end{aligned} \tag{9}$$

where the domain boundary $\Gamma = \Gamma_u + \Gamma_t$.

The Boundary Integral Equation (BIE) can be written as follows:

$$u_j(\boldsymbol{s}) + \int_\Gamma T_{ij}(\boldsymbol{s}, \boldsymbol{x}) u_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) = \int_\Gamma U_{ij}(\boldsymbol{s}, \boldsymbol{x}) t_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) \tag{10}$$

where $\boldsymbol{s}$ is the 'load' point and $\boldsymbol{x}$ the 'field' point, $u_i$ the displacement field, $t_j$ the traction field, $U_{ij}$ the displacement fundamental solution, $T_{ij}$ the traction fundamental solution and $i$, $j$ are the indices running from 1 to 3 in three dimensions. Eq. 10 also can be written as a discretised form which is the traditional BEM formulation

$$\begin{aligned} C_{ij}(\boldsymbol{s}_c) u_j^{e_0 a_0} + \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} u_j^{ea} \int_{\tilde{S}} T_{ij}(\boldsymbol{s}_c, \tilde{\boldsymbol{\xi}}) N_{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \\ = \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} t_j^{ea} \int_{\tilde{S}} U_{ij}(\boldsymbol{s}_c, \tilde{\boldsymbol{\xi}}) N_{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \end{aligned} \tag{11}$$

where $C_{ij}$ indicates the jump term, $c$ the collocation point index, $e_0$ the element in which the collocation point is located, and $a_0$ the local index of the collocation point in element, $e$ is the element index and $a$ the local index of the node in element $e$, $\tilde{\boldsymbol{\xi}}$ the intrinsic coordinates in parent element, $N_{ea}$ the shape function, $J_e$ the Jacobian and $\tilde{S}$ the space spanned by intrinsic coordinates.

The integrals in the above can be written in matrix form

$$\mathbf{H}\boldsymbol{u} = \mathbf{G}\boldsymbol{t} \tag{12}$$

where matrix $\mathbf{H}$ is a square matrix containing a combination of the integrals of the $T_{ij}$ kernel and the jump terms, $\mathbf{G}$ is a rectangular matrix of $U_{ij}$ kernel integrals, $\boldsymbol{u}$ contains the nodal displacements and $\boldsymbol{t}$ the nodal tractions. Both $\boldsymbol{u}$ and $\boldsymbol{t}$ include unknown value and the value prescribed by boundary conditions. By swapping the unknowns and the related coefficients of both sides, Eq. 12 takes the final form

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{f} \tag{13}$$

where $\mathbf{A}$ is a coefficient matrix which is usually non-symmetric and densely populated, the vector $\boldsymbol{x}$ contains all unknown displacement and traction component and the vector $\boldsymbol{f}$ is the summation of all known coefficient. The above equation is a linear system which can be solved to obtain the values of the unknown displacement and tractions.

### 3.2 Isogeometric approximation

The implementation of IGABEM is similar to the traditional BEM with the concept of isoparametric elements; they both discretise the BIE and geometry using the same shape functions. Hence, the displacement and traction fields around the boundary could also be discretised using NURBS, which is the main difference from the traditional BEM,

$$u_j^e(\tilde{\boldsymbol{\xi}}) = \sum_{a=1}^{n_a} R_{ea}(\tilde{\boldsymbol{\xi}}) \tilde{u}_j^{ea} \tag{14}$$

$$t_j^e(\tilde{\boldsymbol{\xi}}) = \sum_{a=1}^{n_a} R_{ea}(\tilde{\boldsymbol{\xi}}) \tilde{t}_j^{ea} \tag{15}$$

where $\tilde{u}_j^{ea}$ and $\tilde{t}_j^{ea}$ are the nodal displacement and traction parameters associated with control points, and $\tilde{\boldsymbol{\xi}}$ the intrinsic coordinates of the field point in the parent element. It should be noted that $\tilde{u}_j^{ea}$ and $\tilde{t}_j^{ea}$ are no longer the nodal displacements or tractions as the NURBS basis functions lack the Kronecker delta property. Eq. 10 can be rearranged by separating the integrals into two parts

$$\begin{aligned} & u_j(\boldsymbol{s})|_{\Gamma_t} + \int_{\Gamma_t} T_{ij}(\boldsymbol{s},\boldsymbol{x}) u_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) - \int_{\Gamma_u} U_{ij}(\boldsymbol{s},\boldsymbol{x}) t_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) \\ = & -\bar{u}_j(\boldsymbol{s})|_{\Gamma_u} - \int_{\Gamma_u} T_{ij}(\boldsymbol{s},\boldsymbol{x}) \bar{u}_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) + \int_{\Gamma_t} U_{ij}(\boldsymbol{s},\boldsymbol{x}) \bar{t}_j(\boldsymbol{x}) d\Gamma(\boldsymbol{x}) \end{aligned} \tag{16}$$

where $\Gamma_t$ denotes the portion of prescribed traction boundary conditions and $\Gamma_u$ displacement boundary conditions. Here, $u_j$ and $t_j$ on the left are unknowns, $\bar{u}_j$ and $\bar{t}_j$ on the right are the known values given by boundary conditions. Then, the discretised form of the BIE could be written by discretising the left side of Eq. 16 with Eq. 14 and Eq. 15,

$$\begin{aligned} \sum_{a_0=1}^{n_{a_0}} C_{ij}(\tilde{\boldsymbol{\zeta}}_c) R_{e_0 a_0}(\tilde{\boldsymbol{\zeta}}_c) \tilde{u}_j^{e_0 a_0} \quad & + \quad \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} \int_{\tilde{S}_t} T_{ij}(\tilde{\boldsymbol{\zeta}}_c, \tilde{\boldsymbol{\xi}}) R_{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \tilde{u}_j^{ea} \\ & - \quad \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} \int_{\tilde{S}_u} U_{ij}(\tilde{\boldsymbol{\zeta}}_c, \tilde{\boldsymbol{\xi}}) R_{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \tilde{t}_j^{ea} \\ = -\sum_{a_0=1}^{n_{a_0}} C_{ij}(\tilde{\boldsymbol{\zeta}}_c) \bar{u}_j^{e_0 a_0}(\tilde{\boldsymbol{\zeta}}_c) \quad & - \quad \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} \int_{\tilde{S}_u} T_{ij}(\tilde{\boldsymbol{\zeta}}_c, \tilde{\boldsymbol{\xi}}) \bar{u}_j^{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \\ & + \quad \sum_{e=1}^{n_e} \sum_{a=1}^{n_a} \int_{\tilde{S}_t} U_{ij}(\tilde{\boldsymbol{\zeta}}_c, \tilde{\boldsymbol{\xi}}) \bar{t}_j^{ea}(\tilde{\boldsymbol{\xi}}) J_e(\tilde{\boldsymbol{\xi}}) d\tilde{S}(\tilde{\boldsymbol{\xi}}) \end{aligned} \tag{17}$$

where $c$ denotes the collocation point index, $\tilde{\boldsymbol{\zeta}}_c$ the intrinsic coordinate of the collocation point, $e_0$ the element of which the collocation point is located, and $a_0$ the local index of the collocation

point in element $e_0$. In addition, contrary to the traditional BEM, the control points in IGABEM may be inappropriate locations for collocation, since they may lie off the physical problem boundary. This requires us to compute feasible locations for the collocation points in parametric space as

$$\zeta_A = \frac{\xi_{A+1} + \xi_{A+2} + \cdots + \xi_{A+p}}{p} \tag{18}$$

By evaluating a series of boundary integrals, Eq. 17 may be assembled into a matrix form

$$\mathbf{A}\boldsymbol{x} = \boldsymbol{f} \tag{19}$$

where matrix $\mathbf{A}$ contains the entries of the left of Eq. 17, $\boldsymbol{f}$ the column vector of the right side and $\boldsymbol{x}$ includes all the unknown displacement and traction. Now, the linear system could be solved directly.

## 4 THE MODEL REDUCTION

### 4.1 The Karhunen-Loève decomposition

This method is also known as the Proper Orthogonal Decomposition (POD), which is a powerful and elegant method for data analysis aimed to obtaining low-dimensional approximate descriptions of a high-dimensional process. The concept is well established and can be seen in different algorithms, including the Karhunen-Loève Decomposition (KLD), the Principal Component Analysis (PCA) and the Singular-Value Decomposition (SVD).

For an arbitrary evolution process, a certain field could be defined by a discrete form written as $u(\boldsymbol{x}_i, s^p)$ , where $\boldsymbol{x}_i$ denotes the nodal displacement and $s^p$ the calculation step. It is also equivalent to $u^p(\boldsymbol{x}_i)$, $\forall p \in [1, \ldots, P]$, $\forall i \in [1, \ldots, N]$. The main idea of the KLD is to obtain the most typical or characteristic structure $\phi(\boldsymbol{x})$ among those $u^p(\boldsymbol{x})$, rather like a small number of vibration modes can encapsulate the vibration characteristics of a structure. This is equivalent to solving for a function $\phi(\boldsymbol{x})$ maximising $\alpha$ where

$$\alpha = \frac{\sum_{p=1}^{P} \left[ \sum_{i=1}^{N} \phi(\boldsymbol{x}_i) u^p(\boldsymbol{x}_i) \right]^2}{\sum_{i=1}^{N} (\phi(\boldsymbol{x}_i))^2} \tag{20}$$

which can be rewritten as

$$\sum_{i=1}^{N} \left\{ \sum_{j=1}^{N} \left[ \sum_{p=1}^{P} u^p(\boldsymbol{x}_i) u^p(\boldsymbol{x}_j) \phi(\boldsymbol{x}_j) \right] \tilde{\phi}(\boldsymbol{x}_i) \right\} = \alpha \sum_{i=1}^{N} \tilde{\phi}(\boldsymbol{x}_i) \phi(\boldsymbol{x}_i) \tag{21}$$

where $\tilde{\phi}(\boldsymbol{x})$ denotes the variation of $\phi(\boldsymbol{x})$. Introducing a vector notation, Eq. 21 takes the following matrix form

$$\tilde{\boldsymbol{\phi}}^{\mathrm{T}} \mathbf{K} \boldsymbol{\phi} = \alpha \tilde{\boldsymbol{\phi}}^{\mathrm{T}} \boldsymbol{\phi}; \qquad \forall \tilde{\boldsymbol{\phi}} \Rightarrow \mathbf{K}\boldsymbol{\phi} = \alpha \boldsymbol{\phi} \tag{22}$$

The two point correlation matrix $\mathbf{K}$ is given by

$$\mathbf{K}_{ij} = \sum_{p=1}^{P} u^p(\boldsymbol{x}_i) u^p(\boldsymbol{x}_j) \tag{23}$$

whose matrix form becomes

$$\mathbf{K} = \sum_{p=1}^{P} \boldsymbol{u}^P (\boldsymbol{u}^P)^{\mathrm{T}} \tag{24}$$

which is symmetric and positive definite. If the matrix $\mathbf{Q}$ can be defined as a discrete field history

$$\mathbf{Q} = \begin{bmatrix} u_1^1 & u_1^2 & \cdots & u_1^P \\ u_2^1 & u_2^2 & \cdots & u_2^P \\ \vdots & \vdots & \ddots & \vdots \\ u_N^1 & u_N^2 & \cdots & u_N^P \end{bmatrix} \tag{25}$$

the matrix $\mathbf{K}$ will be rewritten as

$$\mathbf{K} = \mathbf{Q}\mathbf{Q}^{\mathrm{T}} \tag{26}$$

Thus the characteristic structure of the displacement field $u^p(\boldsymbol{x})$ is presented as the eigenvectors $\phi_k(\boldsymbol{x}) \equiv \boldsymbol{\phi}_k$ associated with the largest eigenvalues.

### 4.2 Reduced model construction

If some direct calculations have been carried out previously, the nodal displacements can be denoted as $u(\boldsymbol{x}_i, s_p) \equiv u_i^p, \forall i \in [1, \ldots, N], \forall p \in [1, \ldots, P]$. The eigenvalues can be considered in a descending sequence, and eigenvectors included in the reduced basis if the corresponding eigenvalue $\lambda_k > 10^{-10}\lambda_1$, $\forall k \in [1, \ldots, n]$, $\lambda_1$ being the largest eigenvalue. Then, those $n$ eigenvectors related to the eigenvalues above could be used for generating an approximating basis for further solutions. The $(N \times n)$ matrix $\mathbf{\Phi}$ can be defined as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1^1 & \phi_1^2 & \cdots & \phi_1^n \\ \phi_2^1 & \phi_2^2 & \cdots & \phi_2^n \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N^1 & \phi_N^2 & \cdots & \phi_N^n \end{bmatrix} \tag{27}$$

where $N$ is the size of the full (i.e. unreduced) system and $n$ is the number of eigenvectors considered. The solution vector in Eq. 19 can be written as

$$\boldsymbol{x} = \sum_{i=1}^{n} y_i \boldsymbol{\phi}_i = \mathbf{\Phi}\boldsymbol{y} \tag{28}$$

which can be substituted into Eq. 19 to obtain

$$\mathbf{A}\mathbf{\Phi}\boldsymbol{y} = \boldsymbol{f} \tag{29}$$

We now premultiply both sides by $\mathbf{\Phi}^{\mathrm{T}}$

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{A}\mathbf{\Phi}\boldsymbol{y} = \mathbf{\Phi}^{\mathrm{T}}\boldsymbol{f} \tag{30}$$

If the original coefficient matrix $\mathbf{A}$ is $N \times N$, this procedure causes the size of the reduced system (30) to decrease to $n \times n$, providing the final low dimensional system.

### 4.3 An 'a priori' model reduction strategy

As a design geometry continuously evolves, perhaps as part of an optimisation scheme, the basis matrix $\mathbf{\Phi}$ should be updated to fit the new condition and maintain the quality of each new solution. If the basis matrix $\mathbf{\Phi}$ is generated from solutions at the first $P$ steps, and another $S$ steps considered with this basis, then after these $S$ steps the residual of the system should be evaluated as

$$\mathbf{R} = \mathbf{A}\boldsymbol{x} - \boldsymbol{f} = \mathbf{A}\mathbf{\Phi}\boldsymbol{y} - \boldsymbol{f} \tag{31}$$

If the norm of residual is sufficiently small, i.e. $\parallel \mathbf{R} \parallel < \epsilon$, with $\epsilon$ a threshold value small enough, the next $S$ steps will be continued efficiently with the existing reduced basis. On the other hand, if $\parallel \mathbf{R} \parallel > \epsilon$, the approximation basis should be enriched and computations repeated for the last $S$ steps. The enrichment is built using a Krylov subspace, the new basis matrix being defined as

$$\mathbf{\Phi}^* = \{\mathbf{\Phi}\mathbf{V}, \mathbf{R}, \mathbf{A}\mathbf{R}, \mathbf{A}^2\mathbf{R}\} \tag{32}$$

where $\mathbf{V}$ is the combination of the most representative eigenvectors which is from the previous reduced result $\boldsymbol{y}$, and a new solution vector could be written as

$$\boldsymbol{y}^* = \left[\mathbf{\Phi}^{*\mathrm{T}}\mathbf{\Phi}^*\right]^{-1}\mathbf{\Phi}^{*\mathrm{T}}\mathbf{\Phi}\boldsymbol{y} \tag{33}$$

## 5 NUMERICAL EXAMPLE

The problem is defined as a single quarter cylinder which contains an internal pressure of 1 MPa, the left, bottom, and back surface are set as $x, y$ and $z$-direction displacement constraint (Fig. 1). The control points for model construction are shown in Fig. 2. The inner radius is $R1 = 0.4\mathrm{m}$, outer radius $R2 = 1\mathrm{m}$, the length $L = 1\mathrm{m}$, and the material properties of steel are used. This problem is solved first for $P = 5$ steps using the full (unreduced) IGABEM system in order to generate the approximation matrix. These five models consider the inner radius $R1 = \{0.4, 0.5, 0.6, 0.7, 0.8\}$. For the set of solution vectors generated from these problems, only 3 eigenvalues satisfy the criterion for inclusion ($\alpha_k > 10^{-10}\alpha_1$), so the full ($78 \times 78$) matrix is reduced through POD to a $3 \times 3$ matrix for solving a large number of problems considering different values of $R1$. We let the inner radius $R1$ continuously increase in 0.01m steps from $0.4\mathrm{m}$ to $0.8\mathrm{m}$, which is 41 calculation steps. After each 5 steps, the quality of the reduction result is checked, and the basis enriched using the Krylov subspace.

Fig. 3 compares the radial displacement at the inner surface between the standard IGABEM and the reduced IGABEM. As the figure shows, the reduced model provides an accurate result with a lower dimensional computation, the error remaining within $0.23\%$.

## 6 CONCLUSIONS

In this article, the Isogeometric Boundary Element Method is combined with a model reduction approach based on the Karhunen-Loève decomposition and Krylov subspace to provide a computational strategy for the repeated analysis of similar geometries. The size of the linear system may be very considerably decreased without significant loss of accuracy. The numerical example proves the potential of this method. Further work, applying this method within a structural optimisation process and for more complicated structures, will continue.

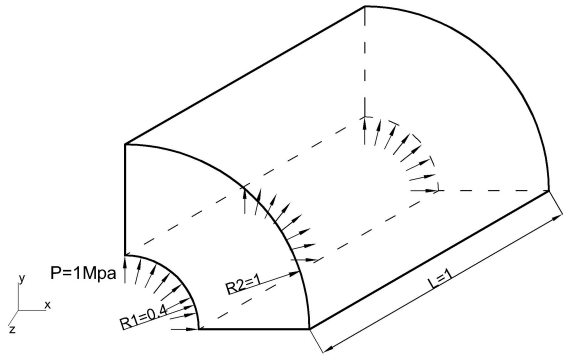



Figure 1: Geometry of the example

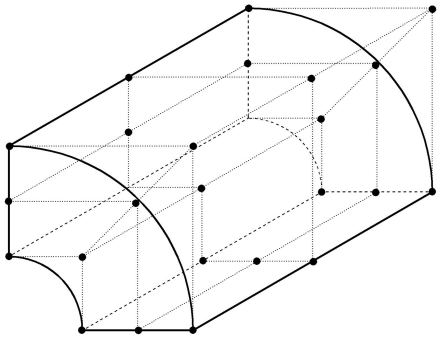


Figure 2: Control points

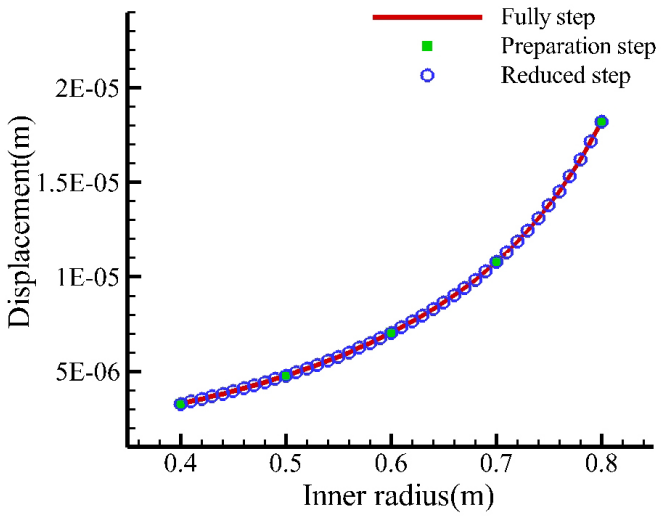



Figure 3: Radial displacement at the inner surface for each step

## REFERENCES

[1] T. A. Cruse, Numerical solutions in three dimensional elastostatics. *International Journal of Solids and Structures*, **5**, 1259-1274, 1969.

[2] T. J. R. Hughes, J. A. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Computer methods in applied mechanics and engineering*, **194**, 4135-4195, 2005.

[3] D. N. Arnold, W. L. Wendland, On the asymptotic convergence of collocation methods. *Mathematics of Computation*, **41**, 349-381, 1983.

[4] D. N. Arnold, J. Saranen, On the asymptotic convergence of spline collocation methods for partial-differential equations. *SIAM Journal on Numerical Analysis*, **21**, 459-472, 1984.

[5] J. A. Ligget, J. R. Salmon, Cubic spline boundary elements. *International Journal for Numerical Methods in Engineering*, **17**, 543-556, 1981.

[6] M. Yu, E. Kuffel, Spline element for boundary element method. *IEEE Transactions on Magnetics*, **30**, 2905-2907, 1994.

[7] E. Turco, M. Aristodemo, A three-dimensional B-spline boundary element. *Computer Methods in Applied Mechanics and Engineering*, **155**, 119-128, 1998.

[8] F. Rivas, L. Valle, M. F. Catedra, A moment method formulation for the analysis of wire antennas attached to arbitrary conducting bodies defined by parametric surfaces. *Applied Computational Electromagnetics Society Journal*, **11**, 32-39, 1996.

[9] R. N. Simpson, S. P. Bordas, J. Trevelyan, T. Rabczuk, A two-dimensional isogeometric boundary element method for elastostatic analysis. *Computer Methods in Applied Mechanics and Engineering*, **209**, 87-100, 2012.

[10] B. Marussig, J. Zechner, G. Beer, T. P. Fries, Fast isogeometric boundary element method based on independent field approximation. *Computer Methods in Applied Mechanics and Engineering*, **284**, 458-488, 2015.

[11] A. K. Jain, A fast Karhunen-Loeve transform for a class of random processes. *NASA STI/Recon Technical Report*, **76**, 42860, 1976.

[12] M. Kirby, L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **12**, 103-108, 1990.

[13] D. Ryckelynck, A priori hyperreduction method: an adaptive approach. *Journal of Computational Physics*, **202**, 346-366, 2005.

[14] D. Ryckelynck, F. Chinesta, E. Cueto, A. Ammar, On the a priori Model Reduction: Overview and Recent Developments. *Archives of Computational Methods in Engineering*, **13**, 91-128, 2006.

[15] M. G. Cox, The numerical evaluation of B-splines. *IMA Journal of Applied Mathematics*, **10**, 134149, 1972.

[16] C. De Boor, On calculating with B-splines. *Journal of Approximation Theory*, **6**, 5062, 1972.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# RESEARCH OF DEFORMATION QUALITIES OF POROUS MATERIALS ENHANCING COMPUTER DESING OF WETSUIT

**Cherunova I.[1], Sirota E.[1], Kornev N.[2] , Paschen M.[2],**

**Schreier S.[2], Lesnikova T.[1] and Cherunov P.[1]**

[1] Don State Technical University
Shevchenko Str. 147, 346500 Shakhty, Russia
e-mail: i_sch@mail.ru

[2] University of Rostock
A. Einstein Str.2, 18059 Rostock, Germany
e-mail: nikolai.kornev@uni-rostock.de

**Keywords:** Design of clothing, CAD systems, Expanded materials., Hydrostatic pressure, Underwater outfits.

**Abstract.** *The article presents the research in order to create a new level of computer tools for effective design of clothing for underwater kinds of activity. This report covers the consideration, sub-stantiation, and proposals relating to the ways for improving the mathematical models built into integrated CAD systems to be used in creation of underwater outfits. In order to ensure precise solutions for stocking the outfit object with materials with respect to the underwater conditions selected, the experimental studies of deformation of modern expanded polymer materials under the action of external pressure and local stretching have been performed in this research. Such multiaxial deformations lead to redistribution of the thickness of clothing being under investigation over the complicated surface of the man's body. A method of autonomous fixing the deformation of expanded material thickness has been developed for conduction of such investigations. Basing on the method proposed, the authors have developed a fundamental solution for a special device for autonomous measurements of the deformation characteristics of expanded polymer materials, in particular, in the conditions of direct hydrostatic pressure on the clothing. The results of experimental studies obtained have allowed to derive relations for the behaviour of deformation characteristics of expanded materials.*

## 1 INTRODUCTION

Rapid development of modern CAD/CAM technologies in clothing industry is increasingly using mathematical models of various type and purpose. Clothing design for underwater environments is based on conventional computer-aided-design (CAD) systems basis of which consists of mathematical models for obtaining construction parts and surface development, which are adequate to the external shape of the man's body under static and dynamic conditions [1]. In doing so, such clothing is a sophisticated system of work of expanded polymer materials aimed at ensuring heat and mechanical protection of the man's body [2].

In the manufacture of wetsuits, 3 types of design are noted:

- "Dry" — these wetsuits almost never pass the water inside. This effect is achieved by the use of packing seals at the hands and neck, and waterproof zippers. Such suits can be made of trilaminate or neoprene.
- "Wet" — these wetsuits are made of neoprene (foam rubber). Thermal insulation is provided by the suit material itself due to the presence of air bubbles. Water trapped under the wetsuit cannot be a heat insulator, therefore, the closer suit is fitting, the less is a circulation of water under it, and the less a heat is spent for heating of new cold-water portion.
- "Semi-dry" is an intermediate type of wetsuits. Despite the presence of seals, a water can penetrate to under-suit space (if air penetrates into it), but to a lesser extent, and, in close suit's fitting, a water does ingress practically, due to which the heat insulating properties is increased.

For wet type suits, a porous neoprene is used. For the manufacture of dry type wetsuits, a pressed neoprene is most commonly used (neoprene, which pores are partially or completely crushed), which increases reliability and reduces the buoyancy of the wetsuit. For spearfishing, an ordinary neoprene is used, but with the cut pore (there are unclosed pores on inside, due to which the material is snag against the body, a water circulation inside the suit is reduced, the man is getting warmer).

Neoprene can be used at -55 °C to + 90 °C temperatures, but the actual temperature range depends on the particular chemical composition of the material. Neoprene is resistant to the sunlight influence and chemically active oil products [3].

The average thickness of the material can range from water temperature and diving purposes (Table.1).

Table 1: Dependence of average thickness of the suit material on the water temperature

| Material thickness, mm | Water temperature, °C | Intended use |
|---|---|---|
| 9-10 | 4—12 | diving operations |
| 7 | 10—18 | spearfishing, diving |
| 5 | 17—23 | surfing, diving |
| 3 | 24 и выше | surfing, wind-kite surfing, other sports |

The thickness of diving suits depends on the structure of set of individual layers of materials. These materials have a feature of compression deformation under external loads. A particular importance is the change in thickness of such materials at the hydrostatic pressure. At that, the presented data of material thickness for underwater clothes lose their effectiveness because of change of thermal resistance of clothes and overall heat loss by the man. These

changes must be taken into account at the stage of engineering analysis in the CAD. This paper presents the engineering developments that allow determining the experimental changes in the geometric characteristics of the foam material under the influence of external pressure. The parts of the diving suits are exposed to the most pressure under the influence of the foot. Therefore, the integration of the results of experimental studies of foam materials for clothing is considered by the example of the design of support structures of the diving suit.

## 2. EXPERIMENTAL STADIES

### 2.1 Justification of special materials for the research

Support structures of diving suits can be made from several types of materials that are part of information provision in CAD-designing of underwater clothes (Fig.1):

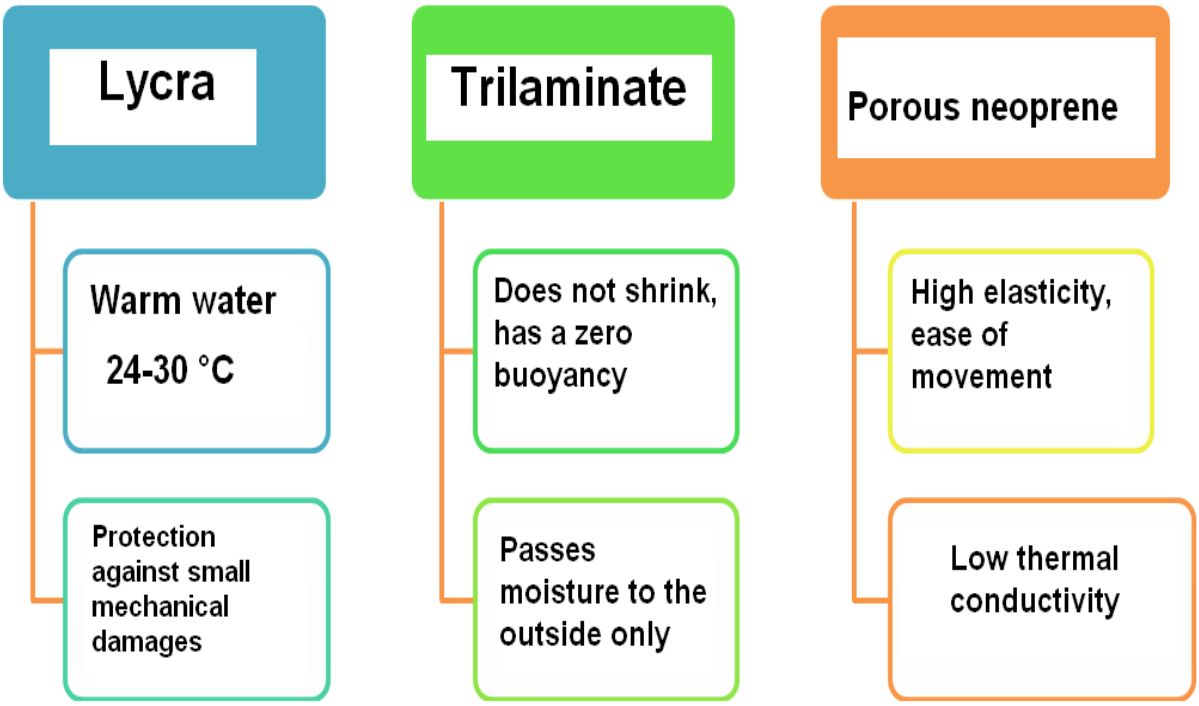



Figure 1: Main types of materials for the supporting structures of diving suits

### 2.2. Features of structure of “sandwich-type” material sets.

To determine the source data required in the automated calculation of clothing, the information registry of necessary materials for the supporting structures of diving suits has been established [4] (Table 2).

**Table 2:** Features of materials for wetsuits

| New materials | Manufacturing country | Composition | Benefits |
|---|---|---|---|
| Radially covered neoprene | Korea | super stretch (latex), neoprene, nano-titanium alpha, jersey nylon | High strength, durability |
| Nano-zirconium lycra | Japan | lycra-nylon, neoprene, nano-titanium alpha and aero-zirconium | Convenience, flexibility. Technology of transmission of solar heat to the human body. Allows to compare the heat of 5 mm fabric with 8 mm one. |
| Nano-super stretch | Japan | super stretch, neoprene, nano-titanium alpha, super stretch | Super stretch material increases the elasticity of neoprene in 2 times. Nano-titanium alpha is used for maximum heating |
| Nano-aqua block, neoprene | Japan | aqua block, neoprene, nano-titanium alpha, aqua block | "Aqua block" is a material that blocks the water from stagnation. Nano-titanium alpha provides maximum heating. |
| Nano-SCS Silver | Japan | super stretch, neoprene, nano-titanium silver | It retains 30 % more heat. |
| Nano-SCS Gold | Japan | super stretch, neoprene, nano-titanium gold | Retains 35 % more heat |
| Breathable fabric | Japan | cordura, polyurethane, polyurethane film, a net-like layer | Removes a moisture, the weight of suit is 3 kg.<br>Suitable for all-season dry suits |
| Butyl-rubber fabric | United Kingdom | lycra nylon, neoprene, nylon | Suitable for light dry suits. More durability and less volume |

As can be seen from the above analysis of materials used for the manufacture of wetsuits, neoprene is the basic material from which the wetsuits are made. The main part of "sandwich structures" is neoprene (more precisely, chloroprene),(Fig.2).

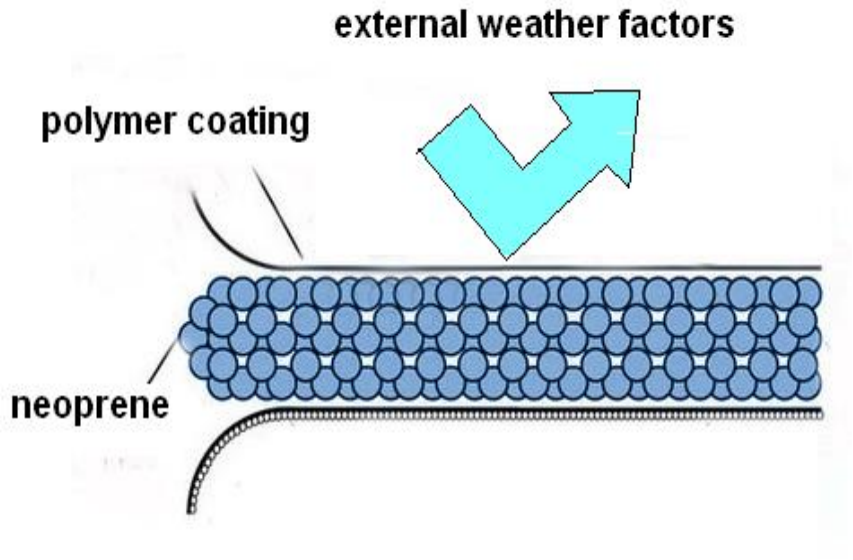



Figure 2: Neoprene

### 2.3 Basic properties of chloroprene materials for information database of CAD-system for designing the wetsuits.

Table 3: Properties of rubbers based on chloroprene rubbers

| Index | Unfilled rubbers | | Filled rubbers | |
|---|---|---|---|---|
| | regulator | | regulator | |
| | sulfur | thiol | sulfur | Thiol |
| Elongation tension 300 %, MPa | 1,0-1,5 | 1,9-2,3 | 8,5-9,5 | 17-18 |
| Relative elongation, % | 880-11000 | 780-900 | 450-550 | 450-550 |
| Tear resistance, kN/m (20 °C) | 30-45 | 25-35 | 55-70 | 55-65 |
| Rebound elasticity, % | 40-42 | 40-42 | 32-35 | 38-40 |
| Brittleness temperature, $^{o}$C | -37 | -37 | -37 | -37 |
| Compression set 20% (120 t, 100 °C), % | 80-90 | 35-40 | 80-85 | 45-53 |

The properties of such materials are determined by the characteristics of composite layers combined into the "sandwich type" general construction of the material. Mechanical and thermophysical properties of clothing from expanded sandwich type materials are greatly dependent upon the conditions of hydrostatic pressure and deformation with respect to the man's body surface as well as upon dynamic deformations during operational service of the suit [2,3]

### 2.4. Study the effect of pressure on the deformation of neoprene.

Material of clothes in service suffers loads and deformation that in general do not exceed rupture stress. At thus, when examining mechanical properties of materials, total strain and components thereof at single loadings below rupture stress are measured besides strength and rupture elongation characteristics.

A method of autonomous fixing the deformation of expanded material thickness has been developed for conduction of such investigations. Basing on the method proposed, the authors have developed a fundamental solution for a special device for autonomous measurements of the deformation characteristics of expanded polymer materials, in particular, in the conditions of direct hydrostatic pressure on the clothing (Fig.3).

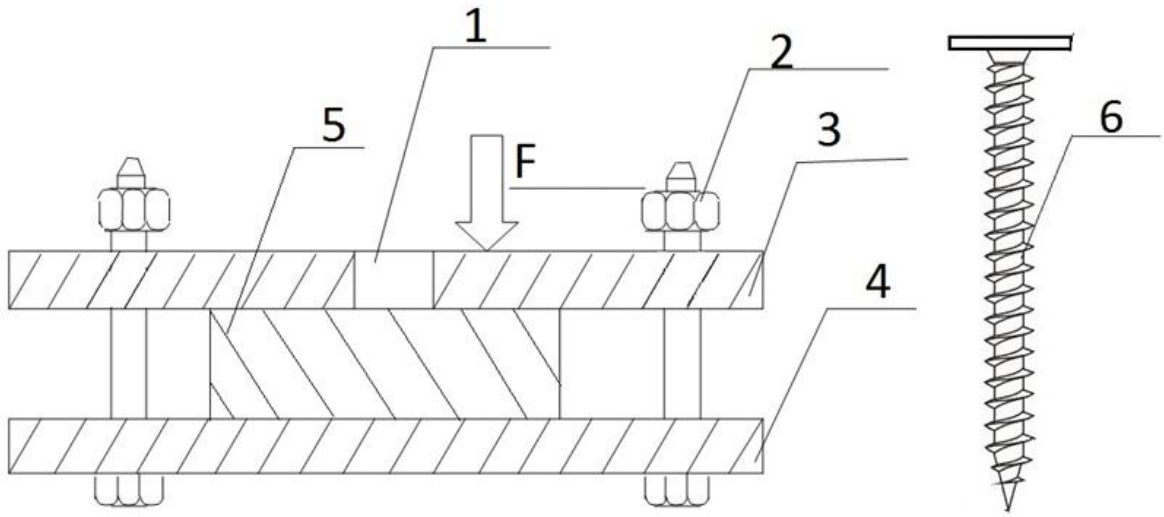



Figure 3: Device for compression of porous materials research (1-axis neoprene fixation; 2 - screw connection; 3 - the top support plate; 4 - a lower base plate; 5 - neoprene; 6 - screw clamp for a porous material).

The results of experimental studies obtained have allowed to derive relations for the behaviour of deformation characteristics of expanded materials. The examination resulted in derived dependency for further use in automated calculation of correlations of the diving dress' parts dimensional parameters:

$$\delta = -0{,}0002P + 0.0098$$

$$R^2 = 0.9943 \quad (1)$$

where: δ - Thickness, m.; $R^2$ - credibility factor.

## 2.5. Results for CAD wetsuit

The thicker the neoprene sample is, the less it is affected with strain changes. As to the plastic strain, the experiment had show that plastic strain depends upon both thickness of the sample and the coating of it.

Accounting derived results there was an algorithm for further integration in CAD systems elaborated [5] (Fig.4).

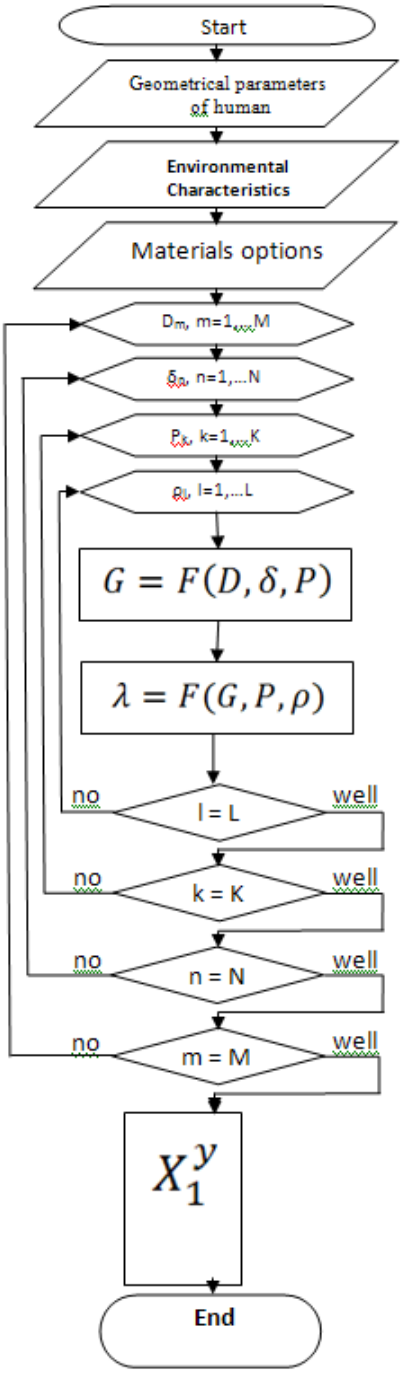



Figure 4: Algorithm CAD-Design wetsuits (D- deformation; δ - thickness, P - pressure, ρ - density, G- geometric index,λ - thermal conductivity, $X_1^y$- design factors).

## 3. CONCLUSIONS

The results allow enlarging of modern capabilities of diving dress' design using computer-aided methods, accounting for stress-strain behavior of materials, being complex physico-chemical systems, thus facilitating development of science and technology.

On the basis of the new relations for the behaviour of material properties in the elements of clothing construction, the authors have proposed a mathematical model and a basic algorithm of the program module for extending and improving the functional life cycle of modern CAD systems for functional clothing design.

**Acknowledgments. The authors express their acknowledgement to the Russian Ministry of Education and Science for the partial financial support of this work within the grant .( state task, performer Don State Technical University ) № 2838.**

## REFERENCES

[1] Wua D., Rosena D., Wangb L., Schaefera D. (2015) Cloud-based design and manufacturing: A new paradigm in digital manufacturing and design innovation // Computer-Aided Design . - Volume 59, February 2015, Pages 1–14

[2] Cherunov P., Cherunova I., Knyazeva S., Stenkina M., Stefanova E., Kornev N. (2015) The Development of the Research Techniques of Structure and Properties of Composite Textile Materials when Interacting with Viscous Fractions of Hydrocarbon Compounds / // The seven conference on "Textile Composites and Inflatable Structures" (Structural Membranes 2015) , 19-21 October 2015: International Center for Numerical Methods in Engineering. Barcelona, Spain.- 2015. – pp.555-564.

[3] Bardy E, Mollendorf J, Pendergast D. (2005) Thermal conductivity and compressive strain of foam neoprene insulation under hydrostatic pressure// Journal of Physics D: Applied Physics. 2005;38(20):3832-40.Cherunova I., Kornev N., Jacobi G., Treshchun I., Gro A., Turnow J., Schreier S. and

[4] MSL Engineering Limited. "Development of Design Guidance for Neoprene-Lined Clamps for Offshore Application". JIP Phase I Final Report, Doc. Ref. CH10010R005, Rev 1, May 1999.

[5] Cherunova I., Kornev N., Jacobi G., Treshchun I., Gro A., Turnow J., Schreier S. and-Paschen M. (2014). Application of computational fluid mechanics for protection clothes design. Proceedings of the 11th World Congress on Computational Mechanics (WCCM XI), 22.-25.07.2014, Barcelona, Spain, 7262–7273.

[6] Cherunova, I.V., Kuznetsov,D.M.,Cherunova,E.S. (2013) Acoustic emission of liquid at flexible porous materials impregnation// Material Sciens, 2013.-Issue:3, pg:37-41

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MATHEMATICAL MODELLING OF CITY AERODYNAMICS

**Svetlana A. Valger**[1], **Natalya N. Fedorova**[1,2], **Alexandr A. Fedorov**[1,2]

[1] Novosibirsk State University of Architecture and Civil Engineering (Sibstrin)
Leningradskaya Str., 113, Novosibirsk (Russia), swetla-ya@mail.ru

[2] Khristianovich Institute of Theoretical and Applied Mechanics SB RAS
Institutskaya Str., 4/1, Novosibirsk (Russia), {nfed, fedorov}@itam.nsc.ru

**Keywords:** city aerodynamics, wind actions, pedestrian comfort, turbulence, CAE, Fluent.

**Abstract.** *The article is devoted to the numerical simulation of turbulent air flows in the vicinity of building complexes. The simulation was performed under isothermal assumption on the basis of 3D URANS equations supplemented with the k-ω SST turbulence model. The ANSYS Fluent Software was used as the main modeling tool. At the first stage, the atmospheric flow in the neighborhood of a building of a complex shape was carried out taking into account the surrounding objects. The numerical simulation was performed under the conditions of the experiments. The 3D structure of the flow in the vicinity of the building was obtained and the comparison of the calculation results with the experimental data on the pressure coefficient distribution on the walls of the building was performed. At the second stage, we made the full-scale numerical modeling of the flow around the complex of high-rise buildings in the Frankfurt/Main, Germany, city center. According to the results of the numerical modeling the transient flow structure in the vicinity of the buildings was obtained and the mean and the fluctuating components of the wind load were estimated. The pedestrian comfortable/uncomfortable zones were also detected for this area of the city.*

## 1 INTRODUCTION

Aerodynamics of buildings is an area of applied research which has emerged in the recent years as an academic discipline examining the wind impacts on civil engineering structures disposed on a ground surface in the atmospheric boundary layer (ABL). In the recent years the density of urban environment tends to increase. When new high-rise objects are constructed within an already densely-built area, it leads to changes in air flow around the already existing buildings through vortex formation or higher wind velocities. The study of wind actions on buildings is one of the most important design stages since such actions affect significantly the mechanical strength of buildings. Wind actions accompanied by the acceleration due to structure vibrations under dynamic gusts may violate the normal service conditions in upper-floor premises of high-rise buildings.

The study of air flow in the vicinity of buildings is especially important for design of additional structural elements and ventilated facades. Caving of facades elements under the strong wind conditions is often observed. To prevent, specific calculation methods have to be developed which take into account the inhomogeneous distribution of wind load and the presence of local zones of negative pressure on the building walls. Another important task is to provide a comfortable environment for human activity, which requires a study of urban pedestrian comfort.

Presently, there are two approaches to solving building aerodynamics problems: the theoretical approach and the experimental one. The theoretical approach is based on rational mechanics methods, in particular, on fluid dynamics, and it applies mathematical models to describe transient turbulent air flows. For solving the initial-boundary-value problems, both theoretical methods and numerical algorithms are used. In the latter half of the XX century, the basic concepts of hydrodynamic processes in atmosphere were formulated [1 – 3] and, in this way, a certain theoretical basis for evaluating the wind actions was developed [4 – 6]. Progress in computing means and numerical methods has enabled the use of computer science in solving hydro- and gas-dynamics problems, including problems in the exterior building aerodynamics. The experimental approach is based on wind-tunnel studies [7 – 11] followed by treatment of the experimental data aimed to get empirical engineering dependences to describe the wind actions on various objects [12 – 14]. It should be noted, that the experimental modeling is time and cost consuming. Moreover, experiments are significantly limited by parameters of wind facilities. For example, it is quite difficult to reproduce the atmospheric boundary layer in the experiment. As for semi-empirical engineering methods, they permit evaluation of the mean and fluctuating wind-load components only for very simple configurations and fail to account for the interference effects between buildings as well as for complex phenomena of self-exciting oscillations and resonances arising under external dynamic actions. It leads to the necessity to develop new approaches that allow considering the interference effects in the air flow to predict wind loads and pedestrian comfort zones location. Thus it is reasonable to use methods based on computational experiments, which allow us to reduce time and cost to optimize engineering project and to dump excessive oscillations. Such methods may be based on the use of CAD/CAE software packages [15], [16].

In [17], the authors carried out the numerical simulation of turbulent flows of incompressible air in the vicinity of a plate-mounted prism of a square cross-section. The comparison with experimental data [18] on the velocity and turbulent kinetic energy (TKE) profiles was performed and satisfactory agreement was obtained. However, it becomes especially important to study the interference effects in flows around more complex configurations, for example for building complexes in cities.

In the paper, the numerical simulation of the turbulent air flows in the vicinity of the buildings complexes was performed. At the first stage, the atmospheric flow in the neighborhood of a building of a complex shape was carried out taking into account the interference with surrounding objects. The numerical simulation was performed under the conditions of the experimental data obtained in 3-AT-17.5/3 aerodynamic facility [19]. The 3D structure of the flow in the vicinity of the building was obtained and the comparison of the calculation results with the experimental data on the pressure coefficient distribution on the walls of the building was performed. At the second stage, the full-scale numerical modeling of the flow around the complex of high-rise buildings in the Frankfurt/Main, Germany, city center was made and pedestrian comfort of the area was investigated.

## 2 METHOD OF COMPUTATIONS

The simulation was performed under isothermal assumption on the basis of 3D URANS equations [20] supplemented with the *k-ω SST* turbulence model [21] and Kato-Launder [22] correction. To solve the initial-boundary problem, the finite-volume method and the method of splitting taking physical processes into account were used [23]. A monotonic solution was obtained using the MUSCL scheme of third-order accuracy for convective terms and the central-difference scheme of second-order accuracy for the viscous terms. For the temporal approximations, an implicit scheme of second-order accuracy was used. The ANSYS Fluent Software [15] supplemented with user-defined functions for settings the initial and boundary conditions was used.

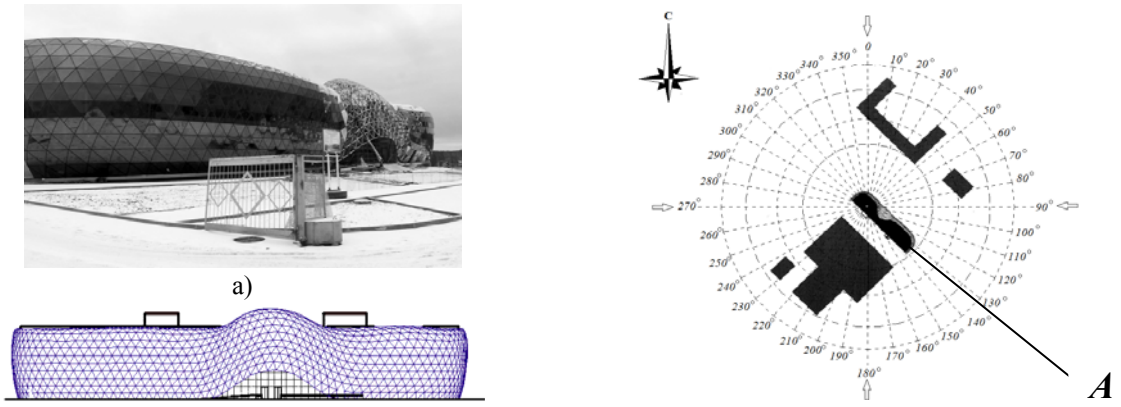
## 3 AIR FLOW IN THE VICINITY OF BUILDING OF COMPLEX SHAPE

### 3.1 Problem statement, boundary conditions and computational grid

Let us consider a turbulent isothermal air flow in the neighborhood of a building of complex shape (Fig. 1, a, b). The numerical simulation was performed under the conditions of the experimental data obtained in 3-AT-17.5/3 aerodynamic facility [19]. The model is of 1:150 to the real scale. The main goal of the investigation was to get pressure coefficient distribution on the walls of the building of a complex shape (building *A*, fig. 1, c). Two cases were examined in the paper: when the building *A* was surrounded by another buildings (*conf. 1*) and when it was isolated (*conf. 2*). The incoming flow velocity is $U_\infty$= 16.6 m/s and the boundary-layer thickness is $\delta$ = 0.92 m for these cases. The Reynolds number calculated by length scale $\delta$ and flow velocity $U_\infty$ is $Re \approx 1.05 \times 10^6$.

The computational domain is shown in Fig. 2, a. The characteristic diameter of the domain is about $3.5\delta$ and the height of the domain is $\approx 2.5\delta$. The sizes of the calculation domain were chosen with regard for the scale $\delta$ so, that the artificial boundaries do not influence the calculated data [18].

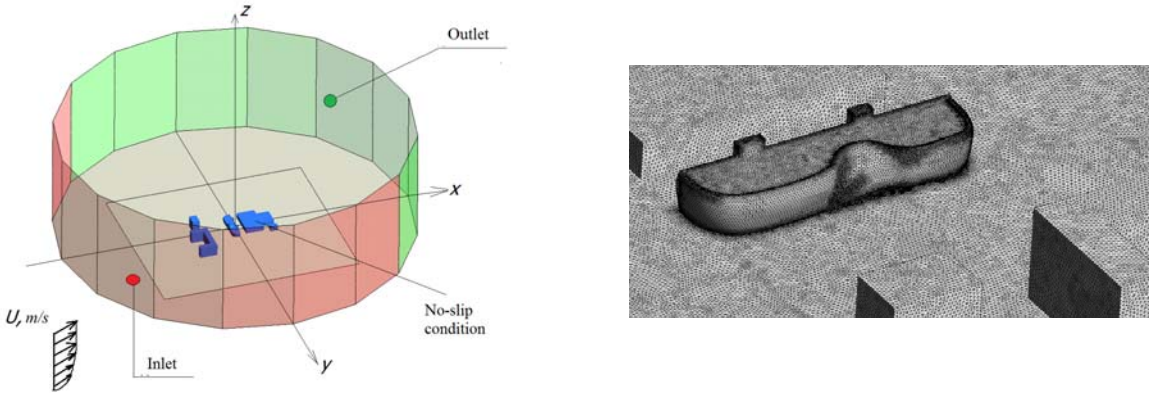
In the computation, the north wind direction ($\alpha$ = 0º, Fig. 1, c) was modeled with the inflow velocity and TKE profiles taken from the experimental data [19]. The "no-slip" boundary condition was used for the substrate and the walls of the buildings. At the exit, the condition of constant static pressure was used, $\Delta P = P_{\text{wall}} - P_0 = 0$. At the upper boundary, we used the symmetry boundary condition, which provided the absence of flow across this boundary.




a) b) c)

Fig. 1: Building of complex shape (a), the geometrical model of the building (b, facade's view) and location of the building *A* relative to the surrounding area (c).

The computational finite-volume grid was constructed with the use of the ANSYS Meshing preprocessor. A fragment of unstructured computational grid on building surfaces is shown on Fig. 2, b. The computational finite-volume grid includes tetra elements in the outer region and the prismatic layers near the solid walls. A series of calculations using various meshes was performed for *conf. 1* to obtain the mesh-independent solution. The parameters of the grids are presented in Table 1. The most detailed numerical grid includes approx. 59 millions of cells. The non-dimensional distance to the wall at the first calculation node was y+= 1 for all cases.




a) b)

Fig.2: The computational domain (a) and a fragment of the mesh on the building's walls (b).

| No | Total number of cells, millions | Characteristic element size near building *A*, m | y+ |
|---|---|---|---|
| 1 | ≈ 9 | ≈ 0.013 | ≤ 1 |
| 2 | ≈ 16,6 | ≈ 0.012 | |
| 3 | ≈ 26 | ≈ 0.01 | |
| 4 | ≈ 46 | ≈ 0.008 | |
| 5 | ≈ 59 | ≈ 0.0066 | |

Table 1: Mesh characteristics.

### 3.2 Results and discussions

Let us consider the flow structure in the vicinity of the building *A* taking into account the surrounding buildings (*conf. 1*). In this case, an steady-state solution with residuals for all equations smaller than $10^{-3}$ was obtained. The flow structure is considerably influenced by the interference effects due to presence of the neighboring buildings. Under the conditions of north wind direction, the building *A* is located behind the building *B*. As we can see in Fig. 3, a, the flow separates from the edge *b1* of the building *B*. The large recirculation zone *V2* is formed behind it. That changes the angle at which the free stream comes on the front side of building *A*. Moreover, the free stream velocity is increased up to 20.5 m/s, when passing around the building *B* and zone *V2*. It leads to an increase in the pressure coefficient values on the facade *a1* of the building *A* (Fig. 3, c). We can also see a horseshoe vortex in front of the building *A*. The large separation zone is formed on the roof of the building *A*, and the negative values of pressure are observed in this zone. It is also important to note that the building *C* is located extremely close to the building *A* and it is the cause the velocity increase up to 19 m/s in the region between them.

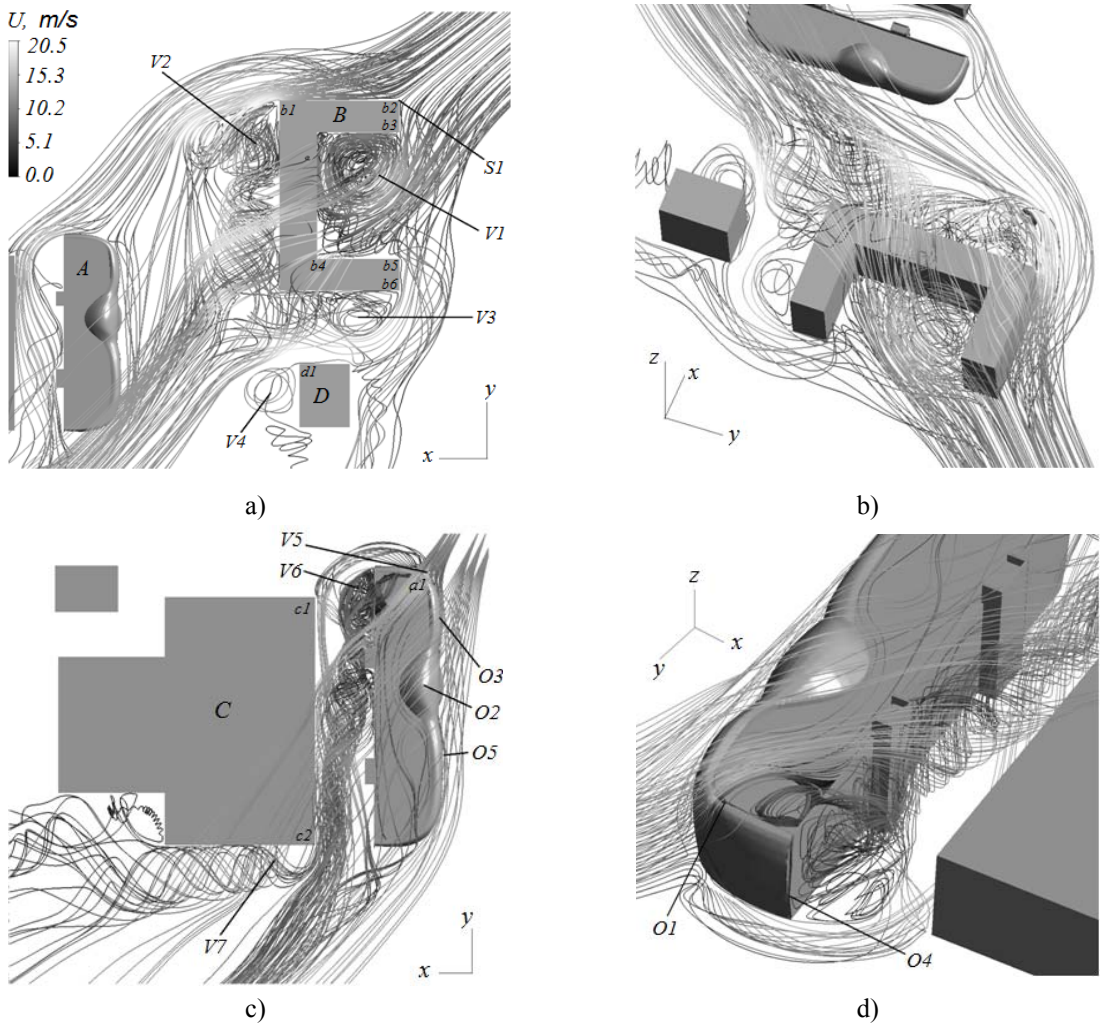
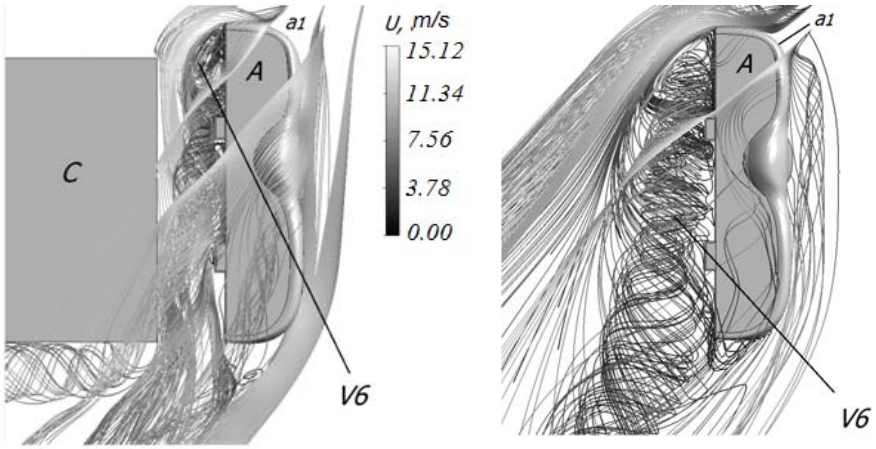



Fig. 3: Air flow structure in the vicinity of building *B* (a,b) and in the vicinity of building *A* (c,d).

The flow structures in the vicinity of building *A* are shown on the Fig. 4 for both the configurations 1 and 2. In the second case, the inflow comes to the facade of the building *A* with a slower speed and at a smaller angle (Fig. 4, b) in comparison with the first case. The maximal

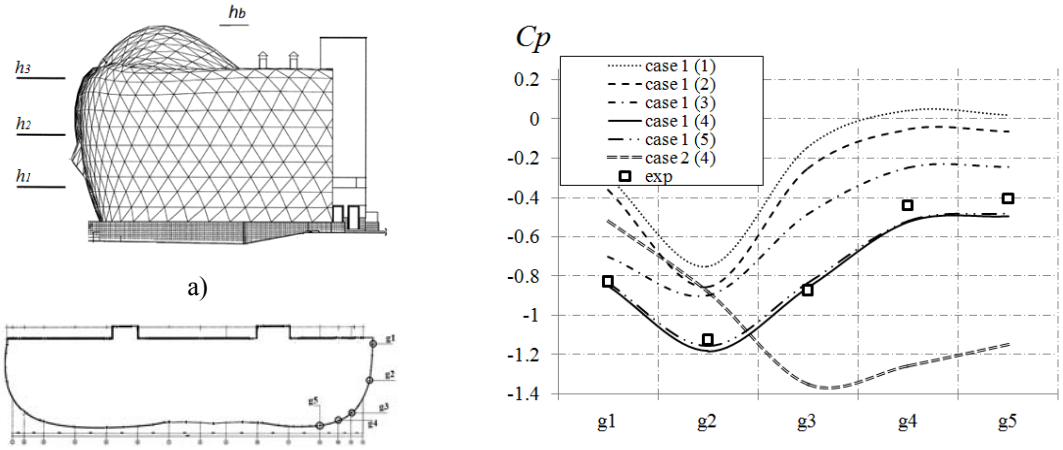
values of positive pressure coefficient are less at this zone for *conf. 2* and reaches values of $C_{p,max,2} \approx 0.83$ ($\approx 0.6 C_{p,max,1}$), where $C_{p,max,1}$ is the maximal values of positive pressure coefficient for *conf. 1*. The maximal absolute values of negative pressure coefficient for configuration 2 are about $C_{p,min,2} \approx -1.68$ ($\approx 0.73 C_{p,min,1}$).

For configuration 2 the absence of obstacles behind the building *A* also leads to the formation of an extensive recirculation zone *V6* on the leeward side. For configuration 1 zone *V6* is smaller "pressed" to the leeward side of the building *A* (Fig 4,a).




a) b)

Fig. 4: Flow structure in the vicinity of building *A* obtained from numerical simulation for *conf. 1* (a) and *conf. 2* (b).




a) b) c)

Fig. 5: The pressure gauges location in experiment [19] (a, b) and pressure coefficient values (c) for gauges g1 - g5 in section $h_3$ obtained from the experiment (exp), from the numerical simulation of *conf. 1* for various meshes (case 1 (1) - (5)) and from the numerical simulation of *conf. 2* for mesh (4) (case 2).

The obtained numerical results for *conf. 1* and *2* were compared with the experimental data [19] on the pressure coefficient $C_p$. Fig. 5 shows the values of the $C_p$, obtained in characteristic points (Fig. 5, b) at the section $h_3 = 0.760 h_b$ (Fig. 5, a) in the experiment and in the numerical simulation for *conf. 2* on the grid No. 4 (see Table 1) and for *conf. 1* on the various grids. As we can see in Fig. 5, negative pressure coefficient values are observed for all the considered points in the section $h_3$, which connected with the fact, that the cross section $h_3$ is located in the separation zone. The good agreement with experimental data [19] was obtained for *conf. 1* on the grid No. 4 and No. 5. For isolated building (*conf. 2*), the numerical solution

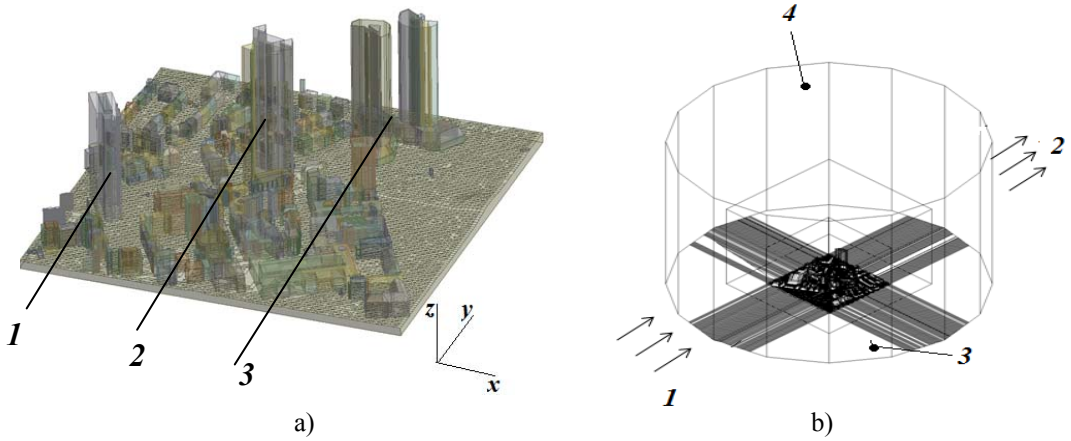
significantly differs from the experimental data. For *conf. 2* the separation zone on the facade is shifted due to the change in free stream flow angle. It testifies of the necessity to consider the interference effects and flow interaction with the surrounding buildings when calculating wind loads.

## 4 AIR FLOW IN THE VICINITY OF THE COMPLEX OF HIGH-RISE BUILDINGS

### 4.1 Problem statement, boundary conditions and computational grid

At the second stage, a turbulent isothermal air flow is considered in vicinity of a high-rise building complex located in Frankfurt/Main city centre (Fig. 6, a). The numerical simulation was performed using full-scale geometrical model[1]. The maximal height of the buildings is $h_{b,max}$≈186 м. The velocity in the flow core is $U_\infty$= 25.4 m/s and the boundary-layer thickness is $\delta \approx$ 500 m [24]. The freestream Reynolds number calculated by length scale $L = \delta$ and flow velocity $U_\infty$ is $Re \approx 3.2 \times 10^8$. The computational domain shown in Fig. 6, b, has a diameter of $d \approx 10h_{bmax}$ and a height of $h \approx 6h_{bmax}$.

The inflow velocity and TKE were calculated using Monin-Obukhov theory considering the nature measurements [24] in some points on Commerz Bank walls for south-west wind direction ($\alpha$ = 225º). The total number of finite-volume cells is ≈$33.6\times10^6$ for the case. The characteristic size of cells near the walls is about ≈3 m.




a) b)

Fig. 6: The geometrical model of buildings complex (a) and the computational domain (b): 1 – inflow conditions, 2 –outlet, 3 – "no-slip" wall condition, 4 – symmetry.

### 4.2 Results and discussions

The air flow near the buildings has a complex 3D vortex structure. In Fig. 7, the instantaneous velocity fields at the moment t=210.7 sec are presented in several horizontal cross-section. Under the conditions of south-west wind direction, the vortexes separate from the side edges of the first-located high-rise building (building 1, Fig. 6, a) and a von-Karman vortex structure is formed. It leads to velocity increasing up to ≈1.3÷1.5 $U_{h,\infty}$, where $U_{h,\infty}$ is a freestream velocity at the height $h$. The building 2 is located almost completely in separation zone formed behind the building 1. It causes the high negative values of pressure coefficients at the front faces of the building 2.

[1] Geobasisdaten:© Stadtvermessungsamt Frankfurt am Main, Stand 20.01.2015

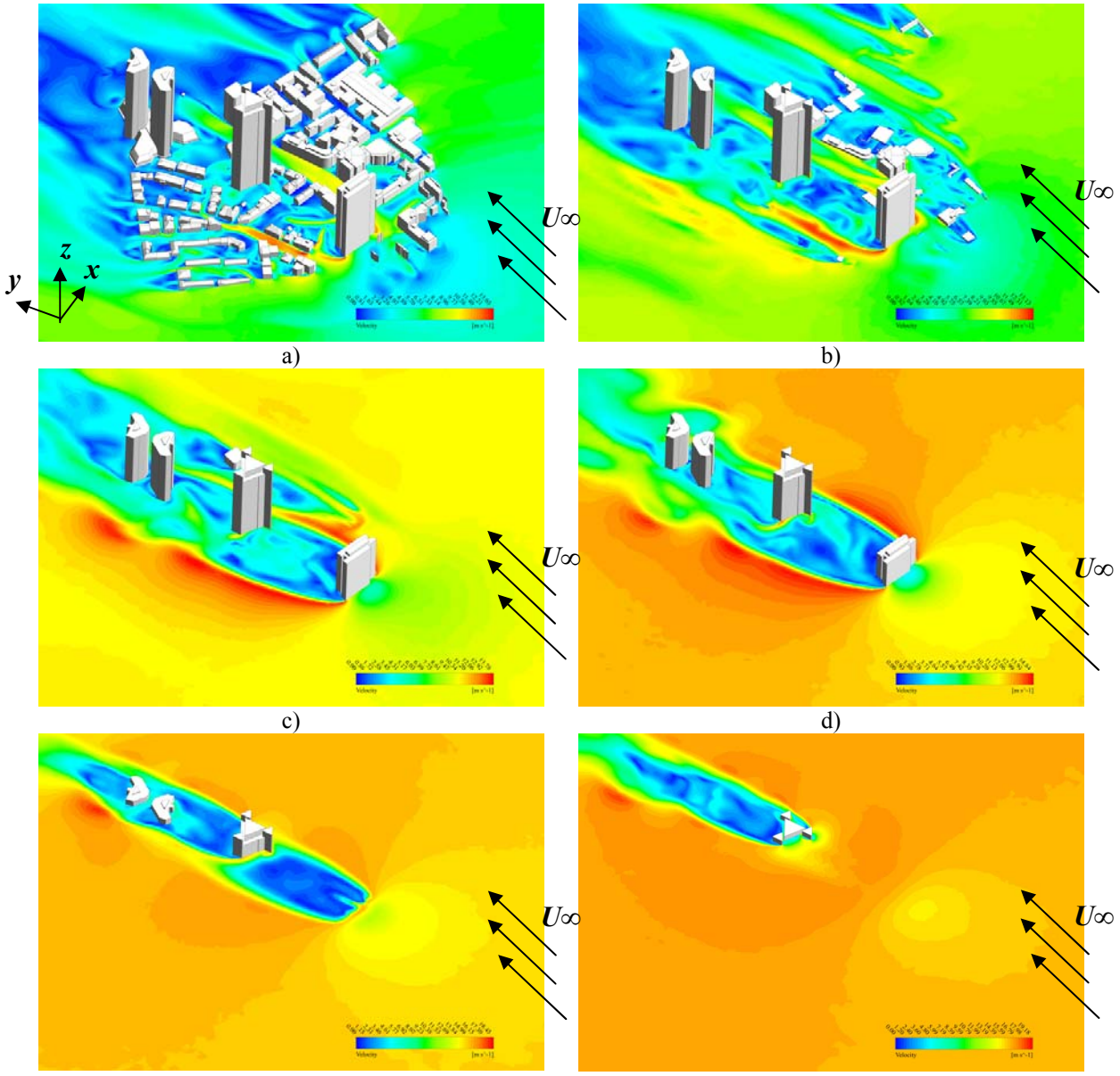
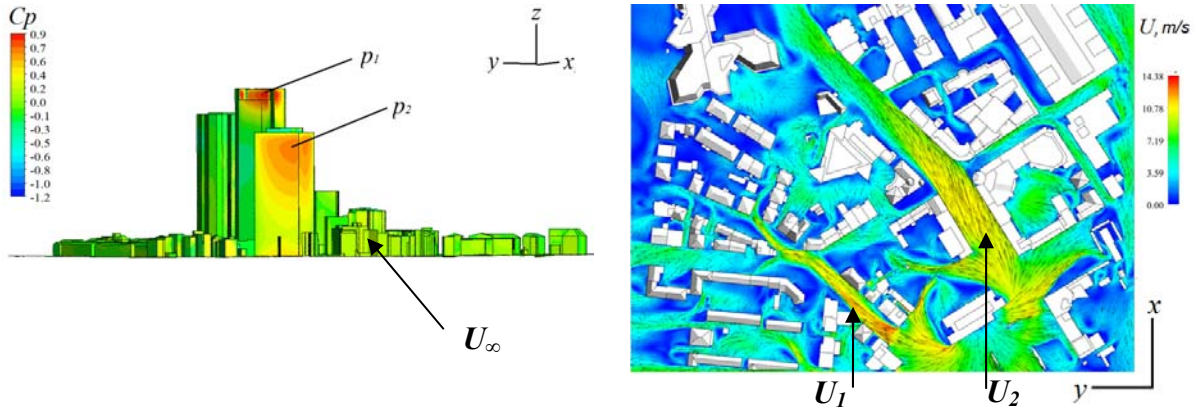


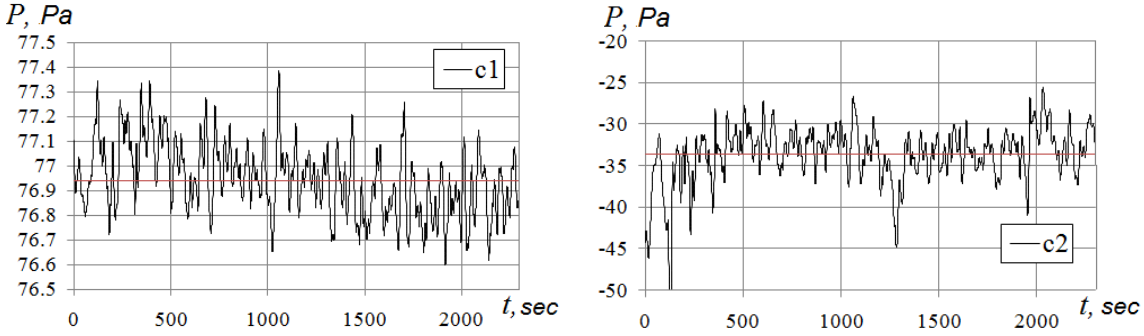

Fig. 7: The instantaneous velocity fields at the moment $t$=210.7 sec. for different characteristic horizontal sections: $z$= 8 m (a), 26 m (b), 56 m (c), 86 m (d), 146 m (e), 176 m (f).

In Fig. 8, a we can see an instantaneous distribution of the pressure coefficient $C_p$ on the walls of the high-rise buildings. The analysis of the $C_p$ distribution has shown that high values of wind pressure are observed in the *p2* zone where the flow comes to the front face of the building 1, and also in the *p1* zone where the flow comes to the upper part of the windward face of the building 2. To estimate the mean and the fluctuating components of the wind load on the building 1, we calculated the pressure time history in several characteristic points on its surface. The Fig. 9, a shows *P(t)* at the point *c1* (-107.8 m;-189.5 m; 60 m) located at the frontal face of building 1. The average pressure at this point is $P_{c1}\approx$76.9 Pa. Pressure fluctuations at this point are not significant ($P'_{c1}\approx$0.5 Pa). We observe a different situation at the frontal surface of building 2 (Fig. 9, b), namely, the fluctuating component of the wind load at the point *c2* (10.3 m; 46.5 m; 60 m) reaches more than 30 % from the average value ($P_{c2}\approx$-33.6 Pa; $P'_{c2}\approx$11.4 Pa).




a) b)

Fig. 8: The instantenous pressure coefficient distribution on the buildings walls at the moment $t$=210.7 sec. (a) and the field of velocity for horizontal section $z$= 2 m (b).




a) b)

Fig. 9: The static pressure - time dependence at the gauge point c1 (-107.8 m;-189.5 m; 60 m) located at the frontal surface of building 1 and the gauge point c2 (10.3 m; 46.5 m; 60 m) located at the frontal surface of building 2.

Low-rise buildings located in the neighborhood of high-rise buildings 1, 2 and 3 do not significantly influence the formation of large vortex flow structures. However, low-rise buildings form roughness of the ground surface and affect the air flow in pedestrian zones. The instantaneous velocity field at the horizontal section $z$=2 m is shown in Fig. 9. The maximal velocity amplitude at this section occurs at zones $U_1$ and $U_2$, that is caused by "restriction" of flow in these areas. The maximal flow velocity at the zone $U_1$ reaches values of about $4.8U_{\infty,h1}$, where $U_{\infty,h1} \approx 2.8$ m/s is a velocity of free stream at the height $z$=2м. The maximal flow velocity at the zone $U_2$ reaches values of about $3.5U_{\infty,h1}$. From the viewpoint of pedestrian comfort such values may be considered as unfavorable. A visualization of the flow structure near the pedestrian zones may allow us to optimize city environment taking into account street topology. It also allows us to design additional protective measures like park zones etc. to control wind flows.

## 5 CONCLUSIONS

- 3D computational analysis of a turbulent separated air flow in the vicinity of the building of complex shape has been performed with a high resolution of the viscous sublayer taking into account the surrounding buildings.
- The flow interference effects were described and their influence on the wind load distribution on the buildings wall was evaluated. It has been demonstrated that local separated zones on the building faces may occur and pressure coefficient may be sign-alternating with high amplitudes that must be taken into account in the design of facades.

- The comparison of the numerical results and the experimental data [19] on the pressure coefficient distribution on the buildings walls was carried out for the two configurations. Satisfactory qualitative and quantitative agreement with the experimental data was obtained for the non-isolated building location (*conf. 1*). We have observed that flow interference effects may lead to a significant increase in the wind load (up to 50%) on a building. The results obtained for the *conf. 2*, when the building was isolated, significantly underpredict the values of the wind pressure coefficient.
- We performed the 3D computational analysis of the flow around the complex of high-rising buildings in the Frankfurt/Main city center, Germany. The flow structure and pressure distribution on buildings walls were obtained. The main features of a wind flow in the vicinity of buildings were investigated and pedestrian comfortable/uncomfortable zones were detected.

## 6 ACKNOWLEDGMENTS

The work was supported by the Ministry of Education of the Russian Federation (Project No. 211, task No. 2014/140 for executing scientific activities within the basic part of government order); German Academic Exchange Service (DAAD) 2014/2015.

The authors would like to thank:

- Prof. Valery M. Mitasov, Novosibirsk State University of Architecture and Civil Engineering (Sibstrin), for providing the wind tunnel experimental data [19] and valuable advices on wind loads modelling;
- CADFEM International GmbH and especially Dr.-Eng. Günter Müller and Dr.-Eng. Stefan Trometer for providing geometry model of Frankfurt/Main city area and the valuable advices on using of CAE ANSYS for cityscape modelling;
- Our colleagues from IAG, Stuttgart University, Prof. Dr.-Eng. Ewald Krämer, Dr.-Eng. Uwe Gaisbauer and Dr.-Eng. Thorsten Lutz for their support of Svetlana Valger during her internship at IAG.

## REFERENCES

[1] L.T. Matveev, Fundamentals of General Meteorology: Physics of Atmosphere, Gidrometeoizdat, Leningrad, 1985 (in Russian).

[2] A. Gill, Atmosphere-Ocean Dynamics, Vol. 1, Mir, Moscow, 1986 (in Russian).

[3] A. Gill, Atmosphere-Ocean Dynamics, Vol. 2, Mir, Moscow, 1986 (in Russian).

[4] E. Simiu and R. Scanlan, Wind Effects on Structures, edited by B.E. Maslov, Stroiizdat, Moscow, 1984 (in Russian).

[5] A.N. Birbraer and A.Yu. Roleder, Extreme Impacts on Buildings, St.-Petersburg Polytechnic University, St. Peters-burg, 2009 (in Russian).

[6] G. Fershing, Fundamentals of Aeroelasticity, edited by G.M. Fomin, Mashinostroenie, Moscow, 1984 (in Russian).

[7] Computational Techniques for Fluid Dynamics, Springer-Verlag, Berlin, 1988.

[8] P.N. Solyanik, Experimental Aerodynamics, Kharkov, National Aero-Space University "Kharkov Aviation Institute", 2007.

[9] N.A. Zaks, Basics of Experimental Aerodynamics, Oborongiz, Moscow, 1953 (in Russian).

[10] S.M. Gorlin, Experimental Aerodynamics, Vysshaya Shkola, Moscow, 1970 (in Russian).

[11] S.V. Nikolaev, V.M. Ostretsov, et al. Aerodynamic tests of high-rise building models, in: Unique and Special Technologies in Construction Activities, Dom na Brestskoi, Moscow, 2005, P. 82-84 (in Russian).

[12] Construction norms and regulations, CN&R 2.01.07-85*, Loads and Impacts, FGUP CPP, Moscow, 2010 (in Russian).

[13] ENV 1991: Eurocode 1: Basis design and actions on structures. Parts 2-4: Wind action, CEN, 1994.

[14] Eurocode 1. Impacts on Building Structures, Parts 1–4, General impacts, Wind impacts, Standard-Practice Engineering Code, Minstroi Arkhitektura, Minsk, 2010.

[15] ANSYS Fluent [Electronic resource] // ANSYS Software [Official web-site],

http:// www.ansys.com

[16] SOFISTIK CFD [Electronic resource] // SOFISTIK [Official web-site], http://www.sofistik.com/

[17] S.A. Valger, N.N. Fedorova, A.V. Fedorov, Structure of turbulent separated flow in the neighborhood of a plate-mounted prism of square section. Thermophysics and Aeromechanics, 2015, T. 22. Vol. 1, P. 29-41 (in Russian).

[18] A. Mochida, Y. Tominaga, S. Murakami, R. Yoshie, T. Ishihara, and R. Ooka, Comparison of various k-ε models and DSM applied to flow around a high-rise building report on AIJ cooperative project for CFD prediction of wind environment, Wind and Structures, 2002, Vol. 5, No. 2-4, P. 227-244.

[19] The science-technical report on experimental measurements of pressure on the buildings walls of model "Biotechnopark", Kol'tcovo city, 2013 (in Russian).

[20] J. Blazek, Computational Fluid Dynamics: Principles and Applications, Elsevier, Amsterdam, 2005.

[21] Menter F.R. Two-Equation Eddy-Viscosity Turbulence Models for Engineering Application // AIAA Jour. 32, No. 8, 1994. P. 1598-1605

[22] Kato, M., Launder B.E. The modelling of turbulent flow around stationary and vibrating square cylinders // Ninth Symposium on "Turbulent Shear Flows". Kyoto, Japan. August 16-18, 1993. –P.10–14.

[23] A.J. Chorin, Numerical solution of Navier—Stokes equations, Math. Comp., 1968, P. 745–762.

[24] Hessisches Ministerium des Innern. Ergänzungserlaß zu DIN 1055, Teil 4, Ausgabe August 1986, Betr. Windlasten bei hohen Hochhäusern im Raum Frankfurt am Main, 1987. – 491 p.

[25] Berneiser A., König G. Wind Loads in City Centres Demonstrated at the New Commerzbank Building in Frankfurt/Main // LACER, 1997. – No. 2. – P. 231–244.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A STABLE AND CONSERVATIVE TIME-DEPENDENT INTERFACE FORMULATION ON SUMMATION-BY-PARTS FORM: AN INITIAL INVESTIGATION

**Samira Nikkar and Jan Nordström**

Department of Mathematics, Computational Mathematics, Linköping University
SE-581 83 Linköping, Sweden
e-mail: {samira.nikkar, jan.nordstrom}@liu.se

**Keywords:** Time-dependent interface, Summation-by-parts, Deforming domains, Stability, Conservation, Multi-blocks.

**Abstract.** *A time-dependent coupling for curved multi-domain problems is considered. First, we transform the problem from Cartesian coordinates into curvilinear coordinates and apply the energy method to derive well-posed and conservative interface conditions.*

*Next, we discretize the problem in space and time by employing finite difference operators that satisfy the summation-by-parts convention. The boundary and interface conditions are imposed weakly by using the simoultaneous approximation term technique as penalty formulations. The discrete version of the energy method is used to derive the stability and conservation requirements. We show how to formulate the penalty operators in such a way that the interface procedure is automatically adjusted to the movements and deformations of the interface, while stability requirements are fulfilled and conservation conditions are respected.*

*Finally, we illustrate the developed techniques by considering an application with the Euler's equations posed on time-dependent curved multi-domains in two space dimensions. The numerical calculations corroborate stability and accuracy of the fully discrete approximation.*

## 1 INTRODUCTION

Multi-block schemes that use Summation-by-Parts (SBP) operators and the Simultaneous Approximation Terms (SAT) technique [2], have previously been investigated in terms of stability, accuracy and conservation [3, 5, 6, 10, 11]. The focus of the SBP-SAT multi-block methodology has been, so far, mostly on time-independent spatial domains with a notable exception being [13].

In this paper, we extend the multi-block technique for moving domains in [13] and construct an interface formulation that couples two curved domains over a moving and/or deforming interface. The new time-dependent interface formulation is provably stable, accurate and conservative.

The rest of this paper proceeds as follows. In section 2, we transform the continuous problem from Cartesian to curvilinear coordinates and derive a conservative and well-posed interface condition. Section 3 deals with the discrete problem where we guarantee the stability and conservation of the interface procedure. In section 4, numerical experiments are performed and corroborate the results regarding the accuracy, stability and conservation of the scheme. Finally, we summarize and draw conclusions in section 5.

## 2 THE CONTINUOUS PROBLEM

Consider the following system of equations on a time dependent deforming domain

$$\begin{aligned} U_t + (\hat{A}U)_x + (\hat{B}U)_y = 0,\ (x,\ y) \in \Omega_L(t),\ t \in [0,\ T], \\ V_t + (\hat{A}V)_x + (\hat{B}V)_y = 0,\ (x,\ y) \in \Omega_R(t),\ t \in [0,\ T], \end{aligned} \tag{1}$$

where $U, V$ are the solutions on the left and right sub-domains, the subscripts $t, x$ and $y$ indicate partial derivatives in their respective direction and $\Omega_{L,R}(t)$ are the left and right spatial domains. Moreover, $\hat{A}$ and $\hat{B}$ are constant and symmetric matrices of size $l$. In (1), $\Omega_{L,R}(t)$ meet at a time-dependent interface denoted by $i(t)$, as shown schematically in Figure 1.

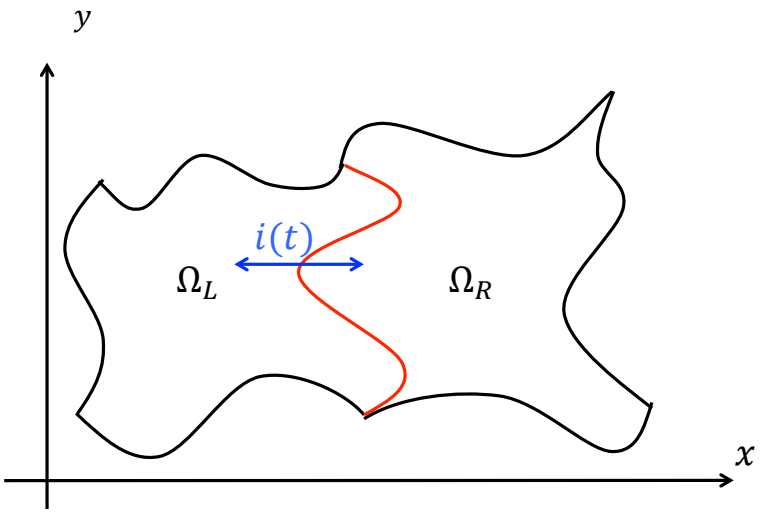



Figure 1: A schematic of the domains $\Omega_{L,R}(t)$ and the time-dependent interface $i(t)$

We transform (1) from the Cartesian coordinates, $x$ and $y$, into curvilinear coordinates $\xi$ and $\eta$ by employing a Lagrangian-Eulerian transformation [22]. The dependence between the

coordinates is described by

$$\begin{aligned} x = x(\xi,\ \eta,\ \tau),\ y = y(\xi,\ \eta,\ \tau),\ t = \tau, \\ \xi = \xi(x,\ y,\ t),\ \eta = \eta(x,\ y,\ t),\ \tau = t. \end{aligned} \tag{2}$$

A schematic of the transformed sub-domains, $\Phi_{L,R}$, and the fixed interface between them, $i$, is shown in Figure 2. For more details about the transformation, see [13].

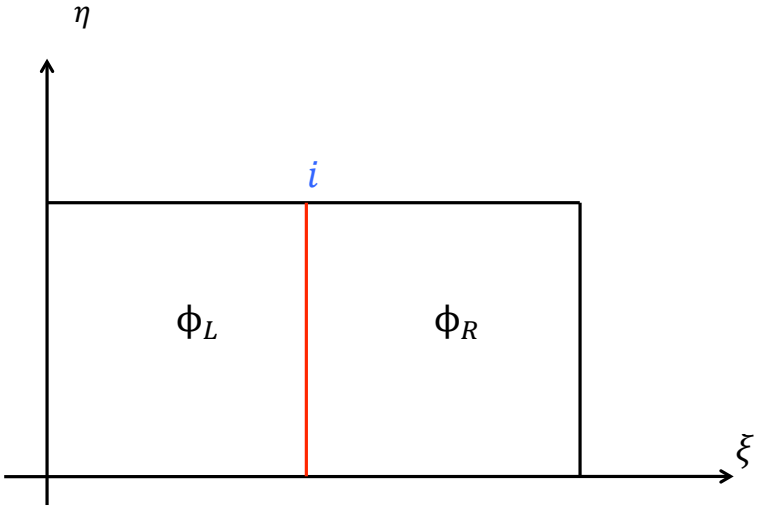



Figure 2: A schematic of the transformed domains $\Phi_{L,R}$ and the time-independent interface $i$.

The governing equations in (1) are then expressed in terms of $\xi$ and $\eta$ by using the chain rule, as

$$\begin{aligned} (J_L U_\tau) + (A_L U)_\xi + (B_L U)_\eta = 0,\ (\xi,\ \eta) \in \Phi_L,\ \tau \in [0,\ T], \\ (J_R V_\tau) + (A_R V)_\xi + (B_R V)_\eta = 0,\ (\xi,\ \eta) \in \Phi_R,\ \tau \in [0,\ T], \end{aligned} \tag{3}$$

where $J_{L,R} > 0$ are the determinants of the Jacobian matrix for the left and right transformations, respectively. Moreover,

$$\begin{aligned} A_{L,R} &= (J\xi_t)_{L,R} I + (J\xi_x)_{L,R} \hat{A} + (J\xi_y)_{L,R} \hat{B}, \\ B_{L,R} &= (J\eta_t)_{L,R} I + (J\eta_x)_{L,R} \hat{A} + (J\eta_y)_{L,R} \hat{B}, \end{aligned} \tag{4}$$

where $I$ is the identity matrix of size $l$. To get the conservative form in (3), we have used the Geometrical Conservation Law (GCL) [12, 13] summarized as $(J_{L_R})_\tau + (A_{L,R})_\xi + (B_{L,R})_\eta = 0$.

### 2.1 Conservation

To derive conservation conditions, we apply the energy method (multiplying the first and second equations in (3) with $\phi_L^T$ and $\phi_R^T$ where $(\phi_{L,R})_j \in H^0$ for $j \in \{1, \dots l\}$ are arbitrary test functions that vanish at the boundaries (not at the interface), and integrate in space and time). The result is

$$\begin{aligned} \int_0^T \iint_{\Phi_L} [(\phi_L^T J_L U)_\tau + (\phi_L^T A_L U)_\xi + (\phi_L^T B_L U)_\eta] d\Phi d\tau = \int_0^T \iint_{\Phi_L} [(\phi_L^T)_\tau J_L U + (\phi_L^T)_\xi A_L U + (\phi_L^T)_\eta B_L U] d\Phi d\tau, \\ \int_0^T \iint_{\Phi_R} [(\phi_R^T J_R V)_\tau + (\phi_R^T A_R V)_\xi + (\phi_R^T B_R V)_\eta] d\Phi d\tau = \int_0^T \iint_{\Phi_R} [(\phi_R^T)_\tau J_R V + (\phi_R^T)_\xi A_R V + (\phi_R^T)_\eta B_R V] d\Phi d\tau. \end{aligned} \tag{5}$$

To derive the total derivatives on the left hand side of both equations in (5), one needs to use the GCL corresponding to the left and right transformations. We add the two relations in (5),

integrate by parts and only consider the terms at the interface. Additionally, at the interface we assume $\phi_L = \phi_R := \phi_i$ to obtain

$$\begin{aligned}\iint_{\Phi_L}(\phi_L^T J_L U)\Big|_{\tau=0}^{\tau=T} d\Phi + \iint_{\Phi_R}(\phi_R^T J_R V)\Big|_{\tau=0}^{\tau=T} d\Phi - \int_0^T \iint_{\Phi_L}\Big[U^T[(J_L\phi_L)_\tau + (A_L\phi_L)_\xi + (B_L\phi_L)_\eta]\Big]^T d\Phi\, d\tau \\ - \int_0^T \iint_{\Phi_R}\Big[V^T[(J_R\phi_R)_\tau + (A_R\phi_R)_\xi + (B_R\phi_R)_\eta]\Big]^T d\Phi\, d\tau \\ = \; - \int_0^T \int_i \underbrace{\phi_i^T(A_{L_i}U_i - A_{R_i}V_i)}_{IT} d\eta\, d\tau\end{aligned} \tag{6}$$

where the subscript $i$ denotes interface and $IT$ indicates interface term. To derive the $IT$, we have used the Green-Gauss theorem. Moreover, we have again applied the GCL in order to arrive at the weak forms on the left hand side of (6). Further, we note that the following terms are included in $IT$

$$\begin{aligned}A_{L_i} &= (J\xi_t)_L I + (J\xi_x)_L \hat{A} + (J\xi_y)_L \hat{B}, \\ A_{R_i} &= (J\xi_t)_R I + (J\xi_x)_R \hat{A} + (J\xi_y)_R \hat{B},\end{aligned} \tag{7}$$

where the following relations between the metric terms and their counterparts corresponding to the inverse transformation hold

$$(J\xi_t)_{L,R} = (x_\eta y_\tau - x_\tau y_\eta)_{L,R}, \;\; (J\xi_x)_{L,R} = (y_\eta)_{L,R}, \;\; (J\xi_y)_{L,R} = (-x_\eta)_{L,R}. \tag{8}$$

The two domains, $\Omega_{L,R}(t)$, are always connected at the interface regardless of the movements and deformations. Therefore, the left and right transformations map the same curve ($i(t)$ in the Cartesian coordinates) to the same line segment ($i$ in the curvilinear coordinates). Hence, although the left and right transformations may differ in general, the left and right metrics terms in $(8)$ will have exactly the same values at the interface. We conclude that

$$A_{L_i} = A_{R_i} := A_i \tag{9}$$

and $U_i = V_i$ removes the interface term. Finally, (6) becomes an integral statement of the original problem and conservation is respected. Condition (9) will need to be respected also in the numerical approximations as we will show below.

### 2.2 Well-posedness

The energy method (multiplying (3) with the transpose of the solution and integrating over the spatial and temporal domains) together with the Green-Gauss theorem results in

$$\begin{aligned}||U(T,\,\xi,\,\eta)||_{J_L}^2 &= ||U(0,\,\xi,\,\eta))||_{J_L}^2 - \int_0^T \oint_{\delta\Phi_L} U^T C_L\, U\; ds\; d\tau, \\ ||V(T,\,\xi,\,\eta)||_{J_R}^2 &= ||V(0,\,\xi,\,\eta))||_{J_R}^2 - \int_0^T \oint_{\delta\Phi_R} V^T C_R\, V\; ds\; d\tau.\end{aligned} \tag{10}$$

In (10), the norm is defined as $||W||_{J_{L,R}}^2 = \iint_{\Phi_{L,R}} W^T J_{L,R} W\; d\xi\; d\eta$, for $W \in \{U, V\}$ and $\delta\Phi_{L,R}$ are the boundaries of $\Phi_{L,R}$. Moreover, $C_{L,R} = (A_{L,R},\; B_{L,R}) \cdot n_{L,R}$ where $n_{L,R} = (\alpha_{L,R},\; \beta_{L,R})$

is the outward pointing normal vector from $\Phi_{L,R}$. Further, $C_{L,R}$ can be expanded as $C_{L,R} = \alpha_{L,R}A_{L,R} + \beta_{L,R}B_{L,R}$ and due to symmetry be decomposed into $C_{L,R} = X_{L,R}\Lambda_{L,R}X_{L,R}^T$ where $\Lambda_{L,R}$ is the matrix of eigenvalues of $C_{L,R}$ and $X_{L,R}$ is the corresponding eigenvector matrix.

To control the growth of the energy of the solution due to boundary terms, we choose the following far-field well-posed boundary conditions

$$\begin{array}{l}(X_L^T U)_j = (X_L^T U_\infty)_j \ \text{ if } \ (\Lambda_L)_{jj} < 0, \\ (X_R^T V)_j = (X_R^T V_\infty)_j \ \text{ if } \ (\Lambda_R)_{jj} < 0,\end{array} \tag{11}$$

where $U_\infty, V_\infty$ are the data to the problem and $j \in \{1, \ldots, l\}$. Additionally, at the interface where $n_L = (1,\ 0)^T$ and $n_R = (-1,\ 0)^T$, we have $C_L = A_i$ and $C_R = -A_i$ and the decomposition $A_i = X_i\Lambda_i X_i^T$.

By substituting (11) in (10), observing that $U_i = V_i$, and considering the initial conditions $U(0,\ \xi,\ \eta) = f_L$ and $V(0,\ \xi,\ \eta) = f_R$, we find

$$\begin{aligned}||U(T,\xi,\eta)||_{J_L}^2 + ||V(T,\xi,\eta)||_{J_R}^2 \ &= \ ||f_L||_{J_L}^2 + ||f_R||_{J_R}^2 \\ &- \int_0^T\!\!\int_{\delta\Phi_L\setminus i} \left[(X_L^T U)^T\Lambda_L^+(X_L^T U) - (X_L^T U_\infty)^T\Lambda_L^-(X_L^T U_\infty)\right] ds\, d\tau \\ &- \int_0^T\!\!\int_{\delta\Phi_R\setminus i} \left[(X_R^T V)^T\Lambda_R^+(X_R^T V) - (X_R^T V_\infty)^T\Lambda_R^-(X_R^T V_\infty)\right] ds\, d\tau,\end{aligned} \tag{12}$$

where the positive and negative superscripts in $\Lambda_L^{+,-}$, $\Lambda_R^{+,-}$ and $\Lambda_i^{+,-}$ restrict $\Lambda_{L,R,i}$ to the non-negative and negative eigenvalues, respectively. Finally, we conclude that the energy of the solution is bounded by data and that the problem is strongly well-posed [21].

## 3 THE DISCRETE PROBLEM

We discretize the left and right sub-domains using $N_{L,R}$ and $M$ grid points in the $\xi$ and $\eta$ directions, respectively. In this article, we consider matching grid points along the interface. For non-conforming grids at the interface, interpolation techniques must be used [24, 25] in order to couple the blocks. However, with the added difficulty of a moving interface we refrain from this technical complication and will consider that in a future paper. We use $K$ time levels from 0 to $T$ and tensor products to arrange the fully discrete numerical solution. As an example, the numerical solution on the left sub-domain is arranged as

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ [\mathbf{U}_k] \\ \vdots \\ \mathbf{U}_K \end{bmatrix};\ [\mathbf{U}_k] = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ [\mathbf{U}_i] \\ \vdots \\ \mathbf{U}_{N_L} \end{bmatrix}_k;\ [\mathbf{U}_i]_k = \begin{bmatrix} \mathbf{U}_1 \\ \vdots \\ \mathbf{U}_j \\ \vdots \\ \mathbf{U}_M \end{bmatrix}_{ki}, \tag{13}$$

in which $\mathbf{U}_{kij} \approx U(\tau_k,\ \xi_i,\ \eta_j)$ for $(\xi,\ \eta) \in \Phi_L$. The fully discrete numerical solution corresponding to the right sub-domain is denoted by $\mathbf{V}$ and arranged in the same way.

The first derivative $u_\xi$ is approximated by $D_\xi\mathbf{u}$, where $D_\xi$ is a so-called SBP operator of the form

$$D_\xi = P_\xi^{-1}Q_\xi, \tag{14}$$

and $\mathbf{u} = [u_0,\ u_1,\ \cdots,\ u_N]^T$ is the solution vector evaluated in each grid point in the $\xi$ direction. $P_\xi$ is a symmetric positive definite matrix, and $Q_\xi$ is an almost skew-symmetric matrix that satisfies

$$Q_\xi + Q_\xi^T = E_1 - E_0 = B = diag(-1, 0, ..., 0, 1). \tag{15}$$

In (15), $E_0 = diag(1, 0, ..., 0)$ and $E_1 = diag(0, ..., 0, 1)$. The $\eta$ and $\tau$ directions are discretized in the same way.

A first derivative SBP operator is a 2s-order accurate central difference operator which is modified close to the boundaries such that it becomes one-sided. Together with a diagonal norm P, the boundary closure is s-order accurate, making a stable first order approximation of a hyperbolic problem s+1 order accurate globally [4, 9]. For more details on non-standard SBP operators see [15, 18, 16, 17].

A finite difference approximation including the time discretization [7, 8], on SBP-SAT form, is constructed by extending the one-dimensional SBP operators in a tensor product fashion as

$$\begin{aligned} D_\tau &= P_\tau^{-1} Q_\tau \otimes I_\xi \otimes I_\eta \otimes I, \\ D_\xi &= I_\tau \otimes P_\xi^{-1} Q_\xi \otimes I_\eta \otimes I, \\ D_\eta &= I_\tau \otimes I_\xi \otimes P_\eta^{-1} Q_\eta \otimes I, \end{aligned} \tag{16}$$

where $\otimes$ represents the Kronecker product [14]. Here and in the remainder of this article, all matrices in the first position are of size $K \times K$, the second position $N_{L,R} \times N_{L,R}$, the third position $M \times M$ and the fourth position $l \times l$. Additionally, the identity matrix, $I$, has a consistent size with its position in the Kronecker product.

Prior to discretizing (3), we use the splitting technique explained in [19]. The discrete version of (3) including only the interface term (the far-field boundaries are addressed in [13] and are not repeated here) is

$$\begin{aligned} &\frac{1}{2}[D_\tau \mathbf{J}_L \mathbf{U} + \mathbf{J}_L D_\tau \mathbf{U} + (\mathbf{J}_L)_\tau \mathbf{U} + D_\xi \mathbf{A}_L \mathbf{U} + \mathbf{A}_L D_\xi \mathbf{U} + (\mathbf{A}_L)_\xi \mathbf{U} + D_\eta \mathbf{B}_L \mathbf{U} + \mathbf{B}_L D_\eta \mathbf{U} + (\mathbf{B}_L)_\eta \mathbf{U}] = \\ &\qquad (I_\tau \otimes P_\xi^{-1} E_1 \otimes I_\eta \otimes I) \Sigma_L (\mathbf{U} - \mathbf{V}) \\ &\frac{1}{2}[D_\tau \mathbf{J}_R \mathbf{V} + \mathbf{J}_R D_\tau \mathbf{V} + (\mathbf{J}_R)_\tau \mathbf{V} + D_\xi \mathbf{A}_R \mathbf{V} + \mathbf{A}_R D_\xi \mathbf{V} + (\mathbf{A}_R)_\xi \mathbf{V} + D_\eta \mathbf{B}_R \mathbf{V} + \mathbf{B}_R D_\eta \mathbf{V} + (\mathbf{B}_R)_\eta \mathbf{V}] = \\ &\qquad (I_\tau \otimes P_\xi^{-1} E_0 \otimes I_\eta \otimes I) \Sigma_R (\mathbf{V} - \mathbf{U}), \end{aligned} \tag{17}$$

where $\Sigma_{L,R}$ are matrices of size $(KNMl) \times (KNMl)$ and operate as penalty parameters corresponding to the weak interface treatments for the left and right sub-domains, respectively, and will be chosen later based on conservation and stability requirements. Moreover, $\mathbf{J}_{L,R}, \mathbf{A}_{L,R}, \mathbf{B}_{L,R}, (\mathbf{A}_{L,R})_\xi$ and $(\mathbf{B}_{L,R})_\eta$ are block diagonal matrices of size $(KN_{L,R}Ml) \times (KN_{L,R}Ml)$, where each $l \times l$ block approximates respectively $J_{L,R}, A_{L,R}, B_{L,R}, (A_{L,R})_\xi$ and $(B_{L,R})_\eta$ on a grid point. Note that the SBP operators are not necessarily the same on the left and right sub-domains. To ease the notation the subscripts $L, R$ on the derivatives $D_{\xi,\eta,\tau}$ are dropped.

### 3.1 Conservation

To show that the scheme is conservative, we multiply (17) by arbitrary vector functions $\phi_L^T P$ and $\phi_R^T P$, where $P = (P_\tau \otimes P_\xi \otimes P_\eta \otimes I)$. Moreover, we use the SBP property stated in (15)

and focus on the interface to obtain

$$\begin{aligned}
&\phi_L^T[(E_1-E_0)\otimes P_\xi\otimes P_\eta\otimes I]\mathbf{J}_L\mathbf{U}+\phi_L^T(P_\tau\otimes E_1\otimes P_\eta\otimes I)\mathbf{A}_L\mathbf{U}\\
&-\frac{1}{2}\Big[\mathbf{U}^TP[D_\tau\mathbf{J}_L\phi_L+\mathbf{J}_LD_\tau\phi_L+(\mathbf{J}_L)_\tau\phi_L]\Big]^T-\frac{1}{2}\Big[\mathbf{U}^TP[D_\xi\mathbf{A}_L\phi_L+\mathbf{A}_LD_\xi\phi_L+(\mathbf{A}_L)_\xi\phi_L]\Big]^T\\
&-\frac{1}{2}\Big[\mathbf{U}^TP[D_\eta\mathbf{B}_L\phi_L+\mathbf{B}_LD_\eta\phi_L+(\mathbf{B}_L)_\eta\phi_L]\Big]^T=\phi_L^T(P_\tau\otimes E_1\otimes P_\eta\otimes I)\Sigma_L(\mathbf{U}-\mathbf{V})\\
&\phi_R^T[(E_1-E_0)\otimes P_\xi\otimes P_\eta\otimes I]\mathbf{J}_R\mathbf{V}-\phi_R^T(P_\tau\otimes E_0\otimes P_\eta\otimes I)\mathbf{A}_R\mathbf{V}\\
&-\frac{1}{2}\Big[\mathbf{V}^TP[D_\tau\mathbf{J}_R\phi_R+\mathbf{J}_RD_\tau\phi_R+(\mathbf{J}_R)_\tau\phi_R]\Big]^T-\frac{1}{2}\Big[\mathbf{V}^TP[D_\xi\mathbf{A}_R\phi_R+\mathbf{A}_RD_\xi\phi_R+(\mathbf{A}_R)_\xi\phi_R]\Big]^T\\
&-\frac{1}{2}\Big[\mathbf{V}^TP[D_\eta\mathbf{B}_R\phi_R+\mathbf{B}_RD_\eta\phi_R+(\mathbf{B}_R)_\eta\phi_R]\Big]^T=\phi_R^T(P_\tau\otimes E_0\otimes P_\eta\otimes I)\Sigma_R(\mathbf{V}-\mathbf{U}).
\end{aligned}\tag{18}$$

In order to arrive at (18), we have used the symmetry property of the matrices as well as the Numerical Geometric Conservation Law (NGCL) summarized as $(\mathbf{J}_\tau)_{L,R}+(\mathbf{A}_\xi)_{L,R}+(\mathbf{B}_\eta)_{L,R}=0$. For more details on NGCL see [13]. Now, we consider the same $P_{\eta,\tau}$ in the left and right subdomains and add the two relations in (18). The result is

$$\begin{aligned}
&\phi_L^T[(E_1-E_0)\otimes P_\xi\otimes P_\eta\otimes I]\mathbf{J}_L\mathbf{U}+\phi_R^T[(E_1-E_0)\otimes P_\xi\otimes P_\eta\otimes I]\mathbf{J}_R\mathbf{V}\\
&-\frac{1}{2}\Big[\mathbf{U}^TP[(D_\tau\mathbf{J}_L\phi_L+\mathbf{J}_LD_\tau\phi_L+(\mathbf{J}_L)_\tau\phi_L)\Big]^T-\frac{1}{2}\Big[\mathbf{U}^TP[D_\xi\mathbf{A}_L\phi_L+\mathbf{A}_LD_\xi\phi_L+(\mathbf{A}_L)_\xi\phi_L]\Big]^T\\
&-\frac{1}{2}\Big[\mathbf{U}^TP[D_\eta\mathbf{B}_L\phi_L+\mathbf{B}_LD_\eta\phi_L+(\mathbf{B}_L)_\eta\phi_L]\Big]^T-\frac{1}{2}\Big[\mathbf{V}^TP[D_\tau\mathbf{J}_R\phi_R+\mathbf{J}_RD_\tau\phi_R+(\mathbf{J}_R)_\tau\phi_R]\Big]^T\\
&-\frac{1}{2}\Big[\mathbf{V}^TP[D_\xi\mathbf{A}_R\phi_R+\mathbf{A}_RD_\xi\phi_R+(\mathbf{A}_R)_\xi\phi_R]\Big]^T-\frac{1}{2}\Big[\mathbf{V}^TP[D_\eta\mathbf{B}_R\phi_R+\mathbf{B}_RD_\eta\phi_R+(\mathbf{B}_R)_\eta\phi_R]\Big]^T\\
&=-\phi_L^T(P_\tau\otimes E_1\otimes P_\eta\otimes I)\mathbf{A}_L\mathbf{U}+\phi_L^T(P_\tau\otimes E_1\otimes P_\eta\otimes I)\Sigma_L(\mathbf{U}-\mathbf{V})\\
&+\phi_R^T(P_\tau\otimes E_0\otimes P_\eta\otimes I)\mathbf{A}_R\mathbf{V}+\phi_R^T(P_\tau\otimes E_0\otimes P_\eta\otimes I)\Sigma_R(\mathbf{V}-\mathbf{U}).
\end{aligned}\tag{19}$$

As in the continuous problem, $\phi_L=\phi_R:=\phi_i$ at the interface, by which the right hand side of (19) becomes

$$\phi_i^T(P_\tau\otimes P_\eta\otimes I)[-\mathbf{A}_{L_i}\mathbf{U}_i+\mathbf{A}_{R_i}\mathbf{V}_i+(\Sigma_{L_i}-\Sigma_{R_i})(\mathbf{U}_i-\mathbf{V}_i)],\tag{20}$$

where the subscript $i$ restricts the vectors and matrices to the interface, i.e. $\mathbf{U}_i$ is now of size $KMl$. Since $\mathbf{A}_L$ and $\mathbf{A}_R$ include pointwise approximations to $A_L$ and $A_R$ in (7), respectively, we have

$$\begin{aligned}
\mathbf{A}_L&=(\overline{J\xi_t})_L(I_\tau\otimes I_\xi\otimes I_\eta\otimes I)+(\overline{J\xi_x})_L(I_\tau\otimes I_\xi\otimes I_\eta\otimes\hat{A})+(\overline{J\xi_y})_L(I_\tau\otimes I_\xi\otimes I_\eta\otimes\hat{B}),\\
\mathbf{A}_R&=(\overline{J\xi_t})_R(I_\tau\otimes I_\xi\otimes I_\eta\otimes I)+(\overline{J\xi_x})_R(I_\tau\otimes I_\xi\otimes I_\eta\otimes\hat{A})+(\overline{J\xi_y})_R(I_\tau\otimes I_\xi\otimes I_\eta\otimes\hat{B}),
\end{aligned}\tag{21}$$

where the bar sign is used to indicate that the metric terms are approximated numerically. The numerical metric terms are evaluated as

$$\overline{J\xi_t}=diag[D_\eta(D_\eta M^{(1)}-D_\xi M^{(2)})],\ \ \overline{J\xi_x}=diag[D_\eta\mathbf{y}],\ \ \overline{J\xi_y}=-diag[D_\eta\mathbf{x}],\tag{22}$$

where **x, y** are the x and y coordinates of the mesh in the Cartesian coordinate system, arranged in a vector, consistent to (13). Moreover, $M^{(1)}=diag(\mathbf{y})D_\xi\mathbf{x}$ and $M^{(2)}=diag(\mathbf{y})(D_\eta\mathbf{x})$. To see more details about the numerical metrics, see [13, 23]. From (21) and (22) one can conclude that if we have matching grid points along the interface, and use the same SBP operators for

the left and right problems, we have $\mathbf{A}_{L_i} = \mathbf{A}_{R_i} := \mathbf{A_i}$, which correspond to the continuous requirement in (9).

We use the decomposition $\mathbf{A}_i = \mathbf{X}_i\mathbf{\Lambda}_i\mathbf{X}_i^T$ and choose $\Sigma_{L_i,R_i} = \mathbf{X}_i\tilde{\Sigma}_{L_i,R_i}\mathbf{X}_i^T$ where $\tilde{\Sigma}_{L_i,R_i}$ are diagonal and rewrite (20) as

$$\phi_i^T(P_\tau \otimes P_\eta \otimes I)\left[\mathbf{X}_i(-\mathbf{\Lambda}_i + \tilde{\Sigma}_{L_i} - \tilde{\Sigma}_{R_i})\mathbf{X}_i^T\right](\mathbf{U}_i - \mathbf{V}_i). \tag{23}$$

In order to obtain a conservative scheme, $\tilde{\Sigma}_{L_i,R_i}$ must be chosen such that

$$-\mathbf{\Lambda}_i + \tilde{\Sigma}_{L_i} - \tilde{\Sigma}_{R_i} = 0 \tag{24}$$

holds.

### 3.2 Stability

To prove stability we apply the discrete energy method (multiplying the left and right sub-problems with $\mathbf{U}^T P$ and $\mathbf{V}^T P$, respectively) on (17). We add the transpose of the result to itself and use the SBP property (15) to arrive at

$$\begin{aligned}
&\mathbf{U}^T[(E_1 - E_0) \otimes P_\xi \otimes P_\eta \otimes I]\mathbf{J}_L\mathbf{U} + \mathbf{U}^T[P_\tau \otimes (E_1 - E_0) \otimes P_\eta \otimes I]\mathbf{A}_L\mathbf{U} + \\
&\mathbf{U}^T[P_\tau \otimes P_\xi \otimes (E_1 - E_0) \otimes I]\mathbf{B}_L\mathbf{U} = \mathbf{U}^T(P_\tau \otimes E_1 \otimes P_\eta \otimes I)\Sigma_L(\mathbf{U} - \mathbf{V}) + \\
&(\mathbf{U} - \mathbf{V})^T\Sigma_L^T(P_\tau \otimes E_1 \otimes P_\eta \otimes I)\mathbf{U}, \\
&\mathbf{V}^T[(E_1 - E_0) \otimes P_\xi \otimes P_\eta \otimes I]\mathbf{J}_R\mathbf{V} + \mathbf{V}^T[P_\tau \otimes (E_1 - E_0) \otimes P_\eta \otimes I]\mathbf{A}_R\mathbf{V} + \\
&\mathbf{V}^T[P_\tau \otimes P_\xi \otimes (E_1 - E_0) \otimes I]B_R\mathbf{V} = \mathbf{V}^T(P_\tau \otimes E_0 \otimes P_\eta \otimes I)\Sigma_R(\mathbf{V} - \mathbf{U}) + \\
&(\mathbf{V} - \mathbf{U})^T\Sigma_R^T(P_\tau \otimes E_0 \otimes P_\eta \otimes I)\mathbf{V}.
\end{aligned} \tag{25}$$

Again, we have used the NGCL in order to obtain (25). We simplify (25) and only keep the terms at the interface which gives

$$\begin{aligned}
&||\mathbf{U}_K||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{L_K}} - ||\mathbf{U}_1||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{L_1}} = \mathbf{U}^T[P_\tau \otimes -E_1 \otimes P_\eta \otimes I]\mathbf{A}_L\mathbf{U} + \\
&\mathbf{U}^T(P_\tau \otimes E_1 \otimes P_\eta \otimes I)\Sigma_L(\mathbf{U} - \mathbf{V}) + (\mathbf{U} - \mathbf{V})^T\Sigma_L^T(P_\tau \otimes E_1 \otimes P_\eta \otimes I)\mathbf{U} \\
&||\mathbf{V}_K||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{R_K}} - ||\mathbf{V}_1||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{R_1}} = \mathbf{V}^T[P_\tau \otimes E_0 \otimes P_\eta \otimes I]\mathbf{A}_R\mathbf{V} + \\
&\mathbf{V}^T(P_\tau \otimes E_0 \otimes P_\eta \otimes I)\Sigma_R(\mathbf{V} - \mathbf{U}) + (\mathbf{V} - \mathbf{U})^T\Sigma_R^T(P_\tau \otimes E_0 \otimes P_\eta \otimes I)\mathbf{V}.
\end{aligned} \tag{26}$$

The norms in (26) are given by $||\mathbf{W}||^2_{H_{L,R}} = \mathbf{W}^T H_{L,R}\mathbf{W}$, where $H_{L,R} = (P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{L,R} > 0$ for $\mathbf{W} \in \{\mathbf{U},\ \mathbf{V}\}$. Moreover, the subscripts $K$ and 1 indicate restrictions to the last and first time levels, respectively. We add the two relations in (26) and again use the decompositions $\mathbf{A}_i = \mathbf{X}_i\mathbf{\Lambda}_i\mathbf{X}_i^T$ and $\Sigma_{L_i,R_i} = \mathbf{X}_i\tilde{\Sigma}_{L_i,R_i}\mathbf{X}_i^T$ to obtain

$$\begin{aligned}
&||\mathbf{U}_K||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{L_K}} + ||\mathbf{V}_K||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{R_K}} = ||\mathbf{U}_1||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{L_1}} + ||\mathbf{V}_1||^2_{(P_\xi \otimes P_\eta \otimes I)\mathbf{J}_{R_1}} + \\
&\begin{bmatrix} \mathbf{X}_i^T\mathbf{U}_i \\ \mathbf{X}_i^T\mathbf{V}_i \end{bmatrix}^T \left[P_\tau \otimes P_\eta \otimes I \otimes \tilde{I}_1\right] \begin{bmatrix} -\mathbf{\Lambda}_i + 2\tilde{\Sigma}_{L_i} & -(\tilde{\Sigma}_{L_i} + \tilde{\Sigma}_{R_i}) \\ -(\tilde{\Sigma}_{L_i} + \tilde{\Sigma}_{R_i}) & \mathbf{\Lambda}_i + 2\tilde{\Sigma}_R \end{bmatrix} \begin{bmatrix} \mathbf{X}_i^T\mathbf{U}_i \\ \mathbf{X}_i^T\mathbf{V}_i \end{bmatrix},
\end{aligned} \tag{27}$$

where $\tilde{I}$ is

$$\tilde{I}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{28}$$

By considering the conservation requirement on the penalty operators in (24), (27) becomes

$$||\mathbf{U}_K||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{L_K}} + ||\mathbf{V}_K||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{R_K}} = ||\mathbf{U}_1||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{L_1}} + ||\mathbf{V}_1||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{R_1}} + \\ \left[\ \mathbf{X}_i^T(\mathbf{U}_i-\mathbf{V}_i)\ \right]^T (P_\tau\otimes P_\eta\otimes I)\left[-\mathbf{\Lambda}_i + 2\tilde{\Sigma}_{L_i}\right]\ \mathbf{X}_i^T(\mathbf{U}_i-\mathbf{V}_i). \tag{29}$$

From (29), we conclude that the following conditions lead to a stable interface treatment

$$\begin{array}{lcll} (\tilde{\Sigma}_{L_i})_{jj} & \leq & (\mathbf{\Lambda}_i)_{jj}/2 & \text{if} \quad (\mathbf{\Lambda}_i)_{jj} < 0, \\ (\tilde{\Sigma}_{L_i})_{jj} & = & 0 & \text{if} \quad (\mathbf{\Lambda}_i)_{jj} \geq 0, \end{array} \tag{30}$$

for $j \in \{1,\ldots,l\}$. As an example if we let $\tilde{\Sigma}_{L_i} = \mathbf{\Lambda}_i^-$, where the negative superscription restricts $\mathbf{\Lambda}_i$ to the negative eigenvalues. Then (29) becomes

$$||\mathbf{U}_K||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{L_K}} + ||\mathbf{V}_K||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{R_K}} = ||\mathbf{U}_1||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{L_1}} + ||\mathbf{V}_1||^2_{(P_\xi\otimes P_\eta\otimes I)\mathbf{J}_{R_1}} - \\ (\mathbf{U}_i-\mathbf{V}_i)^T\ (P_\tau\otimes P_\eta\otimes I)\underbrace{|\mathbf{X}_i\mathbf{\Lambda}_i\mathbf{X}_i^T|}_{=|\mathbf{A}_i|}\ (\mathbf{U}_i-\mathbf{V}_i) \tag{31}$$

which represents a dissipative, stable and conservative interface procedure.

## 4 NUMERICAL EXPERIMENTS

We consider the two-dimensional constant coefficient symmetrized Euler equations [20] described by

$$W_t + \hat{A}W_x + \hat{B}W_y = 0,\ (x,\ y) \in \Omega_{L,R}(t),\ t \in [0,\ T], \tag{32}$$

where

$$W = \left[\frac{\bar{c}\rho}{\sqrt{\gamma}\bar{\rho}},\ u,\ v,\ \frac{\theta}{\bar{c}\sqrt{\gamma(\gamma-1)}}\right]^T. \tag{33}$$

In (33), $W \in \{U,\ V\}$, and $\rho,\ u,\ v,\ \theta,$ and $\gamma$ are respectively the density, the $x$ and $y$ velocity components, the temperature and the ratio of specific heats. An equation of state of the form $\gamma p = \bar{\rho}\theta + \rho\bar{\theta}$, where $p$ is the pressure, completes the system (32). Moreover, the bar sign denotes the reference state around which we have linearized. The matrices in (32) are

$$\hat{A} = \begin{bmatrix} \bar{u} & \bar{c}/\sqrt{\gamma} & 0 & 0 \\ \bar{c}/\sqrt{\gamma} & \bar{u} & 0 & \sqrt{\frac{\gamma-1}{\gamma}}\bar{c} \\ 0 & 0 & \bar{u} & 0 \\ 0 & \sqrt{\frac{\gamma-1}{\gamma}}\bar{c} & 0 & \bar{u} \end{bmatrix},\ \hat{B} = \begin{bmatrix} \bar{v} & 0 & \bar{c}/\sqrt{\gamma} & 0 \\ 0 & \bar{v} & 0 & 0 \\ \bar{c}/\sqrt{\gamma} & 0 & \bar{v} & \sqrt{\frac{\gamma-1}{\gamma}}\bar{c} \\ 0 & 0 & \sqrt{\frac{\gamma-1}{\gamma}}\bar{c} & \bar{v} \end{bmatrix}. \tag{34}$$

We prescribe $\gamma = 1.4$, $\bar{c} = 2$, $\bar{\rho} = 1$ and consider a mean velocity field of the form $(\bar{u},\ \bar{v}) = (1,\ 1)$. The geometries $\Omega_{L,R}(t)$ are described by

$$\Omega_L(t): \begin{cases} x_{w_L}(y) & = -1 - 0.1\ \sin(2\pi y), \\ x_i(t,\ y) & = -0.1\ [\sin(\pi t) + \cos(2\pi t)\sin(2\pi y)], \\ y_{s_L}(t,\ x) & = -0.05\ \sin[2\pi(x_i(t)-x)/(x_i(t)+1)], \\ y_{n_L}(t,\ x) & = 1 - 0.05\ \sin[2\pi(x_i(t)-x)/(x_i(t)+1)], \end{cases} \tag{35}$$

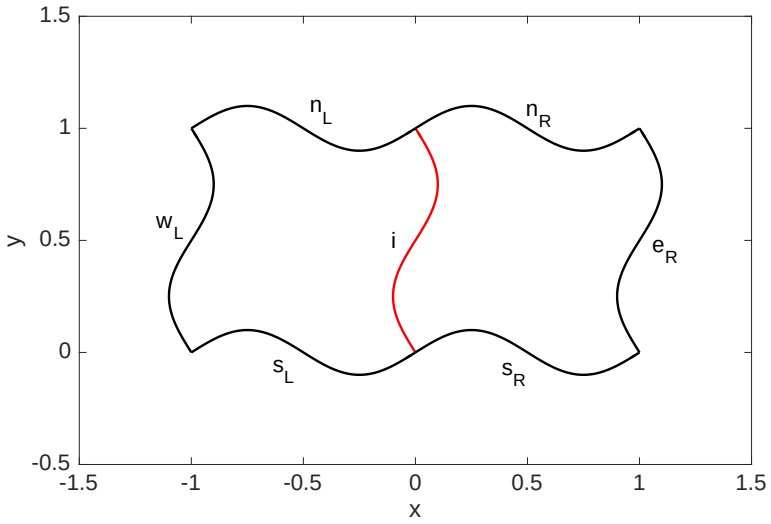



Figure 3: A schematic of $\Omega_{L,R}$ and the boundaries

and

$$\Omega_R(t):\begin{cases} x_{e_R}(y) &= 1-0.1\ \sin(2\pi y), \\ x_i(t,\ y) &= -0.1\ [\sin(\pi t)+\cos(2\pi t)\sin(2\pi y)], \\ y_{s_R}(t,x) &= -0.05\ \sin[2\pi(x-x_i(t))/(1-x_i(t))], \\ x_{n_R}(t,y) &= 1+0.05\ \sin[2\pi(x-x_i(t))/(1-x_i(t))], \end{cases} \tag{36}$$

where $w_L,\ e_R,\ s_{L,R},\ n_{L,R}$ and $i$ are schematically defined in Figure 3.

As seen in (35) and (36), also the four adjacent boundaries $n_{L,R}, s_{L_R}$ to the interface are slightly deforming. To see the details of how to treat deforming boundaries, see [13].

### 4.1 Accuracy

To conclude the accuracy of our numerical approximations, we use the method of manufactured solution with the reference solution $W_\infty$

$$W_\infty = [5\sin(x-t),\ 5\cos(x-t),\ 10\sin(y-t),\ 10\cos(y-t)]^T \tag{37}$$

which is injected as a forcing function to the right hand side of (32). Moreover, characteristic boundary conditions [13] are used.

We examine the scheme for SBP operators of order $2s$ in the interior and $s$ close to the boundaries in space, where $s \in \{1,\ 2,\ 3\}$. The fifth order accurate SBP operator i.e. SBP84, with a sufficiently large $K$, is used in time. The rates of convergence are calculated and shown in tables 1-3.

| $N_{L,R}, M$ | 21 | 31 | 41 |
|---|---|---|---|
| $\rho$ | 2.40229 | 2.08909 | 2.03912 |
| $u$ | 2.13136 | 2.03877 | 2.01409 |
| $v$ | 2.06440 | 2.03268 | 2.01787 |
| $p$ | 2.48489 | 2.12153 | 2.04978 |

Table 1: Convergence rates at T=1, for a sequence of mesh refinements, SBP21 in space, SBP84 in time (K=201)

The convergence results in tables 1-3 are correct according to the theory [4, 2].

| $N_{L,R}, M$ | 21 | 31 | 41 |
|---|---|---|---|
| $\rho$ | 3.08961 | 2.94594 | 2.93760 |
| $u$ | 3.22104 | 3.11838 | 3.06405 |
| $v$ | 2.91411 | 2.99215 | 3.02028 |
| $p$ | 3.16628 | 3.12290 | 3.08363 |

Table 2: Convergence rates at T=1, for a sequence of mesh refinements, SBP42 in space, SBP84 in time (K=201)

| $N_{L,R}, M$ | 21 | 31 | 41 |
|---|---|---|---|
| $\rho$ | 4.68075 | 4.45144 | 4.60620 |
| $u$ | 3.87224 | 4.15980 | 4.48346 |
| $v$ | 3.84525 | 3.94578 | 4.28970 |
| $p$ | 3.73903 | 4.50886 | 4.75789 |

Table 3: Convergence rates at T=1, for a sequence of mesh refinements, SBP63 in space, SBP84 in time (K=201)

### 4.2 An application

We consider an initial pressure of the form $p = e^{-20((x+1)^2+y^2)}$ together with zero velocities and densities everywhere in the left and right sub-domains. Characteristic boundary conditions with data from manufactured solution $p_\infty = e^{-20((x+1-2t)^2+(y-t)^2)}$ is used for far-field boundaries. Moreover, we construct a grid of $41 \times 41$ points in space and $81$ nodes in time. Third and fifth order accurate SBP operators in space and time, respectively, are used.

The pressure distribution at different times (the red and/or dashed curves corresponding to $\Omega_{L,R}(0)$) are shown in Figures 4-11. As shown in the figures, the pressure pulse passes across the interface smoothly and moves out of the domain as time passes.

## 5 SUMMARY AND CONCLUSIONS

We have constructed a numerical scheme that satisfies the summation-by-parts convention in combination with the simultaneous approximation term technique, for general multi-block problems with time-dependent deforming interfaces in several space dimensions.

We have studied well-posedness, stability and as well continuous and discrete conservation by employing the continuous and discrete versions of the energy method as our analytical tools and constructed a dissipative, conservative and stable scheme.

An application using the Euler equations posed on a time-dependent multi-block geometry with a moving and deforming interface, was presented. The correct rates of convergence toward the exact solution, were concluded.

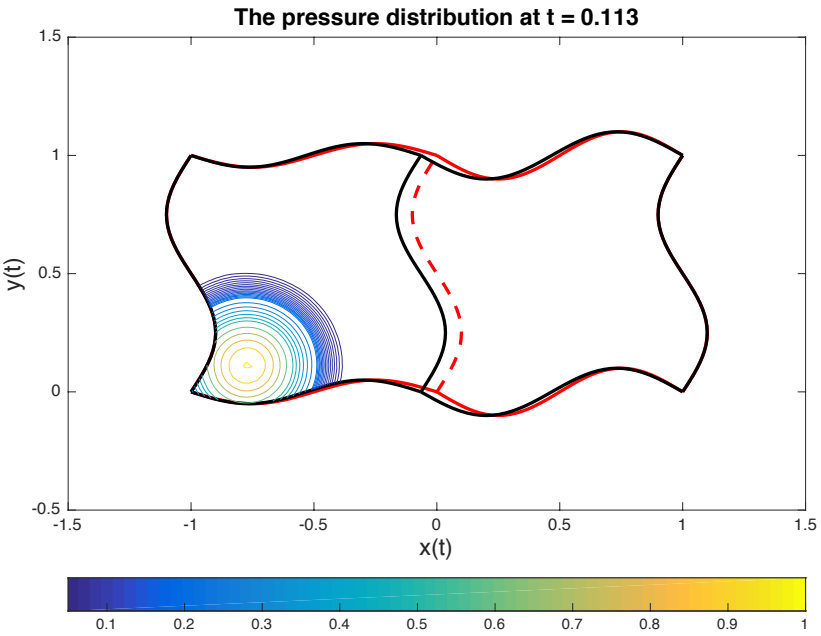



Figure 4: The pressure distribution.

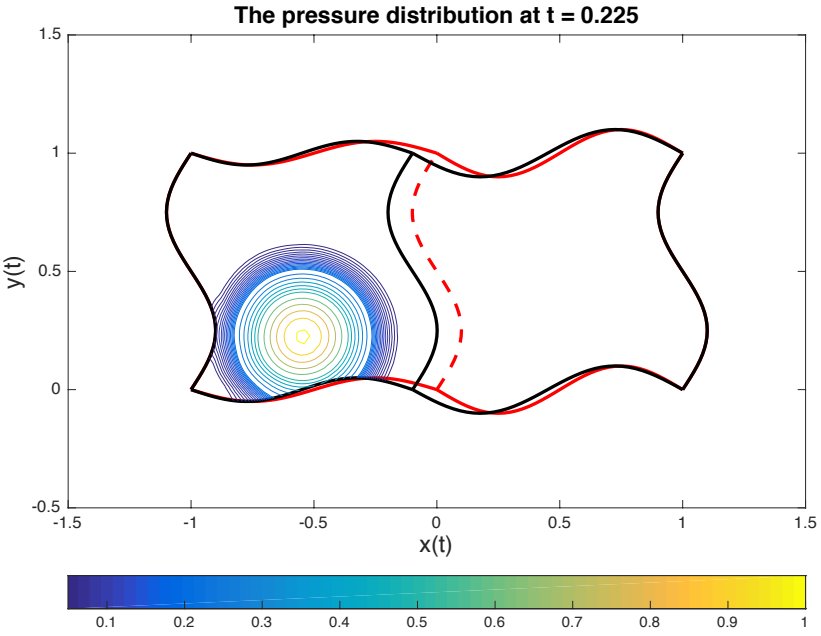



Figure 5: The pressure distribution.

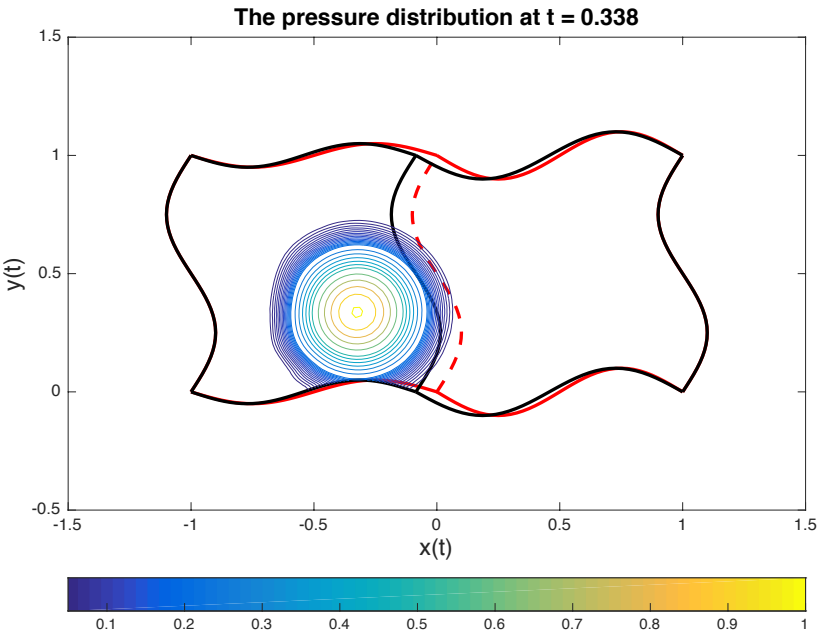



Figure 6: The pressure distribution.

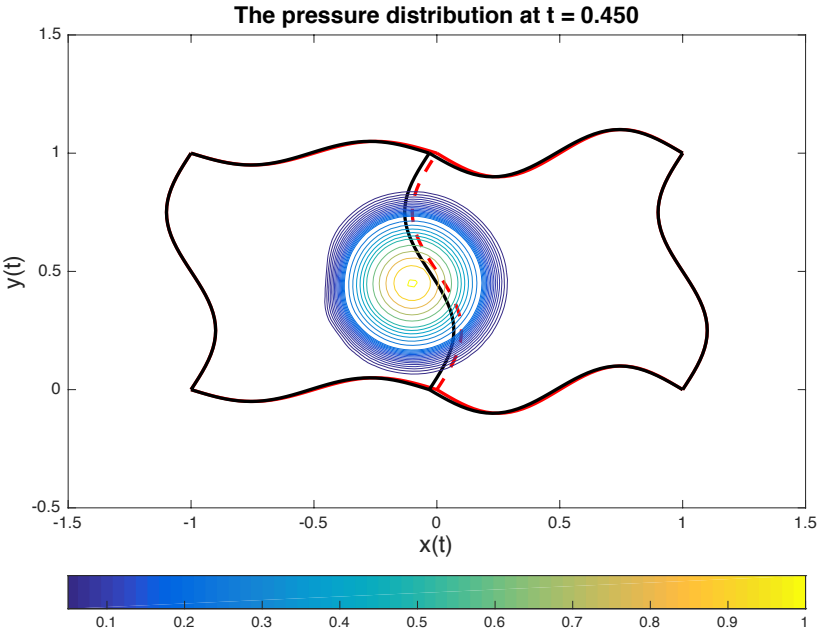



Figure 7: The pressure distribution.

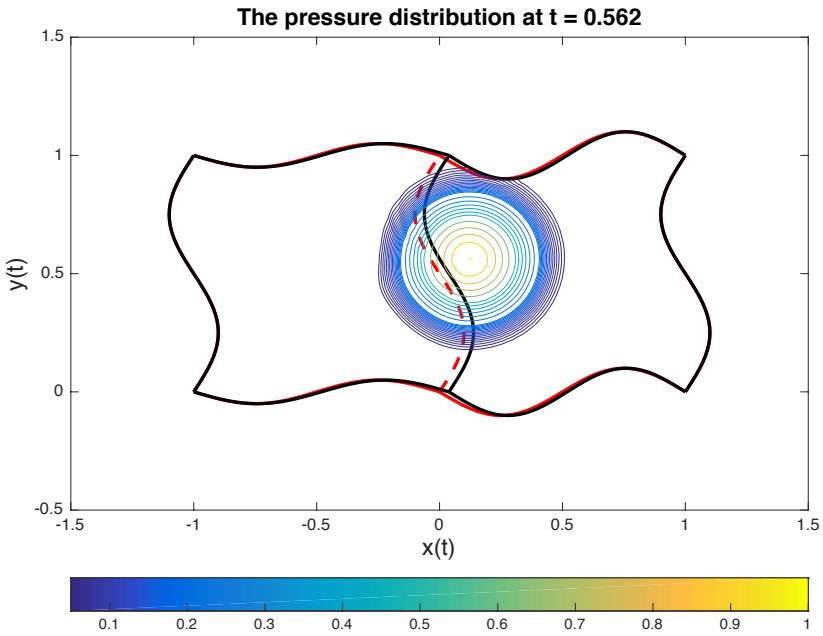



Figure 8: The pressure distribution.

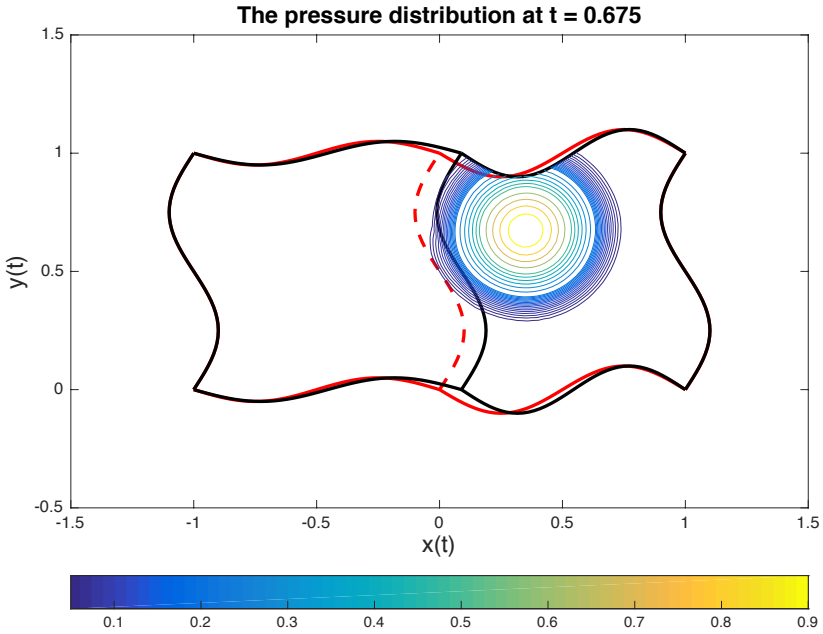



Figure 9: The pressure distribution.

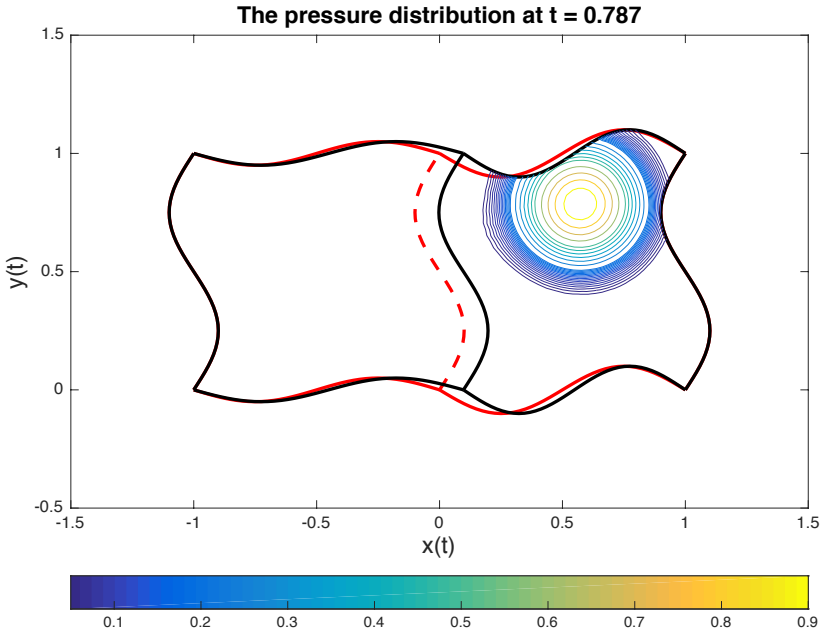



Figure 10: The pressure distribution.

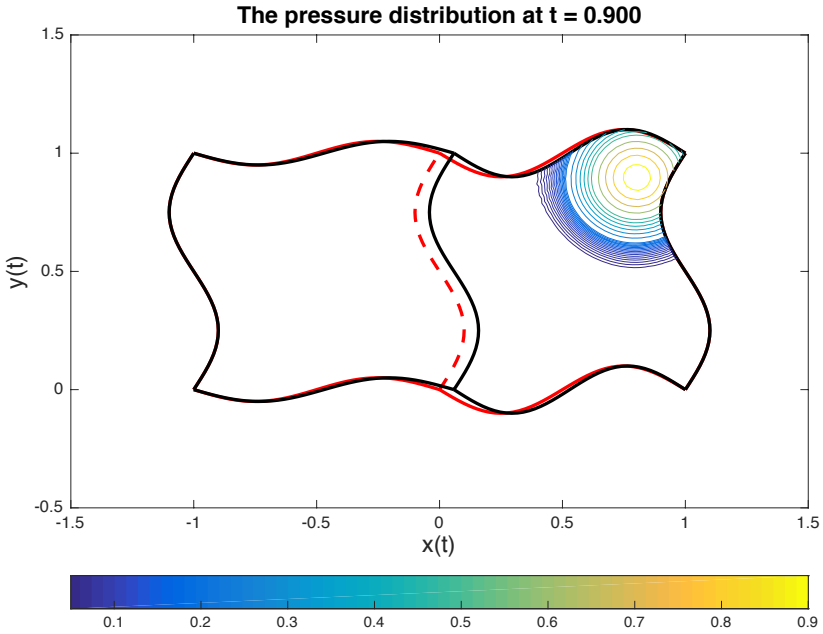



Figure 11: The pressure distribution.

## REFERENCES


[1] S. Nikkar and J. Nordström, Fully discrete energy stable high order

finite difference methods for hyperbolic problems in deforming domains, *Journal of Computational Physics*, **291**:82-98 (2015).

[2] M. Svärd and J. Nordström, Review of summation-by-parts schemes for initial-boundary-value problems, *Journal of Computational Physics*, **268**:17-38, (2014).

[3] J. Nordström and J. Gong and E. van der Weide and M. Svärd, A stable and conservative high order high order multi-block method for the compressible Navier-Stokes equations, *Journal of Computational Physics*, **228**:9020-9035 (2009).

[4] Strand B., Summation by parts for finite difference approximations of d/dx, *Journal of Computational Physics*, **110**:47-67 (1994).

[5] J. Gong and J. Nordström, Interface procedures for finite difference approximations of the advection-diffusion equation, *Journal of Computational and Applied Mathematics*, **236**:602-620 (2011).

[6] K. Mattsson and M. H. Carpenter, Stable and Accurate interpolation operators for high-order multi-block finite-difference methods, *SIAM Journal of Scientific Computing*, **32**(4):22982320 (2010).

[7] Nordström J. and Lundquist. T., Summation-by-parts in time, *Journal of Computational Physics*, **251**:487-499 (2013).

[8] Lundquist, T. and Nordström, J., The SBP-SAT technique for initial value problems, *Journal of Computational Physics*, **270**:86-104 (2014).

[9] Svärd, M. Nordström, J., On the order of accuracy for difference approximations of initial boundary value problems, *Journal of Computational Physics*, **218**:333-352 (2006).

[10] Carpenter M. H., Nordström, J., Gottleib D., A Stable and Conservative Interface Treatment of Arbitrary Spatial Accuracy, *Journal of Computational Physics*, **148**:341-365 (1999).

[11] Carpenter M. H., Nordström, J., Gottleib D., Revising and extending interface penalties for multi-domain summation-by-parts operators, *Journal of Scientific Computing*, **45**:118-150 (2010).

[12] C. Farhat, P. Geuzaine, C. Grandmont, The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids, *Journal of Computational Physics*, **291**:82-98 (2015).

[13] S. Nikkar and J. Nordström, Fully discrete energy stable high order finite difference methods for hyperbolic problems in deforming domains, *Journal of Computational Physics*, **291**:82-98 (2015).

[14] Charles F. Van Loan, The ubiquitous Kronecker product, *Journal of Computational and Applied Mathematics*, **123**:85-100 (2000).

[15] Saul S. Abarbanel and Alina E. Chertock, Strict stability of high-order compact implicit finite-difference schemes: the role of boundary conditions for hyperbolic PDEs, I, *Journal of Computational Physics*, **160**(1):42-66 (2000).

[16] Mark H. Carpenter and David Gottlieb, Spectral Methods on Arbitrary Grids, *Journal of Computational Physics*, **129**:74-86, (1996).

[17] David C. Del Rey Fernández and Pieter D. Boom and David W. Zingg, A generalized framework for nodal first derivative summation-by-parts operators, *Journal of Computational Physics*, **266**:214-239 (2014).

[18] Saul S. Abarbanel and Alina E. Chertock and Amir Yefet, Strict stability of high-order compact implicit finite-difference schemes: the role of boundary conditions for hyperbolic PDEs, II, *Journal of Computational Physics*, **160**:67-87 (2000).

[19] Nordström, J., Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation, *Journal of Scientific Computing*, **29**:375-404 (2006).

[20] Turkel, E., Symmetrization of the fluid dynamics matrices with applications, *Mathematics of Computation*, **27**:729-736 (1973).

[21] B. Gustafsson, H.-O. Kreiss, and J. Oliger, Time dependent problems and difference methods, John Wiley & Sons, New York, 1995.

[22] C.W Hirt and A.A Amsden and J.L Cook, An arbitrary Lagrangian-Eulerian computing method for all flow speeds, *Journal of Computational Physics*, **14**(3):227-253 (1974).

[23] Y. Abe and T. Nonomura and N. Iizuka and K. Fujii, Conservative metric evaluation for high-order finite-difference schemes with the GCL identities on moving and deforming grids, *Journal of Computational Physics*, **232**:14-21 (2013).

[24] K. Mattsson and M. H. Carpenter, table and accurate interpolation operators for high-order multiblock finite difference methods, *SIAM Journal on Scientific Computing*, **32**:2298-2320 (2010).

[25] T. Lundquist and J. Nordström, On the Suboptimal Accuracy of Summation-by-parts Schemes with Non-conforming Block Interfaces, *LiTH-MAT-R–2015/16–SE*.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. J. Martin-Burgos, D. González-Juárez and E. Andrés-Pérez
Crete Island, Greece, 5–10 June 2016

# A LAPLACIAN MESH DEFORMATION TECHNIQUE FOR SIMULATION-DRIVEN DESIGN OPTIMIZATION

**M. J. Martin-Burgos[1], D. González-Juárez[1] and E. Andrés-Pérez[2]**

[1] Fluid Dynamics Branch, Spanish National Institute for Aerospace Technologies (INTA)
Crta. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
e-mail: {martinbj, gonzalezjd}@inta.es

[2] Engineering Department, Ingeniería de Sistemas para la Defensa de España S.A. (ISDEFE-INTA) & Technical University of Madrid (UPM).
Crta. de Ajalvir Km. 4.5. 28850 Torrejón de Ardoz
e-mail: eandres@isdefe.es, esther.andres@upm.es

**Keywords:** Mesh deformation, Aerodynamic shape optimization, Computational Fluid Dynamics, Partial differential equations.

**Abstract.** *A simple method to automatically adapt the computational grid to surface boundary deformations is presented. Mesh adaptation techniques can be used to reduce the number of re-meshing steps in aerodynamic shape optimization problems or during the solution of a coupled fluid-structure interaction problem. The method is inspired on the solution of the Laplace equation and the resulting algorithm resembles a fully explicit Jacobi iterative scheme, avoiding the need to invert a large matrix, in order to solve the resulting system of equations. The proposed method can withstand large deformations and also includes orthogonality correction to maintain the quality of the grid, especially in the boundary layer region. The proposed technique is presented in two-dimensional unstructured hybrid grids, but can be easily extended to three-dimensions.*

## 1 INTRODUCTION

Using Computational Fluid Dynamics (CFD) simulations, the mesh must be updated to conform the boundary surface modifications. The mesh update can be achieved by re-meshing, but the generation of the mesh, for industrial complex configurations, is a time-consuming operation and automatic grid generation is not always feasible. This issue particularly arises in the search of an optimal shape using CFD simulations, aero-elastic analysis, or control surface deflection problems. In addition, the use of different overlapped grids or chimera for the simulation of moving parts is a well established and versatile technique. However, the interpolation between zones is not intrinsically conservative and for some specific problems, such as fluid-structural interaction, the use of mesh deformation would be preferable instead, in order to reduce the numerical noise. To avoid the regeneration of the grid, mesh deformation or mesh update techniques can be considered as a fast and cheap alternative.

In the literature, there are several proposals for mesh deformation. The transfinite interpolation [1,2] is an algebraic method, generally used in structured grids, that interpolates the displacement of mesh points along arc mesh lines. This method is simple but not appropriate for unstructured grids. The advancing front method is another algebraic technique for unstructured grids [3]. In this method, grid points are sorted with increasing distance to the deformation surface. Then, new displacements for each point are scaled average from the displacements of the neighbor points, which have a smaller distance and which are thus already updated. This strategy leads to advancing the front of displacements as iso-surfaces of equal distances, which was found to be of advantage in order to transport the information orthogonal to the surfaces into the interior of the grid. The advancing front algorithm is simple and computationally very efficient. However, the robustness is uncertain for large deformations and it is unclear how to face multiple components, such as wing-flap configurations.

In the tension spring analogy [4], the edges of the computational grid are modeled as springs, whose stiffness are inversely proportional to the length of the edge, making short edges stiffer. This method is mathematically simple and can be solved with an iterative Jacobi solver. However, the spring-edge idealization will eventually fail, in some cases even for small displacements. This method prevents the colliding of the vertices, but there is no mechanism to prevent a vertex to cross and edge and the consequently collapsing element.

In order to mitigate overstretching and negative volume elements, the edge-spring is usually combined with a torsional spring analogy [5]. In two-dimensional grids, each vertex is attached to a torsional spring that prevents the faces from squashing. Unfortunately, the extension to three-dimensional grids is not straightforward, as it may lead to negative volume tetrahedrons, even if all tetrahedron faces are valid elements. There are some attempts to extend it to three-dimensions using more complex formulations and kinematic systems [6]. However, the spring-torsion analogy works under the assumption of relatively small displacements. When elements are nearly flat, the Partial Differential Equations (PDE) that governs the system is ill-conditioned, imposing a limited time step and convergence problems. Despite of its limitations, the spring analogy still remains popular thanks to its easy implementation.

The linear elasticity analogy [7] assumes that the computational mesh obeys the isotropic linear elasticity equations, much like a rubber. The modulus of elasticity is often taken to be inversely proportional to the cell volume or the cell aspect ratio. In this manner high aspect ratio cells, as those close to the boundary layer, are less susceptible to compression. The spring and linear elasticity approaches are compared in [8], where only the linear elasticity method is stable enough for large scale deformations of 3D hybrid Navier-Stokes grids. The linear elasticity approach allows significantly larger geometric deformations and can be naturally extended to three-dimensional grids. However, it requires the solution of a large system of equa-

tions, one for each vertex, and the computational effort is not diminished in comparison with the CFD simulations.

The concept of Radial Basis Functions (RBF) has been also applied to update the mesh [9-11]. In the RBF method, the mesh points are calculated through a global function that depends on the distance of the base points, which are usually the deformed surface. The RBF method does not require connectivity information, making it suitable for unstructured grids. However, without information of the elements shape, there is little control of the grid quality. Some authors propose the incorporation of a second set of base points to limit the change in orthogonality near the wall [12].

Partial Differential Equations (PDE) are often used for generating grids. The use of the Laplace equations is a popular approach to optimize smoothness and control the variation of cell volumes. Techniques based on PDEs can also be applied for updating the grid to a moving boundary [13], which can be naturally extended to three-dimensions [14]. This approach tackles the mesh motion as a boundary value problem by finding the mesh displacement field that emanates from the solution of an Euler-Lagrange equation. This works when the mesh is composed of approximately equal-sized elements and the motion of the interface boundary is in the order of the size of the elements; otherwise, results in an algorithm breakdown. In order to prevent an ill-conditioned system, a constraint condition is applied element wise; so larger elements absorb most of the distortion. Similar to the linear elasticity analogy, methods based on PDE can withstand a large amount of deformation, but require the solution of a very large equation system; often tackled with a gradient conjugate method, or one of its variants.

The tension spring and linear elasticity are closely related to the Laplace equation. In this work, a mesh deformation algorithm based on the Laplace operator is presented. The method is similar to the linear elasticity in terms of robustness and deformation withstanding, while it is as easy to implement as the spring analogy. The method can update very large deformations without losing grid quality and the resulting equation system can be easily solved with an explicit Jacobi algorithm.

## 2 PROPOSED APPROACH

### 2.1 Mesh deformation

The heat equation describes how thermal energy is distributed through a body until a steady state, governed by the Laplace equation

$$\nabla^2 u = 0$$

Let's consider the kinematic system represented in Figure 1, which illustrates a an interior vertex $a$ connected to its neighbors $x_i$ through the edges of an unstructured grid.

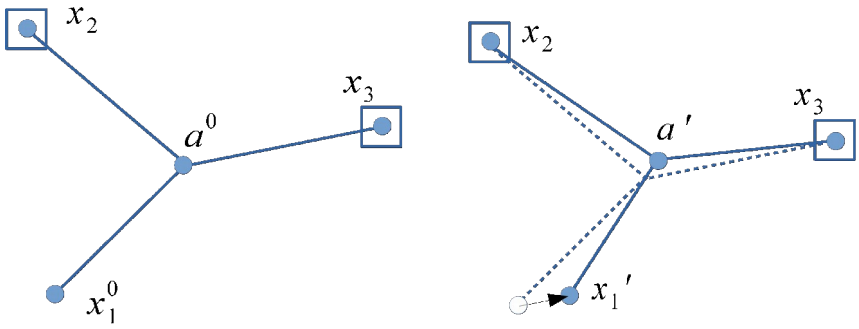



Figure 1: Kinematic system, where the vertex a is updated due to a perturbation of one of its neighbor.

Equivalently to the heat analogy, a mesh deformation rule is to minimize the variation of the relative vertex displacements $Min\{\partial(x_i\text{-}x_a)\}$, compared with the original ones. By imposing Dirichlet boundary conditions and using relative displacements as a field, this can be expressed as a problem to calculate the solution of the Laplace equation, which in the discrete form can be approximated as a linear system of equations:

$$\frac{\sum_i \left[\left(x_i' - x_a'\right) - \left(x_i^0 - x_a^0\right)\right] w_i}{\sum_i w_i} \tag{1}$$

where the term $w_i$ is the norm usually associated to the grid domain, such as the cell dimensions. The superscript $x^0$ indicates the original positions, while the superscript $x'$ denotes the updated vertex positions. The above expression can be rewritten and be solved wth an explicit Jacobi algorithm:

$$x_a^{n+1} = \frac{\sum_i \left(x_i^n - x_i^0\right) w_i}{\sum_i w_i} + x_a^0 \tag{2}$$

where $n$ is the iteration index. The norm selected in the test cases is the inverse squared length of the edge connecting two vertices:

$$w_i = \left| x_i^0 - x_a^0 \right|^{-2} \tag{3}$$

In this way, smaller elements, such as those close to the boundary layer, will remain stiffer, while deformations are absorbed by larger elements, which can sustain more strain. Alternatively, selecting the inverse of the dual area associated to the vertex or a weight calculated from the area of the associated elements, will achieve similar results. However, the edge distance turns out better for avoiding invalid elements in regions with high mesh stretching. In this formulation, displacements in each coordinate direction become decoupled and are solved independently. The norm are calculated from the original grid, before the deformation, which is taken as a reference. So, a good grid quality is expected from the original grid. The above expression corrects the relative position of the vertex, but it is insufficient to correct the skewness of the elements, or the orthogonality, which is the subject of the next section.

## 2.2 Mesh orthogonality correction

The concept of mesh orthogonality relates how a grid element faces an optimal angle. In convection-diffusion simulations a "false diffusion" is an error when the flow is not aligned

with the grid lines, which is accumulated downstream. Orthogonality is a relevant mesh quality metric, in especial at the boundary layer, and might compromise the robustness and accuracy of the simulations.

The mesh deformation algorithm presented in Eq. (2) can sustain large deformations, but there is no information about the shape of the elements. It deals very well with translation transformations, e.g. if the airfoil is moved to the left. On the other hand, rotations present a more challenging problem and the orthogonality is required to be corrected, to maintain the cell lines perpendicular to the wall. This can be achieved by preserving the angle of the elements, as it is illustrated in Figure 2.

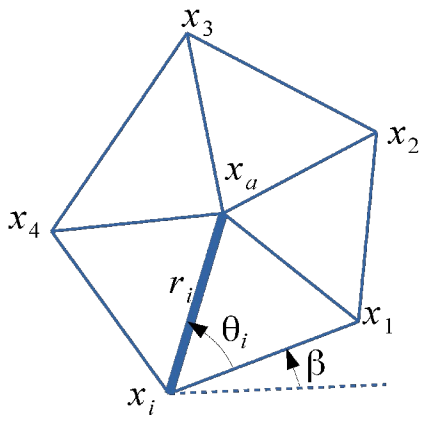



Figure 2: Kinematic system, where the vertex a is updated to maintain the angle θ of the element.

In a similar way to section 2.1, the aim is to minimize the variation of the angles after deformation: $Min\{\partial\theta\}$ (this can be trivially rewritten as $\theta'=\theta^0$). The reference angles typically are 90º for quadrilaterals and 60º for triangles. In our test cases, the angles taken as reference are those of the original grid, before deformation. In two dimensions, this can be expressed as a coupled linear system of equations.

$$
\begin{aligned}
x_a^{n+1} &= \alpha \frac{\sum_i \left[x_i^n + r_i^0 \cos(\theta_i^0 + \beta_i^n)\right] w_i}{\sum_i w_i} + (1-\alpha) x_a^n \\
y_a^{n+1} &= \alpha \frac{\sum_i \left[y_i^n + r_i^0 \sin(\theta_i^0 + \beta_i^n)\right] w_i}{\sum_i w_i} + (1-\alpha) y_a^n
\end{aligned}
\tag{5}
$$

where the weight $w_i$ is the same as Eq. (4), the terms $x$ and $y$ are the Cartesians coordinates, while the angles $\beta$ and $\theta$ are illustrated in Figure 2. The superscript $\beta^n$ indicates the updated value at iteration level and $\beta^0$ the reference value, which in our test cases are the angles of the original mesh elements. The above formulation is ill-defined (there are oscillations between two local equilibrium states), and therefore requires a relaxation coefficient $\alpha$ to ensure the convergence; a value of $\alpha=0.5$ ensures the convergence. The algorithm corrects the orthogonality, but cannot prevent elements from squashing. The final algorithm comes as the combination of both approaches.

$$X^{n+1} = \gamma \cdot X_p^n + (1-\gamma) \cdot X_g^n \tag{5}$$

where the term $X_p$ are the updated coordinates from Eq. (3) that maintains the vertex relative positions, and $X_g$ are the updated coordinates from Eq. (5) that corrects the orthogonality of the mesh. The value of $\gamma$ highly depends on the severity of the deformation. The orthogonality correction has slow convergence, while the algorithm based on the relative positions is very robust. In the test cases presented in this work, the strategy employed is to start with $\gamma=0$, to give more strength to the relative position correction, then gradually increase the orthogonality correction, as the residuals drops to a final value of $\gamma=1$.

## 3 TEST CASE

The test case considered is a two dimensional hybrid grid for a RAE2822 airfoil, as shown in Figure 3, where the geometry is rotated 20º. The proposed grid deformation method is compared with the advancing front algorithm [3].

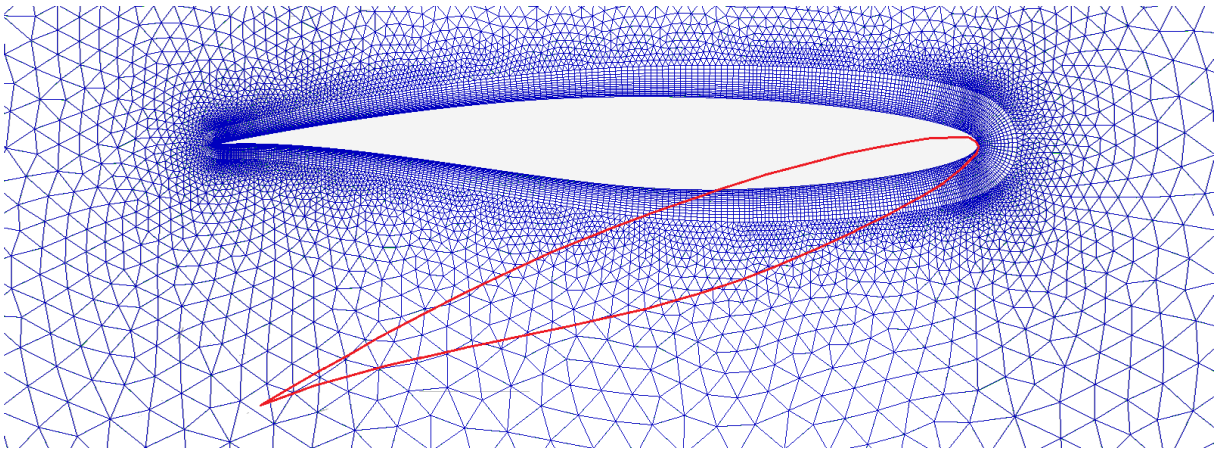


Figure 3: Original RAE2822 hybrid grid and the rotation tested for the mesh deformation.

Figure 4 and Figure 5 compare the mesh updates performed with the two different methods. Even when none of the methods turn out with invalid elements, there are disparities on the grid quality. The first one is the advancing front algorithm, which presents squashing cells for this test case. The Laplacian deformation, in comparison, is able to maintain the integrity of the elements, but the grid lines are not perpendicular to the wall, which might compromise the accuracy of the simulations. Both methods are then post-processed with the orthogonality correction presented in section 2.2, delivering good quality grids. Both post-processed grids lead to almost identical deformations, because the orthogonality correction is based on the Laplace operator, and therefore, the solution is theoretically unique. Disparities can only be explained at numerical level.

The advancing front algorithm is very fast and computationally efficient; in contrast, the Laplace deformation requires a long number of iterations. The orthogonality correction is then applied to both grids; because the starting grid is already close to the optimal, convergence is very fast.

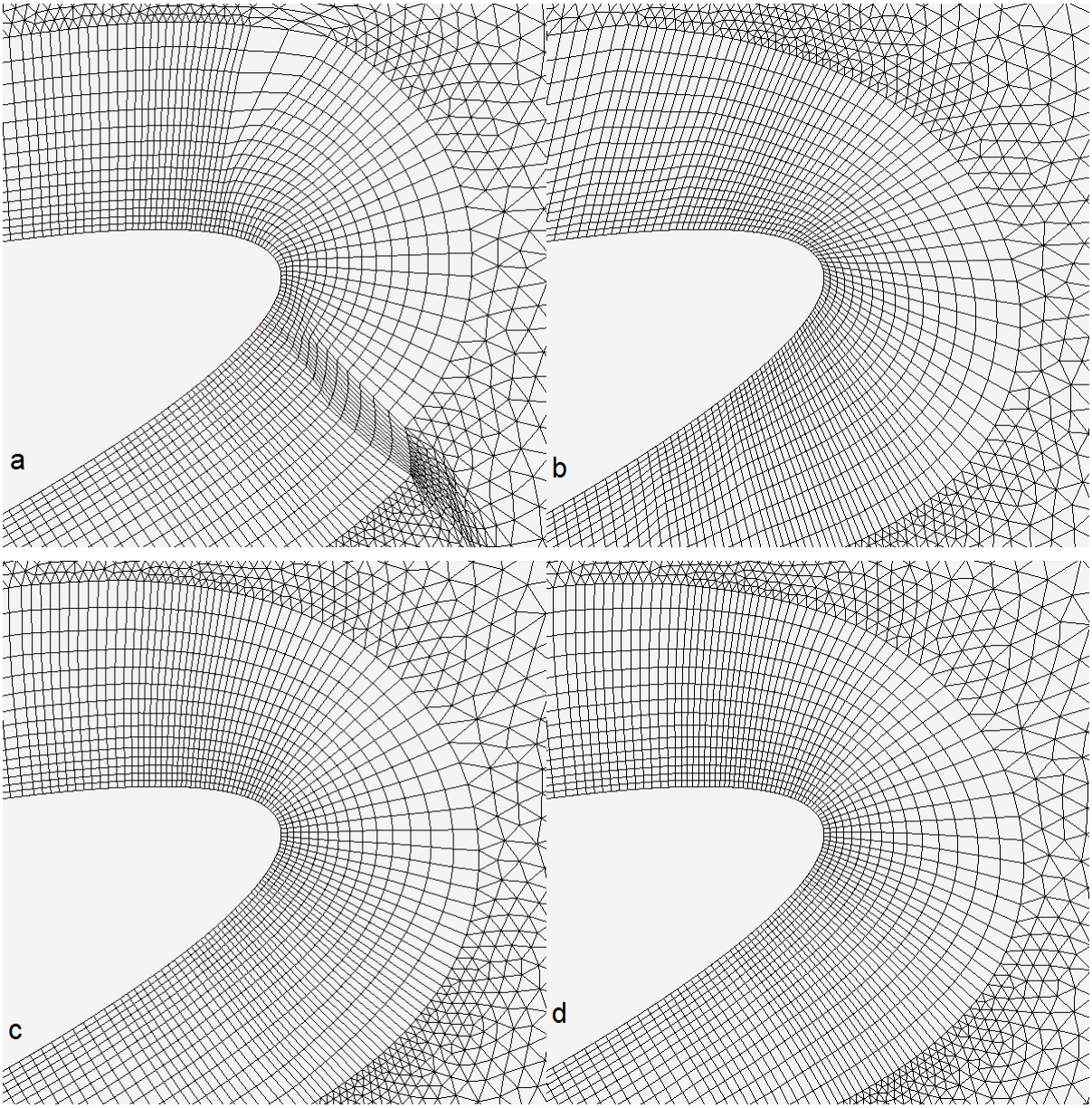



Figure 4: Details at the leading edge of the mesh deformation algorithm. a) Advancing front, b) Only Laplacian deformation, c) Advancing front with orthogonality correction, and d) Laplacian with orthogonality correction.

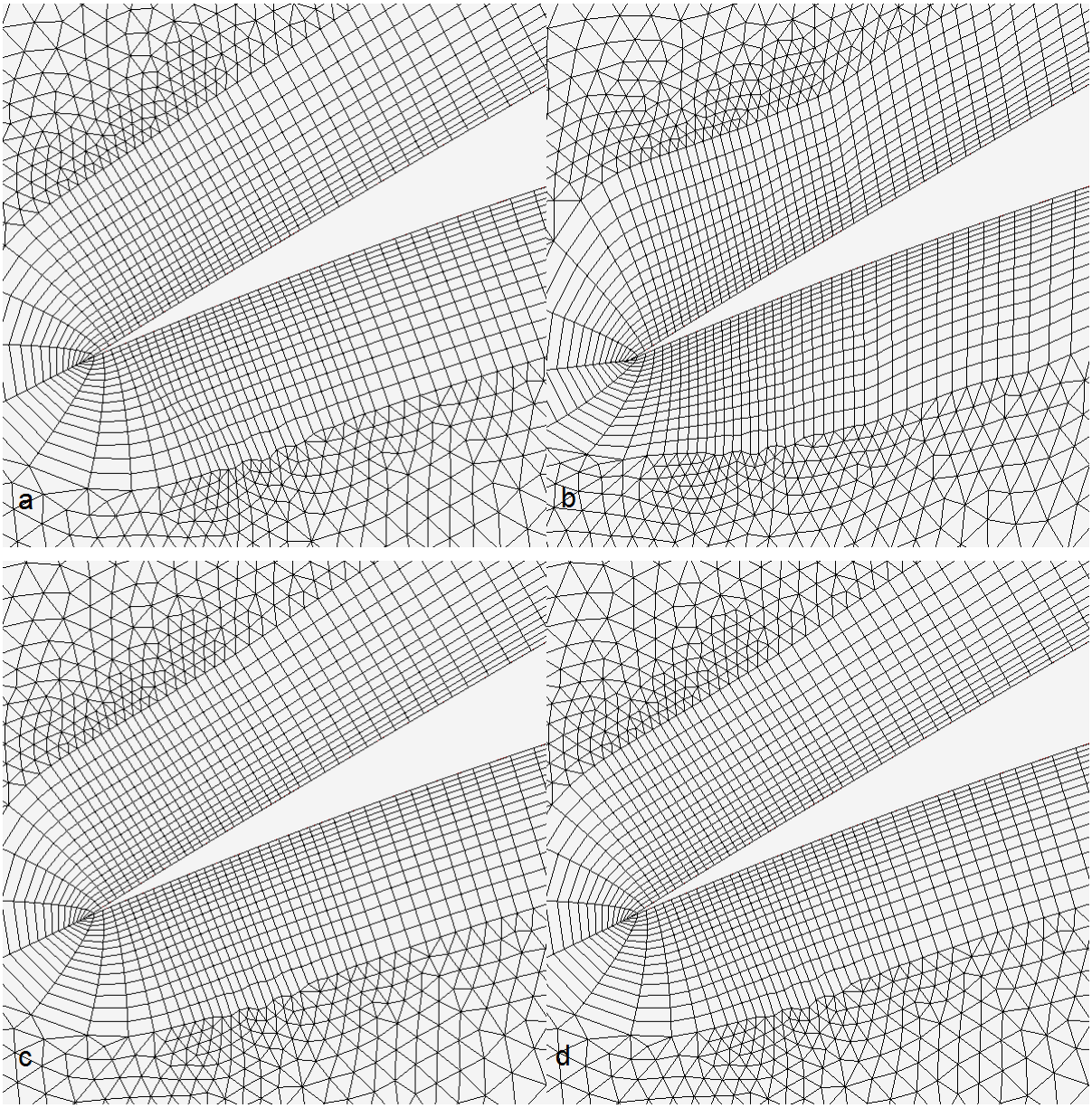



Figure 5: Details at the trailing edge of the mesh deformation algorithm. a) Advancing front, b) Only Laplacian deformation, c) Advancing front with orthogonality correction, and d) Laplacian with orthogonality correction. Note: c and d are almost identical grids, because the orthogonality correction is based on the Laplace operator, and therefore, the solution is unique.

The skewness is used to compare the grid quality with the methods exposed, as shown in Figure 6. As expected, the Laplacian deformation without orthogonality correction is not able to maintain the correct skewness of the elements. Hexahedrons are associated to the boundary layer and a value close to the original grid, previous to the deformation, indicates that most of the deformation is transferred to the prismatic layer, where is less critical. The presented algorithm, with orthogonality correction, shows a grid quality very close to the original.

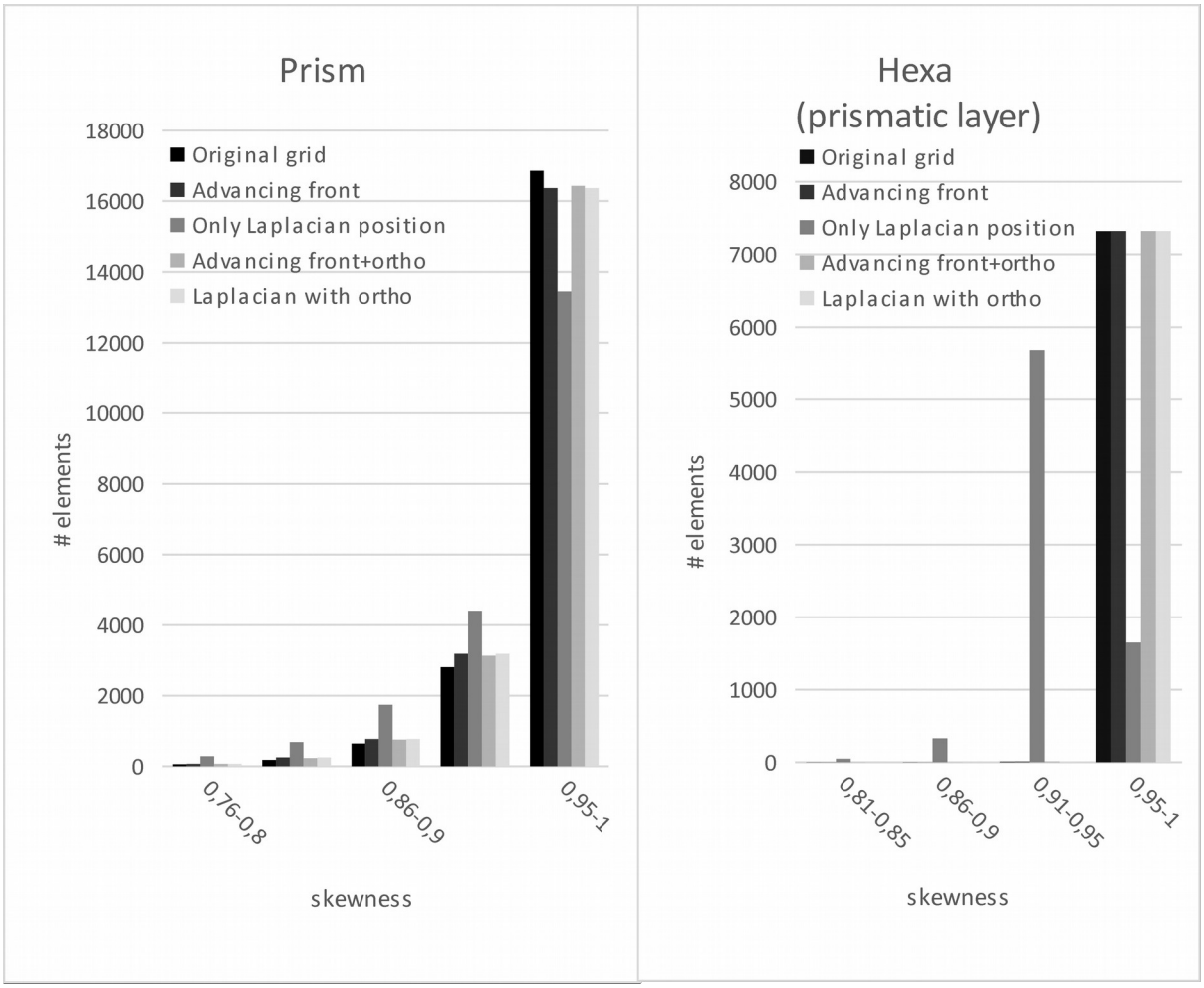



Figure 6: Comparison of the skewness, as a measurement of the quality of the grid.

## 4 CONCLUSIONS

A mesh deformation technique with orthogonality correction is presented in this work, inspired in the Laplace operator. The proposed method can maintain good grid quality, even in the case of large deformations. The method is very easy to implement, resulting into an explicit Jacobi iterative algorithm. Moreover, the algorithm is an element-based data structure ,and it is suitable for parallelization. In comparison, the advancing front algorithm is computationally more efficient, but the grid quality is uncertain. On the other hand, the Laplacian deformation requires an iterative algorithm. Using the grid delivered from the advancing front algorithm, and then applying the Laplacian deformation correction, to fix bad shaped elements, convergence is very quickly reached.

Future works would extend the proposed method to three dimensional grids and testing the method for multi-comonent grids. A further research will tackle how to improve the convergence rate to reduce the computational time; one posibility may be to take advantage of multigrid techniques.

## REFERENCES


[1] L. Dubuc, F. Cantariti, M. Woodgate, B. Gribben, K.J. Badcock, and B.E. Richards, "A grid deformation technique for unsteady flow computations", International Journal of Numerical Methods in Fluids, vol. 32, Issue 3, 2000, pp.285-311.

[2] S.P. Spekreijse, B.B. Prananta, and J.C Kok, "A simple, robust and fast algorithm to compute deformations of multi block structured grids", NLR-TP-2002-105, 2002.

[3] T. Gerhold, J. Neumann, "The parallel mesh deformation of the DLR TAU-Code", vol. 96 of Notes on Numerical Mechanics and Multidisciplinary Design (NNFM), Springer Berlin Heidelberg, 2008, Chapter 20, pp. 162-169.

[4] Batina, J.T.,"Unsteady Euler Airfoil Solutions. Using Unstructured Dynamic Meshes", AIAA Journal,. Vol. 28, No. 8, 1990, pp.1381-1388.

[5] C. Farhat, C. Degand, B. Koobus, and M. Lesoinne, "Torsional Springs for Two-Dimensional Dynamic Unstructured Fluid Meshes", Computer Methods in Applied Mechanics and Engineering, Vol. 163, 1998, pp. 231-245.

[6] C. Degand, C. Farhat, A three-dimensional torsional spring analogy method for unstructured dynamic meshes, Comput. Struct. 80 (2002) 305.

[7] A. Masud, M. Bhanabhagvanwala, R. A. Khurram, "An adaptive mesh rezoning scheme for moving boundary flows and fluid-structure interaction", Computers & Fluids 36 (2007) 77-91.

[8] Z. Yang, D.J. Mavriplis, "Unstructured dynamic meshes with higher order time integration schemes for the unsteady Navier-Stokes equations", AIAA 2005-1222, 41th AIAA Aerospace Sciences Meeting and Exhibit, January 10-13, 2005/Reno, NV.

[9] A. de Boer, M.S. van der Schoot, and H. Bijl, "Mesh deformation based on radial basis function interpolation", Computers and Structures, Vol. 85, 2007, pp. 784{795.

[10] T.C.S. Rendall, and C.B. Allen, "Unified Fluid-Structure Interpolation and Mesh Motion using Radial Basis Functions", International Journal for Numerical Methods in Engineering. Vol. 74(10), 2008, pp1519-1559.

[11] S. Jacobsson, and O. Amoignon, "Mesh deformation using radial basis functions for gradient-based aerodynamic shape optimization", Computers & Fluids 36 (2007) 1119-1136.

[12] T. Gillebaart, A.H. van Zuijleny, and H. Bijlz, "Radial basis function mesh deformation including surface orthogonality", 54th AIAA Aerospace Sciences Meeting, 4-8 January 2016, San Diego, California, USA, AIAA 2016-1674.

[13] A. Masud, M. Bhanabhagvanwala, R. A. Khurram, "An adaptive mesh rezoning scheme for moving boundary flows and fluid-structure interaction", Computers & Fluids 36 (2007) 77-91.

[14] H. Kanchi, A. Masud, "A 3D adaptive moving scheme", International Journal for Numerical Methods, in Fluids 54 (6-8), 923-944.

[15] Nielsen, E. J., and Anderson, W. K., "Recent Improvements in Aerodynamic Optimization on Unstructured Meshes". AIAA Journal, Vol. 40, No. 6, June 2002, pp. 1155–1163

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN INTEGRATED FRAMEWORK FOR WRAPPING AND MESH GENERATION OF COMPLEX GEOMETRIES

**David G. Martineau**[1], **Jeremy D. Gould**[1], **and Jacques Papper**[1]

[1] ICON Technology & Process Consulting Ltd
Berkshire House, Thames Side Windsor, SL4 1QN
{d.martineau,j.gould,j.papper}@iconcfd.com

**Keywords:** Mesh Generation, Surface wrapping, Complex geometry.

**Abstract.** *This paper presents a combined mesh generation and surface wrapping capability which is able to create manifold watertight surfaces for complex industrial model assemblies. The wrapping functionality is able to join disconnected and overlapping surfaces. Large holes in the assembly are closed by performing the wrapping at intermediate stages of the mesh refinement. The ability to perform wrapping in combination with mesh generation ensures that holes are closed without loss of geometric fidelity. The wrapping capability is demonstrated on a range of test cases including an engine block and a highly detailed automotive case.*

## 1 INTRODUCTION

Thanks to significant advances in computational fluid dynamics (CFD) over the last few decades, and increasing availability of high performance computing resources, engineers are able to apply CFD to more and more complex flow simulations. However, robust and automatic mesh generation still represents a significant challenge in the overall process from CAD to CFD solution in industrial engineering analysis. A major bottleneck in the mesh generation process is the creation of a watertight representation of the geometry from the original CAD model. A watertight geometry representation is required as input to traditional structured and unstructured mesh generation methods, which first discretize the surface of the flow domain before meshing the interior. The translation of the geometry from the native CAD data into another format can often result in problems such as missing or duplicate parts, small gaps and overlaps [1]. Often the input to the simulation process is a model which has been designed for manufacture, not for the purpose of simulation, and therefore contains an unnecessary level of detail and geometry which should not be part of the fluid domain. Resolving these issues, and creating a geometry model suitable for mesh generation requires labour-intensive and time-consuming geometry repair and modification.

Mesh generation approaches such as octree or Cartesian-based methods [3, 4], which start with a volume mesh in the interior of the domain, and then obtain a boundary conforming surface mesh through cell-cutting or snapping procedures, have therefore become increasingly popular in recent years, in part due to their inherent tolerance to poor quality or "dirty" geometry [5]. However, although such meshing approaches do not need a watertight geometry representation, and are able to tolerate geometry issues such as overlaps and poor quality triangulations, they cannot handle so-called 'fully-resolved gaps', i.e. gaps in the geometry larger than the local element size [5]. There have been numerous attempts over the last decade to develop methods for automatic repair of holes in discrete geometry models. These methods broadly fall into two main categories: surface-based approaches and voxel or volume-based approaches.

Surface-based approaches operate directly on the discrete geometry input data and resolve defects in the geometry such as holes with open boundary edges. The closing of such holes is a relatively mature field of study, and there are various algorithms described in the literature [6-10]. Holes in relatively planar regions of the model can be easily patched via planar triangulation. To handle more complex holes, Jun [6] proposes a piecewise scheme which divides them into several simpler holes which are filled with a planar triangulation, before applying smoothing and sub-division techniques to improve the quality. Zhao et al [7] employ an advancing front method to close complex holes, and then solve the Poisson equation to reposition the new vertices. Alternative algorithms use radial basis functions [8] or the Level Set Method [10] to construct an implicit surface which covers the hole. Since surface-based methods operate only locally, they can be very efficient and incur only minimal perturbation of the input data. However, they can only close holes with an open boundary, and are not able to handle semantic holes caused by missing components, or gaps between parts in a model assembly.

Volume-based approaches operate by generating a volumetric representation of the surface, and then classifying the cells in the volume as either inside, outside, or intersecting the surface. The shrink wrapping technique developed by Lee et al [1] is one such example of this approach, and is based on the concept of the interior-to-boundary Cartesian mesh generation described by Wang [5]. Volume approaches such as this are typically fully automatic and can be very robust. However, large holes in the geometry can lead to mesh 'leaking' into areas of the domain which should not be modelled. Juretić and Putz [12] identify the holes lead-

ing to such mesh leaks by solving the heat diffusion equation and looking at regions of high heat flux intensity. However it is left to the user to then manually close these holes and repeat the process until no leaks are present. Kumar and Shih [13] propose a hybrid method, which applies surface-based repair techniques to close holes with simple topologies, and then obtains a watertight surface using the Marching Cubes algorithm [14] to extract the zero-set surface of the numerical solution to the diffusion equation. Whilst they demonstrate good quality results on models with complex topology, their approach relies on consistent orientation of the input geometry, and is not able to deal with semantic holes in an assembly.

This paper describes an integrated approach to wrapping and mesh generation implemented in iconCFD® [15], which is able to handle large gaps in model assemblies, including holes with open boundaries and semantic holes, without significant loss of geometric fidelity. The problems of resampling a geometry (e.g. loss of model features) that are typically encountered with volume-based geometry repair methods are also avoided in our approach by exploiting an adaptively-refined Cartesian grid to simultaneously perform the wrapping of the geometry and generate a boundary conforming, hexahedral-dominant mesh of the flow domain. Section 2 of this paper describes the various stages of the existing mesh generation process which formed the basis for this work. The extension of this capability to include surface wrapping is then described in section 3, and in section 4 the two different modes of operation are described. In section 5 the capability is applied to a number of models to demonstrate the ability of the approach to handle industrial geometries used in external aerodynamic and under-hood thermal management simulations. Finally, the paper finishes with conclusions and a discussion of directions for future work.

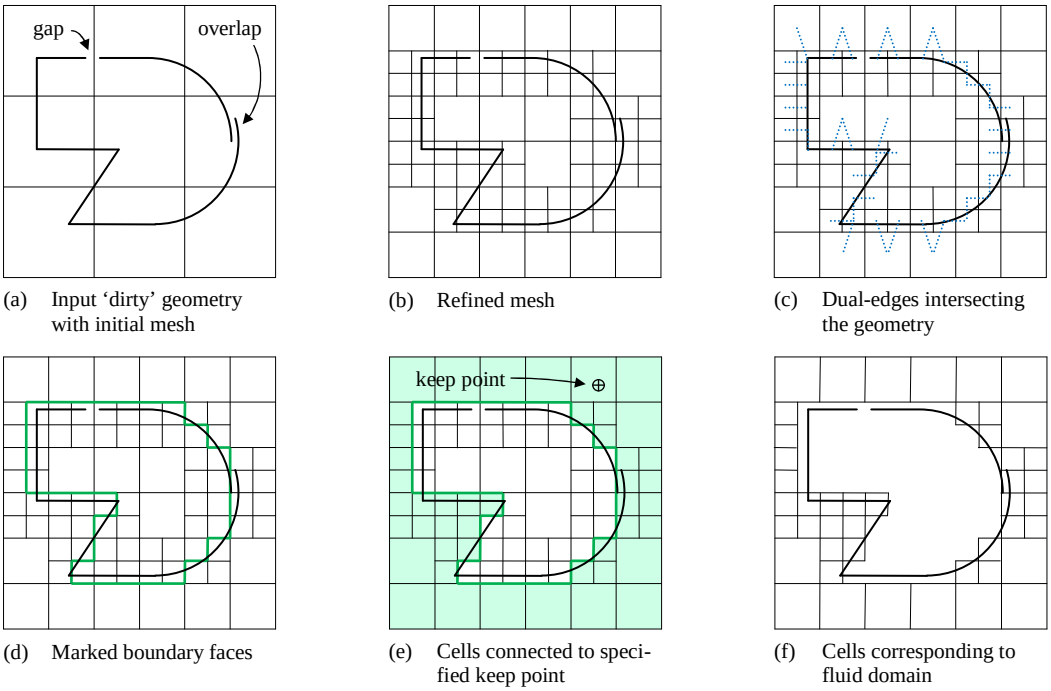
## 2 OVERVIEW OF MESHING APPROACH

The wrapping capability has been implemented within an existing parallel hexahedral-dominant mesh generator, iconHexMesh, which is part of the iconCFD process, an open source-based CFD software suite developed by ICON, using OpenFOAM® [16] technology. The mesh generation process in iconHexMesh comprises three main stages: (1) refinement of the initial block mesh based on prescribed refinement levels on geometry surfaces and feature lines, (2) creation of a conformal mesh by snapping to geometry surfaces, and (3) insertion of a layer mesh to capture viscous boundary layers. At various points during the mesh generation, the mesh is dynamically re-partitioned to achieve good load-balancing. The refinement and snapping stages are common requirements for both meshing and surface wrapping, and are therefore described in more detail in the following sections.

### 2.1 Refinement

The input to the meshing process is a discretized model of the geometry. In the automotive industry, the STL (stereo lithography) format is commonly used as input to the simulation process as it is a simple portable format which can be generated by most CAD systems and is easy to visualize [5]. The meshing process begins by reading one or more STL files describing the assembly to be meshed, and creates an initial coarse mesh of the domain, typically using a uniform grid of Cartesian-aligned cells which completely encloses the entire geometry, as illustrated in Figure 1(a). Successive iterations of local mesh refinement are then applied, until the required level of mesh refinement is reached (Figure 1(b)). The mesh generator supports a number of different refinement criteria to capture the details of the geometry, including:

- Uniform refinement of geometry surfaces and feature lines
- Curvature-based surface refinement
- Proximity-based refinement of surfaces and feature lines
- User-defined volumes of refinement




(a) Input 'dirty' geometry with initial mesh
(b) Refined mesh
(c) Dual-edges intersecting the geometry
(d) Marked boundary faces
(e) Cells connected to specified keep point
(f) Cells corresponding to fluid domain

Figure 1: Illustration of the refinement process in iconHexMesh.

Once the refinement criteria have all been met, 'dual-edges' of the mesh are tested for intersection with the geometry (Figure 1(c)). A 'dual-edge' corresponding to an interior face of the mesh is the line joining the centres of the two cells neighbouring the face. If this edge intersects the geometry, then the face is marked as a boundary face (Figure 1(d)). By painting cells connected to one or more user-specified keep points without crossing boundary faces, the fluid region to be kept is identified (Figure 1(e)). Any cells on the other side of the set of boundary faces from the keep point are then discarded (Figure 1(f)).

## 2.2 Snapping

The outer surface of the remaining mesh, which will be castellated at this stage, is then snapped to the geometry to provide a body-conforming volume mesh. This process involves a combination of smoothing of the nodes on the boundary, and projection to the geometry, and as such shares a degree of similarity with the shrink wrapping approach described by Kobbelt et al [17]. A full description of the snapping algorithm is beyond the scope of this paper, but the overall process is summarized as follows:

1. Apply a few iterations of global Laplacian smoothing to the initial castellated surface. This helps to ensure that the subsequent projection onto the geometry does not result in self-intersection of the mesh.
2. Project selected nodes in the surface mesh to feature lines in the geometry. These feature lines are automatically detected based on discontinuities (i.e. sharp edges) in the discretized geometry model, or region boundaries. For each feature line, a

corresponding path of connected nodes is found in the surface mesh, which is then projected to the feature line. The deformation field arising from this projection is smoothed into the surrounding mesh, minimizing the distortion of the surrounding mesh elements.

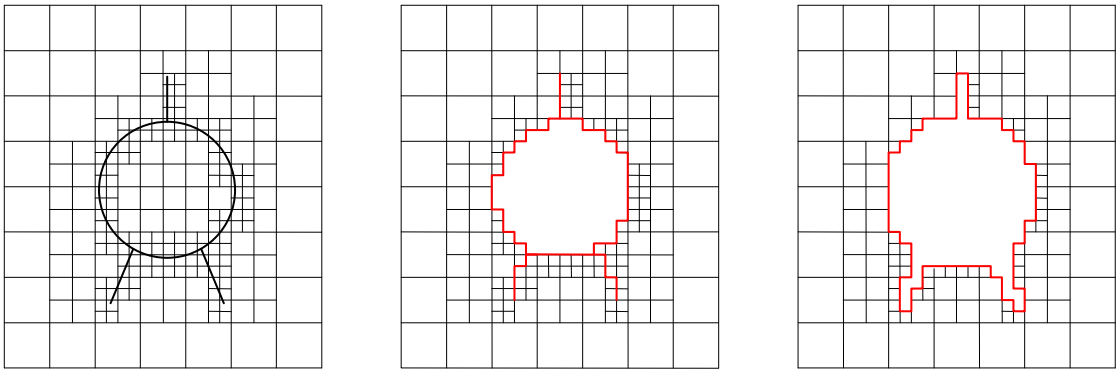
3. Project all remaining boundary nodes to the geometry. Rather than simply projecting each node to the closest point on the geometry, this process is performed iteratively, and attempts to gradually move and rotate the faces of the surface mesh to match the local normal direction of the geometry.
4. Smooth the final surface locally where the distance from the face centre to the geometry is large. This relaxation of the surface mesh where the snapping has failed to accurately capture the geometry improves the local quality of the surface mesh, and aids in the subsequent insertion of layer mesh.

Throughout the snapping process, the geometric quality of the volume mesh is monitored. Any deformation which would result in degradation of the mesh quality below user-specified limits is locally scaled back to ensure that the mesh quality constraints are satisfied. Similarly, any topological changes to the mesh which could violate these mesh quality constraints can be reversed to guarantee a resulting volume mesh of acceptable quality.

## 3 WRAPPING IMPLEMENTATION

### 3.1 Improving tolerance to gaps

The method by which the boundary faces of the mesh are determined in iconHexMesh is quite different to conventional Cartesian and octree mesh generation. Typically, Cartesian or octree mesh generation defines the boundary faces as the exposed faces remaining after removal of the intersected cells and cells outside the flow domain, as illustrated in Figure 2(c). The advantage that the approach described in section 2.1 has over this latter approach is that thin non-manifold surfaces in the geometry can be modelled with a ‘sheet’ of faces in the volume mesh. In general, the approach allows the topology of the initial set of boundary faces to more closely match the topology of the geometry, as shown in Figure 2(b). This enables much more accurate capture of the non-manifold parts of the geometry, such as the open edges of thin surfaces.



(a) Non-manifold geometry and refined mesh

(b) Boundary faces obtained using intersecting dual-edges

(c) Boundary faces obtained using boundary of intersected cells

Figure 2: Capture of thin surfaces.

There is a disadvantage to the ‘dual-edge’ intersection method for defining boundary faces however. With this approach, there exists the possibility that a dual-edge passes through a

gap in the geometry, thus leaving a hole in the boundary faces which permits the volume mesh to 'leak' into parts of the volume which are not intended to be part of the flow domain.

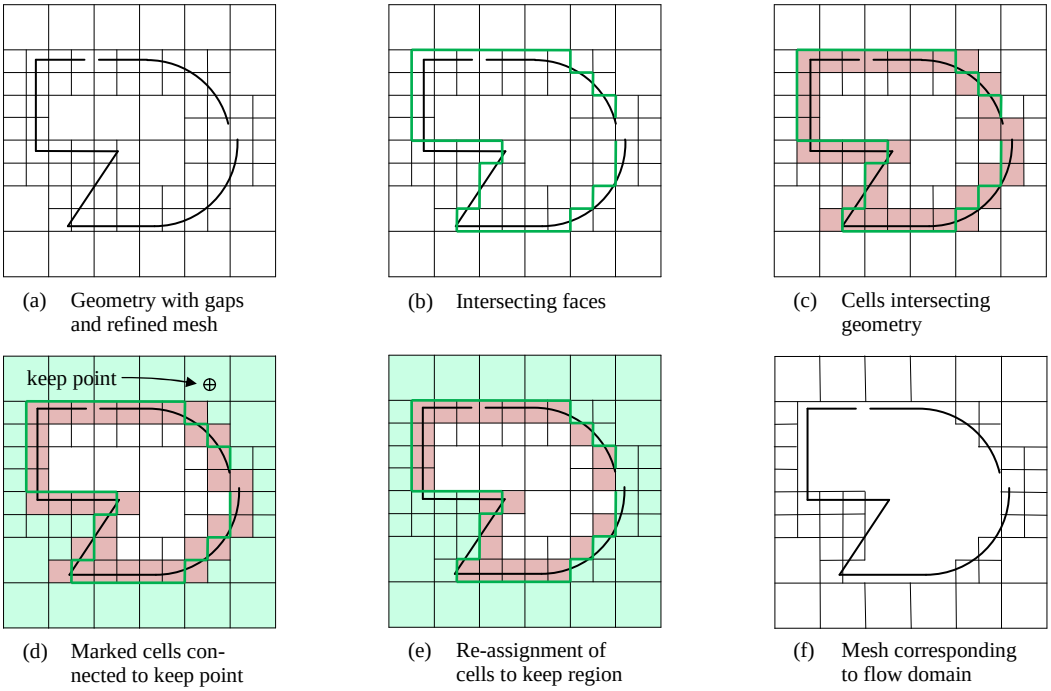



Figure 3: Modified process for identification of fluid region.

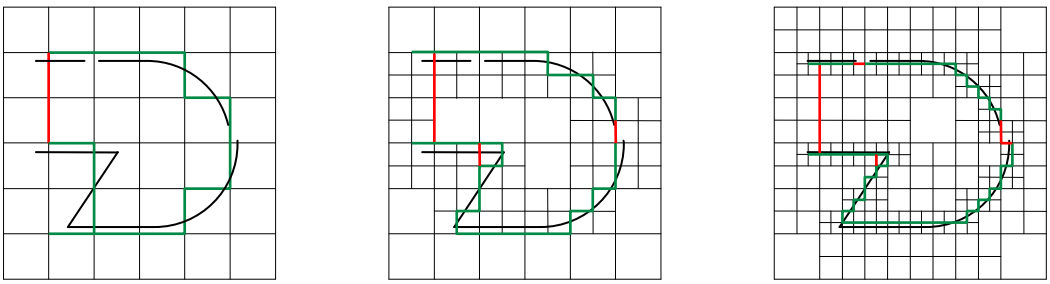
This limitation is overcome by combining the existing method for detecting boundary faces with the more conventional method. The new method is depicted in Figure 3 and starts, as in the original method, with marking of faces whose dual-edges intersect the geometry. These are defined as 'intersecting' faces. The nodes belonging to these intersecting faces are also marked as boundary nodes. The algorithm then marks all cells which intersect the geometry, and paints all the remaining cells connected to the keep point (the keep region). The intersecting cells which share a non-boundary node with the cells in the keep region are added to this region, and finally all cells not in the keep region are discarded. The boundary faces are then defined as the faces on the outer surface of the keep region.

Although the modified process described above improves the tolerance to dirty geometry, it cannot handle the situation of fully resolved gaps, as discussed in [5], where the size of the gap is larger than the local element size. Unfortunately, as computing resources increase and engineers attempt to improve the accuracy of flow simulations by resolving finer and finer detail, the possibility of creating a fully resolved gap becomes more and more likely. In the automotive industry for example, small gaps between body panels or paths through the heating, ventilation and air-conditioning (HVAC) system can connect the region outside the car to the region inside the car cabin. Although tools are available in iconCFD® to detect and visualize leaks in a volume mesh, this still requires labour-intensive visual inspection and manual repair of the model, and prevents the formation of a fully automated analysis process.

### 3.2 Wrapping at coarser levels

To address the issue of fully resolved gaps in the meshing process, an iterative gap-filling methodology is proposed. The iterative gap-filling methodology starts by applying the modified boundary identification process, described in the previous section, at an intermediate

stage of the refinement process, as illustrated in Figure 4(a). This intermediate stage of the refinement process is achieved by refining the mesh until a user-defined refinement level has been reached on all surfaces of the geometry. This is referred to as the wrap level for a surface and is either equal to or less than the maximum refinement level associated with a surface. If the wrap levels have been set correctly, then the set of boundary faces identified in this coarse mesh should form a closed surface enclosing the region to be modelled.



(a) Gap faces (red) identified at level 1

(b) Gap faces (red) identified at level 2

(c) Gap faces (red) identified at level 3

Figure 4: Iterative identification of gap faces during refinement process.

Having created a closed set of boundary faces, a subset of these faces which close the gaps in the geometry can be identified as those whose corresponding dual edge does not intersect the geometry. The refinement process now continues, with the mesh being refined from the wrap level to the maximum level required. The set of gap faces is updated during the mesh refinement to take into account changes to the mesh topology. At each step in the refinement, the process of identifying the boundary faces must be repeated as shown in Figure 4(b) and (c), since the additional refinement may otherwise cause a gap to be created.

The set of gap faces is also filtered at each step of the refinement process, to minimize the area of the faces closing a hole and improve the quality of the final wrapped surface.

### 3.3 Surface snapping

After the refinement process, and removal of the mesh outside of the flow domain, the boundary nodes of the mesh are snapped to the surface of the geometry as described in section 2.2. However, in the original implementation, it could be assumed that all the boundary nodes were within a relatively small distance from the geometry. With the creation of gap faces spanning potentially large gaps in the geometry, this assumption is no longer valid. However, by maintaining a list of mesh faces corresponding to the gaps in the geometry, it is possible to apply a special treatment to these nodes. Firstly, more aggressive smoothing is applied to the gap faces to improve the quality of the surface mesh in these regions. After this initial smoothing, and snapping to feature lines on the geometry, the dual-edges corresponding to the gap faces are tested for intersection with the geometry. Any intersection of these dual edges indicates that the gap face likely corresponds to a crack in the geometry, rather than a fully resolved gap, and is therefore un-marked as a gap face. During the rest of the snapping process, the displacement of the boundary nodes to conform to the geometry is propagated to the remaining gap faces from the neighboring boundary nodes.

Once the snapping stage has been completed, the gap faces are assigned to neighbouring patches. This allows a consistent layer mesh to be inserted on boundary patches, irrespective of the presence of any cracks or gaps.

### 3.4 Pure Wrapping Mode

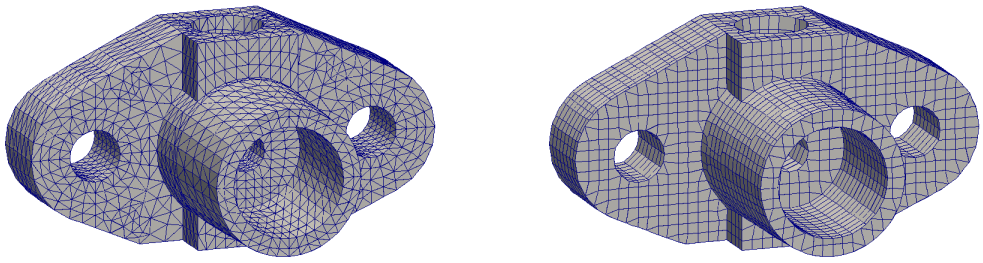
If a wrapped surface is the only output required, then the quality of the volume mesh can be disregarded. This allows for a much faster snapping process, since no checks on the volume mesh quality need to be performed, and there is no need for scaling of the mesh deformation when moving the boundary points onto the surface of the geometry.

Finally, in pure wrapping mode, the surface mesh is triangulated and output to STL format, preserving any patch IDs assigned to the original geometry. The gap faces can either be assigned to a separate patch, or assigned to the nearest adjacent patch.

## 4 RESULTS

### 4.1 Flange

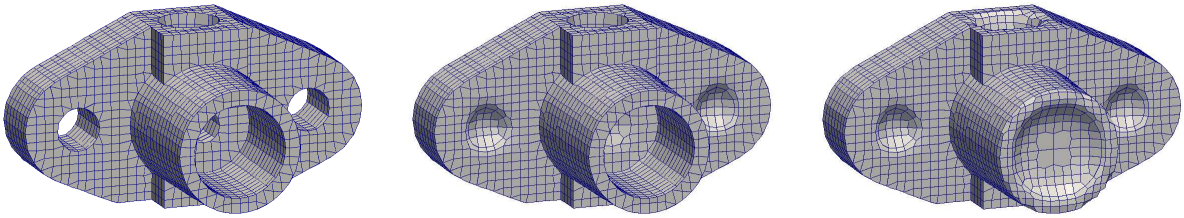
To demonstrate the ability to wrap holes larger than the local element size, the wrapping capabilities of iconHexMesh are applied to the Flange geometry from the OpenFOAM® tutorial [18]. This is a simple geometry, as shown in Figure 5(a), which features several holes and recesses of varying diameter. An initial uniform Cartesian mesh of element size 0.02 is created which encloses the geometry, as shown in Figure 5(b). The iconHexMesh mesh generator is then used to apply 4 levels of surface refinement, resulting in the final surface mesh shown in Figure 5(c), which accurately captures the sharp features of the geometry.



(a) STL geometry (b) Surface mesh

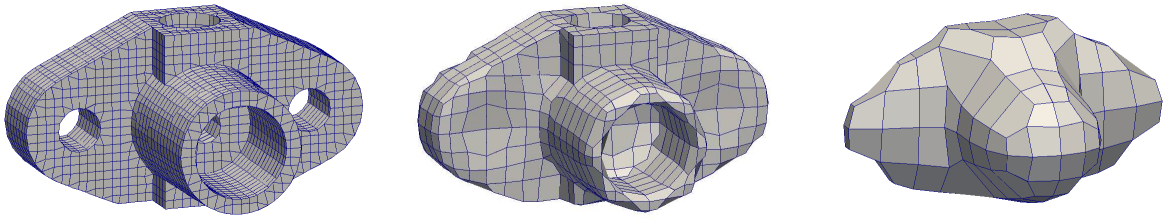
Figure 5: Flange geometry and corresponding surface mesh.

The series of images in Figure 6 show how iconHexMesh is able to close off different size gaps in the geometry using varying wrap levels, but with the same final refinement level in all cases, enabling the geometric fidelity of the model to be maintained as much as possible. In contrast, Figure 7 shows how use of varying refinement level to close off the holes in the mesh results in considerable loss of geometric fidelity.



(a) Wrap level = 4 (b) Wrap level = 3 (c) Wrap level = 2

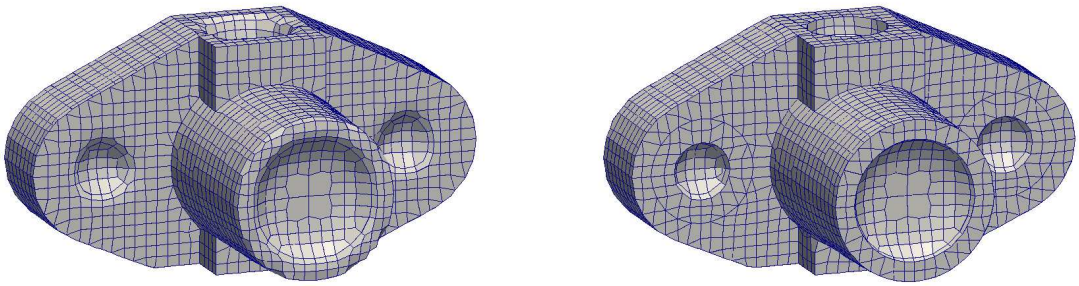
Figure 6: Flange surface mesh at different wrap levels.



(a) Refinement level 4 (b) Refinement level 3 (c) Refinement level 2

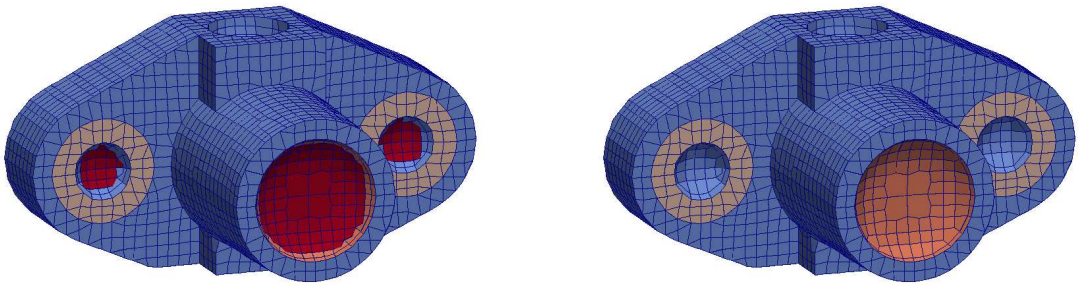
Figure 7: Flange surface mesh at different refinement levels.

The geometric fidelity of the surface mesh can be further improved by enabling the feature line snapping functionality in iconHexMesh, as illustrated in Figure 8. The feature line snapping is able to improve the capture of sharp edges on the geometry, as well as properly resolving the region boundaries on the flat surface of the model. Finally, Figure 9 shows how the gap faces can be kept in a separate patch, or assigned to the same patch as neighbouring faces in the surface mesh.



(a) Surface mesh with basic surface snapping (b) Surface mesh with feature line snapping

Figure 8: Improved geometry capture using feature line snapping.



(a) Gap faces (red) put in separate patch (b) Gap faces assigned to neighbouring patches

Figure 9: Surface mesh shaded by patch ID.

## 4.2 External Aerodynamics Case

In this section a few holes have been deliberately added to a model of the Koenigsegg Agera hypercar to illustrate how the combined wrapping and meshing capability is able to handle geometry defects commonly encountered in automotive applications. The geometry of the car is shown in Figure 10.

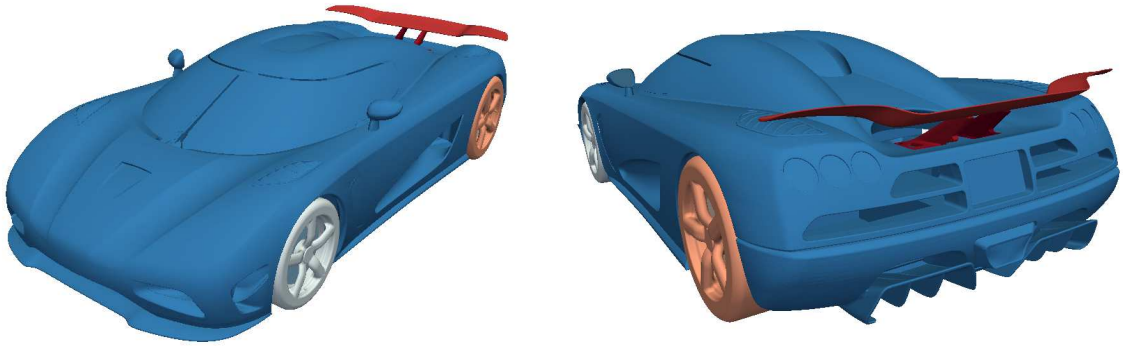


Figure 10: Koenigsegg Agera geometry.

To create a mesh on the Koenigsegg Agera vehicle, an initial Cartesian mesh was generated around the car and wind tunnel with a base size of 3 m. A minimum refinement level of 8 was applied to all of the car surfaces, giving a maximum element size of approximately 12 mm on the car surface. Between one and three levels of curvature refinement were applied to capture the feature edges and highly curved regions of the car, such as the rear wing blade, leading to a minimum element size of 1.5 mm. To capture the viscous boundary layer, 6 layers of cells are inserted at the car surface, with a first layer thickness of 1 mm.

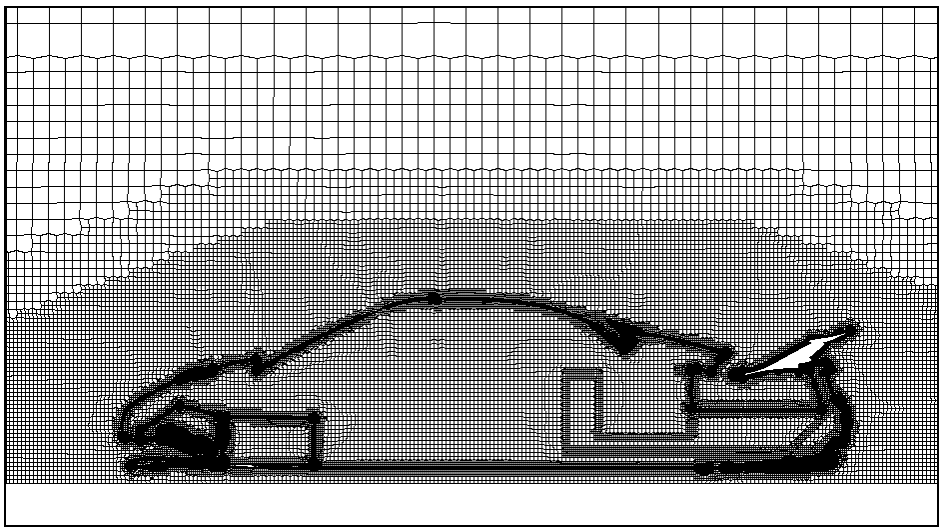


Figure 11: Slice through volume mesh with leakage into cabin region.

A slice through the volume mesh generated without wrapping is shown in Figure 11, clearly indicating that mesh has leaked into the car interior, resulting in a mesh which is not suitable for an external aerodynamic flow simulation. A flow simulation on this mesh will waste valuable computational resources, and may fail to converge due to the stagnant region inside the car.

A leak detector utility is provided within the iconCFD software suite, which can be run in order to help find gaps in geometries in this type of situation. Figure 12 shows a path (in red) of cells detected using this tool which connect a user-specified orphan point inside the car to the keep point. It can easily be seen in this case that the leakage path corresponds to a gap around the wind screen highlighted in blue in Figure 12. However, there may be many leaks in the geometry and the process of finding and closing them all can be slow and labour-intensive on a complex industrial model.

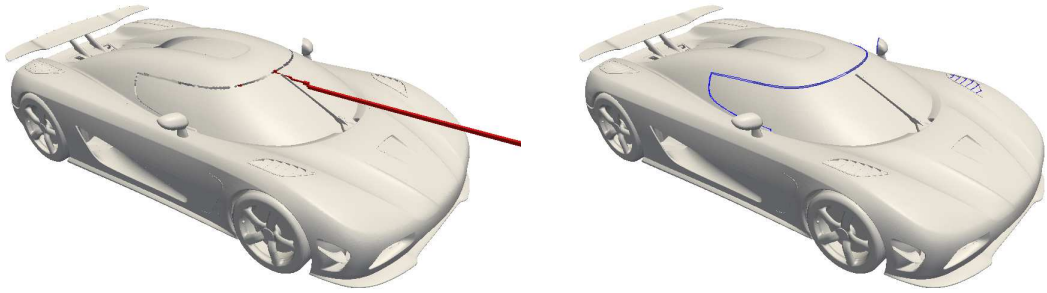


Figure 12: Leakage path detected (red) and corresponding hole in geometry (edges in blue).

In order to close the holes in the geometry without the need for any leak detection and manual repair, a wrap level of 7 was applied to all the car surfaces during the meshing process. This should close any holes of size 23 mm or smaller. All other meshing parameters were kept the same. Several views of the gaps in the geometry and the surface mesh generated in these regions are shown in Figure 13. A slice through the resulting surface volume mesh is shown in Figure 14, indicating that the wrapping has managed to successfully close all the gaps and prevent the mesh leaking into the interior.

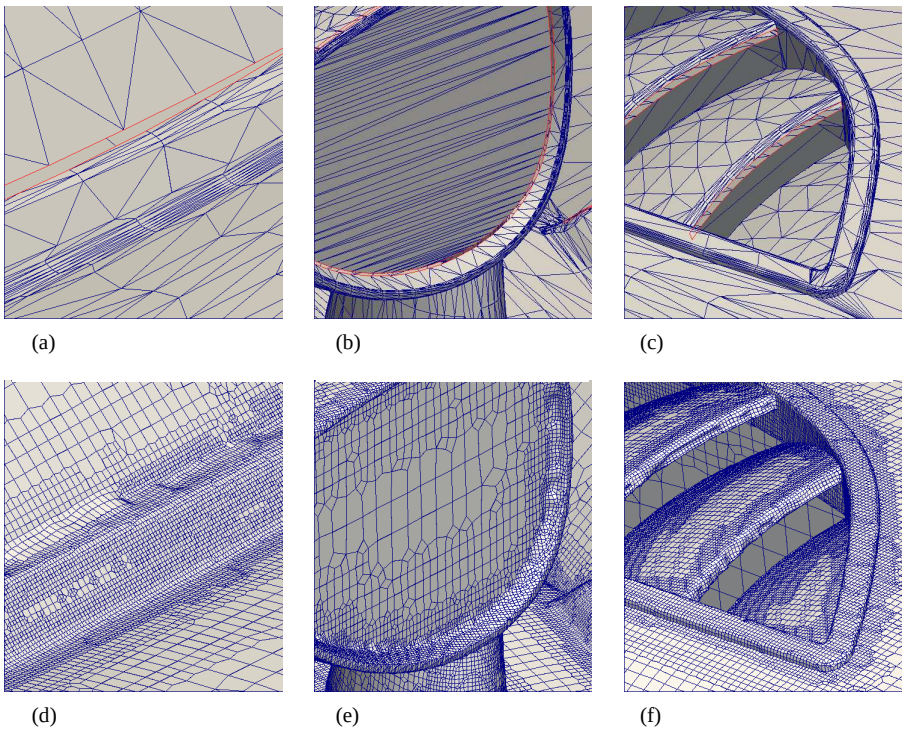



Figure 13: Details of car geometry (a), (b), (c) and corresponding surface mesh (d), (e), (f).

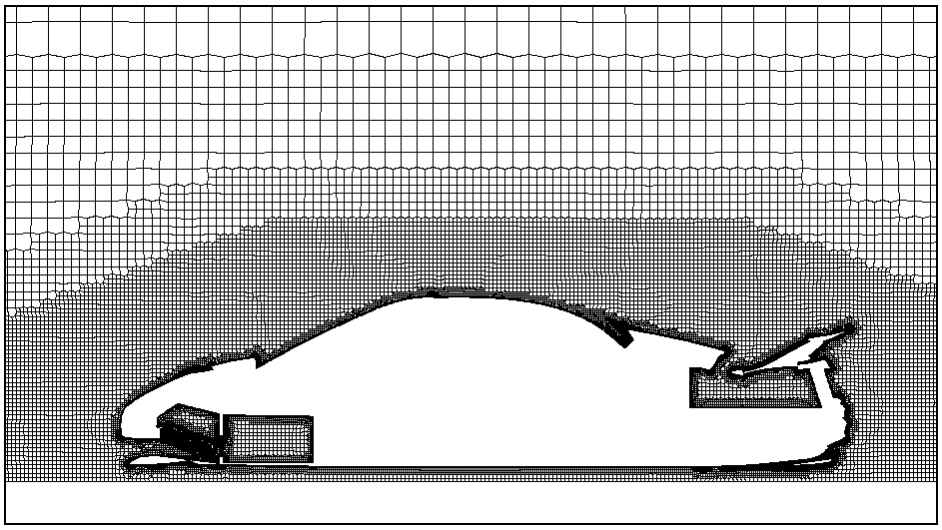


Figure 14: Mesh of car with wrapping activated.

Table 1 gives some statistics of the meshes obtained with and without wrapping activated during the meshing process. These indicate that the mesh leakage into the cabin incurs a 79% increase in element count, which could result in a significant waste of computational resources if not detected early in the analysis process. Both meshes were generated using 32 cores of a Linux cluster with Intel Xeon E5-2670 (2.60GHz) processors.

| | Mesh 1 (unwrapped) | Mesh 2 (wrapped) |
|---|---|---|
| Time taken (s) | 6,593 | 6,907 |
| Number of cells | 66,826,035 | 37,366,550 |
| Number of faces | 215,589,109 | 120,713,107 |
| Number of points | 83,252,580 | 46,732,723 |

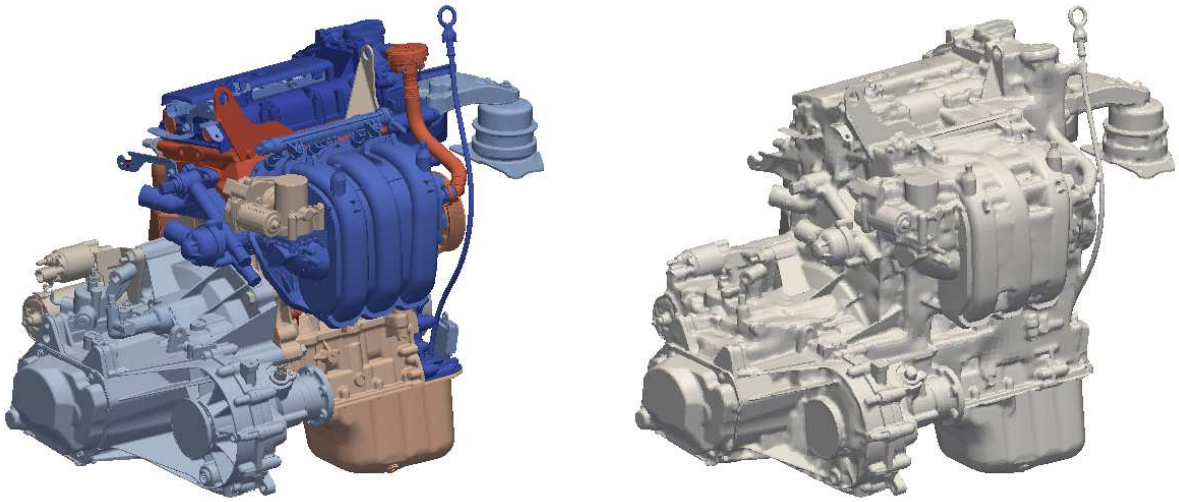
Table 1: Comparison of mesh statistics.

### 4.3 Engine Block

In the automotive industry, under-hood thermal management (UHTM) deals with the design and verification of critical components in the engine compartment to ensure proper cooling and acceptable operating temperatures in a wide range of driving conditions. Although accurate capture of the highly detailed geometry in the engine compartment is not important for UHTM simulations, the full car geometry must be included, since the blockage effects of the engine and other components must be modelled properly. The engine geometry is however, highly detailed, and manual cleanup of the CAD geometry to provide a watertight surface can be very complex and time-consuming [19]. For these reasons, surface wrapping tools are often employed in the automotive industry to wrap engine geometries prior to mesh generation.

This test case therefore demonstrates the use of iconHexMesh in pure ‘wrapping’ mode, to create a watertight surface wrap of a complex engine block geometry. The model, shown in Figure 15(a), comprises 887 separate components grouped into 57 solids defined with a total of 4.2 million triangles. An initial Cartesian mesh was created with an element size of 1.25 m. A uniform surface refinement of level 9 was applied to the engine, giving a smallest element size of 2.44 mm. A wrap level of 5 was applied to fill any holes larger than 39 mm, resulting

in the surface wrap shown in Figure 15(b). The entire wrapping process, including projection of the wrap surface to the geometry and capture of feature lines, took 278 seconds on a Linux workstation using 2 Intel Xeon X5650 (2.67GHz) processors. The final wrapped surface consists of a single closed manifold surface containing 762,028 triangles.

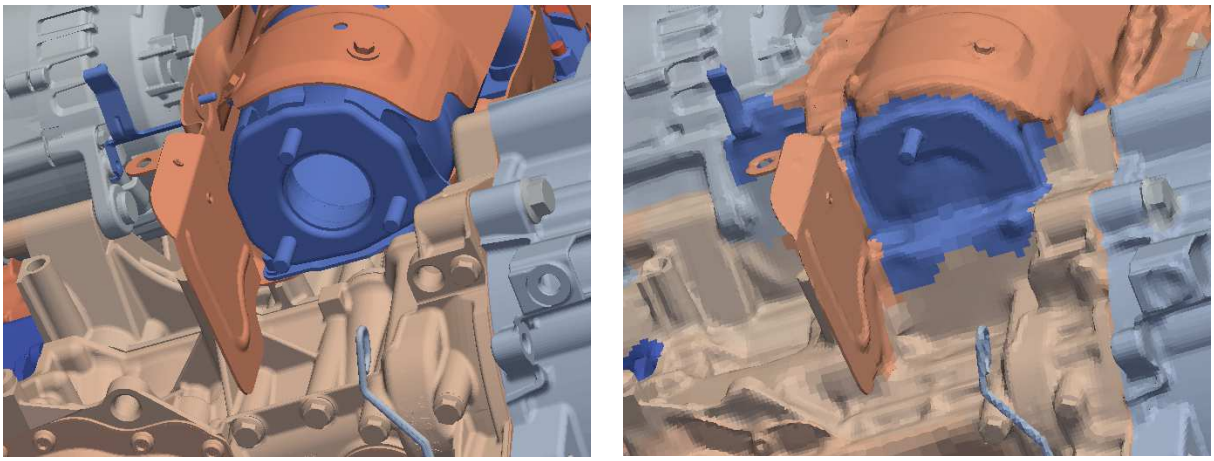


(a) Engine geometry shaded by PID

(b) Watertight surface wrap of engine

Figure 15: Engine geometry and wrap.

In the detailed view of the wrap surface shown in Figure 16, the ability of the wrapping functionality to close large holes in the model, whilst still capturing fine details, is clearly demonstrated. A large circular hole in the geometry can be seen in the centre of Figure 16(a), which the wrap surface closes off in Figure 16(b), whilst still capturing the thin plate and narrow rod structures nearby.



(a) Engine geometry

(b) Surface wrap

Figure 16: Detail of engine geometry and wrap, shaded by PID.

### 4.4 UHTM Case

The final example in this section is used to demonstrate the wrapping capability on a fully complex industrial application. The model is a detailed geometry of the Skoda Fabia II containing the complete geometry of the engine compartment including power-train, suspension, cooling, exhaust system, and electrical components. The geometry consists of 14 STL files, which together comprise 382 patches and 36 million triangles. A clip through the geometry is shown in Figure 17, which illustrates the complex topology of the engine compartment. Manual clean-up of such a geometry and subsequent processing to allow volume mesh generation can involve several man weeks of CAD repair and preparation [17].

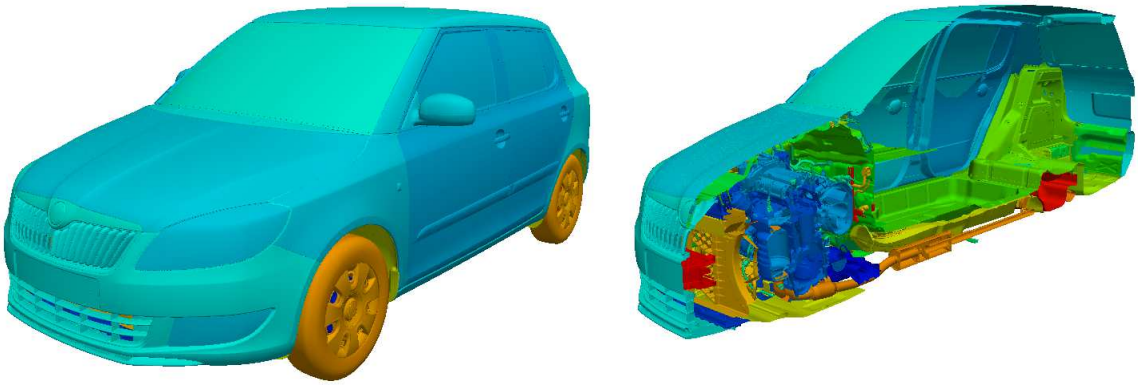


Figure 17: Geometry and clip through geometry of UHTM model

The ‘pure wrapping’ mode was used to generate a watertight surface wrap of this model suitable for subsequent analysis. Firstly, an initial background mesh of element size 0.625 m was generated enclosing the entire car geometry. A wrap level of 4 was applied to the entire geometry to close any holes smaller than 40 mm in the assembly. The background mesh was refined to level 8 (2.44 mm) on all surfaces, and an additional 2 levels of curvature refinement were applied to provide accurate capture of geometry features. The resulting surface wrap is shown in Figure 18. A slice through the background mesh in Figure 19 shows that the wrapping capability has managed to close any gaps which connect the outside of the car to the cabin.



Figure 18: Wrap and clip through wrapped surface of UTHM model

The wrapping process took 2 hours and 42 minutes on 32 cores of a Linux cluster with Intel Xeon E5-2670 (2.60GHz) processors. The final wrapped surface consists of 12.5 million triangles.

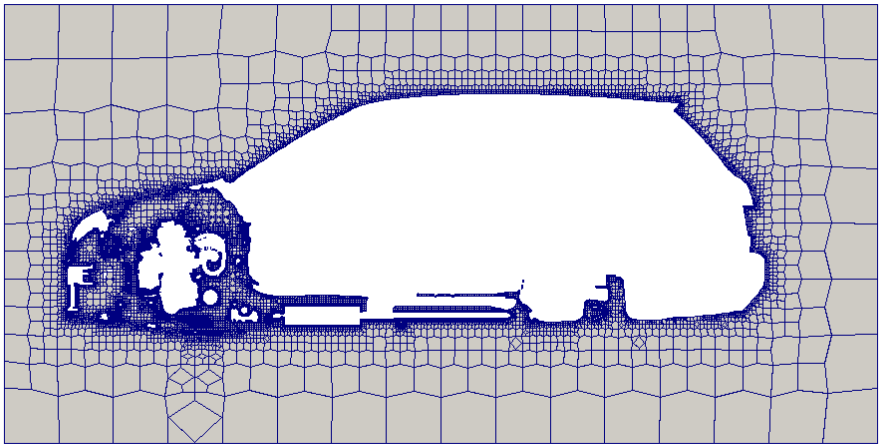


Figure 19: Slice through background mesh on UHTM model

## 5 FUTURE WORK

Although the combined wrapping and meshing approach presented here offers a significant improvement in automation of the simulation process, the user is still required to prescribe suitable refinement levels for the wrapping based on an upper bound on the size of any gaps which need to be filled in the geometry. Determining the size of gaps in the geometry is not a trivial task, and in future it would be preferable to remove this burden from the user. Instead of specifying wrap levels, it is envisaged that the user specifies one or more orphan points, similar to the keep points, which would indicate volumes of the model which are not to be included in the simulation. As the refinement process progresses, as soon as a path is created connecting an orphan point to a keep point, the refinement level at which wrapping is required can be identified.

In order to help keep selected components separate during the wrapping and meshing process, proximity refinement can be employed. However, this is currently applied between the selected surface and any other surface. In future, this will be extended to allow users to specify proximity refinement between selected pairs of surfaces and therefore enable highly focused contact prevention.

The primary focus of this work has been the combined wrapping and meshing functionality within iconCFD. However, as mentioned previously, it is possible to operate the mesher in a pure ‘wrapping’ mode, and output the resulting surface mesh as a modified set of STL files. In this scenario, it may be desirable to further improve the quality of the wrapped surface for the upstream applications. Topological operations such as edge swapping, edge splitting and edge collapsing [20], as well as node smoothing [20] are a few ways in which to achieve better surface mesh quality. If the size of the wrapped STL is a concern, then there are various surface mesh decimation techniques [21] which could be employed to reduce the number of triangles. There is also potential to improve the performance of the pure wrapping mode, by avoiding certain operations which are only required when the quality of the resulting volume mesh is of importance.

## 6 CONCLUSIONS

- A combined wrapping and mesh generation capability has been described which allows the original geometry to be retained to whatever fidelity is required in the simulation. This capability eliminates the de-featuring of geometry associated with conventional surface wrapping techniques.
- The problem of fully-resolved gaps encountered in other interior-to-boundary mesh generation approaches is avoided by wrapping at intermediate refinement levels during the meshing process.
- Implementing wrapping within the framework of an existing parallel mesh generation process means that wrapping can be performed quickly and efficiently with minimal overhead. The wrapping is also able to re-use a substantial amount of software developed in the mesh generator for snapping to geometry lines and surfaces.
- The ability to wrap geometry within the meshing process allows special treatment to be applied to the gap-closure faces, leading to a better quality volume mesh for flow simulation.
- The wrapping capability is designed to handle extremely complex industrial models containing many parts with complex topology. No assumptions are made regarding the quality or orientation of the input geometry, or the shape or nature of any holes present in the assembly. The wrapping is controlled simply by specifying the minimum size of hole which should be closed in the geometry.

## 7 ACKNOWLEDGEMENTS

The authors would like to thank Koenigsegg Automotive AB for their kind permission to use the Agera vehicle geometry, and Skoda for their kind permission to use the Fabia II vehicle geometry.

## REFERENCES

[1] Y. K. Lee, C.K. Lim, H. Ghazialam, H. Vardham, E. Eklund, Surface Mesh Generation for Dirty Geometries by Shrink Wrapping using Cartesian Grid Approach, *Engineering with Computers*, August 2010, Vol 26, Issue 4, pp 377-390.

[2] M.W. Beall, J. Walsh, M. S. Shephard, Accessing CAD Geometry for Mesh Generation, *Proceedings of the 12th International Meshing Roundtable*, pp 33-42, Sept. 2003.

[3] L. Marechal, Advances in Octree-Based All-Hexahedral Mesh Generation: Handling Sharp Features, *Proceedings of the 18th International Meshing Roundtable*, pp 65-84, 2009.

[4] A.A. Demargne, R.O. Evans, P.J. Tiller, W.N. Dawes, Practical and Reliable Mesh Generation for Complex, Real-World Geometries, *Proceedings of the 52nd Aerospace Sciences Meeting*, National Harbour, Maryland, January 2014.

[5] Z. J. Wang, K. Srinivasan, Complex "Dirty" Geometry Handling With An Interior-to-boundary Grid Generation Method, AIAA 2001-2538, June 2001.

[6] Y. Jun, A Piecewise Hole Filling Algorithm In Reverse Engineering, *Computer Aided Design 37*, 263-270, 2005.

[7] W. Zhao, S. Gao, H. Lin, A Robust Hole-Filling Algorithm For Triangular Mesh, *International Journal of Computer Graphics*, Vol. 23, Issue 12, November 2007.

[8] J. Carr, R. Beatson, J. Cherrie, T. Mitchell, W. Fright, B. McCallum, Reconstruction and representation of 3D objects with radial basis functions, *Proceedings of SIGGRAPH*, Los Angeles, CA, 12–17 August 2001, pp. 67–76. ACM Press, New York, 2001.

[9] J. Branch, F. Prieto, P. Boulanger, A Hole-Filling Algorithm for Triangular Meshes using Local Radial Basis Function, *Proceedings of the 15th International Meshing Roundtable*, pp. 411-431. Springer, Heidelburg, 2007.

[10] R.T. Whitaker, A Level-Set Approach to 3D Reconstruction From Range Data, *Int. J. Computer Vision 29*, 203-231, 1998.

[11] H.K. Zhao, S. Osher, R. Fedkiw, Fast Surface Reconstruction Using the Level Set Method, *Proceedings of the IEEE Workshop on Variational and Level Set Methods in Computer Vision (VLSM 2001)*, pp. 194-201, 2001.

[12] F. Juretić, N. Putz, A Surface-Wrapping Algorithm with Hole Detection Based on the Heat Equation, *Proceedings of the 20th International Meshing Roundtable*, pp. 405-418, 2011.

[13] A. Kumar, A.M. Shih, Hybrid Approach for Repair of Geometry with Complex Topology, *Proceedings of the 20th International Meshing Roundtable*, pp. 387-403, 2011.

[14] E.W. Lorenson, E.H. Cline, Marching Cubes: A High Resolution 3D Surface Construction Algorithm, *ACM Computer Graphics 21(4)*, 1987.

[15] D. Martineau, J. Gould, J. Papper, Towards an Efficient Distributed Geometry for Parallel Mesh Generation, Research Note, *22nd International Meshing Roundtable*, 2013.

[16] H.G. Weller, G. Tabor, H. Jasak, C. Fureby, A tensorial approach to computational continuum mechanics using object-oriented techniques, *Computers In Physics, Vol. 12*, No. 6, Nov/Dec 1998.

[17] L.P. Kobbelt, J. Vorsatz, U. Labsik, H. Seidel, A Shrink Wrapping Approach to Remeshing Polygonal Surfaces. *Computer Graphics Forum, Proceedings of Eurographics '99*, 18(3): 119–130, 1999.

[18] Flange case geometry. Retrieved on 15/01/2016 from https://github.com/OpenFOAM/OpenFOAM-3.0.x/tree/master/tutorials/resources/geometry

[19] K. Srinivasan, Z.J. Wang, W. Yuan, R. Sun, Vehicle Thermal Management Simulation using a Rapid Omni-Tree Based Adaptive Cartesian Mesh Generation Methodology, *Proceedings of ASME Heat Transfer/Fluids Engineering Summer Conference*, Charlotte, North Carolina, USA, July 11-15, 2004.

[20] D. Wang, O. Hassan, K. Morgan, N. Weatherill, Enhanced remeshing from STL files with applications to surface grid generation. *Commun. Numer. Meth. Engng.*, 23: 227–239, 2007.

[21] L. Kobbelt, S. Campagna and H. Seidel, A General Framework for Mesh Decimation, *Proceedings of Graphics Interface 1998 Conference*, June 18-20, 1998.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE VORTICITY CREATION PROCESS AT PHYSICAL SURFACES

**Horia Dumitrescu[1] and Vladimir Cardos[2]**

"Gheorghe Mihoc – Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy

Calea 13 Septembrie no. 13, 050711 Bucharest, Romania

[1]e-mail: dumitrescu.horia@yahoo.com

[2]e-mail: v_cardos@yahoo.ca

**Keywords:** Boundary layer, Vorticity waves, Laminar-turbulent transition.

**Abstract.** *The fluids as deformable bodies without own shape, when starting from rest, experience interactions between the flowing fluid and the physical surfaces marking the bounds of flow. These interactions are a kind of impact process where there is a momentum exchange between two colliding bodies, i.e. the flow and its boundary surfaces. Within a short time of contact a post-impact shear flow occurs where two main effects are triggered off by the flow-induced collision: dramatic redistribution of the momentum and the boundary vorticity followed by the shear stress/viscosity change in the microstructure of the fluid which at the beginning behaves as linear reactive medium and latter as nonlinear dispersive medium. The disturbance of the starting flow induces the entanglement of the wall-bounded flow in the form of point-vortices or concentrated vorticity balls whence waves are emitted and propagated through flow field. The paper develops a wave mechanism for the transport of the concentrated boundary vorticity, directly related to the fascinating turbulence phenomenon, using the torsion concept of vorticity filaments associated with the hypothesis of thixotropic/nonlinear viscous fluid.*

## 1 INTRODUCTION

Generally, the impact/collision is a process of momentum exchange between two colliding bodies within a short time of contact. With respect to a single impacted body or structure, the loading in such a process acts with high intensity during this short period of time. As a result, the initial velocity distribution is rapidly changed (even pressure wave loadings). Such rapid loading in the contacting area is a source where waves are emitted that propagates with finite speed through the body. In the case of sufficiently small amplitudes, the linear elastic body waves propagate with the speed of sound waves; the distinction between the fast longitudinal L – wave (carrying wave) and the slower transverse T – wave complicates the pattern as it will be seen in the sequel.

When starting from rest the fluids as deformable bodies without own shape, experience interactions between the flowing fluid and the physical surfaces enclosing the flow. These interactions are a kind of impact process where there is a momentum exchange between the flow and its boundary surfaces. Within a short time of contact a post-impact shear flow occurs where two main effects are triggered off by the flow-induced collision: dramatic redistribution of the momentum and the boundary vorticity creation followed by the shear stress/viscosity change in the microstructure of the fluid which at the beginning behaves as linear reactive medium and latter as nonlinear dispersive medium. The disturbances of the starting flow ($t \to 0$) cause the entanglement of the wall-bounded flow (stream function $\Psi = 0, t \to 0$) inducing a wall torsion pressure (suction) in the form of point-vortices or concentrated vorticity balls whence waves are emitted and propagated through the flow field [1], [2]. Such vorticity concentrations and their potentially devastating rotary motions are observed in the natural case of tornados. The above phenomenology of the impact process of a starting fluid shows that at the beginning the vorticity lines are subjected to a torsion pressure [2] being concentrated at the boundaries; after that the vorticity concentrations are dispersed in a shear layer by vorticity shear waves. Thus, the description of the complicate dynamics of the concentrated boundary vorticity, in fact a kind of variable angular velocity, requests a new fluid medium able to adjust itself continuously and to respond to the flow/stress. Such a fluid playing a role of variable mass is called the thixotropic/nonlinear viscous fluid [3]. In contrast to the Newtonian fluid model, too restrictive (linear shear stress and constant viscosity), the thixotropic model is a flexible one, able to describe even intricate turbulent motions.

The boundary vorticity dynamics, (i.e. its creation, growing and finally fading under the form of the full turbulent flow) is presented in the paper through the results concerning the plate boundary layer flow and Couette flow.

## 2 GENERAL FORMULATION

We consider an infinite region of fluid, which is viscous and incompressible. The Navier-Stokes equations expressed in the vorticity-stream function formulation for two-dimensional motion of this fluid is recasted in the vorticity transport/transfer equation

$$\frac{\partial \omega}{\partial t} + u\frac{\partial \omega}{\partial x} + v\frac{\partial \omega}{\partial y} = \nu\left(\frac{\partial^2 \omega}{\partial x^2} + \frac{\partial^2 \omega}{\partial y^2}\right), \tag{1}$$

where $\omega = \frac{1}{2}\left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)$ is the vorticity at the fluid point (rotation of fluid), t is the time, x is the coordinate in the direction of mean flow and y is the coordinate normal to that direction; x and y velocity components are $u = \frac{\partial \Psi}{\partial y}$ and $v = -\frac{\partial \Psi}{\partial x}$ and $\Psi$ is the stream function.

The vorticity transport equation is classified as a parabolic equation with the unknown being the vorticity ω. The definition of vorticity yields

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} = -\omega, \tag{2}$$

known as the stream function equation, classified as an elliptic partial differential equation.

The unknown is the stream function Ψ, whose ω is provided from the solution of Eq. (1). Once the stream function has been computed, the velocity component may be determined from its derivatives. By introduction of new variables, namely the vorticity and the stream function, the incompressible Navier-Stokes equations are decoupled into one elliptic equation and one parabolic equation which can be solved sequentially. The main drawback of the method consists in the poor knowledge of the initial and boundary conditions that must be obtained by physical considerations. The initial condition at fluid-boundary impact ( $\Psi = 0, t = \frac{x}{U_e} \to 0$ ) given by Stuart's solution for vorticity $\omega_w = e^2$ [1] associated with the thixotropic fluid model [2] renders Eq. 1 at the boundary enclosing the flow (Ψ = 0, t) into a non-linear ordinary differential equation for the boundary vorticity, $\omega_w \left( \Psi = 0, t = \frac{x}{U_e} \right)$

$$\frac{d\omega_w}{dx} = \nu \frac{d^2 \omega_w}{dx^2}, \text{ - vorticity boundary conditions}, \tag{3}$$

$$\omega_w \nu = U_e^2 \text{ - shear compliance relation}, \tag{4}$$

where the ν(*x*) is the variable viscosity of the thixotropic viscous fluid able to respond to flow/stress and Eq. (4) represents a kind of the conservation of flow angular momentum. The initial/starting and boundary conditions (3) for vorticity are a crucial part of the mechanism by which a laminar flow becomes transient and then turbulent one.

## 3 ON PHYSICS OF VORTICITY

Formally, the problem of turbulence is to solve the Navier-Stokes equations subject to initial and boundary conditions. At present, it is possible to obtain fully resolved solutions at moderate Reynolds numbers via direct numerical simulations of the Navier-Stokes equations. However, the common mathematical conjecture that the turbulent flows can be correctly described in an asymptotical manner (i.e. at large Re with $\nu \to 0$) by some sort of specially selected weak solution of Euler equations has little to do with real physics at any large Reynolds number. But, beside the difficulties of a formal/technical nature, there is another difficulty of a more general nature. It is lack of knowledge about the basic physical processes of turbulence and its generation and origin, and poor understanding of the processes which are already known.

Here, in contrast with the common view that the origin of turbulence lies in the instability of some basic laminar flows, the turbulence is the result of the boundary vorticity dynamics of the flow-induced collision and its consequences where shear waves are emitted and propagated in the flow field from the vibrating concentrated vorticity at the boundaries.

This is understood in the sense that any flow starts from rest at some moment in time, and as long as the Reynolds number, or the reduced frequency of vorticity is small, the flow remains laminar (creeping motion of vorticity). As the Reynolds number/frequency of vorticity increases a wide instability range sets in, which is followed by transition and then a

fully developed turbulent state. This is the visible face or large scale of the flow field, i.e. the velocity fluctuations, whereas the poorly known origin of turbulence is the invisible face or small scale of the flow field, i.e. the vorticity fluctuations.

Therefore, the origin of turbulence is a problem of boundary vorticity dynamics that needs a holistic approach of the motion process containing the linked up events: the flow surface colliding and starting vorticity creation, the non-dispersive creeping motion of vorticity (laminar flow) and the dispersive vibrating motion of vorticity (turbulent flow).

Each perturbation, small or large, is in fact a kind of impact between the flow and its boundaries exchanging the momentum between bodies within a short time of contact. As a result, the initial velocity distribution is rapidly skewed/squeezed, the vorticity is created and organizes itself into more and more concentrated structures, thus at the boundary there is a set of point-vortices. The exact solutions of the equations of inviscid motion found by Stuart [1] ($e^{\tau}$ - solutions, $\tau \in [0,2]$) can describe strong vorticity concentrations developing in skewed shear layers. The concentration level of vorticity is estimated on a natural logarithmic scale $e^{\tau}$ from $e^{0}$ - sparse/weak vorticity, up to $e^{2}$ – concentrated vorticity, where the index $\tau$ is a measure of the concentration of vorticity. In contrast to the sparse vorticity transported by the ideal fluid flow according to the laws given by Helmholtz [4], the concentrated vorticities are transported by waves, their shape depending mostly on the concentration of vorticity, Fig. 1.

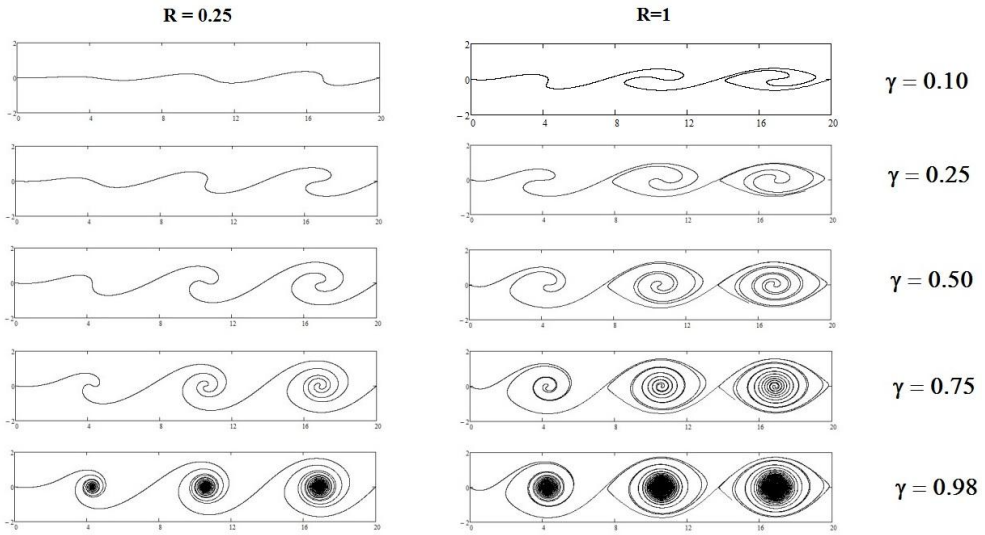



Figure 1: Effect of γ (concentration), R (circulation) parameters on streakline patterns [5].

Physically, the concentration of vorticity at boundaries is a local compression of flow inducing by the torsion of vorticity wires. The concept of torsion of the concentrated vorticity allows a better understanding of the boundary vorticity creation and its dynamics, which becomes an active one just after impact. Essentially, the boundary vorticity dynamics contains small amplitude vibrational motions generating vorticity weak waves that create the covered/hidden field of flow. The self-sustained vorticity wave onset is the mechanical origin of turbulence.

Concomitantly with the vorticity creation the impact process induces microstructure changes of the flow properties resulting in a time dependent shear stress, $\nu = \nu(t)$, known as the thixotropic behavior of the flowing fluid. It is experimentally shown that the transient viscosities follow the line of the complex viscosity versus angular frequency [5]. This behavior can be described by a Klein-Gordon like wave equation [6]

$$\frac{d^2\nu(x)}{dx^2}=\frac{1}{U_e^2}\left(\omega_0^2-\omega^2\right)\nu(x), \tag{5}$$

where $\nu_0$ is steady shear viscosity and $\omega_0=\nu_0^{-1}$ is the natural angular frequency of fluid, defined as the first zero of the Fourier coefficient B(ω) of a square pulse/impact. Thus, the solutions of Eq. 5, describe well enough the dual behavior of the thixotropic fluid, as a reactive medium, $\omega<\omega_0$, at impact inducing exponential waves (without energy dissipation), and as a dispersive medium which can support sinusoidal waves for ω above the natural frequency $\nu_0^{-1}$. Equation (5) also shows that the microstructure takes time to respond to the flow/stress and its elastic response at low frequency is faster as the flow velocity $U_e$ increases.

The disturbed post-impact flow is a boundary-layer type flow which is relaxed through a complicated wave system, which transports concentrated vorticity from boundaries to the flow field and rebuilds the flow microstructure. There is a non-dispersive transport of vorticity performed by exponential waves in the form of the laminar flows dominated by the frictional shear stress and a dispersive one which involves lightly damped sinusoidal waves by dry friction in turbulent flows. Hence, it is evident that the analysis of the impact-relaxation process requests another constitutive relation to describe the intricate behavior of viscous fluid. For the thixotropic fluid, such a relationship is a shear compliance defined as

$$\frac{1}{\rho}p_{torsion,w}\equiv\omega_w\nu=U_e^2 \text{ on } \partial\mathrm{B}, \tag{6}$$

where $p_{torsion,w}$ is the torsion pressure at wall, $\omega_w=e^{\tau}$ is the vorticity at a two-dimensional wall ($\partial$B), τ is a torsion/concentration index $\tau\in[0,2]$, and ν(*t*) denotes the change of viscosity during the post-impact flow which is able to adjust itself continuously.

A non-steady fluid system involves an oscillating behavior of its opposite, intrinsic properties (vorticity and viscosity) and suddenly excited it decays as a big damped harmonic oscillator. The evolution is slow and to visualize its full way a plotting on suited scales is necessary. Using the exponential scales and measure units *e* and $\nu_0^{-1}$, the shear compliance, Eq. 6, can be written as

$$\begin{aligned} &e^{\tau}\left(\nu_0^{-1}\right)^{1/\tau}=\mathrm{Re}_x \text{ for } \mathrm{Re}_x<\mathrm{Re}_c \text{ (laminar flow)}\\ &e^{1/\tau}\left(\nu_0^{-1}\right)^{\tau}=\mathrm{Re}_x \text{ for } \mathrm{Re}_x\geq\mathrm{Re}_c \text{ (turbulent flow)} \end{aligned}, \tag{7}$$

where the critic Reynolds number $\mathrm{Re}_c=e^2\nu_0^{-1}$ is the non-rolling condition for concentrated vorticity, which separates the non-periodic creeping motion of vorticity inducing laminar flow, from the torsional vibration motion of vorticity generating turbulent flow, and $\nu_0^{-1}$ is the natural frequency of the thixotropic fluid. Equations (7) express the conservation of boundary vorticity in laminar flow, its dispersion in turbulent flow, respectively.

For the above $\mathrm{Re}_x=\nu_0^{-1}$ the transient flow in the neighborhood of the wall vibrates as a continuous and homogeneous string carrying transverse vorticity waves which permanently disperse vorticity. Figure 2 illustrates the wave system induced by the flow-boundary impact and the dispersion mechanism of concentrated vorticity that displays a wide frequency/Reynolds number spectrum from the low indifference Reynolds number $\mathrm{Re}_{lind}=e^2\nu_0^{-1/2}$ - the onset of the weakest waves (TSW-Tolmien-Schlichting waves) up to the high indifference

Reynolds number $\mathrm{Re}_{hind} = e^{1/2}\nu_0^{-2}$. For the last Re the wave system becomes a slightly damped one with the resonance close to the natural frequency $\nu_0^{-1}$, emitting acoustic waves. For above $\mathrm{Re}_x = e^2\nu_0^{-1}$ the transitional flow displays a strong beat phenomenon with varicose aspect of vorticity where its frequency $\mathrm{Re}_x^{1/2}$, is far from the resonance frequency. For above $\mathrm{Re}_x = e^{2/3}\nu_0^{-3/2}$, the vorticity is broken down in contra-rotating fragments/flocks and its frequency approaches the resonance frequency where the flow is full turbulent and it can be statistically described. The Reynolds number controls the wave system playing a role of tuning button that switches the frequency band from concentrated vorticity and small Re/low frequency and long wavelength to dispersed vorticity and high Re/high frequency and short wavelength.

The essential difference between laminar flow and turbulent flow is given by the difference between the behaviors of fluid as linear viscoelastic-reactive medium and nonlinear thixotropic-dispersive medium. That is, while both are time effects, the former is in the linear region, where the microstructure responds but remains unchanged and the latter takes place in the nonlinear region where the microstructure is broken down by deformation as well as responding to it.

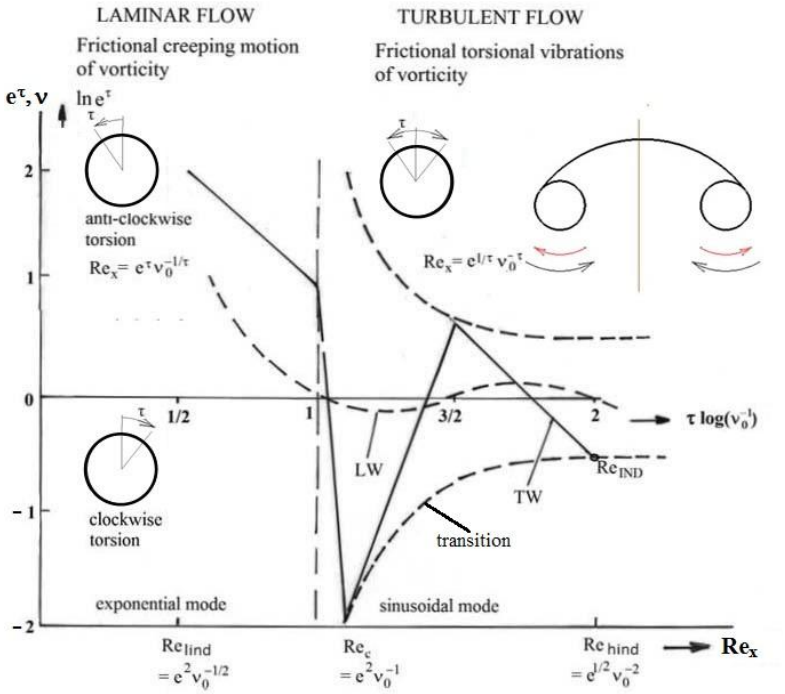



Figure 2: Transverse waves (TW) modulated in amplitude on a longitudinal carrying wave (LW) and dispersion mechanism of concentrated vorticity by shear waves [8].

## 4 RESULTS FOR ZERO-PRESSURE GRADIENT FLOW

The zero-pressure gradient flows are simple shear flows (Prandtl boundary-layer flow and Couette flow) where the shear stress has a prevailing role everywhere in the flow field. The reason to treat these flows is because of fundamental importance for turbulent flows close to walls in general, far beyond these particular cases. It will be seen that the flow regions of the chosen turbulent flows close to wall have universal importance, so that, up to some yet to be specified conditions, the results can be carried over to the regions close to the wall of general turbulent flows. That is, the Prandtl and Couette flows are the extreme examples concerning

the visual perception of turbulence: one sidewall-bounded flows (boundary-layers flows) with visible velocity fluctuations, and two sidewall-bounded flows (shear flows) involving invisible velocity fluctuations. As a matter of fact, the Couette flow is also called no-fluctuation shear flow.

In this context, the motion equations contain an irrotational potential $\nabla\varphi = U_\infty x$ (outside a shear layer) and a vector potential $\mathbf{A} = \psi(x,y)\mathbf{k}$ (inside and normal to shear layer) where $\psi$ is the stream function defined by

$$u = \frac{\partial\psi}{\partial y}, v = -\frac{\partial\psi}{\partial x}, \tag{8}$$

Using the approximations and notations of the boundary-layer flow [7]

$$\psi = \sqrt{2\nu x U_\infty}\, f(\eta), \delta(x) \approx \left(\frac{x\nu}{U_\infty}\right)^{1/2}, \eta = \frac{y}{\delta(x)}, \tag{9}$$

where $f(\eta)$ is the dimensionless stream function, $\delta(x)$ is a scaled measure of the boundary-layer thickness (up to the approximation $u = 0.99U_\infty$) and $\frac{u}{U_\infty} = f'(\eta)$ is the similarity law of the velocity profile, the boundary layer equations and their boundary equations become one ODE for the stream equation

$$\text{Prandtl flow}\begin{cases} k_w f''' + ff'' = 0 \\ \eta = 0: f = 0, f' = 0 \\ \eta \to \infty: f' = 1 \end{cases} \qquad \text{Couette flow}\begin{cases} k_w f''' + ff'' = 0 \\ \eta \to +\infty: f' = 1 \\ \eta = 0: f = 0 \\ \eta \to -\infty: f' = 0 \end{cases}. \tag{10}$$

This nonlinear third order equation and the three boundary conditions completely determine its solution, if the mute constant $k_w$ can be known a priori. Blasius found a solution of Eq. 10 for $k_w = 2$, known as Blasius equation, in the form of a series expansion for $\eta \to 0$ and an asymptotic expansion for $\eta \to \infty$, the two forms being matched at a suitable value of $\eta$. But, Eq. 10 describes a more general phenomenon, that of the transverse standing vorticity/shear waves, called solitons, which retain their identity upon a collision.

The vorticity soliton identified here as Blasius soliton, depends on the function $k_w$ directly related to $\mathrm{Re}_x$ via the shear compliance, Eq. 7 , which is responsible for coupling between the autonomous fast motion of vorticity and the velocity field of the non-autonomous slower flow,

$$\begin{aligned} &k_w = e^{\tau} \text{ for } \mathrm{Re}_x < \mathrm{Re}_c \text{ (weak coupling)} \\ &k_w = (\log \mathrm{Re}_x)^{-1} \text{ for } \mathrm{Re}_x \geq \mathrm{Re}_c \text{ (strong coupling)} \end{aligned}. \tag{11}$$

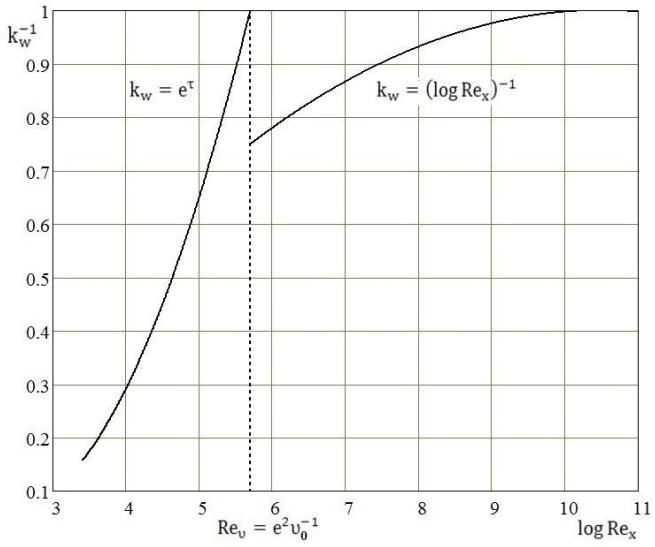



Figure 3: Coupling function $k_w\left(\mathrm{Re}_x\right)$ [8].

Figure 3 shows the variation of the coupling function $k_w\left(\mathrm{Re}_x\right)$, that as a matter of fact is a vorticity creation boundary condition, with a physical support, for any wall-bounded flow. This natural boundary condition can be interested in the numerical methods using vorticity-stream function/velocity formulations [8]. Now, justly $\mathrm{Re}_x$ known the solutions of Blasius soliton can be easy computed by a standard shooting technique.

Figure 4 shows the solution for the vorticity waves close to the ends of the Reynolds spectrum, i.e. flow field at small scale for Prandtl flow. In contrast to the strong shock/pressure waves, the weaker vorticity waves propagate under the form of the three wave packet: shear stress wave *f''*, elastic wave *f'''* and dispersion wave $f^{iv}$. The vorticity wave packet is a superimposing of three waves, each wave having different roles depending on the Reynolds number (the flow type), that is while the laminar flows are dominated by the shear stress waves induced by the creeping motion of vorticity (without energy dissipation), in turbulent flows the dispersive sinusoidal waves, induced by torsional vibrations of vorticity, are the key mechanism of the turbulence phenomenon.

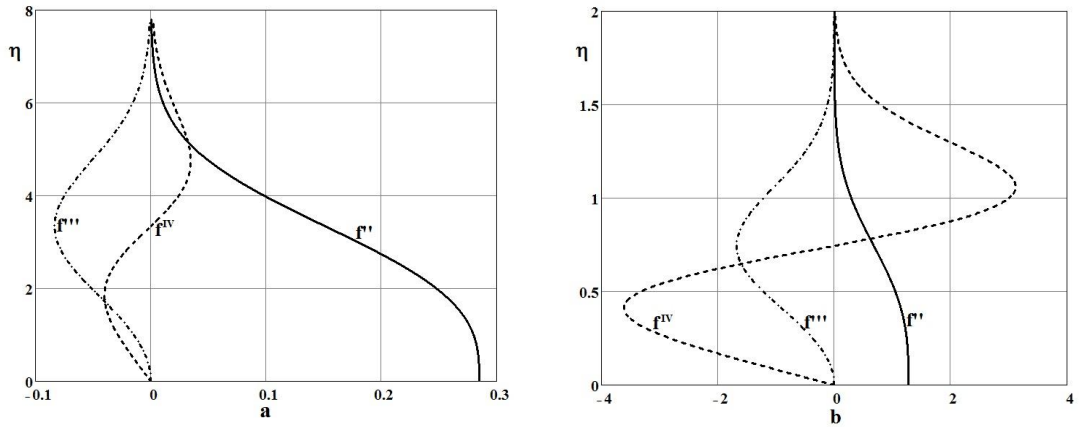



Figure 4: Blasius soliton for T – waves, (Prandtl flow): *f''* - shear wave, *f'''* - elastic wave, $f^{iv}$ –dispersion wave for a) $k_w = e$ and b) $k_w = e^{-2}$.

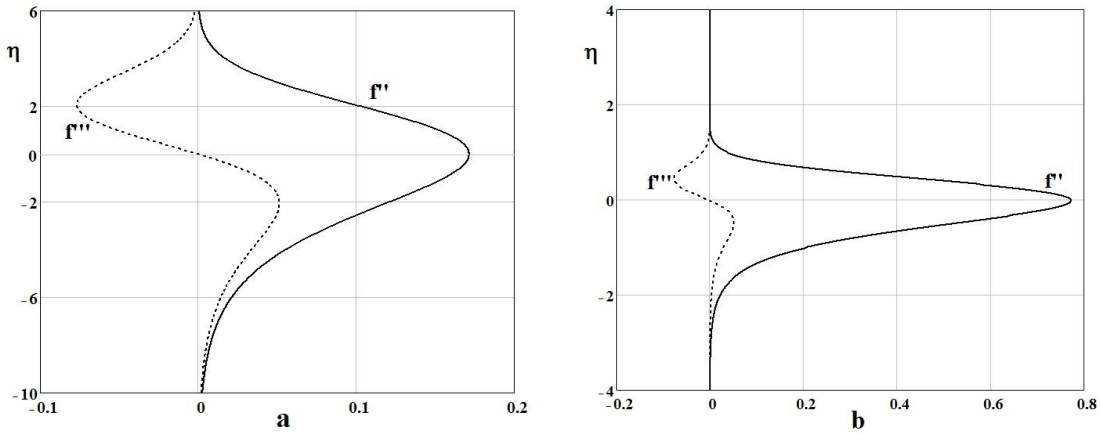



Figure 5: Blasius soliton for T – waves, (Couette flow): *f''* - shear wave, *f'''* - elastic wave, for a) $k_w = e$ and b) $k_w = e^{-2}$.

Figure 5 shows the vorticity field for Couette flow, where the wave packet contains only shear stress wave (f") and elastic wave (f"') because of the stronger restriction on the flow bounded by two sidewalls. The suppression of the dispersion wave leads to a pressure increase in the flow field as the Reynolds number increases. At $\mathrm{Re} = 2\mathrm{Re}_c^{1/2} \simeq 1400$ the concentrated vorticity ($e^2$) is broken down in two contra-rotating fragments, the wall torsion pressure is recovered in the flow field and the velocity field is regularized, Fig. 6b2. The onset of turbulence is the vorticity jump itself where the vorticity (its concentration) is halved and the energy storage at impact as torsion pressure (suction) is recovered as dynamic pressure of flow, Fig. 6.

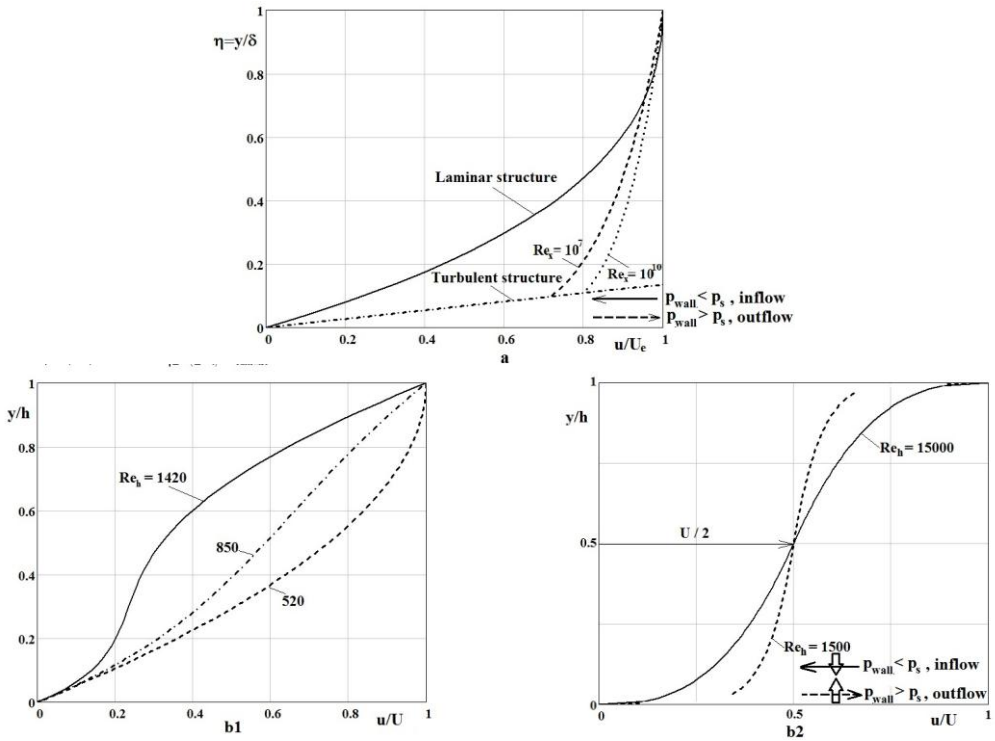



Figure 6: Average velocity profiles and dynamic features of a wall-bounded flow : a) Prandtl boundary layer flow: ——— laminar, – – – – – turbulent two layer model ; b) Couette flow, b1: laminar, b2: turbulent.

Figures 6(a, b) show the flow field at large scale marking also the vibration tendencies of flow near wall when the Reynolds number exceeds its critic value. The analysis of the wall-bounded flow both at small scale (vorticity field) and large scale (velocity field) points out the self-sustained wave mechanism of turbulence, that is similar to synthetic jets generated by pulsed jets with zero net-mass flux and a small flow-momentum consumption. The thixotropic fluid in the turbulent wall-bounded flows operates as a diaphragm that alternatively sucks inflow and ejects outflow in a periodic manner, creating discrete/concentrated vortical structures followed by their dispersion by wall friction and transport by flow.

## 5 CONCLUSIONS

Using the concept of torsion of the concentrated vorticity associated with the hypothesis of thixotropic/nonlinear viscous fluid a vorticity wave mechanism is devised for dispersion/transport of concentrated boundary vorticity. The main results obtained from the investigation of the vorticity wave system, triggered off by the flow-boundary collision, can be summarized as follows:

- The boundary vorticity creation is result of the flowing fluid - boundary collision that is a weak viscous shock of its evolution is governed by solitary shear waves called solitons, called in the paper Blasius solitons, analogous to KdV (Korteweg-de Vries) solitons formed in shallow water.
- There is a wide frequency/Reynolds spectrum from $\mathrm{Re}_x = e^2 \nu_0^{-1/2}$ (TSW) up to $\mathrm{Re}_x = e^{1/2} \nu_0^{-2}$ (acoustic waves) where the wave system operates like a big slightly damped oscillator close to the natural frequency of the fluid.
- The interested region in research lies in the audible frequency range, i.e. below $\mathrm{Re}_x = e^{2/3} \nu_0^{-3/2}$ where the breaking down process of concentrated vorticity by torsion vibrations remains a black box of turbulence.
- The origin of turbulence is one of mechanical nature where vibrations of concentrated boundary vorticity (transverse vorticity waves) are induced by the slightly damped oscillations (longitudinal compressing/expanding waves) of the fluid excited at the beginning of motion. The self-sustaining mechanism of fluctuations in turbulent shear flows consisting of creation, dispersion and transport of vorticity is similar to a typical device generating synthetic jets.
- The dual concept of the torsional concentrated vorticity – thixotropic fluid is vital for any paradigm in research of turbulence.

## ACKNOWLEDGEMENT

This work was realized through the Partnership programme in priority domains – PN II, developed from ANCS CNDI – UEFISCDI, project no. PN-II-PT-PCCA-2011-32-1670.

## REFERENCES

[1] T. J. Stuart, On finite amplitude oscillations in laminar mixing layers. *Journal of Fluid Mechanics*, **29**, part 3, 417-440, 1967.

[2] H. Dumitrescu, V. Cardos, I. Malael, Boundary vorticity dynamics at very large Reynolds numbers. *INCAS Bulletin* **7**, 3, 89-100, 2015.

[3] K. T. Nijenhuis, G. H. McKinley, S. Spiegelberg, H. A. Barnes, N. Aksel, L. Heymann, J. A. Odell, *Thixotropy, Rheopexy, Yield Stress*, In Springer Handbook of Experimental Fluid Mechanics, Eds. C. Tropea, A. L. Yarin, J. L. Foss, Part C, Cp. 9.2, 661-679, Springer Verlag Berlin Heidelberg, 2007.

[4] H. von Helmholtz, Über Integrale der Hydrodynamischen Gleichungen, welche den Wirbelbewegungen entsoprechen. *J. Reine Angew. Math*., **55**, 25-55, 1858.

[5] H. Dumitrescu, V. Cardos, Vorticity waves in shear flows. *INCAS Bulletin* **7**, 1, 51-56, 2015.

[6] F. S. Crawford, Waves, *Berkeley Physics Course*, Vol. 3, Education Development Center, Inc., Newton, Massachusetts, (1968).

[7] H. Schlichting, *Boundary-Layer Theory*, McGraw-Hill Book Company, New York, 1968.

[8] H. Dumitrescu, V. Cardos, On the paradigmatic crisis of turbulence. *Physics of Fluids*, (submitted to publication) (2015).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# HEAT RECOVERY BY CROSS FLOW

**K. Adámek[1], J. Kolář[1], and P.Peukert[1]**

[1] VÚTS Liberec
Svárovská 619, CZ-46001 Liberec
{karel.adamek, jan.kolar, pavel.peukert}@vuts.cz

**Keywords:** Heat recovery, cross flow, numerical flow simulation.

**Abstract.** *Important problem of residential buildings, social centers etc. is the increased concentration of pollutants during use of the building / room.*
*The air-tight outer housing is typical for new or reconstructed buildings, due to tight windows, outer thermal insulation, etc. It is good for the reduced heat consumption in winter period but there are some disadvantages, in general, too, known in general as the "sick building syndrome". For instance, the higher air humidity could result in water vapor condensation in the colder outer corners of the observed room and consequently in mould growth, or in assembling rooms (classrooms, etc.) with poor ventilation, the concentration of $CO_2$ is increasing quickly over the hygienic limits and is the reason of fatigue, headache etc., or due to internal heat sources, the room temperature is increasing and it is necessary to keep it on a suitable level. Many other pollutants can be present in observed buildings, as for instance formaldehyde, radon, bacteria etc., but it is over the extent of this short paper.*
*In standard rooms without ventilation or air conditioning, the natural or forced ventilation is used, only. Results of numerical simulations describe, how to simply remedy the so-called sick building syndrome and without any expenses to get and keep healthy living environment.*
*The design of air conditioning units, taking into account all such pollutant sources, is well-known. Many producers offer many different units for various kind of use. In general, the costs are significant – not only investment, but operational, too. In reconstructed buildings is not space enough for such equipment.*
*In each system of ventilation / air exchange is a part of polluted inner air replaced by fresh outer one. The exhausted inner warm air contains some heat, which must be in winter period added again into the fresh and cold outer air. To reduce the extent of such thermal energy, various heat recovery systems are used.*
*The paper deals with numerical simulations of flow and heat exchange between warm and cold air, using heat exchanger with crosswise flows. It contains two systems of crosswise oriented channels, in the first one the outlet warm air is cooled by inlet cold air, situated in the second channel. The aim is to get the maximum of recovery heat to reduce costs for reheat of outer air, at minimum pressure losses during air flow.*

## 1 INTRODUCTION

Important problem of residential buildings, social centers etc. is the increased concentration of pollutants during use of the building / room.

The air-tight outer housing is typical for new or reconstructed buildings, due to tight windows, outer thermal insulation, etc. It is good for the reduced heat consumption in winter period but there are some disadvantages, in general, too, known in general as the "sick building syndrome" [1]. For instance, the higher air humidity could result in water vapor condensation in the colder outer corners of the observed room and consequently in mould growth, or in assembling rooms (classrooms, etc.) without ventilation, the concentration of $CO_2$ is increasing quickly over the hygienic limits and is the reason of fatigue, headache etc., or due to internal heat sources, the room temperature is increasing and it is necessary to keep it on a suitable level [2]. Many other pollutants can be present in observed buildings, as for instance formaldehyde, radon, bacteria etc. [3], [4], but it is over the extent of this short paper.

In standard rooms without ventilation or air conditioning, the natural or forced ventilation is used, only. Results of numerical simulations [5] describe, how to simply remedy the so-called sick building syndrome and without any expenses to get and keep healthy living environment. Similar data about problems of ventilation are presented for instance in [6] to [11].

The design of air conditioning units, taking into account all such pollutant sources, is well-known. Many producers offer many different units for various kind of use. In general, the costs are significant – not only investment, but operational, too. In reconstructed buildings is not space enough for such equipment.

In each system of ventilation / air exchange is a part of polluted inner air replaced by fresh outer one. The exhausted inner warm air contains some heat, which must be in winter period added again into the fresh and cold outer air. To reduce the extent of such thermal energy, various heat recovery systems are used.

The paper deals with numerical simulations of flow and heat exchange between warm and cold air, using heat exchanger with crosswise flows. It contains two systems of crosswise oriented channels, in the first one the outlet warm air is cooled by inlet cold air, situated in the second channel. The aim is to get the maximum of recovery heat to reduce costs for reheat of outer air, at minimum pressure losses during air flow.

## 2 MODEL OF ONE CHANNEL FLOW

The simple introductory model simulates the influence of various kinds of inner ribs. It is clear that ribs increase the heat transfer surface in general, but in the same time the pressure losses are increasing, too. For comparison of various designs of ribs the simple model of one channel, only, is created, flowed by warm air (as an example). The warm air is cooled by outside cold walls, defined as constant cold temperature. Really, on outside surfaces there is flowing the second cold air, warming along the heat transfer surface, and then its temperature is not constant.

This model is idealization, only, for simple and quick evaluation of various designs.

Some results are presented here for several kinds of triangular ribs of various pitches in the same channel height. The Figure 1 shows the pressure field in inlets cross-sections for three simultaneously solved cases (signed as 7-8-9, where 3-6-9 ribs are used on the same distance).



Figure 1: Pressure field in inlet cross sections (cases signed 7-8-9)

On the Figure 2 to Figure 4 there are presented the fields of pressure, temperature and velocity, all in the middle height of observed channels. Used example is for warm air at the inlet. In narrow channels there is faster temperature change, comparing with large ones. Probably, for specified channel length and temperatures difference in the recuperation, in the narrows channel the maximum possible temperature gradient was reached. It should be to check this case for higher specified temperature difference.

In narrow channels the pressure decreasing in the flow direction is slower – the flow resistance in narrow channels is higher in general, the pressure decreasing is not so fast.

The velocity in narrow channels is smaller in general, due to higher flow resistance.

From realized numerical simulations all necessary flow parameters were evaluated and the relation between thermal outputs received by recovery is determined as

$$Q = m \,.\, cp \,.\, \Delta T \; [W]$$

and mechanical input, necessary for overcoming of flow resistance, is

$$P = V \,.\, \Delta p \; [W].$$

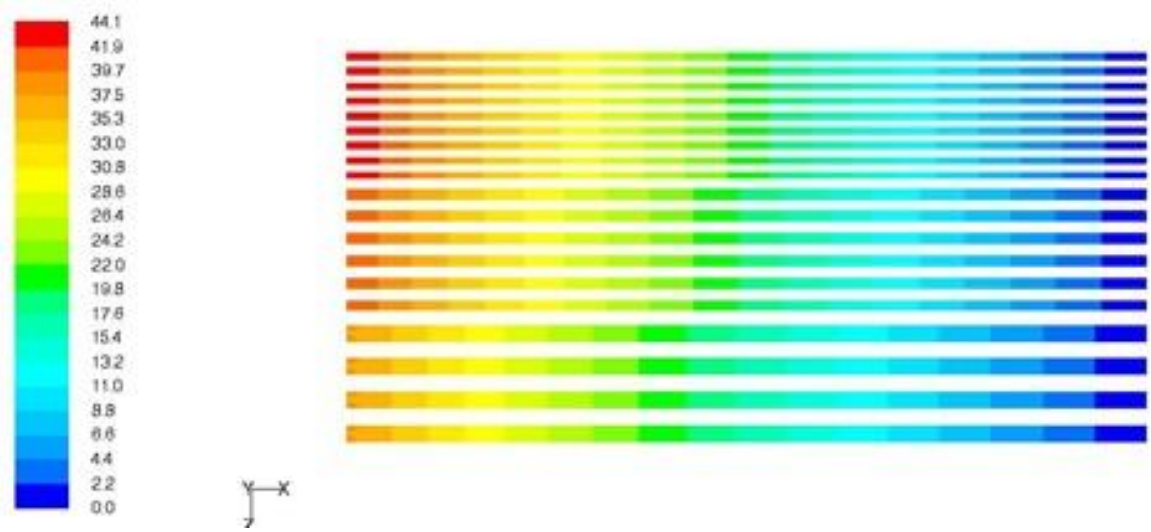



Figure 2: Pressure field

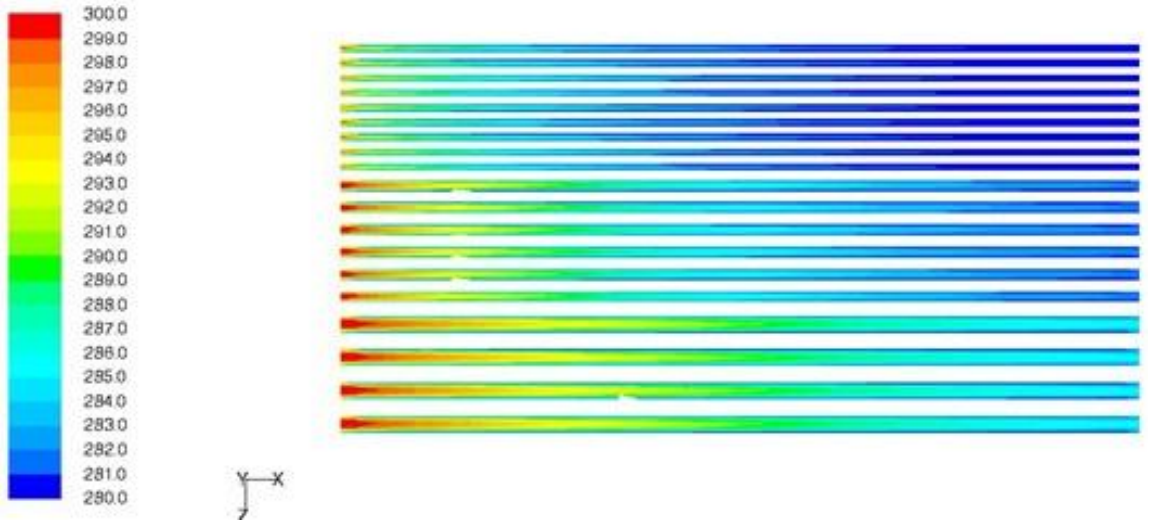



Figure 3: Temperature field

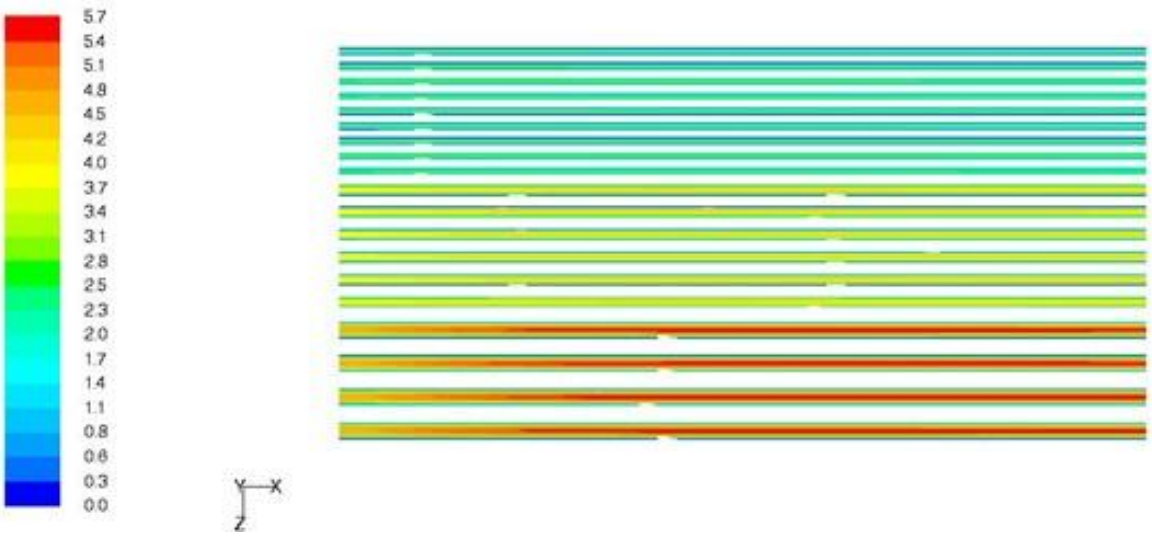



Figure 4: Velocity field

Received values for simulated cases of different ribs arrangements are in the Table 1, as the first information, only, because channel dimensions and their arrangement are not identical with reality.

| case | Q/P | |
|---|---|---|
| **1** | **435.6** | **100%** |
| 2 | 456.3 | 105% |
| 3 | 477.6 | 110% |
| 4 | 481.9 | 111% |
| 5 | 468.6 | 108% |
| 6 | 471.1 | 108% |
| 7 | 490.7 | 113% |
| 8 | 496.4 | 114% |
| 9 | 528.5 | 121% |

Table 1: Effectivity of various shapes of ribs – summary

The cases 1-2-3 contain 1-2-3 channels. The effectivity is increasing with channel density (i.e. with decreased channel pitch), cases (1-2-3-7-8-9). The influence of various detailed modifications of the case 3 (cases 4-5-6) on the value of effectivity is practically none. The greater heat transfer surface influences markedly the greater thermal output. The influence of greater flow resistance due to the higher flow velocity in smaller cross-section of channel is not so important.

Probably, with decreased ribs pitch (increased ribs density) the recuperation effectivity is increasing theoretically. Practical restriction is given by production feasibility – which slimness of ribs (their density and height) is possible to produce using the actual technology.

## 3 MODEL OF TWO CROSS FLOWS

The channel modifications above, tested on single channel, only, are verified on the more complicated model of cross flow of both warm and cold flows. The model contains two flat channels of the crosswise orientation of flows, equipped by ribs, arranged after the previous paragraph. The heat transfer is realized on the common surface between them, while the outer

surfaces are defined as thermaly insulated. Really, on outside surfaces there are connected next crosswise channels, the modeled situation is really repeated in each pair of neighboring layers.

Real heat exchanger contains many pairs of such channels of warm and cold air, to get an adequate total air flow. Simulation of such real design is not possible due to large extent – large outer dimensions, many fine details of ribs etc.

Temperatures are specified after [12], warm air (25°C, dew point 5°C) and cold air (5°C, dew point 1°C), i.e. without humidity condensation.

Features of the model:

1).The simulated temperature distribution is the same as in the reality. The heat exchange is realized between two layers of warm and cold air. On the outer surfaces there is defined perfect thermal insulation – really there is next heat transfer into next air layer, but for the evaluation of designed geometry of heat transfer surface here is used such simplification.

2) The simulated flow or velocity distribution is real as in the reality. The complete exchanger consists from many pairs of such channels, in each of them some partial air volume is flowing and the flow resistance is the same in each layer.

From results of simulations we can state that the transferred heat output ($m.cp.\Delta T$) is much higher than the mechanical input consumed by flow resistance ($V.\Delta p$). Using more complicated shape or higher number of ribs, the gain of thermal output from heat recovery is probably more important than the increased consumption of mechanical energy due to higher flow resistance in narrow channels among ribs.

As an information here is presented one of possible solution procedure and received results. As mentioned above, due to the task extent the heat transfer between two adjacent layers of warm and cold air is solved, only, the outer surfaces are thermally insulated (really there are next heat transfer surfaces here). This modeled situation is repeated in each pair of adjacent layers.

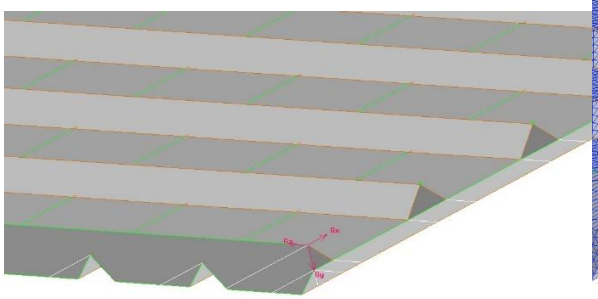


Figure 5: Geometry of 1 layer

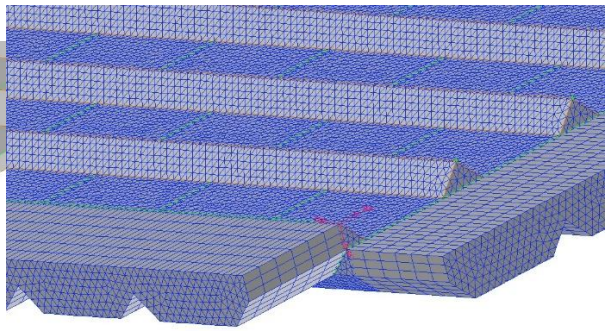


Figure 6: Mesh of 2 layers

**Geometry**

Created after the Figure 5, the model contains two layers situated across. Along outlines some inlet and outlet channels are added to suppress an influence of possible backflows on the total mass balance. In the model should be removed all details, necessary for production (rounding, foil thickness etc.), because such details extremely complicate the model creation and solution, too. The foil thickness is equal zero in the geometry; the real value is defined later, together with other material specifications.

By laying of individual surfaces in crosswise manner, inside arise complicated shapes of individual channels, with high turbulence of the flow, which increases the heat transfer coefficient of observed case.

**Mesh**

Created from tetrahedrons of 1.5 mm, see the Figure 6 - together 5.6 millions of elements, over 0.5 GB of data respectively. Regarding fine geometric details and heat exchange, the mesh should be even finer, but restrictions are given by the extent of actual operational memory of standard PC and the time of solution.

**Boundary and initial conditions**

After [12] the cold air temperature is defined +5°C (278 K) for lower body and warm air temperature +25°C (298 K) for upper body.

Defined pressure gradient of 100 Pa is identical for both bodies. Thus in the model of certain geometry arises flow of certain velocity, here in the specified geometry 1.4 – 1.5 m/s approx. For several specified pressures several operational points are simulated for prescribed velocity range 1 to 4 m/s approx. and so-called operational characteristics $V = f(\Delta p)$ is evaluated. From the simulation the characteristics of thermal input, mechanical output, thermal efficiency etc. are evaluated, too.

**Results**

As an illustration, only, there are presented some basic figures of the flow field for one solved case, only, see the Figure 7 to Figure 12.

Values in individual points of observed surface are quite different, so it should be to remind here, that the standard deviation of evaluated data is quite high. Nevertheless, the errors of energy balances for heat recovery and flow resistance, using averaged values, are not important.

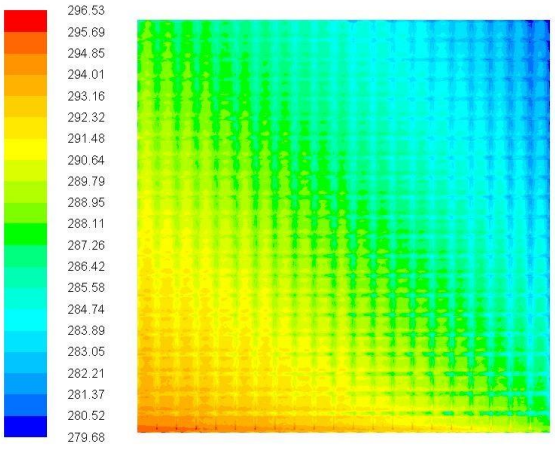



Figure 7: Temperature on the middle (heat transfer) surface – warm air from down, cold air from right

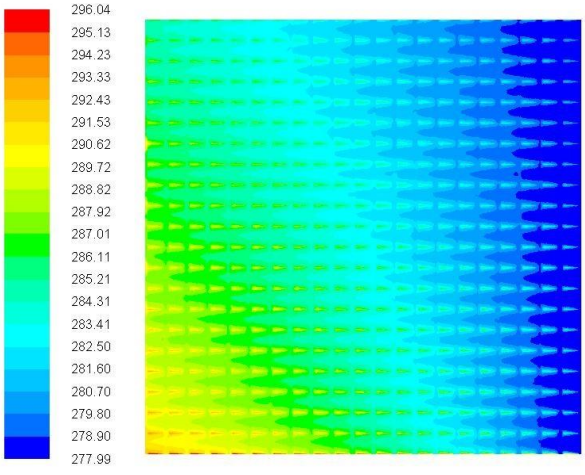



Figure 8: Temperature on the lower (insulated) surface – cold air inlet from right

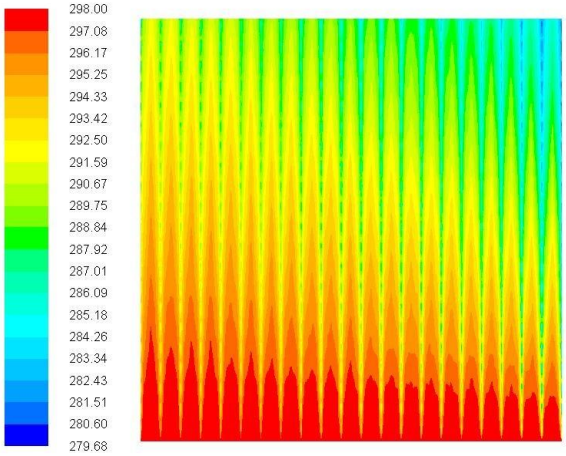



Figure 9: Temperature on upper (insulated) surface – warm air inlet from down

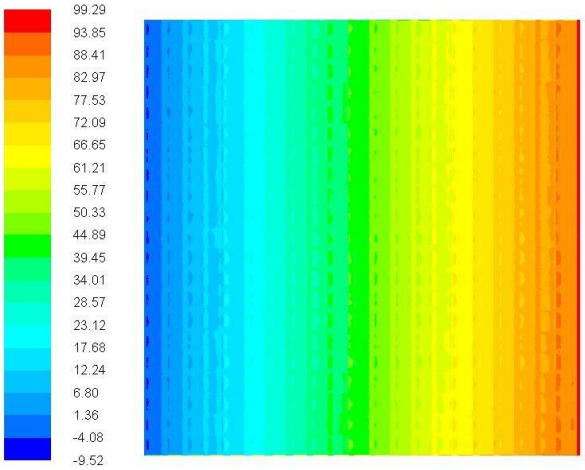



Figure 10: Pressure field in the lower layer (inlet of 100 Pa from right)

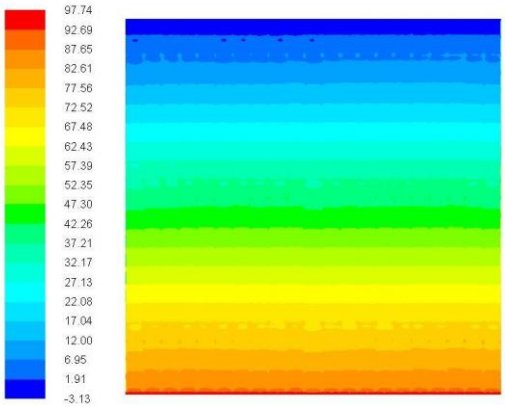



Figure 11: Pressure field in the upper layer (inlet of 100 Pa from down)

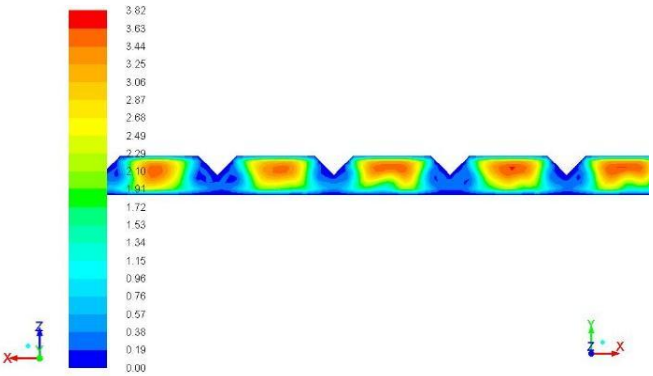



Figure 12: Velocity in the outlet cross section

## 3.1 Rectangular channels

Above mentioned model contains ribs of complicated shape, which means the complicated modeling and simulation, too. The advantage of the model with rectangular elementary channels is the smaller extent.

To get the operational characteristics V = f(p) of the model, it is necessary to solve 4-5 different pressure gradients for each geometry. For channel cross sections of 6x6, 6x9, 6x12, 6x18 and 6x36 mm the sum of 20-25 solved cases is necessary. For instance the model of channels with cross section 6x6 mm contains 0.9 millions of elements sized 2 mm, and then the total time of solution for all solved cases will be shortened significantly.

For more detailed temperature and velocity fields it would be better to use finer mesh, but it needs longer time of solution. But higher number of elements in cross section of channels is here on the limit of applicability. Boundary conditions are identical with previous model.

Using the same pressure gradient as in the previous paragraph, the thermal output is decreasing here and the negligible mechanical input is increasing a little. As an illustration the distribution of some flow field parameters is presented on Figure 13 to Figure 15.

Eliminating some partitions (i.e. model 6x36) the pressure losses are decreasing and the flow rate is increasing.

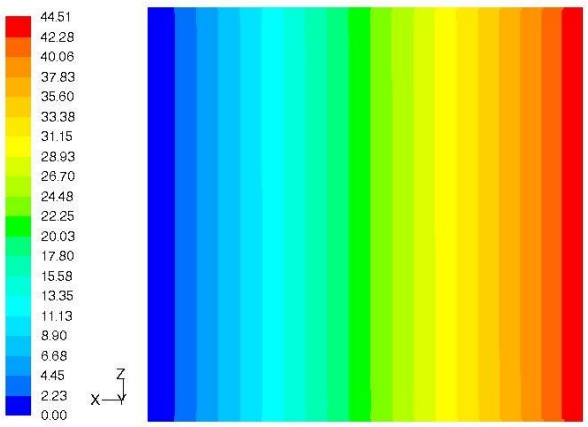



Figure 13: Pressure field of air

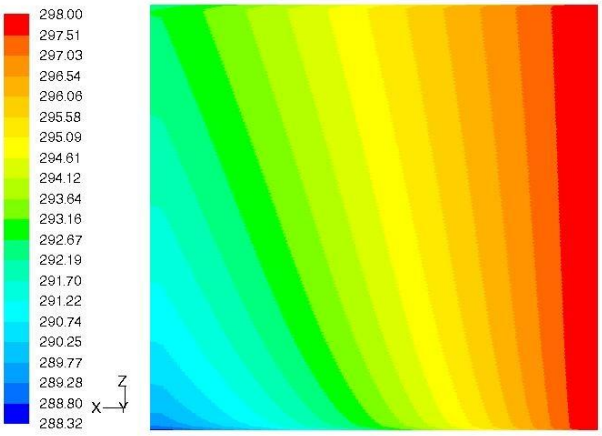



Figure 14: Temperature field of air (flow direction from the right to the left)

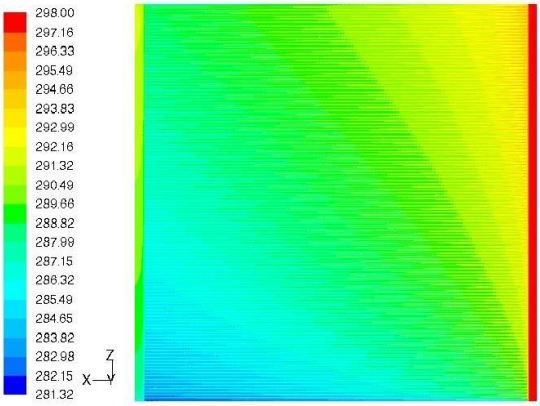



Figure 15: Temperature field on the middle surface (warm air from right, cold air from down)

## 3.2 Influence of wall thickness

Rectangular channels have simple geometry and maybe here should be used cellular boards known as Makrolon. By assembling of such board there arises double thickness of heat transfer wall between adjacent channels. The influence of wall thickness on the heat transfer coefficient is summarized in the Table 2, using well-known formula for heat passage through simple flat wall

$$1/k = 1/\alpha_1 + s/\lambda + 1/\alpha_2.$$

The basic case 1 contains plastic ribs of thermal conductivity $\lambda = 0.2$ W/(m.K) and thickness 0.2 mm. The used heat transfer coefficients between air and wall is preliminary specified as $\alpha = 20$ or 40 W/(m$^2$.K), depending on the flow velocity. For expected velocity range up to 5 m/s the empirical formula [13] can be used

$$\alpha\ [W/(m^2.K)] = 5.8 + 4.1\ .\ w\ [m/s].$$

Using double wall thickness (case 2), the heat passage coefficient is decreasing of 0.6%, only.

Using double heat transfer coefficient (case 3) the heat passage coefficient is increasing twice (198.8%) and the double wall thickness (case 4) means its relative decreasing of 1.9%.

The next wall thickness increasing on the real value of cellular board (case 5) means its relative decreasing of 2.9%, compared with case 3 or for double wall thickness (case 6) decreasing of 2.2%, compared with case 4. The difference between high and low conduction (case 7 - aluminum) is negligible.

Result: Use of standard cellular boards could be possible.

| No. | λ | α | s | k |
|---|---|---|---|---|
| | W/(m.K) | W/(m$^2$.K) | mm | W/(m$^2$.K) |
| 1 | 0.2 | 20 | 0.2 | 9.901 |
| 2 | | | 0.4 | 9.804 |
| 3 | | 40 | 0.2 | 19.608 |
| 4 | | | 0.4 | 19.230 |
| 5 | | 40 | 0.5 | 19.048 |
| 6 | | | 1.0 | 18.182 |
| 7 | 200 | 20 | 0.2 | 9.9000 |
| | | | 0.4 | 9.9800 |

Table 2: Influence of main parameters on resulting heat passage

## 4 COMPARING OF RESULTS

Actual PC enables to simulate the flow in 1+1 layers of cross flow heat exchanger for outlines of 0.8x0.8 m approx. For comparison here is used the heat exchanger with triangular ribs (6 mm height, 12 mm base, marked as Δ6x12t), see the Par. 2 and the exchanger like the cellular board ("Makrolon") with cross sections of rectangular channels 6x6 mm and 6x36 mm (marked as 6x6t, 6x36t), see Par. 3. From above mentioned results here are presented some basic ones.

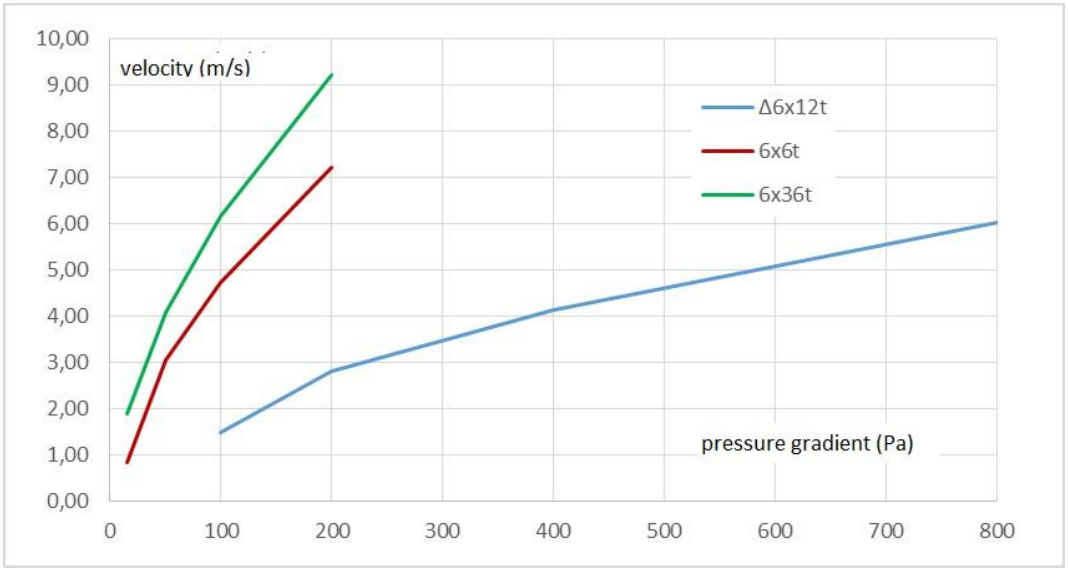



Figure 16: Conversion of pressure gradient on velocity in different geometries (recommended range 1 to 4 m/s)

To create specified range of velocities 1 to 4 m/s approx. it is necessary to specify the pressure gradient. It is different depending on used geometry. The Figure 16 presents the

conversion between specified pressure gradient and velocity. It is visible that in smooth channel of small flow resistance the velocity is much higher than in shaped channel, using the same pressure gradient.

Velocity differences between inlet and outlet of the channel or between channels of warm or cold air can be neglected – it is given by different air density at different temperatures. Those differences here are not displayed.

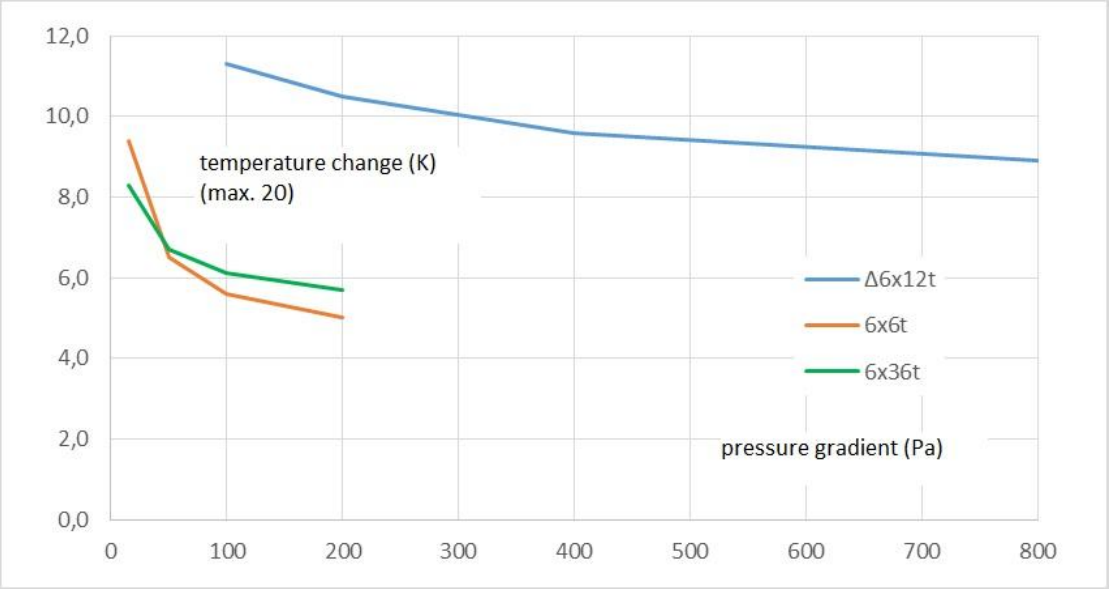



Figure 17: Change of air temperature during passage through heat exchanger

The change of air temperature (i.e. the cooling of warm air or warming of the cold one) after the Figure 17 is typically higher at small flow. With increasing flow the value is going to any value in asymptotic way.

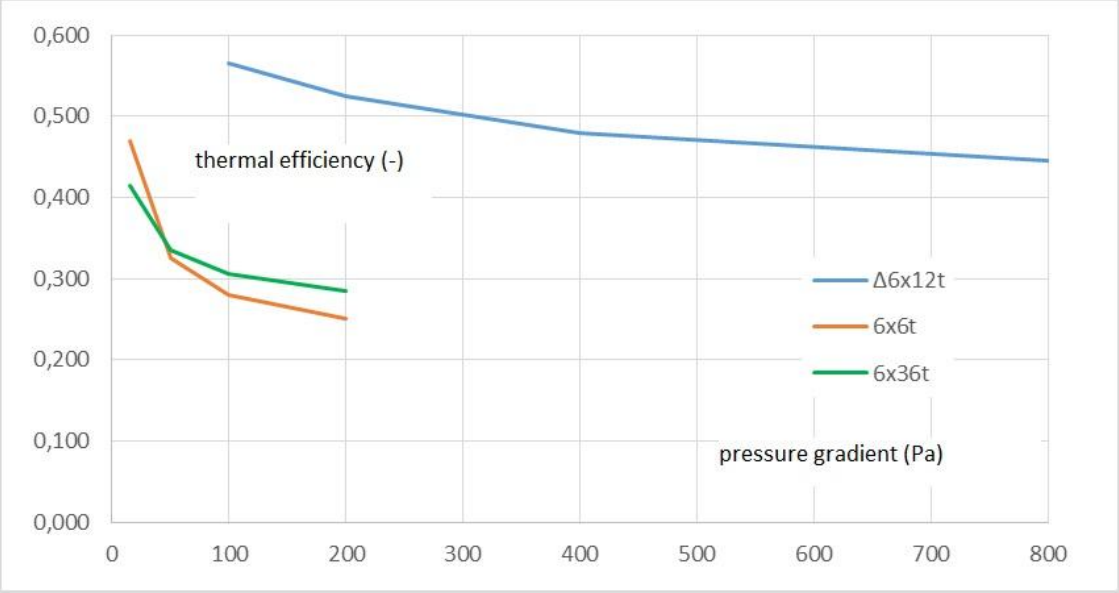



Figure 18: Thermal efficiency of observed heat exchangers

Definition of thermal efficiency: After specified values here is the theoretical maximum temperature change of 20 K (the cold air +5°C would warmed on the temperature of warm air +25°C and vice versa), but practically it is not possible. The rate of real temperature change to

the maximum theoretical one is here defined as **thermal efficiency** of the heat exchanger, see the Figure 18.

Remark: Both characteristics (Figure 17 and Figure 18) are identical; the distinction is given by used proportionality constant, i.e. by specified difference between inlet temperatures of cold and warm air.

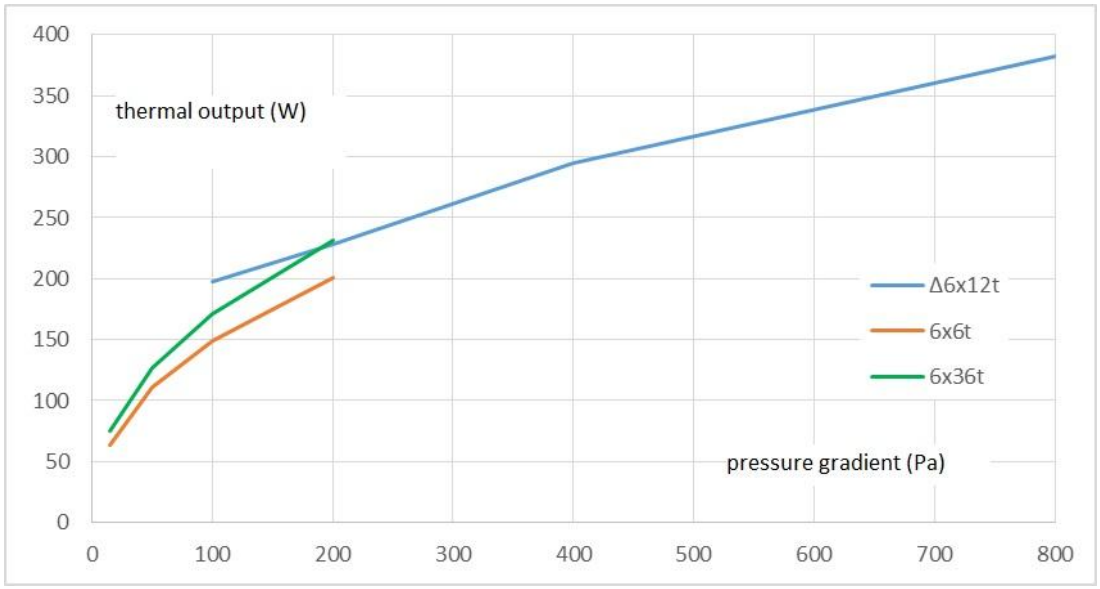



Figure 19: Thermal output of heat exchangers

Thermal output Q = m.cp.ΔT of both kinds of heat exchangers is presented on the Figure 19. The spatially more complicated design (Δ6x12) gives higher thermal output. Its change in dependence on change of flow velocity is slow, compared with smooth channels (6x6t, 6x36t). Their thermal output is smaller, because they have smaller heat transfer surface and higher flow velocity. Thermal output of smooth channels is going near to the output of shaped design until for very high flow velocities (comparing Figure 19 with Figure 16).

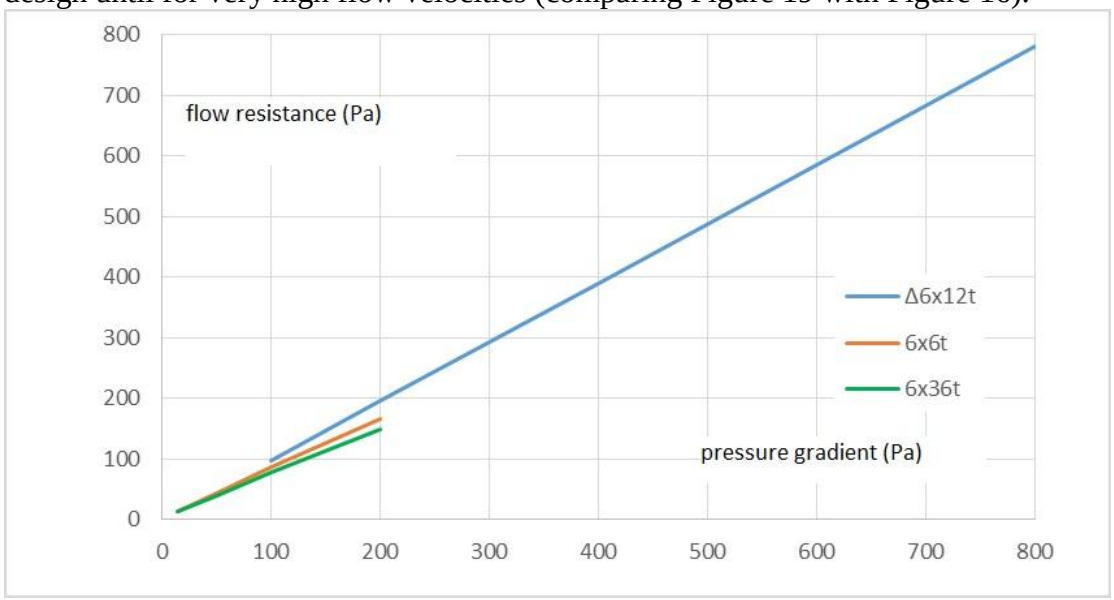



Figure 20: Flow loss of heat exchangers

Flow resistance $\Delta pz = f(w^2)$ on the Figure 20 is proportional to the specified pressure gradient $\Delta p = f(w^2)$. Characteristics of both types of exchangers are quite consecutive, but for smooth channels it is visible absolutely smaller loss.

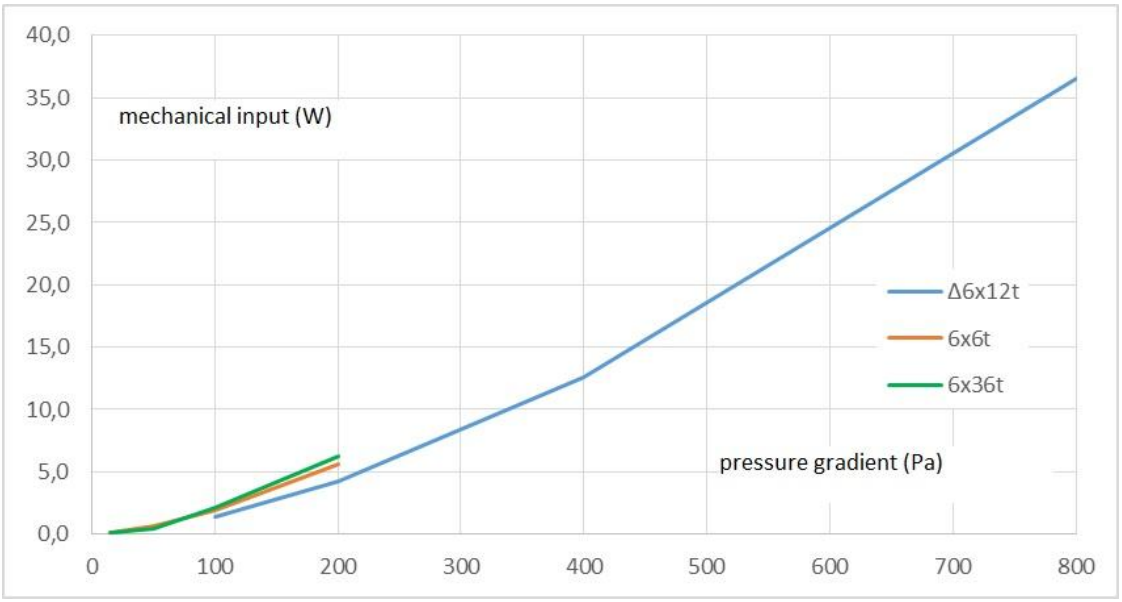



Figure 21: Mechanical input for overcoming of flow resistance

The Figure 21 shows the theoretical mechanical input $P = V. \Delta pz /\eta$, necessary for covering of flow resistance, at the fan efficiency of 100%. Comparing with thermal output, this value can be neglected, even though any real value of fan efficiency would be used, for instance 20%, only.

## 5 CONCLUSION

Models for numerical simulations of flow and heat exchange in cross flow heat exchangers are simplified on 1+1 layer of cold + warm air, only. For simulation of a complete model of any real arrangement is not actually enough of operational memory and number of parallel processes, too. But the system of two cross flows is repeating and so the total air volume is multiplied.

After some testing simulations the simplest model was used as sole channel of different shape of outlines and of ribs, too. The cross flow is simulated by alternative way as inlet temperature at the inlet in channel (for instance warm air) and constant temperature of channel surfaces (for instance cold air) or vice versa.

The simple model can be used for simple and quick testing of several design arrangements of the geometry on two substantial results: recovered heat output between warm and cold air and consumed mechanical input due to the flow resistance. Only the design, which seems to be hopeful, i.e. of relative small flow resistance and of relative high heat recovery, is suitable to verify on enlarged model and finally by experiment, too.

The more complicated model contains 1+1 layers of cross flows with different inlet temperatures (warm + cold) and with heat transfer surface between them. The outside surfaces are defined as thermally insulated; really here are next layers of cross flows, where the heat transfer is realized, too. Such model observes the heat transfer between two adjoining layers of cold and warm air. It is possible to suppose that the really transferred heat could be similar to the reality and the flow resistance is proportional to the lengths of modeled channels.

Actual standard PC enables to simulate the flow and heat exchange in 1+1 layer of cross flow heat exchanger of outlines 0.8x0.8 m approx. Several results of such testing cases are presented as fields of important flow field parameters and operational characteristics as graphs.

Probably the best solution gives the heat transport surface created as narrow corrugated channels. The optimizing of ribs shape and their arrangement by numerical simulation will determine the amplitude and period of the channel axis, the rib pitch and height, where the minimum rib thickness is given by used production technology.

## REFERENCES


[1] "Sick Building Syndrome - Fact Sheet". US EPA

[2] EN 15251 Indoor environmental input parameters for design and assessment of energy performance of buildings-addressing indoor air quality, thermal environment, lighting and acoustics.

[3] Sick Building Syndrome: Modern Day Dilemma - http://www.home-air-purifier-expert.com/sick-building-syndrome.html.

[4] Sick Building Syndrome. *The Environmental Illness Resource Team* ei-resource.org/illness-information/related-conditions/sick-building-syndrome/

K. Adámek, J. Kolář, P. Peukert, Sick Building Syndrome Suppressing by Room Ventilation, *WSEAS Transactions on Fluid Mechanics*, Vol. 11, **2015**

[6] Adámek, K.: Problems of industrial air exhausting, *5th ECCOMAS*, 2014, p. 7454-7465

[7] Ventilating slot „Regel-air“, www.aluplast.net

[8] Pavlů, M., Ventilation for windows suppliers, *Int. Conf. ISOZ*, Liberec, 2015

[9] Adámek, K., Kolář, J., Peukert, P.: Room ventilation, *Proc. of the 13th International Conference on Heat Transfer, Thermal Engineering and Environment (HTE 15*), p. 49-56, Salerno, June, 27-29, 2015

[10] Adámek, K., Kolář, J., Půlpán, P., Pustka, M., Noise Reduction at the Fan Outlet, *Proc. of the EEED 2013, International Conference on Energy, Environment, Ecosystems and Development,* Rhodes, July 16-19, 2013

[11] Adámek, K., Pavlů, M., Bandouch, M., Solar ventilation and tempering, *Proc. of Int. Conf. - XIXth Applic. of Exp. and Num. Meth. in Fluid Mech. and Energ., 2014*, AIP Conf. Proc. 1608, 1 (2014)

[12] ČSN EN 308: Method of laboratory testing for devices of heat recovery air-air

[13] Černoch, S.: Strojně technická příručka, SNTL Praha, 1967

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE INLET AND OUTLET BOUNDARY PROBLEM WITH THE PREFERENCE OF MASS FLOW

**Martin Kyncl, and Jaroslav Pelant**

VZLÚ
Beranových 130, Prague, 19900, CZECH REPUBLIC
e-mail: kyncl@vzlu.cz, pelant@vzlu.cz

**Keywords:** CFD, Compressible Flow, the Navier-Stokes Equations, the Riemann Problem, Boundary Conditions, Mass Flow.

**Abstract.** *We work with the numerical solution of the turbulent compressible gas flow, and we focus on the numerical solution of these equations, and on the boundary conditions. In this work we focus on the inlet and outlet boundary condition with the preference of given mass flow. Usually, the boundary problem is being linearized, or roughly approximated. The inaccuracies implied by these simplifications may be small, but it has a huge impact on the solution in the whole studied area, especially for the non-stationary flow. The boundary condition with the preference of mass flow is sometimes being implemented with the use of some iterative process, guessing the correct values (for the pressure, density, velocity) in order to match the given mass flow through the boundary. In our approach we try to be as exact as possible, using our own original procedures. We follow the exact solution of the initial-value problem for the system of hyperbolic partial differential equations. This complicated problem is modified at the close vicinity of boundary, where the conservation laws are supplied with the additional boundary conditions. We complement the boundary problem suitably, and we show the analysis of the resulting uniquely-solvable modified Riemann problem. The resulting algorithm was coded and used within our own developed code for the solution of the compressible gas flow (the Euler, NS, and RANS equations). The examples show good behaviour of the analyzed boundary conditions.*

## 1 INTRODUCTION

The physical theory of the compressible fluid motion is based on the principles of conservation laws of mass, momentum, and energy. The mathematical equations describing these fundamental conservation laws form a system of partial differential equations. We focus on the numerical solution of these equations. We choose the well-known finite volume method to discretize the analytical problem, represented by the system of the equations in generalized (integral) form. To apply this method we split the area of the interest into the elements, and we construct a piecewise constant solution in time. The crucial problem of this method lies in the evaluation of the so-called fluxes through the edges/faces of the particular elements. We use the analysis of the exact solution of the Riemann problem for the discretization of the fluxes through the boundary edges/faces. We use own algorithms for the solution of the boundary problem, and we use it in the numerical examples.

## 2 THE RIEMANN PROBLEM FOR THE EULER EQUATIONS

For many numerical methods dealing with the two or three dimensional equations, describing the compressible flow, it is useful to solve the Riemann problem for the *3D split* Euler equations. We search the solution of the system of partial differential equations in time $t$ and space $(x, y, z)$

$$\begin{array}{c} \frac{\partial \varrho}{\partial t} + \frac{\partial \varrho u}{\partial x} = 0 \\ \frac{\partial \varrho u}{\partial t} + \frac{\partial (p+\varrho u^2)}{\partial x} = 0 \\ \frac{\partial \varrho v}{\partial t} + \frac{\partial \varrho uv}{\partial x} = 0 \\ \frac{\partial \varrho w}{\partial t} + \frac{\partial \varrho uw}{\partial x} = 0 \\ \frac{\partial E}{\partial t} + \frac{\partial u(E+p)}{\partial x} = 0, \end{array} \tag{1}$$

equipped with the initial conditions

$$\varrho(x,t) = \varrho_L, \boldsymbol{v}(x,t) = \boldsymbol{v}_L, p(x,t) = p_L, \quad x < 0, \tag{2}$$

$$\varrho(x,t) = \varrho_R, \boldsymbol{v}(x,t) = \boldsymbol{v}_R, p(x,t) = p_R, \quad x > 0. \tag{3}$$

Vector $\boldsymbol{v} = (u, v, w)$ denotes the velocity, $\varrho$ density, $p$ pressure, $E = \varrho e + \varrho|\boldsymbol{v}|^2$ is the total energy, with $e$ denoting the specific internal energy. We assume the equation of state for the calorically ideal gas

$$e = \frac{p}{\varrho(\gamma - 1)}.$$

'Split' means here that we still have 5 equations in 3D, but the dependence on $y, z$ (space coordinates) is neglected, and we deal with the system for one space variable $x$. The system (1) is considered in the set $Q_\infty = (-\infty, \infty) \times (0, +\infty)$. The solution of this problem is fundamentally the same as the solution of the Riemann problem for the 1D Euler equations, see [3, page 138]. In fact, the solution for the pressure, the first component of the velocity, and the density is exactly the same as in one-dimensional case. It is a characteristic feature of the hyperbolic equations, that there is a possible raise of discontinuities in solutions, even in the case when the initial conditions are smooth, see [5, page 390]. Here the concept of the classical solution is too restrictive, and therefore we seek a weak solution of this problem. To distinguish physically admissible solutions from nonphysical ones, entropy condition must be introduced, see [5, page 396]. By the solution of the problem (1),(2),(3) we mean the *weak entropy solution* of this problem in $Q_\infty$. The analysis to the solution of this problem can be found in many books, i.e. [3], [5], [6].

Further we are concerned with the one-dimensional Euler equations.

$$\frac{\partial \boldsymbol{q}}{\partial t} + \frac{\partial \boldsymbol{f}(\boldsymbol{q})}{\partial x} = 0, \quad \text{in } Q_\infty = (-\infty, +\infty) \times (0, +\infty) \tag{4}$$

with notation $\boldsymbol{q} = (\varrho, \varrho u, E)^{\mathrm{T}}, \quad \boldsymbol{f}(\boldsymbol{q}) = (\varrho u, \varrho u^2 + p, (E+p)u)^{\mathrm{T}}$. We assume equation of state for ideal gas holds

$$p = (\gamma - 1)(E - \varrho u^2/2) \quad \text{in } Q_\infty.$$

$u$ denotes velocity, $\varrho$ density, $p$ pressure. $E$ denotes total energy. The system is hyperbolic. The Riemann problem for hyperbolic system (4) consists in finding its *entropy weak solution* in $Q_\infty$, which satisfies the initial condition formed by two known constant states $\boldsymbol{q}_L$, $\boldsymbol{q}_R$

$$\boldsymbol{q}(x,0) = \boldsymbol{q}_L = (\varrho_L, \varrho_L u_L, E_L)^{\mathrm{T}}, \qquad x < 0, \tag{5}$$

$$\boldsymbol{q}(x,0) = \boldsymbol{q}_R = (\varrho_R, \varrho_R u_R, E_R)^{\mathrm{T}}, \qquad x > 0. \tag{6}$$

The physical analogue to this problem is so-called *shock-tube problem*.

The general theorem on the solvability of the Riemann problem can be found in [3, page 88]. The solution $\boldsymbol{q} = \boldsymbol{q}(x,t)$ of Riemann problem (4),(5),(6) is piecewise **smooth** and its general form can be seen from figure 1, where a system of half lines is drawn. These half lines define

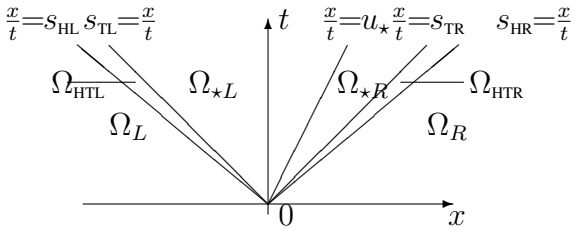



Figure 1: Structure of the solution of the Riemann problem (4),(5),(6)

regions, where $\boldsymbol{q}$ is either constant or smooth. Let us define the open sets called **wedges** (see figure 1):

$$\begin{aligned} \Omega_L &= \text{wedge}(L_{-\infty}, s_{\mathrm{HL}}), & \Omega_{\star R} &= \text{wedge}(u_\star, s_{\mathrm{TR}}), \\ \Omega_{\mathrm{HTL}} &= \text{wedge}(s_{\mathrm{HL}}, s_{\mathrm{TL}}), & \Omega_{\mathrm{HTR}} &= \text{wedge}(s_{\mathrm{TR}}, s_{\mathrm{HR}}), \\ \Omega_{\star L} &= \text{wedge}(s_{\mathrm{TL}}, u_*), & \Omega_R &= \text{wedge}(s_{\mathrm{HR}}, L_{+\infty}). \end{aligned}$$

The solution in $\Omega_L,\ \Omega_{\star L},\ \Omega_{\star R},\ \Omega_R$ is constant (see e.g. [3, page 128])

$$\begin{aligned} \boldsymbol{q}|_{\Omega_L} &= \boldsymbol{q}_L, & \boldsymbol{q}|_{\Omega_{\star R}} &= \boldsymbol{q}_{\star R}, \\ \boldsymbol{q}|_{\Omega_{\star L}} &= \boldsymbol{q}_{\star L}, & \boldsymbol{q}|_{\Omega_R} &= \boldsymbol{q}_R. \end{aligned}$$

Exact solution of the Riemann problem has three waves. The wedges $\Omega_L$ and $\Omega_{\star L}$ are separated by left wave (either 1-*shock wave*, or 1-*rarefaction wave*). There is a *contact discontinuity* in between regions $\Omega_{\star L}$ and $\Omega_{\star R}$. Wedges $\Omega_{\star R}$ and $\Omega_R$ are separated by right wave (either 3-*shock wave*, or 3-*rarefaction wave*). There are four possible wave patterns in the solution.

- ***Contact Discontinuity*** Pressure $p$ and velocity $u$ don't change across **contact discontinuity**, which separates $\Omega_{\star L}$ and $\Omega_{\star R}$ region in general, and moving at speed $u_\star$

$$p|_{\Omega_{\star L} \cup \Omega_{\star R}} = p_\star \ ,$$

$$u|_{\Omega_{\star L}\cup\Omega_{\star R}} = u_\star \ ,$$

In general, there is a discontinuity in density $\varrho$ across half line $\frac{x}{t} = u_\star$

$$\varrho|_{\Omega_{\star L}} \neq \varrho|_{\Omega_{\star R}} \text{ in general.}$$

It is clear, that there must be a jump in temperature, if there is a jump in density, but not in pressure. There is also a jump in entropy. Contact discontinuity is sometimes called *entropy wave*. In reality, contact discontinuities are smeared out due to diffusive effects, which are neglected and not included in hyperbolic system (4).

It is more convenient to use the vector of primitive variables rather then the vector of conservative variables in solving the Riemann problem. Using the theory in [3], [5], [6], we can write the solution for the primitive variables as

$$\begin{aligned}
(\varrho, u, v, w, p)|_{\Omega_L} &= (\varrho_L, u_L, v_L, w_L, p_L),\\
(\varrho, u, v, w, p)|_{\Omega_{\star L}} &= (\varrho_{\star L}, u_\star, v_L, w_L, p_\star),\\
(\varrho, u, v, w, p)|_{\Omega_{\star R}} &= (\varrho_{\star R}, u_\star, v_R, w_R, p_\star),\\
(\varrho, u, v, w, p)|_{\Omega_R} &= (\varrho_R, u_R, v_R, w_R, p_R).
\end{aligned}$$

Here we show the equations for the primitive variables in $\Omega_L \cup \Omega_{HTL} \cup \Omega_{\star L}$. For the solution of the whole system see [3].

- ***1-shock wave***
  One of the possible wave patterns connecting $\Omega_L$ and $\Omega_{\star L}$ is a shock wave. Region $\Omega_{HTL}$ degenerates into single half-line. Primitive variables $\varrho, u, p$ "jump" accross this wave. It is $u_\star < u_L, p_\star > p_L$. Inviscid shock jump relations can be derived, we call them *Rankine-Hugoniot relations*. These leads us to following relations across the 1-shock wave (see 2, or [3, page 125])

$$u_\star = u_L - (p_\star - p_L)\left(\frac{\frac{2}{(\gamma+1)\varrho_L}}{p_\star + \frac{\gamma-1}{\gamma+1}p_L}\right)^{\frac{1}{2}} \tag{7}$$

$$\varrho_{\star L} = \varrho_L \frac{\frac{\gamma-1}{\gamma+1}\frac{p_L}{p_\star} + 1}{\frac{p_L}{p_\star} + \frac{\gamma-1}{\gamma+1}} \tag{8}$$

$$s_1 = u_L - a_L\sqrt{\frac{\gamma+1}{2\gamma}\frac{p_\star}{p_L} + \frac{\gamma-1}{2\gamma}}, \tag{9}$$

  $s_1$ denotes speed of the 1-shock wave. Half line $\frac{x}{t} = s_1$ shapes the boundary between $\Omega_L$ and $\Omega_{\star L}$. It can be shown (see [2]), that (7) can be rewritten in the form

$$p_\star = E_{1Ls}(u_\star), \tag{10}$$

  where

$$\begin{aligned}
E_{1Ls}(u) \ = \ & p_L + \frac{\gamma+1}{4}\varrho_L(u_L-u)^2 + \\
& + \frac{(u_L-u)}{2}\sqrt{4\varrho_L\gamma p_L + \left(\frac{\gamma+1}{2}\right)^2 \varrho_L^2(u_L-u)^2},
\end{aligned} \tag{11}$$

  and $u < u_L$.

- ***1-rarefaction wave***
  Another possible left wave pattern is *rarefaction wave*. It forms $\Omega_{HTL}$ region.Variables changes smoothly within this wave. Across the 1-rarefaction wave $p_\star \leq p_L, u_\star \geq u_L$. Following equations are true.

$$u_\star = u_L + \frac{2}{\gamma - 1} a_L \left[ 1 - \left( \frac{p_\star}{p_L} \right)^{(\gamma-1)/2\gamma} \right] \tag{12}$$

$$\varrho_{\star L} = \varrho_L \left( \frac{p_\star}{p_L} \right)^{\frac{1}{\gamma}} \tag{13}$$

$$s_{TL} = u_\star - a_L \left( \frac{p_\star}{p_L} \right)^{(\gamma-1)/2\gamma}. \tag{14}$$

  Here $a_L = \sqrt{\gamma \frac{p_L}{\rho_L}}$ is the speed of sound in the $\Omega_L$. $s_{TL}$ is speed of the tail of the 1-rarefaction wave. Speed of the head of the rarefaction wave can be expressed

$$s_{HL} = u_L - a_L. \tag{15}$$

  Half line $\frac{x}{t} = s_{TL}$ shapes the boundary between $\Omega_L$ and $\Omega_{HTL}$. Pressure positivity in (12) gives condition $u_\star < u_L + \frac{2}{\gamma-1} a_L$. Equation (12) can be written in the form

$$p_\star = p_L \left( \frac{-u_\star + u_L + \frac{2}{\gamma-1} a_L}{\frac{2}{\gamma-1} a_L} \right)^{\frac{2\gamma}{\gamma-1}}. \tag{16}$$

  State variables in $\Omega_{HTL}$ changes continuously, and can be expressed using equations (see [3], (3.1.97), page 118.)

$$\varrho(x,t) = \varrho_L \left[ \frac{2}{\gamma+1} + \frac{\gamma-1}{(\gamma+1)a_L} \left( u_L - \frac{x}{t} \right) \right]^{\frac{2}{\gamma-1}}, \tag{17}$$

$$u(x,t) = \frac{2}{\gamma+1} \left[ a_L + \frac{\gamma-1}{2} u_L + \frac{x}{t} \right], \tag{18}$$

$$p(x,t) = p_L \left[ \frac{2}{\gamma+1} + \frac{\gamma-1}{(\gamma+1)a_L} \left( u_L - \frac{x}{t} \right) \right]^{\frac{2\gamma}{\gamma-1}}. \tag{19}$$

We combine both possibilities (10), (16) and we get *equation for pressure* $p_\star$ using values $\varrho_L, u_L, p_L$

$$p_\star = \begin{cases} \frac{2p_L + \frac{\gamma+1}{2}\varrho_L(u_L-u_\star)^2 + (u_L-u_\star)\sqrt{4\varrho_L\gamma p_L + \varrho_L^2(\frac{\gamma+1}{2})^2(u_L-u_\star)^2}}{2}, & u_\star < u_L \\ p_L \left( \frac{-u_\star+u_L+\frac{2}{\gamma-1}a_L}{\frac{2}{\gamma-1}a_L} \right)^{\frac{2\gamma}{\gamma-1}}, & u_L \leq u_\star < u_L + \frac{2}{\gamma-1}a_L. \end{cases} \tag{20}$$

Combining (8) and (13) we get equation for density $\varrho_{\star L}$ and pressure $p_\star$

$$\varrho_{\star L} = \begin{cases} \varrho_L \frac{\frac{\gamma-1}{\gamma+1}\frac{p_L}{p_\star}+1}{\frac{p_L}{p_\star}+\frac{\gamma-1}{\gamma+1}}, & p_\star > p_L \\ \varrho_L \left( \frac{p_\star}{p_L} \right)^{\frac{1}{\gamma}}, & p_\star \leq p_L. \end{cases} \tag{21}$$

Let us denote $s_{L?}$ the "wave speed" as follows:

$$s_{L?} = \begin{cases} s_1, & p_\star > p_L \\ s_{TL}, & p_\star \leq p_L. \end{cases}$$

In case of 1-shock wave it denotes $s_1$ wave speed, in the case of 1-rarefaction wave it is the speed of the tail of the wave $s_{TL}$. We combine (9) and (14) to obtain relation between $s_{L?}$ and $p_\star$:

$$s_{L?} = \begin{cases} u_L - a_L\sqrt{\frac{\gamma+1}{2\gamma}\frac{p_\star}{p_L} + \frac{\gamma-1}{2\gamma}}, & p_\star > p_L \\ u_\star - a_L\left(\frac{p_\star}{p_L}\right)^{(\gamma-1)/2\gamma}, & p_\star \leq p_L. \end{cases} \tag{22}$$

There are four unknowns in $\Omega_{\star L}$ to resolve in order to get the solution across left wave and the position of this wave. Relations of these unknowns $\varrho_{\star L}, p_\star, u_\star$ and the wave speed $s_{L?}$ to left state variables $\varrho_L, u_L, p_L$ are given by three equations (20),(21),(22). To get state variables across left wave, i.e. in $\Omega_L \cup \Omega_{STL} \cup \Omega_{\star L}$, and position of the wave, we have to add another equation into the system (20),(21),(22). There is a jump in density across contact discontinuity. This brings another unknown $\varrho_{\star R}$ to unknowns $\varrho_{\star L}, p_\star, u_\star$,$s_{L?}$. We have to add two properly chosen equations into this system of equations in order to get uniquely solvable system for 5 unknowns sought in $\Omega_L \cup \Omega_{STL} \cup \Omega_{\star L} \cup \Omega_{\star R}$ region.

## 3 OUTLET BOUNDARY CONDITION BY THE PREFERENCE OF MASS FLOW

In this section, we attempt to solve the incomplete modified Riemann problem for the Euler equations (4),(5), seeking the solution in the general form of the Riemann problem solution, consisting of 4 constant states separated by 3 waves. Variables $\varrho_L, u_L, p_L$ are known from the initial condition, also the speed of sound $a_L$ is known. We add the boundary condition for the mass flow in $\Omega_{\star L}$ given as

$$u_\star \varrho_{\star L} = G_\star, \tag{23}$$

where $G_\star \geq 0$ is given constant ($G_\star = \frac{massflow}{face\ area}$). Here $u_\star$, $\varrho_{\star R}$ denotes unknown parts of the solution in the region $\Omega_{\star R}$. Velocity $u(x,t)$ and density $\varrho(x,t)$ are constant in this region, as was mentioned in 2. We are interested in the solution with $u_\star > 0$, and $s_{L?} < 0$, which guarantees the possibility of the values to be prescribed at the boundary. In general, there are two possibilities of the wave pattern, which may interest us.

- **1-shock wave**
  Here, we are interested in the solution with $s_1 < 0, u_\star > 0$. Let us construct the function $F_S(u)$, using the relations (20),(8)

$$F_S(u) = u\varrho_{\star L}(u) - G_\star, \tag{24}$$

  where

$$\varrho_{\star L}(u) = \varrho_L \frac{(\gamma-1)p_L + (\gamma+1)p(u)}{(\gamma+1)p_L + (\gamma-1)p(u)},$$

  with

$$p(u) = p_L + \frac{\gamma+1}{4}\varrho_L(u_L-u)^2 + \frac{(u_L-u)}{2}\sqrt{4\varrho_L\gamma p_L + \varrho_L^2(\frac{\gamma+1}{2})^2(u_L-u)^2}.$$

  The sought velocity $u_\star$ is the root of this function $F_S(u)$. It is $s_1 < 0$ for $u_\star < u_X = u_L\frac{2+(\gamma-1)M_L^2}{(\gamma+1)M_L^2}$, with $M_L^2 = u_L^2\varrho/\gamma p$, see [2]. The first derivative is $F_S' > 0$ for

$u < min\{u_L, u_X\}$. The iterative process can be used to compute the solution for the velocity $u_\star$ to any desired accuracy. The problem has a solution with the 1-shock wave if $F_S(min\{u_L, u_X\}) > 0$. Knowing the velocity $u_\star$, we use the relations shown above in (24) to compute the pressure $p_\star$ and the density $\varrho_{\star L}$.

**Remark 1.** Similarly to (24), it is possible to construct the function for the pressure. Let us use the Rankine–Hugoniot conditions across the 1-shock wave $s_1(\varrho_{\star L} - \varrho_L) + u_L\varrho_L = u_\star\varrho_{\star L}$. We construct the function with the use of (8),(9)

$$H(p) = s_1(p)\left(\varrho_{\star L}(p) - \varrho_L\right) + u_L\varrho_L - G,$$

where

$$\varrho_{\star L}(p) = \varrho_L\frac{(\gamma-1)p_L + (\gamma+1)p}{(\gamma+1)p_L + (\gamma-1)p},\ s_1(p) = u_L - a_L\sqrt{\frac{\gamma+1}{2\gamma}\frac{p}{p_L} + \frac{\gamma-1}{2\gamma}}.$$

It is $s_1'(p) < 0$, $\varrho_{\star L}'(p) > 0$ in the interval $\langle p_L, \infty)$. Further it is $(\varrho_{\star L}(p) - \varrho_L) > 0$ in this interval. We are interested in the solutions with $s_1(p) < 0$, which is satisfied for $p > p_X = p_L\left[1 + \frac{2\gamma}{\gamma+1}(M_L^2 - 1)\right]$, see [2], or analysis in [4, page 59]. The first derivative is $H'(p) = s_1'(p)\left(\varrho_{\star L}(p) - \varrho_L\right) + s_1(p)\varrho_{\star L}'(p) < 0$ in $\langle min\{p_L, p_X\}, \infty)$, and $H(min\{p_L, p_X\}) = u_L\varrho_L - G$. Now we may write

$$H(p(u)) = u\varrho_{\star L}(u) - G = F_S(u).$$

The first derivative is then $F_S'(u) = H'(p)p'(u) > 0$. Figure 2 shows the graph of the function $u\varrho_{\star L}$ for the chosen left-hand side initial conditions $\varrho_L, u_L, p_L$.

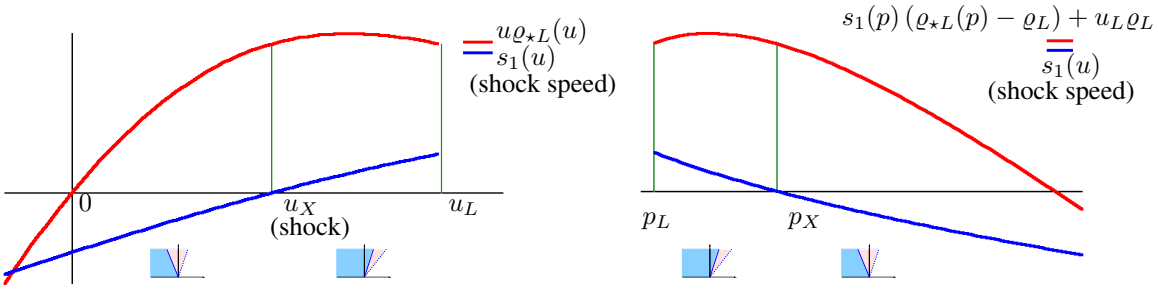



Figure 2: Example of the function $u\varrho_{\star L}(u)$ (representing the massflow in the $\Omega_{\star L}$ region), here $\gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = 500$. Then it is $u_X < u_L$. The function for the pressure is shown at right.

- **1-rarefaction wave**
  Let us assume the solution with the 1-rarefaction wave. From (12),(13),(23) it is

$$u_\star\varrho_L\left[1 - \frac{\gamma-1}{2a_L}(u_\star - u_L)\right]^{\frac{2}{\gamma-1}} - G_\star = 0.$$

The velocity $u_\star$ is the root of the function

$$F_R(u) = u\varrho_L\left[1 - \frac{\gamma-1}{2a_L}(u - u_L)\right]^{\frac{2}{\gamma-1}} - G_\star, \tag{25}$$

defined in the interval $u_L < u < u_L + \frac{2a_L}{\gamma-1}$. It is $F_R(u_L) = \varrho_L u_L - G_\star$. The first derivative $F_R'(u) = \varrho_L \left[1 - \frac{\gamma-1}{2a_L}(u-u_L)\right]^{\frac{2}{\gamma-1}} \left[1 - \frac{u}{a_L - \frac{\gamma-1}{2}(u-u_L)}\right]$ is zero at the points $u_1 = u_L + \frac{2a_L}{\gamma-1}$, $u_0 = \frac{\gamma-1}{\gamma+1}\left(u_L + \frac{2a_L}{\gamma-1}\right)$. Derivative $F_R'(u) > 0$ in the interval $(0, u_0)$ and $F'(u) < 0$ for $u \in (u_0, u_1)$. The maximum of the function $F_R$ is at the point $u_0$. We are interested in the solution with $s_{TL} < 0$. This is satisfied only fo $u_\star < u_0$.

The problem has a solution with the 1-rarefaction wave, if $u_L < u_0$ and $F_R(u_L) \leq 0$ and $F_R(u_0) \geq 0$. It is possible to compute the value $u_\star$ to any desired accuracy using the iterative method. With the velocity $u_\star$ known, we may use the relations (16), and (13). It is

$$p_\star = p_L \left[1 - \frac{\gamma-1}{2a_L}(u-u_L)\right]^{\frac{2\gamma}{\gamma-1}}, \quad \varrho_{\star L} = \varrho_L \left(\frac{p_\star}{p_L}\right)^{\frac{1}{\gamma}}.$$

## Examples

Let $\gamma = 1.4, \varrho_L = 1.25, p_L = 100000$. Figures 3, 4 show the graph of the function $u\varrho_{\star L}$ for various chosen velocities $u_L$.

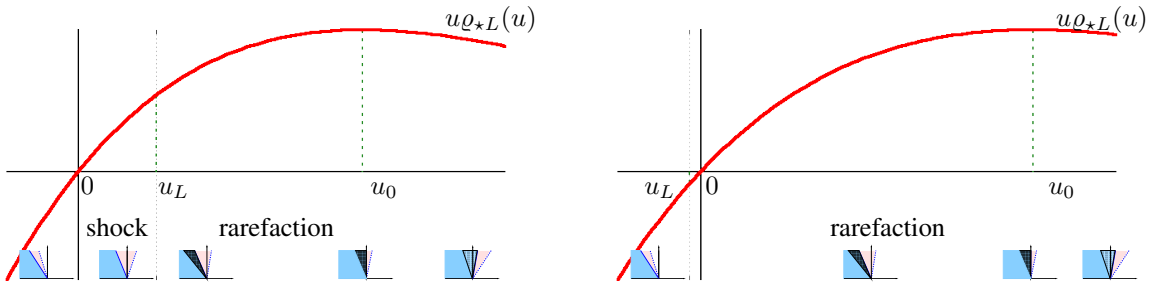



Figure 3: Examples of the function $u\varrho_{\star L}(u)$, here $u_L = 20$ (left), and $u_L = -10$ (right)..

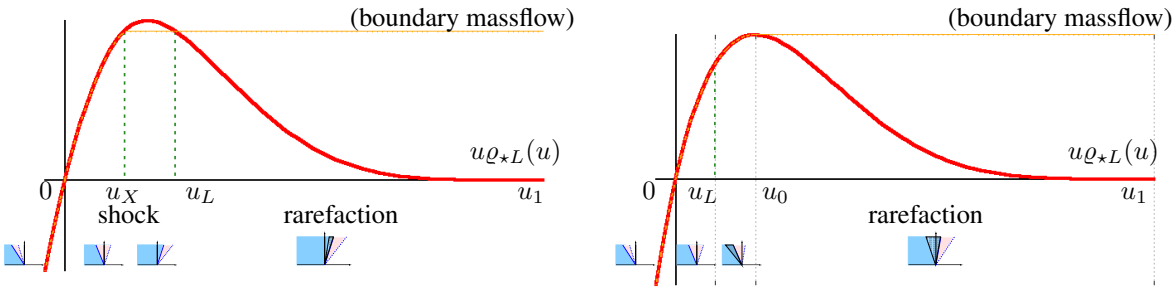



Figure 4: Examples of the function $u\varrho_{\star L}(u)$ for $u_L = 500$ (left), and $u_L = 100$ (right).

## Algorithm 1D

Here we present the possible algorithm for the solution of the state values $U, P, R$ at the boundary.

Input :

$G_\star$ . . . massflow
state inside the boundary element (as a limit to the boundary face)
$\varrho_L$ . . . density , $u_L$ . . . velocity , $p_L$ . . . pressure
Output :
state values at the boundary
$R$ . . . density , $U$ . . . velocity , $P$ . . . pressure

Algorithm :
if $F(u_L) > 0$ then
$U =$ root of $F_S(u),\ u \in (0, min\{u_X, u_L\}),\ (24)$. Use $(20), (8)$.
else
if $u_L > u_0$ then
$U = u_L, P = p_L, R = \varrho_L$
else
if $(u_1 < u_0)$
$U = u_1 - \varepsilon,$ use $(16), (13)$
else
if $F(u_0) < 0$ then
$U = u_0,$ use $(16), (13)$
else
$U =$ root of $F_R(u),\ u \in (u_L, u_0),\ (25)$. Use $(16), (13)$.
end
endif
end
end

**Algorithm 3D**

Here we present the possible algorithm for the solution of the state values at the boundary in 3D.

Input :
$G_\star$ . . . massflow
$\vec{\boldsymbol{n}}$ . . . unit outer normal to the face,
state inside the boundary element (as a limit to the boundary face)
$\varrho_L$ . . . density
$\vec{\boldsymbol{v}} = (v_1, v_2, v_3)$ . . . velocity in global coordinate system
$p_L$ . . . pressure
Output :
state values at the boundary
$R$ . . . density
$\vec{\boldsymbol{V}} = (V_1, V_2, V_3)$ . . . velocity in global coordinate system
$P$ . . . pressure

Algorithm :
1. $u_L = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{n}} = v_1 n_1 + v_2 n_2 + v_3 n_3$
2. use **Algorithm 1D** to compute $U, R, P$
3. $\vec{\boldsymbol{V}} = (n_1(U - u_L) + v_1, n_2(U - u_L) + v_2, n_3(U - u_L) + v_3)$

## 4 INLET BOUNDARY CONDITION BY THE PREFERENCE OF MASS FLOW AND TOTAL TEMPERATURE

In this section, we analyze the incomplete modified Riemann problem for the Euler equations (4),(5), seeking the solution in the general form of the Riemann problem solution, consisting of 4 constant states separated by 3 waves. Variables $\varrho_L, u_L, p_L$ are known from the initial condition, also the speed of sound $a_L$ is known. We add the complementary condition for the mass flow and the total temperature in $\Omega_{\star R}$ given as

$$u_\star \varrho_{\star R} = G_\star, \tag{26}$$

$$\theta_{\star R} = \theta_0 \left(1 - \frac{\gamma - 1}{2} \frac{u_\star^2}{a_0^2}\right), \text{ with } u_\star > -\sqrt{\frac{2a_0^2}{\gamma - 1}}, \tag{27}$$

where $G_\star \leq 0$ is given constant ($G_\star = \frac{massflow}{face\ area}$). Here $u_\star,\ \varrho_{\star R}$ denotes unknown parts of the solution in the region $\Omega_{\star R}$. Further $\theta_0$ is the prescribed total temperature, and $a_0^2 = \gamma R \theta_0$, with $\gamma$ and $R$ being gas constants. We are interested in the solution with $u_\star < 0$ (inlet boundary condition), then the computed values can be prescribed at the boundary. In general, there are two possibilities of the wave pattern, which may interest us.

Let us use the equation of state in the $\Omega_{\star R}$, giving

$$\varrho_{\star R} = \frac{p_\star}{R \theta_{\star R}}, \tag{28}$$

where $p_\star$ is the unknown pressure in the $\Omega_{\star R}$. The equation (20), introduced in section 2, can be used for this pressure. Now equation (26) can be written as

$$u_\star \frac{p_\star}{R \theta_{\star R}} = G_\star, \tag{29}$$

where $p_\star, \theta_{\star R}$ satisfy (20), (27). We seek the unknown velocity $u_\star$ as a root of the function

$$F(u) = u \frac{p(u)}{R\theta(u)} - G_\star \text{ defined for } u \in \left(-\infty, u_L + \frac{2}{\gamma - 1} a_L\right) \cap \left(-\sqrt{\frac{2a_0^2}{\gamma - 1}}, 0\right), \tag{30}$$

where

$$\theta(u) = \theta_0 \left(1 - \frac{\gamma - 1}{2} \frac{u^2}{a_0^2}\right), \quad u \in \left(-\sqrt{\frac{2a_0^2}{\gamma - 1}}, 0\right)$$

$$p(u) = \begin{cases} \frac{2p_L + \frac{\gamma+1}{2}\varrho_L (u_L - u)^2 + (u_L - u)\sqrt{4\varrho_L \gamma p_L + \varrho_L^2 (\frac{\gamma+1}{2})^2 (u_L - u)^2}}{2}, & u < u_L \\ p_L \left(\frac{-u + u_L + \frac{2}{\gamma-1} a_L}{\frac{2}{\gamma-1} a_L}\right)^{\frac{2\gamma}{\gamma-1}}, & u_L \leq u < u_L + \frac{2}{\gamma-1} a_L. \end{cases}$$

The first derivatives of these functions are $\theta'(u) = -\theta_0 \frac{\gamma-1}{2} \frac{2u}{a_o^2} > 0, p'(u) < 0$ in the its definition interval. Then it is

$$F'(u) = \frac{p(u)}{R\theta(u)} + u \frac{p'(u) R\theta(u) - p(u) R\theta'(u)}{R^2 \theta^2(u)} > 0.$$

Further it is $\lim_{u -> -\sqrt{\frac{2a_0^2}{\gamma-1}}} F(u) < 0$ ,and $F(min\{u_L + \frac{2}{\gamma-1} a_L, 0\}) > 0$. The function $F(u)$ is monotone, and the equation (29) has a unique solution if $u_L + \frac{2}{\gamma-1} a_L > -\sqrt{\frac{2a_0^2}{\gamma-1}}$.

**Examples**

Let $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000$. Figures 5-8 show the graph of the function $F(u)$ and the function $u\varrho_{\star R(u)}$ for various chosen velocities $u_L$.

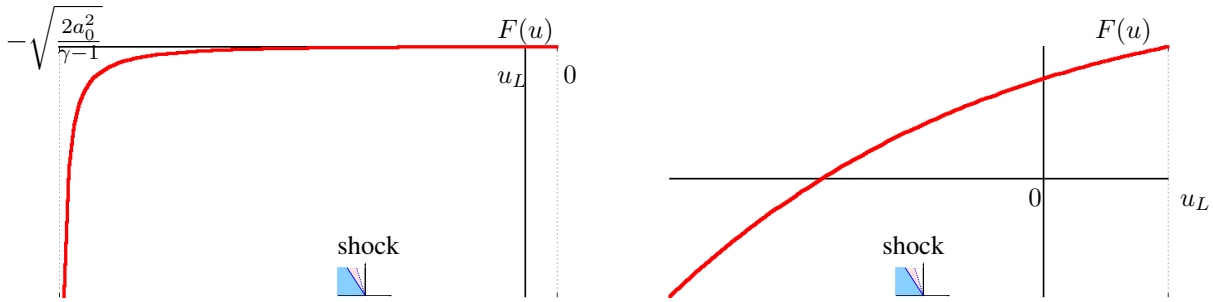



Figure 5: Graph of the function $F(u)$ for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = 50, G_\star = -200$. The right picture shows detail in the restricted interval.

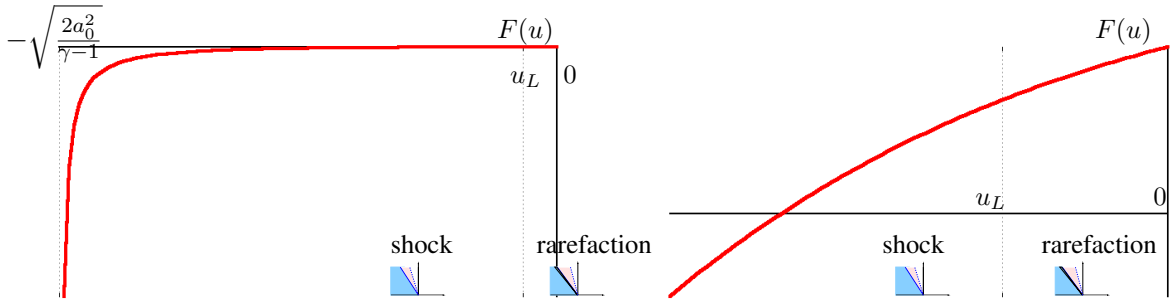



Figure 6: Graph of the function $F(u)$ for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = -50, G_\star = -200$. The right picture shows detail in the restricted interval.

**Remark.**

The equation (26) can be written as

$$u_\star p_\star = G_\star R\theta_0 \left(1 - \frac{\gamma-1}{2}\frac{u_\star^2}{a_0^2}\right). \tag{31}$$

Now we may use (20), and seek the velocity $u_\star$ as a root of the function

$$F(u) = up(u) - G_\star R\theta_0 \left(1 - \frac{\gamma-1}{2}\frac{u^2}{a_0^2}\right),$$

with

$$p(u) = \begin{cases} \frac{2p_L + \frac{\gamma+1}{2}\varrho_L(u_L-u)^2 + (u_L-u)\sqrt{4\varrho_L\gamma p_L + \varrho_L^2(\frac{\gamma+1}{2})^2(u_L-u)^2}}{2}, & u < u_L \\ p_L\left(\frac{-u+u_L+\frac{2}{\gamma-1}a_L}{\frac{2}{\gamma-1}a_L}\right)^{\frac{2\gamma}{\gamma-1}}, & u_L \le u < u_L + \frac{2}{\gamma-1}a_L. \end{cases}$$

The first derivative is

$$F'(u) = p(u) + up'(u) + G_\star R\theta_0 \frac{\gamma-1}{2}\frac{2u}{a_0^2}.$$

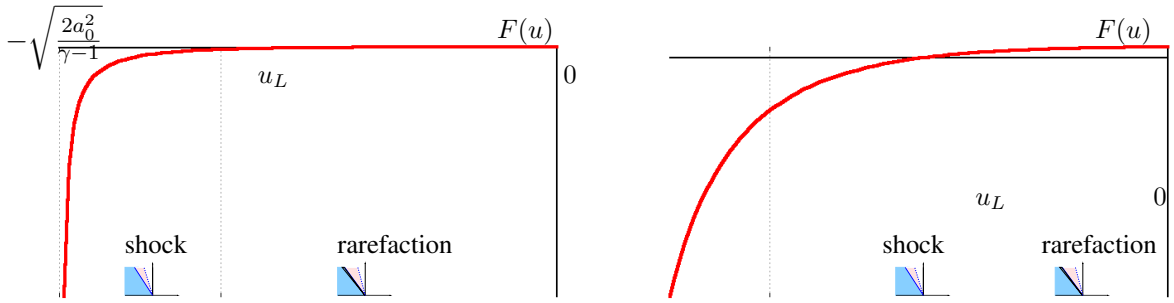



Figure 7: Graph of the function $F(u)$ for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = -500, G_\star = -200$. The right picture shows detail in the restricted interval.

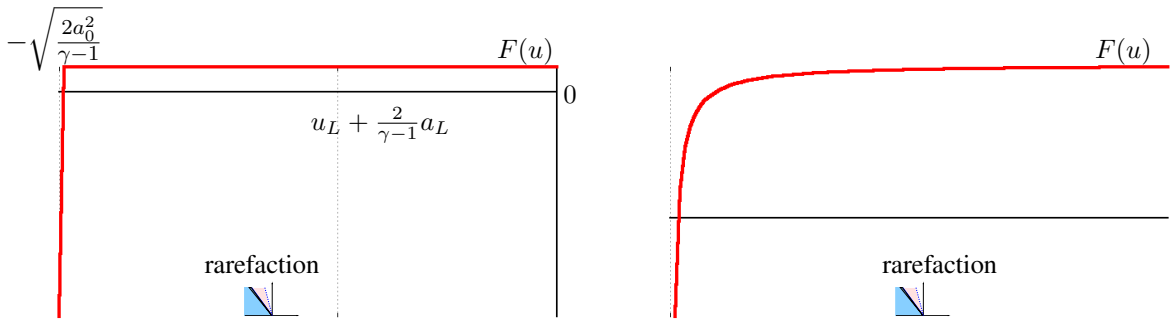



Figure 8: Graph of the function $F(u)$ for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = -2000, G_\star = -200$. The right picture shows detail in the restricted interval.

It is $p(u) > 0$, $p'(u) < 0$, $u < 0$, $G_\star < 0$, and therefore it must be $F_S'(u) > 0$. Further it is $F(0) = -G_\star R\theta_0 > 0$ for $G_\star < 0$, and $F(-\sqrt{\frac{2a_0^2}{\gamma-1}}) < 0$. We seek the root of this function in the interval $u \in \left(max\{u_L, A\}, min\{0, u_L + \frac{2}{\gamma-1}a_L\}\right)$. Figures 9, 10 show the graph of the function $F(u)$ and the function $u\varrho_{\star R(u)}$ for chosen velocities $u_L$ and input data.

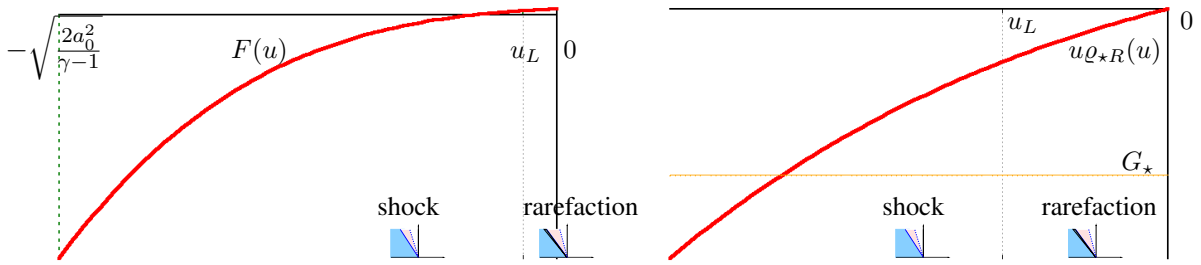



Figure 9: Graph of the function $F(u)$ (left) for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = -50, G_\star = -200$. The right graph shows the massflow function $u\varrho_{\star R}(u)$.

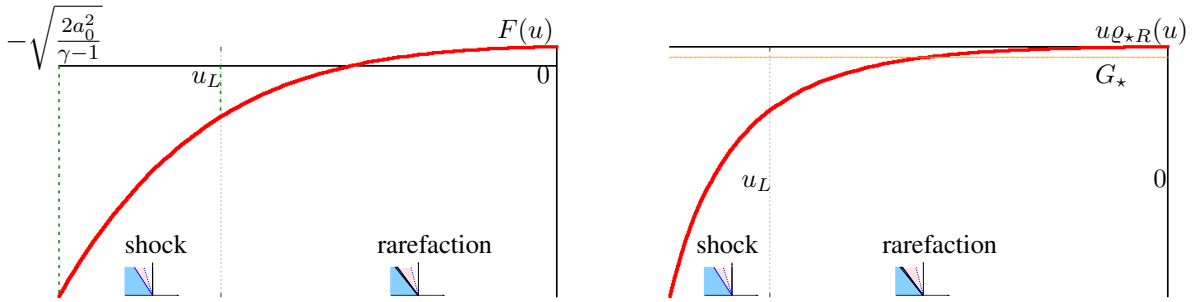



Figure 10: Graph of the function $F(u)$ (left) for given values $\theta_0 = 273.15, \gamma = 1.4, \varrho_L = 1.25, p_L = 100000, u_L = -500, G_\star = -200$. The right graph shows the massflow function $u\varrho_{\star R}(u)$.

**Algorithm 1D**

Here we present the algorithm for the solution of the state values $U, P, R$ at the boundary.

Input :

$G_\star$ . . . massflow

$\theta_0$ . . . total temperature

state inside the boundary element (as a limit to the boundary face)

$\varrho_L$ . . . density

$u_L$ . . . velocity

$p_L$ . . . pressure

Output :

state values at the boundary

$R$ . . . density , $U$ . . . velocity , $P$ . . . pressure

Algorithm :

1. $a_0^2 = \gamma R\theta_0,\ a_L = \sqrt{\frac{\gamma p_L}{\varrho_L}},\ A = -\sqrt{\frac{2a_0^2}{\gamma-1}}$.

if $(u_L + \frac{2}{\gamma-1}a_L < A)$ then

ERROR, NO SOLUTION. bad input data - low temperture at the inlet

else

$F(u_L) = u_L p_L - G_\star R\theta_0 \left(1 - \frac{\gamma-1}{2}\frac{u_L^2}{a_0^2}\right)$

if $(F(u_L) > 0)$ then

$U =$ root of $F_S(u) = up(u) - G_\star R\theta_0 \left(1 - \frac{\gamma-1}{2}\frac{u^2}{a_0^2}\right),\ u \in (A, u_L\rangle\,,$

$p(u) = \frac{2p_L + \frac{\gamma+1}{2}\varrho_L(u_L-u)^2 + (u_L-u)\sqrt{4\varrho_L\gamma p_L + \varrho_L^2(\frac{\gamma+1}{2})^2(u_L-u)^2}}{2}.$

else

$U =$ root of $F_R(u),\ u \in \left(max\{u_L, A\}, min\{0, u_L + \frac{2}{\gamma-1}a_L\}\right),$

$F_R(u) = up(u) - G_\star R\theta_0 \left(1 - \frac{\gamma-1}{2}\frac{u^2}{a_0^2}\right),\ p(u) = p_L\left(\frac{-u+u_L+\frac{2}{\gamma-1}a_L}{\frac{2}{\gamma-1}a_L}\right)^{\frac{2\gamma}{\gamma-1}}.$

endif

use (26), (20): $R = G_\star/u_\star,\ P = p(u_\star)$.

endif

**Algorithm 3D + massflow + total temperature + tangential velocity**

Here we present the possible algorithm for the solution of the state values at the boundary face in 3D.

Input:

$G_\star$ . . . massflow

$\tilde{\theta}_0$ . . . total temperature

$\vec{\boldsymbol{n}}$ . . . unit outer normal to the face, $\vec{\boldsymbol{o}}, \vec{\boldsymbol{p}}$ are tangential directions $(|\vec{\boldsymbol{o}}| = 1, \vec{\boldsymbol{p}} = \vec{\boldsymbol{n}} \times \vec{\boldsymbol{o}})$

$v_t, w_t$ . . . tangential velocity components, $v_t$ in the direction $\vec{\boldsymbol{o}}$, $w_t$ in the direction $\vec{\boldsymbol{p}}$

state inside the boundary element (as a limit to the boundary face)

$\varrho_L$ . . . density

$\vec{\boldsymbol{v}} = (v_1, v_2, v_3)$ . . . velocity in global coordinate system

$p_L$ . . . pressure

Output:

state values at the boundary

$R$ . . . density

$\vec{\boldsymbol{V}} = (V_1, V_2, V_3)$ . . . velocity in global coordinate system

$P$ . . . pressure

Algorithm:

1. $\theta_0 = \tilde{\theta}_0\left(1 - \frac{\gamma-1}{2\gamma R\tilde{\theta}_0}(v_t^2 + w_t^2)\right),\ a_0^2 = \gamma R\theta_0,\ u_L = \vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{n}} = v_1n_1 + v_2n_2 + v_3n_3$
2. use **Algorithm 1D** to compute $U, R, P$
3. $\vec{\boldsymbol{V}} = (n_1U + o_1v_t + p_1w_t, n_2U + o_2v_t + p_2w_t, n_3U + o_3v_t + p_3w_t)$

## 5 EXAMPLES

The developed software with presented boundary conditions was used for the simulation of the compressible turbulent flow in the 3D axis-symmetrical channel. Axis $x$ is the axis of symmetry. At the geometry crosscut shown in figure 11., the inlet is located at the upper part of the boundary, the outlet is located left. We used the following setup in the computation:

- initial condition - constant state in the whole domain,
  $\theta^o = 288.15,\ v_1^o = 0,\ v_2^o = 0,\ v_3^o = 0,\ p^o = 101325.$
- inlet boundary condition - total quantities and velocity direction
  The boundary condition conserving the total pressure, total temperature, and zero tangential velocity, with $\theta_o = 288.15, p_o = 101325.$
- outlet boundary condition (left) - preference of the mass flow
  The boundary condition shown in Section 3, was used, with $G_\star = 4.301$ in average. At each face, the value $G_\star$ was computed (in each iteration) in order to match the average (across the whole boundary).
- walls - wall boundary condition
  The boundary condition preffering the zero normal velocity was used in the case of the inviscid flow. For the viscous flow, this condition was modified by the zero velocity at the wall, and wall temperature $\theta_{WALL} = 288.15$ was set. Further $k_{WALL} = 0$ and $\omega_{WALL} = c_\omega \frac{6\mu}{\beta \varrho y_s^2}$, $c_\omega \frac{6}{\beta} = 120$. Here by $y_s$ we mean the distance between the face and the center of the neighbouring element.

The computational mesh in 2D crosscut consisted of 73x89 quadrilaterals.

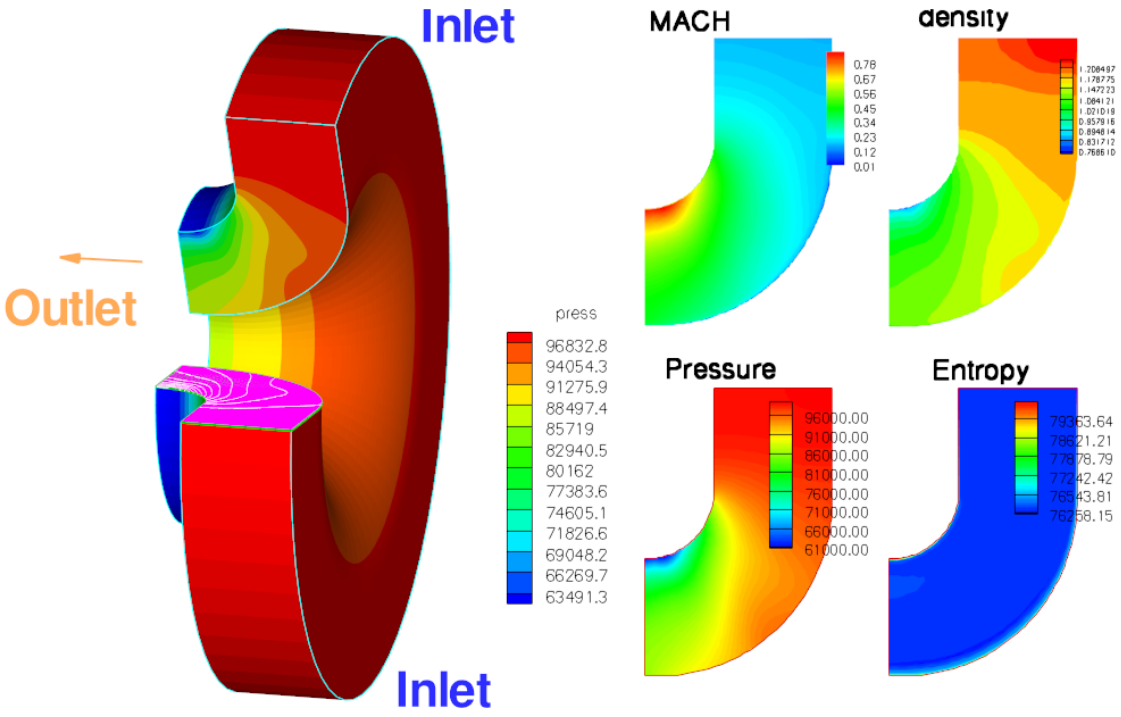



Figure 11: 3D axis-symmetrical turbulent flow, 3D geometry shape, results at chosen 2D crosscuts.

Another example with for different setup was computed, $x$ is the axis of symmetry. At the geometry crosscut shown at figure 12., the inlet is located at the upper part of the boundary, the outlet is located left. We used the following setup for the computation:

- initial condition - constant state in the whole domain,
  $\theta^o = 273.15,\ v_1^o = 0,\ v_2^o = 0,\ v_3^o = 0,\ p^o = 101325.$
- inlet boundary condition - total quantities and velocity direction
  The boundary condition conserving the total pressure, total temperature, and zero tangential velocity, with $\theta_o = 273.15, p_o = 101325.$
- outlet boundary condition (right) - preference of the mass flow
  The boundary condition shown in Section 3, was used, with $G_\star = 4.0$ in average. At each face, the value $G_\star$ was computed (in each iteration) in order to match the average (across the whole boundary).
- walls - wall boundary condition
  The boundary condition preffering the zero normal velocity was used in the case of the inviscid flow. For the viscous flow, this condition was modified by the zero velocity at the wall, and wall temperature $\theta_{WALL} = 273.15$ was set. Further $k_{WALL} = 0$ and $\omega_{WALL} = c_\omega \frac{6\mu}{\beta \varrho y_s^2}$, $c_\omega \frac{6}{\beta} = 120$. Here by $y_s$ we mean the distance between the face and the center of the neighbouring element.

The computational mesh in 2D crosscut consisted of 73x97 quadrilaterals.

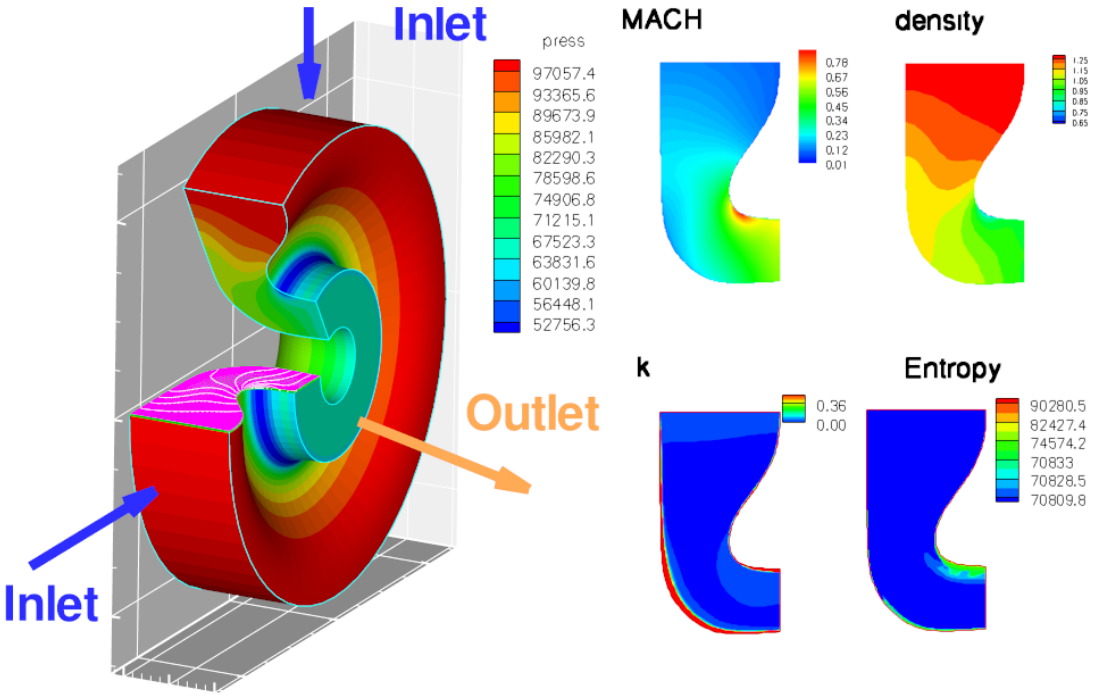



Figure 12: 3D axis-symmetrical turbulent flow, 3D geometry shape, results at chosen 2D crosscuts.

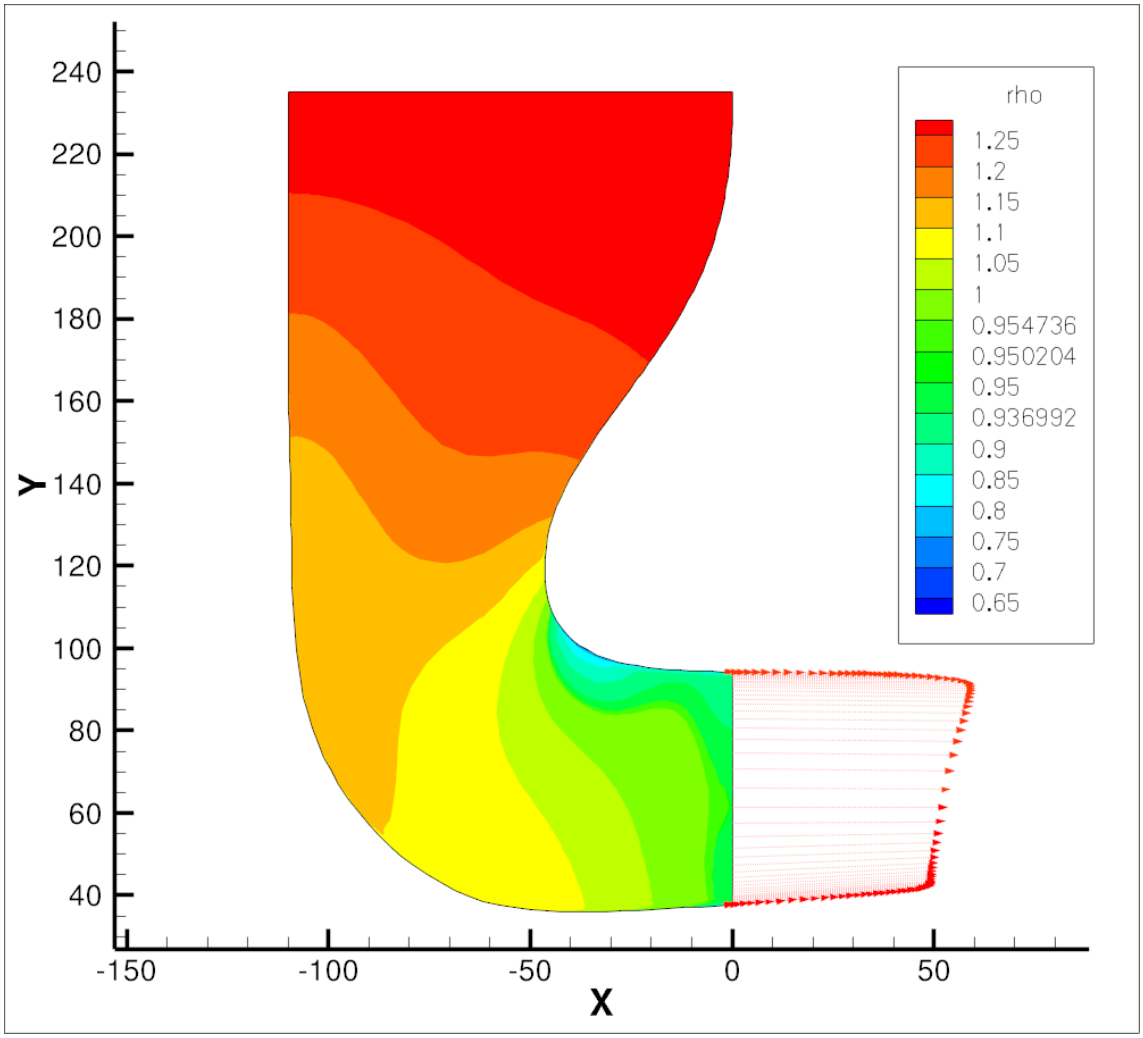



Figure 13: 3D axis-symmetrical turbulent flow, density isolines and velocity profile at the outlet, 2D crosscut.

The following example demonstrates the capabilities of the new boundary condition in the use with the unstationary gas flow. Here $x$ is the axis of symmetry. At the geometry crosscut shown at figure 14., the inlet is located at the left part of the boundary, the outlet points away from axis $x$.

We used the following setup for the simulation:

- initial condition - constant state in the whole domain,
  $\theta^o = 875.00,\ v_1^o = 0,\ v_2^o = 0,\ v_3^o = 0,\ p^o = 101905.$
- inlet boundary condition - total quantities and velocity direction
  The boundary condition conserving the total pressure, total temperature, and zero tangential velocity, with $\theta_o = 875.00, p_o = 101905.$
- outlet boundary condition (right) - preference of the mass flow
  The boundary condition shown in Section 3, was used, with $G_\star = 0.75$ in average. At each face, the value $G_\star$ was computed (in each iteration) in order to match the average (across the whole boundary).
- walls - wall boundary condition
  The boundary condition preffering the zero normal velocity was used in the case of the inviscid flow. For the viscous flow, this condition was modified by the zero velocity at the wall, and wall temperature $\theta_{WALL} = 875.00$ was set. Further $k_{WALL} = 0$ and $\omega_{WALL} = c_\omega \frac{6\mu}{\beta \varrho y_s^2},\ c_\omega \frac{6}{\beta} = 120$. Here by $y_s$ we mean the distance between the face and the center of the neighbouring element.

The computational mesh in 2D crosscut consisted of 89x87 quadrilaterals.

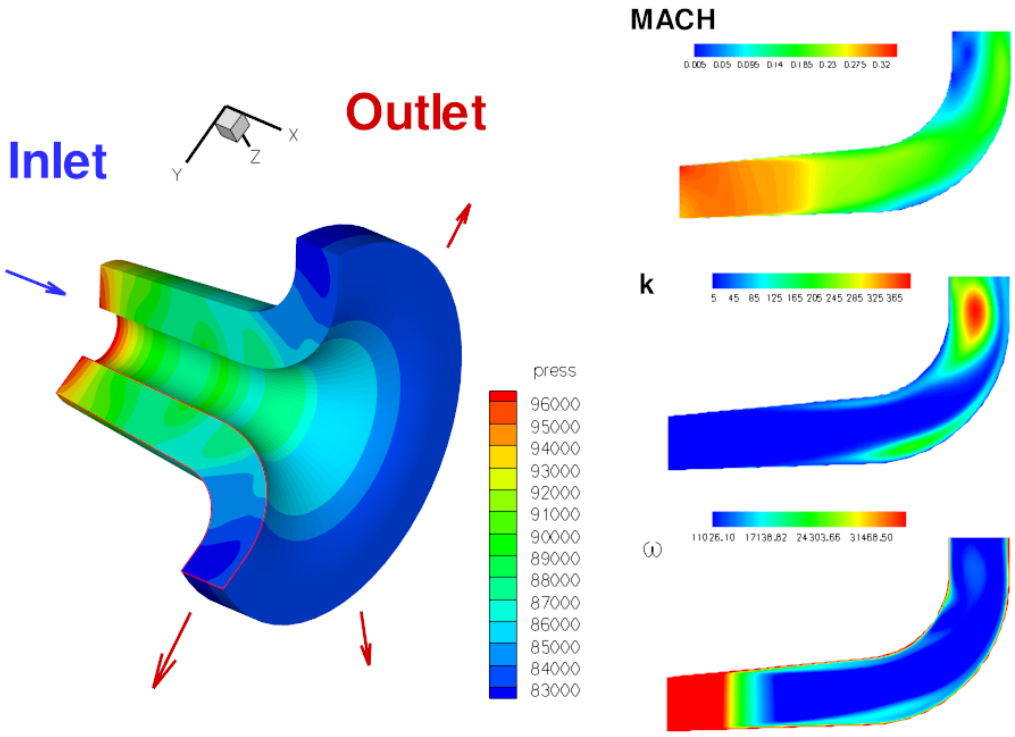



Figure 14: 3D axis-symmetrical turbulent flow, 3D geometry shape, results at chosen 2D crosscuts.

## 6 CONCLUSIONS

In this paper we worked with the system of equations describing the copressible gas flow in 3D. The so-called Riemann problem for the 3D split Euler equations was analyzed, together with the algorithms for the solution of the boundary problems by the preference of mass flow. The suggested algorithms were encoded into own-developed software for the solution of the compressible gas flow. Numerical examples show the use of these original boudary conditions.

**Acknowledgment**
This result originated with the support of Ministry of Industry and Trade of Czech Republic for the long-term strategic development of the research organisation.

## REFERENCES

[1] C. J. Kok. Resolving the dependence on free-stream values for $k - \omega$ turbulence model. *AIAA Journal, Vol. 38., No. 7.*, 2000.

[2] M. Kyncl. *Numerical solution of the three-dimensional compressible flow.* Master's thesis, Prague, (2011). Doctoral Thesis.

[3] M. Feistauer, J. Felcman, and I. Straškraba. *Mathematical and Computational Methods for Compressible Flow*. Oxford University Press, Oxford, (2003).

[4] H. W. Liepmann and A. Roshko. *Elements of Gasdynamics*. John Wiley & Sons, Inc., New York, (1957).

[5] M. Feistauer. *Mathematical Methods in Fluid Dynamics*. Longman Scientific & Technical, Harlow, (1993).

[6] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer, Berlin, (1997).

[7] M. Kyncl and J. Pelant. Applications of the Navier-Stokes equations for 3d viscous laminar flow for symmetric inlet and outlet parts of turbine engines with the use of various boundary conditions. *Technical report R3998, VZLÚ, Beranových 130, Prague, 2006.*

[8] M. Kyncl and J. Pelant. Applications of the Navier-Stokes equations for 2d viscous, compressible turbulent flow on steady grids with the EARSM turbulent model. *Technical report R4300, VZLÚ, Beranových 130, Prague, 2007.*

[9] M. Kyncl and J. Pelant. Outlet Boundary Condition by the Preference of Mass Flow, *Technical report R6155, VZLÚ, Beranových 130, Prague, 2014.*

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# BUDGET ANALYSIS OF TURBULENT KINETIC ENERGY IN CORNER SEPARATION : RANS VS LES

**Jean-François Monier[1,3], Feng Gao[2], Jérôme Boudet[1], Liang Shao[1], Lipeng Lu[4]**

[1]Univ Lyon, École Centrale de Lyon, Laboratoire de Mécanique des Fluides et d'Acoustique,
UMR CNRS 5509, 36, avenue Guy de Collongue F-69134, Écully cedex, France
e-mail: jean-francois.monier@doctorant.ec-lyon.fr, jerome.boudet@ec-lyon.fr, liang.shao@ec-lyon.fr

[2]University of Surrey
Guildford, Surrey GU2 7XH, United Kingdom
e-mail: f.gao@surrey.ac.uk

[3]SNECMA
Rond-point René Ravaud, 77550 Moissy-Cramayel
e-mail: jean-francois.monier@doctorant.ec-lyon.fr

[4]Beihang University
37 Xueyuan Road, Haidian District, Beijing 100191, China
e-mail: lulp@buaa.edu.cn

**Keywords:** Linear compressor cascade, Corner separation, Large-eddy simulation, Reynolds-averaged Navier-Stokes, Reynolds stresses, Turbulent kinetic energy budget

**Abstract.** *The aim of the study is to compare the LES of corner separation, considered as a reference, in the LMFA-NACA65 linear compressor cascade to RANS simulations using Wilcox two-equation models ($k-\omega$). The LES was carried-out by Gao et al. [1] using the Shear-improved Smagorinsky model (SISM), and has been validated by comparison with the experimental data obtained by Ma et al. [2]. Two steady RANS simulations are carried out using $k-\omega$ models: Wilcox 1988 model and Wilcox 1988 model with quadratic constitutive relation.*

*First, the analysis will consider the total pressure losses on a two-dimensional section downstream of the cascade. The Reynolds stresses and the turbulent kinetic energy budget are analysed within the separation region, on a two-dimensional section at the trailing edge of the blade. The RANS simulations mostly overestimate the total pressure losses and the size of the corner separation, compared to the LES and the experiment. The RANS simulations also overestimate the level of the Reynolds stresses and the terms of the turbulent kinetic energy budget, compared to the LES. The quadratic constitutive relation yields some modifications, but not significant enough to have results comparable to LES.*

## 1 INTRODUCTION

The economical and environmental cost reduction in jet engines is linked with the capacity to reduce their size and mass, i.e. increasing the pressure ratio. This leads to an increase of the blade loading, which is known to worsening the three-dimensional flow phenomena. One of those phenomena is the corner separation, i.e. the separation occurring when two orthogonal boundary-layers interact. For a compressor it occurs at the junction of the end-wall and the blade suction side. The corner separation has been investigated experimentally on linear compressor cascades [3, 4, 2, 5], and numerically [6], but its physics is not yet entirely understood.

Numerical simulations give access to a lot of information on the flow behaviour, and are widely used for turbo-machinery development. Reynolds-averaged Navier-Stokes (RANS) simulations are commonly used by industries. In the case of corner separation, RANS simulations tend to over-predict its size and intensity [7, 8, 9]. Bordji *et al.* [10] found that using a non-Boussinesq constitutive relation (the quadratic constitutive relation, referred to as QCR [11]) yields better results for external flows. Large-eddy simulation (LES) is able to simulate three-dimensional secondary flows [12] and is better at simulating corner separation [1, 13] but at a much greater computational cost, which keep industries from using it for conception.

The turbulence physics simulated by both RANS and LES (validated against experimental results) are investigated in order to understand why RANS simulations fail to predict correctly the corner separation characteristics. Reynolds stresses and turbulent kinetic energy budgets from LES and RANS with both original Wilcox $k-\omega$ model and QCR version are compared at the trailing edge of the blade.

## 2 EXPERIMENTAL SET-UP

Ma *et al.* [2, 5] lead an experimental campaign on the LMFA-NACA65 linear cascade. The sketch of the experimental set-up is given figure 1. The blade has a chord $c = 150.0$mm, a stagger angle $\gamma = 42.7°$, a pitch $s = 134.0$mm and a span $h = 370.0$mm. Inlet velocity is set to $40$m/s, which gives a chord-based Reynolds number of $3.82 \times 10^5$. Experimental results used in this paper were acquired for an incidence angle $i = 4°$, where a corner separation has been found. More information concerning the experiment can be found in Ref.[2, 5, 14].

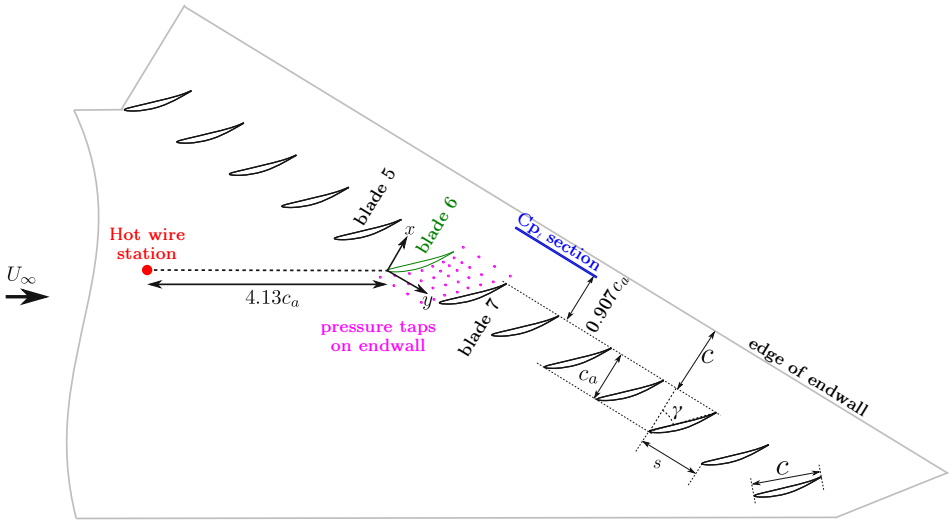



Figure 1: Sketch of the LMFA-NACA65 experimental set-up

# 3 NUMERICAL METHODS

## 3.1 Notations

In order to avoid any ambiguity on the notations, the symbols denoting the ensemble average and the low frequency filter are kept in the equations. The ensemble average of a quantity $\varphi$ is noted $\langle\varphi\rangle$, the Favre ensemble average is noted $[\varphi]$ ($[\varphi] = \langle\rho\varphi\rangle \,/\, \langle\rho\rangle$), the low frequency filter is noted $\overline{\varphi}$ and the Favre low frequency filter is noted $\widetilde{\varphi}$ ($\widetilde{\varphi} = \overline{\rho\varphi}/\overline{\rho}$). Moreover, the fluctuating part of the low-frequency filtered quantity is noted $\varphi'$ (thus, $\overline{\varphi} = \langle\overline{\varphi}\rangle + \varphi'$) and the Favre fluctuating part of the Favre low frequency filtered quantity is noted $\varphi''$ (thus, $\widetilde{\varphi} = [\widetilde{\varphi}] + \varphi''$).

The coordinate system used is the same as in the experiment (Fig. 1). By convention, $x$ and $x_1$ both represent the stream-wise direction, $y$ and $x_2$ both represent the pitch-wise direction and $z$ and $x_3$ both represent the span-wise direction. Consequently, $u$ and $u_1$ both represent the stream-wise velocity, $v$ and $u_2$ both represent the pitch-wise velocity and $w$ and $u_3$ both represent the span-wise velocity.

## 3.2 General specifications

The numerical simulations have been carried out with the solver *Turb'Flow*, a compressible code developed at LMFA relying on finite volume discretization. *Turb'Flow* specifications are given in Ref.[15].

## 3.3 Large-Eddy Simulation

The LES, done by Gao [1], uses a mesh well defined close to the walls ($\simeq 200 \times 10^6$ points, $y^+ \simeq 1$). The spatial scheme is a 4-point Jameson centered scheme [16] with an artificial viscosity coefficient of $0.002$ [17]. The temporal scheme is a 3-step Runge-Kutta scheme with a constant time step of $2.5 \times 10^{-8}$s, corresponding to a CFL (Courant-Friedrichs-Lewy condition) number close to $0.95$. The LES relies on the SISM (shear-improved Smagorinsky model) subgrid scale model of Lévêque *et al.* [18]:

$$\mu_{SGS} = \overline{\rho}\,(C_s\Delta)^2\left(\left|\widetilde{\mathbf{S}}\right| - \left|\left\langle\widetilde{\mathbf{S}}\right\rangle\right|\right), \text{ with } \left|\widetilde{\mathbf{S}}\right| = \left(2\widetilde{S}_{ij}\widetilde{S}_{ij}\right)^{1/2} \text{ and } C_s = 0.18 \qquad (1)$$

where $\mu_{SGS}$ is the sub-grid scale viscosity, $\overline{\rho}$ is the density, $C_s$ is the constant of the original Smagorinsky model, $\Delta$ is a measure of the grid spacing (practically, the cubic root of the grid cell volume) and $\widetilde{\mathbf{S}}$ is the low frequency strain rate tensor. The average $\langle\ \rangle$ is estimated with an exponential smoothing [19], with a cut-off frequency of 533 Hz based on the ratio between twice the free-stream velocity and the chord. Complete details of the LES specifications are given in Ref. [1].

The Reynolds stresses extracted from the LES results are defined as follow:

$$\boldsymbol{\tau}_{\mathbf{t}} = -\left\langle\overline{\rho}\mathbf{u}'' \otimes \mathbf{u}''\right\rangle \qquad (2)$$

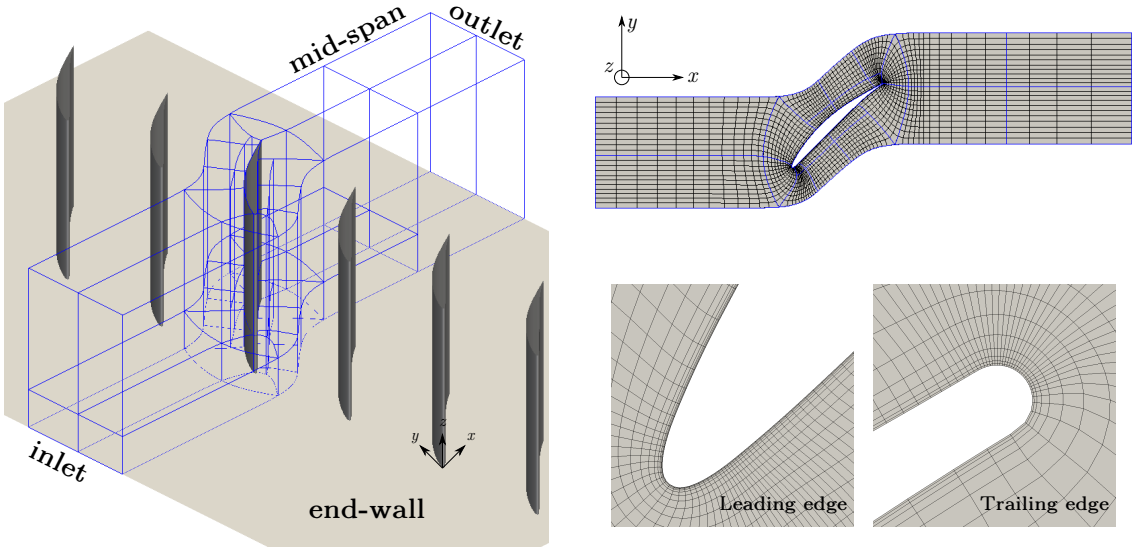
A Reynolds stress budget can be extracted, following the method of Gao [20]. The turbulent kinetic energy budget (referred to as TKE budget) is then deduced from the Reynolds stress budget diagonal terms addition ($k = \frac{1}{2}u_i''u_i''$). Theoretically, the TKE budget is balanced. However, there is practically a residual, with a dissipative effect. This residual represents the various numerical dissipations, such as numerical scheme dissipation. A term of numerical dissipation, noted $\Xi$ and calculated as the opposite of the residual, is added in the budget to close it.

$$\frac{\partial \langle \overline{\rho} k \rangle}{\partial t} = 0 = \underbrace{-\frac{\partial}{\partial x_j}\left(\langle \overline{\rho} k \rangle \left[\widetilde{u_j}\right]\right)}_{\text{Advection}} \underbrace{-\left\langle \overline{\rho} u_i'' u_j'' \right\rangle \frac{\partial \left[\widetilde{u_i}\right]}{\partial x_j}}_{\text{Production}} \underbrace{-\frac{\partial}{\partial x_j}\left(\left\langle \overline{\rho} k u_j'' \right\rangle\right)}_{\text{Turbulent diffusion}} \underbrace{-\frac{\partial \left\langle u_i'' p' \right\rangle}{\partial x_i}}_{\text{Pressure diffusion}} \underbrace{+\left\langle p' \frac{\partial u_i''}{\partial x_i} \right\rangle}_{\text{Pressure dilatation}}$$
$$\underbrace{-\left\langle u_i'' \right\rangle \frac{\partial \langle \overline{p} \rangle}{\partial x_i}}_{} \underbrace{+\frac{\partial \left\langle \overline{\tau_{ij}} u_i'' \right\rangle}{\partial x_j}}_{\text{Viscous diffusion}} \underbrace{-\left\langle \overline{\tau_{ij}} \frac{\partial u_i''}{\partial x_j} \right\rangle}_{\text{Viscous dissipation}} \underbrace{+\frac{\partial \left\langle \overline{\Pi_{ij}} u_i'' \right\rangle}{\partial x_j}}_{\text{SGS diffusion}} \underbrace{-\left\langle \overline{\Pi_{ij}} \frac{\partial u_i''}{\partial x_j} \right\rangle}_{\text{SGS Dissipation}} \underbrace{+\Xi_{ii}}_{\text{Numerical dissipation}} \quad (3)$$

The LES TKE budget, presented in equation (3), shows more terms than the RANS equation of $k$. In order to compare these two, some terms are grouped. Production and numerical dissipation are kept apart. The dissipation term is the sum of the viscous dissipation and the SGS dissipation. All the other terms are summed to form the transport term.

### 3.4 Reynolds-Averaged Navier-Stokes

The RANS simulations use the reference mesh from Gao [20] for the incidence $i = 4°$. This mesh is a coarser version of the LES mesh ($\simeq 2.8 \times 10^6$ points) with no trip bands (cf. Fig. 2). The spatial scheme is a 4-point Jameson centered spatial scheme with an artificial viscosity coefficient of $0.02$. The temporal scheme is a 5-step Runge-Kutta scheme with a CFL number of $0.7$. Two turbulence models are investigated, the original Wilcox $k-\omega$ model [21] and a modified version using the QCR from Spalart [11]. The inlet conditions (density, velocity and turbulent variables) are extracted from a 2-D flat plate boundary layer simulation at the position where momentum boundary layer thickness $\delta_1$ matches the experiment. The outlet is set to a partially non reflective pressure condition. The walls boundary conditions are set to non-slip adiabatic. The lateral boundary conditions are set to periodic. The mid-span boundary condition is set to symmetry, given that only half the channel is simulated.




(a) View of the computational domain

(b) View of the blade-to-blade mesh

Figure 2: Views of the computational domain and the mesh. The mesh is coarsened in order to be displayed.

The original Wilcox $k-\omega$ model [21] calculates the Reynolds stress tensor $\boldsymbol{\tau_t}$ using Boussinesq constitutive relation:

$$\tau_{t,ij} = 2\mu_t S_{ij}^{(0)} - \frac{2}{3}\delta_{ij}\left\langle\rho\right\rangle[k] \text{ , where } \mu_t = \left\langle\rho\right\rangle\frac{[k]}{[\omega]} \tag{4}$$

with $\delta_{ij}$ the Kroeneker symbol and $S^{(0)}$ the zero-trace strain rate tensor:

$$S_{ij}^{(0)} = S_{ij} - \frac{1}{3}\delta_{ij}\frac{\partial[u_k]}{\partial x_k} \ ; \ S_{ij} = \frac{1}{2}\left(\frac{\partial[u_i]}{\partial x_j} + \frac{\partial[u_j]}{\partial x_i}\right) \tag{5}$$

The $k$ and $\omega$ budget equations are :

$$\frac{\partial\left\langle\rho\right\rangle[k]}{\partial t} = 0 = \underbrace{\left\langle\tau_{t,ij}\right\rangle\frac{\partial[u_i]}{\partial x_j}}_{\textbf{Production}} \underbrace{-\frac{\partial\left\langle\rho\right\rangle[k][u_j]}{\partial x_j}}_{\textbf{Advection}} \underbrace{+\frac{\partial}{\partial x_j}\left(\mu\frac{\partial[k]}{\partial x_j}\right)}_{\substack{\textbf{Molecular}\\\textbf{diffusion}}} \underbrace{+\frac{\partial}{\partial x_j}\left(\frac{\mu_t}{\sigma_k}\frac{\partial[k]}{\partial x_j}\right)}_{\substack{\textbf{Turbulent transport \&}\\\textbf{pressure diffusion}}} \underbrace{-c_k\left\langle\rho\right\rangle[k][\omega]}_{\textbf{Dissipation}} \tag{6}$$

$$\frac{\partial\left\langle\rho\right\rangle[\omega]}{\partial t} = 0 = \underbrace{c_{\omega 1}\frac{[\omega]}{[k]}\left\langle\tau_{t,ij}\right\rangle\frac{\partial[u_i]}{\partial x_j}}_{\textbf{Production}} \underbrace{-\frac{\partial\left\langle\rho\right\rangle[\omega][u_j]}{\partial x_j}}_{\textbf{Advection}} \underbrace{+\frac{\partial}{\partial x_j}\left(\left(\mu+\frac{\mu_t}{\sigma_\omega}\right)\frac{\partial[\omega]}{\partial x_j}\right)}_{\substack{\textbf{Turbulent transport,}\\\textbf{pressure diffusion \&}\\\textbf{molecular diffusion}}} \underbrace{-c_{\omega 2}\left\langle\rho\right\rangle[\omega]^2}_{\textbf{Dissipation}} \tag{7}$$

where $c_k = 0.09$, $c_{\omega 1} = 5/9$, $c_{\omega 2} = 3/40$, $\sigma_k = 2.0$ and $\sigma_\omega = 2.0$. Advection, molecular diffusion and turbulent transport terms are summed to form a transport term in the following analysis.

The QCR, proposed by Spalart [11], is a modification of Boussinesq constitutive relation in order to take into account the important effect of vorticity. The version we implemented is the modified one from Mani [22] in order to take into account the compressible part, given that *Turb'Flow* is a compressible solver:

$$\tau_{t,ij}^{QCR} = \tau_{t,ij} - c_{QCR}\left(O_{ik}\tau_{t,jk} + O_{jk}\tau_{t,ik}\right) \tag{8}$$

where $\boldsymbol{\tau_t}$ is the Boussinesq Reynolds stress tensor (Eq. (4)) and $\mathbf{O}$ is the normalised rotation tensor:

$$O_{ij} = \frac{\frac{\partial[u_i]}{\partial x_j} - \frac{\partial[u_j]}{\partial x_i}}{\sqrt{\frac{\partial[u_k]}{\partial x_l}\frac{\partial[u_k]}{\partial x_l}}} \text{ and } c_{QCR} = 0.3 \tag{9}$$

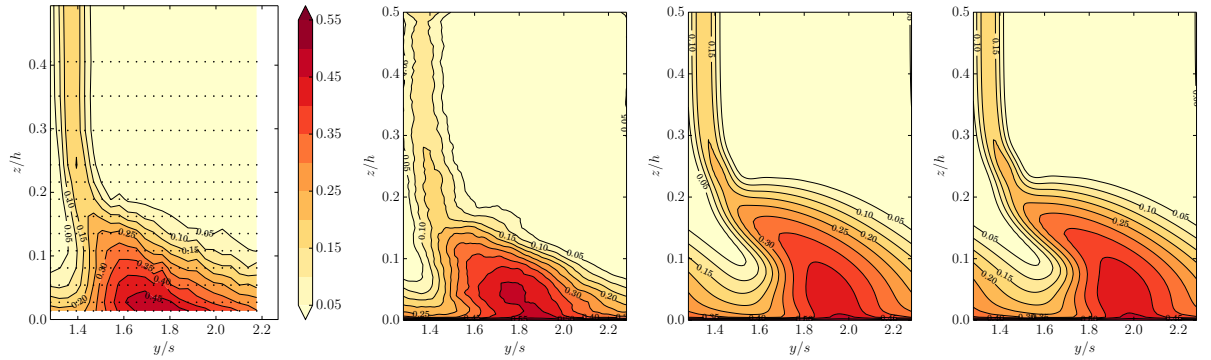
The $k$ and $\omega$ equations remain the same as in the original model except that $\boldsymbol{\tau_t^{QCR}}$ is used instead of $\boldsymbol{\tau_t}$ in the production terms.

## 4 TOTAL PRESSURE LOSS

Total pressure losses are an important criterion for compressor efficiency. The total pressure loss coefficient $Cp_t$ is defined as:

$$Cp_t = \frac{P_{t\infty} - P_t}{P_{t\infty} - P_{s\infty}} \tag{10}$$

where $P_t$ is the local total pressure, $P_{t\infty}$ is the reference total pressure and $P_{s\infty}$ is the reference static pressure. For the RANS simulations, the reference pressures where taken $100$mm upstream the leading edge, at a height of $h/4$.




(a) Experiment (b) LES (c) RANS Wilcox 88 (d) RANS Wilcox 88 - QCR

Figure 3: Total pressure loss coefficient on the outlet plane at $1.907c_a$. (cf. Fig.1).

Experimental total pressure losses (Fig. 3a) and LES total pressure losses (Fig. 3b) are very similar. They both exhibit a triangular form for the corner separation wake, and display levels of pressure loss very similar. These results legitimate the usage of LES as a reference.

The RANS simulations over-predict the losses in each case. The use of a non-linear constitutive relation (QCR) yields almost no effect on the results (Fig. 3d) compared to the original Wilcox model (Fig. 3c). The wake of the corner separation is notably twisted in the RANS simulations. This could be an effect of the over-prediction of the size and intensity of the corner separation.

## 5 REYNOLDS STRESS TENSOR

The Reynolds stresses give a lot of information on the physics simulated. They are by nature a measure of the effect of the turbulence, thus they give a measure of the quality of the turbulence model for RANS simulations when compared with LES. Both normal stresses and shear stresses are affected by constitutive closure, so both have to be investigated against LES.

The Reynolds stresses are extracted at the trailing edge of the blade ($x = 1.0c_a$), on a rectangular surface covering the pitch. The plane position is given in Fig. 4.

In the following subsections, the Reynolds stresses are normalised by the product of a reference density ($\langle\overline{\rho_\infty}\rangle$ for LES, $\langle\rho_\infty\rangle$ for RANS simulations) times a reference velocity squared ($[\widetilde{u_\infty}]^2$ for LES, $[u_\infty]^2$ for RANS). The reference densities and velocities are taken $100$mm upstream the leading edge, at a height of $h/4$.

### 5.1 Reynolds normal stresses

The Reynolds normal stresses ($\tau_{t,11}$, $\tau_{t,22}$ and $\tau_{t,33}$) are the main terms of the Reynolds stress tensor concerning intensity.

Figure 5 shows that both RANS simulations (Fig. 5b and 5c) over-predict the intensity of the first Reynolds normal stress compared to LES (Fig. 5a). The modification of the constitutive relation yields almost no effect in this case, $\tau_{t,11}$ and $\tau_{t,11}^{QCR}$ are similar.

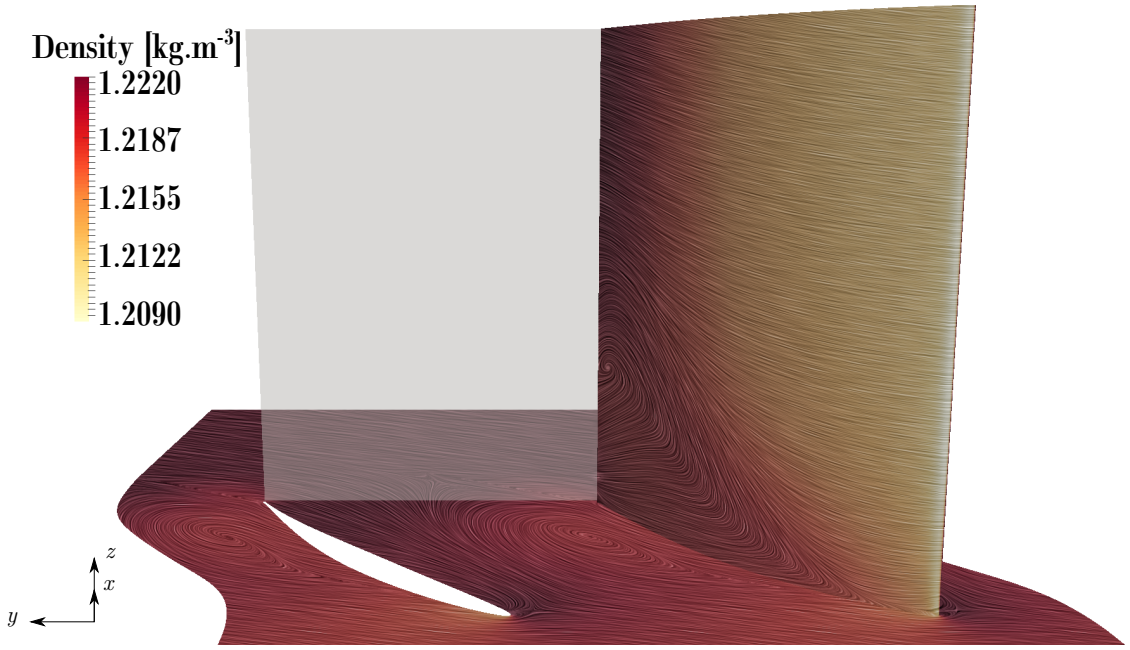
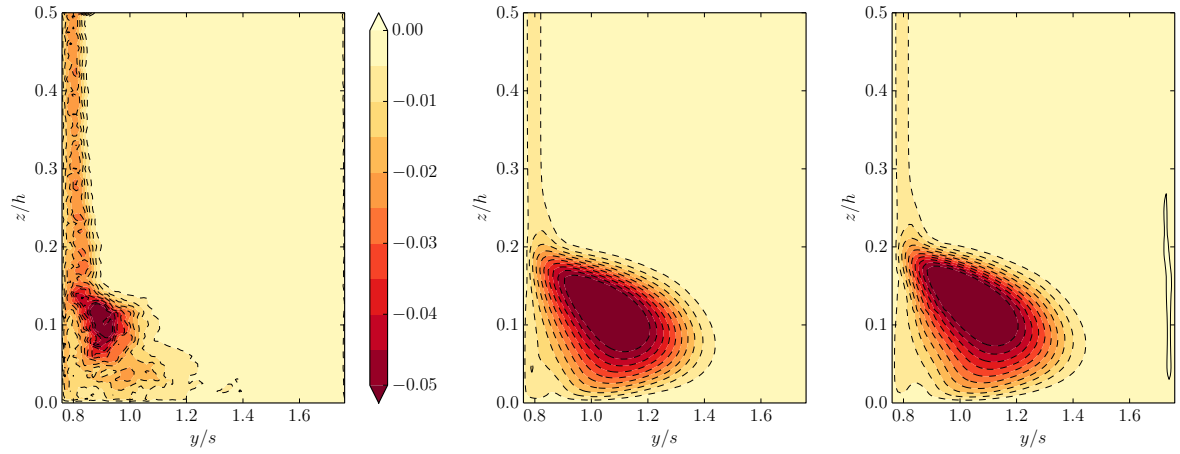



Figure 4: Extraction plane for the Reynolds stresses. Limiting streamlines are plotted for the RANS case, coloured by density.

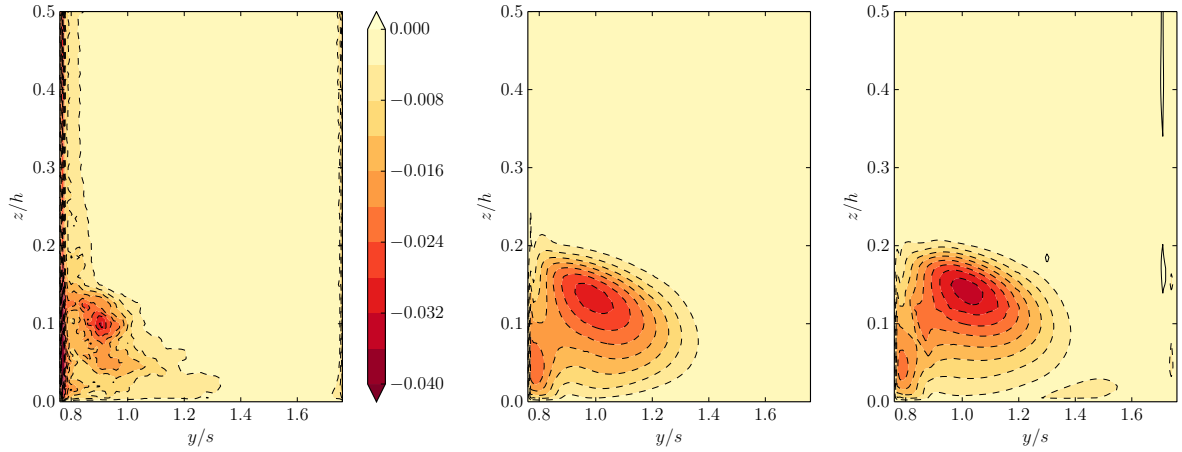



(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 5: $\tau_{t,11}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,11}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)

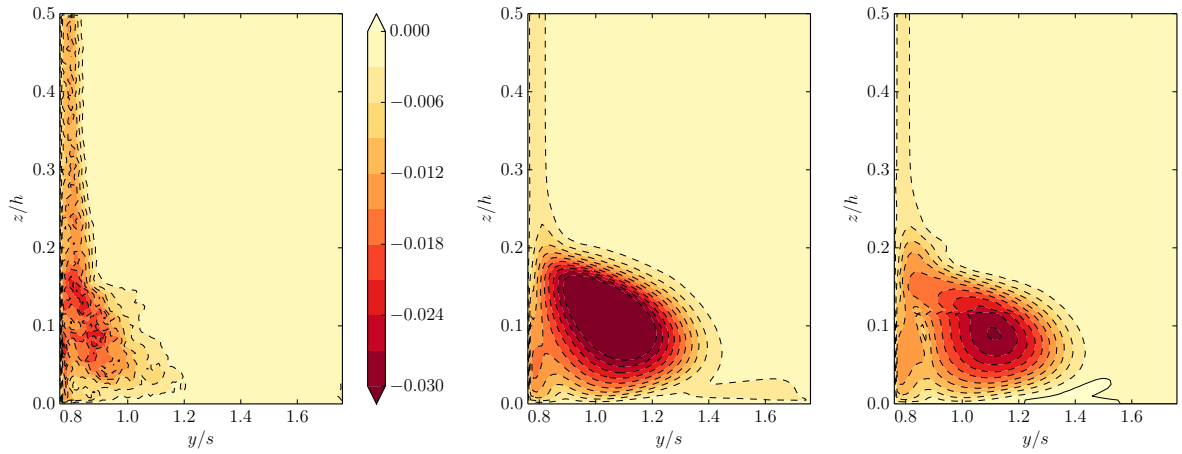
Figure 6 shows that the second Reynolds normal stress follows the same trend as the first one. The RANS simulations (Fig. 6b and 6c) still over-predict the intensity of $\tau_{t,22}$ compared to LES (Fig. 6a). However, the QCR has a more visible impact than previously. Compared to the Boussinesq closure, the QCR yields a lower minimum, more comparable with the minimum found by LES. However, both RANS simulations fail to represent the boundary-layer of the blade, visible in the LES for $y/s \simeq 0.8$.

Figure 7 shows that the third Reynolds normal stress follows the same trends as the first and second ones. The RANS simulations (Fig. 7b and 7c) over-predict the intensity of $\tau_{t,33}$ compared to LES (Fig. 7a). The impact of the QCR is noticeable on this term, compared with




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 6: $\tau_{t,22}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,22}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

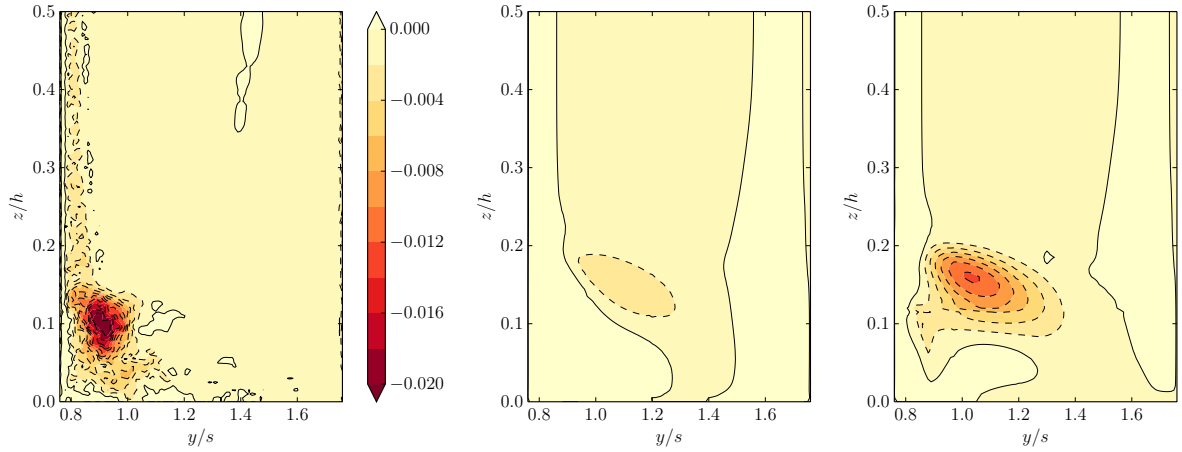
Figure 7: $\tau_{t,33}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,33}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)

the original Wilcox model. It weakens the minimum of the Reynolds stress in the middle of the separation, but not enough to yield a value comparable with the LES.

## 5.2 Reynolds shear stresses

Reynolds shear stresses ($\tau_{t,12}$, $\tau_{t,13}$ and $\tau_{t,23}$) are of a lower intensity than the normal stresses.

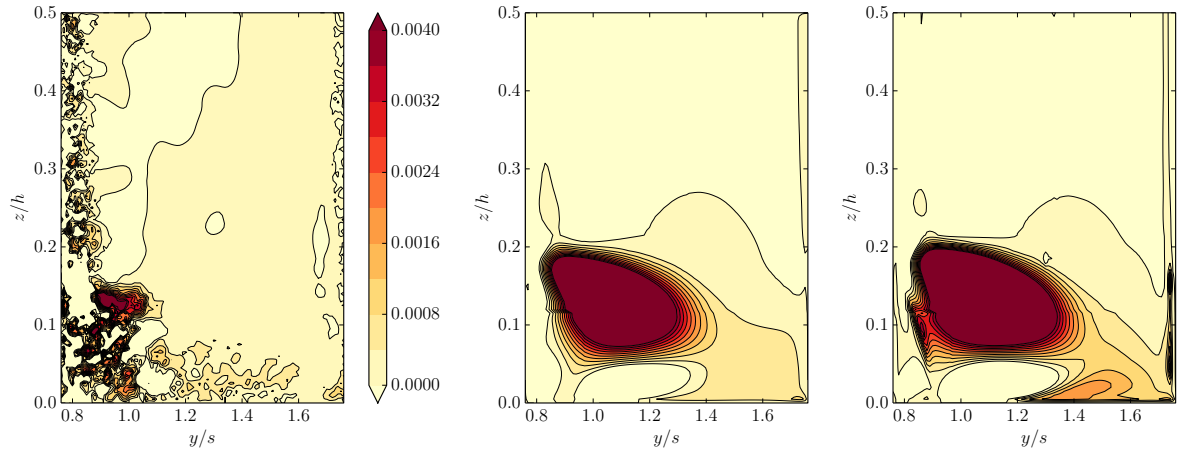
Figure 8 shows that the first Reynolds shear stress is totally mispredicted by the RANS simulation with the original Wilcox turbulence model (Fig. 8b), compared to the LES (Fig. 8a). The minimum of this Reynolds shear stress should be close to the corner according to LES, and it should be more intense. The QCR modification improves the results by reducing the minimum and widening the area of negative stresses. It is still not enough to have the same levels, position, or the same physical behaviour as LES. Besides, both RANS simulations fail to predict the existence of stresses in the boundary-layer of the blade, visible in the LES for




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 8: $\tau_{t,12}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,12}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)

$y/s \simeq 0.8$.

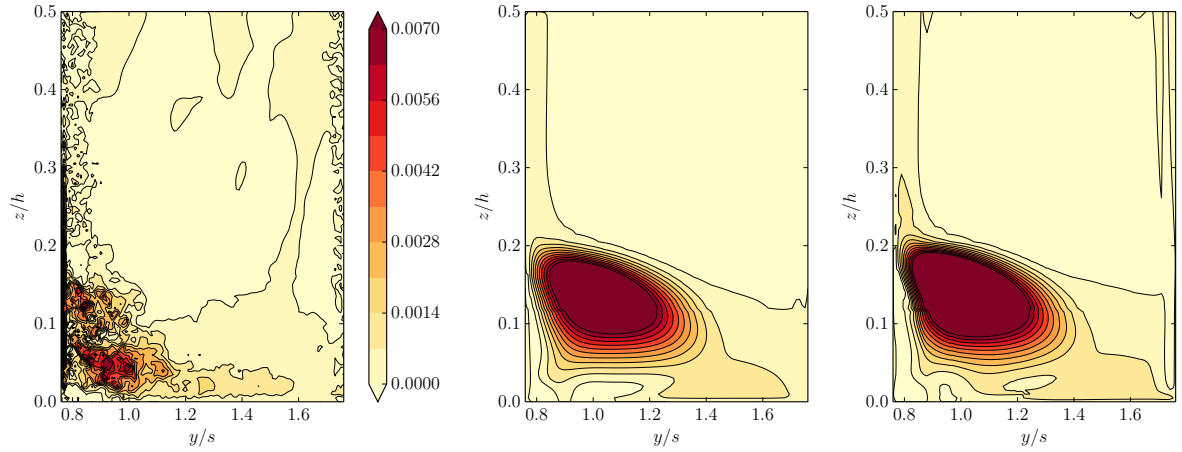



(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 9: $\tau_{t,13}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,13}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)

At the opposite, figure 9 shows that the second Reynolds shear stress is totally over-predicted by both RANS simulations (Fig. 9b and 9c) compared to LES (Fig. 9a). The area of the maximum plotted level is about ten times larger for RANS simulations than for LES, and QCR yields almost no differences from Boussinesq constitutive relation. The Wilcox model fails completely to predict this shear stress.

The third Reynolds shear stress follows the same trends as the second one, as shown in Fig. 10. The area of the maximum plotted level is about ten times larger for RANS simulations (Fig. 10b and 10c) than for LES (Fig. 10a), and QCR yields almost no differences with Boussinesq constitutive relation. Moreover, the stress maximum for both RANS simulations is further from the corner than for the LES.




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 10: $\tau_{t,23}$ for LES (a), for RANS simulation with original Wilcox $k-\omega$ model (b) and $\tau_{t,23}^{QCR}$ for RANS simulation with QCR Wilcox $k-\omega$ model (c)

### 5.3 Synthesis

The comparisons between the RANS simulations and the LES have shown that the RANS simulations lead to a strong over-prediction of the Reynolds stresses compared to LES, except for $\tau_{t,12}$ whose amplitude is under-estimated. The Boussinesq constitutive relation and the QCR yield similar results in many cases, except for the Reynolds normal stress $\tau_{t,33}$ and the Reynolds shear stress $\tau_{t,12}$, where the influence of the QCR is beneficial. It is still not sufficient to correct the Wilcox $k-\omega$ model. The effect of turbulence is badly predicted in both RANS cases. The Reynolds stresses over-prediction can be put in parallel with the over-prediction of the total pressure loss. However, we do not know yet if the over-prediction of the separation leads to the over-prediction of the Reynolds stresses, or if the opposite occurs. The informations gathered do not allow to conclude yet.

## 6 TURBULENT KINETIC ENERGY BUDGET

The turbulent kinetic energy budget shows how the turbulent kinetic energy is created, transported and dissipated. For RANS, it measures the capacity of the turbulence model to represent correctly the physics of the local turbulent kinetic energy equilibrium.

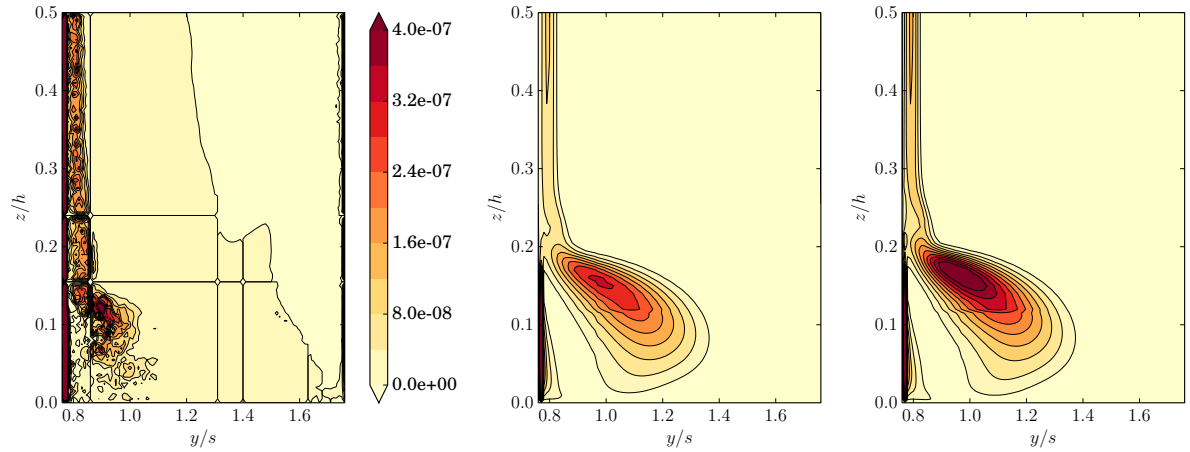
The TKE budget is extracted at the trailing edge of the blade ($x = 1.0c_a$), on the same rectangular surface as for the extraction of the Reynolds stresses. The plane position is given in figure 4.

In the following subsections, the TKE budget terms are normalised by the ratio of the dynamic viscosity on the reference density squared ($\langle\overline{\rho_\infty}\rangle^2$ for LES, $\langle\rho_\infty\rangle^2$ for RANS simulations) times a reference velocity to the fourth power ($[\widetilde{u_\infty}]^4$ for LES, $[u_\infty]^4$ for RANS). The reference densities and velocities are taken $100$mm upstream the leading edge, at a height of $h/4$.

In some of the LES figures, horizontal and vertical lines are visible. These do not represent any physical phenomenon, but are a consequence of the interpolation over many blocks of the computational domain for visualization.

### 6.1 Production

The production term represents the way turbulent kinetic energy is produced. Given its meaning and its expression, it is positive.




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 11: Turbulent kinetic energy production term.

The main area of production is in the separation, close to $z/h = 0.1$ (Fig. 11a). In the RANS simulations, all the Reynolds normal stresses are over-predicted. Consequently, the production term is over-predicted (Fig. 11b and 11c). The maximum is located upper in the span-wise direction for the RANS simulations than for the LES. The effect of the QCR is clearly visible, the intensity of the production term is more important and the area of high production is wider than for the Boussinesq constitutive relation.

### 6.2 Dissipation

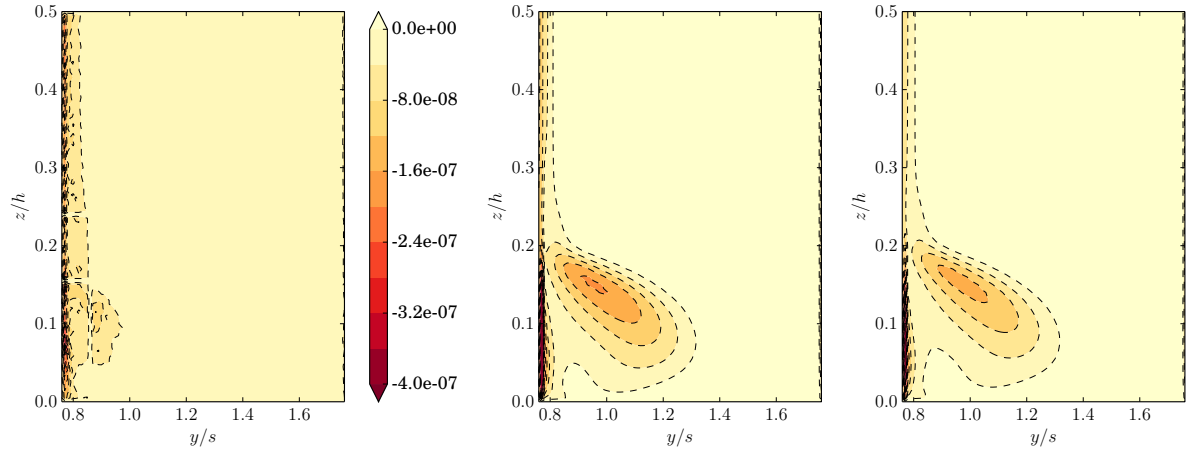
The dissipation term of the TKE budget represents the damping of the turbulent kinetic energy.

For the LES, the major part of the dissipation occurs close to the blade, in the blade boundary-layer (Fig. 12a). A small amount of dissipation occurs also in the area of maximal production, on the corner-separation upper boundary. For the RANS simulations, similar levels of dissipation are observed in the blade boundary layers, but a strong dissipation is observed on the area of maximal production (Fig. 12b and 12c), in the corner separation. This is a known problem of RANS turbulence models, production and dissipation are superposed, with a small amount of transport (Ref. [23]). It is a non-physical behaviour. QCR slightly reduces the intensity of the dissipation term compared to Boussinesq constitutive relation.

### 6.3 Transport

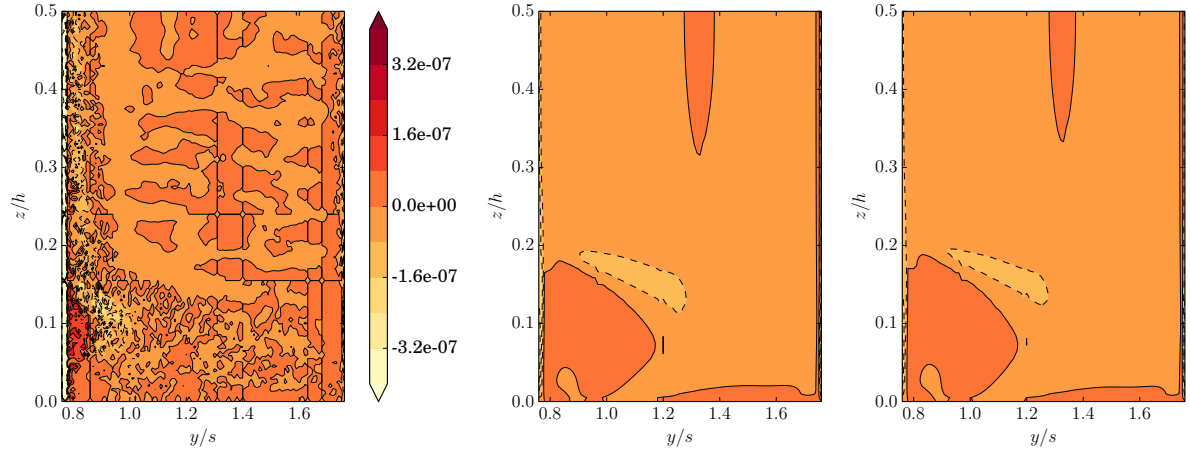
The transport term is constructed to take into account all the transport effects. It can be positive (energy is added) or negative (energy is taken).

The LES transport term shows a physical behaviour (Fig. 13a). The TKE is taken from the area of production to be transported to the dissipative place, mostly close to the blade wall. The RANS models show no significant transport (Fig. 13b and 13c). The Wilcox turbulence model




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 12: Turbulent kinetic energy dissipation term.




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

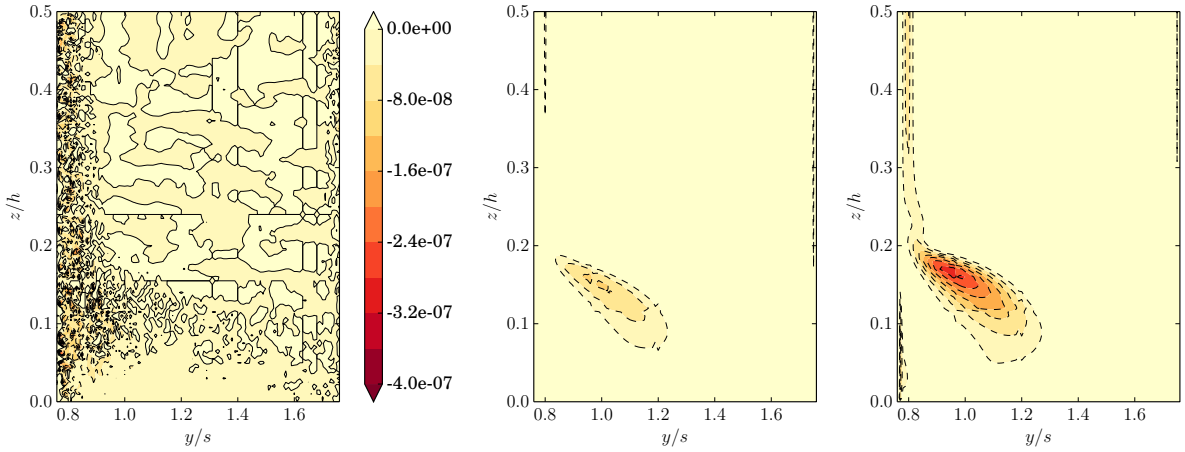
Figure 13: Turbulent kinetic energy transport term.

does not transport turbulent kinetic energy but produces and dissipates it in the same place. The QCR yields no significant effect on the transport term.

### 6.4 Numerical dissipation

The numerical dissipation is a non physical term accounting for the purely numerical effects on the TKE budget. For LES, it gives a measure of the quality of the simulation and budget extraction. For RANS, the modelled TKE equation is directly solved, so the residual should be driven to zero.

The LES numerical dissipation term is not negligible compared to the other budget terms (Fig. 14a) but remains moderate. It is negative and thus corresponds to dissipation. Its minimum is almost four times smaller in absolute value than extrema of production term or dissipation term in absolute value. The RANS simulations (Fig. 14b and 14c) show a numerical dissipation term not negligible, contrary to what was expected. This is not due to a default of the budget reconstruction, but to the use of a production term limiter. The original Wilcox




(a) LES (b) RANS Wilcox 88 (c) RANS Wilcox 88 - QCR

Figure 14: Turbulent kinetic energy numerical dissipation term.

$k-\omega$ model tends to over-predict $k$ production at some points where anisotropy is important, such as leading-edge stagnation point or inside the corner separation. This default is inherent to two-equation models using Boussinesq constitutive relation [24]. To counterpart it, a strategy of $k$ production limitation is implemented in *Turb'Flow*, in the form of a clipping of the production term [25]. Because the clipping is not implemented in the TKE budget extraction, the production term is more intense, and a non-null numerical dissipation term appears. The QCR closure introduces anisotropy in the Reynolds stresses. Far from decreasing the $k$ production limiter effect, it increases it, by increasing the value of production and decreasing the value of dissipation (Fig. 14c).

## 7 CONCLUSIONS

- Large-eddy simulation yields good results according to the experiment, and was used as a reference to analyse RANS simulations.
- Large-eddy simulation yields results physically conclusive for the fine behaviour of the turbulence, for Reynolds stresses and turbulent kinetic energy budget.
- The RANS simulations analysed failed to represent correctly the Reynolds normal and shear stresses. An important over-prediction or under-prediction was observed.
- The QCR closure model yields some modifications, mainly on two terms, $\tau_{t,33}$ and $\tau_{t,12}$, but not significant enough to have results comparable to LES.
- The RANS simulations fail to represent the turbulent kinetic energy budget. The main failure comes from the impossibility to represent correctly the transport of turbulent kinetic energy.
- The QCR tends to activate more the $k$-production limiter in the corner separation than the Boussinesq closure.

## ACKNOWLEDGEMENTS

This study was supported by the Franco-Chinese project VortexFlowCFD, funded by the French National Research and Technology Agency (ANRT) and the SNECMA industry.

The simulations were performed using HPC resources from GENCI-CINES (Grant 2015-2a5039).

## REFERENCES

[1] F. Gao, W. Ma, G. Zambonini, J. Boudet, X. Ottavy, L. Lu, and L. Shao, "Large-eddy simulation of 3-D corner separation in a linear compressor cascade," *Physics of Fluids*, vol. 27, p. 085105, Aug. 2015.

[2] W. Ma, X. Ottavy, L. Lu, F. Leboeuf, and F. Gao, "Experimental Study of Corner Stall in a Linear Compressor Cascade," *Chinese Journal of Aeronautics*, vol. 24, pp. 235–242, June 2011.

[3] A. M. Yocum and W. F. O'Brien, "Separated Flow in a Low-Speed Two-Dimensional Cascade: Part I-Flow Visualization and Time-Mean Velocity Measurements," *Journal of Turbomachinery*, vol. 115, pp. 409–420, July 1993.

[4] A. M. Yocum and W. F. O'Brien, "Separated Flow in a Low-Speed Two-Dimensional Cascade: Part II-Cascade Performance," *Journal of Turbomachinery*, vol. 115, pp. 421–434, July 1993.

[5] W. Ma, X. Ottavy, L. Lu, and F. Leboeuf, "Intermittent corner separation in a linear compressor cascade," *Experiments in Fluids*, vol. 54, pp. 1–17, June 2013.

[6] C. Hah and J. Loellbach, "Development of hub corner stall and its influence on the performance of axial compressor blade rows," *Journal of turbomachinery*, vol. 121, no. 1, pp. 67–77, 1999.

[7] Y. Liu, L. Lu, L. Fang, and F. Gao, "Modification of Spalart–Allmaras model with consideration of turbulence energy backscatter using velocity helicity," *Physics Letters A*, vol. 375, pp. 2377–2381, June 2011.

[8] Y. Liu, H. Yan, L. Fang, L. Lu, Q. Li, and L. Shao, "Modified k-$\omega$ model using kinematic vorticity for corner separation in compressor cascades," *Science China Technological Sciences*, Jan. 2016.

[9] F. Gao, W. Ma, J. Sun, J. Boudet, X. Ottavy, Y. Liu, L. Lu, and L. Shao, "Parameter study on numerical simulation of corner separation in lmfa-naca65 linear compressor cascade," Submitted to *Chinese Journal of Aeronautics*.

[10] M. Bordji, F. Gand, S. Deck, and V. Brunet, "Investigation of a Nonlinear Reynolds-Averaged Navier–Stokes Closure for Corner Flows," *AIAA Journal*, pp. 1–13, Sept. 2015.

[11] P. R. Spalart, "Strategies for turbulence modelling and simulations," *International Journal of Heat and Fluid Flow*, vol. 21, no. 3, pp. 252–263, 2000.

[12] J. Boudet, A. Cahuzac, P. Kausche, and M. C. Jacob, "Zonal Large-Eddy Simulation of a Fan Tip-Clearance Flow, With Evidence of Vortex Wandering," *Journal of Turbomachinery*, vol. 137, p. 061001, June 2015.

[13] F. Gao, G. Zambonini, J. Boudet, X. Ottavy, L. Lu, and L. Shao, "Unsteady behavior of corner separation in a compressor cascade: Large eddy simulation and experimental study," *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, vol. 229, pp. 508–519, Aug. 2015.

[14] G. Zambonini and X. Ottavy, "Unsteady Pressure Investigations of Corner Separated Flow in a Linear Compressor Cascade," p. V02CT44A002, ASME, June 2015.

[15] J. Boudet, J. Caro, L. Shao, and E. Lévêque, "Numerical studies towards practical large-eddy simulation," *Journal of Thermal Science*, vol. 16, pp. 328–336, Nov. 2007.

[16] A. Jameson, W. Schmidt, E. Turkel, and others, "Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes," *AIAA paper*, vol. 1259, p. 1981, 1981.

[17] J. Boudet, J. F. Monier, and F. Gao, "Implementation of a roughness element to trip transition in large-eddy simulation," *Journal of Thermal Science*, vol. 24, pp. 30–36, Feb. 2015.

[18] E. Lévêque, F. Toschi, L. Shao, and J.-P. Bertoglio, "Shear-improved Smagorinsky model for large-eddy simulation of wall-bounded turbulent flows," *Journal of Fluid Mechanics*, vol. 570, p. 491, Jan. 2007.

[19] A. Cahuzac, J. Boudet, P. Borgnat, and E. Lévêque, "Smoothing algorithms for mean-flow extraction in large-eddy simulation of complex turbulent flows," *Physics of Fluids*, vol. 22, no. 12, p. 125104, 2010.

[20] F. Gao, *Advanced numerical simulation of corner separation in a linear compressor cascade*. PhD Thesis, Ecole Centrale de Lyon, Ecully, 2014.

[21] D. Wilcox, C., "Reassessment of the scale-determining equation for advanced turbulence models," *AIAA Journal*, vol. 26, no. 11, pp. 1299 – 1310, 1988.

[22] M. Mani, D. Babcock, C. Winkler, and P. Spalart, "Predictions of a Supersonic Turbulent Flow in a Square Duct," American Institute of Aeronautics and Astronautics, Jan. 2013.

[23] D. C. Wilcox, *Turbulence Modeling for CFD*. D C W Industries, 3rd edition ed.

[24] W. C. STRAHLE, "Stagnation point flows with freestream turbulence - The matching condition," *AIAA Journal*, vol. 23, pp. 1822–1824, Nov. 1985.

[25] F. R. Menter, "Two-equation eddy-viscosity turbulence models for engineering applications," *AIAA Journal*, vol. 32, pp. 1598–1605, Aug. 1994.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A SHALLOW WATER EQUATION BASED ON DISPLACEMENT AND PRESSURE AND THE ZU-CLASS SYMPLECTIC METHOD

**Feng Wu**[1], **Wan-Xie Zhong**[2]

[1] State key laboratory of structural analysis of industrial equipment,
Dalian university of technology, Dalian, Liaoning 116024, P.R. China
wufeng_chn@163.com

[2] State key laboratory of structural analysis of industrial equipment,
Dalian university of technology, Dalian, Liaoning 116024, P.R. China
zwoffice@dlut.edu.cn

**Keywords:** Shallow water equation, Constrained Hamilton variational principle, Wet-dry interface, Sloping water bottom, Zu-class method

**Abstract.** *In this paper, the shallow water problem with sloping water bottom is discussed. By treating the incompressible condition as the constraint, a constrained Hamilton variational principle is presented for the shallow water problem. Based on the constrained Hamilton variational principle, a shallow water equation based on displacement and pressure (SWE-DP) is developed. A hybrid numerical method combining the finite element method for spatial discretization and the Zu-class method for time integration is constructed for the SWE-DP. The correctness and effectiveness of the proposed SWE-DP are verified by using three classical numerical examples. Numerical examples show that the proposed method performs well in the simulation of the shallow water with sloping depths and can correctly deal with wet-dry interfaces.*

## 1 INTRODUCTION

The theory of shallow water is widely applied in engineering such as the offshore engineering and the environment engineering; because of its importance, it has been studied extensively[1-3]. At present, a great number of theoretical models have been proposed for simulating shallow water problems, such as the KdV equation, Saint-Venant equation (SVE) and Boussinesq equation[2, 4-7]. However, all these equations are based on the Euler coordinate. The water equations constructed using the Euler coordinate consist of the convective and source terms. The existence of the convective term make the numerical scheme should be designed carefully. The most widely used method may be the finite volume method[1], because its discrete scheme can preserve the mass and momentum. However the source term often destroy the conservativeness of the finite volume method[2], which results in the harmonious problem of numerical solutions, i.e., the numerical still water cannot keep motionless[3, 4]. The harmonious problem of numerical solutions is one of the two most difficult problems in computational shallow water dynamics. Another one is the wetting/drying condition for shallow water flow which often lead to some numerical problems such as the mass loss and the negative water depth[5].

Generally, an efficient and modern numerical scheme for the nonlinear shallow water problem has to satisfy two requirements: 1) preserve the motionless steady states; 2) allow the occurrence of dry states[3, 4, 6]. However, in the Euler framework it is very difficult to satisfy these two requirements simultaneously. Fortunately, the shallow water problem can also be studied using the Lagrange coordinate, and these two requirement can be satisfied easily in the Lagrange coordinate. In Ref. [9], Tao applied the Lagrange coordinate to discuss the sudden starting of a floating body in deep water. In Ref. [10], Tao and Shi applied the Lagrange coordinate to discuss the problem of hydrodynamic pressure on a suddenly starting vessel. In Ref. [11], Shi employed the Lagrange coordinate to discuss the nonlinear wave induced by the acceleration of a cylindrical tank. In the Lagrangian method, the displacements are viewed as unknown variables. One advantage of the Lagrangian method is that the nonlinear boundary condition on the free surface can be exactly satisfied[9-11]. However, it should be mentioned that in Refs. [9-11], the authors obtained the hydrodynamics equations based on Newton's law rather than the Hamilton variational principle. Undoubtedly, it is important to find a variational principle for the hydrodynamics problem. And it is very hard to find this variational principle in the Euler description. However, it is easy to obtain a Hamilton variational principle by means of the Lagrangian method. Based on the Hamilton variational principle, numerical methods that have been successfully developed and widely applied in structural dynamics, such as the finite element method[12] and symplectic method[13], can be applied to simulate the shallow water problem. The symplectic method preserves the symplecticity and energy of the system and, hence, performs better than the traditional non-symplectic method, especially for problems that require extensive numerical simulation[13].

Recently, Wu and Zhong[7] proposed a constrained Hamilton variational principle for the shallow water problem. In the constrained Hamilton variational principle, the incompressible condition is treated as the constraint, and the pressure is seen as the Lagrangian multiplier. According to the constrained Hamilton variational principle, a shallow water equation based on displacement and pressure (SWE-DP) is developed. To numerically solve the SWE-DP, they developed a hybrid numerical method combining the finite element method for spatial discretization and the Zu-class method for time integration. Their method is symplectic and can preserve the volume numerically. However, in Ref. [7], authors only considered the shallow water with even bottom. The wetting/drying condition and the sloping water bottom was not discussed. In this study, we extend the method proposed in Ref. [7] to the shallow water

with sloping water bottom and wet-dry interface. In Section 2, a SWE-DP for the shallow water flow with sloping bottom and wet-dry interface is constructed. In Section 3, a hybrid numerical method combining the finite element method for spatial discretization and the Zu-class method for time integration is established for the proposed SWE-DP. In Section 4, three classical numerical examples are applied to verify the correction of the proposed SWE-DP and the effectiveness of the proposed numerical method. At last Section, some conclusions are presented.

## 2 HAMILTON VARIATIONAL PRINCIPLE OF SHALLOW WATER WAVES

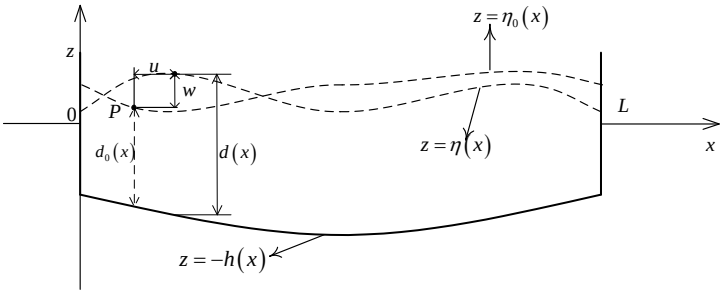



Fig. 1 The considered model

Consider a shallow water domain shown by Fig. 1. The water bed profile is defined by $z=-h(x)$, in which $0\le x\le L$. The initial water surface is $z=\eta_0(x)$. $u(x,z,t)$ and $w(x,z,t)$ represent the displacements of the particle, which is localized at $(x,z)$ initially, at time $t$ in $x$ and $z$, respectively, and the location of the particle at time $t$ is denoted by $(\xi,\zeta)$, in which

$$\xi=x+u(x,z,t),\ \zeta=z+w(x,z,t). \tag{1}$$

According to Ref. [7], the action integral of the two-dimensional water can be shown as

$$S=\int_0^t (T-U+R)\mathrm{d}s \tag{2}$$

where $T$, $U$ and $R$ are kinetic energy, potential energy and constrained term, i.e.,

$$T=\int_0^L\int_{-h(x)}^{\eta_0(x)}\frac{1}{2}\rho(\dot{u}^2+\dot{w}^2)\mathrm{d}x\mathrm{d}z, \tag{3}$$

$$U=\int_0^L\int_{-h(x)}^{\eta_0(x)}\rho g(z+w)\mathrm{d}x\mathrm{d}z, \tag{4}$$

$$R=\int_0^L\int_{-h(x)}^{\eta_0(x)}p\theta\mathrm{d}x\mathrm{d}z, \tag{5}$$

$p$ is the pressure, and $\theta$ is the water volume strain, which can be written as

$$\theta=u_x+w_z+u_xw_z-u_zw_x. \tag{6}$$

For the shallow water flow, it can be assumed that the horizontal displacement is independent of the vertical coordinate $z$, i.e, $u(x,t)=u(x,z,t)$, hence the water volume strain can be simplified as

$$\theta = u_x + w_z + u_x w_z . \tag{7}$$

According to the incompressible condition, the water volume strain should be rigorously zero $\theta = 0$, hence we have

$$\left(1+w_z\right)\left(1+u_x\right)=1, \quad \text{or} \quad w_z\left(x,z\right)=\frac{-u_x\left(x\right)}{1+u_x\left(x\right)}. \tag{8}$$

Equation (8) shows that the vertical displacement distributes linearly along the vertical coordinate $z$. Let the vertical displacements at the bottom and the surface be denoted by

$$w\left(x,-h,t\right)=\varphi\left(x,t\right), \quad w\left(x,0,t\right)=\eta\left(x,t\right), \tag{9}$$

the vertical displacement can be written as:

$$w\left(x,z,t\right)=\frac{z+h}{d_0}\eta\left(x,t\right)+\frac{\eta_0-z}{d_0}\phi\left(x,t\right). \tag{10}$$

Combining Eq. (8) and Eq. (10) gives

$$w_z=-\frac{u_x}{1+u_x}=\frac{\eta\left(x,t\right)-\phi}{h}. \tag{11}$$

At the bottom, $z=-h\left(x\right)$, $w\left(x,-h,t\right)=\varphi\left(x,t\right)$, and the horizontal displacement is $u\left(x,t\right)$, hence

$$\varphi\left(x,t\right)=h\left(x\right)-h\left(x+u\right). \tag{12}$$

Substituting Eq. (12) into Eq. (10) yields

$$w\left(x,z,t\right)=\frac{z+h}{d_0}d-z-h\left(x+u\right) \tag{13}$$

where $d\left(x,t\right)$ is the water depth which can be expressed as

$$d\left(x,t\right)=\eta\left(x,t\right)+h\left(x+u\right)+d_0\left(x\right)-h\left(x\right). \tag{14}$$

Substituting Eq. (13) into Eq. (4), the potential energy is obtained:

$$\begin{aligned} U &= \int_0^L \int_{-h(x)}^{\eta_0(x)} \rho g\left(z+w\right)\mathrm{d}z\mathrm{d}x \\ &= \frac{1}{2}\int_0^L \rho g d_0 d\mathrm{d}x - \int_0^L \rho g d_0 h\left(x+u\right)\mathrm{d}x \end{aligned}. \tag{15}$$

In terms of Eq. (13), the vertical velocity is

$$\dot{w}\left(x,z,t\right)=\frac{z+h}{d_0}\dot{d}\left(x,t\right)-h_x\left(x+u\right)\dot{u}. \tag{16}$$

For the shallow water problem, the nonlinear effect of the vertical velocity can be ignored, hence we have

$$\dot{w}\left(x,z,t\right)=\frac{z+h}{d_0}\dot{d}\left(x,t\right)-h_x\dot{u}. \tag{17}$$

Then substituting Eq. (17) into Eq. (3) yields

$$\begin{aligned} T &= \frac{1}{2}\int_0^L \int_{-h(x)}^{\eta_0(x)} \rho\left(\dot{u}^2 + \dot{w}^2\right) \mathrm{d}x\mathrm{d}z \\ &= \frac{1}{2}\int_0^L \left(\rho d_0 \dot{u}^2\right) \mathrm{d}x + \frac{1}{2}\int_0^L \rho\left(\frac{d_0}{3}\dot{d}^2 - d_0 h_x \dot{u}\dot{d} + d_0 h_x^2 \dot{u}^2\right) \mathrm{d}x \end{aligned} . \tag{18}$$

The pressure at the water surface is zero. Let the pressure at the bottom be denoted by

$$p\left(x, -h(x), t\right) = \beta\left(x,t\right). \tag{19}$$

We assume that the pressure distributes linearly along the vertical coordinate $z$, i.e.,

$$p\left(x,z,t\right) = \frac{\eta_0 - z}{d_0}\beta\left(x,t\right). \tag{20}$$

Substituting Eq. (13) into Eq. (7), the water volume strain can be written as

$$\theta = u_x + \left(1+u_x\right)\left(\frac{d}{d_0} - 1\right). \tag{21}$$

Substituting Eqs. (20) and (21) into Eq. (5) yields

$$R = \frac{1}{2}\int_0^L \beta\left[d\left(1+u_x\right) - d_0\right] \mathrm{d}x. \tag{22}$$

In terms of Eqs. (15), (18) and (22), the action integral of the shallow water wave can be written as

$$\begin{aligned} S = {} & \frac{1}{2}\int_0^t \int_0^L \rho\left(d_0\dot{u}^2 + \frac{d_0}{3}\dot{d}^2 - d_0 h_x \dot{u}\dot{d} + d_0 h_x^2 \dot{u}^2\right) \mathrm{d}x\mathrm{d}s \\ & - \int_0^t \frac{1}{2}\int_0^L \rho g d_0 d \,\mathrm{d}x\mathrm{d}s + \int_0^t \int_0^L \rho g d_0 h\left(x+u\right) \mathrm{d}x\mathrm{d}s \\ & + \frac{1}{2}\int_0^t \int_0^L \beta\left[d\left(1+u_x\right) - d_0\right] \mathrm{d}x\mathrm{d}s \end{aligned} . \tag{23}$$

Taking the first variation of the action integral, i.e., $\delta S = 0$, gives

$$\begin{aligned} \delta S = {} & \frac{1}{2}\int_0^t \int_0^L \rho\left(-2\delta u d_0 \ddot{u} - 2\delta d \frac{d_0}{3}\ddot{d} + d_0 h_x \delta u \ddot{d} + d_0 h_x \ddot{u}\delta d - 2\delta u d_0 h_x^2 \ddot{u}\right) \mathrm{d}x\mathrm{d}s \\ & - \int_0^t \frac{1}{2}\int_0^L \rho g d_0 \delta d \,\mathrm{d}x\mathrm{d}s + \int_0^t \int_0^L \rho g d_0 h_x \left(x+u\right)\delta u \,\mathrm{d}x\mathrm{d}s \\ & + \frac{1}{2}\int_0^t \int_0^L \beta\delta d\left(1+u_x\right) \mathrm{d}x\mathrm{d}s + \frac{1}{2}\int_0^t \int_0^L \delta\beta\left[d\left(1+u_x\right) - d_0\right] \mathrm{d}x\mathrm{d}s \\ & - \frac{1}{2}\int_0^t \int_0^L \frac{\partial\left(\beta d\right)}{\partial x}\delta u \,\mathrm{d}x\mathrm{d}s + \frac{1}{2}\int_0^t \beta d\,\delta u\Big|_{x=0}^{x=L} \mathrm{d}s \end{aligned} . \tag{24}$$

At the boundaries $x = 0$ and $L$, the displacement are often known or the water depths are rigorously zeros, hence, we have

$$d\left(x\right) = 0 \text{ or } \delta u\left(x\right) = 0, \quad x = 0, L \ . \tag{25}$$

Combining Eq. (24) with Eq.(25) yields

$$\begin{cases} \rho d_0 \ddot{u} - \dfrac{1}{2}\rho d_0 h_x \ddot{d} + \rho d_0 h_x^2 \ddot{u} - \rho g d_0 h_x \left(x+u\right) + \dfrac{1}{2}\dfrac{\partial\left(\beta d\right)}{\partial x} = 0 \\ \rho \dfrac{d_0}{3}\ddot{d} - \dfrac{1}{2}\rho d_0 h_x \ddot{u} + \rho g d_0 - \dfrac{1}{2}\beta\left(1+u_x\right) = 0 \\ d\left(1+u_x\right) = d_0 \end{cases} \tag{26}$$

which is the shallow water equation based on displacement and pressure (SWE-DP).

## 3 NUMERICAL TREATMENT

It is necessary to develop a numerical method for the SWE-DP. In this section, a hybrid method combining the finite element for the spatial discretization and Zu-class method for the time integration is presented in detail.

### 3.1 Discretization in space

The proposed SWE-DP (26) is derived in terms of the constrained Hamilton variational principle. Hence, it is a natural choice to use the finite element method for spatial discretization. Let the region $\left[0, L\right]$ be divided into $N_{\mathrm{e}}$ basic elements with $N_{\mathrm{d}}$ nodes; see Fig. 2.




Fig. 2 The finite element mesh

On the *n*th element, the horizontal displacement $u\left(x\right)$ is approximated by the linear function, and $d\left(x\right)$ and $\beta\left(x\right)$ are treated as constant values, i.e.,

$$u\left(x\right) = \frac{x_{n+1}-x}{\Delta x_n}u_n + \frac{x-x_n}{\Delta x_n}u_{n+1},\ d\left(x\right) = d_n,\ \beta\left(x\right) = \beta_n, \tag{27}$$

where $x_n$ is the node location in $x$, $\Delta x_n$ is the length of the *n*th element, $u_n$ is the node horizontal displacement, $d_n$ is the water depth evaluated at the mid-point of the *n*th element, and $\beta_n$ is the pressure of the water bottom evaluated at the mid-point of the *n*th element. In terms of Eq. (27), the incompressible condition on the *n*th element can be approximated as

$$d_n \Delta x_n \left(1 + \frac{u_{n+1}-u_n}{\Delta x_n}\right) = \Delta x_n d_{0,n}. \tag{28}$$

The kinetic energy can be approximated as

$$T = \frac{1}{2}\dot{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{M}_u\dot{\boldsymbol{u}} + \frac{1}{2}\dot{\boldsymbol{d}}^{\mathrm{T}}\boldsymbol{M}_d\dot{\boldsymbol{d}} - \dot{\boldsymbol{d}}^{\mathrm{T}}\boldsymbol{M}_{du}\dot{\boldsymbol{u}} \tag{29}$$

in which

$$\rho_u = \rho d_0\left(1+h_x^2\right),\quad \rho_d = \frac{\rho d_0}{3},\quad \rho_{ud} = \rho h_x d_0, \tag{30}$$

$$\begin{aligned}
&\boldsymbol{u} = \begin{pmatrix} u_1 & u_2 & \cdots & u_{N_\mathrm{d}} \end{pmatrix}^{\mathrm{T}},\quad \boldsymbol{d} = \begin{pmatrix} d_1 & d_2 & \cdots & d_{N_\mathrm{e}} \end{pmatrix}^{\mathrm{T}} \\
&\boldsymbol{M}_u = \operatorname{diag}\begin{Bmatrix} M_{u,1} & \cdots & M_{u,n} & \cdots & M_{u,N_\mathrm{d}} \end{Bmatrix} \\
&\boldsymbol{M}_d = \operatorname{diag}\begin{Bmatrix} \rho_{d,1}\Delta x_1 & \rho_{d,2}\Delta x_2 & \cdots & \rho_{d,N_\mathrm{e}}\Delta x_{N_\mathrm{e}} \end{Bmatrix} \\
&\boldsymbol{M}_{du} = \frac{1}{4}\begin{bmatrix} \Delta x_1\rho_{ud,1} & \Delta x_1\rho_{ud,1} & 0 & \cdots & 0 \\ 0 & \Delta x_2\rho_{ud,2} & \Delta x_2\rho_{ud,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Delta x_{N_\mathrm{e}}\rho_{ud,N_\mathrm{e}} & \Delta x_{N_\mathrm{e}}\rho_{ud,N_\mathrm{e}} \end{bmatrix},
\end{aligned} \tag{31}$$

and

$$\begin{aligned}
&M_{u,1} = \frac{\Delta x_1\rho_{u,1}}{2},\quad M_{u,n} = \frac{\Delta x_{n-1}\rho_{u,n-1}+\Delta x_n\rho_{u,n}}{2},\quad M_{u,N_\mathrm{d}} = \frac{\Delta x_{N_\mathrm{e}}\rho_{u,N_\mathrm{e}}}{2} \\
&\rho_{u,n} = \rho_u\left(\frac{x_n+x_{n+1}}{2}\right),\quad \rho_{d,n} = \rho_d\left(\frac{x_n+x_{n+1}}{2}\right),\quad \rho_{ud,n} = \rho_{ud}\left(\frac{x_n+x_{n+1}}{2}\right).
\end{aligned} \tag{32}$$

The potential energy can be approximated as

$$U = \boldsymbol{g}_d^{\mathrm{T}}\boldsymbol{d} - \boldsymbol{g}_h^{\mathrm{T}}\boldsymbol{h} \tag{33}$$

where

$$\begin{aligned}
&\boldsymbol{g}_h^{\mathrm{T}} = \left(\frac{\Delta x_1}{2}g_{u,1}, \left(\frac{\Delta x_n}{2}+\frac{\Delta x_{n-1}}{2}\right)g_{u,n}, \frac{\Delta x_{N_e}}{2}g_{u,N_d}\right)^{\mathrm{T}} \\
&\boldsymbol{g}_d^{\mathrm{T}} = \left(\frac{\left(g_{d,1}+g_{d,2}\right)\Delta x_1}{2} \quad \frac{\left(g_{d,2}+g_{d,3}\right)\Delta x_2}{2} \quad \cdots \quad \frac{\left(g_{d,N_\mathrm{d}-1}+g_{d,N_\mathrm{d}}\right)\Delta x_{N_\mathrm{e}}}{2}\right)^{\mathrm{T}}. \\
&\boldsymbol{h}^{\mathrm{T}} = \left(h\left(x_1+u_1\right), h\left(x_n+u_n\right), h\left(x_{N_d}+u_{N_d}\right)\right)^{\mathrm{T}} \\
&g_{d,n} = \frac{\rho g}{2}d_0\left(x_n\right),\quad g_{u,n} = \rho g d_0\left(x_n\right)
\end{aligned} \tag{34}$$

The Lagrangian modified term can be approximated as：

$$R = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Delta d} + \boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Lambda C u} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{s} = \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Delta d} + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{D C u} - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{s} \tag{35}$$

in which

$$
\begin{aligned}
&\boldsymbol{s}^{\mathrm{T}} = \left(\Delta x_1 d_{0,1} \quad \Delta x_n d_{0,2} \quad \Delta x_{N_e} d_{0,N_e}\right) \\
&\boldsymbol{D} = \mathrm{diag}\left\{\boldsymbol{d}\right\}, \quad \boldsymbol{\Delta} = \mathrm{diag}\left\{\Delta x_1, \Delta x_2, \cdots\right\} \\
&\boldsymbol{\beta}^{\mathrm{T}} = \left(\beta_1 \quad \beta_2 \quad \ldots \quad \beta_{N_e}\right), \quad \boldsymbol{B} = \mathrm{diag}\left\{\boldsymbol{\beta}\right\} \\
&\boldsymbol{C} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -1 & 1 \end{bmatrix}
\end{aligned}
\tag{36}
$$

In terms of Eqs. (29), (33) and (35), the action integral can be approximated as

$$
S = \int_0^t \begin{bmatrix} \frac{1}{2}\dot{\boldsymbol{u}}^{\mathrm{T}}\boldsymbol{M}_u\dot{\boldsymbol{u}} + \frac{1}{2}\dot{\boldsymbol{d}}^{\mathrm{T}}\boldsymbol{M}_d\dot{\boldsymbol{d}} - \dot{\boldsymbol{d}}^{\mathrm{T}}\boldsymbol{M}_{du}\dot{\boldsymbol{u}} - \boldsymbol{g}_d^{\mathrm{T}}\boldsymbol{d} + \boldsymbol{g}_h^{\mathrm{T}}\boldsymbol{h} \\ + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{d} + \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{C}\boldsymbol{u} - \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{s} \end{bmatrix} \mathrm{d}\tau , \tag{37}
$$

in which $\boldsymbol{u}$ and $\boldsymbol{d}$ are the horizontal displacement vector and the water depth vector, respectively, and $\boldsymbol{\beta}$ is the pressure vector.

In the integrand of Eq. (37), the summation of the first three terms represents the kinetic energy, the summation of the fourth and fifth terms represents the potential energy and the summation of the last three terms represents the constrained term. Calculating the first variation of $S$ gives

$$
\begin{cases} \boldsymbol{M}_u\ddot{\boldsymbol{u}} - \boldsymbol{M}_{du}^{\mathrm{T}}\ddot{\boldsymbol{d}} - \boldsymbol{h}_x\boldsymbol{g}_h - 0.5\boldsymbol{C}^T\boldsymbol{B}\boldsymbol{d} = 0 \\ \boldsymbol{M}_d\ddot{\boldsymbol{d}} - \boldsymbol{M}_{du}\ddot{\boldsymbol{u}} + \boldsymbol{g}_d - 0.5\boldsymbol{\Delta}\boldsymbol{\beta} - 0.5\boldsymbol{B}\boldsymbol{C}\boldsymbol{u} = 0 \\ \boldsymbol{\Delta}\boldsymbol{d} + \boldsymbol{D}\boldsymbol{C}\boldsymbol{u} - \boldsymbol{s} = 0 \end{cases} \tag{38}
$$

which constitutes a system of differential algebraic equations (DAES).

### 3.2 The Zu-class method

The nonlinear DAES (38) is obtained from the first variation of Eq. (37), which corresponds to a constrained Hamilton system. For the Hamilton system, the symplecticity is a characteristic property. As the symplectic method can preserve the property of symplecticity and the approximate energy necessary for a long time computation, it is often used to simulate the Hamilton dynamical system. However, for the constrained Hamilton system, it is required for a time integration method to preserve not only the symplectic structure of the Hamilton system but also all the constraints. In Ref [8], a method preserving all the constraints was developed by Zhong et al. The numerical experiment about the double pendulum presented in [8] shows this method can preserve energy well. In Ref [9] it was named as the Zu-type method. In Ref [10], the Zu-type method was proved to be symplectic. In the Zu-type method, the constraint conditions are satisfied strictly at the integration points, and the orbit is treated as the geodesic in the state space and is, therefore, approximated by using the time finite element method. In this sub-section, the Zu-class method will be employed to solve the DAES(38). Let the time domain be discretized as

$$
t = t_0, t_1, \cdots, t_k, \cdots, \qquad t_k = k \times \Delta t , \tag{39}
$$

where $\Delta t$ is the time step. Let the velocities in $[t_k, t_{k+1}]$ be approximated as

$$\dot{\boldsymbol{u}}_k(t)=\frac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t},\quad \dot{\boldsymbol{d}}_k(t)=\frac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t},\quad t\in[t_k,t_{k+1}], \tag{40}$$

where $\#_k=\#(t_k)$, and the displacements be approximated as

$$\boldsymbol{u}(t)=\frac{\boldsymbol{u}_{k+1}+\boldsymbol{u}_k}{2},\quad \boldsymbol{d}(t)=\frac{\boldsymbol{d}_{k+1}+\boldsymbol{d}_k}{2},\quad t\in[t_k,t_{k+1}]. \tag{41}$$

The pressure is approximated to be a constant, i.e.,

$$\boldsymbol{\beta}(t)=\boldsymbol{\beta}_k,\ t\in[t_k,t_{k+1}]. \tag{42}$$

By substituting Eqs. (40)-(42) into Eq. (37), the kinetic energy, potential energy and constrained term can be approximated as

$$\begin{aligned}
T&=\frac{1}{2}\left(\boldsymbol{u}_{k+1}^{\mathrm{T}}-\boldsymbol{u}_k^{\mathrm{T}}\right)\boldsymbol{M}_u\left(\frac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t}\right)+\frac{1}{2}\left(\boldsymbol{d}_{k+1}^{\mathrm{T}}-\boldsymbol{d}_k^{\mathrm{T}}\right)\boldsymbol{M}_d\frac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t}\\
&\quad-\left(\boldsymbol{d}_{k+1}^{\mathrm{T}}-\boldsymbol{d}_k^{\mathrm{T}}\right)\boldsymbol{M}_{du}\left(\frac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t}\right)\\
U&=\frac{\Delta t}{2}\boldsymbol{g}_d^{\mathrm{T}}\left(\boldsymbol{d}_k+\boldsymbol{d}_{k+1}\right)-\frac{\Delta t}{2}\boldsymbol{g}_h^{\mathrm{T}}\left[\boldsymbol{h}(\boldsymbol{u}_k)+\boldsymbol{h}(\boldsymbol{u}_{k+1})\right]\\
R&=\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Delta}\frac{\Delta t}{2}\left(\boldsymbol{d}_k+\boldsymbol{d}_{k+1}\right)+\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\frac{\Delta t}{2}\left(\boldsymbol{D}_k\boldsymbol{C}\boldsymbol{u}_k+\boldsymbol{D}_{k+1}\boldsymbol{C}\boldsymbol{u}_{k+1}\right)-\frac{1}{2}\Delta t\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{s}
\end{aligned}. \tag{43}$$

Substituting Eq. (43) into the action integral and taking the first variation gives

$$\begin{cases}
\dfrac{\boldsymbol{M}_u}{\Delta t}\left(\boldsymbol{u}_{k+1}-\boldsymbol{u}_k\right)-\boldsymbol{M}_{du}^{\mathrm{T}}\left(\dfrac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t}\right)+\dfrac{\Delta t}{2}\boldsymbol{h}_{\boldsymbol{u}_{k+1}}\boldsymbol{g}_h+\dfrac{\Delta t}{4}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}_{k+1}\boldsymbol{\beta}-\boldsymbol{p}_{u,k+1}=\boldsymbol{0}\\
-\dfrac{\boldsymbol{M}_u}{\Delta t}\left(\boldsymbol{u}_{k+1}-\boldsymbol{u}_k\right)+\boldsymbol{M}_{du}^{\mathrm{T}}\left(\dfrac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t}\right)+\dfrac{\Delta t}{2}\boldsymbol{h}_{\boldsymbol{u}_k}\boldsymbol{g}_h+\dfrac{\Delta t}{4}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{D}_k\boldsymbol{\beta}+\boldsymbol{p}_{u,k}=\boldsymbol{0}\\
\boldsymbol{M}_d\dfrac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t}-\boldsymbol{M}_{du}\left(\dfrac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t}\right)-\dfrac{\Delta t}{2}\boldsymbol{g}_d+\dfrac{\Delta t}{4}\boldsymbol{\Delta}\boldsymbol{\beta}+\dfrac{\Delta t}{4}\boldsymbol{B}\boldsymbol{C}\boldsymbol{u}_{k+1}-\boldsymbol{p}_{d,k+1}=\boldsymbol{0},\\
-\boldsymbol{M}_d\dfrac{\boldsymbol{d}_{k+1}-\boldsymbol{d}_k}{\Delta t}+\boldsymbol{M}_{du}\left(\dfrac{\boldsymbol{u}_{k+1}-\boldsymbol{u}_k}{\Delta t}\right)-\dfrac{\Delta t}{2}\boldsymbol{g}_d+\dfrac{\Delta t}{4}\boldsymbol{\Delta}\boldsymbol{\beta}+\dfrac{\Delta t}{4}\boldsymbol{B}\boldsymbol{C}\boldsymbol{u}_k+\boldsymbol{p}_{d,k}=\boldsymbol{0}\\
\boldsymbol{\Delta}\boldsymbol{d}_{k+1}+\boldsymbol{D}_{k+1}\boldsymbol{C}\boldsymbol{u}_{k+1}-\boldsymbol{s}=\boldsymbol{0}
\end{cases} \tag{44}$$

where $\boldsymbol{p}_u$ and $\boldsymbol{p}_d$ are the momentum vectors

$$\boldsymbol{p}_u=\boldsymbol{M}_u\dot{\boldsymbol{u}}-\boldsymbol{M}_{du}^{\mathrm{T}}\dot{\boldsymbol{d}},\quad \boldsymbol{p}_d=\boldsymbol{M}_d\dot{\boldsymbol{d}}-\boldsymbol{M}_{du}\dot{\boldsymbol{u}}. \tag{45}$$

Eq. (44) is a system of nonlinear algebraic equations that can be solved by using the Newton iteration method. If $\boldsymbol{M}_d$ and $\boldsymbol{M}_{du}$ in Eq. (44) are replaced with $\boldsymbol{M}_d\varepsilon$ and $\boldsymbol{M}_{du}\varepsilon$ ($\varepsilon\ll 1$), Eq. (44) can be used to analyze the shallow water problem ignoring the effect of the vertical acceleration. In this case, the results are in agreement with the SVE solution.

## 4 NUMERICAL EXAMPLES

In this section, three classical numerical examples, which have been used frequently by many researchers to test different numerical methods[11-13], are used to verify the correction of the proposed SWE-DP and the effectiveness of the proposed numerical method.

### 4.1 Evolution of shorelines over a parabolic topography

Analytical solutions of the nonlinear shallow water equations were derived by Sampson et al.[11] for a perturbed flow in 1-dimensional container with a parabolic bed profile. This provides a perfect test for the present method in dealing with bed sloping source term and wetting/drying condition. The initial water domain is shown by Fig. 3.

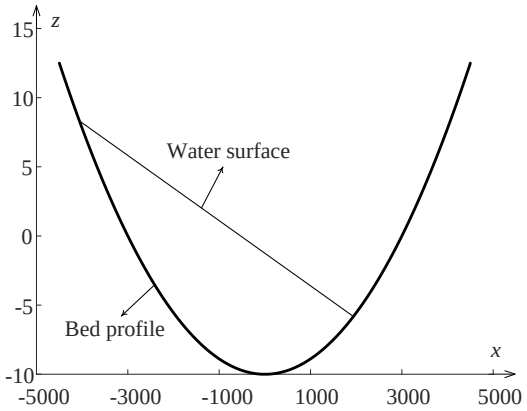



Fig.3 The water profile at $t = 0$ s

The Bed profile is defined by

$$h(x) = h_0 - h_0\left(\frac{x}{a}\right)^2 \tag{46}$$

where $h_0 = 10$ m, $a = 3000$ m. The analytical horizontal displacement and water surface are shown as

$$u(x,t) = \frac{a^2}{2gh_0} B\omega\left(1-\cos\omega t\right) \tag{47}$$

and

$$\eta(x,t) = -\frac{1}{g} B\omega\cos\omega t\left(x+u(x,t)\right) - \frac{a^2B^2}{8g^2h_0}\omega^2\cos 2\omega t - \frac{B^2}{4g}, \quad \omega = \frac{\sqrt{2gh_0}}{a}, \tag{48}$$

respectively, in which $B = 5\,\mathrm{m/s}$, $g = 10\,\mathrm{m/s^2}$, $t \in [0\,\mathrm{s}, 6000\,\mathrm{s}]$.

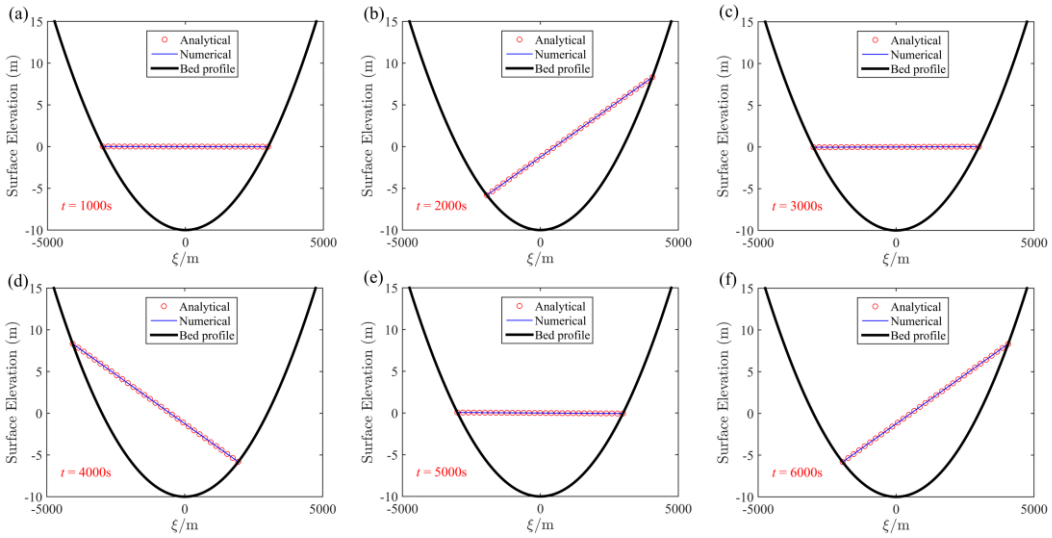



Fig.4 Water surface elevation at different times: (a) $t=1000\,\mathrm{s}$; (b) $t=2000\,\mathrm{s}$; (c) $t=3000\,\mathrm{s}$; (d) $t=4000\,\mathrm{s}$; (e) $t=5000\,\mathrm{s}$; (f) $t=6000\,\mathrm{s}$;

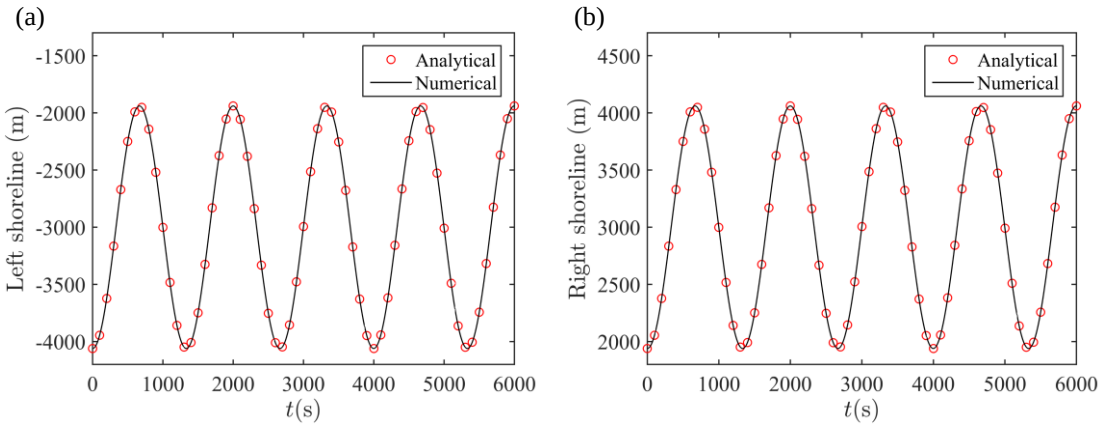



Fig.5 Location of shorelines: (a) Left shoreline; (b) Right shoreline.

Numerical simulation is performed with a time step $\Delta t=1\mathrm{s}$ and a uniform grid with $N_{\mathrm{e}}=300$ elements. Fig. 4 shows the predicted water surface elevation at different times. Fig. 5 shows the predicted locations of left and right shorelines. Excellent agreements are observed between the numerical prediction and analytical solutions from Figs. 4-5, which validates that the proposed method is able to deal with the shallow water with complex bed topographies and wet-dry interfaces.

## 4.2 Spreading of a drop of shallow water

Consider the spreading of a parabola-shaped two-dimensional drop of shallow water on a horizontal plane. The drop is initially confined to $|x|<1$ according to

$$z=\eta_0(x)=\begin{cases}1-x^2 & \text{for } |x|<1\\ 0 & \text{for } |x|\geq 1\end{cases}, \tag{49}$$

and at rest. Upon releasing the drop, it spreads under the effect of gravity. The gravitational acceleration is $g=1\,\mathrm{m/s^2}$. The temporal evolution of this system has been analytically inves-

tigated by Frei[14]. He noted that the parabolic shape is always retained and that the velocity across the drop is a linear function, i.e.,

$$\eta(x,t)=\lambda^{-1}\left[1-\left(\frac{\xi}{\lambda}\right)^2\right],\ \dot{u}(x)=\xi\left(\frac{\lambda_t}{\lambda}\right) \tag{50}$$

in which, $\xi$ is defined by Eq. (1), $\lambda$ describes the half-width of the drop, and $\lambda_t=\partial\lambda/\partial t$ is the velocity of the leading edge. Following Refs. [14, 15], the $\lambda(t)$ is obtained numerically as the root of the follow equation

$$t=\frac{1}{2}\left[\sqrt{\lambda(\lambda-1)}+\ln\left(\sqrt{\lambda-1}+\sqrt{\lambda}\right)\right], \tag{51}$$

and

$$\lambda_t=2\sqrt{1-\lambda^{-1}}\,. \tag{52}$$

The proposed SWE-DP is used to model this problem, and the numerical simulation is performed with $N_e=100$ elements. The time step is selected as $\Delta t=0.01\,\mathrm{s}$, and the numerical simulation lasts for $2\,\mathrm{s}$. To ignore the effect of the vertical acceleration, $\boldsymbol{M}_d$ and $\boldsymbol{M}_{du}$ in Eq. (44) are replaced with $0.001\boldsymbol{M}_d$ and $0.001\boldsymbol{M}_{du}$, respectively. The analytical solutions and the numerical results computed using the proposed method are compared in Figs. 6-7. The example used here was also discussed in Ref. [16] where the SVE was used to model the drop and the second order high resolution algorithm was applied to solve the SVE. The numerical velocities given in Ref. [16] is also displayed in Fig. 7.

It can be observed from Figs. 6-7 that the numerical results computed using the proposed method are in excellent agreement with the analytical solution. However the numerical velocities given in Ref. [16] are different from the analytical solutions at the left and right moving wet-dry interfaces. Hence it can be concluded that the proposed model can correctly handle the problem of wet-dry interfaces, and perform better than the SVE.

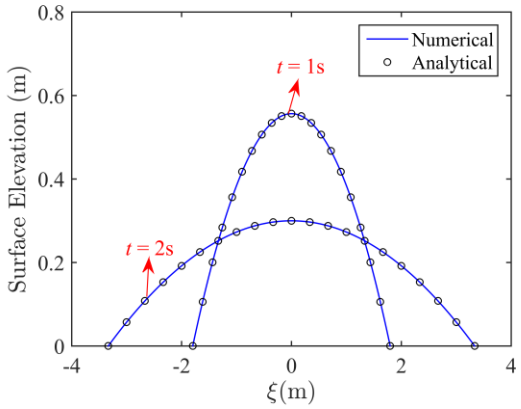



Fig.6 Water free surface

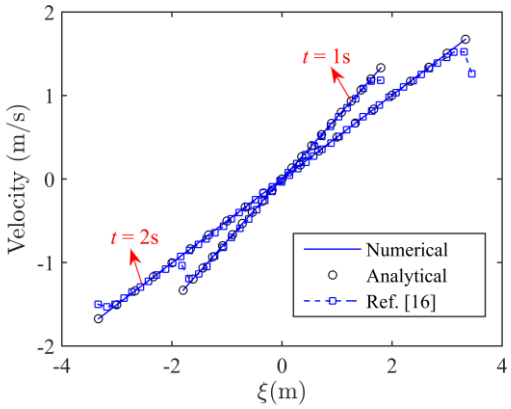



Fig.7 Comparison of velocity

### 4.3 Dam break on dry bed

Consider a reservoir with length 150 m and water depth 10 m; see Fig. 8. Suppose that the right dam is broken suddenly and the water pour out of the reservoir. The right side of the dam is the dry bed. Initially, the water surface is defined as

$$z = \eta_0(x) = \begin{cases} h_0, & \text{if } x \le x_0 \\ 0, & \text{if } x > x_0 \end{cases} \tag{53}$$

where $x_0 = 150\,\text{m}$ and $h_0 = 10\,\text{m}$. The initial velocity is $\dot{u}(x,0) = 0\,\text{m/s}$. The gravitational acceleration is $g = 9.81\,\text{m/s}$.

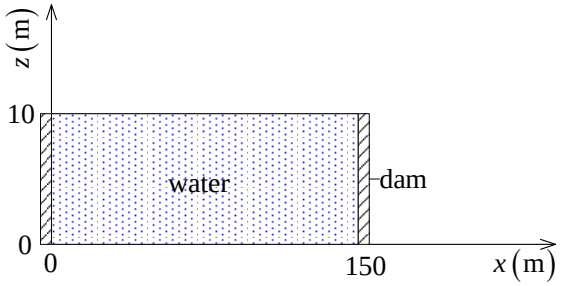



Fig. 8 The dam

The analytical solution of this problem is given as[13]

$$\eta(x,t) = \begin{cases} h_0, & \text{if } \xi \le x_0 - t\sqrt{gh_0} \\ \dfrac{1}{9g}\left(2\sqrt{gh_0} - \dfrac{\xi - x_0}{t}\right)^2, & \text{if } x_0 - t\sqrt{gh_0} < \xi \le x_0 + 2t\sqrt{gh_0} \\ 0, & \text{if } \xi > x_0 + 2t\sqrt{gh_0} \end{cases} \tag{54}$$

and

$$\dot{u}(x,t) = \begin{cases} 0, & \text{if } \xi \le x_0 - t\sqrt{gh_0} \\ \dfrac{2}{3}\left(\sqrt{gh_0} + \dfrac{\xi - x_0}{t}\right), & \text{if } x_0 - t\sqrt{gh_0} < \xi \le x_0 + 2t\sqrt{gh_0} \\ 0, & \text{if } \xi > x_0 + 2t\sqrt{gh_0} \end{cases} \tag{55}$$

where $\xi$ is defined by Eq. (1). Numerical simulation is performed with a time step $\Delta t = 0.01\,\text{s}$ and a uniform grid with $N_e = 300$ elements. The numerical simulation lasts for 6 s. To ignore the effect of the vertical acceleration, $\boldsymbol{M}_d$ and $\boldsymbol{M}_{du}$ in Eq. (44) are replaced with $0.001\boldsymbol{M}_d$ and $0.001\boldsymbol{M}_{du}$, respectively. Fig. 9 shows the predicted water surface elevation at $t = 6\,\text{s}$. Fig. 10 shows the predicted discharges $\eta\dot{u}$ at $t = 6\,\text{s}$. Excellent agreements are observed between the numerical prediction and analytical solutions from Figs. 9-10, which validates that the proposed method is able to deal with the shallow water with wet-dry interfaces.

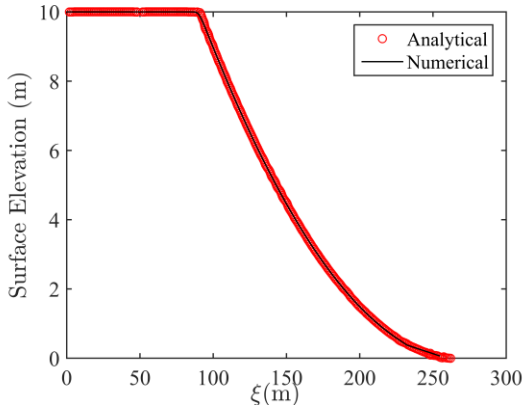



Fig. 9 Water surface at 6 s

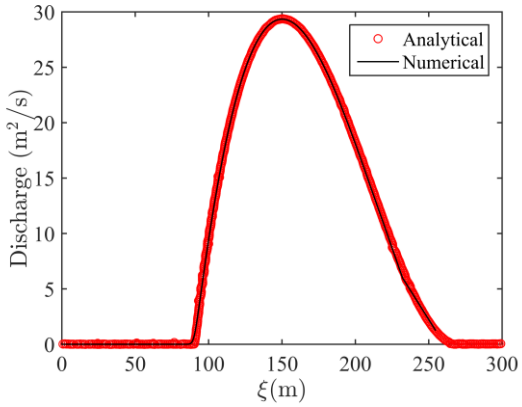



Fig. 10 Water discharge at 6 s

## 5 CONCLUSIONS

A constrained Hamilton variational principle for shallow water with sloping bottom is developed. Based on the constrained Hamilton variational principle, a shallow water equation based on displacement and pressure (SWE-DP) is developed. A hybrid numerical method combining the finite element method for spatial discretization and the Zu-class method for time integration is constructed for the SWE-DP. Three numerical examples are used to test correctness of the proposed SWE-DP and the effectiveness of the hybrid numerical method proposed for the SWE-DP. Numerical examples show that the proposed method performs well in the simulation of the shallow water with sloping depths and wet-dry interfaces.

In this paper, we only consider 1-dimensional shallow water. However, the basic ideas could be expanded to other hydrodynamics problems. We believe that the displacement method will play a major role in accurately and efficiently simulating hydrodynamics problems. In our next work, the proposed method will be expanded to shallow water with two dimensions.

## 6 ACKNOWLEDGEMENTS

The authors are grateful for the support of the Natural Science Foundation of China (No. 11472067).

## REFERENCES

[1] Y. Guo, R. Liu, Y. Duan, Y. Li, A characteristic-based finite volume scheme for shallow water equations. *Journal of Hydrodynamics*, **21**, 531-540, 2009.

[2] L.H. Bo, S. Jin, Study on high resolutio scheme for shallow water equation with source terms. *Journal of Hydrodinamics Ser. A*, **24**, 22-28, 2009.

[3] P. Ortiz, Non-oscillatory continuous FEM for transport and shallow water flows. *Computer Methods in Applied Mechanics and Engineering*, **223-224**, 55-69, 2012.

[4] Q. Liang, F. Marche, Numerical resolution of well-balanced shallow water equations with complex source terms. *Advances in Water Resources*, **32**, 873-884, 2009.

[5] P. Brufau, P. Garcia-Navarro, M.E. Vázquez-Cendón, Zero mass error using unsteady wetting-drying conditions in shallow flows over dry irregular topography. *International*

*Journal for Numerical Method in Fluids*, **45**, 1047-1082, 2004.

[6] A. Duran, Q. Liang, F. Marche, On the well-balanced numerical discretization of shallow water equations on unstructured meshes. *Journal of Computational Physics*, **235**, 565-586, 2013.

[7] F. Wu, W.X. Zhong, Constrained Hamilton variational principle for shallow water problems and Zu-class symplectic algorithm. *Applied Mathematics and Mechanics (English Edition)*, **1**, 1-14, 2016.

[8] W.X. Zhong, Q. Gao, Integration of constrained dynamical system via analytical structural mechanics. *Journal of Dynamics and Control*, **4**, 193-200, 2006.

[9] W.X. Zhong, Q. Gao, H.J. Peng, *Classical Mechanics - Its symplectic description*. Dalian University of Technology Press, 2013.

[10] F. Wu, W.X. Zhong, The Zu-type method is symplectic. *Chinese Journal of Computational Mechanics*, **32**, 447-450, 2015.

[11] J. Sampson, A. Easton, M. Singh, Moving boundary shallow water flow above parabolic bottom topography. *ANZIAM Journal*, **47**, 373-387, 2005.

[12] J. Murillo, P. Garc á-Navarro, Augmented versions of the HLL and HLLC Riemann solvers including source terms in one and two dimensions for shallow flow applications. *Journal of Computational Physics*, **231**, 6861-6906, 2012.

[13] N. Izem, M. Seaid, M. Wakrim, A discontinuous Galerkin method for two-layer shallow water equations. *Mathematics and Computers in Simulation*, **120**, 12-23, 2016.

[14] C. Frei, Dynamics of a two-dimensional ribbon of shallow water on a f-plane. *Tellus a*, **45**, 44-53, 1993.

[15] C. Sch R, P.K. Smolarkiewicz, A synchronous and iterative Flux-Correction formalism for coupled transport equations. *Journal of Computational Physics*, **128**, 101-120, 1996.

[16] L.H. Bo, *Study on the shallow water euqations o high resolution algorithm*. Dalian University of Technology, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# IMMERSED BOUNDARY METHODS FOR COMPRESSIBLE LAMINAR FLOWS

**Rakesh Ramakrishnan**[1], **Anant Girdhar**[2] **and Santanu Ghosh**[3]

[1] Department of Aerospace Engineering
Indian Institute of Technology - Madras
Chennai 600036
e-mail: ae11b026@smail.iitm.ac.in

[2] Department of Aerospace Engineering
Georgia Institute of Technology
Atlanta, GA 30332
e-mail: anant.girdhar@gatech.edu

[3] Department of Aerospace Engineering
Indian Institute of Technology - Madras
Chennai 600036
e-mail: sghosh1@iitm.ac.in

**Keywords:** Immersed Boundary Method, Compressible Flows, Laminar Flows

**Abstract.** *Immersed boundary methods (IB) are a set of methods to deal with non-body conforming grids. This requires forcing the boundary conditions in the vicinity of the immersed surface. In this work, the embedded object is represented as a set of line segments along with their outward unit normal vectors. The flow domain is categorised into field cells, band cells or interior cells using a signed distance based approach. Although IB methods have been widely used for incompressible flows, it's application to high speed flows is relatively less. The proposed work attempts to construct an IB method suitable for compressible flow applications. Velocity boundary conditions are applied using an inverse distance based approach near no-slip walls and a signed distance based approach for slip walls. Different flow problems are simulated: expansion fan, supersonic flow past ramp channel and supersonic flow past NACA0012 airfoil which are inviscid simulations, subsonic viscous laminar flow past a ramp channel and flow past a cylinder (to simulate Von Karman vortex streets) which are viscous simulations. In each case, a comparison is made with simulations using body fitted grids or experimental data to validate the solution.*

## 1 Introduction

Immersed boundary methods (IB) are a set of methods to deal with non-body conforming grids. For simulating flow past bodies of arbitrary shape, generation of a curvilinear structured mesh can be very difficult in many cases and lead to loss of grid quality (high skewness). An alternative approach is to use unstructured grids for complex geometries or cases with multiple or moving bodies. In practice, development of a 3-D solver takes reasonable amount of effort and time and switching to an unstructured solver may not be feasible. Further, structured solvers are more amenable to higher order upwinding schemes, which are commonly used in high-speed flow solvers, and also typically easier to develop compared to their unstructured counterparts. In this context, another approach which retains structured grids but can deal with complex geometries with less computational effort (compared to construction of a new solver) is the immersed-boundary method. In this method, the flow variables are re-constructed near the vicinity of the body so that the flow feels the presence of the same. This requires the application of solution forcing in the vicinity of the immersed body. In this work a novel forcing mechanism suitable for the finite volume approach is used. The exact methods adopted and its effectiveness is the focus of this paper.

Although IB methods have been widely used for incompressible flows, it's application to high speed flows is relatively less. Peskin C S proposed the immersed boundary method [1, 2, 3] to simulate flow past heart valves. Goldstein et al. [4] extended Peskin's work to model no slip boundaries in IB using an external force field. Mohd Yusof [5] combined IB with B-splines to allow simulation of complex moving geometries. Faldun et al. [6] proposed an IB method for simulating unsteady three-dimensional incompressible flows in complex geometries. Choi et al. [7] developed an IB method for incompressible methods based on a signed distance based approach. For compressible flow simulations, Ghias et al. [8] proposed a sharp interface based IB method. Ghosh et al. [9] extended the signed distance based approach of Choi et al to compressible flows.

In this work, the embedded object is represented as a set of line segments. The flow domain is categorised into *field cells* or *interior cells* using a signed distance function [7]. The proposed work attempts to construct an IB method suitable for compressible flow applications. The flow variables in the immediate vicinity of the immersed surface (at the interfaces between *interior cells* and *field cells*) are re-constructed such that the flow around the embedded object feels the presence of the same. Tangential velocities are reconstructed to ensure no-slip, and heat and mass fluxes are set to zero at the interface between such cells and *field cells* for a stationary embedded surface. The IB method is integrated into a finite volume solver for structured grids.

A 3-D Navier-Stokes laminar flow solver with explicit time stepping has been developed based on finite volume methods for structured grids. The solver has a modular design with support for multiple flux reconstruction schemes and higher order extensions to space and time-integration. All simulations performed as a part of this work uses the AUSM flux-splitting scheme of Liou and Steffen [10] to construct the interface fluxes. The Piecewise Parabolic Method (PPM [11]) has been used to reconstruct the state at the cell interfaces. For steady state simulations, the convergence of the simulation was ascertained by the normalised residue (resnorm). The resnorm was calculated as the L2 norm of the residue from flux computations. For the steady-state problems simulated in this paper, a resnorm drop of at least 5 orders of mag-

nitude was observed. A python backend has also been developed to automate post-processing using the VisIt visualisation package [12].

## 2 Immersed Boundary Formulation: Geometric Aspects

This section deals with representation of the immersed surface and the method of classification of fluid cells based on proximity to the immersed body. The method is currently implemented for 2D objects and the following implementation is valid for 2D flows.

### 2.1 Representation of the immersed surface

The immersed surface is represented by a set of line segments (Immersed Boundary (IB) line) along with their outward unit normal vectors. The manner in which the IB surface is represented has a bearing on the manner in which the velocity forcing is imposed. The use of line segments for representing the IB makes the process less memory intensive compared to methods in which immersed surfaces are represented as points [7, 9]. For example, to simulate the flow past a ramp, treated as an IB, the present method requires two line segments in place of a dense cloud of points. This treatment can be extended for moving surfaces as well with velocities provided at the IB lines.

### 2.2 Classification of cells

Given an immersed boundary, the cells in the computational domain are divided as:

- *field cells*: Cells with their centre external to the immersed boundary. Fluid flow equations are integrated in these cells.
- *interior cells*: Cells with their centre internal to the immersed boundary. The state variables are not updated with time at these cells.
- *band cells*: Interior cells which have at least one field cell as neighbour. Such cells are reclassified as band cells. Modified boundary conditions are applied for these cells.
- *band faces*: A face lying between a field and band cell. The state reconstruction at these faces is done in order to achieve a forcing on the solution in the field cell straddling it.

The classification of internal or external is based on the relative position of the cell centre with respect to the outward normal vectors of the immersed boundary. There are algorithms which can be used from computational geometry to find if a point lies inside or outside a polygon, but a different approach is sought here. For each cell the following information is needed: the nearest IB line, external or internal to the IB, and the distance to the nearest IB line.

Given a cell centre and the $i^{th}$ IB line (as shown in figure 1), the cell centre is projected to the IB line along the normal with intersection point $\vec{\boldsymbol{r}}_{NC}$. Equation 1 can be used to solve for $\vec{\boldsymbol{r}}_{NC}$ by solving for $\alpha_i$ and $\beta_i$ and the intersection point can be found using equation 2.

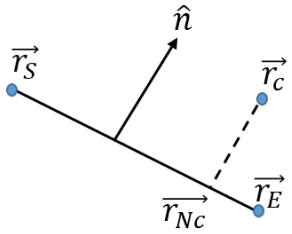



Figure 1: Cell centre C and IB line

$$\vec{\boldsymbol{r}}_{S_i} + \alpha_i \left(\vec{\boldsymbol{r}}_{E_i} - \vec{\boldsymbol{r}}_{S_i}\right) = \vec{\boldsymbol{r}}_C + \beta_i \hat{\boldsymbol{n}}_i \quad (1)$$

$$\vec{\boldsymbol{r}}_{NC} = \vec{\boldsymbol{r}}_C + \beta_i \hat{\boldsymbol{n}}_i \quad (2)$$

$$d = |\vec{\boldsymbol{r}}_C - \vec{\boldsymbol{r}}_{NC}|$$

$$d_{signed} = \frac{\beta_i}{|\beta_i|} d \quad (3)$$

The algorithm is as follows:

- A bounding circle is determined first by finding the centroid of the IB surface and determining a radius based on the maximum X and Y extents of the IB. All cells whose center lie outside the bounding circle are deemed as field cells. The cells inside the circle are classified as per the procedure explained below.
- Solve for $\alpha_i$ and $\beta_i$ as above and find intersection point $\vec{\boldsymbol{r}}_{NC}$ for every IB line.
- Find closest IB line using $d$.
- if $\alpha_i \notin [0, 1]$, set $\vec{\boldsymbol{r}}_{NC}$ as $\vec{\boldsymbol{r}}_{S_i}$ or $\vec{\boldsymbol{r}}_{E_i}$ which ever is closer and reset $d$.
- if $\beta_i < 0$, $\vec{\boldsymbol{r}}_C$ is marked as **exterior**, else **interior**.
- The signed distance $d_{signed}$ is evaluated as per equation 3.
- Interior cells with at least one exterior cell as its neighbour are marked as **band cells**.
- Faces which lie between a field cell and a band cell are marked as **band faces**.

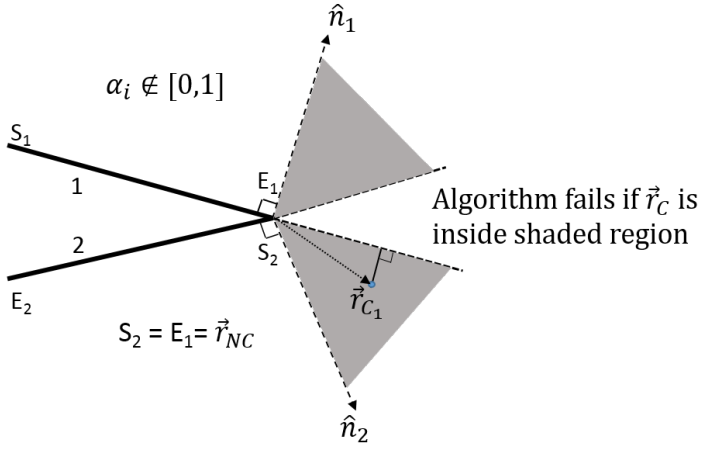



Figure 2: Sample concave region where algorithm fails

The signed distance based classification **fails** at concave junctions of IB or at sharp corners for cell centres with $\alpha_i \notin [0, 1]$. At such cases, some external points are wrongly marked as interior. The situation is shown in figure 2 where the cell centre $\vec{\boldsymbol{r}}_{C_1}$ would be marked as internal as it lies in the opposite direction of the normal of IB line 1 (i.e., $\hat{\boldsymbol{n}}_1.(\vec{\boldsymbol{r}}_{C_1} - \vec{\boldsymbol{r}}_{NC}) < 0$) and the

vertex E1 is chosen as $\vec{\boldsymbol{r}}_{NC}$. The thick lines are the IB lines 1 and 2. Hence the above algorithm is modified as follows:

- if $\alpha_i \notin [0, 1]$, set $\vec{\boldsymbol{r}}_{NC}$ as $\vec{\boldsymbol{r}}_{S_i}$ or $\vec{\boldsymbol{r}}_{E_i}$ which ever is closer and reset $d$
  - Find neighbouring IB line 'j' which shares common vertex $\vec{\boldsymbol{r}}_{NC}$
- if $(\vec{\boldsymbol{r}}_C - \vec{\boldsymbol{r}}_{NC}) \,.\, (\hat{\boldsymbol{n}}_i + \hat{\boldsymbol{n}}_j) \geq 0$: mark cell as **external**. Else mark as **internal**
- In figure 2, to evaluate the status of cell centre $\vec{\boldsymbol{r}}_{C_1}$, IB line 2 becomes the neighbouring IB line and $\hat{\boldsymbol{n}}_2 = \hat{\boldsymbol{n}}_j$

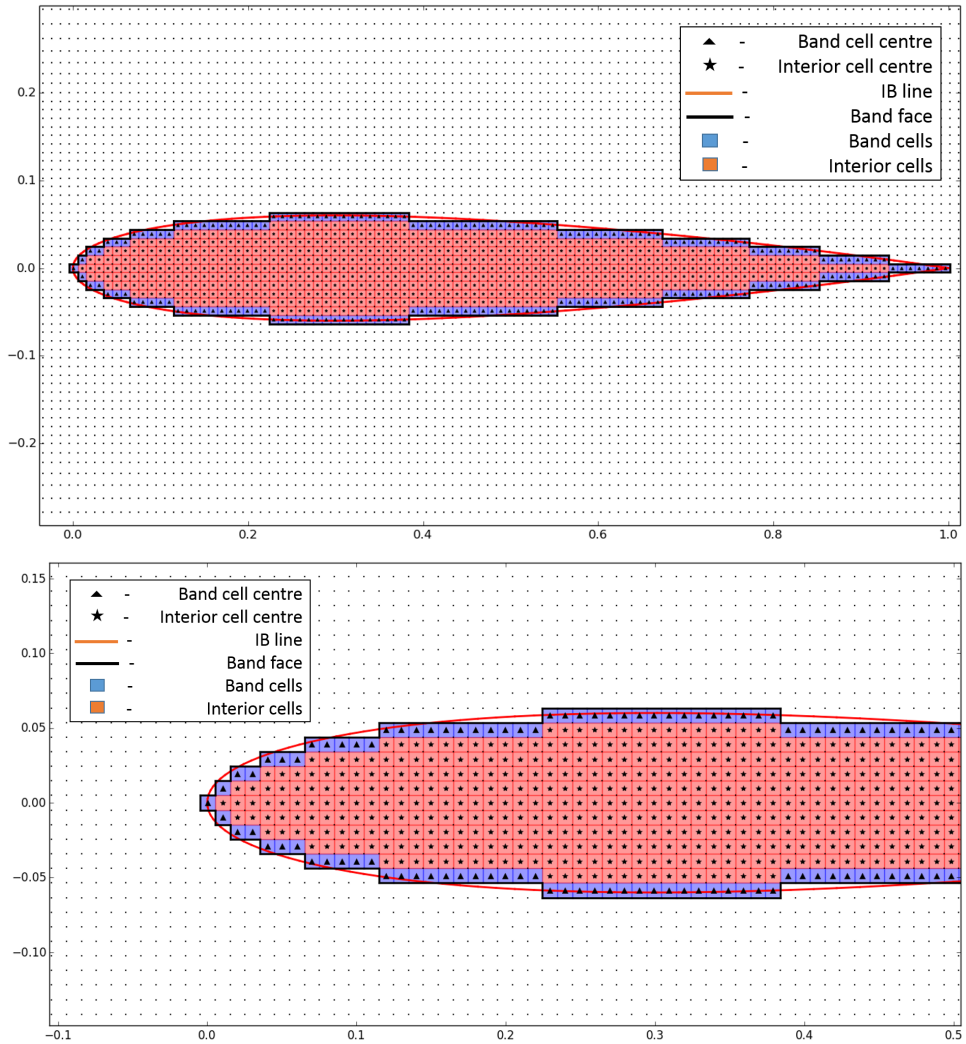



Figure 3: Cell classification around NACA-0012 airfoil. Top - full airfoil. Bottom - zoomed near the leading edge

The classification of field (or exterior), interior and band cells is shown in figure 3. The white cells are field cells, blue are band cells, red are interior cells and the thick black lines are the band faces. The NACA-0012 airfoil is shown as a red line.

## 3 Immersed Boundary Formulation: Solution Forcing

In this work, all boundary conditions are applied at the **band faces** as the values reconstructed at the faces will be used for flux computations. This will affect the evolution of the solution at the field cells neighboring the IB. The idea is to implement the state / flux reconstruction at the band face such that the flow adapts to the embedded surface in physically correct manner. The details of this procedure are explained in the following sub-sections.

The left and right values at a face are reconstructed depending on the spatial accuracy desired. For a first order upwind scheme, the states at the left element is directly taken as the left state of the face and vice-versa. However, higher order accurate methods use a wider stencil to interpolate the values at a face. For example, the Piecewise Parabolic Method (PPM [11]), uses the cell centred values of 4 cells to reconstruct the state at a face. However, near the boundary, the stencil is narrower and the reconstructed state is the average of the values at the neghbouring cell centres.

- Band faces are boundary faces which separate the band cells and field cells. Hence the band cells are like the ghost cells of body fitted grids.
- The boundary conditions are imposed at these faces but only using the field cell values.

### 3.1 Velocity forcing

In order to arrive at the velocity forcing, the boundary conditions imposed for body conforming grids are studied. For body fitted grids, it is assumed that the ghost cell centre is a mirror of the interior cell centre about the face. As a result, the following can be said: $q_{face} = \frac{1}{2}\left(q_{ghost} + q_{interior}\right)$. It should be noted that for body fitted grids, $q_{ghost}$ is enforced and $q_{face}$ is calculated using the former. For the present immersed boundary method, the reverse is done.

For body fitted grids, the following velocity boundary conditions are applied:

- Slip wall: $\vec{\boldsymbol{u}}_G = \vec{\boldsymbol{u}}_I - 2\left(\vec{\boldsymbol{u}}_I.\hat{\boldsymbol{n}}\right)\hat{\boldsymbol{n}}$, i.e., the normal component is flipped
- No slip wall: $\vec{\boldsymbol{u}}_G = -\vec{\boldsymbol{u}}_I$

It is assumed that the parallel and normal components of velocity (with respect to the boundary face) vary linearly or remain constant. Based on the above boundary conditions, the following velocity profile is arrived at as shown in figure 4 and 5. It should be noted that in both the cases, the normal component of velocity vanishes at the face.

Using the above assumption of linear variation for the normal component of velocity (with respect to nearest IB surface), a similar velocity profile can be imposed near the IB surface (at the band face). The following points are taken into consideration for this:

- The distance between the IB and the band face ($d_{face}$), the distance between the field cell centre and IB ($d_F$).
- The normal component of velocity should flip in sign if the face is interior to the IB. Hence signed distance is used.

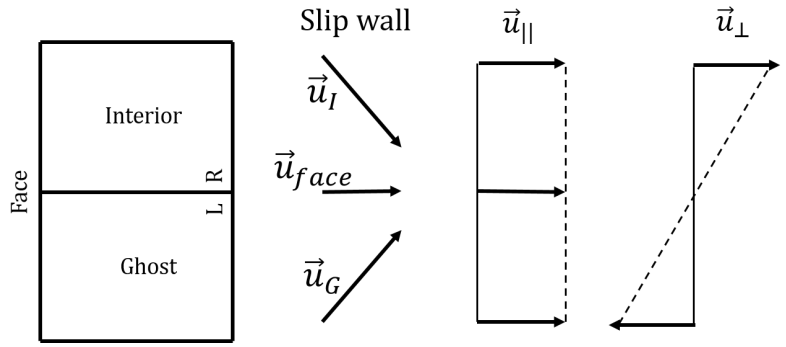



Figure 4: Slip wall velocity boundary condition (body fitted grid)

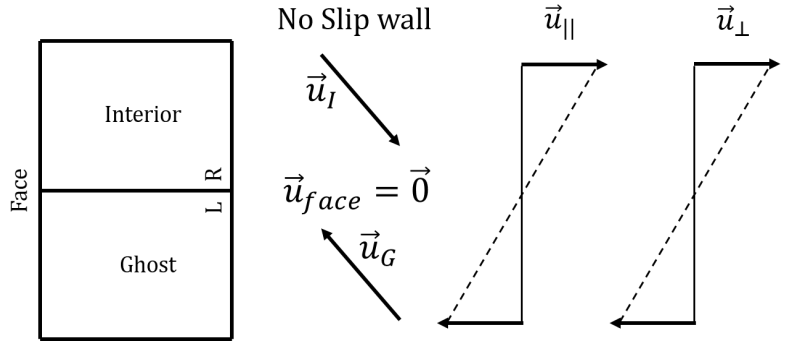



Figure 5: No Slip wall velocity boundary condition (body fitted grid)

- The normal and parallel directions are with respect to the nearest IB line. The normal vectors are the outward normal vectors of the IB and not the face of a finite volume cell.
- It is found that when the band face is below the IB, the velocity reconstruction at the band face using signed distances causes numerical instability. Hence a limiter is added as follows: $d_{face} = min(0., d_{face})$
- For **no slip walls**, an inverse distance based approach is used as per equation 5. Only the parallel component of the velocity (with respect to the nearest IB line) is used. This results in zero normal component of $u_{face}$ (with respect to nearest IB line).
- The velocity reconstruction at the band face is schematically shown in figure 6.

The equations for velocity forcing are summarised below. Equation 4 describes the velocity reconstruction for slip wall and equation 5 for no slip wall:

$$\begin{aligned} \text{Slip wall: } \vec{\boldsymbol{u}}_{face} &= \vec{\boldsymbol{u}}_F - \left( \left( 1 - \frac{d_{face}}{d_F} \right) \vec{\boldsymbol{u}}_F . \hat{\boldsymbol{n}}_{IB} \right) \hat{\boldsymbol{n}}_{IB} \qquad (4) \\ d_{face} &= \min(0, d_face) \\ \text{No Slip wall: } \vec{\boldsymbol{u}}_{face} &= \frac{\frac{\vec{\boldsymbol{u}}_{||_F}}{d_F} + \frac{\vec{\boldsymbol{u}}_{||_{IB}}}{d_{IB}}}{\frac{1}{d_F} + \frac{1}{d_{IB}}} \qquad (5) \\ \vec{\boldsymbol{u}}_{||_F} &= \vec{\boldsymbol{u}}_F - (\vec{\boldsymbol{u}}_F \,.\, \hat{\boldsymbol{n}}_{IB})\, \hat{\boldsymbol{n}}_{IB} \\ \vec{\boldsymbol{u}}_{||_{IB}} &= \vec{\boldsymbol{u}}_{IB} - (\vec{\boldsymbol{u}}_{IB} \,.\, \hat{\boldsymbol{n}}_{IB})\, \hat{\boldsymbol{n}}_{IB} \end{aligned}$$

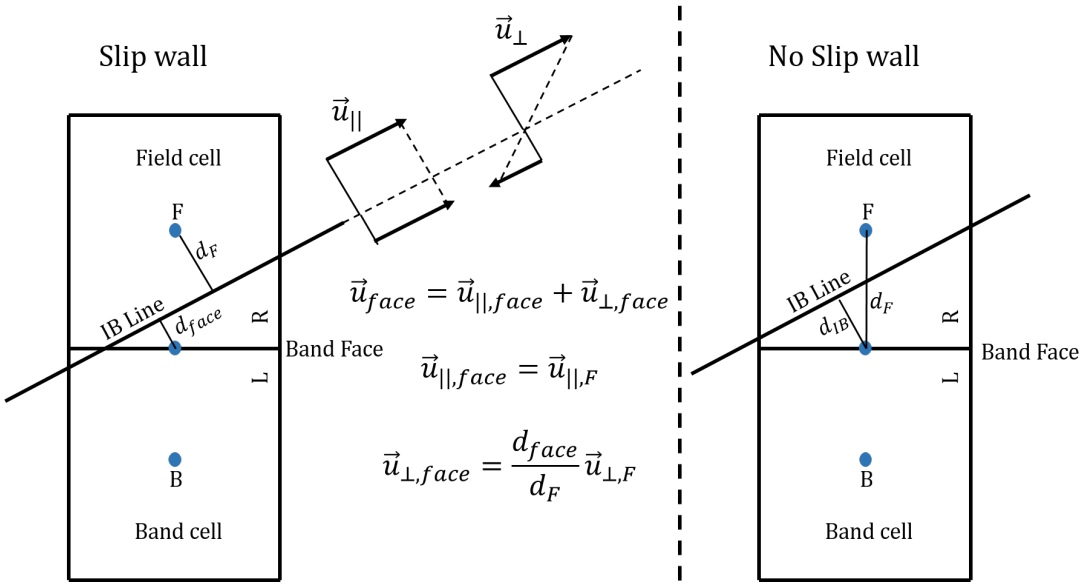



Figure 6: Velocity reconstruction at band face

### 3.2 Gradients at band faces

For viscous flows, the gradients at band faces are recomputed for the velocity components and temperature. The gradient is constructed as:

$$\begin{aligned} \vec{\boldsymbol{\nabla}}\phi &= \frac{2\left(\phi_F - \phi_{face}\right)}{0.5 * \left(V_F + V_B\right)} A_{face}\, \hat{\boldsymbol{n}}_{face} \\ \phi &= \{u, v, T\} \\ V_F &= \text{Volume of field cell} \\ V_B &= \text{Volume of band cell} \end{aligned} \tag{6}$$

In equation 6, the normal vector should be in the same direction as a vector from the band cell centre to field cell centre. If not, care should be taken to enforce the same.

### 3.3 Pressure and density forcing

The fluid flow solver, in which the IB formulation is currently implemented in, uses pressure, density and velocity as the primitive variables. The forcing for density will change if the wall is to be an isothermal wall or an adiabatic wall. Equation 7 describes the pressure forcing at the band faces and equations 8 and 9 describe the density forcing for adiabatic and isothermal IB surfaces. The equations are described below:

$$P_{face} = P_F \tag{7}$$

$$\text{Adiabatic wall} : \rho_{face} = \rho_F \tag{8}$$

$$\begin{aligned} \text{Isothermal wall} : \rho_{face} &= \frac{P_{face}}{RT_{face}} \\ &= \frac{P_F}{RT_{wall}} \\ &= \frac{\rho_F T_F}{T_{wall}} \end{aligned} \tag{9}$$

$$T_{wall} = \text{Temperature at nearest IB surface}$$

Using equations 7 and 8 in equation 6, it is observed that the gradient of temperature across a band face becomes **zero** for an adiabatic wall. Currently, pressure at a band face is set to be the same as that at the corresponding field cell. The current work studies the effect of such an assumption in the flow solution. Future work will involve other forms of interpolation in order to improve the accuracy of pressure near the surface.

## 4 Results

### 4.1 Expansion fan

Supersonic flow past a convex corner of angle $-10^\circ$ was simulated. The resulting expansion fan was studied for ratio of flow properties across it. The simulation is part of "AIAA code validation project" [13], where an expansion fan is simulated for free-stream Mach number of 2.0 with grid size of 120x60 cells. The simulation is higher order in space with PPM [11] and uses HLLE scheme [14] for flux splitting. The underlying grid for the Immersed Boundary (IB) method uses 120x120 cells while the immersed surface contains 2 lines to represent it. The Mach number contours are compared with the solution from body fitted grid (BFG) in figure 7 and a table of results is shown in table 1. The inlet conditions are:

- $M_\infty = M_1 = 2.0$
- $P_\infty = P_1 = 101325$
- $\rho_\infty = \rho_1 = 1.225$
- $V_\infty = V_1 = 680.588$

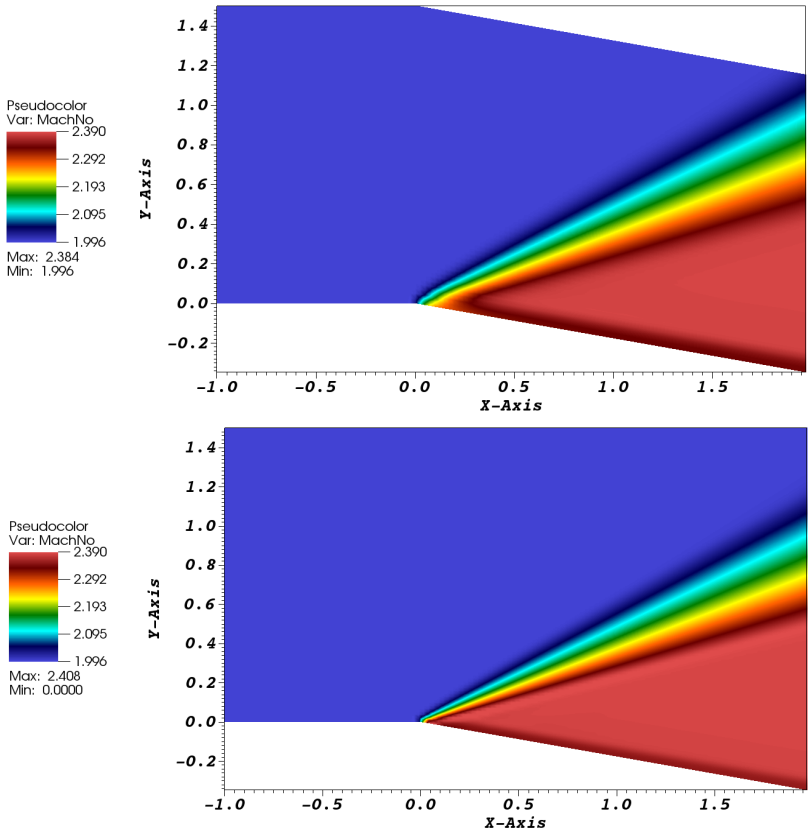



Figure 7: $M_\infty = 2.0$: Expansion Fan Mach number contours - BFG (top) and IB (bottom)

| State ratio | Theoretical | BFG | IB |
|---|---|---|---|
| $\frac{M_2}{M_1}$ | 1.1915 | 1.1919 | 1.192 |
| $\frac{P_2}{P_1}$ | 0.5471 | 0.5479 | 0.5479 |
| $\frac{\rho_2}{\rho_1}$ | 0.65 | 0.6504 | 0.6504 |
| $\frac{V_2}{V_1}$ | 1.0931 | 1.0941 | 1.0941 |

Table 1: $M_\infty = 2.0$: Comparison of state variables across an expansion fan

### 4.2 Ramp channel inviscid supersonic flow

A supersonic, inviscid flow over a ramp channel was simulated at free stream Mach number $M_\infty = 2.2$. The ramp starts from (0.25, 0) and ends at (0.5, 0.067). The channel has a height of 0.2 before the ramp and is of total length 1.0. This problem is simulated to check the accuracy with which the IB method captures attached oblique shocks which is very important for any high-speed flow solver. The top and bottom part of the domain are slip walls. A grid size of 181 x 65 cells was used for both the body fitted and IB methods. The Mach number contours are compared in figure 8. The pressure plot along a streamline very close to the bottom wall is shown in figure 9.

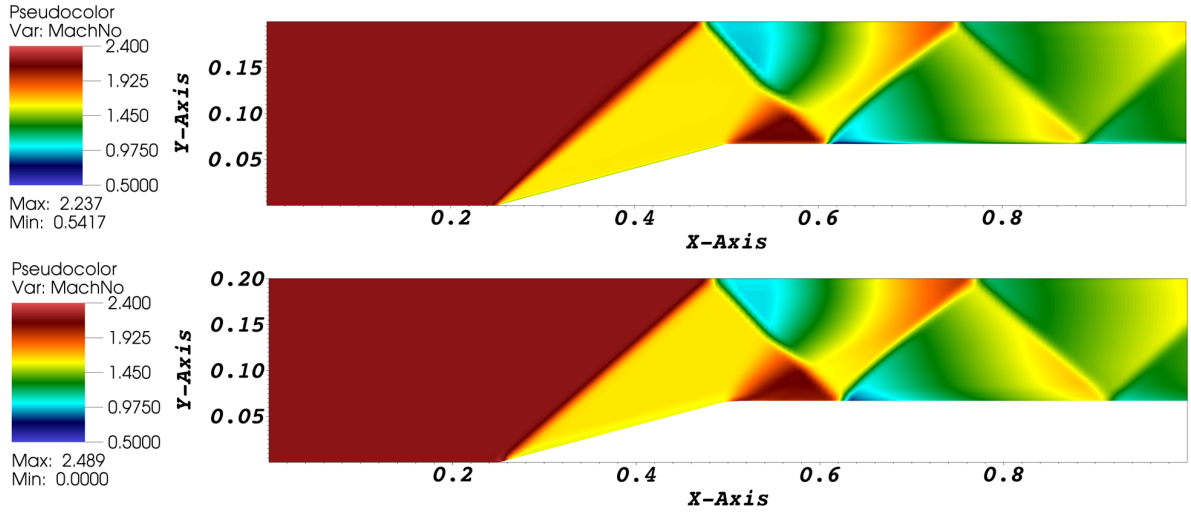



Figure 8: $M_\infty = 2.2$: Ramp Channel Mach number contours - BFG (top) and IB (bottom)

It is seen that while there is a good agreement in the contour plot, the surface pressure is not so accurately captured by the IB method. The surface pressure is presently calculated in VisIt but should ideally not use any interior cell data which has not been ensured at present. Ideally this needs to be extracted from the solution using suitable interpolation procedure which neglects interior cell data and will be developed. Hence, a more thorough surface pressure extraction method is to be developed for the IB.

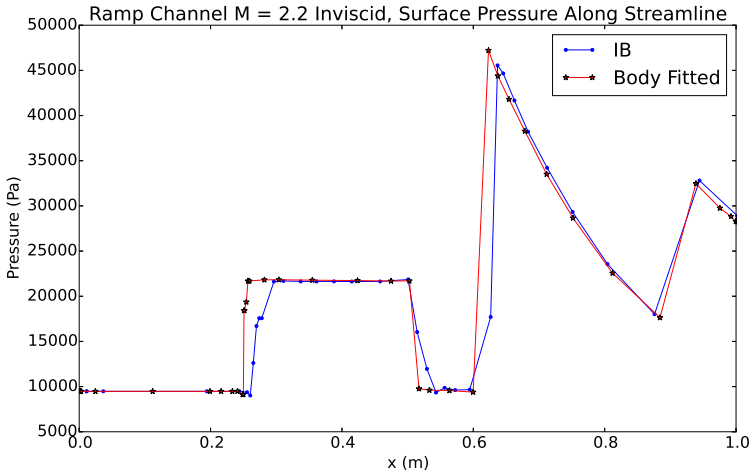



Figure 9: $M_\infty = 2.2$: Ramp Channel Pressure Along Streamline

### 4.3 NACA-0012 Inviscid Supersonic Flow

An inviscid supersonic flow of Mach number $M_\infty = 1.2$ was simulated past NACA-0012 airfoil to determine the ability of the IB method in capturing detached shocks. The underlying grid size is 381 x 141 cells with 200 IB lines representing the airfoil surface. The Mach number contours are compared with results from Oscar et al. (figure 7 in [15]). The Mach number contours from IB is shown in figure 10 and that from Oscar et al. is shown in figure 11. The Mach number contours show good agreement with the results presented by Oscar et al.

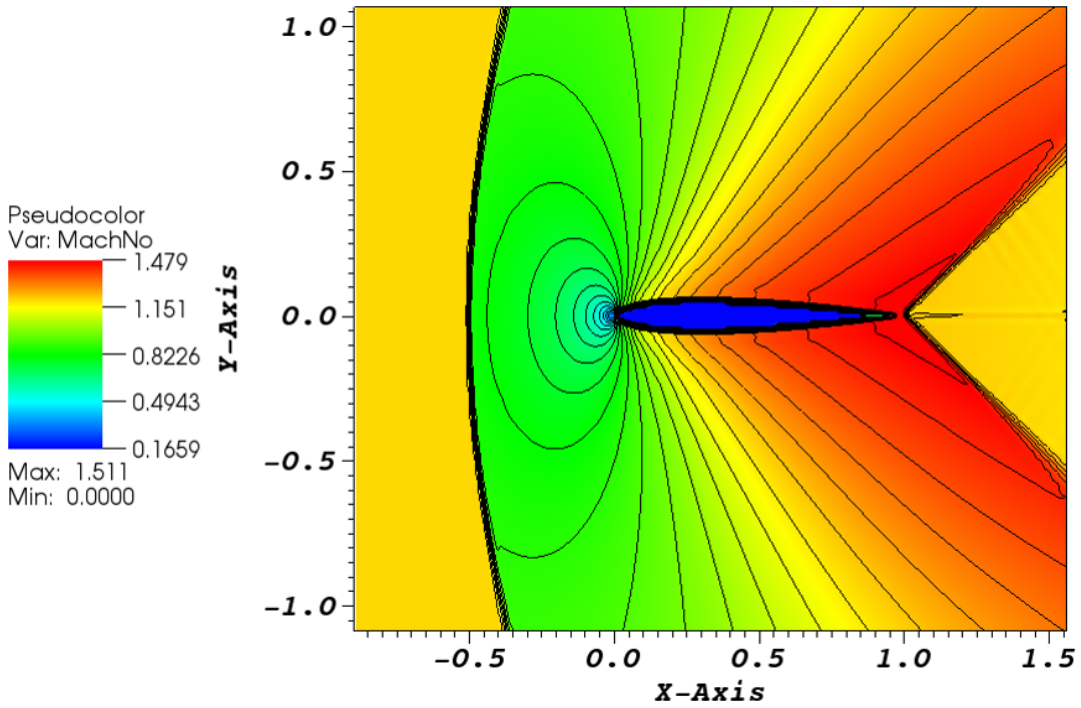



Figure 10: $M_\infty = 1.2$: Mach number contours for flow past NACA-0012 (IB method)

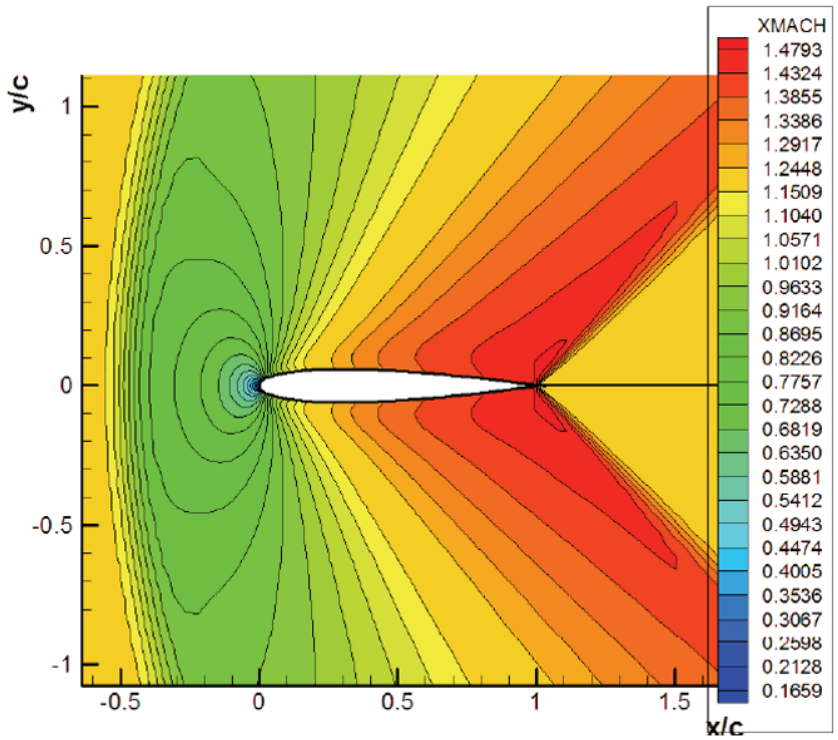



Figure 11: $M_\infty = 1.2$: Mach number contours for flow past NACA-0012 (Oscar et al. [15])

The shock position is mentioned as -0.46 from the leading edge of the airfoil in Oscar et al., but only the location of the end of the shock is mentioned. The location of the start of the shock is -0.512 from the leading edge of the airfoil in the current simulation, which agrees well with the result from Oscar et al. The shock simulated in the present simulation by the IB method is thinner.

### 4.4 Ramp channel viscous subsonic flow

A viscous flow of Re = 6829 past the same ramp channel (Re based on channel height of 0.2) was simulated. The Mach number and pressure contours for the body-fitted grid (BFG) case and immersed boundary (IB) case are compared in figures 12 and 13. The surface pressure plot for the IB method is compared with BFG, result from REACTMB code [9] and Abaqus (commercial FEA package from Dassault Systems) in figure 14.

The Mach number contours look very similar, and except for the noticeable difference in the local pressure minima at the end of the ramp, the pressure contours compare well. It can be seen that the pressures from the BFG and REACTMB code are almost the same while the FEA package predicts slightly lower pressures, but the difference is only **0.6%**. As mentioned earlier, a more thorough surface pressure extraction method is to be developed for the IB.

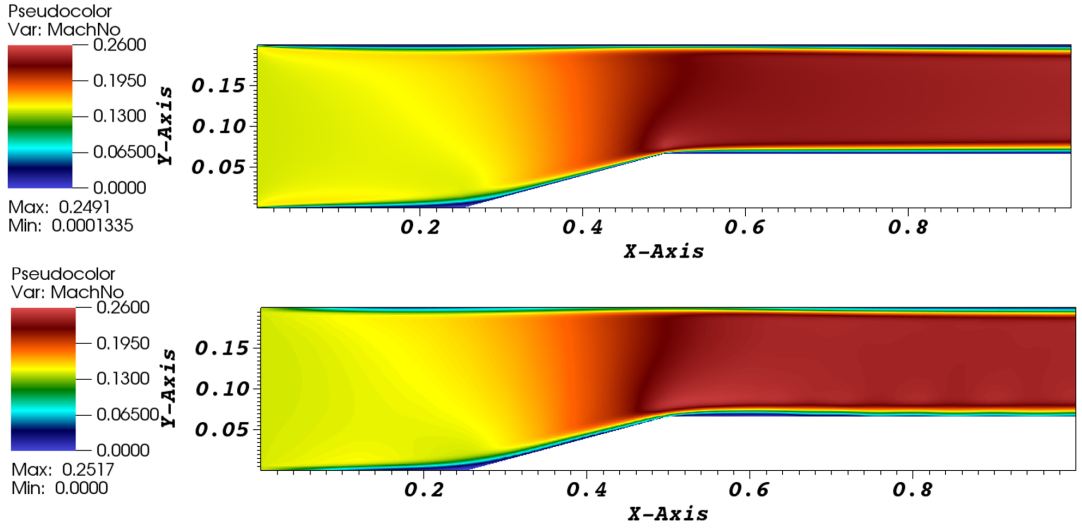



Figure 12: $Re = 6829$: Ramp Channel Mach number contours - BFG (top) and IB (bottom)

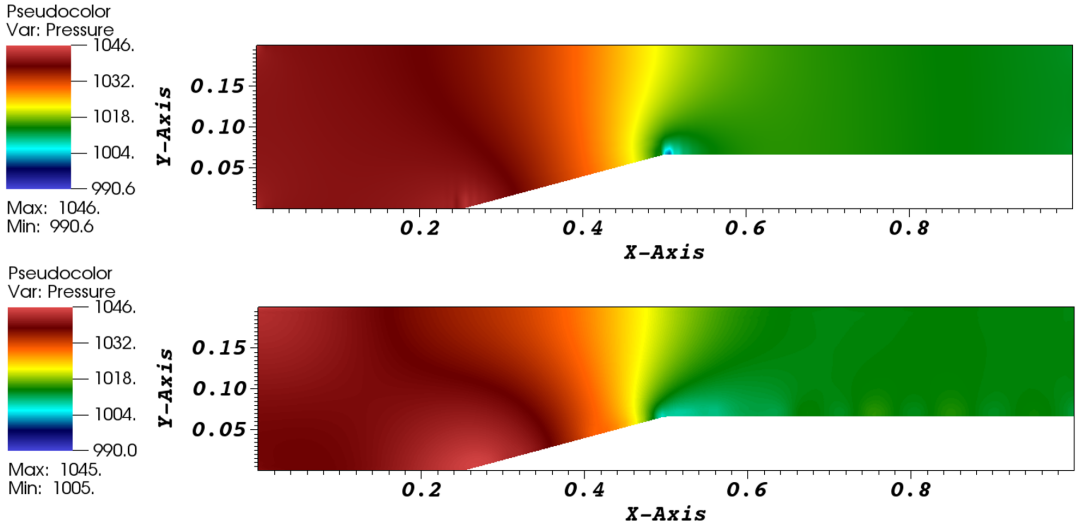



Figure 13: $Re = 6829$: Ramp Channel Pressure contours - BFG (top) and IB (bottom)

### 4.5 Viscous laminar flow past a cylinder

A viscous laminar flow past a circular cylinder is simulated for Reynolds numbers 100 and 478. There occurs a periodic shedding of vortices from the cylinder leading to the formation of the Von Karman vortex street. The domain is a rectangular with a circular diameter of 0.01 as shown schematically in the documentation of Caelus software (CFD solver from Applied CCM). The underlying grid size is 351 x 401 cells with 50 IB lines representing the cylinder surface. The instantaneous vorticity contours are plotted in figure 15.

Pressure was probed downstream of the cylinder for both the cases as a function of time. The frequency of shedding was calculated by taking the FFT (Fast Fourier Transform) of the above data. The Strouhal number (St) was then computed using the relation in equation 10 and compared with (i)experimental data obtained from figure 4 of [16] and (ii)a similar IB simulation from [7]. The results are tabulated in table 2.

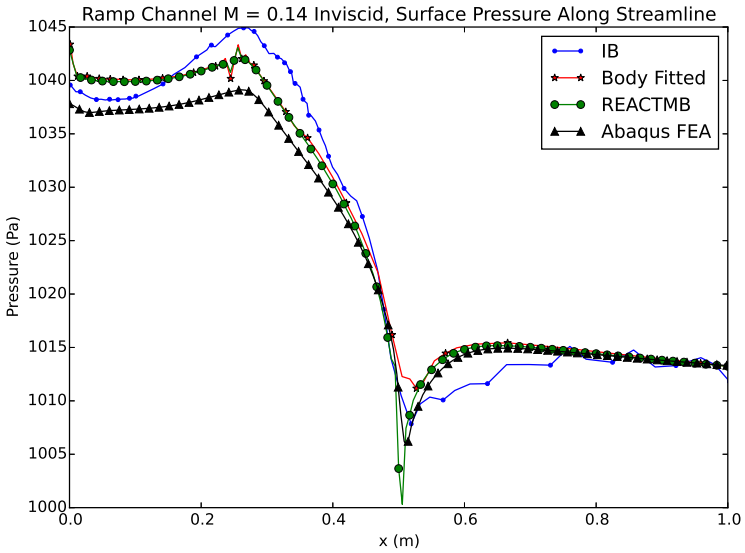



Figure 14: $Re = 6829$:: Ramp Channel Pressure Along Streamline

$$\begin{aligned} St &= \frac{fd}{V_\infty} \\ f &= \text{frequency of vortex shedding} \\ d &= \text{diameter of cylinder} \\ V_\infty &= \text{free stream velocity} \end{aligned} \tag{10}$$

| **Reynolds Number** | **Experimental** | **Other Simulations** | **IB** |
|---|---|---|---|
| 100 | 0.164 to 0.175 | 0.164 [7] | 0.164 |
| 478 | 0.205 to 0.21 | - | 0.222 |

Table 2: Comparison of Strouhal number for flow past cylinder

It is seen that the IB results agree very well with experimental data and previous IB simulations.

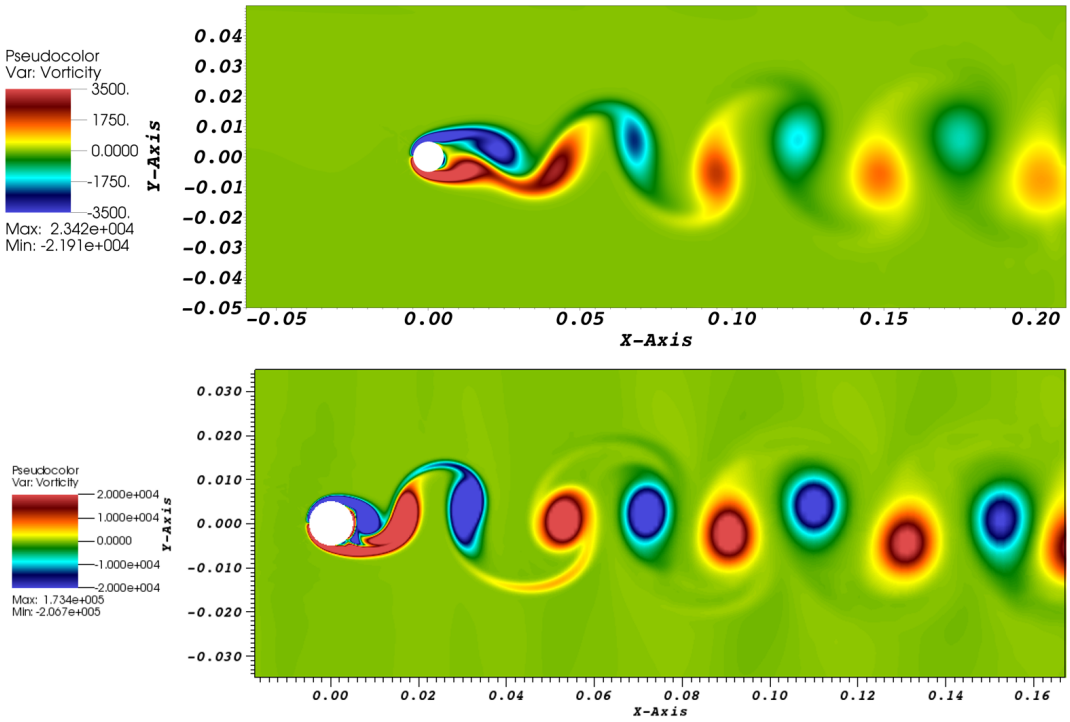



Figure 15: Instantaneous vorticity contours for Re = 100 (top) and Re = 478 (bottom)

## 5 Conclusions

A novel immersed boundary method suitable for finite volume methods is presented and tested for (i) compressible and incompressible (low speed) flows, (ii) inviscid and viscous flows, (iii) steady state and unsteady simulations with mostly good agreement with computations using body fitted grids and experiments. It is seen that while the boundary shape is captured effectively, the surface pressure is not smooth as compared to the surface pressures obtained from body fitted grids. A better method needs to be developed to extract the surface pressures for IB simulations. However, considering the potential of IB methods in solving a wide class of problems involving flow past objects of arbitrary shape while still employing structured grids, the results obtained are encouraging.

## REFERENCES

[1] Peskin, C. S., "Flow patterns around heart valves: A numerical method," *Journal of Computational Physics*, Vol. 10, No. 2, oct 1972, pp. 252–271.

[2] Peskin, C. S., "Numerical analysis of blood flow in the heart," *Journal of Computational Physics*, Vol. 25, No. 3, 1977, pp. 220–252.

[3] Peskin, C. S. and McQueen, D. M., "A three-dimensional computational method for blood flow in the heart I. Immersed elastic fibers in a viscous incompressible fluid," *Journal of Computational Physics*, Vol. 81, No. 2, 1989, pp. 372–405.

[4] Goldstein, D., Handler, R., and Sirovich, L., "Modeling a No-Slip Flow Boundary with an External Force Field," *Journal of Computational Physics*, Vol. 105, No. 2, apr 1993, pp. 354–366.

[5] Mohd-Yusof, J., "Combined immersed-boundary/B-spline methods for simulations of flow in complex geometries," *CTR annual research briefs*, 1997.

[6] Fadlun, E., Verzicco, R., Orlandi, P., and Mohd-Yusof, J., "Combined Immersed-Boundary Finite-Difference Methods for Three-Dimensional Complex Flow Simulations," *Journal of Computational Physics*, Vol. 161, No. 1, jun 2000, pp. 35–60.

[7] Choi, J., Oberoi, R., Edwards, J., and Rosati, J., "An Immersed Boundary Method for Complex Incompressible Flows," *Journal of Computational Physics*, Vol. 224, No. 2, June 2007, pp. 757–784.

[8] Ghias, R., Mittal, R., and Dong, H., "A sharp interface immersed boundary method for compressible viscous flows," *Journal of Computational Physics*, Vol. 225, No. 1, jul 2007, pp. 528–553.

[9] Ghosh, S., Choi, J.-I., and Edwards, J. R., "Numerical Simulations of Effects of Micro Vortex Generators Using Immersed-Boundary Methods," *AIAA Journal*, Vol. 48, No. 1, Jan 2010, pp. 92–103.

[10] Liou, M.-S. and Christopher J Steffen, J., "A New Flux Splitting Scheme," *Journal of Computational Physics*, Vol. 107, No. 1, May 1993, pp. 23–39.

[11] Colella, P. and Woodward, P. R., "The Piecewise Parabolic Method (PPM) for gas-dynamical simulations," *Journal of Computational Physics*, Vol. 54, No. 1, April 1984, pp. 174–201.

[12] Childs, H. and et. al., E. B., "VisIt: An End-User Tool For Visualizing and Analyzing Very Large Data," *High Performance Visualization–Enabling Extreme-Scale Scientific Insight*, Oct 2012, pp. 357–372.

[13] Urmila Ghia, Sami Bayyuk, S. H. C. R. e. a., "The AIAA code verification project - Test cases for CFD code validation," Vol. 48, AIAA, 2010.

[14] Einfeldt, B., "On Gudonov Type Methods for Gas Dynamics," *SIAM Journal on Numerical Analysis*, Vol. 25, No. 2, April 1988, pp. 294–318.

[15] Oscar Arias, Oscar Falcinelli, e. a., "Finite volume simulation of a flow over a NACA 0012 using Jameson, MacCormack, SHU and TVD schemes," Vol. 26, Mecanica Computational, 2007, pp. 3097–3116.

[16] Roshko, A., "On the development of turbulent wakes from vortex streets," *NASA Technical Report*, Vol. 1911, 1954.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# IMPACT OF EXTERNAL SURROUNDINGS ON NATURAL CONVECTION IN A VERTICAL CHANNEL ASYMMETRICALLY HEATED (ECCOMAS CONGRESS 2016)

**D. Ramalingom**[1], **A. Bastide**[1]

[1]PIMENT Université de la Réunion
117 Avenue du Général Ailleret, 97430 Le Tampon, France
e-mail: delphine.ramalingom@univ-reunion.fr
e-mail: alain.bastide@univ-reunion.fr

**Keywords:** natural convection, vertical channel, thermal stratification, radiation

**Abstract.** *This paper reports a numerical study of natural convection problem between vertical plates with asymmetric heating. Influence of perturbation conditions outside of the channel on flow structure and on the heat transfer rate are investigated. The effect of temperature consists of considering a gradient of temperature between the bottom and top of the channel in order to obtain a thermal stratification. The effect of surface radiation on the laminar air flow with a thermal stratification is investigated by considering temperature of grey bodies. Results show that these weak perturbations outside the channel are a real influence on flow and that the influence of thermal stratification is more important than surface radiation. Numerical simulations have been carried out at modified Rayleigh number $Ra=5.10^5$ (laminar regime) and with Prandtl number $Pr=0.71$.*

## Nomenclature

| | | | |
|---|---|---|---|
| $A$ | surface of the heated wall | $d_w$ | width of the downward flow |
| $q_{r_i}$ | net radiative heat flux | $l_w$ | length of the downward flow |
| $Ra$ | Rayleigh number | $H$ | height hot plate of channel |
| $Ra_b$ | modified Rayleigh number | $\tilde{J}_i$ | radiosity linear system |
| $\dot{m}_{in}$ | mass flow rate entering through bottom-end | $Nu_1$ | Nusselt number on the heated plate |
| $\dot{m}_{out}$ | mass flow rate exiting through top-end | $\overline{Nu_1}$ | Mean Nusselt number on the heated plate |
| $\dot{m}_{es}$ | mass flow rate entering through top-end | $p$ | pressure |
| $T$ | temperature | $Pr$ | Prandtl number |
| $T_0$ | reference temperature | **Greek Symbols** | |
| $T_{bottom\ body}$ | temperature of grey surface bottom-end | $\Delta T$ | temperature difference scale |
| $T_{top\ body}$ | temperature of grey surface top-end | $\epsilon$ | emissivity |
| $b$ | width of the channel | $\kappa$ | thermal conductivity |
| $u, v$ | velocity components | $\nu$ | kinematic viscosity |
| $t$ | time | $\Phi$ | heat flux |
| $U_{CN}$ | characteristic velocity | $\Psi$ | stream function |
| $\vec{e_y}$ | coordinates axes | $\sigma$ | Stefan-Boltzmann constant |
| $\vec{v}$ | velocity vector | $\theta$ | dimensionless temperature |

## 1 INTRODUCTION

The heated vertical open-ended channel is representative of practical interest such as the chimney, the solar panel, or the Trombe wall. The problem of natural convection in vertical channels has been the focus of extensive investigations [10]. Authors studied free convection between vertical flat plates with symmetric or asymmetric heating, with uniform heat fluxes or constant temperatures[4]. There are several ways to take into account the surrounding conditions. The channel can have adiabatic extensions, be closed in cavities or open domain [7, 1, 2, 17, 16]. These different strategies lead to an increase in size of the computational domain which proves to be expensive, both in memory and in computational time [19]. Moreover, interactions between the channel and surroundings are not correctly estimated by these strategies [3, 12, 13, 14]. Indeed, thermal stratifications are often not considered although experimental studies show its existence. As the same, except some studies in literature [6, 15], major of numerical investigations overlooked surface radiation in natural convection case. Moreover, several studies [6, 20, 21] confirmed that radiation transfer represent significant percent in heat transfer global.

The influence of thermal stratification on the laminar air flow induced by natural convection in vertical, asymetrically-heated channels is discussed. Studies conducted by the French Research Group AMETh [9] on natural convection in open-ended channels are taken as the base case for carrying out the computations. The thermal stratification is created by setting a constant bottom temperature different from the top temperature, in order to obtain a weak gradient of temperature between the entrance and the exit of the channel. The influence of external surface radiation is also investigated and temperature of grey bodies (bottom and top channel) are set to values in agreement with thermal stratification outside the channel.

## 2 DESCRIPTION OF TEST CASE

A vertical parallel plate channel of width $b$ and height $H$ is formed by two walls, one partially heated at a constant and uniform heatflux $\Phi$ on its half middle section and the remaining walls are adiabatic (see Figure 1 ).

The fluid flow is assumed laminar and two-dimensional. Accounting for the small relative temperature difference occurring between the heated wall and the aperture, the Navier-Stokes and energy equations are expressed with the Boussinesq approximation. The viscous dissipation term in the energy equation is neglected. The energy equation and the one dealing with radiant interchanges amongst surfaces is coupled through the thermal boundary conditions.

### 2.1 Heat and fluid flow equations

The governing flow and heat transfer equations, written in dimensionless and conservative form, read:

$$\nabla \cdot \vec{v} = 0 \tag{1}$$

$$\frac{\partial \vec{v}}{\partial t} + \nabla \cdot (\vec{v} \otimes \vec{v}) = -\nabla p + Pr Ra_b^{-1/2} \nabla^2 \vec{v} + Pr\theta \overrightarrow{e_y} \tag{2}$$

$$\frac{\partial \theta}{\partial t} + \nabla \cdot (\vec{v}\theta) = Ra_b^{-1/2} \nabla^2 \theta \tag{3}$$

with the reduced dimensionless temperature $\theta = \frac{T - T_0}{\Delta T}$. The reference temperature $T_0$ is set to 298.15 K. The dimensionless parameters governing the fluid flow and heat transfer are the Prandtl number $Pr = \frac{\nu}{\kappa}$ set to 0.71, the Rayleigh number $Ra_b = \frac{g\beta\Delta T b^3}{\nu\kappa}$ set to $Ra_b = 5 \cdot 10^5$ and characteristic velocity $U_{CN} = \kappa \frac{Ra_b^{1/2}}{b}$.

### 2.2 Surface radiation

For a given temperature distribution on the channel internal surfaces, the surface radiation problem is fully described by the linear system for radiosity $\tilde{J}_i$ $(W/m^2)$. The net radiative heat flux resulting from surface radiation, which is defined in the hemisphere of a surface element, can be calculated by:

$$q_{r_i} = \frac{\epsilon_i}{1 - \epsilon_i} (\sigma T_i^4 - \tilde{J}_j) \ (i = 1, 2, ..., m) \tag{4}$$

where $\sigma$ is the Stefan-Boltzmann constant, $m$ is the total number of surface elements, $\epsilon_i$ is the emissivity of the surface element $i$ and $\tilde{J}_i$ $(W/m^2)$, the linear system for the radiosity.

In this paper, we consider that the emissivity of internal walls is equal to $0.1$ and the emissivity of the surfaces corresponding to the bottom-end and top-end, are set to $0.9$. The originality of this study is temperature of bottom body and top body which are different to reference temperature. Indeed, surface radiation is only considered in thermal stratification case. As a consequence, bottom and top body can have a temperature as if they are located at a distance from the aperture of the channel, in a stratified ambient air.

### 2.3 Boundary Conditions

Boundary conditions have been the focus of several studies [5], [9] [22] [14]. According to latest results produced by [5], pressure boundary conditions at the top and bottom sections improved results when they are based on Local Bernoulli relation.

Three configuration cases are referenced in this paper. In each case, the conditions at the solid walls are the same as those in [9]; only the thermal boundary conditions are modified.

- Reference case: it's the case of AMETh [9] with Local Bernoulli (LB-LB) type of pressure boundary conditions at bottom and top of the channel. Thermal boundary conditions at bottom/top are:
  $\theta(x,0) = 0$ and $\theta(x,2H) = 0$

- Thermal stratification case: it consists of setting a low dimensionless temperature at the bottom, set to $-0.01$, in order to obtain a gradient of temperature at the outside of the channel. Thermal boundary conditions at the bottom/top are:
  $\theta(x,0) = -0.01$ and $\theta(x,2H) = 0$.

- Radiation surface and thermal stratification: the channel consists of 4 gray-diffuse, vertical surfaces (the three parts of left-hand side wall and the adiabatic right-hand side wall) and two horizontal surfaces regarded as grey radiators at an effective temperature of $T_{rad} = -0.03$ and $T_{rad} = 0.03$, respectively.
  The emissivity of vertical surfaces are set to $\epsilon = 0.1$ and emissivity of horizontal surfaces are setted to $\epsilon = 0.9$. Thermal boundary conditions at the bottom/top:
  $\theta(x,0) = -0.01$ and $\theta(x,2H) = 0$
  and horizontal surface temperature are set to:
  $T_{bottom\ body} = -0.03$ and $T_{top\ body} = 0.03$

### 2.4 Monitored variables

#### 2.4.1 Dynamic quantities

- Mass flow rate entering into the channel at $y = 0$:

$$\dot{m}_{in} = \int_0^1 v(x,0)\,\mathrm{d}x \tag{5}$$

- Mass flow rate entering into the top section of the channel $y = 2H$:

$$\dot{m}_{es} = \int_0^1 \frac{|v(x,2H)|v(x,2H)}{2}\,\mathrm{d}x \tag{6}$$

- Mass flow rate exiting the channel at $y = 0$:

$$\dot{m}_{out} = \int_0^1 v(x,2H)\,\mathrm{d}x \tag{7}$$

#### 2.4.2 Thermal quantities

- Local nusselt number corresponds to the inverse of temperature at the left wall:

$$Nu_1(y) = \frac{1}{\theta(0,y)} \tag{8}$$

- Average of Nusselt number on the heated wall:

$$\overline{Nu_1} = \frac{2}{A}\int_{\frac{H}{2}}^{\frac{3H}{2}} Nu_1(y)\,\mathrm{d}y \tag{9}$$

  where A is the surface of the heated wall.

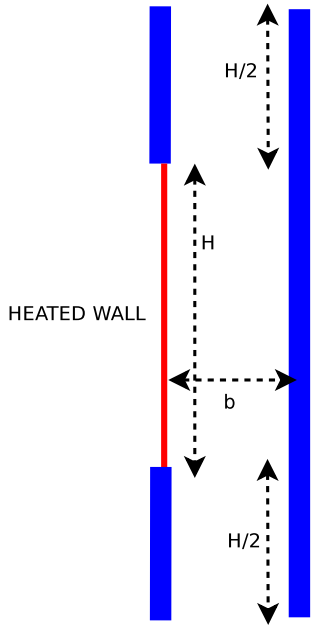



Figure 1: Geometry of the parallel-plate channel

#### 2.4.3 Characterization of flow patterns

- The width of downward flow at the top-end of the channel ($y = 2H$) is calculated from :

$$d_w = b - x_1 \text{ with } \Psi(x_1, 2H) = \Psi_w \tag{10}$$

  $\Psi_w$ the value of the stream function along the right plate arbitrary set to zero.

- The length of downward flow in top-end section is also deduced from the stream function set to zero: $l_w = 2H - y_1$ where $y_1$ is the such that $v(y_1) = 0$ along the right plate.

### 2.5 The channel geometry

## 3 RESULTS

### 3.1 Impact on dynamic quantities and on flow field

Thermal stratification has a more significant influence on dynamic quantities than surface radiation. Figure 3 displays the vertical component of the velocity in different horizontal sections of the channel. Major differences are shown in the entrance section at $y = 0$. Indeed, in thermal stratification case, the velocity at $y = 0$ consists essentially of a flat profile and it turns out to be parabolic without thermal stratification. The taking account of surface radiation do not modify the profile of vertical component of the velocity. This flat velocity profile at the entrance makes the mass flow rate in the channel decline. As reported in Table 1, the mass flow rate at the bottom of channel significantly decrease of 73% in thermal stratification case, and 65% with surface radiation. As a consequence, a plug-flow in the channel is produced. This result is in good agreement with the experimental studies realized by [8] and [14]. Therefore, thermal stratification and surface radiation reduces the mass flow rate at the bottom.

However, flow penetrates deeper into the channel in thermal stratification case and also with surface radiation. The development of the dynamic layer along the left plate wall at the entrance

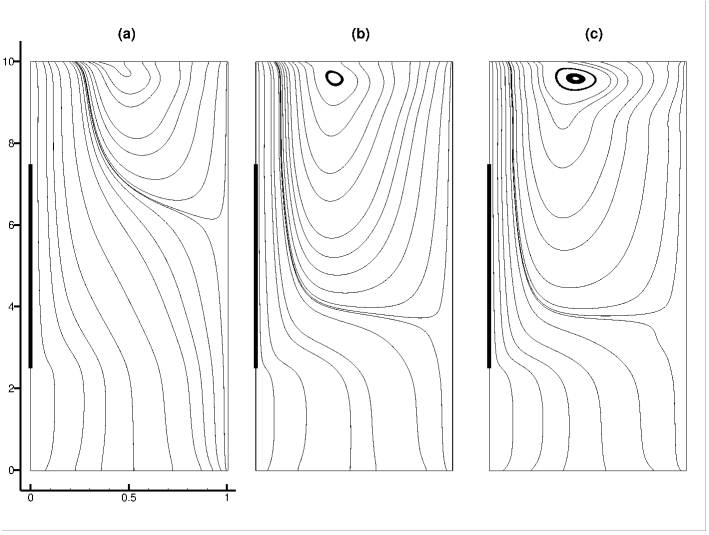



Figure 2: Streamtraces for (a) Reference case, (b) Thermal stratification case and (c) Thermal stratification with radiation

of the heated region at $y = \frac{H}{4}$ is visible in Figure 3. The return flow penetrating through the top section at $y = H$ is also visible with a negative value of vertical component of velocity. This observation is confirmed by the accurate measurements of $d_w$ and $l_w$ which are reported in Table 1. The depth of the pocket of recirculation increases of 67%. As the same, the mass flow rate in entrance at the top of the channel significantly increases: +144% in thermal stratification case and +207% when considering surface radiation. Figure 2 shows the pocket-like streamlines.The taking account of surface radiation modifies the flow structure much at the top section than in the remainder of the channel. This recirculation, fed by fluid drawn into the top section of the channel adjacent to the adiabatic wall, creates a vortex at the top section of the channel. Moreover, the size of the vortex increases by considering surface radiation and setting a temperature to grey bodies surface. Observation of vortex at the exit of channel has been already made by [11] and [18]. So, thermal stratification produces a vortex-like structure at the top region of the channel. The taking account of surface radiation alters significantly the structure of the flow at the top region of the channel. This can be explained by the influence of the temperature of the surface body of the top aperture of the channel.

Although all these effects, the flow stays in steady-state in thermal stratification case and with surface radiation.

| Dynamic quantities | $\dot{m}_{in}$ | $\dot{m}_{out}$ | $\dot{m}_{es}$ | $d_w$ | $l_w$ |
|---|---|---|---|---|---|
| Reference Case | 5.91 | 5.91 | 1.61 | 0.61 | 4.05 |
| Thermal stratification Case | 1.58 | 1.58 | 3.93 | 0.68 | 6.78 |
| Radiation and thermal stratification case | 2.05 | 2.05 | 4.94 | 0.90 | 6.22 |

Table 1: Dynamic quantities

### 3.2 Impact on thermal quantities

Temperature field is not significantly altered by thermal stratification and surface radiation. Figure 4 shows the effect of thermal stratification on the temperature distributions along the

two vertical walls. As it can be seen, dimensionless temperature of right plate increases range from $-0.01$ below the channel bottom-end to $0.002$ above the channel top-end when a low thermal stratification. Right wall temperature remains low and close to that of the fluid entering at the bottom of the channel. The taking into account of surface radiation slows down its heating. In the reference case, the maximum temperature is reached at the upper region of the channel instead of the mid-height in other configuration cases. So, as observed by [8] in their experimentation, thermal stratification increases slightly plates temperature. However local and mean Nusselt indicated in Table 2 decrease up to -2.5% in presence of thermal stratification and surface radiation. As a consequence, heat thermal transfer is reduced. This result is in good agreement with [8].

| Thermal quantities | $Nu_1(\frac{H}{2})$ | $\overline{Nu_1}$ |
|---|---|---|
| Reference case | 6.27 | 6.74 |
| Thermal stratification case | 6.16 | 6.62 |
| Radiation and thermal stratification case | 6.16 | 6.56 |

Table 2: Thermal quantities

## 4 CONCLUSIONS

In presence of a small thermal stratification outside of the channel, the flow structure at the top end of the channel is strongly impacted : the depth of the recirculation flow and the mass flow in entrance at the top section increase. The principal influence is the creation of a vortex penetrating the channel at the top end. The impact of surface radiation is essentially observed at the top section of the channel, where the dimensionless temperature of grey body has been set to 0.03. The vortex observed is larger and streamlines are tighter near the left wall.

Thermal quantities are slightly altered by thermal stratification and surface radiation. Temperature of the right wall increases but remains low and close to the temperature of the fluid entering at the bottom of the channel.

In conclusion, thermal stratification outside the channel has a significant effect on the flow structure and on dynamic quantities. In present study, thermal stratification has been investigated with a weak constant variation at the bottom of the channel. It could be interesting to more quantify the thermal stratification effect on flow and identify others sensible parameters like pressure or aspect ratio.

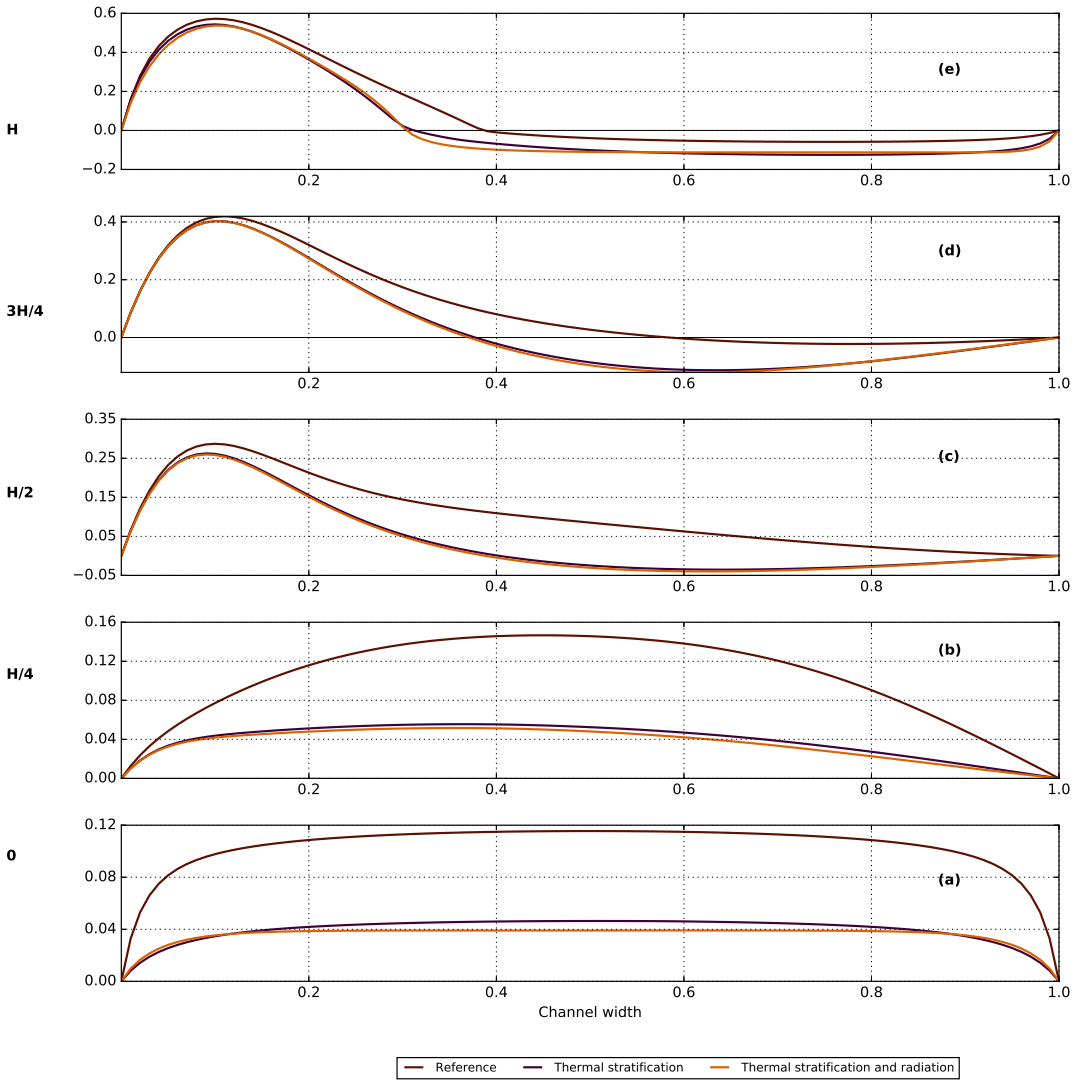



Figure 3: Vertical velocity profiles (a) at the bottom of the channel(y=0), (b) at the inlet of heated wall section (y=H/4), (c) at the mid-height of the channel (y=H/2), (d) at the end of the heated wall section (3H/4), (e) at the top section of the channel (y=H)

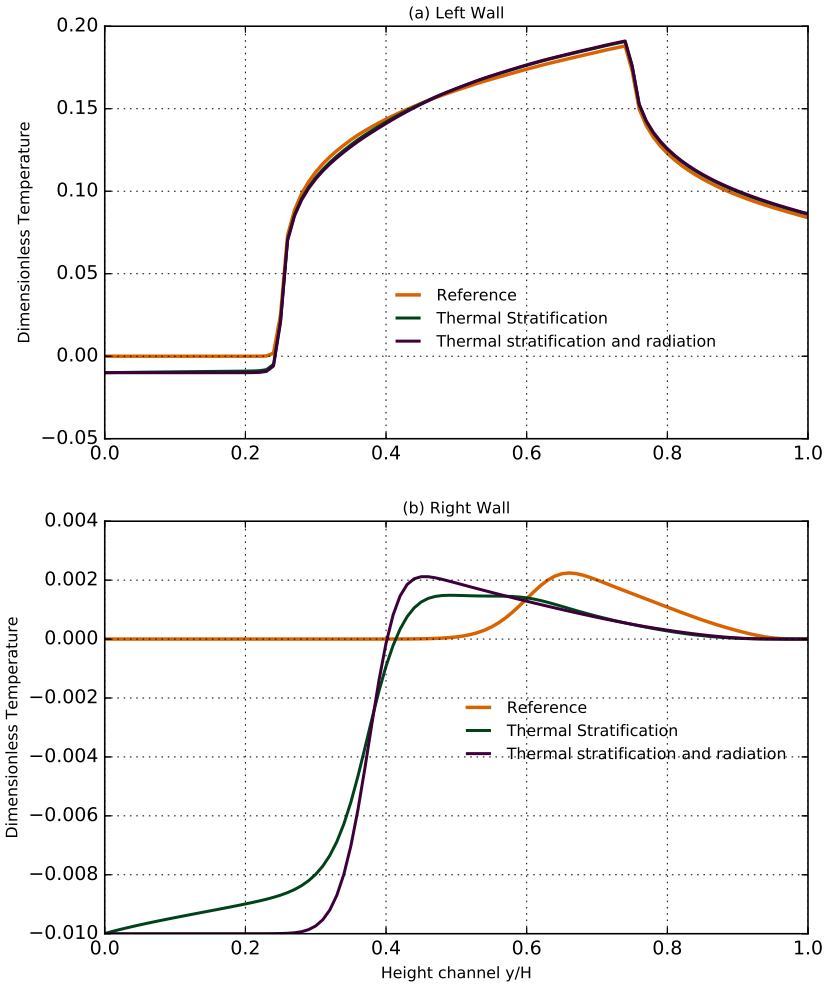



Figure 4: Dimensionless temperature distribution along the vertical channel walls: (a) Adiabatic wall, (b) Heated wall

## REFERENCES


[1] Toshio Aihara. *Effects of inlet boundary-conditions on numerical solutions of free convection between vertical parallel plates*. Tohoku University, 1973.

[2] A. Auletta, O. Manca, B. Morrone, and V. Naso. Heat transfer enhancement by the chimney effect in a vertical isoflux channel. *International Journal of Heat and Mass Transfer*, 44(22):4345–4357, 2001.

[3] G S Barozzi, M A Corticelli, and E Nobile. Numerical simulation of time-dependent buoyant flows in an enclosed vertical channel. *Heat and mass transfer*, 35(2):89–99, 1999.

[4] J R Bodoia and J F Osterle. The Development of Free Convection Between Heated Vertical Plates. *Journal of Heat Transfer*, 84(1):40–43, feb 1962.

[5] B. Brangeon, P. Joubert, and A. Bastide. Influence of the dynamic boundary conditions on natural convection in an asymmetrically heated channel. *International Journal of Thermal Sciences*, 95:64–72, sep 2015.

[6] J R Carpenter, D G Briggs, and V Sernas. Combined Radiation and Developing Laminar Free Convection Between Vertical Flat Plates With Asymmetric Heating. *Journal of Heat Transfer*, 98(1):95–100, feb 1976.

[7] A-M Dalbert, F Penot, and J-L Peube. Convection naturelle laminaire dans un canal vertical chauffé à flux constant. *International journal of heat and mass transfer*, 24(9):1463–1473, 1981.

[8] C Daverat, H Pabiou, H Bouia, S Xin, and Moret Loing. Convection naturelle dans un canal vertical en eau avec chauffage pariétale : influence de la stratification. In *Congrès Français de Mécanique*, pages 1–6, 2011.

[9] G. Desrayaud, E. Chénier, a. Joulin, a. Bastide, B. Brangeon, J.P. P. Caltagirone, Y. Cherif, R. Eymard, C. Garnier, S. Giroux-Julien, Y. Harnane, P. Joubert, N. Laaroussi, S. Lassue, P. Le Quéré, R. Li, D. Saury, a. Sergent, S. Xin, and a. Zoubir. Benchmark solutions for natural convection flows in vertical channels submitted to different open boundary conditions. *International Journal of Thermal Sciences*, 72(January 2009):18–33, oct 2013.

[10] W. Elenbaas. Heat dissipation of parallel plates by free convection. *Physica*, 9(1):1–28, jan 1942.

[11] Dupont Frédéric, Ternat Fabien, Samot Sandrine, and Blonbou Ruddy. Two-dimension experimental study of the reverse flow in a free convection channel with active walls differentially heated. *Experimental Thermal and Fluid Science*, 47:150–157, 2013.

[12] Guohui Gan. Impact of computational domain on the prediction of buoyancy-driven ventilation cooling. *Building and Environment*, 45(5):1173–1183, 2010.

[13] Guohui Gan. Simulation of buoyancy-driven natural ventilation of buildingsImpact of computational domain. *Energy and Buildings*, 42(8):1290–1300, 2010.

[14] Charles Garnier. *Modélisation numérique des écoulements ouverts de convection naturelle au sein d'un canal vertical asymétriquement chauffé*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2014.

[15] R. Li, M. Bousetta, E. Chénier, and G. Lauriat. Effect of surface radiation on natural convective flows and onset of flow reversal in asymmetrically heated vertical channels. *International journal of thermal sciences*, pages 9–27, 2013.

[16] J P Liu and W Q Tao. Numerische Ermittlung der natürlichen Konvektion um einen vertikalen Kanal in einer rechteckigen Kammer, 1996.

[17] O Manca, B Morrone, and V Naso. A numerical study of natural convection between symmetrically heated vertical parallel plates. In *Atti del XII Congresso Nazionale sulla Trasmissione del Calore*, pages 379–390. WIT Press, 1994.

[18] Dan Ospir, Catalin Popa, Cristian Chereches, Guillaume Polidori, and Stéphane Fohanno. Flow visualization of natural convection in a vertical channel with asymmetric heating. *International Communications in Heat and Mass Transfer*, 2012.

[19] S Ramanathan and R Kumar. Correlations for natural convection between heated vertical plates. *Journal of heat transfer*, 113(1):97–107, 1991.

[20] M. Sandberg and B. Moshfegh. Investigation of fluid flow and heat transfer in a vertical channel heated from one side by PV elements, part II - Experimental study. *Renewable Energy*, 8(1-4):254–258, may 1996.

[21] B W Webb and D P Hill. High Rayleigh Number Laminar Natural Convection in an Asymmetrically Heated Vertical Channel. *Journal of Heat Transfer*, 111(3):649–656, aug 1989.

[22] Amine Zoubir. *Etude des transferts thermo-convectifs dans un canal semi-ouvert : Application aux façades type double-peau*. PhD thesis, L'institut national des sciences appliquées de Lyon, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# INFLUENCE OF ADJACENT ROOMS ON THE DEVELOPMENT OF GAS EXPLOSION

**Polandov Iu.Kh.*[1], Dobrikov S.A. [1]**

[1] Orel State University, Scientific - educational center "Fluid Mechanics. Combustion", Naugorskoe shosse 29, Orel, Russia, 302020
E-mail; polandov@yandex.ru

**Keywords**: room gas explosion, the place of ignition.

**Abstract**. *The physical experiment conducted at the Moscow State University of Civil Engineering (MSUCE) has revealed that the pressure of a gas explosion in a room with a window and a hole in an adjacent room in the more than two times higher than a similar explosion in a single room. In the experiment, the ignition is located in the center of the room and there was no gas in the adjacent room.*

*Our numerical experiments confirm this fact. Calculations were performed using the software "Vulkan-2M", which is based on the method of large particles. Simulated room consisted of two chambers size of 0.5m x 0.5m x 0.5m each with a hole size of 0.15m x 0.15m between them. The first chamber was filled with a stoichiometric mixture of propane-air and had a window size of 0.15m.*

*Pressure increasing effect reaches a maximum when the ignition is in the area between the window and the center of the first chamber. The pressure is increased more than twice in this case in comparison with a single blast chamber. The effect is minimal when the ignition is closer to a window or a hole between the chambers. These results can be used to assess the risk of explosion.*

## 1 SHORT REWIEV

Researchers from Moscow State University of Civil Engineering (MSUCE) (Mishuev A.V., Kazionnov V.V., Gromov N.V. and others*)* [1] described the effect of increasing the pressure of the explosion in the presence of an adjacent chamber, connected to the vented explosion chamber. They found an increase in overpressure of 2.5 times compared to the same explosion without a hole between the chambers. These results are consistent with earlier studies of Molkov V.V. [2].

The rules of the European Union [3, 4], the USA [5, 6, 7] and Russia [8, 9, 10] does not take into account the effect of the adjacent cameras on the development of explosion. In spite of the different standards, the style and the different calculation formulas, all these documents have a common drawback. Their results based on the quasi-static pressure in the room that is not taken into account the distribution of the parameters of the gas in the chamber volume. The method of quasi-static pressure makes it possible to determine the shape and the flame front area based only on the phenomenological models, in which there is no turbulence. It is not possible to use such a model to account for the maximum pressure in the event of an adjacent room. Only experimental researches or CFD-methods can solve this problem. However, such studies have not been conducted in Europe, even in such research centers as the FM Global (Toronto, USA) [11], known for his experimental work.

Article by software developers FLACS [12] close to the theme of this work. This article describes the development of simulation of a gas flame in the two adjacent chambers filled with gas. However, the authors are more interested in the accuracy of numerical model than in the mechanics of explosion and hazard assessment.

Vodyanik V.I. [13] also considered the development of a flame inside two communicating vessels. In his experiment, the gas fills both of the vessel before the explosion, so that the results are not relevant to the subject.

To this is the closest Molkov's [2] study in which he analyzes pilot burst in the tank from which the volume of 2 $m^3$ gas out through line into another tank volume 3.5 $m^3$ which was filled with air. He found that in some experiments the pressure in the receiving vessel during the explosion becomes greater than the pressure in the first tank.

In addition, he established the dependence of explosion from the ignition location of the device in the first tank, which allowed him to explain the reason for this effect. Unfortunately, the studies were not specific for the explosion in the room such as the kitchen where there is a window through which the gases vented to the atmosphere apart holes between rooms.

Thus, the research of scientists from MSUCE are new in this issue.

## 2 THE HYPOTHESYS

It is known that the place of ignition of gas in the room greatly affect the development of the explosion [3, 5, 11]. These researches show that the farther the place of gas ignition from vent, the greater the maximum explosion pressure. Experimental studies were conducted which confirmed earlier [14, 16] these data, it was found that the impact of this effect (from 0 to 20 times with the difference from maximum pressure) is more complex in dependence on the window size and shape of the room. In this context, and based on the experience Molkov V.V. We can expect that somewhere a source of ignition in explosions in two adjoining rooms, one of which is filled with gas, and the window will also affect the development of the explosion.

We offer three versions of the explosion in this case.

The first case. The source is located near the hole in the next room. The products of combustion will flow into the next room, not gas-air mixture. In this case, the adjacent room will act as a damper.

The second case. Upon ignition of the mixture near the center of the room, it can be expected that the gas mixture will fall into the next room, where it will burn. However, hole will turbulence mixture, it should dramatically increase the area of the flame front and explosion pressure.

The third case. The gas is ignited near the window. In this case, the maximum amount of mixture enters the next room. But it enters into force known pressure lowering effect of an explosion if ignited near a window. The explosion pressure will depend on the ratio of these two factors.

Therefore, we decided to investigate the relationship between the maximum pressure of gas explosion and ignition of the gas mixture space.

## 3 RESEARCH TOOLS

In drawing up the mathematical model of the process, we have made the following assumptions concerning the simulated environment:

1. The initial mixture of propane-air is a homogeneous and stoichiometric;
2. The difference between the thermodynamic characteristics of the original mixture and the combustion products is negligible;
3. The gases in the physical process is inviscid and are ideal fluid;
4. The combustion reaction occurs at the boundary of the original mixture and the combustion products.

Given the assumptions, the problem is reduced to modeling the dynamics of the gas with uniform properties by using one of the methods for the unsteady multidimensional problems of fluid mechanics (CFD). The choice of a particular method is limited by arbitrary geometry of the computational domain, as well as the possibility of taking into account the availability of features in the simulated currents. As a basic system of equations to describe the dynamics of the medium was used known system of Euler equations in divergence form, closable equation of state.

On the domain of integration is superimposed Euler (fixed) grid of rectangular cells with sides $\Delta x$, $\Delta y$ и $\Delta z$. Numerical solution of the system is carried out by large particles method, LPM [16], which is based on the idea of Harlow n particles in the cell, allowing splitting of physical processes. However, in the LPM instead small solids particles replaced by a single a drop of liquid, fill the entire volume of the cell. Method of large particles as well as other modern methods such as Godunov method [17], FLACS [12] et al., allow us to study the gasdynamic flow without a priori information about the structure of the solution. The calculation consists of repetitive time steps. In turn, each such step includes three steps:

1. "Euler" stage, when neglected all effects associated with the movement of the fluid (mass flow through the faces of the cells is not);
2. "Lagrangian" stage, where the calculated mass flow through the cell boundaries;
3. The final stage, which determines the final flow parameters based on conservation laws for each cell and the entire system as a whole.

The system includes equations that describe the process of heat and mass transfer with the environment and the process of propagation the flame. Cooling processes on the chamber walls are estimated based on physical experiments carried out according to pressure drop in the explosion in a closed chamber. To calculate the flow through the open border to border

pressure cell is assumed equal to the average between the pressures in the chamber and atmospheric.

Two additional parameters, "the mass fraction of combustion products in the chamber," and "the mass fraction of the formation of the products of combustion", introduced for the simulation of flame propagation. The first parameter in the cell determined by the position of the flame front, which moves toward the initial mixture at a rate equal to the sum of the gas flow velocity and the normal combustion rate. The second parameter, the mass fraction of newly formed products of combustion is determined by taking into account only the normal burn rate, and takes into account only if the flame front is located within the cell.

In the system of equations in the equation of conservation of momentum, we have taken into account the effect of buoyancy.

The computer model uses 7 types of cells:

– calculated cells - which are carried out of the equation;
– border cells provide impermeability condition;
– flow cells - which connect the cells of calculated area with the atmosphere;
– flame front - cells in which combustion occurs;
– cells with the products of combustion;
– air cells - which are filled with air;
– non-calculated cells - which are located behind the border and fix their properties.

Fig. 1 shows a diagram of the calculated volume that is represented by two cameras, one of which (left) filled with gas, and the second - the air. Overall dimensions make 0.5x0.5x1.0 m. Designated ignition of the gas mixture are in the left chamber. Ribs cells taken equal $\Delta x=\Delta y=\Delta z=0.01$ m, so that only received about 100,000 cell, the time step $\Delta t=5\cdot10\text{-}7$ s with that by a wide margin meets the criterion of stability Courant - Friedrichs - Lewy. This margin is accepted, because the equations describing the propagation of flame, adversely affect the stability of the account.

On fig. 2 shows a typical picture of the explosion. In this case, the vector lines are shown as lines Bezier, flame front - red cells, and other gases are painted on a scale that the higher the temperature, the lighter background.

## 4 THE SIMULATION RESULTS

*a) An explosion in solitary chamber*

In the first stage, experiences were conducted with a gas explosion in a solitary chamber with disconnecting adjacent chamber. The simulation results shown in figure 2.

It is seen that with the appearance of the flame front starts expiration of the initial mixture to the atmosphere. The flame front initially increases as a sphere, and then starts to be pulled towards the vent. As soon as the flame front will come to the vent, begin to expire the flame cell. In the last step, expire only products of combustion. Released source mixture of gas burns outside the chamber in the atmosphere.

According to the data, is shown in figure 3, it is evident that the time of fire is not exceed 0.6 seconds at any position of the igniter in chamber. In addition, it traced unique relation between the parameters of the explosion and the place of ignition of gas.

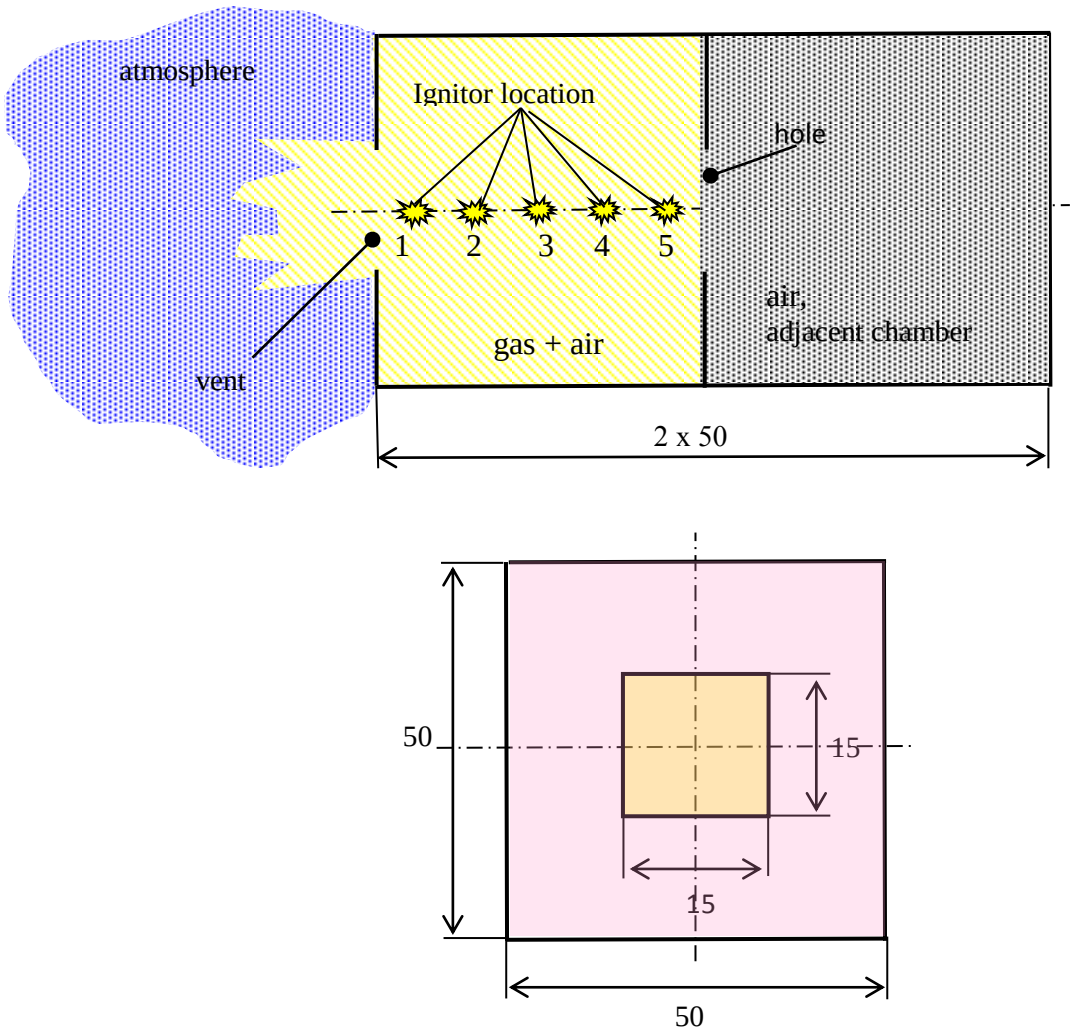



Figure 1: Calculated scheme 2-chamber unit (dimensions in cm)

Firstly, the farther from the window is igniter, the shorter the time of the explosion, the more the primary mixture is ejected in atmosphere out and flame particles. Second, the farther from the window ignition occurs, the greater the pressure of an explosion (fig. 4), in this case a factor of 2, which confirms the above results known. In a graph, this result is shown in fig. 7 (curve 1).

*b) explosion in chamber in the presence of adjacent chamber*

At first glance, the presence of the adjacent chamber with the explosion (fig. 4) should lead to a simple damping process in its dynamic development. It is certainly true, but not entirely so. At first, with increasing pressure in the first chamber starts flowing gas mixture it into the next. Due to the leak, mixture begins to increase the pressure in the adjacent cell. Therefore, the amount of the mixture received into an adjacent chamber can be estimated by the pressure, which is reached in an explosion. In this case, the excess pressure is from 6 to 20 barg. Therefore, in the adjacent volume chamber gases arrive from 6 to 20% of the chamber volume. This, of course, raises the question, what kind of gas is supplied to an adjacent

camera? The answer is quite simple. Composition of gas entering in the adjacent chamber it depends on whether approached to the hole of the flame front or not. If you have not yet approached the combustion front, only the initial mixture is fed into an adjacent chamber when approached, they begin to enter the combustion products.

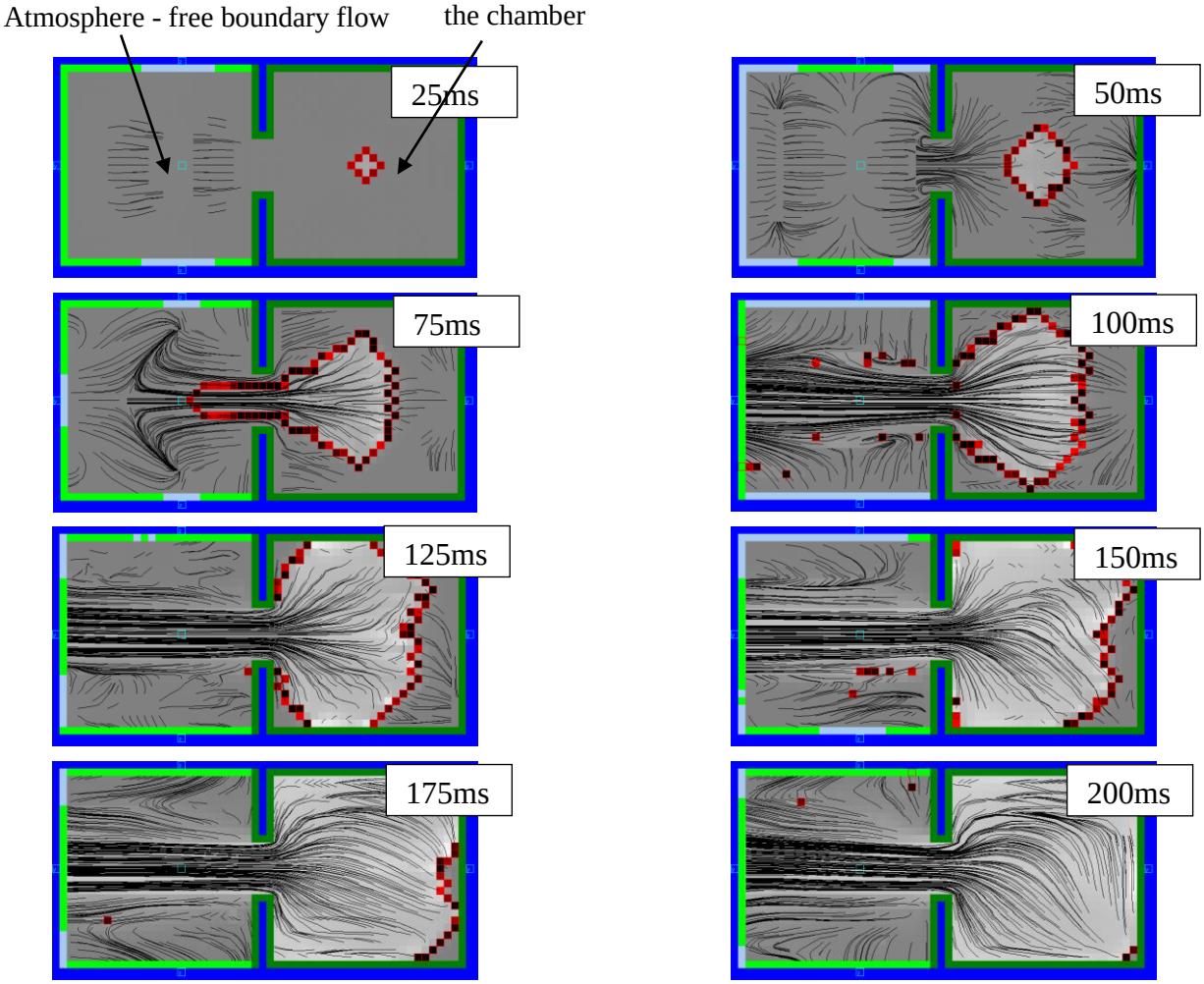



Figure 2: The picture of gas explosion in solitary chamber when ignition in position 3

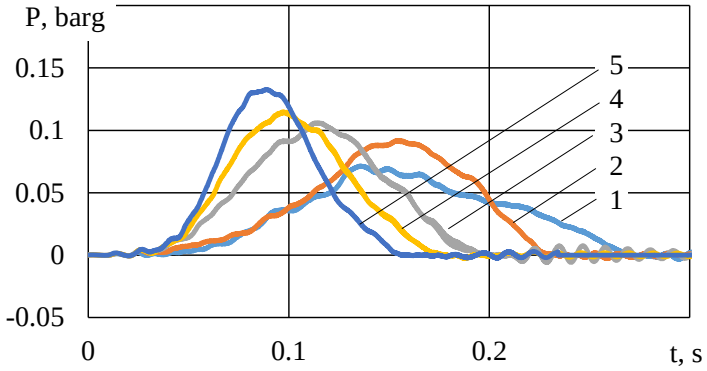



Figure 3: Dynamics of explosion pressure depending on the position igniter in the solitary chamber

The change in time of pressure in the explosion chambers is shown in fig. 5 and 6. The number of each curve corresponds to the number igniter place according to fig. 1. The greatest pressure reaches values in position 2. When you install the ignition at the passage into the adjoining chamber, the pressure of the explosion nearly corresponds to the pressure in solitary confinement. In addition, when installing igniter close to the vent, pressure is lower than the maximum, but higher than in solitary confinement.

It is understood that the earlier the ignition of the mixture in the adjacent cell, the higher pressure will be in the first chamber. This effect is particularly strong if the ignition happens when the pressure in the first chamber increases. However, the closer to the passage of ignition occurs, in the adjacent room is less than the initial mixture. In this case, the "help" the adjacent room to increase the pressure will decrease.

*c) Comparing the results of numerical and physical experiments*

Comparison of numerical simulation results and the results that have been obtained researchers at the MSUCE shown in figure 7.

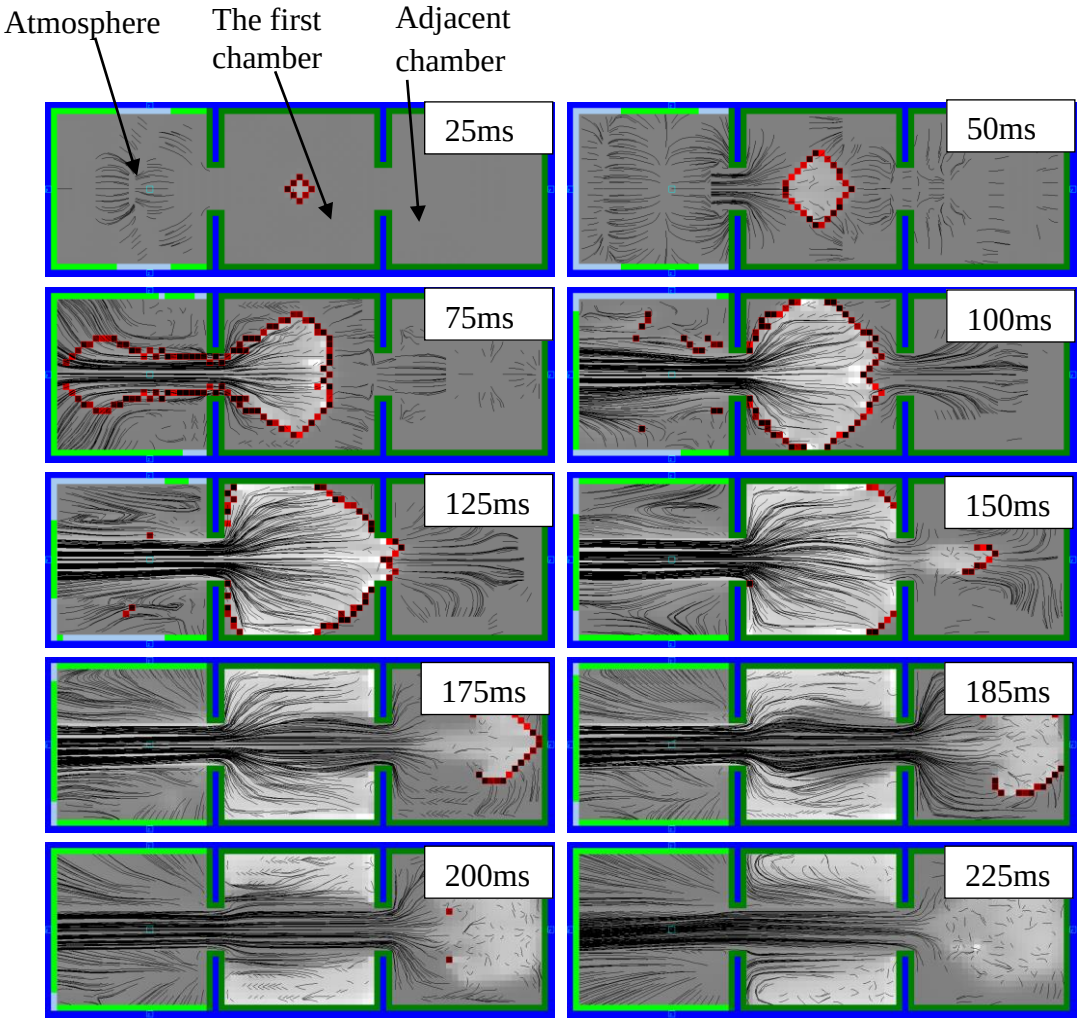



Figure 4: The picture of gas explosion in solitary confinement at position igniter 3

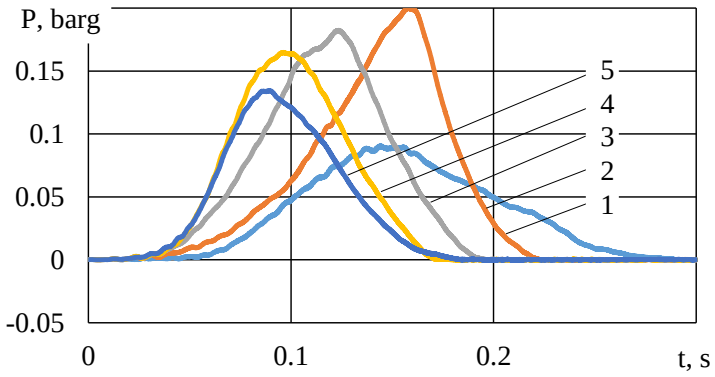



Figure 5: The development of pressure in the explosion in the chamber together with the adjacent chamber at different positions ignition

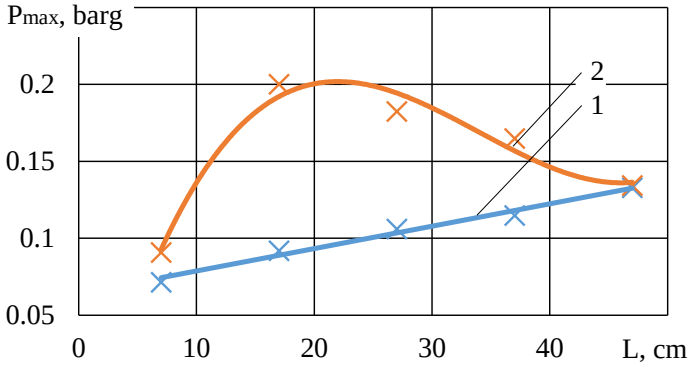



Figure 6: Regression curves $P_{max} = P(L)$,
where $P_{max}$ - maximum pressure in the explosion;
L - the distance from the vent, cm;
1 - single chamber; 2 - together with the adjacent

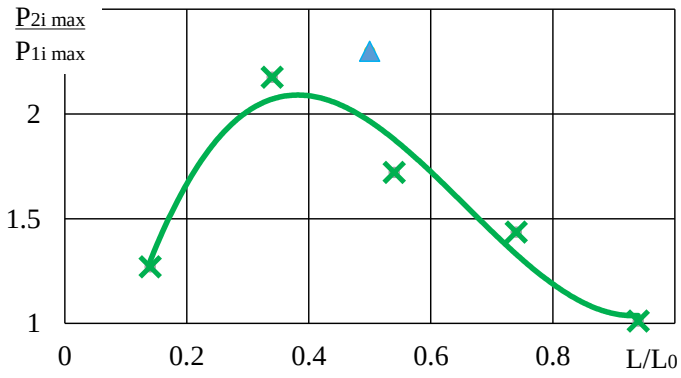



Figure 7: Comparison of the results of computing and physical experiments
Symbol Δ denotes the result of a physical experiment in MSUCE.

It must be borne in mind that the installation had different geometrical dimensions, although the ratio $(V^{2/3} / F)_c = 11.1$, and $(V^{2/3} / F)_e = 12$, where $V$ - volume of the first chamber, $m^3$ and $F$ - $m^2$ vent area, both systems have similar values. If in addition to this, use the relative pressure and the distance from the ignition space by entering them in the calculations

and the physical experiment, it seems a formal opportunity to make such a comparison. It is best to do this by plotting $P_{2i\ max}/P_{1i\ max}=f\ (L/L_{0)}$, where $P_{2i\ max}$ - maximum explosion pressure with the adjacent chamber in 1th position; $P_{1i\ max}$ - the maximum pressure in the explosion in solitary chamber in the 1 - position; and $L/L_0$, where $L$ - distance from Vent to the place of installation of igniter and $L_0$ - distance from vent to pass into the adjoining chamber. We see an acceptable coincidence. It is understood that the proposed dimensionless quantities not fully correspond to the concept of criteria but they allow somehow compare with each other the results obtained under various conditions.

## 5 CONCLUSION

These results confirms the hypothesis about the influence of gas ignition locations for the development of the explosion in the chamber. It was found that when ignited gas near passage into the adjoining chamber explosion pressure is almost the same as in an explosion in a separate room. Influence of combustion is shown at center of the chamber and, especially, in the space between the center and vent. However, when ignited near the vent begins to affect other effect: release of combustion products with a high temperature of the flame, resulting in the effect of the adjacent rooms on the development of the process is reduced.

## REFERENCES

[1] Mishuev A. V., Kazyennov V. V., Komarov A. A., Gromov N. V., Lukyanov A. V., Prozorovskiy D. V. Osobennosti avariynykh vzryvov vnutri zhilykh gazifitsirovannykh zdaniy i promyshlennykh obyektov [Peculiar properties of emergency blasts inside gasified buildings and industrial objects]. Pozharovzryvobezopasnost – Fire and Explosion Safety, 2012, vol. 21, no. 3, pp. 49–56.

[2] Molkov V. V. Dinamika sgoraniya gaza v negermetichnom sosude. Dis. kand fiz.-mat. nauk [The dynamics of the combustion gas in the sealed container. [Cand. phys. and math. sci. diss.]. Moscow, Moscow Engineering Physics Institute Publ., 2006. 211 p.

[3] EN 14994:2007. Gas explosion venting protective systems. BSI, 2007. Available at: http://shop.bsigroup.com/ProductDetail/?pid=000000000030117073 (Accessed 25 October 2015).

[4] EN 14797:2006. Explosion venting devices. BSI, 2007. Available at: http://shop.bsigroup.com/ProductDetail/?pid=000000000030104794 (Accessed 25 October 2015).

[5] NFPA 68:2007, Standard on explosion protection by deflagration venting. USA, Quincy, National Fire Protection Association, 2007.

[6] Ismaila A., Andrews G. E., Abdullahi I., Nasiru R, Abdullahi Y. A. Venting as a means of mitigating explosions: The need to revised European and USA (NFPA68) guidance for explosion venting. *Archives of Applied Science Research*, 2012, vol. 4, no. 1, pp. 155-168.

[7] Bauwens C. R., Chaffee J., Dorofeev S. Experimental and numerical study of methane-air deflagrations in a vented enclosure. *Proceedings of the Ninth International Symposium "Fire Safety Science"*, vol. 9, pp. 1043–1054. DOI: 10.3801/IAFSS.FSS.9-1043.

[8] State standard 12.3.047–2012. Occupational safety standards system. Fire safety of technological processes. General requirements. Methods of control. Moscow, Standartinform Publ., 2014. 65 p. (in Russian).

[9] Set of rules 12.13130.2009. Determination of categories of rooms, buildings and external installations on explosion and fire hazard. Moscow, All-Russian Research Institute for Fire Protection of Emercom of Russia Publ., 2009. 27 p. (in Russian).

[10] Set of rules 89.13330.2012. Combustion boiler systems of heating generation. Moscow, FAU "FCC" Publ., 2013.

[11] Bauwens C. R., Chaffee J., Dorofeev S. Effect of ignition location, vent size, and obstacles on vented explosion overpressures in propane-air mixtures. *Combustion Science and Technology*, 2010, vol. 182, issue 11-12, pp. 1915–1932. DOI: 10.1080/00102202.2010.497415.

[12] Helene Hisken Pedersen, Gary Tomlinc, Prankul Middhab, Herodotos Phylaktouc, Gordon Andrewsc. Comparison of FLACS Simulations against large-scale vented gas explosion experiments in a twin compartment enclosure.

[13] Vodyanik V. I. Vzryvobezopasnost tekhnologicheskogo oborudovaniya [Explosion technological equipment]. Moscow, Khimiya Publ., 1991. 254 p.

[14] Polandov Yu. H., Barg M. A., Vlasenko S. A. Modelirovaniye protsessa goreniya gazovozdushnoy smesi metodom krupnykh chastits [Simulation of combustion gas mixture by large particles]. *Pozharovzryvobezopasnost – Fire and Explosion Safety*, 2007, vol. 16, no. 3, pp. 6–9.

[15] Belocerkovskii O. M., Davidov Yu. M. Method of large particles in the gas dynamics. Moscow, Science, 1982. 370.

[16] Polandov Yu. Kh., Babankov V. A. Vliyaniye mesta raspolozheniya istochnika vosplameneniya v pomeshchenii na razvitiye vzryva gaza Effect of location source of fire in the room on the development of gas explosion]. *Pozharovzryvobezopasnost – Fire and Explosion Safety*, 2014, vol. 23, no. 3, pp. 68–74.

[17] Godunov S. K., A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics, *Mat. Sb. (N.S.)*, 47(89):3 (1959), 271–306

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DOMAIN DECOMPOSITION APPROACH FOR NEAR-WALL TURBULENCE MODELING

**Sergey V. Utyuzhnikov** [1], [2]

[1] School of MACE, University of Manchester,
Manchester, M13 9PL, UK
e-mail: s.utyuzhnikov@manchester.ac.uk

[2] Moscow Institute of Physics and Technology,
Dolgoprudny 141700, Russia

**Keywords:** Near-wall Turbulence Modeling, Domain Decomposition, Wall Functions, Reynolds Models.

**Abstract.**

*Turbulence modeling leads to a multiscale problem in which there is a very thin boundary layer that requires most computational time for its resolution. The paper addresses a novel approach based on non-overlapping domain decomposition. It allows us to avoid calculations of the region with high gradients in the vicinity of the wall while retaining sufficient overall accuracy. The technique is introduced in application to low-Reynolds number RANS models. The domain decomposition is achieved via the transfer of the boundary condition from the wall to an interface boundary. If the governing equations in the inner domain are simplified, then the interface boundary conditions are of Robin type. These boundary conditions can be obtained in an analytical form despite the fact that they are nonlinear. Possible ways to achieve a reasonable trade-off between efficiency and accuracy are discussed.*

*The obtained interface boundary conditions are mesh-independent. They can be used to avoid computationally expensive resolution of a high-gradient region near the wall. Moreover, once the solution is constructed in the outer region, the near-wall profile can be restored if required. In two extreme cases, if the interface boundary is too close to the wall or too far from it, the so-constructed solution to the problem automatically corresponds to the low- and high-Reynolds number RANS models, respectively. Different applications are considered including unsteady problems and complex geometries. It is shown that in comparison to the low-Reynolds number models the near-wall domain decomposition approach allows the computational time to be reduced by one order of magnitude whilst practically retaining the accuracy.*

*The developed approach proved to be quite robust and relatively universal. It does not contain any tuning parameters. The technique might be extended to other multiscale problems.*

## 1 INTRODUCTION

Numerical modeling turbulent flows near walls is still computationally very expensive. The problem is that nearby the wall there is a very thin laminar boundary layer due to the no-slip boundary condition and damping effect of the wall. The thickness of the laminar sublayer is about 1% of the entire thickness of the boundary layer. However, resolution of this region takes up to 90% of the total computational time [17] because of high gradients of the solution. One possible way to avoid this problem is the use of high Reynolds number (HRN) models. In this case the flow structure above the laminar sublayer is deliberately not resolved. Instead of this the solution is represented by approximate boundary conditions which are usually set at the nearest to the wall cell. Such Dirichlet boundary conditions are often called the wall functions. In the core flow the governing Reynolds Averaged Navier-Stokes rquations (RANS) do not depend on the distance to the wall in contrast to the original low Reynolds number (LRN) RANS models. Thus, all information on the wall is supposed to be included in the wall functions. As noted in [11], this approach should be treated as a domain decomposition in which the solution in the inner near-wall layer is replaced by the wall functions.

As mentioned above, the wall functions represent an approximate solution in the inner layer. They can be obtained in one way or another one. First wall functions were based on the analytical solution for a plate [18], [12] and its semi-empirical modifications (see e.g. [13], [14]) to make them more universal. More advanced wall functions such as [15], [16] are not restricted by approximate local solutions. They take into account the solution in the outer (core) region via iterations. The solution in the inner region is obtained by solving boundary layer equations either numerically or analytically. In the latter case the solution of the governing equations is supposed to be approximate. In this case a domain decomposition can be realized with a Dirichlet-Dirichlet (D-D) iterative algorithm [19].

As shown in [9], the interface boundary condition (IBC) between the inner and outer regions can be obtained via transfer the boundary condition from the wall to the interface boundary. In this approach, the wall boundary conditions and governing equations are equivalently replaced by the IBC [10]. If the governing equations in the inner layer are of boundary-layer type and locally one-dimensional, the transfer of boundary condition can be exactly realized [10]. One can prove that in the case of a Dirichlet boundary condition set at the wall the IBC is always of Robin type. In a more general formulation the IBC must be nonlocal in both space and time [7], [3]. In [7] for a model equation it is demonstrated that if the flow along the wall changes rapidly enough, then the IBC must be nonlocal, otherwise essential effects can be lost. In turn, in [3] for a model unsteady equation it is shown that if the solution is essentially unsteady, then the IBC must be nonlocal in time and contain a memory term. In [4] the approach was successfully applied for modeling laminar-turbulent transition. In general the opportunity to replace the solution in the inner region by an interface pseudo-differential equation follows from the theory of generalized Calderón–Ryaben'kii's potentials that are proved to be projections [8]. The technique of boundary condition transfer can be applied to the LRN models straightforward [5]. In this case the distance between the interface boundary and wall is the parameter reflecting the trade-off between the accuracy and time consumption [1]. As shown in [1], in comparison with the LRN models the near-wall domain decomposition approach can reduce the computational time by one order of magnitude whilst the accuracy dropping by about 1%. In [1] the near-wall domain decomposition technique was implemented with different RANS models and successfully applied to a few industrial problems.

In the current paper the approach is described in a general formulation. The algorithm of

realization is given in detail. The approach is compared against standard methods. Examples of practical applications are provided. Some prospectives for future realization are also discussed.

## 2 NEAR-WALL DOMAIN DECOMPOSITION

Consider first a domain decomposition for a 1D non-linear equation defined in an interval $\Omega := [0,\ y_e]$:

$$\mathcal{L}_y(U) = f, \tag{1}$$

$$U\,(0) = U_0, \tag{2}$$

$$U\,(y_e) = U_1.$$

where $\mathcal{L}_y$ is a non-linear differential operator of second order.

Next, we split the interval $\Omega$ into two intervals $\Omega^- := [0,\ y^*]$ and $\Omega^+ := [y^*,\ y_e]$. To transfer the boundary condition from the wall ($y = 0$) to $y = y^*$ one can use the nonlinear potentials introduced in [6]. The IBC is formulated as follows:

$$U_y = \Psi(U, f). \tag{3}$$

Although being exact, such a boundary condition is essentially nonlinear and not easy to be used.

To simplify the IBC, presume that in the inner region the governing equation is of boundary-layer type: $\mathcal{L}_y(U) \equiv \frac{d}{dy}(\mu U)$, where $\mu$ is the efficient viscosity coefficient, then the IBC is reduced to a Robin boundary condition formulated at $y^*$ [10]:

$$U' = \frac{U - U_0}{I_1} + \frac{I_1 I_{S1} - I_{S2}}{\mu^* I_1}, \tag{4}$$

where $I_1 = \int_0^{y^*} \frac{\mu^*}{\mu} dy$, $I_{S1} = \int_0^{y^*} f dy$, $I_{S2} = \int_0^{y^*} \frac{\mu^*}{\mu} \int_0^{y'} f dy dy'$, $\mu^* = \mu(y^*)$.

In 1D case the boundary equation (4) fully replaces the solution of the problem in the inner region $\Omega^-$ for the outer region $\Omega^+$. This means the solution of boundary value problem (1), (2) must satisfy equation (4). In turn, it is easy to prove that condition (4) is unique. In the linear case equation (4) would be totally independent of the solution in the outer region. In our consideration the dependance is weak. It is only revealed via the turbulent viscosity coefficient that depends on the solution in the entire region. It is worth noting that it is impossible to obtain the solution in the outer region without the solution in the inner region. However, it is possible to transfer the boundary condition from the wall to the interface boundary without knowledge of the solution in the outer region even in the nonlinear formulation. In turn, it is impossible to obtain the solution in $\Omega^-$ with boundary condition (4) because this condition is the consequence of the governing equation and wall boundary condition.

The IBC are formulated in a universal form (4) for all variables but the normal to the wall velocity. The problem is that the conservation law of mass is not directly used in the IBC. It might be violated if the approximation does not guarantee the integral conservation law in the inner region as a whole. In this case a Dirichlet boundary condition for the normal velocity suggested in [1] can be used.

To obtain the solution in the entire region via the near-wall domain decomposition, the procedure is the following. First, one should obtain the solution in the outer region $\Omega^+$ with boundary

condition (4) at the left-hand side. Next, a boundary value problem (BVP) is to be solved in the inner region $\Omega^+$ with the Dirichlet boundary condition at the right-hand side. This boundary condition becomes known once BVP in the outer region has been solved. It is not always necessary to solve the BVP in the inner region since the friction at the wall is immediately related with the friction $\tau_w$ at the interface boundary:

$$\tau_w = \mu(y^*)\frac{dU}{dy} - \int_0^{y^*} f dy. \tag{5}$$

In boundary equation $(4)$, $\mu$ represents the sum of laminar and turbulent viscosity coefficients. It is to be noted that the IBC (4) is exact even in the nonlinear case when the turbulent viscosity coefficient, $\mu_t$, depends on the solution. To simplify the use of boundary condition (4), one can use available approximations for $\mu_t$ nearby the wall. Some examples are given in [1]. Apart from a piece-wise linear approximation used in [15] for the analytical wall functions, the others are nonlinear because they depend on the wall friction.

Thus, the entire algorithm is as follows:

$1^0$. Select an approximation for the near-wall turbulent viscosity coefficient $\mu_t$.

$2^0$. Select an initial approximation for $\tau_w$ (for nonlinear approximations of $\mu_t$).

$3^0$. Calculate IBC (4).

$4^0$. Solve BVP in the outer region with the IBC.

$5^0$. Update $\tau_w$ via (5)

$6^0$. Update the near-wall turbulent viscosity profile.

$7^0$. Repeat the procedure from stage $3$.

If necessary, in stage 5 BVP in the inner region can be solved with Dirichlet boundary conditions at $y^*$ obtained from the solution in the outer region. The governing equations in the inner region are sufficient to calculate the turbulent viscosity coefficient. In this case, stage 1 can be avoided. In multidimensional case, the IBC can be nonlocal [7]. Alternatively, in the inner region the governing equations can be of boundary-layer type because nearby the wall the normal derivatives are predominant. In this case, the transfer of boundary conditions can be locally one-dimensional.

It is interesting to compare the approach against the advanced-wall-function techniques [15], [16]. First, one should note that the wall functions are only applied to the HRN models. The key point is that the interaction between the inner and outer regions is carried out via D-D iterative methods. As noted above, the IBC (4) is unique regardless the approach. This means in the D-D algorithms with wall functions the derivative $U'$ is taken from the previous iteration. Since it is a leading term, the number of iterations should significantly increase. In turn, this damages the robustness. In the linear case the near-wall domain decomposition would be exact and not require iterations in contrast to the D-D algorithm. In the linear case the near-wall domain decomposition would be exact and not require any iterations in contrast to the D-D algorithm. It is to be noted that the IBC are mesh independent. They are formulated in a differential form. The coefficients in the IBC can be calculated either analytically or numerically. In the latter case the mesh in the inner region is not related with the mesh in the outer region at all. Therefore, the IBC can be easily implemented in a code via a separate routine.

As mentioned above, the introduced technique can be applied to LRN RANS models straightforward. If in the inner region the problem is solved approximately, then there is a trade-off between the computational time and accuracy. Consider two extreme cases. If $y^*$ tends to zero, then the solution model becomes equivalent to the LRN model. However, the effect of the approach vanishes. In turn, if $y^*$ is big enough, then the computational time should significantly

drop by the cost of accuracy. In this case the approach becomes similar to HRN models with respect to the accuracy of prediction because the laminar sublayer region is not resolved.

The approach represents an approximate non-overlapping near-wall domain decomposition (NDD). It is natural to expect that the computational time with the NDD should significantly be less than that without it. The justification is the following. The computational time to inverse an operator is proportional to its condition number. One can show that in the case of the Navier-Stokes equations the condition number is approximately proportional to the ratio between the maximal and minimal space steps squared. For the sake of simplicity, consider a mesh which is adaptive in the direction normal to the wall. Obviously, the condition number in the outer region should decrease rapidly as $y^*$ increases so should do the computation time. In turn, since the inner region is only approximately resolved, the error of prediction should inevitably decrease.

The trade-off between the computational time and accuracy was investigated in [1] for a two-dimensional asymmetric diffuser test case. The sketch of the diffuser is given in Fig. 1.

In Figure 2 a relative computational time is shown for the SST model [20]. Here, $e$ represents a relative error in computation of the skin friction. One can see that the near-wall domain decomposition can reduce the computation time by one order of magnitude whilst the error is just about one percent. The approach with the wall functions requires approximately the same time but the error is above $100\%$ in this case. The results are shown for the SST model (LRN), NDD and scalable wall functions [13] (HRN).

The skin friction distribution along the inclined wall at different $y^*$ is given in Figure 3. The solid dot line represents the experimental results by Buice and Eaton [21]. Computational results demonstrate that dependence of the solution on $y^*$ is not essentially sensitive to $y^*$ [10], [5], [1], [2] if $y^*$ is small enough. The variation of $y^*$ by one order of magnitude leads to the change of the solution as much as a few percent. Thus, a reasonable value of $y^*$ can be easily evaluated from either general information of the solution to be obtained or preliminary calculations on a coarse mesh.

In the future the approach can be extended to unsteady problems and nonlocal effects. Preliminary development of this technique was carried out in [3] and [5], respectively.

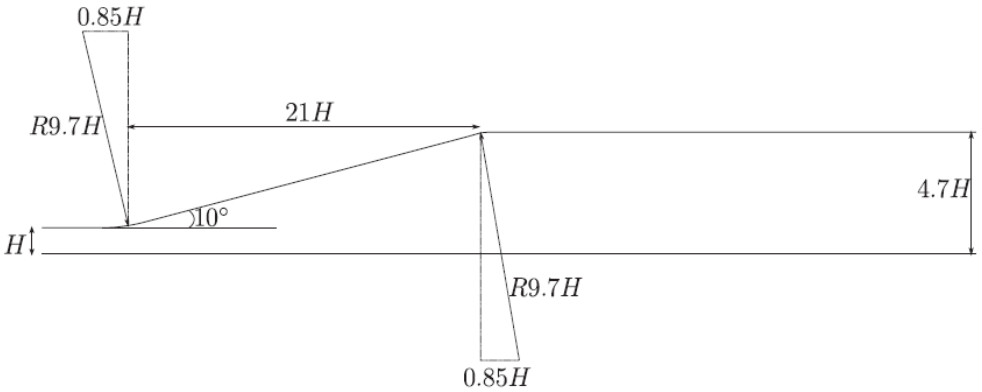



Figure 1: Two-dimensional asymmetric diffuser.

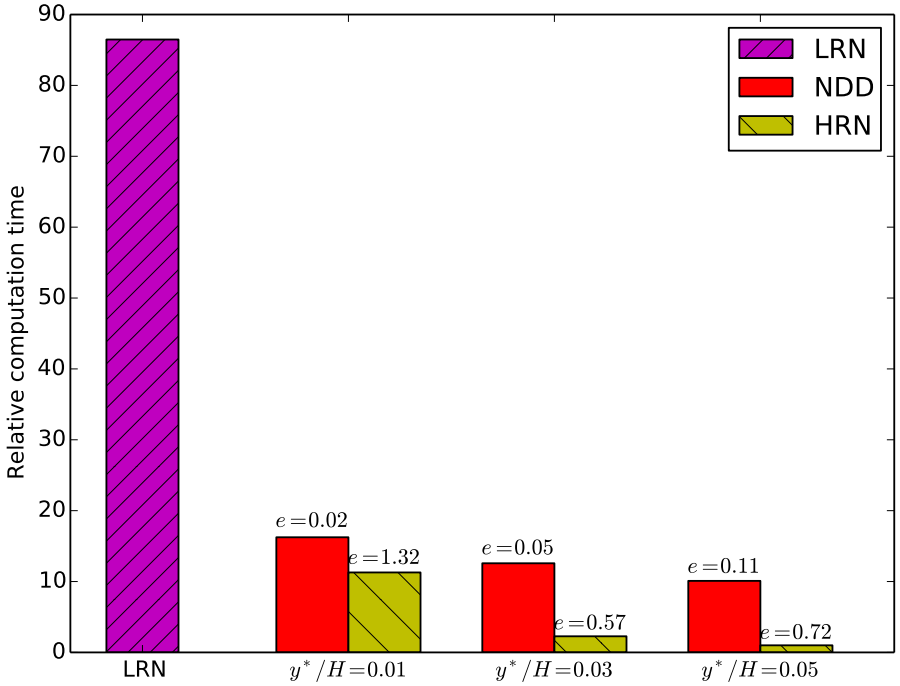



Figure 2: Trade-off between the accuracy and computational time for a two-dimensional asymmetric diffuser [1]. Low-Reynolds number SST model (LRN) [20], near-wall domain decomposition (HRN) and scalable wall functions (HRN) are used. $e$ is the relative error of prediction with respect to LRN.

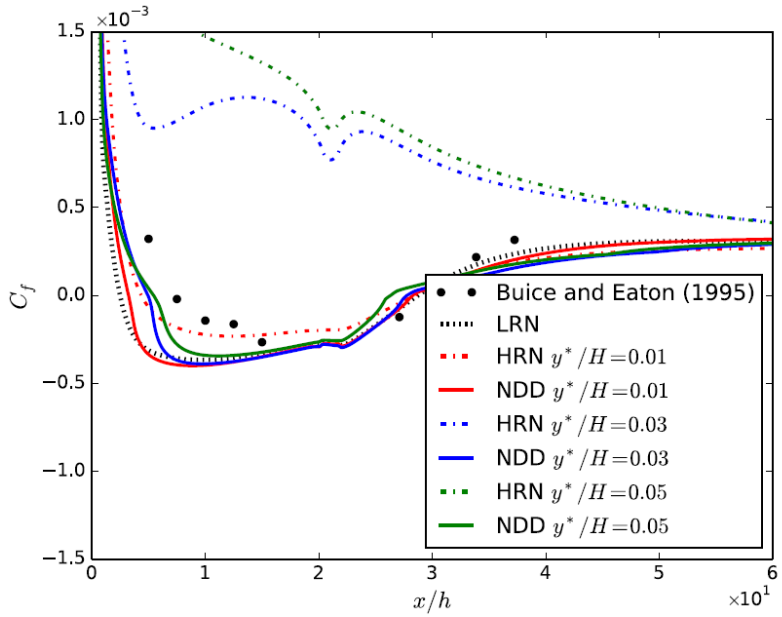



Figure 3: The skin friction distribution along the inclined wall of the diffuser for different $y^*$ [1].

## References


[1] A. Jones, S.V. Utyuzhnikov, Application of a near-wall domain decomposition method to turbulent flows with heat transfer. *Computers & Fluids*, **119**, 87–100, 2015.

[2] A. Jones, S.V. Utyuzhnikov, A near-wall domain decomposition approach in application to turbulent flow in a diffuser. *Applied Mathematical Modelling*, **40**, (1): 329–342. 2016.

[3] S.V. Utyuzhnikov, Towards development of unsteady near-wall interface boundary conditions for turbulence modelling. *Comp. Phys. Communic.*, **185**, 11, 2879–2884, 2014.

[4] Aleksin, V.A. Aleksin, S.V. Utyuzhnikov, Implementation of near-wall boundary conditions for modelling boundary layers with free-stream turbulence. *Applied Mathematical Modeling*, **38**, 14, 3591–3606, 2014.

[5] S.V. Utyuzhnikov, Interface boundary conditions in near-wall turbulence modeling. *Computers & Fluids*, **68**, 186–191, 2012.

[6] S.V. Utyuzhnikov, Nonlinear Problem of Active Sound Control. *J. of Computational and Applied Mathematics*, textbf234 (1): 215–223, 2010.

[7] S.V. Utyuzhnikov, Domain decomposition for near-wall turbulent flows. *Computers & Fluids*, **38**, (9), 1710–1717, 2009.

[8] S.V. Utyuzhnikov, Generalized Calderon-Ryabenkiis potentials. *IMA J. of Applied Mathematics*, **74**, (1): 128–148, 2009.

[9] S.V. Utyuzhnikov, The method of boundary condition transfer in application to modeling near-wall turbulent flows. *Computers & Fluids*, **35**, (10), 1193–1204, 2006.

[10] S.V. Utyuzhnikov, Robin-type wall functions and their numerical implementation. *J. Applied and Numerical Mathematics*, **58**, (10), 1521–1533, 2008.

[11] P.A. Durbin, Limiters and wall treatments in applied turbulence modeling. *Fluid Dynamics Research*, **41**, 1–8, 2009.

[12] B.E. Launder, B.I. Spalding, The numerical computation of turbulent flows. *Computer Methods in Applied Mechanics and Engineering*, **3**, 269–289, 1974.

[13] H. Grotjans, F.R. Menter, Wall functions for industrial applications. K.D. Papailiou eds. *Proceedings of Computational Fluid Dynamics'98, ECCOMAS, 1(2)*, Chichester, UK: John Wiley & Sons, 1112–1117, 1998.

[14] G. Kalitzin, G. Medic, G. Iaccarino, P.A. Durbin, Near-wall behavior of RANS turbulence models and implications for wall functions. *J. Computational Physics*, **204**, 1, 269–289, 2005.

[15] T.J. Craft, A.V. Gerasimov, H. Iacovides, B.E. Launder, Progress in the generalization of wall-functions treatments. *Heat and Fluid Flow*, **23**, 148–160, 2002.

[16] T.J. Craft, S.E. Gant, H. Iacovides, B.E. Launder, A new wall function strategy for complex turbulent flows. *Numerical Heat Transfer*, **45**, 301–318, 2004.

[17] T. Kopp, On grid-independence of RANS predictions for aerodynamic flows using model-consistent universal wall-functions. *Proceedings of ECCOMAS CFD'2006*, The Netherlands, 2006.

[18] S.V. Patankar, D.B. Spalding, *Heat and mass transfer in boundary layers*. Morgan-Grampian Press, London, 1967.

[19] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*. Clarendon Press, Oxford, 1999.

[20] F.R. Menter, M. Kuntz, R. Langry, Ten years of industrial experience with the SST turbulence model. *Turbulence, Heat and Transfer*, 4, 625–632, 2003.

[21] C.U. Buice, J.K. Eaton, Experimental investigation of flow through an asymmetric plane diffuser. Centre of Turbulence Research, Annual Research Brief 1995, 117–120, 1995.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# LARGE EDDY SIMULATION OF SALTATION OVER GAUSSIAN HILLS

**G. Huang , C. Le Ribault, S. Simoëns, I. Vinkovic, J.M. Vignon**

LMFA, ECL, CNRS UMR 5509, UCB Lyon 1
36, avenue Guy de collongue, F 69134 Ecully Cedex, FRANCE
e-mail : catherine.le-ribault@ec-lyon.fr

**Keywords:** Large eddy simulation, Lagrangian models, saltation, hills

**Abstract.** *Soil erosion plays an important role for climate, health and environmental issues. In this study, we focus on the transport of solid particles over a dry rough bed and their interaction with Gaussian hills. For this aim, we use large-eddy simulations (LES) performed with the atmospheric code ARPS. To follow the trajectories of the solid particles, a Lagrangian model is coupled with the LES, according to Vinkovic [1]. Different empirical models have been developed to take into account the interactions of particles with the rough bed, especially the takeoff, the rebound and the deposition.*

*A rough boundary layer at different external velocities is first simulated. The solid particle mass flux as function of the friction velocity is plotted and compared to classical analytical and empirical results in the Shields diagram.*

*Then, LES velocity and concentration profiles are compared to the experiments on solid particle dispersion downstream of Gaussian hills, by Simoëns et al. [2]. Special attention is given to the trapping of solid particles inside the recirculation bubbles. Regions of high particle deposition or erosion are identified. The influence of the distance between two consecutive hills is evaluated.*

## 1 INTRODUCTION

Desertification and soil degradation are major issues for semi arid regions. Erosion is a very complex process which involves large spatial scales and their interaction with soils. Numerical simulations can play a very important role for detailed studies of erodible zones.

Obstacles and hills greatly modify flow characteristics of the incoming boundary layer and in some occasions create large recirculation zones where particles can be trapped and deposited. Bi-dimensional hills are widely used as idealized topographical elements in studies of flows over wall-mounted obstructions. Flows around various kinds of 2D hills have been studied in the past using analytical, numerical, laboratory and in-situ investigations ([3], [4], [5], [6]). Using wind-tunnel measurements, Cao and Tamura [4] highlighted the importance of the lee-side flow separation behind a steep hill.

Our aim is to evaluate the effects of recirculation zones generated by hills on the solid particle mass flux. To this purpose, an experimental campaign was conducted in the frame of the NFS/ANR Sino-French program PEDO-COTESOF "Particle EMission and Deposition Over Complex Terrain for Soil Fixation (PC09). Particle dispersion over one or two hills was investigated [2]. At our best knowledge, this is the only experimental case that produces both aerodynamic description, recirculation zones and solid particle concentration. Our purpose is to further study the link between the recirculation zone and particles deposition through numerical simulations.

The large eddy simulation approach allows the computation of instantaneous large turbulent structures able to produce ejections and sweeping events responsible for the takeoff and deposition of solid particles. This method, coupled with Lagrangian tracking of solid particles is particularly well suited to study saltation. LES coupled with a Lagrangian model has already been used for the study of sand saltation over a flat surface by [1] or [7].

Since the work of [8], research on saltation has concentrated on finding a description of the physical processes of interactions between the particles and the ground. The threshold velocity at which saltation is initiated was studied through wind-tunnel and in-situ experiments ([8], [9], [10], [11]). The behavior of rebounding and ejected particles was investigated through collision experiments between propelling solid particles and a static bed of similar particles ([12], [13]). Most of these models are presented in Shao [14]. Those studies have come to the parametrization of the ejection [15] or the rebound of particles from the surface [13]. From these models, a numerical method has been developed here to simulate solid particle transport and interaction with a rough bed with obstacles.

In the section 2, the flow governing equations and the numerical method are presented. Then, physical models related to particle motion are described. Computational results in the turbulent boundary layer and over one or several hills are discussed in the last section.

## 2 LARGE EDDY SIMULATION AND NUMERICAL DETAILS

Simulations are performed with The Advanced Regional Prediction System (ARPS version 5.15) code, developed at the university of Oklahoma for atmospheric flows

[16]). The Navier-Stokes equations in the LES approach are filtered in order to separate the small scales from the large scales. The semi-compressible filtered Navier Stokes equations, including momentum, heat (potential temperature), mass (pressure) and the equation of state are solved. Subgrid scale turbulence is modeled through a $1.5$ order turbulence closure scheme that requires the resolution of a subgrid scale kinetic energy equation (Yoshizawa & Horiuti [17]).

A fourth-order finite difference method is used for the spatial integration of the equations. The time discretization is performed using a second-order implicit Crank-Nicholson method. The code is parallelized by a domain decomposition method. The full model equations and the numerical method are detailed in [16].

Since this study focuses on particle take-off and deposition, accurate wall modeling is very important for the quality of the simulations. A wall model, based on the equilibrium flow assumption have been initially developed in ARPS. Boundary conditions on the solid wall are imposed through surface fluxes of momentum. The roughness sublayer above rough surfaces presents complex interaction between wakes developing behind roughness elements. Based on direct numerical simulations and wind -tunnel experiments available in the literature, a generalized wall function for the roughness sublayer over periodical roughness elements has been developed by Huang *et al.* [18]. This law is used to estimate the friction velocity in the surface fluxes of momentum.

A method for generating three-dimensional, time-dependent turbulent inflow data for simulations of complex spatially developing boundary layers [19] is used to for the inlet conditions. The approach is based on extraction/rescaling techniques that produce instantaneous planes of velocity data from a downstream station far from the inlet.

## 3 MOTION OF SOLID PARTICLES

Equations for the motion of solid particles have been presented in [1]. Only the main features are summarized in this article.

### 3.1 Lagrangian solid particles tracking

The acceleration of the particle is determined by the forces acting on the particle. For the range of particles considered in this study, simplified equations including only the aerodynamic drag $F_d$ and the gravity forces $F_g$ are considered:

$$\frac{d\vec{x}_p}{dt} = \vec{v}_p(t) \tag{1}$$

$$\frac{d\vec{v}_p}{dt} = \frac{\vec{v}(\vec{x}_p(t),t) - \vec{v}_p(t)}{\tau_p} f(Re_p) + \vec{g} \tag{2}$$

where $\vec{v}_p$ is the velocity of the particle, $\vec{v}(\vec{x}_p(t),t)$ is the velocity of the fluid at the particle position and $\vec{g}$ is the acceleration of gravity. The particle relaxation time is given by:

$$\tau_p = \frac{\rho_p d_p^2}{18\rho_f \nu} \tag{3}$$

and

$$Re_p = \frac{|\vec{v}_p - \vec{v}|d_p}{\nu} \tag{4}$$

is the particle Reynolds number. Here, $d_p$ is the particle diameter, $\rho_p$ the particle density, $\rho_f$ the fluid density and $\nu$ the fluid viscosity. Effects of nonlinear drag are incorporated through $f(Re_p)$. In this work, an empirical relation is used [20].

$$f(Re_p) = \begin{cases} 1 + 0.15Re_p^{0.687} & \text{if } \mathrm{Re_p} < 1000 \\ 0.0183Re_p & \text{otherwise} \end{cases} \tag{5}$$

The Lagrangian equations and the Navier Stokes equations are solved simultaneously and a second order Runge-Kutta scheme is used for the time integration. The different variables of the Navier-Stokes equations are only available at the discrete mesh nodes. They need to be interpolated on the trajectories of the fluid particles. A tri-linear scheme of quadratic Lagrange polynomials developed by Casulli & Cheng [21], is used for the interpolation.

### 3.2 Two-way coupling

A two-way coupling model is used to take into account the influence of the solid particles on the fluid. Small particles, with much larger density than the surrounding fluid, act as if they were an extra burden to the fluid. The momentum transfer from particles to fluid is modeled by adding a drag force to the fluid momentum equation. This force is modeled as follows:

$$\vec{f} = -\frac{1}{\rho_f V_{grid}} \sum_{p=1}^{N_p} m_p \frac{\vec{v}(\vec{x}_p(t), t) - \vec{v}_p(t)}{\tau_p} f(Re_p) \tag{6}$$

where $V_{grid}$ is the volume of the fluid grid cell, $N_p$ is the number of numerically resolved particles in the grid cell and $m_p$ the mass of a particle.

### 3.3 Take-off model

To evaluate the aerodynamic entrainment rate, a new take-off criterion has been developed based on the instantaneous evaluation of the different forces exerted on the particle. The threshold friction velocity is the minimum friction velocity required for wind erosion to occur.

The driving force for the lift-off of sand particles is the aerodynamic lift $F_l$. According to Mollinger & Nieuwstadt [22], the mean lift force can be expressed as:

$$F_l = 15.5\rho\nu^2 \left( \frac{u_*^2 d_p}{\nu} \right)^{1.87} \tag{7}$$

where $u_*$ is the friction velocity and $d_p$ is the diameter of the particle.

The gravity force $F_g$ and the inter-particle cohesive force $F_{ad}$ act as retarding forces. The gravity force is equal to:

$$F_g = \frac{1}{6}(\rho_p - \rho_f) g d_p^3 \tag{8}$$

According to Zimon [23] the cohesive forces can be modeled as:

$$F_{ad} = C d_p \tag{9}$$

where $C$ is a parameter set to $C = 1.43 \times 10^{-5} Nm^{-1}$.

The role of turbulent structures on the initiation of particle motion has to be accounted for as well. There is experimental evidence of the presence of sweep events at the time of grain entrainment. In our model, the particle lifts off when it is located within a sweep event and when the driving force is greater than the retarding forces.

$$F_l >> P + F_{ad} \tag{10}$$

The friction velocity is estimated locally from the velocity fluctuations $u'$ and $w'$ by:

$$u_* = \sqrt{u'w'} \tag{11}$$

The experimental results of Diplat [10] showed that the duration of energetic near-bed turbulent events is also relevant in predicting grain entrainment. Thus, the product of the force by its duration, or impulse, is a more appropriate parameter for identifying flow conditions leading to particle take-off. Based on these conclusions, our criterion has been adapted to satisfy the following equation at each time step $t$:

$$\int_t^{t+\tau_p} (F_l - F_{adh} - P)dt > m_p v_{p0} \tag{12}$$

where the threshold lift-off velocity $v_{p0}$ is equal to $v_{p0} = \sqrt{2gd_p}$.

### 3.4 Rebound model

Due to gravity, sand particles fall and impact the ground. Some of them remain on the ground, others rebound on the soil and can eject several new grains from the bed. In this study, the splash effects are not accounted for according to the experimental set-up.

The velocity after the rebound is characterized by its rebound angle $\theta_r$ and its norm $v_r$. They are modeled following a formulation proposed by Beladjine *et al.* [13]. The effect of local hill bed slope is also taken into account in the modeling of particle-wall collisions.

## 4 EXPERIMENTAL SET-UP

The flow field around one or two hills with various spacings was studied using particle image velocimetry (PIV) measurements. A Gaussian hill shape was used with a maximal slope on the lee side equal to $31^o$. With a model/field scale of $1 : 10000$, the experimental configuration is consistent with an atmospheric turbulent boundary layer of $150m$ and an equivalent dune height of $20m$. In the case of the experiments over two hills, two different spacings $3H$ and $8H$, $H$ being the hill height, were used. Three regimes were studied with external velocities ranging from $5.6m/s$ to $11.2m/s$. Experiments have been performed on both smooth and rough walls.

Powered by an upward-moving piston at a fixed place, particles were continuously introduced into the flow via a $20cm \times 10cm$ slot upstream of the Gaussian hills. The grains have a mean diameter of $200\mu m$ with variations between $170\mu m$ and $250\mu m$ and a density of $1000kg/m^3$. No splash effects were considered. Particle concentration and velocity profiles were obtained using digital image treatment. For more details on the experimental set-up, readers are referred to Simoëns *al.*[2].

## 5 PARTICLE SALTATION IN A BOUNDARY LAYER

To validate the particles transport models, simulations of turbulent boundary layers at different external velocities ranging from $5.6m/s$ to $20m/s$ are first performed. The mean diameter of the particles and their density are the same as in the experimental set-up.

Rough boundary conditions are applied at the wall. To mimic the experiment, the splash effect is not accounted for in our simulations and particles are lift through the take-off process. The saltation process is highly dependent on the friction velocity. The global behavior of particles in a turbulent flow can be characterized by the Shields number:

$$Sh = \frac{u_*^2}{\sigma_p g d_p} \tag{13}$$

where $\sigma_p = \rho_p/\rho_f - 1$. The Shields number measures the transport capacity of solid particles by the flow.

The saltation flux $Q$ is a crucial parameter for the study of wind erosion The streamwise mass flux is computed as follows:

$$q_x(x,z) = \frac{1}{\Delta x_p L_y \Delta z_p} \Sigma_p m_p u_p \tag{14}$$

where $m_p$ and $u_p$ are the mass and the streamwise velocity of the particles present in the volume $\Delta x_p L_y \Delta z_p$. $L_y$ is the length of the domain in the transversal direction. $\Delta x_p$ and $\Delta z_p$ are taken as $0.1H$ where $H$ is the height of the boundary layer. The saltation mass flux, Q, is the vertical integration of the streamwise flux $q_x$:

$$Q(x) = \int_0^\infty q_x(x,z)dz \tag{15}$$

Various saltation models that relate the saltation flux to the characteristics of the flow and of the solid particles through empirical formulas are available in the literature (Bagnold [8], Cresseyls *et al.* [25], Owen [24], Sorensen [9]). Relationships of Bagnolds and Owen predict cubic dependence of $Q$ on the friction velocity whereas the models of Sorensen (1991) [9] , and Creyssels *et al.* [25] predict a dependence on $u*^2$. On figure 1, simulation results are compared with the experiments and the different empirical formulas. The flux is normalized by the total flux $Q_0 = \rho_p d_p \sqrt{g d_p}$. For low Shields numbers ($Sh < 0.1$) a transitional zone appears starting from the threshold range of particle saltation. For high Shields numbers ($Sh > 0.1$), our results exhibit a similar behavior to the empirical formula of Sorensen [9]. The absence of the splash effect in our simulation could explain the shift between our results and the other saltation models.

## 6 PARTICLE SALTATION BEHIND GAUSSIAN HILLS

Simulation results for the solid particles transport over one or two hills are now presented. Results are compared with the experimental results obtained with the external velocity of $U_\infty = 11.2m/s$ on a rough wall. The computation domain includes an upstream Gaussian hill, $80H$ after the inlet and two Gaussian hills separated by a distance of $3H$ or $8H$. The center of the valley between the double hills

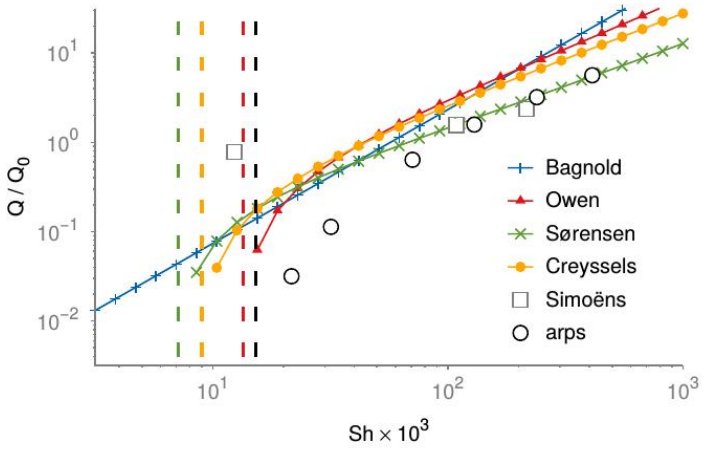



Figure 1: Saltation mass flux as a function of the Shields parameter

is located at $71H$ after the first hill. Table 1 summarizes the domain size and the mesh resolution for all the simulations.

| $N_x \times N_y \times N_z$ | $L_x/H$ | $L_y/H$ | $L_z/H$ | $\Delta x/H$ | $\Delta y/H$ | $\Delta z_{min}/H$ |
|---|---|---|---|---|---|---|
| $1763 \times 63 \times 100$ | 176 | 6 | 15 | 0.1 | 0.1 | 0.05 |

Table 1 Computational parameters

### 6.1 Shear stress and recirculation zones

The wall shear stress, which is of primary importance for the entrainment of solid particles, is presented in figure 2. A negative wall shear stress indicates the presence of a recirculation bubble. The friction velocity in regions where flow adheres to the wall is marked in red and backflow regions inside the recirculation zones are marked in blue. For the single hill configuration, a recirculation zone of approximatively $6H$ appears on the lee side of the hill. This is in agreement with the experimental results of Simoëns *et al.* [2]. The flow evolution before the first hill is similar in the three cases.

For the double hill configuration, two recirculation zones, respectively between and behind the two Gaussian hills appear. As the hill spacing increases from $3H$ to $8H$, we can see an expansion of the inter-hill vortex and a deflation of the lee-side recirculation bubble. The size of the recirculation zone between the two hills is limited by the valley size. For the $8H$ case, the second recirculation zone is smaller than for the $3H$ case which is itself smaller than the recirculation zone behind the single hill. The double hill case with $3H$ spacing behaves as much as a whole to the incoming flow and the effects of the double-hill configuration are negligible starting from $X/H = 6$.

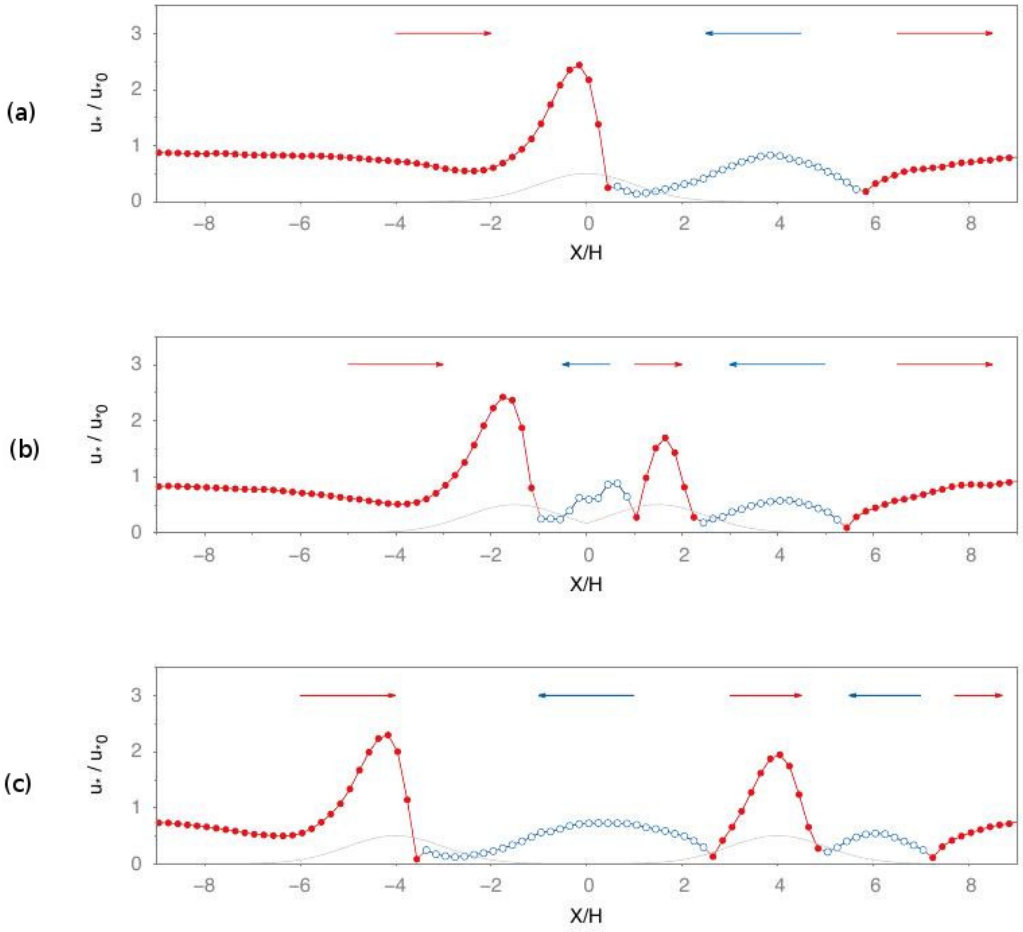



Figure 2: Shear velocity $u_*$ along the hill surface, (a) single hill, (b) 3H, (c) 8H

## 6.2 Particles concentration

Results for the particle concentration are presented for the single hill and for the two hills. For all the plots, the solid lines represent LES results whereas symbols represent experimental results. Profiles are normalized by the maximal concentration at each section. For the single hill, the coordinate origin is fixed at the center of the Gaussian hill and for the two hills, at the middle distance between the two hills. The concentration profiles describe how the particles react to the change of terrain from a flat surface to a steep, transverse hill.

The figure 3 presents the concentration profiles around one hill at different sections. The first profile at $x/H = -25$ corresponds to the concentration profile near the downstream edge of the sandbox. The other profiles present the concentration evolution around the hill. For all the profiles, the agreement between simulation and experimental results for the locations of the concentration peak is reasonable. The simulation seems to overestimate the saltation height growth behind the Gaussian hill. In the simulation results, a higher percentage of solid particles is moved into the upper part of the boundary layer behind the hill at $x/H = 1.5$ and $x/H = 3$. This upward migration is less pronounced in the experiments. Besides, in those zones, the simulation apparently underestimates the quantity of solid particles near

the ground inside the recirculation zone compared to the experiments. This result could be explained by the fact that solid particles, impacting the windward side of the hill, are abruptly redirected upwards due to high hill slopes. As an inertial effect, these rebounding grains maintain this upward motion and reach high layers of rapid flow motion, which immediately evacuate them downstream of the hill.

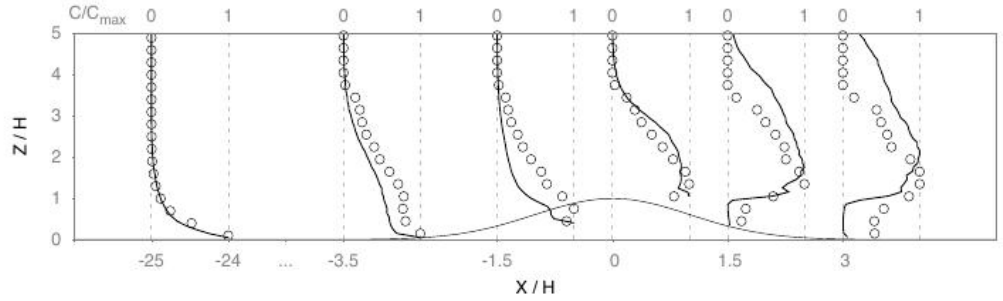



Figure 3: Concentration profiles at the end of the sandbox and around a single hill. Symbols - experiments, lines - LES.

Figure 4 presents the same plots for the case of two consecutive hills separated by $3H$. A better collapse with the experimental profiles is achieved around the second hill. Compared to experimental results, a smaller fraction of particles is present inside the wall vortex between the crests. The simulation overestimates the quantity of the incoming particles over the top of the first hill.

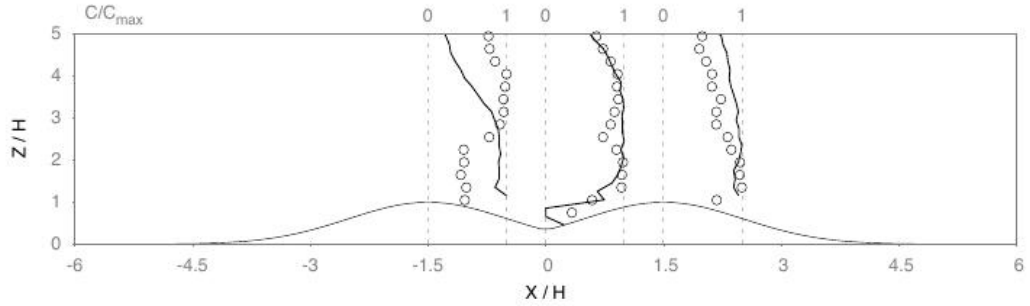



Figure 4: Concentration profiles for the $3H$ case. Symbols - experiments, lines - LES.

In figure 5, the concentrations profiles around the two hills with $8H$ distance are plotted. Similar to figure 4, the simulation apparently underestimates the quantity of solid particles passing between the two hills compared to the experimental results. We note that the global evolution of concentration peaks is similar between the $3H$ and the $8H$ cases. This implies a weak influence of the hill spacing on the overall evolution of mean particle concentration.

### 6.3 Near wall particle transport and deposition

Our goal is to explore the link between particle trapping and the flow characteristics inside the recirculation zones. To this purpose, regions of preferential deposition

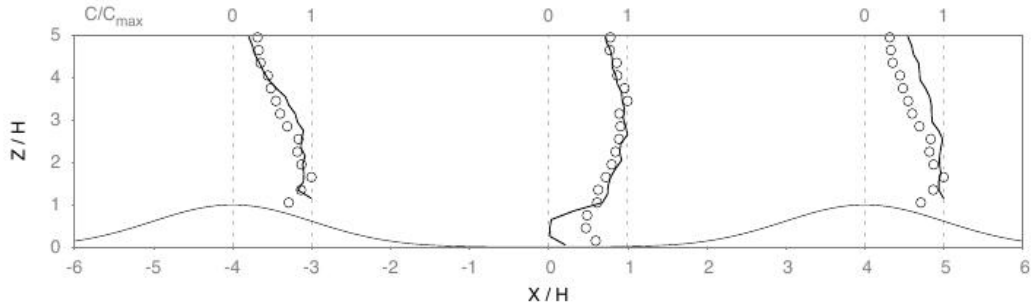



Figure 5: Concentration profiles for the $8H$ case. Symbols - experiments, lines - LES.

in the recirculation zones between and after the hills are identified. The net deposition of particles is defined as the accumulation $\lambda_D$ scaled by the packing density at the sandbox $\lambda_{box}$. The net deposition is computed during a period of $\Delta T = 30\delta_0/u_*$, $\delta_0$ being the depth of the incoming boundary layer. Regions with $\lambda_D < 0$ indicate net erosion, whereas $\lambda_D > 0$ indicate net deposition. An instantaneous map of the net accumulation is shown in Figure 6 for the three different hill spacings. Areas of erosion are marked in red and deposition zones are marked in yellow.

The sporadic areas of net erosion are weak and more likely to occur on the windward face of the hill. The net deposition of particles is more important and most of the time located in the recirculation zones behind and between hill crests. The accumulations due to the deposited grains are in general one, even two orders of magnitude sparser than that at the sandbox. Not surprisingly, the regions of high deposition are located at the junction of positive and reverse transport, where the friction velocity and the mean lift force are weak. This could be explained by the fact that instantaneously frequent alterations take place in the direction of particle transport in these areas. For the particles already trapped by the recirculation zones, this back and forth motion may lead to a final deposition. Due to a weak lift force, particles spend a longer time immobilized on the wall and therefore contribute to the net accumulation in these regions.

## 7 CONCLUSIONS

The paper presents a numerical study of the physical problem of soil erosion by wind. The particle transport inside a turbulent boundary layer in the presence of obstacles is studied to improve our knowledge on wind erosion and particle deposition over complex terrains.

The simulation results are compared to the PC09 experiments of Simoëns *et al.* [2]. The flow between and behind Gaussian hills is dominated by large recirculation zones. The predicted particle concentration is in overall agreement with the experiments. A mapping of particle accumulation on the wall reveals preferential particle deposition at the junction between regions with mean and adverse particle transport.

**Acknowledgments**

We acknowledge NFSC/ANR Chinese/French program PEDO-COTESOF. This work was granted access to the HPC resources of CINES. Numerical simulations

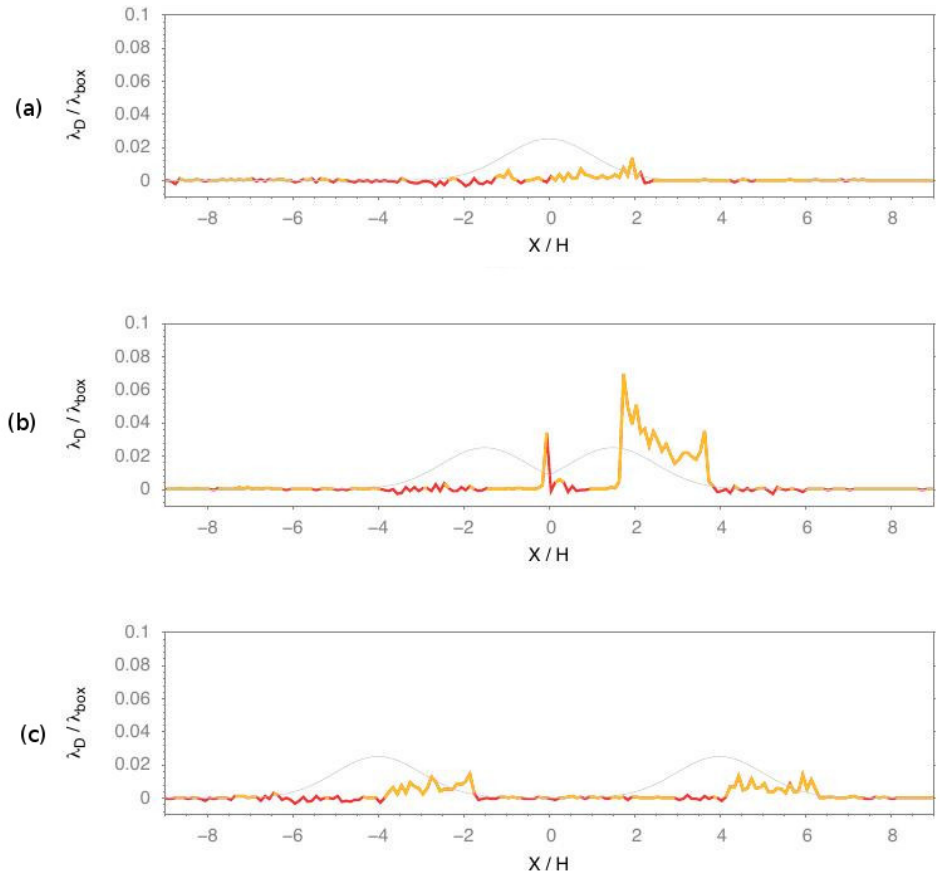



Figure 6: Instantaneous map of the net accumulation of solid particles. Red - erosion, yellow - deposition., (a) single hill, (b) 3H, (c) 8H

were also performed on the P2CHPD parallel cluster.

## REFERENCES

[1] Vinkovic I., Aguirre C., Simoëns S., Ayrault M. Large eddy simulation of the dispersion of solid particles in a turbulent boundary layer. *Boundary Layer Meteoroly* **121,** 283-311, 2006

[2] Simoëns S. , Saleh S., Le Ribault C. , Belmadi M., Zegadi R. , Allag F., Vignon J.M., Huang G., "Influence of Gaussian hill on concentration of solid particles in suspension inside Turbulent Boundary Layer", *Procedia IUTAM*, **17** , pp. 110-118, 2015

[3] Jackson P.S., Hunt J.C.R. Turbulent wind flow over a low hill, *Quarterly Journal of the Royal Meteorological Society*, **101** , 929-955, 1975

[4] Cao S., Tamura T. Effects of roughness blocks on atmospheric boundary layer flow over a two-dimensional low hill with/without sudden roughness change, *Journal of wind engineering and industrial aerodynamics*, **95**, 679-695, 2007

[5] Chapman, C., Walker, I.J., Hesp, P.A., Bauer, B.O., Davidson-Arnott, R.G.D. Oller- head, J. (2013). Reynolds stress and sand transport over a foredune. *Earth Surface Processes and Landforms*, **38**, 17351747, 2013

[6] Almeida G.P., Durao D.F.G., Heitor M.V. Wake flows behind two-dimensional model hills, *Experimental Thermal and fluid Science*, **7**, 87-101, 1993

[7] Dupont S., Bergametti G., Marticorena B., Simoëns S. Modeling saltation intermittency, *Journal of Geophysical Research: Atmospheres*, **118**, 7109-7128, 2013

[8] Bagnold R.A. *The physics of wind blown sand and desert dunes*, Methuen, London, 1941

[9] Sorensen M. An analytic model of wind-blown sand transport *Acta Mechanica Supplementum*, 67-81, Springer Vienna, Vienna, 1991

[10] Diplat P., Dancey C.L. Coherent flow structures, initiation of motion, sediment transport and morphological feedbacks in rivers. J.G. Venditti, J.L. Best, M. Church & R.J. Hardy eds., *Coherent Structures at Earth's surface*, 289-307, John Wiley & Sons, Ltd, Chichester,UK, 2013

[11] Foucaut J.M., Stanislas M. Take-off threshold velocity of solid particles lying under a turbulent boundary layer, *Experiments in Fluids*, **20**, 1996

[12] Anderson R.S., Haff P.K, *Wind modification and bed response during saltation of sand in air*, Aeolian Grain Transport 1, Springer Vienna, Vienna, pp. 21-51, 1991

[13] Beladjine D., Ammi M., Oger L., Valance A. Collision process between an incident bed and a three-dimensional granular packing, *Physical Review E.*, **75**, 061305, 2007

[14] Shao Y. *Physics and modeling of wind erosion*, **37**, Atmospheric and Oceanographic Sciences Library, Springer Netherlands, Dordrecht, 2009

[15] Descamps I., Erosion éolienne d'un lit de particules à large spectre granulométrique, Thèse de doctorat, Université de Valenciennces et du Hainaut-Cambrésis, France, 2004

[16] Xue M., Droegemeier K., Wong V., Shapiro A., Brewster K. ARPS version 4.0 Users Guide, 1995

[17] Yoshizawa A., Horiuti K. A statistically-derived subgrid-scale kinetic models for large eddy simulation of turbulent, *J. Phys. Soc. Japan* **54**(8), 2834-2839, 1985

[18] Huang G., Simoëns S., Vinkovic I., Le Ribault C., Dupont S., Bergametti G. Law-of-the-wall in a boundary-layer over regularly distributed roughness elements, *Journal of Turbulence*, 2016

[19] Lund T., Wu X., squires K Generation of turbulent inflow data for spatially developing boundary layer simulations, *Journal of Computational Physics*, **140**, 233-258, 1998

[20] Clift R., Grace J.R., Weber M.E., *Bubbles, drops and particles*, Academic Press, New York 1978

[21] Casulli V., Cheng R.T. Semi-implicit finite difference methods for the three-dimensional shallow water flow, *International Journal for Numerical in Fluids*, **15**, pp. 629-648, 1992

[22] Mollinger A.M., Nieuwstadt F.T.M. Measurement of the lift force on a particle fixed to the wall in the viscous sublayer of a fully developed turbulent boundary layer,*Journal of Fluid Mechanics*, **316**, 285-306, 1996

[23] Zimon A.D. *Adhesion of dust and powder*, Springer, 1969

[24] Owen P.R. Saltation of uniform grains in air, *Journal of fluid Mechanics*, **20**, 225-242, 1964

[25] Creyssels M., Dupont P., Ould El Moctar, Valance A., cantat A., Jenkins J.T., Pasini J.M., Rasmussen K.R., Saltating particles in a turbulent boundary layer: experiment and theory, *Journal of Fluid Mechanics*, **625**, 47-28, 2009

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EXPERIMENTAL AND NUMERICAL STUDY ON THE PERFORMANCE OF AN INNOVATIVE VERTICAL-AXIS WIND TURBINE

**Katarzyna Kludzinska[1], Krzysztof Tesch[1], and Piotr Doerffer[2]**

[1]Fluid Mechanics Department
Gdansk University of Technology
ul. G. Narutowicza 11/12, 80-233 Gdansk, Poland
e-mail: kkludzinska@o2.pl

[2] Institute of Fluid Flow Machinery
ul. Fiszera 14, 80-231 Gdansk, Poland
e-mail: doerffer@imp.gda.pl

**Keywords:** Wind Turbines, Savonius wind turbine, CFD.

**Abstract.** *This paper introduces the innovative modification of the Savonius wind turbine being able to significantly increase efficiency in comparison with the classic design. This innovative design is equipped with a stator directing the flow. The presence of the stator increases the active surface area and generates higher torques acting on a shaft. Additionally, it makes it possible to take better advantage of wind energy and compensate the effect of larger active surface area. The larger the stator angle the higher the efficiency. One of the biggest problem of the classic design, namely its relatively low efficiency, is overcome.*

*The results of an experimental investigation, carried out in the closed return wind tunnel, are presented for various conditions and configurations. These include the most important characteristic such as the torque and power coefficient as a function of the tip speed ratio. Furthermore, a three dimensional numerical analysis of the transient phenomena occurring in the innovative turbine is investigated. Evaluation and comparison criteria are proposed in order to evaluate and compare various designs and solutions. This makes it possible to explain the increased efficiency of the innovative design.*

## 1 INTRODUCTION

The innovative vertical-axis wind turbine [2], shown in figure 1, is a modification of the Savonius classic design. It is equipped with an adjustable stator directing the flow and increasing the active surface area. This makes it possible to generate higher torques acting on a shaft and to take better advantage of wind energy and compensate the effect of larger active surface area. Thus one of the biggest problem of the classic design, namely its relatively low efficiency, is overcome.

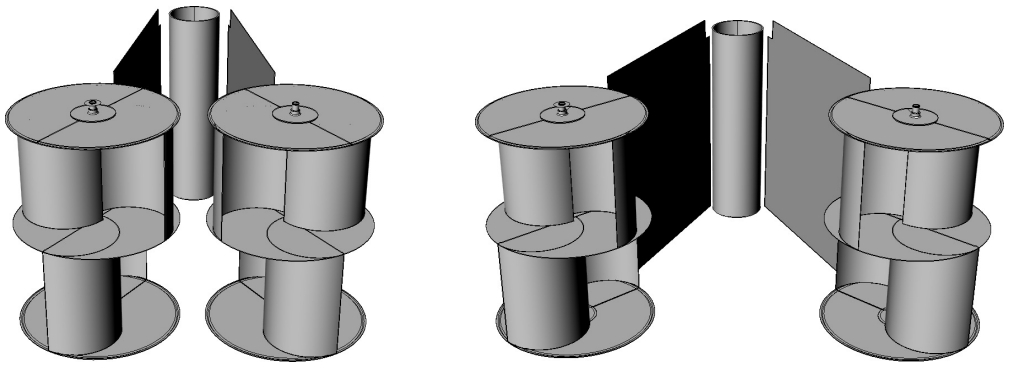


Figure 1: Innovative design.

The classic Savonius rotor has the simplest design of all devices converting wind into other energy forms. Despite numerous advantages, such as low noise, simplicity of design, applicability for a wide range of wind velocities, one of drawbacks of the classic design is its relatively low efficiency. It appears that the proposed modifications of Savonius like rotors in available literature have been developed using a trial and error method. This is because the flow structure inside the rotor is complicated and that is why in past investigations were limited only to laboratory tests [1, 6, 11, 10]. Recently attempts were made to analyse the structure of the flow through the Savonius rotor numerically [7, 9]. Most of data presented on the numerical aspect are mostly two-dimensional.

The main goal of this paper it to present results of an experimental investigation and to compare them with a three dimensional numerical analysis of the transient phenomena occurring in the innovative turbine. Furthermore, evaluation and comparison criteria are proposed in order to evaluate and compare various designs and solutions. An attempt to explain the increased efficiency of the innovative design has been also undertaken.

## 2 EXPERIMENTAL RESULTS

### 2.1 Coefficients

The most important characteristics of the wind turbines include the torque $C_T$ and power $C_P$ coefficients. Both of them are expressed as a function of the rotor tip speed ratio $\lambda$

$$\lambda = \frac{\omega D}{2\,U}. \tag{1}$$

The tip speed ration represents the ratio between the tangential speed of the tip of a blade and the velocity of the wind $U$. In the above definition $\omega$ stands for angular velocity and $D$ is the diameter of the rotor.

The former of the mentioned non-dimensional quantities, i.e. $C_T$, is defined as

$$C_T = \frac{\bar{T}}{\frac{1}{4}\rho U^2 D^2 H} \tag{2}$$

where $H$ is the rotor's height and $\rho$ represents the density. For transient flows the average value of time dependent torque $T(t)$ is considered. This means that we deal with the distribution of $C_T$ as a function of the angular position of the rotor $\alpha$. The average value $\bar{T}$ is given by

$$\bar{T} = \frac{1}{\Delta t} \int_{t}^{t+\Delta t} T(t)\mathrm{d}t \tag{3}$$

where $\Delta t$ stands for the time of one revolution. If the time step of the transient CFD calculations is constant we can approximate the integral average with the arithmetic average. The formal definition of the former non-dimensional quantity, i.e. power coefficient, is

$$C_P = \frac{\omega \bar{T}}{\frac{1}{2}\rho U^3 H D}. \tag{4}$$

Again, this is valid for transient flows, which is typical for the Savonius rotor operation. The non-dimensional quantities distributions as a function of $\lambda$ are the basis for comparison the innovative design with the traditional solution or other studies.

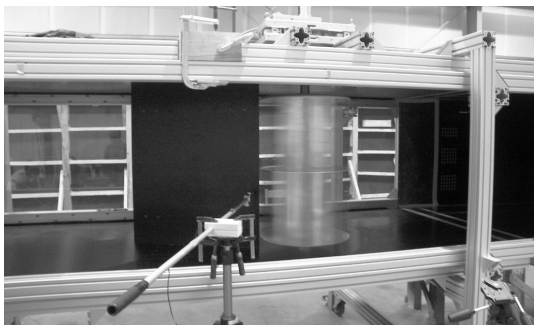


Figure 2: Experimental setup.

## 2.2 Results

The experimental setup is shown in figure 2. This figure presents the test section of the closed return wind tunnel with one side of the section open. The two-stage Savonius rotor equipped with an adjustable stator is visible. A pitot tube located at the inlet to the test section is utilised in order to determine the wind speed. Furthermore, the rotational speed of the rotor and the torque acting on the shaft are also measured.

The wind tunnel can attain maximal velocity of about $50\,\mathrm{m\,s^{-1}}$. Experimental results for $U \approx 10\,\mathrm{m\,s^{-1}}$ are shown in figures 3 and 4. Three different configurations are discussed here, namely the classic Savonius design and the innovative design equipped with a stator at $40^\circ$ and $60^\circ$ angle of attack. Figure 3 presents the power $C_P$ and figure 4 torque coefficient $C_T$ distributions as a function of tip speed ratios $\lambda$. In order to indicate the uncertainty in measurements absolute errors are superimposed.

Figure 4 shows that the torque coefficient distributions of the modified design are lower in comparison with the original Savonius wind turbine even that the higher torques acting on a shaft. This may be explained by the larger active surface area $HD$ for the modified design in equation (2). Furthermore, the effect of larger active surface area $HD$ can is compensated by the same stator directing the flow and generating higher torques. Figure 3 shows the distribution of $C_P$ as a function of $\lambda$. The larger the stator angle of atack the higher the efficiency. Obviously, the innovative design with the stator at $40^\circ$ and $60^\circ$ is always more efficient for the same wind speed $\sim 10\,\mathrm{m\,s^{-1}}$. This shows that even for low wind speeds the innovative design if more efficient and may be regarded as an advantage of the proposed design over the Savonius wind turbine.

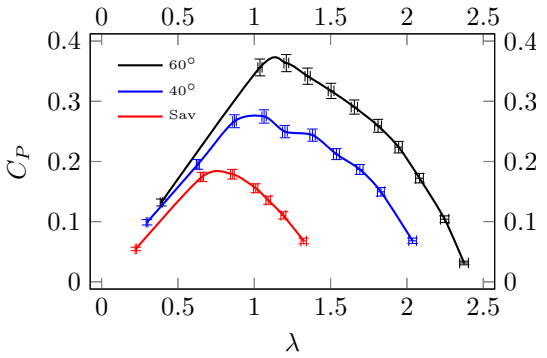



Figure 3: $C_P$ distribution as a function of $\lambda$.

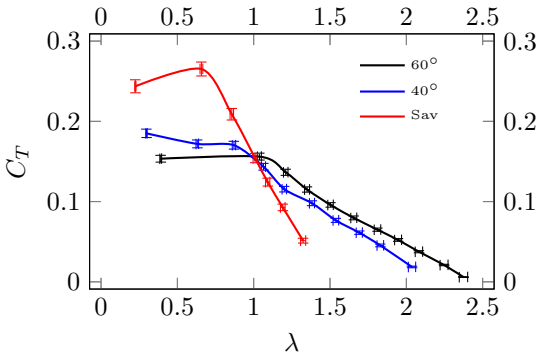



Figure 4: $C_T$ distribution as a function of $\lambda$.

## 3 NUMERICAL CALCULATIONS

### 3.1 Governing equations

The numerical calculations have been performed by means of CFD. The turbulent flow of air is regarded as an incompressible medium. RANS approach has been utilised in order to model the turbulence, namely the two-equation Shear Stress Transport turbulence model [8]. The average form of mass conservation equation has the form

$$\nabla \cdot \bar{\mathbf{U}} = 0. \tag{5}$$

The Reynolds equation is

$$\rho \frac{\mathrm{d}\bar{\mathbf{U}}}{\mathrm{d}t} = \rho \vec{g} - \nabla p_e + \nabla \cdot \left(2\mu_t \bar{\mathbf{D}}\right), \tag{6}$$

where the effective pressure $p_e := \langle p \rangle + 2/3\rho k$ and the effective viscosity consists of the eddy and molecular components $\mu_e := \mu_t + \mu$. Two additional transport equations are those for modelled kinetic energy of velocity fluctuation which arises from Reynolds stress transport equation

$$\rho \frac{\mathrm{d}k}{\mathrm{d}t} = 2\mu_t \bar{\mathbf{D}}^2 + \nabla \cdot \left( \left( \frac{\mu_t}{\sigma_{k3}} + \mu \right) \nabla k \right) - C_\mu \rho k \omega \tag{7}$$

and the turbulent frequency $\omega$. This is analogous to $k$ transport

$$\rho \frac{\mathrm{d}\omega}{\mathrm{d}t} = \alpha_3 \frac{\omega}{k} 2\mu_t \bar{\mathbf{D}}^2 + \nabla \cdot \left( \left( \frac{\mu_t}{\sigma_{\omega 3}} + \mu \right) \nabla \omega \right) - \beta_3 \rho \omega^2 + (1 - F_1) \frac{2}{\omega} \rho \sigma_{\omega 3} \nabla k \cdot \nabla \omega. \tag{8}$$

The eddy viscosity is defined as $\mu_t = \rho k \omega^{-1}$. Constants marked with the subscript 3, namely $\sigma_{k3}$, $\sigma_{\omega 3}$, $\alpha_3$, $\beta_3$ are linear combinations of constants from the component models.

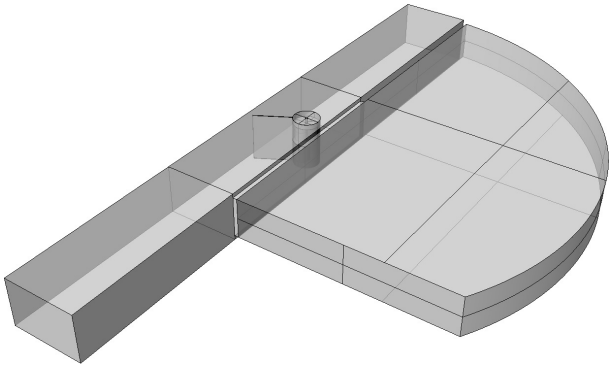


Figure 5: Flow domain.

### 3.2 Boundary conditions and mesh

The flow domain resembles a wind tunnel with one side partially open to the atmosphere as shown if figure 5. It is divided into parts merged by means of interfaces. These include the rotating rotor and the steady wind tunnel and surrounding area. The boundary conditions selected here include tunnel inlet with a prescribed velocity $10.5\,\mathrm{m\,s^{-1}}$ and medium turbulent intensity conditions. Furthermore, the so called 'far field' condition is chosen with prescribed constant atmospheric pressure. The surrounding area is limited by symmetry planes and wind tunnel walls are modelled as no slip wall. The same concerns the rotor. The only difference being the rotating frame of reference.

The domains have an unstructured grid consisting of mostly tetrahedral elements. The total number of elements covering the flow area is about $9$ million. In order to properly resolve flow near the wall special elements around the blades are generated. The wall function approach has been used to provide near wall boundary conditions for the mean flow. The average value of $y^+$ distribution does not exceed $2$ for all the time steps. Finally, the time step of the transient calculations corresponds to $4^\circ$ of revolution and the angular velocity of the rotor corresponds to $4$ revolutions per second.

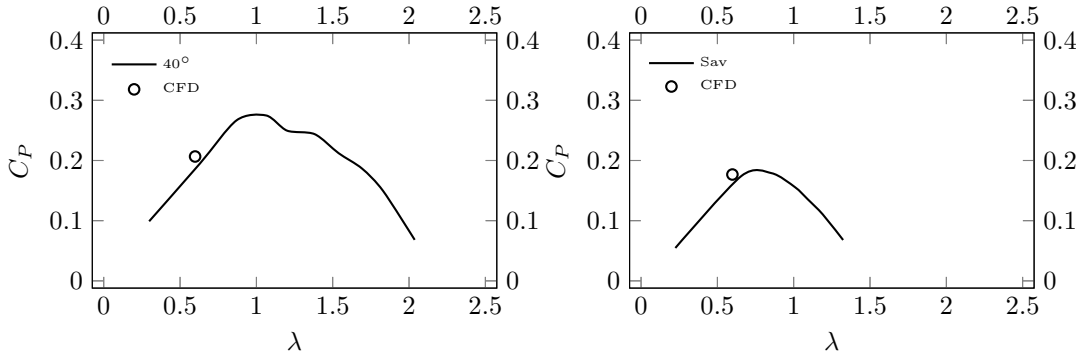



Figure 6: Experiment vs numerical prediction.

### 3.3 Results of calculations

Figures 6 show the experimental values of $C_P$ versus CFD calculations for two selected designs as a function of $\lambda$. These include the classic Savonius rotor and the innovative design equipped with a stator at $40^\circ$ angle of attack. Circles correspond to the numerical prediction. Good agreement is visible.

Figures 7 present the torque coefficient $C_T$ distribution as a function of revolution angle $\alpha$ for three different configurations, namely Savonius rotor and an innovative design equipped with a stator at $40^\circ$ and $60^\circ$ angle of attack. These are results of transient calculations. The four peaks per curve are visible due to the two rotors rotated by an angle relative to each other. The torque coefficient distributions of the modified design are lower in comparison with the original Savonius wind turbine. This is because the presence of the stator increases the active surface area $HD$.

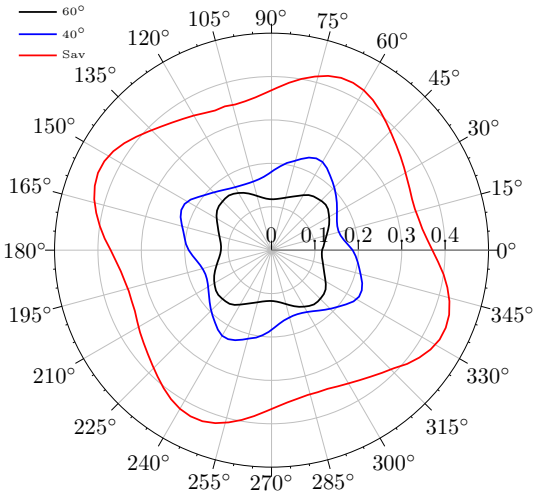



Figure 7: $C_T$ distribution as a function of the angular position of the rotor $\alpha$.

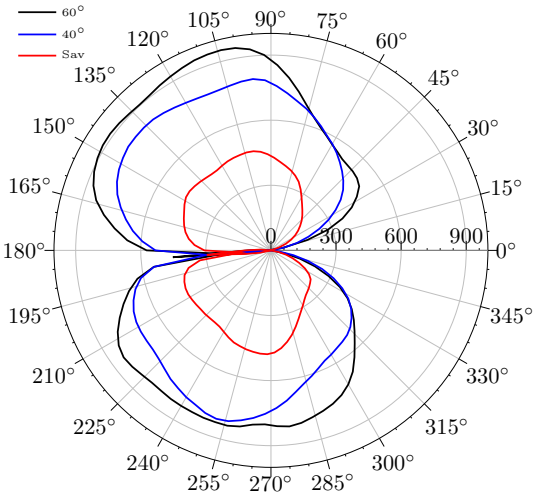



Figure 8: $\Lambda_2$ distribution as a function of the angular position of the rotor $\alpha$.

## 4 COMPARISON CRITERIA

The proposed criteria allow for evaluation and comparison of various designs and solutions. Invariants of velocity gradient tensors $\nabla\mathbf{U}$, or its decomposition

$$\nabla\mathbf{U} = \mathbf{D} + \mathbf{A} \tag{9}$$

are often used [3] in turbulence modelling. This is because they contain all the necessary information that are responsible for kinetic energy dissipation and vortex stretching [13]. In the above equation $\mathbf{D}$ is the rate-of-strain tensor and $\mathbf{A}$ represents the spin (vorticity) tensor.

The vorticity measure $\mathcal{Q}$, firstly investigated in [4], is defined by means of the integral of the velocity gradient tensor second invariant in the following form

$$\mathcal{Q} = \iiint_V \mathrm{tr}\,(\nabla\mathbf{U})^2\,\mathrm{d}V = \iiint_V \left(\mathrm{tr}\mathbf{D}^2 - \|\mathbf{A}\|^2\right)\mathrm{d}V. \tag{10}$$

The above criterion represents a global balance between the power of energy dissipation $N_d$ and

vortex structures by means of the vorticity magnitude or enstrophy $\mathcal{E}^*$

$$\mathcal{Q} = \frac{N_d}{2\mu} - \mathcal{E}^*. \tag{11}$$

These structures are generated by the rotor and may affect the efficiency [12]. Another criterion $\lambda_2$ [5] is investigated in this paper. The vorticity measure $\Lambda_2$ is defined by means of the integral of the $\lambda_2$ in the following form

$$\Lambda_2 = \iiint_V \lambda_2 \left(\mathbf{D}^2 + \mathbf{A}^2\right) \mathrm{d}V \tag{12}$$

where $\lambda_2(\mathbf{D}^2 + \mathbf{A}^2)$ represents the second ordered eigenvalue of tensor $\mathbf{D}^2 + \mathbf{A}^2$. The three eigenvalues are ordered as follows $\lambda_1 \geq \lambda_2 \geq \lambda_3$.

Figure 8 shows the distribution of the vorticity measure $\Lambda_2$ defined by means of equation (12) which can be used to compare various designs. The higher the efficiency the higher the values of $\Lambda_2$. Equation (12) is appropriate for individual time steps, while for the total revolution it is necessary to use the time average of $\Lambda_2$ instead

$$\langle \Lambda_2 \rangle = \frac{1}{\Delta t} \int_t^{t+\Delta t} \Lambda_2 \, \mathrm{d}t. \tag{13}$$

The above time integral (13) can be used directly to evaluate specific turbine in terms of a vorticity measure. This is summarised in table 1 where the ratio of $\langle \Lambda_2 \rangle$ to $\langle \Lambda_{2S} \rangle$ of the respective Savonius design is listed. It follows previous observations and makes it possible to anticipate a relationship between vorticity measure and efficiency of a turbine.

| Cinfiguration | $\langle \Lambda_2 \rangle / \langle \Lambda_{2S} \rangle$ |
|---|---|
| Savonius | 1.000 |
| 40° | 2.543 |
| 60° | 2.834 |

Table 1: $\langle \Lambda_2 \rangle / \langle \Lambda_{2S} \rangle$ values for different configurations.

## 5 CONCLUSIONS

Conclusions can be summarised in the following points:

- Usually, all data presented on the numerical aspect in the available literature are mostly two-dimensional and steady state. A three dimensional numerical analysis of the transient phenomena occurring in the innovative turbine is investigated here.
- Wind tunnel experiments prove that the innovative modification of the classic Savonius shows higher efficiency for the same wind speed.
- The increased efficiency is due to presence of the stator which directs the air and makes it possible to take better advantage of its energy.

- A relationship between vorticity measure and efficiency of a turbine can be anticipated. The proposed criteria are related to the vortex structures being generated by the rotor and affecting pressure distribution around it and thus the performance. The time integral of this measures can be used directly to evaluate specific turbine.

## REFERENCES

[1] B.F. Blackwell, R.E. Sheldahl, L.V. Feltz, Wind Tunnel Performance Data for Two- and Three-Bucket Savonius Rotor, *Sandia Laboratories Report SAND 76-0131*, **105**, 1977.

[2] P. Doerffer, *International patent application PCT/PL2012/000125 and national patent application*

[3] G. Haller, An objective definition of a vortex. *Journal of Fluid Mechanics*, **525**, 1–26, 2005.

[4] J. C. R. Hunt, A. Wray, P. Moin, Eddies, stream, and convergence zones in turbulent flows. *Center for Turbulence Research Report CTR-S88*, 1988.

[5] J. Jeong, F. Hussain, On the identification of a vortex. *Journal of Fluid Mechanics*, **285**, 69–94, 1995.

[6] M.A. Kamoji, S.B. Kedare, S.V. Prabhu, Experimental investigations on single stage, two stage and three stage conventional Savonius rotor, *International Journal of Energy Research*, **32**, 877–895, 2008.

[7] J. Menet, Prediction of the aerodynamics of a new type of vertical axis wind turbine: the reverse bladed rotor. In: *Proc. of the European Wind Energy Conference*, 2008.

[8] F. R. Menter, Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications, *AIAA Journal*, **32** 8, 1598–1605, 1994.

[9] M.H. Mohamed, G. Janiga, E. Pap, D. Thevenin, Optimisation of Savonius turbines using an obstacle shielding the returning blade, *Renewable Energy*, **35**, 2618–2626, 2010.

[10] M. Nakajima, S. Iio, T. Ikeda, Performance of Double-Step Savonius Rotor for Environmentally Friendly Hydraulic Turbine, *Journal of Fluid Science and Technology*, **3**, 410–419, 2008.

[11] U.K. Saha, S. Thotla, D. Maity, Optimum design configuration of Savonius rotor through wind tunnel experiments, *Journal of Wind Engineering and Industrial Aerodynamics*, **96**, 1359–1375, 2008.

[12] K. Tesch, K. Kludzinska, P. Doerffer, Investigation of the Aerodynamics of an Innovative Vertical-Axis Wind Turbine. *Flow, Turbulence and Combustion*, **95** 4, 739–754, 2015.

[13] K. Tesch, On invariants of fluid mechanics tensors. *Task Quarterly* **17** 3–4, 1000–1008, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF NONLINEAR FREE SURFACE WATER WAVES USING A COUPLED POTENTIAL FLOW-URANS/VOF APPROACH

**Bülent Düz[1*], Tim Bunnik[1], and Geert Kapsenberg[1]**

[1]Maritime Research Institute (MARIN)
P.O. Box 28, 6700AA, Wageningen
The Netherlands
e-mail: {b.duz,t.bunnik,g.k.kapsenberg}@marin.nl

**Keywords:** Nonlinear free surface water waves, One-way coupling of potential and viscous flow models, Focusing wave events, Waves on sloping bottom, Survey on wave models.

**Abstract.** *This paper first presents a comparison particularly in terms of accuracy between four nonlinear potential flow models for simulation of nonlinear free surface water waves. The performance of the wave models is compared in five test cases in 2D for which measurement data was also available. Three cases involved wave groups propagating over a flat bottom and producing a focusing event, and two cases involved irregular waves propagating over sloping bottom. After the comparison study, one of the potential flow solvers was selected for one-way coupling with a viscous flow solver. Details of this coupling are presented along with some information about the viscous flow solver in question. Here strong suits of both solvers are utilized to simulate nonlinear water waves in an accurate and efficient manner. Even though only simulating nonlinear waves is discussed in the paper, this work was considered as a stepping stone to later focus on hydrodynamic wave loading on structures and wave-structure interaction using this coupling. In the end results showed that such a coupling can indeed be a valuable tool to study wave motion. In all the simulations, comparison with the measurements demonstrated good agreement, and the convergence behavior of the coupled solver was acceptable.*

*Address all correspondence to this author.

## 1 INTRODUCTION

Numerical simulation of nonlinear wave-wave, wave-bottom and wave-structure interaction is a challenging endeavor. Even though a myriad of numerical models have been available to investigate these problems, current levels of computational power require making smart decisions to have a balance between accuracy and efficiency. In typical wave-structure interaction applications, viscous effects are generally important near the structure and in regions of strong wave breaking. Naturally a domain decomposition strategy comes to one's mind, where viscous flow solvers are used only in areas where viscous effects may be important, and potential flow models (or other economical but sufficiently accurate models) are used in the rest of the domain. For fully nonlinear, multi-phase and unsteady CFD computations ReFRESCO [1] has been extensively used at the Maritime Research Institute in the Netherlands (MARIN) (for information visit http://www.refresco.org). In order to find a wave model that can be used in combination with ReFRESCO, we first conducted an inventory study on some of the existing codes. The functionalities and performances of the codes were analyzed in various test cases. Then one of these codes was chosen for one-way coupling with ReFRESCO, and the results of this coupling are presented.

Domain decomposition is not the only application area for the potential flow models. These models can also be used to study statistics of extreme waves, reproduce basin waves in a numerical environment, and calculate kinematics underneath waves which then can be used to calculate forces using Morison type equations.

After a detailed survey on the existing codes, four were selected for consideration. Below some information about these codes is given, and some of their features are listed in Tab. 1.

1. HAWASSI (LabMath Indonesia - University of Twente)
   The acronym HAWASSI stands for Hamiltonian Wave-Ship-Structure Interaction, and the code was developed at LabMath Indonesia and the University of Twente. Programming language is MATLAB. The website http://hawassi.labmath-indonesia.org contains information about the software and its functionalities. HAWASSI is not open source, but a demo version with limited functionalities can be downloaded from the website. The software comes in two versions: HAWASSI-AB (Analytic Boussinesq) and HAWASSI-VBM (Variational Boussinesq Model). In this work only HAWASSI-AB was used, and it is simply referred to as HAWASSI in the remainder of this document. For more information about the model, see [2, 3, 4, 5].

2. SWASH (Delft University of Technology)
   The acronym SWASH stands Simulating WAves till SHore. This open source code has been developed based on the work of [6, 7, 8, 9]. The source code was written in FORTRAN 90 and can be downloaded from http://swash.sourceforge.net, which also contains a detailed description of the software and its functionalities. SWASH can be used freely under the terms of the GNU General Public License, and it can be implemented on Microsoft Windows, Linux, Unix and Mac OS/X, provided a Fortran 90 compiler is available.

3. OceanWave3D (Technical University of Denmark)
   OceanWave3D is a fully nonlinear and dispersive potential flow model. It is distributed under the GNU General Public License. The base code was developed at DTU Mechanics between 2006 and 2008 by Allan P. Engsig-Karup, and was entirely rewritten at DTU

Informatics between 2008 and 2011 (current version) by Allan P. Engsig-Karup with contributions from various researchers. The model is described in [10]. Even though the software exists in several versions (FORTRAN 90 and C++), only the FORTRAN 90 version has been made available to public. The source code can be downloaded from https://github.com/apengsigkarup/OceanWave3D-Fortran90.

4. HOS-NWT (École Centrale de Nantes)
   HOS-NWT is a Numerical Wave Tank based on the High-Order Spectral (HOS) method. It has been developed at École Centrale de Nantes, and is currently used in several research projects. The source code is written in FORTRAN 90 and can be downloaded from https://github.com/LHEEA/HOS-NWT. The software is distributed under the terms of the GNU General Public License. The description of the model can be found in [11, 12].

## 2 RESULTS FROM THE COMPARISON OF THE WAVE MODELS

In this section we present the results from the simulations with HAWASSI, SWASH, OceanWave3D, and HOS-NWT. We consider five test cases for which measurement results are also available from MARIN. The tests consist of three wave groups propagating over a flat bottom, and two irregular waves propagating over a slope. The details of the tests will be given in the corresponding section.

In the experiments a piston-type wavemaker is used at one end of the basin to generate waves. The motion, velocity and acceleration of the wavemaker was available from the experiment. At the other end of the basin a beach is installed to minimise reflections. The wave elevation was measured at various stations in the basin. All the data was provided at 50Hz.

The numerical simulations were all carried out in 2D. The waves are generated using the information from the piston-type wavemaker in SWASH (velocity of the wavemaker), OceanWave3D (velocity of the wavemaker) and HOS-NWT (motion of the wavemaker). With HAWASSI

| | Kinematics | Bathymetry | Workaround for wave breaking | Parallelization | Wave generation | | |
|---|---|---|---|---|---|---|---|
| | | | | | Piston | Flap | Time trace |
| SWASH | | | | (MPI) | | | |
| OceanWave3D | | | | Not yet public! (MPI) | | | |
| HOS-NWT | Available after post-proc. | | Work is ongoing! | | | | |
| HAWASSI | Available after post-proc. | | | | | | |

Table 1: Features of the codes selected for comparison. The version numbers of the codes are: SWASH2.00, HOS-NWT1.1, OceanWave3D0.99, and HAWASSI1.0. SWASH, HOS-NWT and OceanWave3D are open source, and capable of running 3D simulations. Using HAWASSI requires a licence, and only 2D simulations can be performed with this software. Green indicates that the code has the corresponding feature while red indicates that it does not or the work is ongoing. Kinematics: The code calculates the kinematics at each time level during the simulation in the entire domain. Bathymetry: Effects of bathymetry are included in the model. Workaround for wave breaking: In case of wave breaking, the code does not stop the simulation, but rather activates an *ad hoc* numerical workaround for continuation of simulation. Parallelization: A parallel version of the code is available and open source. Piston: It is possible to use information from a piston type wavemaker. Flap: It is possible to use information from a flap type wavemaker. Time trace: It is possible to use time trace of wave elevation from a measurement station.

the wave elevation history at the measurement location closest to the wavemaker was used. Additionally, we carried out grid and time step refinement with the codes. The results presented in this paper were obtained from the convergence study. For the sake of brevity, details of the convergence study are not presented here, but some information about the grid and time step for each code in each test case will be given in the corresponding section.

In order to compare the performance of the codes, we will use both qualitative and quantitative information. For qualitative assessment, we will provide wave elevation histories at various measurement stations. For quantitative assessment, we will resort to analyzing wave elevation histories in order to calculate characteristics of the waves. The significant wave height $H_{m_0}$, variance quotient, correlation, and relative error in $L_2$ will be provided in tables. $H_{m_0}$ is approximated as 4 times the square root of the variance. The correlation between the simulation and measurement is calculated by the 'cosine' correlation,

$$corr\left(\eta_e, \eta_s\right) = \frac{\eta_e \cdot \eta_s}{\|\eta_e\| \, \|\eta_s\|}. \tag{1}$$

This measures how well the experiment and the simulation align linearly, that is how well they point in the same direction. Therefore, this is especially useful to measure the phase shift errors in the simulation. The maximum value of 1 indicates that the simulation and the experiment are in phase, whereas the minimum value of -1 shows that the simulation is in counter phase with the experiment.

Additionally we will also use the relative error in $L_2$ between the measurement and the simulation,

$$\|\varepsilon\|_2 = \frac{\|\eta_e - \eta_s\|_2}{\|\eta_e\|_2}. \tag{2}$$

It is important to note that the relative error in $L_2$ may not be the ideal measure to compare the wave elevation from the simulation and experiment, since even small phase shifts can cause large error values. Nonetheless, it will give us an indication of the total error in the results.

We will also provide quantitative information particularly about the focused wave. In that respect we will pay attention to the crest height, trough depth, wave height, and time instance where the maximum amplitude occurs.

### 2.1 Wave groups propagating over a flat bottom

In this section we investigate the performance of the four codes in tests of propagating wave groups over a flat bottom. The tests consist of two focusing wave events, and a model scale of an irregular wave widely known as the Draupner wave or New Year's wave. The water depth is 1m and the length of the basin (from piston to the point where the beach intersects with the calm water surface) is 195.4m. The wave elevation was measured by six wave probes which were all located within the first 54m of the basin, see Tab. 2.

| | W1 | W2 | W3 | W4 | W5 | W6 |
|---|---|---|---|---|---|---|
| Distance to wave generator (m) | 10 | 20 | 40 | 49.5 | 50 | 54 |

Table 2: Location of the wave probes in the basin for the tests with a flat bottom.

### 2.1.1 Small-amplitude focusing wave (Experiment run number:202002)

Results for the test case 202002 were obtained using the following numerical settings. With SWASH, the grid resolution in the horizontal direction was $\Delta x = 0.01m$, and 20 layers were used in the vertical direction. CFL number was allowed to be between 0.2 and 0.5. With OceanWave3D, $\Delta x = 0.01m$, and 20 layers were used in the vertical direction with clustering towards the free surface. Half-width stencils were set to 3 and maximum CFL number was set to 0.5. With HOS-NWT, the number of point/modes were $N_x = 2048$, $N_z = 65$, and the HOS order was set to $M = 5$. Finally with HAWASSI, $2^{12}$ points were used in the simulations with a time step of $0.02s$.

At the measurement location W2, we observe that the wave signals from the four codes are overall similar as shown in Fig. 1a. However, it is possible to observe that HAWASSI, compared to the other three codes, slightly overestimates the wave heights, and results in slightly larger phase shifts especially for shorter waves.

At the measurement location W4 which is half a meter upstream of the focal point W5, we start to observe occurrence of a focussed wave as depicted in Fig. 1b. Focusing takes place at W5 as illustrated in Fig. 1c at $t = 109.34s$, and information about the focused waves at W4 and W5 is listed in Tab. 3. The results show that SWASH, OceanWave3D and HOS-NWT perform somewhat similarly. SWASH and HOS-NWT produce the smallest phase shift error. SWASH produces the smallest wave heights at both W4 and W5 underestimating both the crest height and trough depth of the focused waves. HAWASSI yields the largest phase shift and overestimates the trough depths at W4 and W5, however it estimates the crest heights more accurately than the other three codes at both locations. Considering both amplitude and phasing HOS-NWT performs better than SWASH and OceanWave3D.

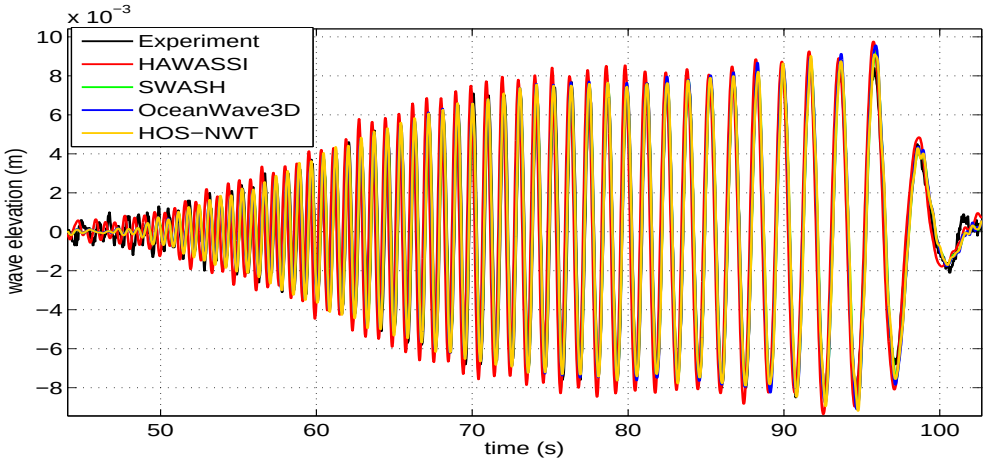
| | | Experiment | HAWASSI | SWASH | OceanWave3D | HOS-NWT |
|---|---|---|---|---|---|---|
| W4 | Crest (m) | 0.046 | 0.044 | 0.042 | 0.041 | 0.043 |
| | Trough (m) | -0.037 | -0.04 | -0.035 | -0.038 | -0.037 |
| | Wave height (m) | 0.083 | 0.084 | 0.077 | 0.079 | 0.08 |
| | $t_{peak}$ (s) | 109.04 | 108.91 | 109.0 | 108.97 | 109.0 |
| W5 | Crest (m) | 0.055 | 0.055 | 0.049 | 0.052 | 0.052 |
| | Trough (m) | -0.025 | -0.029 | -0.024 | -0.027 | -0.025 |
| | Wave height (m) | 0.08 | 0.084 | 0.073 | 0.079 | 0.077 |
| | $t_{peak}$ (s) | 109.34 | 109.22 | 109.33 | 109.3 | 109.33 |

Table 3: Maximum crest height, minimum trough depth, wave height, and time instance of the maximum amplitude from the measurement and simulations at W4 and W5 for the case 202002.

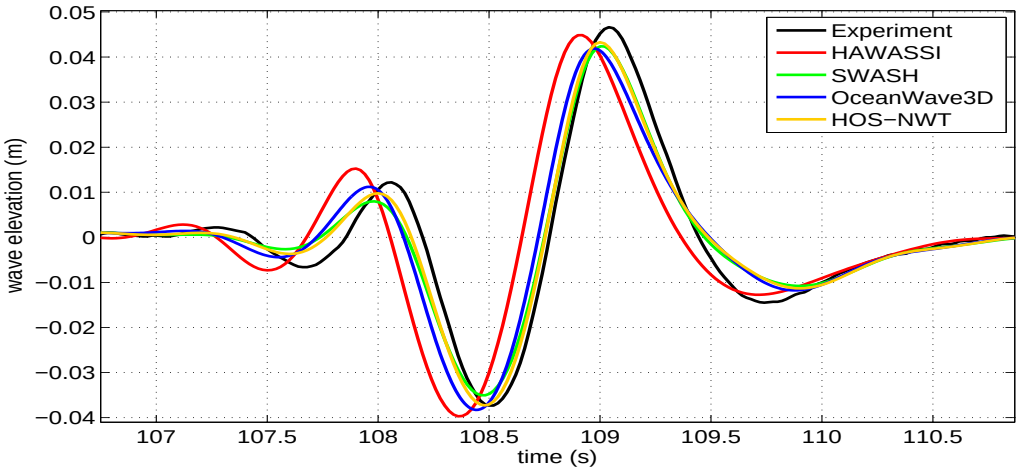
### 2.1.2 Large-amplitude focusing wave (Experiment run number:203001)

In the test case 203001, the same settings as the case 202002 were used for the four codes. Figure 2a displays wave signals from the experiment and the three codes at W2. Here, close observation of the wave signals reveals that SWASH performs slightly better than the other three codes yielding smaller phase shift, and estimating wave heights more accurately.

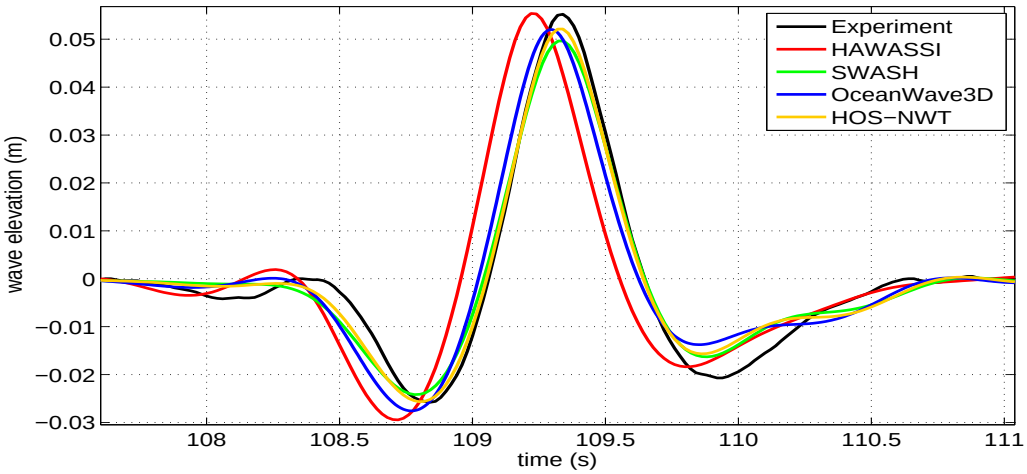
At the measurement locations W4 and W5 which are only half a meter apart from each other, we observe the focusing event as demonstrated in Figs. 2b and 2c. At both locations we detect




(a) W2




(b) W4




(c) W5

Figure 1: Comparison of the wave signals at the measurement location W2, W4 and W5 for the case 202002.

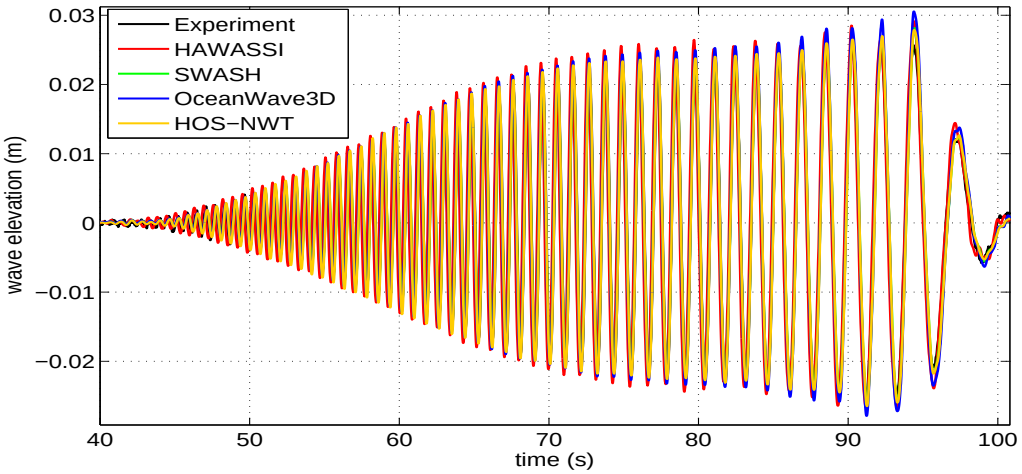
three peaks in the wave signals. Table 4 lists the information about the three peaks at W5, which occur at $t = 105.75s$, $t = 106.56s$, and $t = 107.86s$. For these three peaks HAWASSI results in the largest phase shift compared to the other codes. Analyzing the peaks occurring at $t = 106.24s$ and $t = 107.6s$ at W4 we observe that the wave crests are distorted in the wave signal from HOS-NWT. When we look at the wave profile of the highest peak occurring at $t = 107.86s$ at W5, we also notice some distortion in the wave crest from SWASH. For the highest peaks at both W4 and W5 the agreement between the OceanWave3D result and the experiment is good. Overall, OceanWave3D shows a better performance than the other three codes in this test case. In the simulations with HAWASSI and OceanWave3D the mechanisms to prevent wave breaking from happening were activated. With SWASH more than 10 layers in the vertical direction were used, and in the SWASH manual [13] it is mentioned that no additional measure needs to be taken in such a situation to account for wave breaking. With HOS-NWT no *ad hoc* measure was taken for wave breaking because HOS-NWT did not yet have a mechanism for this purpose. All these choices influenced the results inevitably, and the extent of this influence should not be underestimated. When the mechanisms for prevention of wave breaking are activated, the user is usually left to find a value for a *strength* parameter. Since there is not a clear guideline for the value of this parameter (typically there are only minimum and maximum limits), different values were chosen and their effects were observed. The results shown in this paper were those which were in the best agreement with the measurements.

| | | Experiment | HAWASSI | SWASH | OceanWave3D | HOS-NWT |
|---|---|---|---|---|---|---|
| P1 | Crest (m) | 0.029 | 0.054 | 0.012 | 0.038 | 0.040 |
| | Trough (m) | -0.051 | -0.062 | -0.023 | -0.059 | -0.061 |
| | Wave height (m) | 0.08 | 0.116 | 0.035 | 0.097 | 0.101 |
| | $t_{peak}$ (s) | 105.75 | 105.5 | 105.62 | 105.64 | 105.68 |
| P2 | Crest (m) | 0.08 | 0.071 | 0.073 | 0.072 | 0.088 |
| | Trough (m) | -0.079 | -0.082 | -0.102 | -0.087 | -0.072 |
| | Wave height (m) | 0.159 | 0.153 | 0.175 | 0.159 | 0.16 |
| | $t_{peak}$ (s) | 106.56 | 106.32 | 106.6 | 106.42 | 106.44 |
| P3 | Crest (m) | 0.106 | 0.102 | 0.093 | 0.099 | 0.113 |
| | Trough (m) | -0.039 | -0.043 | -0.029 | -0.038 | -0.045 |
| | Wave height (m) | 0.145 | 0.145 | 0.122 | 0.137 | 0.158 |
| | $t_{peak}$ (s) | 107.86 | 107.66 | 107.92 | 107.79 | 107.8 |

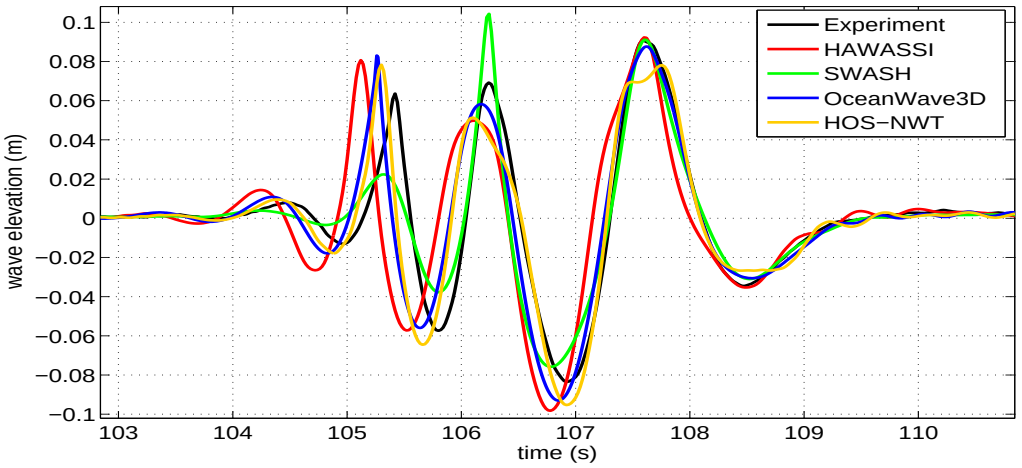
Table 4: Maximum crest height, minimum trough depth, wave height, and time instance of the three peaks at W5 for the case 203001.

### 2.1.3 The Draupner Wave (Experiment run number:204001)

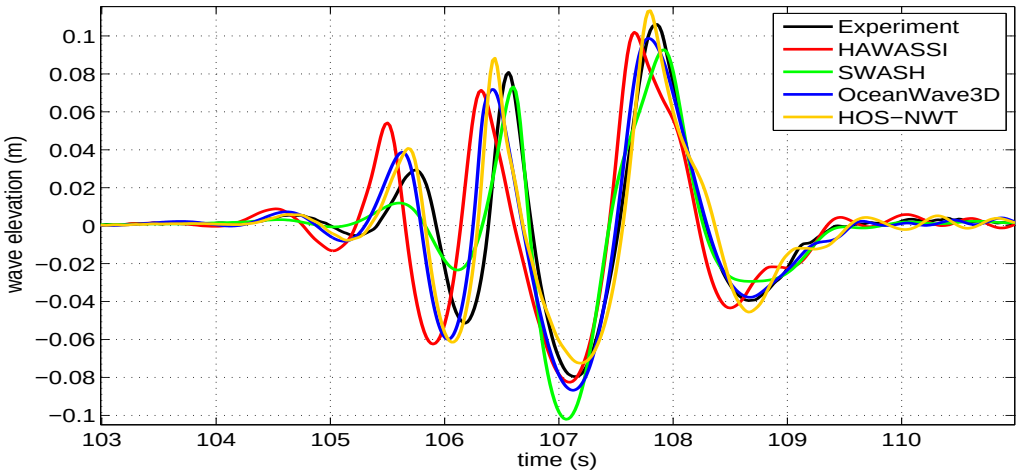
The last test case over a flat bottom is the model scale of an irregular wave widely known as the Draupner Wave or New Year's wave. For this test case, the same settings as the case 202002 were used with SWASH and HOS-NWT, but we had several difficulties with OceanWave3D. The results from OceanWave3D were obtained on a grid with $\Delta x = 0.2m$, and $5$ layers in the vertical direction. The time step was $0.02s$. Any attempt to refine the grid or change the time step was unsuccessful, as the solution diverged in all those attempts. This problem was reported




(a) W2




(b) W4




(c) W5

Figure 2: Comparison of the wave signals at the measurement location W2, W4 and W5 for the case 203001.

back to the developers of the OceanWave3D code. With HAWASSI, $2^{11}$ points were used in the simulations with a time step of $0.02s$.

At W2, W3, W4, W5 and W6, the wave signals from the three codes were analyzed, and the quantitative information about simulation accuracy is listed in Tab. 5. Similar to the test cases 202002 and 203001, a focussing event takes place in the case of 204001, and the wave probe W5 records the most extreme event occurring at $t = 164.6s$. The information about the focussed wave in particular is also provided in Tab. 6.

At the measurement station W2, we observe a good performance from all the codes resulting in at least $90\%$ correlation with the experiment. Especially in the bottom plot of Fig. 3 where more energetic part of the wave group is illustrated, we observe a good agreement between the four simulations and the experiment. In the less energetic part of the wave group, however, we notice a clear difference between OceanWave3D and the other codes as shown in the middle plot of Fig. 3. This difference can be attributed to the low grid resolution in the OceanWave3D simulation, since these short and small waves are possibly not resolved on the grid. This also causes OceanWave3D to generate the highest relative error in $L_2$.

Among all the measurement stations, W5 records the most extreme event taking place at $t = 164.6s$. When we look at the values in Tab. 5, we see that OceanWave3D and HOS-NWT correspond with the experiment in a better fashion than SWASH and HAWASSI at W5. We notice this outcome also in the close-up view of the focusing event shown in the bottom plot of Fig. 4.

### 2.2 Wave groups propagating over a slope

In this section, we will present results from simulations with irregular waves propagating over a slope. These tests were not carried out with HOS-NWT since bathymetry effects had not been included in the model of HOS-NWT when this study was conducted. The geometrical setup, the details of the bathymetry, and the positions of the measurement stations are shown in Figure 5. The waves were generated with a piston-type wavemaker making only horizontal translations. 143.41m away from the wavemaker the slope with a steepness of $5\%$ starts. The water depth in the deeper part is 0.6m, and the water depth in the shallower part is 0.3m. In the experiment, a beach is located at 173.41m away from the wavemaker. We will consider two test cases listed in Table 7. The duration of the experiment in both cases was nearly 1840s, but we carried out the simulations only for 500s which contains approximately 316 waves.

#### 2.2.1 Test case:103001

Results for the test case 103001 were obtained using the following numerical settings. With SWASH, the grid resolution in the horizontal direction was $\Delta x = 0.025m$, and 20 layers were used in the vertical direction. CFL number was allowed to be between 0.2 and 0.5. With OceanWave3D, $\Delta x = 0.025m$, and 20 layers were used in the vertical direction with clustering towards the free surface. Half-width stencils were set to 2 and maximum CFL number was set to 0.5. Finally with HAWASSI, $2^{11}$ points were used in the simulations with a time step of $0.08s$.

At the measurement stations the wave signals were analyzed, and the quantitative information about simulation accuracy is listed in Table 8. We observe that HAWASSI underestimates wave heights at all measurement stations, while SWASH and OceanWave3D overestimate them. This can be seen in both $H_{m_0}$ and variance quotient values from the three simulations. For this test case OceanWave3D generally shows better performance than the other codes producing

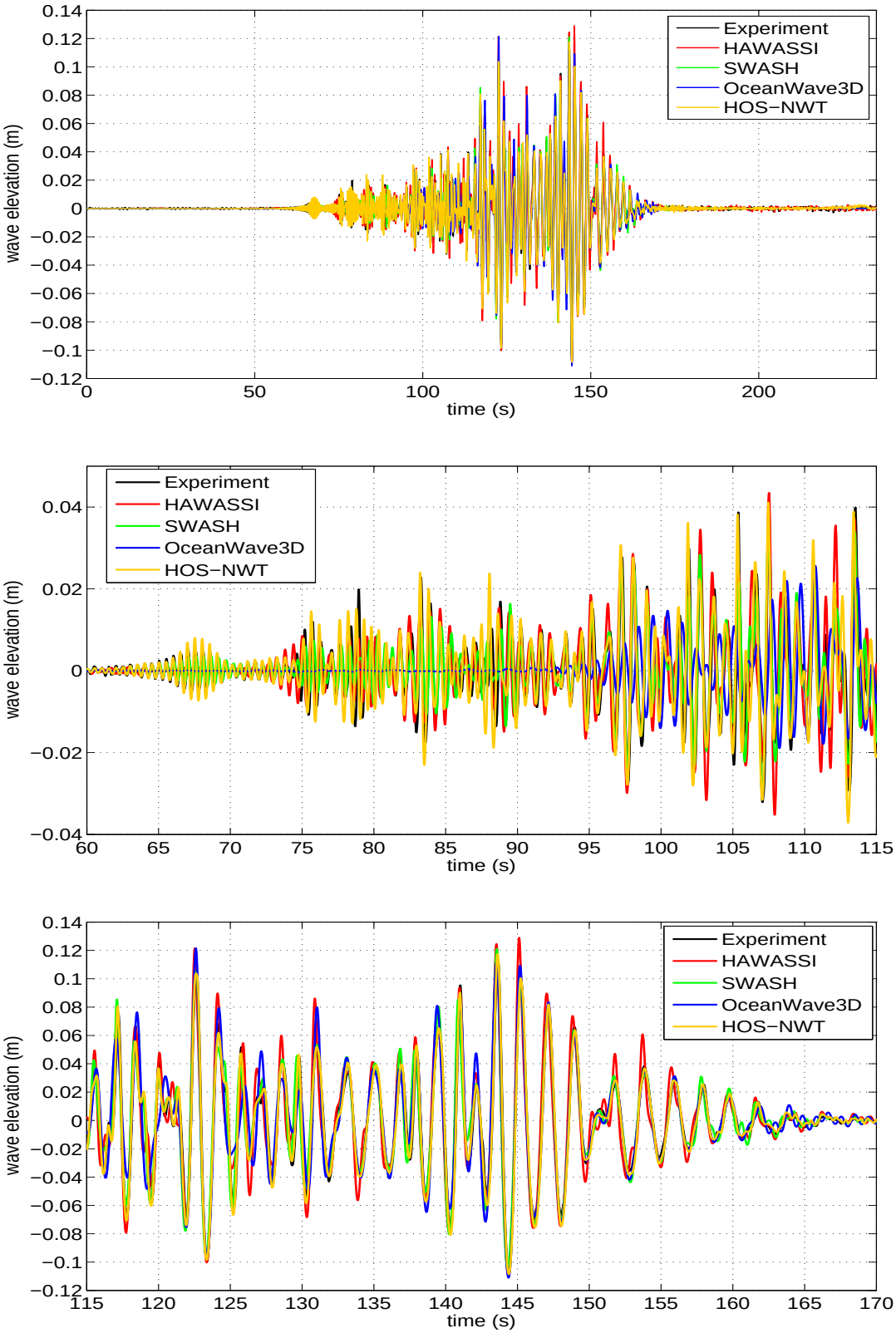



Figure 3: Comparison of the wave signals at the measurement location W2 for the case 204001. Top plot shows the complete time history, and the other two plots show close-up views of the two parts of the top plot.

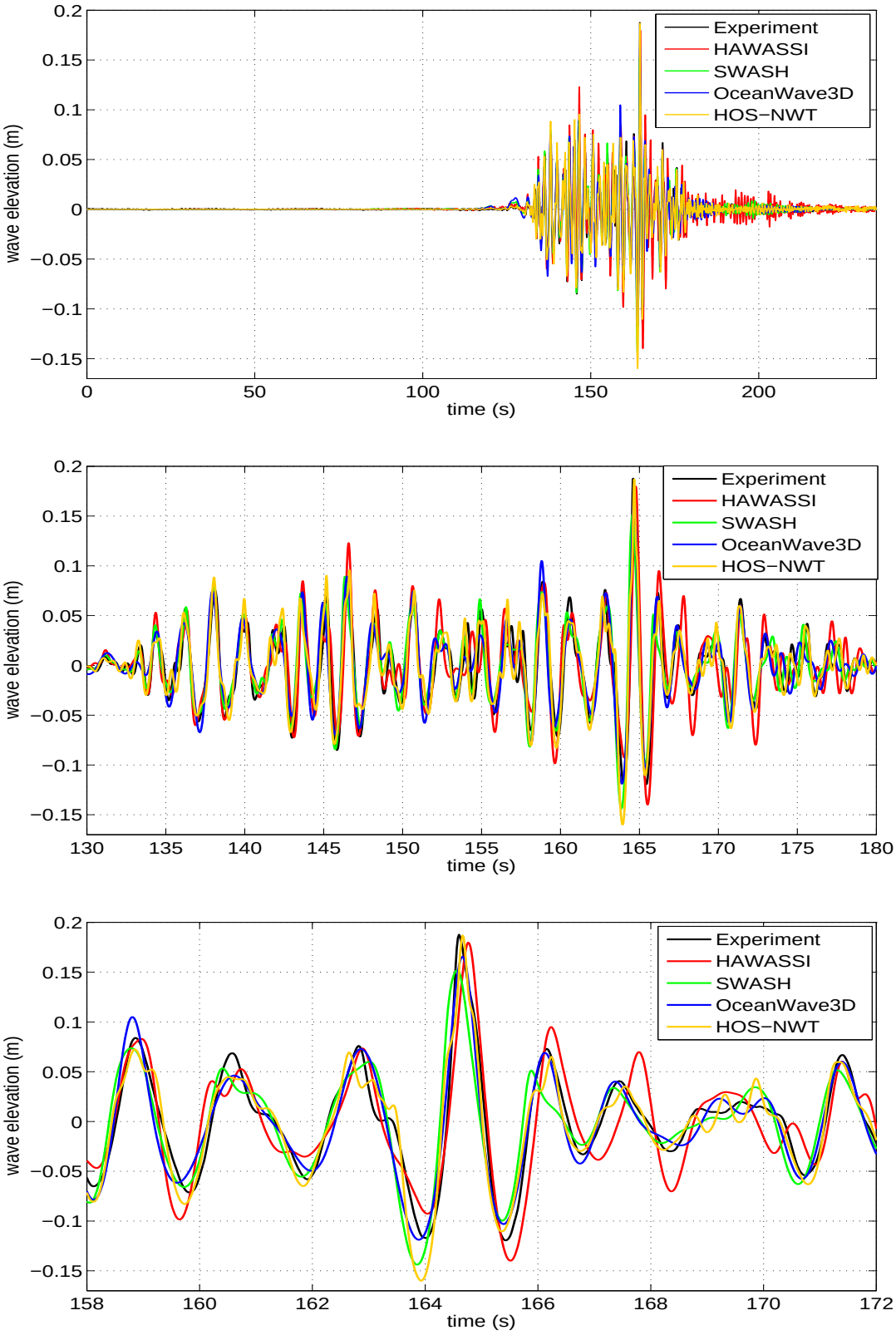



Figure 4: Comparison of the wave signals at the focal point W5 for the case 204001. Top plot shows the complete time history, middle plot shows the close-up view of the wave group, and the bottom plot shows the freak-like wave which occurs at $t = 164.6s$.

| | | HAWASSI | SWASH | OceanWave3D | HOS-NWT |
|---|---|---|---|---|---|
| W2 | $H_{m_0}$ (sim, exp) (m) | (0.075, 0.069) | (0.069, 0.069) | (0.070, 0.069) | (0.071, 0.069) |
| | Variance quotient | 1.17 | 1.0 | 1.02 | 1.06 |
| | Correlation | 0.91 | 0.94 | 0.90 | 0.97 |
| | $L_2$ norm | 0.44 | 0.34 | 0.45 | 0.21 |
| W3 | $H_{m_0}$ (sim, exp) (m) | (0.075, 0.072) | (0.067, 0.072) | (0.069, 0.072) | (0.070, 0.072) |
| | Variance quotient | 1.07 | 0.87 | 0.92 | 0.95 |
| | Correlation | 0.87 | 0.87 | 0.88 | 0.93 |
| | $L_2$ norm | 0.52 | 0.50 | 0.48 | 0.36 |
| W4 | $H_{m_0}$ (sim, exp) (m) | (0.075, 0.071) | (0.068, 0.071) | (0.069, 0.071) | (0.070, 0.071) |
| | Variance quotient | 1.11 | 0.93 | 0.96 | 0.98 |
| | Correlation | 0.85 | 0.85 | 0.84 | 0.89 |
| | $L_2$ norm | 0.55 | 0.54 | 0.56 | 0.45 |
| W5 | $H_{m_0}$ (sim, exp) (m) | (0.075, 0.07) | (0.068, 0.07) | (0.069, 0.07) | (0.071, 0.07) |
| | Variance quotient | 1.15 | 0.94 | 0.97 | 1.03 |
| | Correlation | 0.84 | 0.87 | 0.89 | 0.91 |
| | $L_2$ norm | 0.58 | 0.50 | 0.45 | 0.42 |
| W6 | $H_{m_0}$ (sim, exp) (m) | (0.074, 0.071) | (0.067, 0.071) | (0.069, 0.071) | (0.071, 0.071) |
| | Variance quotient | 1.11 | 0.91 | 0.95 | 1.00 |
| | Correlation | 0.81 | 0.86 | 0.88 | 0.89 |
| | $L_2$ norm | 0.64 | 0.51 | 0.47 | 0.46 |

Table 5: Quantitative information concerning simulation accuracy for the case 204001.

| | Experiment | HAWASSI | SWASH | OceanWave3D | HOS-NWT |
|---|---|---|---|---|---|
| Crest (m) | 0.187 | 0.18 | 0.152 | 0.165 | 0.187 |
| Trough (m) | -0.117 | -0.092 | -0.143 | -0.119 | -0.16 |
| Wave height (m) | 0.304 | 0.272 | 0.295 | 0.284 | 0.347 |
| $t_{peak}$ (s) | 164.6 | 164.76 | 164.56 | 164.66 | 164.66 |

Table 6: Maximum crest height, minimum trough depth, wave height, and time instance of the maximum amplitude for the freak-like wave at W5 from the case 204001.

| Run number | $H_s$ (m) | $T_p$ (s) |
|---|---|---|
| 103001 | $\cong 0.06$ | 1.697 |
| 104001 | $\cong 0.06$ | 2.121 |

Table 7: Characteristics of the irregular waves propagating over a slope.

better correlation and smaller relative error in $L_2$. Measurement station W12 attracts our attention because this is where the three codes show their worst performance. Further investigation may be worthwhile to find an explanation for this observation.

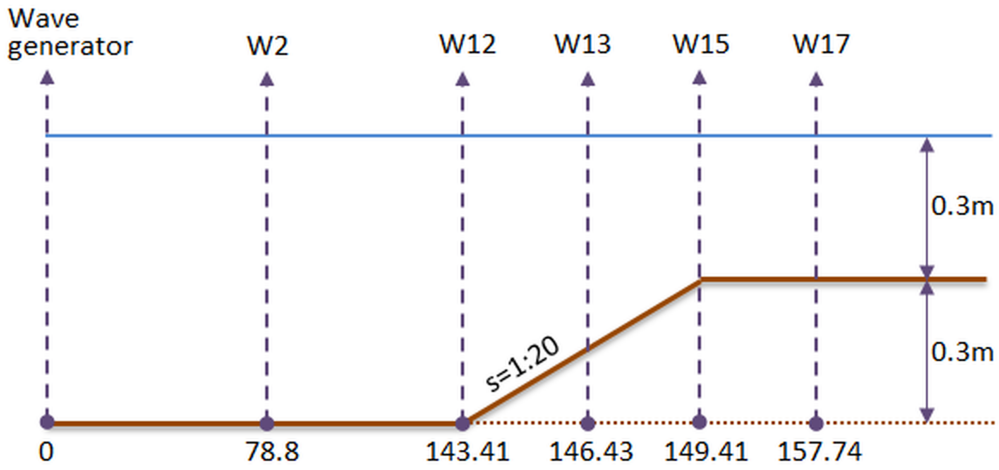



Figure 5: Setup of the experiment for the tests with irregular waves propagating over a slope. Piston-type wave generator is located at the left end of the basin. Wave elevations are measured at W2, W12, W13, W15 and W17. Slope has a steepness of 1:20. A beach is located at 173.41m away from the wavemaker (not shown in the figure). Distances are given in meters, and are not on scale.

| | | HAWASSI | SWASH | OceanWave3D |
|---|---|---|---|---|
| W2 | $H_{m_0}$ (sim, exp) (m) | (0.056, 0.06) | (0.063, 0.06) | (0.064, 0.06) |
| | Variance quotient | 0.88 | 1.15 | 1.16 |
| | Correlation | 0.91 | 0.85 | 0.88 |
| | $L_2$ norm | 0.41 | 0.57 | 0.52 |
| W12 | $H_{m_0}$ (sim, exp) (m) | (0.053, 0.054) | (0.059, 0.054) | (0.058, 0.054) |
| | Variance quotient | 0.95 | 1.18 | 1.17 |
| | Correlation | 0.52 | 0.45 | 0.80 |
| | $L_2$ norm | 0.97 | 1.10 | 0.67 |
| W13 | $H_{m_0}$ (sim, exp) (m) | (0.054, 0.056) | (0.06, 0.056) | (0.059, 0.056) |
| | Variance quotient | 0.92 | 1.11 | 1.09 |
| | Correlation | 0.84 | 0.75 | 0.90 |
| | $L_2$ norm | 0.55 | 0.73 | 0.46 |
| W15 | $H_{m_0}$ (sim, exp) (m) | (0.055, 0.06) | (0.062, 0.06) | (0.061, 0.06) |
| | Variance quotient | 0.86 | 1.07 | 1.05 |
| | Correlation | 0.83 | 0.73 | 0.91 |
| | $L_2$ norm | 0.57 | 0.75 | 0.43 |
| W17 | $H_{m_0}$ (sim, exp) (m) | (0.054, 0.058) | (0.054, 0.058) | (0.06, 0.058) |
| | Variance quotient | 0.88 | 0.87 | 1.09 |
| | Correlation | 0.67 | 0.70 | 0.89 |
| | $L_2$ norm | 0.79 | 0.76 | 0.48 |

Table 8: Quantitative information concerning simulation accuracy for the case 103001.

### 2.2.2 Test case:104001

In the test case 104001, the same settings as the case 103001 were used for the three codes. Table 9 lists quantitative information about simulation accuracy based on the analysis of the wave signals at the measurement stations. HAWASSI, SWASH and OceanWave3D show a good agreement with the experiment at the measurement station W2. Here the correlation with the experiment is $95\%$ with HAWASSI, $94\%$ with SWASH, and $90\%$ with OceanWave3D. As the wave propagates, we observe that both HAWASSI and SWASH fail to achieve these high correlation values at the other measurement stations. OceanWave3D, however, maintains high correlation values at all the five stations: $90\%$ at W2, $88\%$ at W12, $90\%$ at W13, $91\%$ at W15, and $89\%$ at W17. Also, at W12, W13, W15 and W17 OceanWave3D yields much smaller errors in $L_2$ than the other two codes.

| | | HAWASSI | SWASH | OceanWave3D |
|---|---|---|---|---|
| W2 | $H_{m_0}$ (sim, exp) (m) | (0.06, 0.062) | (0.065, 0.062) | (0.064, 0.062) |
| | Variance quotient | 0.92 | 1.10 | 1.06 |
| | Correlation | 0.95 | 0.94 | 0.90 |
| | $L_2$ norm | 0.32 | 0.37 | 0.45 |
| W12 | $H_{m_0}$ (sim, exp) (m) | (0.058, 0.063) | (0.062, 0.063) | (0.061, 0.063) |
| | Variance quotient | 0.84 | 0.98 | 0.96 |
| | Correlation | 0.64 | 0.60 | 0.88 |
| | $L_2$ norm | 0.82 | 0.89 | 0.49 |
| W13 | $H_{m_0}$ (sim, exp) (m) | (0.06, 0.067) | (0.064, 0.067) | (0.063, 0.067) |
| | Variance quotient | 0.80 | 0.91 | 0.90 |
| | Correlation | 0.85 | 0.78 | 0.90 |
| | $L_2$ norm | 0.53 | 0.64 | 0.43 |
| W15 | $H_{m_0}$ (sim, exp) (m) | (0.063, 0.072) | (0.068, 0.072) | (0.067, 0.072) |
| | Variance quotient | 0.76 | 0.88 | 0.87 |
| | Correlation | 0.85 | 0.76 | 0.91 |
| | $L_2$ norm | 0.53 | 0.67 | 0.42 |
| W17 | $H_{m_0}$ (sim, exp) (m) | (0.062, 0.064) | (0.068, 0.064) | (0.067, 0.064) |
| | Variance quotient | 0.96 | 1.13 | 1.11 |
| | Correlation | 0.71 | 0.66 | 0.89 |
| | $L_2$ norm | 0.76 | 0.85 | 0.48 |

Table 9: Quantitative information concerning simulation accuracy for the case 104001.

## 2.3 Summary and discussion

* For a comparative study in simulation of free surface water waves, four codes were selected: SWASH, OceanWave3D, HOS-NWT and HAWASSI. SWASH, OceanWave3D and HOS-NWT are open source and capable of running simulations in 3D, while HAWASSI requires a licence and can only run simulations in 2D. Nevertheless, a demo version of HAWASSI with fewer functionalities can be downloaded from its website.

* To assess their performance, a set of wave measurements is used; three wave groups propagating on a flat bottom and two irregular waves propagating on a slope. Since the design of HOS-NWT does not address varying bathymetry, this code is only used in the cases with flat bottom.

* The performance of HAWASSI in terms of computational cost is superior by a clear margin; in many cases the duration of simulation was close to the duration of real-time. This suggests that HAWASSI can be an ideal candidate to investigate statistics of extreme waves and wave groups. Computational cost of the other three codes was in the same order of magnitude. Therefore, a detailed analysis on this subject is not provided here. Even though grid and time step refinement studies were conducted, these were not presented here for the sake of brevity.

* In some cases, there were divergence issues with the OceanWave3D solution. At the time this paper was compiled, these issues were still not resolved. An important note is that when the solution diverges, the code does not stop the simulation, but rather continues until the end of the pre-specified simulation time. This is inconvenient, and it may inevitably waste computational resources.

* OceanWave3D allows using the velocity of only piston type wavemaker when the reproduction of basin waves is in question. This unfortunately limits the applicability of the code. Through personal communication OceanWave3D developers noted that there is currently no plan to include the use of flap type wavemaker information. This suggests that if this functionality is needed in a test case, either SWASH or HOS-NWT can be used.

* Various robustness issues were detected with SWASH. The code sometimes stops the simulation with a warning which does not give a clear indication of what went wrong. On the other hand, SWASH has some unique features, for example, it is the only code which can run a simulation on varying bathymetry using the information from a flap type wavemaker. SWASH also contains models for conservative transport of salinity, temperature and suspended sediment. The software has a well-written manual and an active user-community.

* HOS-NWT has some important limitations in functionality, for example, the code lacks a numerical workaround for wave breaking (through personal communication we were informed that the work on this subject is ongoing), and bathymetry effects are not incorporated into the model. Nevertheless the code showed a good performance in the test cases, and future improvements on its design will be monitored closely.

* Running simulations in parallel is an important functionality especially in 3D problems. One of the objectives in this work is to couple one of these codes with ReFRESCO which is mostly used in parallel. When ReFRESCO is run in parallel and the other code in serial, the efficiency of parallel computation may be compromised. At the moment only SWASH and OceanWave3D have parallel versions, and only SWASH has parallel open source version. Through personal communication OceanWave3D developers stated that the licence for the parallel version of OceanWave3D is not yet available.

* SWASH and OceanWave3D have the kinematics available in the entire domain at every time step. HOS-NWT and HAWASSI calculate the kinematics during a post-processing

step. This feature is an important requirement especially when coupling one of these codes with ReFRESCO is in question.

* Results indicate that none of the codes is clearly superior to the other ones especially in terms of accuracy. After an overall assessment of the codes based on their functionalities, accuracy, robustness and efficiency, OceanWave3D is selected for one-way coupling with ReFRESCO. However, the improvements on all the four models in the future will be monitored closely.

## 3 ONE-WAY COUPLING WITH REFRESCO

### 3.1 ReFRESCO Code

ReFRESCO (http://www.refresco.org) is a community based open-usage CFD code targeted for Maritime World problems. It solves multiphase (unsteady) incompressible viscous flows using the Navier-Stokes equations, complemented with turbulence models, cavitation models and volume-fraction transport equations for different phases [1] (in this particular case at hand both the air and water phases are modelled). The equations are discretized using a finite-volume approach with cell-centered collocated variables, in strong-conservation form, and a pressure-correction equation based on the SIMPLE algorithm is used to ensure mass conservation. The implementation is face-based, which permits grids with elements consisting of an arbitrary number of faces and if needed h-refinement (hanging nodes). Time integration is performed implicitly with first or second-order backward schemes. This means that there are no Courant number stability limitations, contrary to other codes used for tackling free-surface and wave problems [14]. However, small Courant numbers are important for a good balance between spatial and temporal discretizations, and also help the iterative convergence of the non-linear problem.

In a free-surface capturing method, such as the one here used, the discretization of the convective terms, both for momentum and volume-fraction transport equations, have an important effect in the robustness and in the accuracy of the solutions [15]. For the momentum a standard 2nd order unstructured-grids QUICK scheme is used. For the volume-fraction transport equation, it is known that in order to keep the free-surface interface sharp, compressive or anti-diffusive schemes have to be used [1]. In ReFRESCO one has several of these schemes, such as the well-known HRIC [16] and CICSAM [17]. However, these are Courant-number dependent and keep the compressive character only for Courant numbers lower than 1, which for an time-implicit code and for steady calculations is troublesome if not questionable. Therefore, in ReFRESCO, a compressive scheme ReFRICS is used for which the compressive character of the scheme is only based on the direction of the grid lines and the free-surface normal, and where there are no restrictions on the Courant number. For the rest of the discretization, the diffusion terms are treated using central schemes and the time derivatives using a three-time-level implicit scheme.

ReFRESCO (v2.2.0) is currently being developed, verified and its several applications validated at MARIN (in the Netherlands), in collaboration with IST (in Portugal), USP-TPN (University of Sao Paulo, Brasil), TUDelft (Technical University of Delft, the Netherlands), Soton (University of Southampton, UK), UTwente (University of Twente, the Netherlands) and Chalmers (Chalmers University, Sweden).

### 3.2 Coupling Procedure

For the sake of brevity we will not present the mathematical and numerical designs of OceanWave3D and ReFRESCO. For this purpose several references were previously mentioned. For one-way coupling of OceanWave3D with OpenFOAM®, see [18].

The information transfer from OceanWave3D to ReFRESCO takes place only at the inflow boundary of ReFRESCO, see Fig. 6. This transfer can be carried out in two ways:

1. One-way coupling of the two codes: Solutions in both codes advance in time simultaneously, and ReFRESCO calls OceanWave3D at every time step to receive the flow information. This information is then interpolated onto the inflow boundary of ReFRESCO.

2. Running simulation in each code separately and consecutively: First the simulation is performed using the potential flow solver, and the flow information in the relevant part of the domain is stored to a file. Secondly this file is read by ReFRESCO to assign values to the flow variables at the inflow boundary. Any one of the three codes (SWASH, OceanWave3D and HOS-NWT) can very well be used in the first step as long as all the necessary flow information is outputted and stored to a file. This method allows ReFRESCO to have a very flexible environment in order to generate waves. Furthermore, since the complete solution history is at the disposal of the user before starting the simulation in ReFRESCO, a certain part of that history can be cropped and sent into ReFRESCO instead of the complete history.

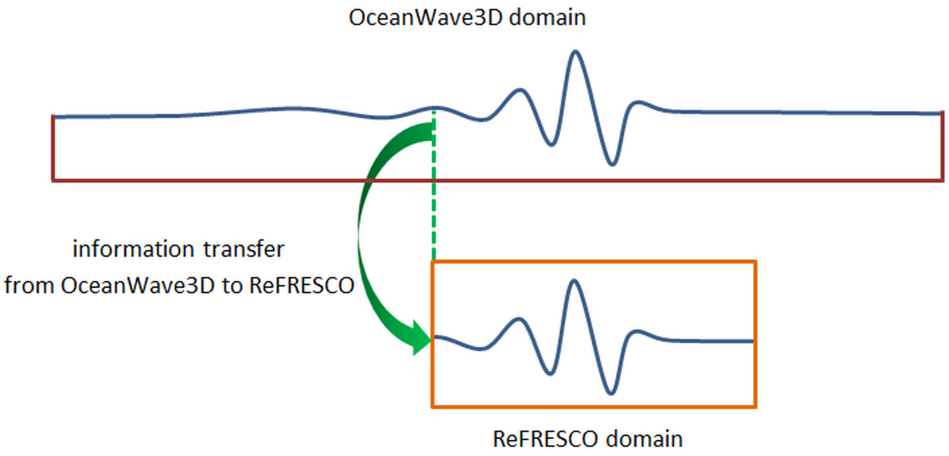



Figure 6: An example layout of the OceanWave3D and ReFRESCO domains, and the one-way coupling.

Both methods should produce the same result since the information that is transferred into ReFRESCO is essentially the same, only the way it is transferred changes. Therefore, in the remainder of this paper we will show results from only the first method.

In this method solutions in both codes advance in time simultaneously, and ReFRESCO calls OceanWave3D at every time step to receive the flow information. This information is then interpolated onto the inflow boundary of ReFRESCO. Both codes use the same time step, therefore no interpolation in time is necessary between the two solutions. At the beginning of each time step, the velocity and free surface elevation values are received from OceanWave3D, and imposed at the inflow boundary of the ReFRESCO domain. Since both codes use different

computational grids, the information from OceanWave3D has to be interpolated onto the grid of ReFRESCO. This is the only interpolation procedure used in the entire one-way coupling of both codes. The interpolation is done in the computational domain of OceanWave3D. The target value of a flow quantity $\psi_{target}\left(\boldsymbol{x}+\delta \boldsymbol{e}_i\right)$ at the location $\boldsymbol{x}+\delta \boldsymbol{e}_i$ whose nearest neighbour point located at $\boldsymbol{x}$ is calculated using the following finite Taylor series

$$\psi_{target}\left(\boldsymbol{x}+\delta \boldsymbol{e}_i\right)=\sum_{n=0}^{2a} \frac{\delta^n}{n!} \frac{\partial^n \psi}{\partial x_i^n}(\boldsymbol{x}) \tag{3}$$

where $a$ is user-defined. The gradients $\partial^n \psi / \partial x_i^n$ were computed previously during the discretization of the governing equations. The number of terms in the Taylor series, $2a+1$, is consistent with the size of the finite difference stencil of the solver. Therefore both the Taylor series and the computed wave solution have a consistent numerical discretization error.

### 3.3 Results from the ReFRESCO-OceanWave3D coupling

In order to refer to OceanWave3D/ReFRESCO coupling, simply ReFRESCO will be used in the remainder of this document. A close-up view of a typical grid used in the simulations with ReFRESCO is shown in Fig. 7. The grid is finely spaced near the undisturbed interface between water and air, but becomes coarser towards the bottom and top boundaries of the domain. Along the horizontal direction no grid stretching was applied. The size of the grid cells near the undisturbed interface is given below in each test case.

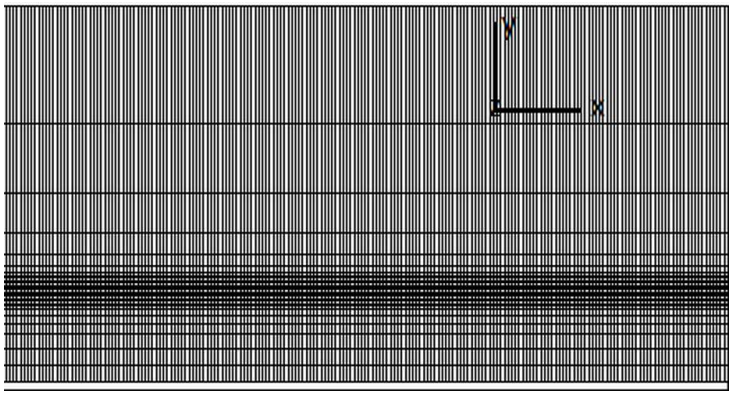



Figure 7: Close-up view of a typical grid used for the wave simulations.

In all the test cases, a three-level time discretization scheme is used. The time step was fixed throughout the simulations for all equations. Convergence tolerance is set to $10^{-4}$ for the infinity norm of the residuals of all the flow variables. In order to achieve such a residual decrease, the maximum value for the number of outer loops was set to 150.

The momentum and volume fraction equations are solved using GMRES with BJACOBI preconditioner, while the pressure correction equation is solved using CG with BJACOBI preconditioner. No turbulence model is used.

Dynamic viscosity of water is set to $1.002 \times 10^{-3} kg \cdot m^{-1} \cdot s^{-1}$, and air to $1.824 \times 10^{-5} kg \cdot m^{-1} \cdot s^{-1}$. For the density values of water and air, $998 kg \cdot m^{-3}$ and $1.118 kg \cdot m^{-3}$ are used, respectively.

The same boundary conditions are used in all the simulations with ReFRESCO, which are shown in Fig. 8. The outflow boundary, where the Sommerfeld boundary condition [19] is

applied, is positioned in such a way that reflected waves do not disturb the solution in the part of the domain where wave probes are located.

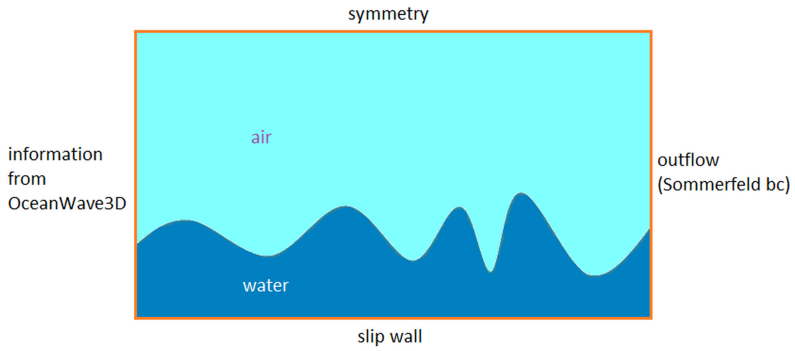



Figure 8: Boundary conditions for the simulations with ReFRESCO.

### 3.3.1 Small-amplitude focusing wave (Experiment run number:202002)

The wave probe W3 is the inflow boundary of ReFRESCO, where the quantities from OceanWave3D are imposed, see Tab. 2 for the locations of the wave probes. Three grid resolutions are considered with ReFRESCO: $(\Delta x = 0.04m,\ \Delta z = 0.016m)$, $(\Delta x = 0.02m,\ \Delta z = 0.008m)$, and $(\Delta x = 0.01m,\ \Delta z = 0.004m)$. Time step size is equal to 0.005s, and the simulation is carried out for 120s.

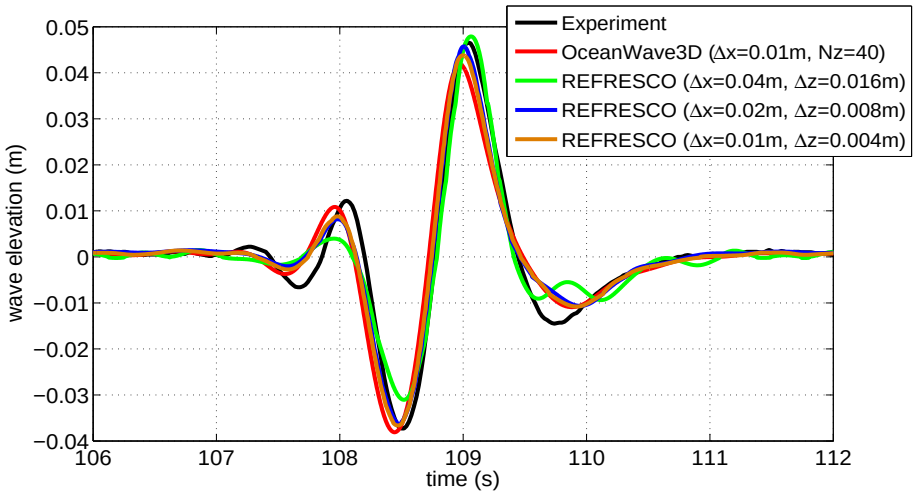
Figure 9 shows the wave signals at the measurement stations W4 and W5. We observe that the result on the coarse grid shows relatively larger deviations in terms of both wave height and phase shift compared to the medium and fine grids. The results on the medium and fine grids are, however, considerably similar and in good agreement with the measurement.

The residual history with time is plotted in Fig. 10 for the three grid resolutions. Note that only the last residual value for each time step is plotted for all the flow variables. The residuals for all the flow variables generally decrease around 5 orders of magnitude on all the three grids. The only exception to this is observed on the fine grid between $t = 107s$ and $t = 111s$, where 3.5 orders of magnitude decrease is achieved for some time levels, see Fig. 10c. When we look at the wave elevation histories from the wave probes W4 and W5, we notice that this time window coincides with the focusing event. During the focusing event the flow exhibits the highest velocities, which results in a considerable increase in the CFL number, see Fig. 11. This can be the cause for the relatively higher residuals on the fine grid. Small CFL numbers observed for the large part of the simulation result in small discretization errors and help the iterative convergence of the non-linear problem.

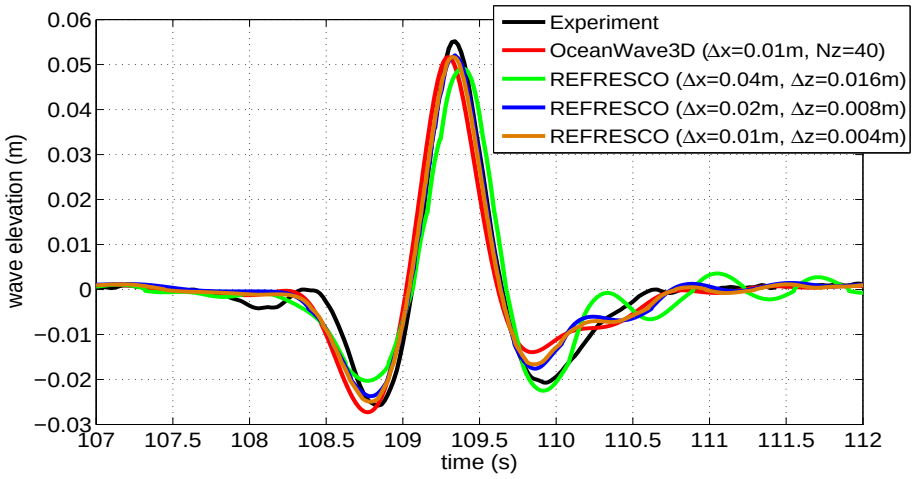
### 3.3.2 Large-amplitude focusing wave (Experiment run number:203001)

Similar to the previous test case, the wave probe W3 is the inflow boundary of ReFRESCO, where the quantities from OceanWave3D are imposed, see Tab. 2 for the locations of the wave probes. Three grid resolutions are considered with ReFRESCO: $(\Delta x = 0.04m,\ \Delta z = 0.024m)$, $(\Delta x = 0.02m,\ \Delta z = 0.012m)$, and $(\Delta x = 0.01m,\ \Delta z = 0.006m)$. Time step size is equal to 0.005s, and the simulation is carried out for 120s.

Similar to the test case 202002, the coarse grid seems inadequate as deviations from the experiment are relatively larger compared to the other two grids, see Fig. 12. It can also




(a) W4

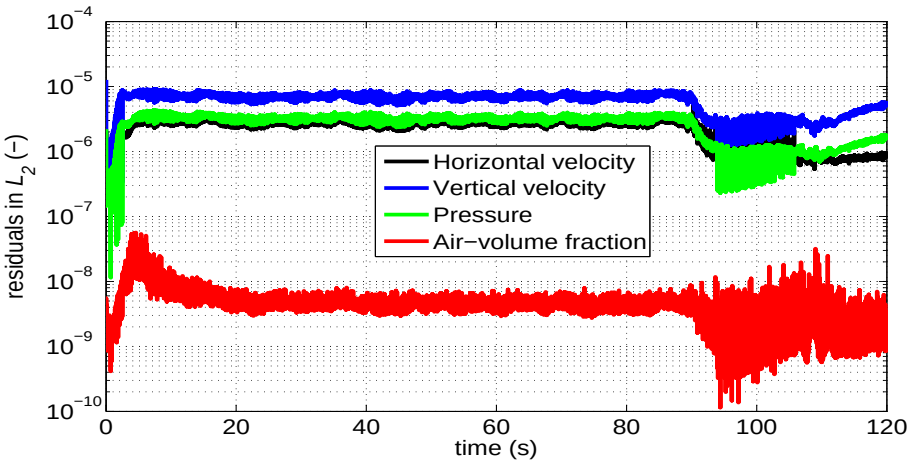



(b) W5

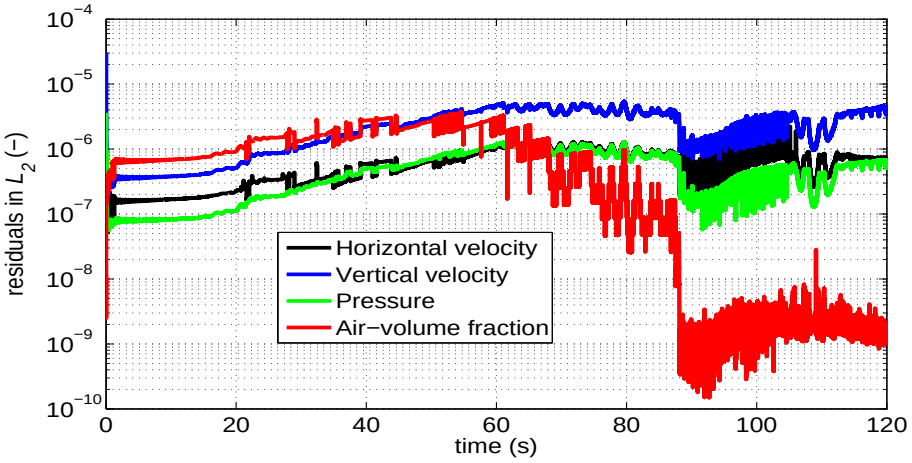
Figure 9: Comparison of the wave signals at the measurement locations W4 and W5 for the case 202002.

be observed that no further grid refinement is necessary as the results from the medium and fine grids are considerably close. The agreement between the experiment and ReFRESCO simulations on the medium and fine grids is acceptable.

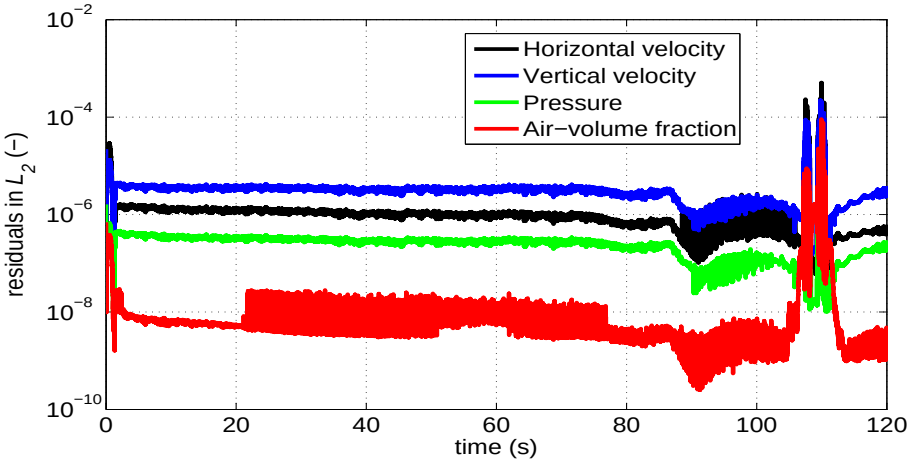
The residual history with time is plotted in Fig. 13 for the coarse grid with the resolution of $(\Delta x = 0.04m,\ \Delta z = 0.024m)$. On this grid we observe that nearly 5 orders of magnitude decrease in error is achieved for almost the entire simulation. However, for some time levels, we observe a temporary increase in residuals. Even though not shown here, this increase becomes more pronounced on the medium and fine grids, and the residuals decrease at approximately 3 order of magnitude on these two grids. These relatively high errors take place around the focusing event where high velocities are observed in the simulation. This results in a sharp increase in the CFL number which almost reaches the value of 6 on the fine grid, see Fig. 14. Even though the order of magnitude of the residuals can be acceptable for wave simulations, an investigation into possible modifications in the settings of the solvers for a better convergence behavior can be worthwhile.




(a) Coarse grid ($\Delta x = 0.04m, \Delta z = 0.016m$)




(b) Medium grid ($\Delta x = 0.02m, \Delta z = 0.008m$)




(c) Fine grid ($\Delta x = 0.01m, \Delta z = 0.004m$)

Figure 10: Residuals in $L_2$ on the three grid resolutions for the case 202002.

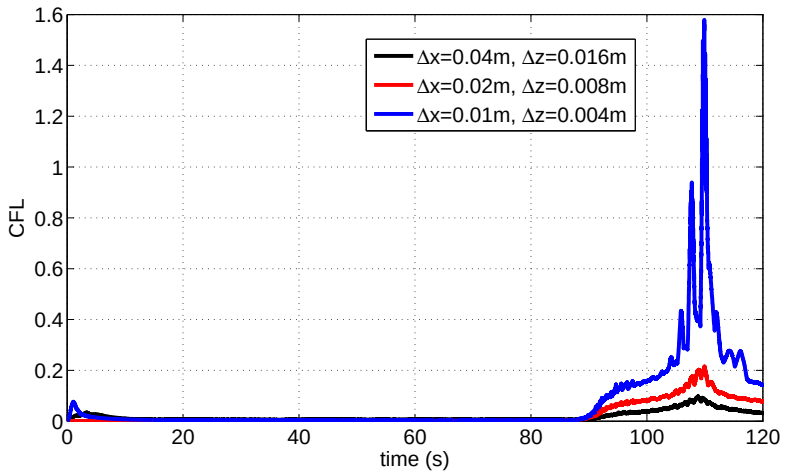



Figure 11: History of CFL number on the three grid resolutions for the case 202002.

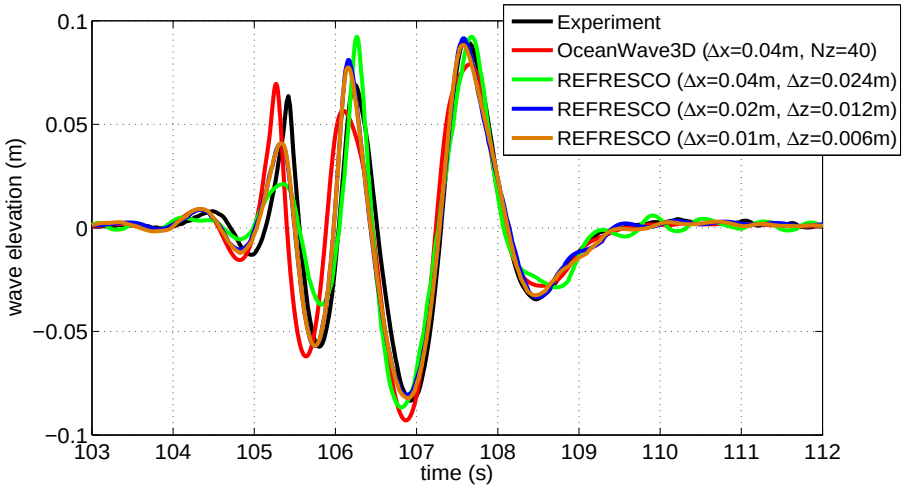
## 4 CONCLUSIONS

* In this study several objectives were set and realized. The first objective was to conduct an inventory study on the existing potential flow codes which can be used for simulating nonlinear waves. Then an overall assessment of these codes had to be made in order to explore their features and investigate their performances. The final objective was to use these codes in order to generate waves in ReFRESCO.

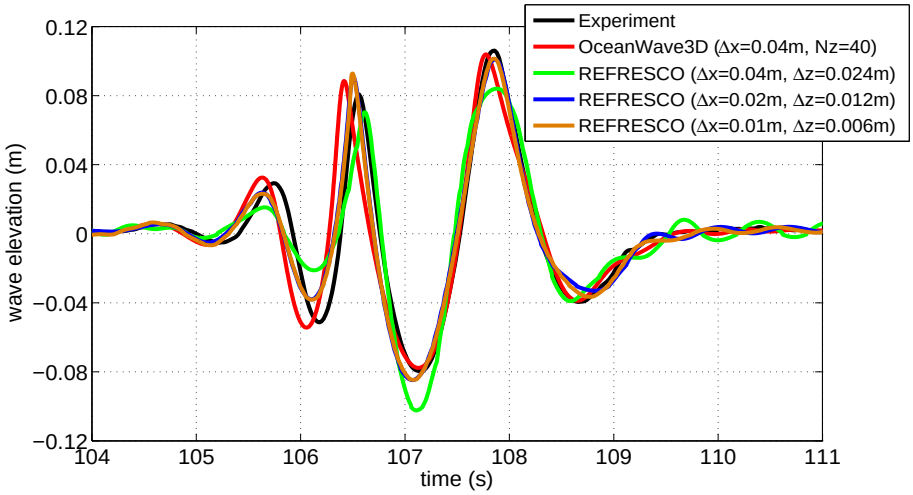
* After a survey on the existing models, four codes were selected: SWASH, OceanWave3D, HOS-NWT and HAWASSI. The functionalities and performances of these codes were analyzed. After careful consideration, OceanWave3D was selected for one-way coupling with ReFRESCO.

* Various simulations were carried out with the ReFRESCO\OceanWave3D coupling. Grid refinement study was conducted and convergence behavior of ReFRESCO was monitored. Results indicated that the ReFRESCO\OceanWave3D coupling is a valuable tool to investigate wave motion in an accurate and efficient manner. In all the simulations, comparison with the measurements demonstrated good agreement, and the convergence behavior of ReFRESCO was acceptable.

* Future work will include testing the ReFRESCO\OceanWave3D coupling in various 3D problems particularly focusing on hydrodynamic wave loading on structures and wave-structure interaction.




(a) W4




(b) W5

Figure 12: Comparison of the wave signals at the measurement locations W4 and W5 for the case 203001.

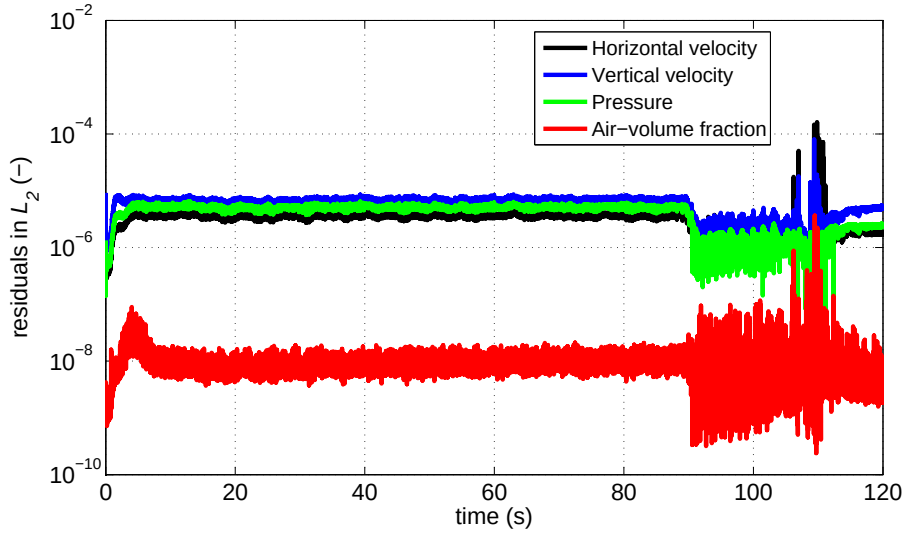



Figure 13: Residuals in $L_2$ on the coarse grid with the resolution of $(\Delta x = 0.04m, \Delta z = 0.024m)$ for the case 203001.

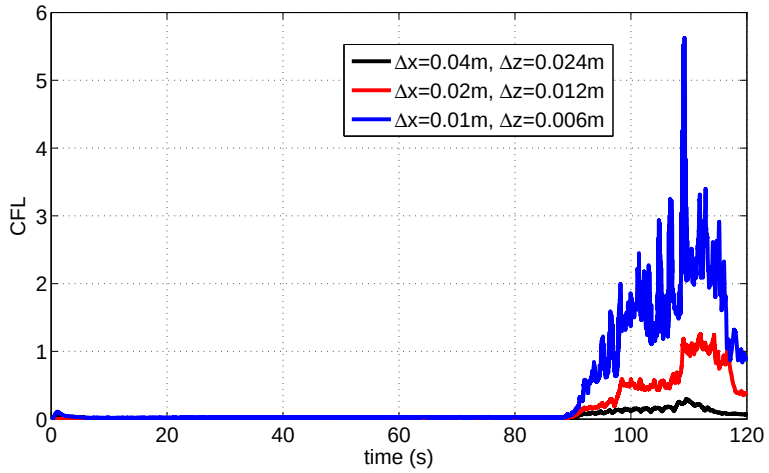



Figure 14: History of CFL number on three grid resolutions for the case 203001.

## REFERENCES


[1] Vaz, G., Jaouen, F., and Hoekstra, M., 2009. "Free-surface viscous flow computations. Validation of URANS code FRESCO". In Proc. ASME 2009 28th International Conference on Ocean, Offshore and Arctic Engineering OMAE2009, Vol. 5, paper OMAE2009-79398.

[2] Liam, L. S., Adytia, D., and van Groesen, E., 2014. "Embedded wave generation for dispersive surface wave models". *Ocean Engineering,* ***80***, pp. 73 – 83.

[3] Kurnia, R., and van Groesen, E., 2014. "High order Hamiltonian water wave models with wave-breaking mechanism". *Coastal Engineering,* ***93***, pp. 55 – 70.

[4] van Groesen, E., and Andonowati, 2011. "Fully dispersive dynamic models for surface water waves above varying bottom, Part 1: Model equations". *Wave Motion,* ***48***(7), pp. 658 – 667.

[5] van Groesen, E., and van der Kroon, I., 2012. "Fully dispersive dynamic models for surface water waves above varying bottom, Part 2: Hybrid spatial-spectral implementations". *Wave Motion,* ***49***(1), pp. 198 – 211.

[6] Stelling, G., and Zijlema, M., 2003. "An accurate and efficient finite-difference algorithm for non-hydrostatic free-surface flow with application to wave propagation". *International Journal for Numerical Methods in Fluids,* ***43***(1), pp. 1–23.

[7] Stelling, G. S., and Duinmeijer, S. P. A., 2003. "A staggered conservative scheme for every Froude number in rapidly varied shallow water flows". *International Journal for Numerical Methods in Fluids,* ***43***(12), pp. 1329–1354.

[8] Zijlema, M., and Stelling, G. S., 2005. "Further experiences with computing non-hydrostatic free-surface flows involving water waves". *International Journal for Numerical Methods in Fluids,* ***48***(2), pp. 169–197.

[9] Zijlema, M., and Stelling, G., 2008. "Efficient computation of surf zone waves using the nonlinear shallow water equations with non-hydrostatic pressure". *Coastal Engineering,* ***55***(10), pp. 780 – 790.

[10] Engsig-Karup, A., Bingham, H., and Lindberg, O., 2009. "An efficient flexible-order model for 3D nonlinear water waves". *Journal of Computational Physics,* ***228***(6), pp. 2100 – 2118.

[11] Ducrozet, G., Bonnefoy, F., Touzé, D. L., and Ferrant, P., 2012. "A modified High-Order Spectral method for wavemaker modeling in a numerical wave tank". *European Journal of Mechanics - B/Fluids,* ***34***, pp. 19 – 34.

[12] Bonnefoy, F., Ducrozet, G., Touzé, D. L., and Ferrant, P., 2009. "Time-domain simulation of nonlinear water waves using spectral methods". In *Advances in numerical simulation of nonlinear water waves*, Q. Ma, ed. World Scientific, ch. 11, pp. 129–164.

[13] SWASHteam, 2010. *SWASH User Manual*. Delft University of Technology.

[14] Düz, B., 2015. "Wave generation, propagation and absorption in CFD simulations of free surface flows". PhD thesis, Delft University of Technology, The Netherlands.

[15] Kerkvliet, M., Vaz, G., Carette, N., and Gunsing, M., 2014. "Analysis of U-type anti-roll tank using URANS. sensitivity and validation". In Proc. ASME 2014 33rd International Conference on Ocean, Offshore and Arctic Engineering OMAE2014, Vol. 2, paper OMAE2014-23483.

[16] Muzaferija, S., Peric, M., Sames, P., , and Schelin, T., 1998. "A two-fluid Navier-Stokes solver to simulate water entry". In Proc. 22nd Symposium on Naval Hydrodynamics.

[17] Ubbink, O., and Issa, R., 1999. "Method for capturing sharp fluid interfaces on arbitrary meshes". *Journal of Computational Physics,* ***153***, pp. 26–50.

[18] Paulsen, B. T., Bredmose, H., and Bingham, H. B., 2014. "An efficient domain decomposition strategy for wave loads on surface piercing circular cylinders". *Coastal Engineering,* ***86***, pp. 57 – 76.

[19] Sommerfeld, A., 1949. *Partial differential equations in physics*. Academic Press.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE OPTIMAL CONTROL OF A MULTI-MASS VIBRATION PROPULSION SYSTEM IN A VISCOUS INCOMPRESSIBLE FLUID

**Artem Nuriev**[1] **and Olga Zakharova**[2]

[1] Lobachevsky State University of Nizhni Novgorod
Russia, Nizhni Novgorod,
e-mail: nuriev_an@mail.ru

[2] Kazan Federal University
Russia, Kazan
e-mail: zakharovaos.mex@gmail.com

**Keywords:** Vibration propulsion system, viscous fluid, optimal control, numerical modeling, Navier-Stokes equations

**Abstract.** *The present work is devoted to the study of the optimal control of the two-mass vibration propulsion system in a viscous incompressible fluid. The study of the motion is carried out in two stages. At the first stage the simplified model of a viscous fluid is considered. On the basis of this model, the problem of the optimal control of the vibration system is solved in terms of minimizing energy consumption. The obtained optimal laws are studied at the second stage on the basis of the direct numerical simulation.*

## 1 INTRODUCTION

The design of the effective vibration propulsion systems is one of the key vectors of the development of the modern micro-robotics. Currently, there are several popular implementation concepts of such propulsion systems for moving micro-devices in a resistant medium. One of them is based on the use of the mobile internal mass. This type of devices usually called vibration-driven robots. They represent a multi-body system that are consisted of a closed shell, placed in a resistant medium, and moving internal parts. The periodic motion of the inner system of the bodies (the internal mass) causes motion of the shell, which can be used (in the case of a nonlinear medium resistance) for directional motion in space. The obvious advantages of such a propulsion are the shell tightness and lack of moving external parts, which allow to use such micro-devices for non-destructive inspection of miniature technical objects.

Research of movement capabilities of multi-body vibration propulsion systems were conducted earlier for medium with different resistance laws. In [1-3] the possibility of motion in an ideal fluid with related deformations of the outer shell was considered, articles [4, 5] were devoted to the study of the movement on a rough plane in the presence of a Coulomb friction, in [6] the movement on the liquid interface was studied. Movement in the Newtonian fluid was considered in [7-11].

Problems of optimal motion control of vibration multi-body systems in a viscous fluid were discussed in studies [9, 10]. In [9] the problem of optimization of the movement of the robot was solved in the presence of an arbitrary power law relationship between the velocity and the resistance forces, including the square law, which is often used as an approximation for the resistance forces resulting from the motion of a body in a Newtonian fluid. In [10] vibration-driven robot motion in a viscous fluid was optimized for the case, when the resistance law was designed on the basis of experimental data for a sphere moving in a viscous fluid.

However, all these problems of optimization were solved on the basis of quasi-stationary models, when the drag force is uniquely determined by the velocity of the body movement. In fact, the hydrodynamic resistance is defined by fluid flows that have been formed around the body during the whole time of motion. In general, the forces cannot be described solely in terms of instantaneous velocity and should be determined by the history of the entire movement. Developing the correct model of the drag description requires a large-scale study of the interaction of the vibration system with a viscous fluid.

In this paper, the research in this direction is carried out for the two-mass model of vibration-driven robot consisting of the outer spherical shell and a movable internal mass, which makes periodic oscillations along the axis passing through the center of the body.

As a first approximation, a simplified model of a viscous medium is considered, in which the hydrodynamic force is represented as the sum of the viscous and Basset history components. The viscous component is responsible for the viscous friction force. The history component is responsible for accounting of the movement history is selected in the form proposed in [12]. On the basis of this model the problem of optimal control of the vibration system is solved in terms of minimizing energy consumption.

The obtained optimal laws are studied in the second phase with on the basis of the direct numerical simulation. The complete problem of interaction of vibration system with a viscous incompressible fluid motion is considered, when the fluid flow is described by the full system of Navier-Stokes equations. Axisymmetric and three-dimensional formulations are discussed. Numerical model are implemented in the OpenFOAM package.

## 2 MODEL OF TWO-MASS VIBRATION SYSTEM

Let us consider a system of two bodies. The body of a spherical shape (shell) with mass $M$ is in a viscous liquid, and a shell with mass $m$ (hereinafter, the internal mass) moves inside it. The longitudinal periodic movements of the internal mass relative to the shell, in which the whole system moves as a whole, are investigated. Let us denote the body velocity through $u$ and the movement and velocity of the internal mass relative to the shell as $x$ and $v = \dot{x}$. The basic equation which describes the motion's velocity $u(t)$ of the shell under the given law $x(t)$ of the motion of the internal mass has the form

$$(m+M)\dot{u} + R(u) = -m\ddot{x} \tag{1}$$

Here, $R$ is the liquid resistance force to the shell's movement. For any given periodic law $x(t)$ with period $T$ (1) uniquely determines with the same period the periodic function $u(t)$, and $x(t)$ plays the role of the kinematic control.

Energy consumptions for the movement of the body with an internal propulsion (movable internal mass) is conveniently characterized by a the energy coefficient (EC)

$$\eta = \frac{N_0}{N_{vbr}}, \quad N_{vbr} = \min_{\langle R(u)\rangle=0, \langle u\rangle=U} \langle N(u)\rangle, \quad N_0 = \min_{\langle u\rangle=U} \langle N(u)\rangle$$

as the ratio of the minimum power $N_0$ required for movement of the body at an average velocity $U$ to the power $N_{vbr}$ consumed for moving at the same velocity of vibration-driven robot.

## 3 OPTIMIZATION PROBLEM

The formulation of the optimization problem consists of finding such a periodic law $x(t)$ of internal mass oscillations which for a fixed period $T$ of oscillations and given average velocity $U$ of the shell movement will minimize the power of the internal propulsion $\langle N(u)\rangle$. The convenience of this formulation is that the original problem is thus reduced to the problem of finding the periodic function $u(t)$ with period $T$ that provides a minimum for function $\langle N(u)\rangle$ within the constraints $\langle u\rangle = U$ and $\langle R\rangle = 0$.

As an approximation of the resistance law let us consider widely used in the case of high-frequency oscillations [16] dependence

$$R[u] = \frac{1}{2} C_x \pi \rho a^2 |u| u + 6\pi\rho\nu a^2 \int_{-\infty}^{t} \frac{du/d\tau}{\sqrt{\pi\nu(t-\tau)}} d\tau \tag{2}$$

The difference between the quasi-stationary approximation (used earlier in [10, 11]) and (2) consist in the addition Basset forces. The resistance is determined by not only the current value of the velocity, but also by the whole history of the motion. We consider the important special case $C_x = \text{const}$ of the quadratic viscous resistance.

By normalizing velocity $u$ on $U$ and time $t$ on period $T$, we write down the problem of the optimal control of the shell movements in the following form:

$$N_{\min} = \min\left(N_V[u] + sN_H[u]\right) \tag{3}$$

$$\langle u\rangle = 1 \tag{4}$$

$$\langle u|u|\rangle = 0 \tag{5}$$

$$N_V = \left\langle |u|^3 \right\rangle, \quad N_H = \langle uR_H \rangle, \quad R_H = \int_{-\infty}^{t} \frac{\dot{u}(\tau)d\tau}{\sqrt{(t-\tau)}}$$

The minimization in (3) is carried out on a set of periodic functions with a unit period that satisfy constraints (4) and (5). When writing (5), it is further taken into account that $\langle R_H \rangle = 0$ for any periodic function $u$. The only dimensionless parameter of problem (3)

$$s = \frac{12}{C_x U}\sqrt{\frac{\nu}{\pi T}} \tag{6}$$

sets the degree of the nonstationarity of the shell's motion by characterizing the ratio of Basset forces to viscous forces. If $s = 0$ the problem (3) - (5) of optimal control of body motion is quasi-stationary problem with the quadratic resistance law, the decision of which is described in [10].

The problem (3) - (5) was solved numerically with grid methods. The obtained optimal laws of motion are biphasic. They consist of a slow forward movement phase and a fast backward movement phase.

Figure 1 shows that the most significant change in the optimal motion law occurs in the range of $s$ from 0.1 to 1. If $s$ is less than 0.1, the law of motion is close to the quasistationary $u_0(t)$ obtained by neglecting the Basset forces. At $s$ larger than 3, in contrast, viscous friction forces can be ignored. Here $u(t;s)$ practically coincides with $u_\infty(t) = u(t;\infty)$.

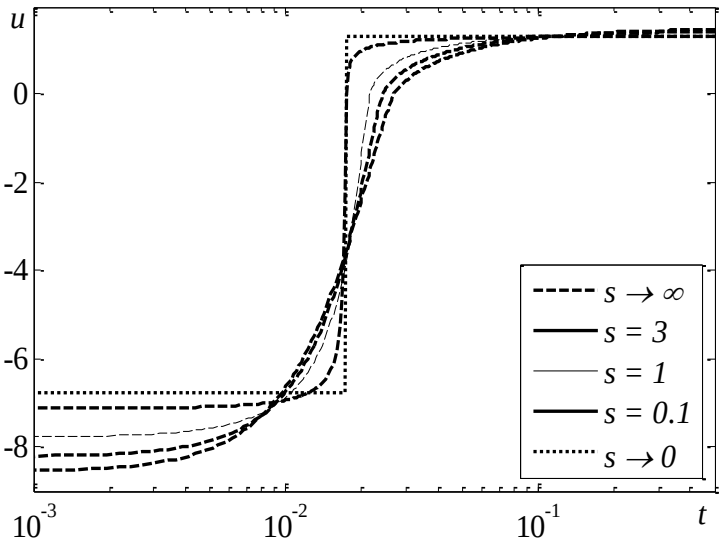



Figure 1. The optimal law of motion $u(t)$ for different values $s$.

In Figure 2 the solid line shows the dependence of the main integral characteristic, which is the energy coefficient $\eta_{\max}$, on parameter s. The dashed lines in this figure indicate $\eta_{\max}^{(0)} = 0.079$ and the asymptotic $\eta_{\max}(s) = \eta_\infty s^{-1}$ ($s \to \infty$). The value $\eta_\infty = 0.056$ is calculated by the power of Basset forces for $u_\infty(t)$ according to the formula $\eta_\infty = \left(N_H(u_\infty)\right)^{-1}$. As might be expected, the energy coefficient decreases monotonically with an increase of parameter $s$, which corresponds to additional power losses of the propulsion to overcome Basset forces.

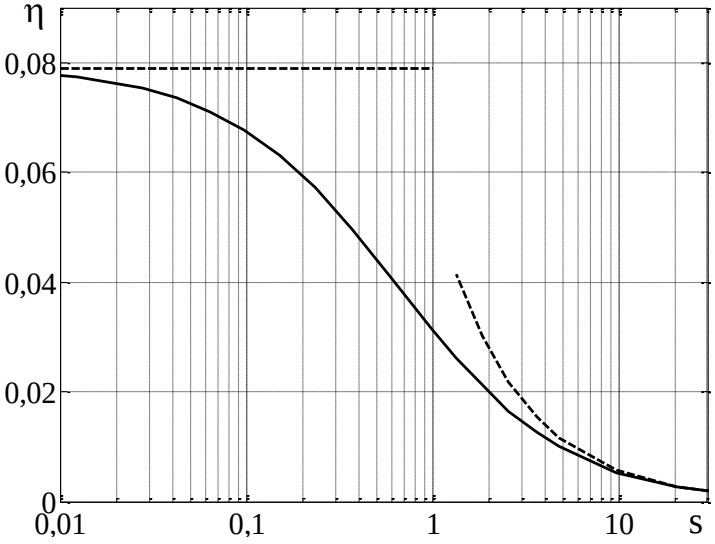



Figure 2. Dependence of the energy coefficient $\eta_{\max}$ on the parameter $s$ - a solid line (dashed line - asymptotic behavior at the $s \to 0$ and $s \to \infty$).

## 4 NUMERICAL MODELING

### 4.1 Problem description

Approximation of the hydrodynamic forces that was proposed in the previous section contains only the simplest representation of Basset history term. To develop more accurate model of the interaction of the vibration system with a viscous fluid a more detailed analysis of the components of the hydrodynamic force is required. For this purpose in the second part of the research a numerical simulation of the movement of a vibration-driven robot in a viscous fluid is carried out on the basis of the full Navier-Stokes equations.

Motion laws are selected from the results of the optimization problem presented in the previous section. The velocity of the shell is defined up to a constant, which is determined from the conditions of periodicity of the internal mass and the shell motion.

The resulting problem depends on two dimensionless parameters: the Reynolds number (Re) and the dimensionless period of the movement ($T_0$) that are defined as follows

$$\mathrm{Re} = \frac{U_m D}{\nu},\ T_0 = \frac{U_m T}{D}.$$

Here $U_m$ is the maximal absolute value of velocity of the shell, $D$ is the diameter .of the shell. The dimensionless parameters $(\mathrm{Re}, T_0)$ have the following interconnection with the parameter $s$ of the optimization problem

$$s = \frac{12}{C_x}\frac{U}{U_m}\sqrt{\frac{1}{\pi T_0 \,\mathrm{Re}}}.$$

A numerical model for the simulation of the interaction of the vibration system with a viscous fluid is constructed in OpenFOAM package on the basis of the numerical schemes presented in [13, 14].

### 4.2 Results

The numerical modeling was carried out in the range 300 <Re <2500 of Reynolds numbers for the fixed value of dimensionless period $T_0 = 40$. The law of the shell motion was chosen

on the basis of the data of analytical model, so that it provides a theoretically high efficiency in the whole study area. The corresponding dependence $u(t)$ is shown in Figure 3.

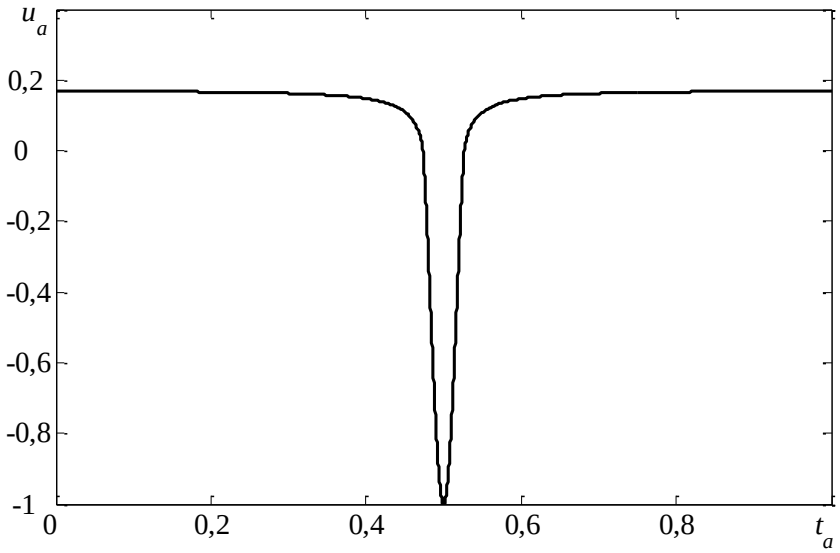



Figure 3. Motion law of the shell

Despite the high value of the maximum Reynolds number the flow structure around the vibration-driven robot retains axisymmetric in the investigated range of Re. This is due to the fact that the maximum speed of motion is reached in the time interval (see. Figure 3) that is too short for asymmetry development. At the same time, the velocity values of the direct motion phase are an order of magnitude smaller than the maximum, so all asymmetric disturbances are damped during this phase. In Figure 4 the main flow patterns observed in the investigated range of Re are presented. Visualization is made with colored weightless particles. In the first case (Figure 4, left), in the wake of the body there is irrotational fluid motion. Such a regime is observed in the range Re <2000. In the second case, the wake consists of gradually dissipating vortex rings (Figure 4, right).

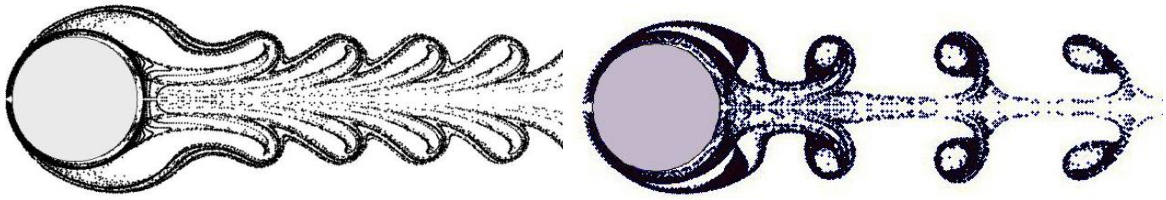


Figure 4. Flow patterns for parameter values $T_0 = 40$ , Re=500 (left) and $T_0 = 40$ , Re=2500 (right).

The efficiency of motion grows with increasing Reynolds number. Dynamics of change of the energy coefficient $\eta$ is shown in Figure 5. However, values of $\eta$ obtained in numerical simulation are more than two times lower than predicted by the theoretical model. To analyze the reasons for this discrepancy, the structure of the forces acting on the vibration-driven robot is studied.

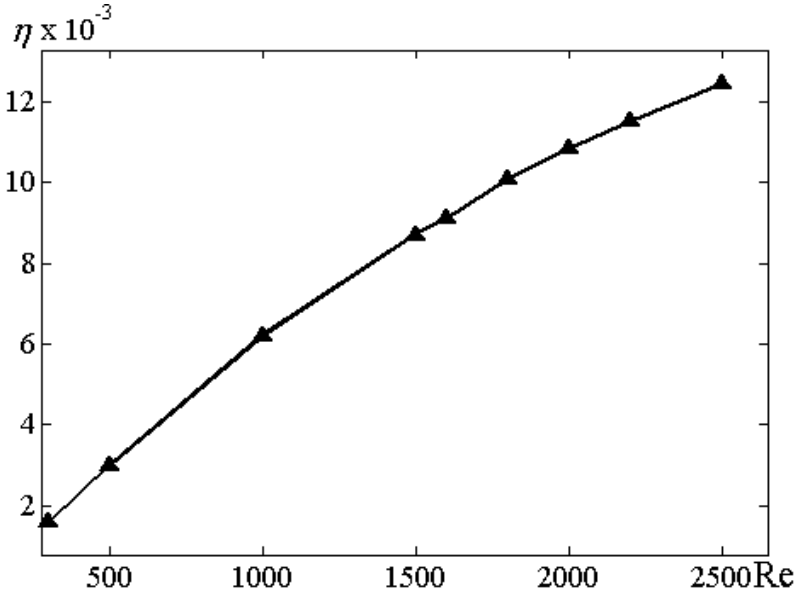



Figure 5. Dependence of the energy coefficient $\eta$ on the parameter Re.

The resulting force is represented as the decomposition into viscous, inertial and history components in the following form

$$R(C_D, C_m, C_b, t) = C_D u(t)|u(t)| + C_m a(t) + C_b \int_{-\infty}^{t} \frac{a(\tau)}{\sqrt{t-\tau}} d\tau \,. \tag{7}$$

To do such decomposition using the assumptions $C_D = \text{const}$, $C_m = \text{const}$, $C_b = \text{const}$, that were applied in optimization problem, is impossible. Since we are dealing with two-phase regimes of motion, it is possible to assume that the coefficient of viscous forces $C_D$ can take two values $C_{d+}$ if $u(t) > 0$ and $C_{d-}$ if $u(t) < 0$, and the coefficients of the inertia and Basset forces ($C_m$, $C_b$) are constants.

In order to determine the unknown coefficients the minimization problem is solved:

$$L(C_{d+}, C_{d-}, C_b, C_m) = \sqrt{\sum_{i=1}^{N} \left( R(C_{d+}, C_{d-}, C_b, C_m, t_i) - R_i^{\text{num}} \right)^2} \longrightarrow \min \,,$$

$$\langle R \rangle = 0 \,.$$

Here $L$ is a standard deviation function and $R^{\text{num}}$ is computed force. Finally we find the combination $(C_{d+}, C_{d-}, C_b, C_m)$ for each Reynolds number. The proposed relationship (7) makes it possible to construct a good approximation of force $R^{\text{num}}$ in the entire range.

Coefficient $C_m$ almost independent on the Reynolds number. The values of coefficient $C_b$ are quite small, so the third term of decomposition (3.21) has a very small influence. The values $C_{d+}$ and $C_{d-}$ change with increasing Reynolds number, they are shown on Figure 6. As we can see they are close to the experimental dependence [15] for the forces acting on a sphere in the stationary viscous flow.

The resulting force approximation differs from the one used in the section 3. This explains the differences in the estimates of the effectiveness of the resulting motion. The next step is to use it to solve the optimization problem. As the first results show, a new representation of forces does not introduce significant changes in the form of optimal laws of motion. That

probably will allow to achieve a good consistency in the results of analytical and numerical models.

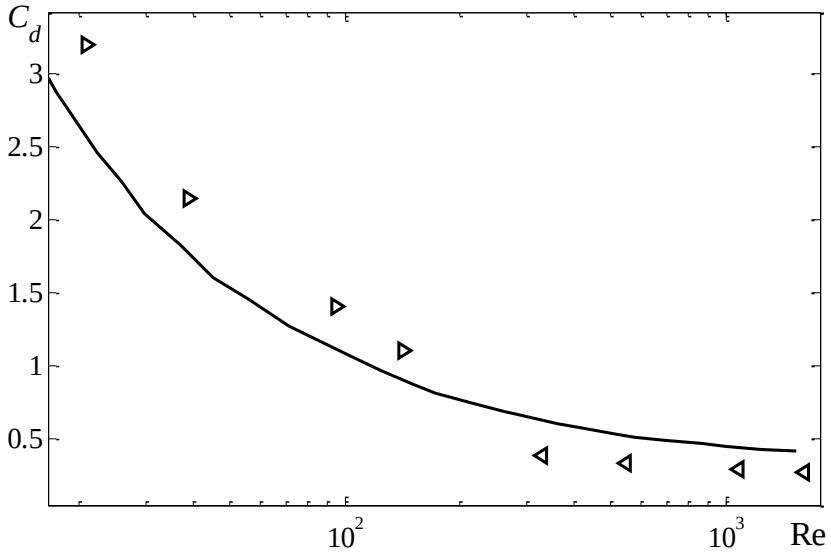



Figure 6 Resistant force. Markers are $C_{d+}$ and $C_{d-}$ values, solid line is the experimental dependence [15].

The reported study was partially supported by RSF, research project No. 15-19-10039. Numerical part of the research was carried out in the Nizhny Novgorod State University.

## REFERENCES

[1] Lighthill, M. J., On the Squirming Motion of Nearly Spherical Deformable Bodies through Liquids at Very Small Reynolds Numbers. *Comm. Pure Appl. Math*, **5(2)**, 109–118, 1952.

[2] Saffman, P. G., The Self-Propulsion of a Deformable Body in a Perfect Fluid. *J. Fluid Mech.*, **28(2)**, 385–389, 1967.

[3] Ramodanov, S. M., Tenenev, V. A., and Treschev, D. V., Self-propulsion of a Body with Rigid Surface and Variable Coefficient of Lift in a Perfect Fluid. *Regul. Chaotic Dyn.*, **17(6)**, 547–558, 2012.

[4] Chernous'ko F. L., On the motion of a body containing a movable internal mass. Dokl. *Phys.*, **50**, 593-597, 2005.

[5] Chernous'ko F. L., Analysis and optimization of the motion of a body controlled by means of a movable internal mass. *J. Appl. Math. Mech.,* **70**, 819-842, 2006.

[6] Volkova, L. Yu. and Jatsun, S. F., Control of the Three-Mass Robot Moving in the Liquid Environment. *Rus. J. Nonlin. Dyn.*, **7(4)**, 845–857, 2011.

[7] Childress, S., Spagnolie, S. E., and Tokieda, T., A Bug on a Raft: Recoil Locomotion in a Viscous Fluid. *J. Fluid Mech.,* **669**, 527–556, 2011.

[8] Auziņš, J., Beresņevičs, V., Kaktabulis, I., Kuļikovskis, G. Dynamics of Water Vehicle with Internal Vibrating Gyrodrive. *Vibration Problems ICOVP 2011. Supplement: The*

*10th International Conference on Vibration Problems*, Czech Republic, Prague, 5-8 September, 2011

[9] Vetchanin E. The Self-propulsion of a Body with Moving Internal Masses in a Viscous Fluid. *Regular and Chaotic Dynamics*, **18**, 100–117, 2013.

[10] Egorov A. G., Zakharova O. S. The energyoptimal motion of a vibrationdriven robot in a resistive medium. *J. Appl. Math. Mech*. **74**, 443, 2010.

[11] Egorov A. G., Zakharova O. S. Optimal quasistationar motion of vibrationdriven robot in a viscous liquid. *Izv. Vyssh. Uchebn. Zaved., Mat.*, **2**, 57–64, 2012.

[12] Mei R. Velocity fidelity of flow tracer particles. *Experiments in Fluids.* **22**, 13, 1996.

[13] Nuriev A.N., Zaytseva O.N. Solution to the problem of oscillatory motion of a cylinder in a viscous fluid in the OpenFOAM package. *Heald of Kazan Technological University* **8**, 116-123, 2013.

[14] Egorov A. G., Kamalutdinov A. M., Nuriev A. N., Paimushin V. N. Theoretical-experimental method for determining the parameters of damping based on the study of damped flexural vibrations of test specimens. 2. Aerodynamic Component of Damping *Mech. Compos. Mater*. **50**, 267-275, 2014

[15] Schlichting, *Boundary Layer Theory*, McGraw-Hill, 1979.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# HIGH-RESOLUTION SIMULATION OF INTERNAL WAVES ATTRACTORS AND IMPACT OF INTERACTION OF HIGH AMPLITUDE INTERNAL WAVES WITH WALLS ON DYNAMICS OF WAVES ATTRACTORS

**Ilias Sibgatullin**[1,2], **Michael Kalugin**[1]

[1]Institute of System Programming
Moscow, Solzhenitsina str., 25
e-mail: sibgat@ocean.ru

[2] Moscow State University, Faculty of Mechanics and Mathematics, Russia, 119991, Moscow, GSP-1,
1 Leninskiye Gory, Main Building
e-mail: sibgat@imec.msu.ru

**Keywords:** internal waves, attractors, wall interaction

**Abstract.**

*Internal wave attractors have received great attention since its discovery in 1995 by Leo Maas ([1, 2]). Now convectional theory describing the formation of attractors is generally accepted, and the principal interest of researchers in recent years is focused on nonlinear interactions. A number of theories had been proposed and now nonlinear interaction due to triadic resonance is accepted as a principle cause of instability of attractors, in which case the parent wave of large amplitude gives birth to two daughter waves, such that conditions of triadic resonance are fulfilled for all the three waves [3]. All this presumes that wave-wall interaction participate in the process only by focusing-defocusing of energy on the inclined boundaries. Here we address interactions of large amplitude internal waves with the boundaries in real laboratory conditions, where Prandtl-Schmidt number is equal to 700. We show that large amplitude waves produce folded structures which are clearly visible on density-gradient images, as if produced by "kneading the dough". These structures are not quickly dissipated due to high Schmidt number and with time they propagate to the interior of the domain, as was shown by our numerical simulation. These structures have visible impact on the instability of attractor and the whole picture of turbulent motion for large amplitudes. The mechanism of interaction of these structures with internal gravity waves structures is the subject of further research. Numerical simulation is quite challenging due to high Prandtl-Schmidt number and small scale of the folded structures in highly nonlinear regimes. Also study is performed on large time scales. As a consequence most of conventional computational approaches give unreliable results. We have applied spectral element approach base on code nek5000 by Paul Fischer. It allowed us to carefully follow the fine space structures on large time scales.*

## 1 INTRODUCTION

Global and local atmospheric motions continue to be pivotal subjects of research, since only deep understanding of such processes may give insights on climate change, pollutants concentrations, and forecasts, short-term as well long-term. In contrast to the atmosphere of the Earth, the driving force of the Ocean is not a classical heat engine. Vertical transfer mechanisms of energy due to thermal processes and wind over the surface play significant role mostly in the vicinity of the surface of the Ocean. Meanwhile the global dynamics of the Ocean is greatly affected by deep water processes and mixing [4, 5, 6, 7, 8], which were studied in much a lesser extent, experimentally and theoretically.

Internal waves in stratified media give an important class of energy transfer in the Ocean. They may form due to tidal motions or flows past orography [9]. In 1995 Leo Maas discovered an amazing feature of internal gravity waves: in some geometries the waves, emitted by source of constant frequency, may form after multiple reflections from boundaries certain pathways [1]. These pathways, which were named *internal waves attractors*, accumulate almost all the energy of the wave motions in the containers. The works by Leo Maas were followed by a great amount of works devoted to formations and behavior of the attractors, so now we have conventional theory for the appearance of small amplitude internal wave attractors.

This is why the principal focus of researches is now on the nonlinear properties of large amplitude internal wave attractors. For such cases the motion can not be described with the help of linearised equation. Internal wave attractors may become unstable, turbulent, they may change structure with time, change background stratification and manifest other nonlinear properties.

In recent years in ENS de Lyon a number of methods for analysis of experimental results have been developed ([10, 11, 12, 13, 3]). Despite remarkable success in experimental study of wave attractors there are significant intrinsic constraints in convectional experimental approaches like PIV or Schlieren. These constraints can be overcomed by direct numerical simulation, provided that it is reliable. The experimental setup possess very large scales interval because of very high Prandtl-Schmidt number. This is why till recently such numerical research were not possible, and numerical modelling dealt with linear modes only, as in ([14, 15]). In these works finite volume method was used for numerical simulation, and realisation of finite volume method was based on MIT code of general circulation model. These papers supported the theory of formation of internal wave attractors. At the same time the authors confessed that their numerical approach could only give insight on formation of the attractor but it is not reliable on large time intervals or in strongly nonlinear modes due to high numerical viscosity. Only recently careful 3D direct numerical simulation was carried out, and for the first time hydrodynamical fields of the experiments and direct numerical simulation were in correspondence within 10% for linear, as well for nonlinear modes ( [16, 17]). Now direct numerical simulation is planned to model more complicated configurations and highly nonlinear modes with overturning and mixing.

Internal waves in continuously stratified fluid differ from more “conventional” types of waves, such as electromagnetic or acoustic waves, by very peculiar dispersive relations: first – they are highly anisotropic, and second – temporal frequency depends only on direction (i.e. angle with vertical direction), and don’t depend on magnitude of the waves vector. Internal waves may seem unimportant from “practical point” of view since usually they are considered in linear approximation with quite small amplitude. But actually the may have influence on background stratification and on transport of admixtures and pollutants. The influence on background stratification may occur due to interaction with the boundaries, which is always nonlinear. If the amplitude of the waves is high enough, the overturning may happen. Prob-

ability of such events may be estimated with the help of Miles – Howard necessary condition. In recent years understating of internal waves dynamics was greatly improved due to numerous experimental laboratory works. Usually internal waves are modelled with the density stratification due to salt solution in a tank filled using double-bucket method. In 1995 [1] and 1997 [2], first theoretically, and next experimentally the internal waves attractors were discovered. Later numerous works were devoted to theoretical investigation, laboratory and numerical experiments. Most of them were devoted to very small amplitudes almost corresponding to linear equations. In internal waves of larger amplitudes vorticity becomes significant. Interaction with the boundaries becomes much more complicated. In this work we show that this interaction produce folded structures which may propagate inside the domain and interact with the main flow.

## 2 MATHEMATICAL SETUP

In mathematical setup we will try to model as close as possible the experimental setup [3]. The sketch of the experimental installation is shown in Figure 1.

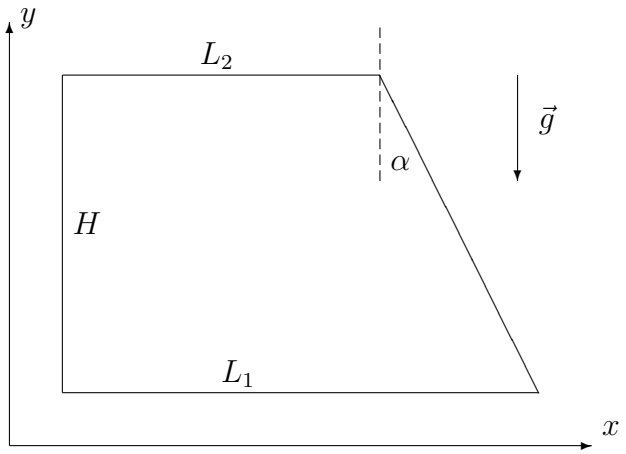



Figure 1: Computational domain in 2D case.

A salt solution is confined in the shown trapezoidal domain, all the boundaries are fixed except for the left one, which oscillates around mid-height point and has the shape of demi-cosine of vertical coordinate.

We will represent density as $\rho = \rho_m + \rho_s(t,x,y,z) = \rho_m + \rho_{s,0}(z) + \rho'(t,x,y,z)$, where $\rho_m$ – density with minimal salinity and $\rho_s = \rho_{s,0}(z) + \rho'_s(t,x,y,x)$ – density increase due to increase of salinity, $\rho_m + \rho_{s,0}(z)$ – initial state. It can be easily shown that such conditions can be modeled by Boussinesq approximation of full Navier – Stokes equations, so we will use the following system of equation for mass, impulse a salt conservation:

$$\frac{\partial \vec{v}}{\partial t} + v^k \nabla_k \vec{v} = -\nabla \frac{\tilde{p}}{\rho_0} + \nu \Delta \vec{v} + s\vec{g} \tag{1}$$

$$\frac{\partial s}{\partial t} + v^k \nabla_k s = \lambda \Delta s \tag{2}$$

$$\mathrm{div}\vec{v} = 0 \tag{3}$$

The most important assumption for Boussinesq approximation $\rho_s/\rho_m << 1$ is obviously satisfied.

Viscosity and diffusion are assumed to be constant.

Initially, the fluid is in rest, and concentration of salt decreases with height.

If fluid density decrease with height, fluid particle that does get displaced vertically tends to be restored to its original level; it may then overshoot inertially and oscillate about this level.

Let the unperturbed state of the layer have the density

$$\rho(x, y, z, t) = \rho_m + \rho_0(z), \tag{4}$$

where $\rho_m$ is the space average of the unperturbed state. Then the net gravitational force on a fluid particle after vertical displacement $\zeta$ is $-\frac{\mathrm{d}\rho_0}{\mathrm{d}z}\zeta\vec{\mathrm{g}}$, so

$$\frac{\mathrm{d}^2\zeta}{\mathrm{d}t^2} = \frac{\mathrm{d}\rho_0}{\mathrm{d}z}\frac{\mathrm{g}}{\rho_0}\zeta, \tag{5}$$

which is actually an equation of oscillator so there naturally appears a notion of buoyancy frequency, (or Brunt-Väisälä frequency), the angular frequency at which a vertically displaced parcel will oscillate within a statically stable environment $\rho(z) = \rho_m + \rho_0(z)$:

$$\mathrm{N(z)} = \sqrt{-\frac{\mathrm{g}}{\rho(z)}\frac{\mathrm{d}\rho(z)}{\mathrm{d}z}} \tag{6}$$

To understand the basic properties of the system we may linearise to see how it behaves when the amplitudes of oscillations are small.

In case of constant buoyancy frequency the linearized system of equations is:

$$\rho_m\frac{\partial\vec{\mathrm{v}}}{\partial t} = -\nabla\mathrm{p} + \rho'\vec{g} \tag{7}$$

$$\frac{\partial\rho'}{\partial t} + \mathrm{v_z}\frac{\mathrm{d}\rho_{s,0}(z)}{\mathrm{dz}} = 0 \tag{8}$$

$$\mathrm{div}(\vec{\mathrm{v}}) = 0 \tag{9}$$

$\rho = \rho_m + \rho_s(t, x, y, z) = \rho_m + \rho_{s,0}(z) + \rho'(t, x, y, z)$.

If we will concider wavelike solution, periodic in space and time

$$f = F\exp i(\omega t - (\vec{k}, \vec{r})) \tag{10}$$

$$\omega^2 = N^2\left(1 - k_z^2/k^2\right) = (N\left|\sin\theta\right|)^2, \tag{11}$$

where $\theta$ is the angle between the *vertical* axis and $\vec{k} = e_x k_x + e_y k_y + e_x k_z = e_x k\sin\theta + \ldots$ Waves exist for any value of the angular frequency from zero up to N.

Here we see two most important features of linear waves in weakly stratified fluid of constant bouyancy frequency: direction of propagation of waves is derermined by frequency of perturbations, and after reflection from rigid surface the modulus of angle with vertical is preserved.

A tank is filled with salty water using double-bucket method. The stratification is almost perfectly linear, except for the 2 cm. layer near the top of tank. The size of tank is 46 cm length, 32 cm. height, 17 cm. In our recent work [17] we examined 3D effects on the dynamics of attractor and we showed very good correspondence of experiments and 3D numerical simulation, and we explained discrepancies observed in previous 2D numerical simulation [14].

### 2.1 Boundary conditions

The upper boundary is stress free, the bottom and right boundary are rigid with no-slip condition, the left vertical boundary oscillate according to a given law.

$$x_b(0, y, t) = a\cos(\pi y/H)\cos(\omega_0 t). \tag{12}$$

If amplitudes are small this condition can be replaced by condition on velocity on fixed boundary:

$$u(0, y, t) = a\omega_0 cos(\pi y/H) sin(\omega_0 t), v = 0. \tag{13}$$

At $t = 0$ the fluid is in rest and stratified with constant buoyancy frequency $N$.

## 3 DIRECT NUMERICAL SIMULATION

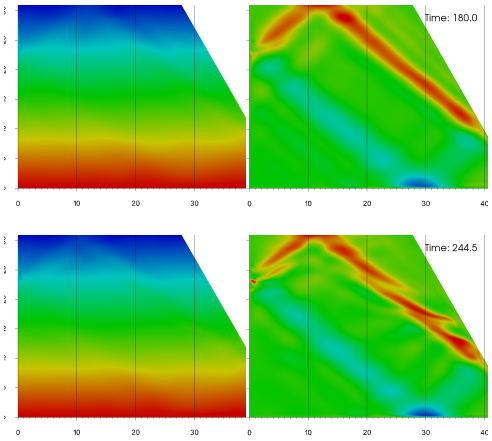



Figure 2: Results of DNS with moderate amplitude $a = 0.25cm$. Density field and horizontal component of velocity before and after PSI.

Numerical computations are performed with the help of spectral element methods [18, 19]. The geometry of the numerical setup closely reproduces the experimental one. The full system of equations being solved consists of the Navier-Stokes equation in the Boussinesq approximation, the continuity equation and the equation for the transport of salt. Typical meshes used in calculations consist of 50 thousands to half-million elements, with 8 to 10-order polynomial decomposition within each element, time discretization from $10^{-4}$ to $10^{-5}$ of the external forcing period.

As shown in [17], comparison of our experimental and numerical results are in beautiful agreement, not only qualitative but also quantitative.

Small amplitude behaviour is well predicted by linear theory [1, 2, 20]. The resulting motion is similar to the shown in first row of Figure 2: the internal wave attractor is formed. With higher amplitude we will get unstable wave attractor, which is shown in Figure 2: the first row shows the attractor just after it's formation, and the pictures in second row show the developing triadic resonance. At the same time no significant interaction with the walls, except for well predicted reflections, can be noted.

In Figure 3 you can see sequence of snapshots for twice higher amplitude $a = 0.5\mathrm{cm}$. Here you can see appearance of folding in the wall regions. Layers of high concentrations are first formed due impermeability of the walls and rotational motion of waves. Next they begin to propagate inside the domain, where they interact with the background flow.

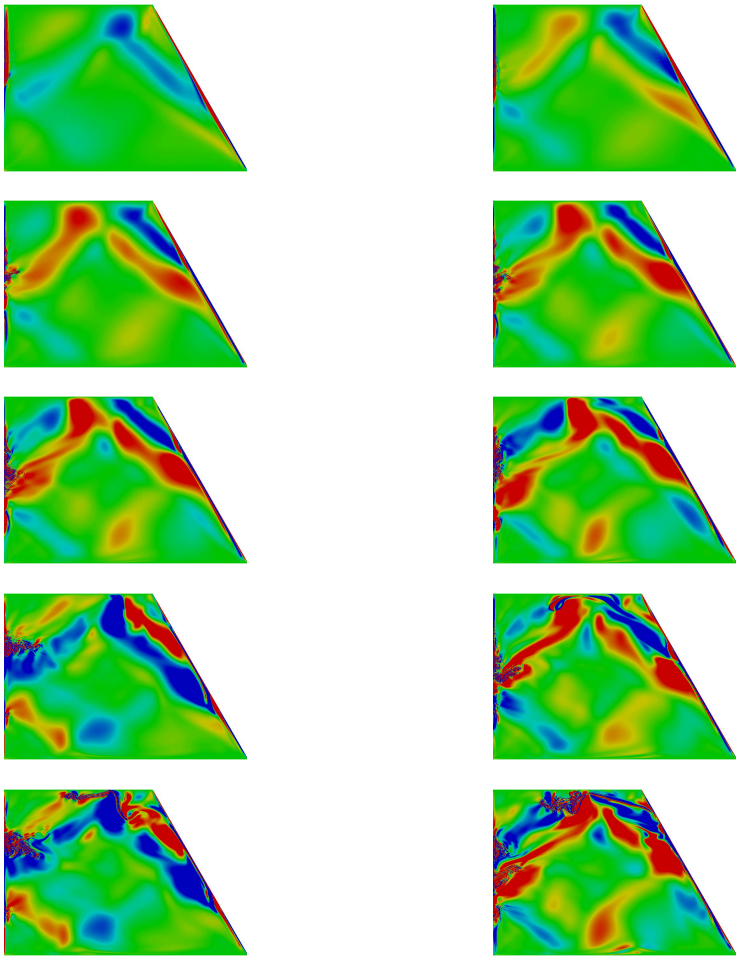


Figure 3: Results of DNS with large amplitude $a = 0.5cm$. Time-evolution of horizontal component of density gradient $\partial\rho/\partial x$

## 4 CONCLUSIONS

Direct numerical simulation of large amplitude internal wave attractors have shown the presence of folding structures in the wall regions of the container. They appear as a result of oscillating wave-like rotating motions of fluid near the boundaries. Since the boundaries are impermeable for salt and ratio of diffusion of motion and of concentration is very high (Sm=700), these folding structures are not quickly dissipated and propagate to the interior of the domain, where they interact with the background flow and influence mean stratification on long time periods. Since the precess is highly nonlinear and application conventional numerical methods is often criticized, careful experimental investigation is now needed fully describe this phenomena. Qualitative assessment of vertical transport and impact on background stratification is needed to describe interaction of the wave attractor and folded structures.

## 5 ACKNOWLEDGEMENTS

The work is supported by Russian ministry of education (RFMEFI60714X0090).

## REFERENCES

[1] L. R. M. Maas and F.-P. A. Lam. Geometric focusing of internal waves. *Journal of Fluid Mechanics*, 300:1–41, 1995.

[2] L. R. M. Maas, D. Benielli, J. Sommeria, and F.-P. A. Lam. Observation of an internal wave attractor in a confined, stably stratified fluid. Nature, 388:557–561, August 1997.

[3] H. Scolan, E. Ermanyuk, and T. Dauxois. Nonlinear Fate of Internal Wave Attractors. *Physical Review Letters*, 110(23):234501, June 2013.

[4] W. Munk. Abyssal recipes. *Deep-Sea Res.*, 13:707–730, 1966.

[5] W. Munk and C. Wunsch. Abyssal recipes II: energetics of tidal and wind mixing. *Deep-Sea Res.*, 45:1977–2010, 1998.

[6] D.J. Webb and N. Suginohara. Oceanography: Vertical mixing in the ocean. *Nature*, 409:37, 2001.

[7] J. Marshal and K. Speer. Closure of the meridional overturning circulation through Southern Ocean upwelling. *Nature Geoscience*, 5:171–180, 2012.

[8] M. Nikurashin and G. Vallis. A theory of the interhemispheric meridional overturning circulation and associated stratification. *J. Phys. Oceanogr.*, 42:1652–1667, 2012.

[9] C. Garrett and E. Kunze. Internal tide generation in deep ocean. *Annu. Rev. Fluid Mech.*, 39:57–87, 2007.

[10] M. Mathur, M. Mercier, T. Dauxois, and T. Peacock. New (Internal) Wave Generation - Laboratory Experiments. *ArXiv e-prints*, October 2009.

[11] M. Mercier, T. Peacock, and T. Dauxois. Experimental Study of Internal-Tide Scattering by 2D Topography. In D. N. Arabelos and C. C. Tscherning, editors, *EGU General Assembly Conference Abstracts*, volume 11 of *EGU General Assembly Conference Abstracts*, page 6213, April 2009.

[12] M. Mercier, N. Garnier, and T. Dauxois. Analyzing emission, reflection and diffraction of internal waves using the Hilbert transform. In *APS Division of Fluid Dynamics Meeting Abstracts*, page G2, November 2008.

[13] M. J. Mercier, N. B. Garnier, and T. Dauxois. Reflection and diffraction of internal waves analyzed with the Hilbert transform. *Physics of Fluids*, 20(8):086601, August 2008.

[14] N. Grisouard, C. Staquet, and I. Pairaud. Numerical simulation of a two-dimensional internal wave attractor. *Journal of Fluid Mechanics*, 614:1, October 2008.

[15] J. Hazewinkel, N. Grisouard, and S. B. Dalziel. Comparison of laboratory and numerically observed scalar fields of an internal wave attractor. *European Journal of Mechanics B Fluids*, 30:51–56, January 2011.

[16] C. Brouzet, S Joubaud, E. Ermanyuk, I Sibgatullin, and T. Dauxois. Energy cascade in internal wave attractors. *Europhysics Letters*, Accepted 2016.

[17] C. Brouzet, Sibgatullin, E.I Ermanyuk, H Scolan, and T. Dauxois. Internal wave attractors: experiments and numerical simulations. *Journal of Fluid Mechanics*, Accepted 2016.

[18] P. Fischer. An overlapping Schwarz method for spectral element solution of the incompressible Navier-Stokes equations. *Journal of Comp. Physics*, 133:84–101, 1997.

[19] P. Fischer, J. Lottes, D. Pointer, and A. Siegel. Petascale algorithms for reactor hydrodynamics. *Journal of Physics: Conference Series*, 125:012076, 2008.

[20] J. Hazewinkel, P. van Breevoort, S. Dalziel, and L.R.M. Maas. Observations on the wavenumber spectrum and evolution of an internal wave attractor. *J. Fluid Mech.*, 598:373–382, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF HEAT AND MASS TRANSFER PROCESSES IN LARGE-SCALE FLUIDIZED BED COMPLEX STRUCTURE APPARATUS AS AN EXAMPLE OF THE REACTOR OF ISOPARAFFINS DEHYDROGENATION

**Sergei S. Solovev**[1], **Svetlana R. Egorova**[2], **Alexander A. Lamberov**[2], **Olga V. Soloveva**[3]

[1] Kazan (Volga region) Federal Univ., Inst. of Mathematics and Mechanics
Kremlevskaya St. 35, 420008 Kazan, Russia
e-mail: serguei_s349@mail.ru

[2] Kazan (Volga region) Federal Univ., Inst. of Chemistry
Kremlevskaya St. 29, 420111 Kazan, Russia
segorova@rambler.ru, lamberov@list.ru

[3] Kazan State Power Engineering Univ., Inst. of Heat Power Engineering
Krasnoselskaya St. 51, 420066 Kazan, Russia
rara_avis86@mail.ru

**Keywords:** Fluidized bed, dehydrogenation of isoparaffins, industrial reactor, mathematical model, numerical simulation.

**Abstract.** *In the chemical industry are widely used fluidized bed apparatus. The advantage of them is the high speed of heat and mass transfer between components of the reaction, which are in different aggregation states. Studies of large-scale apparatus are hindered big sizes and plurality of structural elements. Often such apparatus operate at high temperatures (500-900 C), which further complicates the study. In this paper we consider a fluidized bed reactor block intended for the dehydrogenation of isobutane. In numerical simulation of fluidization was extended Eulerian-Eulerian approach. Differential equations that describe the hydrodynamic and thermal processes in the field of computational model of the reactor were solved in ANSYS Fluent CFD for axisymmetric unsteady flow scheme. At full simulation of the unit of the reactor in differential equations for the mass fraction of components of the gas mixture is necessary to consider changes related with chemical reactions. In the model used for this purpose it is necessary to add terms to the equations of mass transfer and absorption of heat depending primarily on the gas temperature and catalyst concentration. In this paper we'll restrict considering the minimum number of components of the reaction (raw materials – isobutane, product - isobutylene). For a given chemical reaction is written User Defined Functions (UDF). The influence of the ambient gas, the catalyst and the time step on the progress of chemical reaction in the volume element is studied. Numerical calculations were carried out, due to them circulating streams in the apparatus, the temperature field distribution of the catalyst and the conversion of the feeding gas-raw were analyzed.*

## 1 INTRODUCTION

In the chemical industry fluidized bed apparatus are widely used. The advantage is the high rate of heat and mass transfer between the reaction participants being in different aggregate states. It is known a large number of works devoted to the experimental results and constructing on their basis simplified models of fluidization and recommendations for the fluidized bed apparatus operations (e.g. [1-5]). Although such apparatus are used in the industry for over fifty years, interest in the study of the properties and possibilities of the fluidized bed use is not dried up due to the advent of new and modernization of existing apparatus and technologies.

In numerical simulation of fluidization Euler-Euler approach was extended, when the carrier (gas, liquid) and discrete (solid) phase are considered as continuous. To account for features of movement in a fluidized bed, by analogy with the kinetic theory of gases, is added the equation describing the change in the kinetic energy of the granules due to their collision (e.g. [3]). Ability of numerical implementation of the equations, describing the fluidized bed, allows calculations of different variations of the fluidization process, that are of interest both in terms of fundamental research and practical application.

Each structural element in particular the feed supplying device to the reactor contributes to the apparatus efficiency. Changes in the character of the movement and concentration of particles along the wall with deflectors are shown in [6]. In works [7, 8] by numerical simulation the effect of ring baffles located on the walls of the cylindrical apparatus on the hydrodynamics of the fluidized bed was analyzed. Experimental studies with internal elements in a liquid-solid fluidized bed are presented in [9]. Investigations of fluidized bed processes in the column with an inner cylinder were experimentally made in [10]. Modelling and comparative analysis of two large fluidized bed units with various feed gas supply devices is carried out in [11]. Investigations of [12] describe a numerical simulation of the fluidized bed and the kinetics process of olefins production from methanol. There are various mathematical models, for example, to the forces of resistance, viscosity and heat transfer, analysis of the applicability of the investigated parameters, for example, in [13, 14], performing numerical simulations.

A mathematical model and a numerical study of the chemical parameters of the fluidized bed reactor unit, intended for the dehydrogenation of isobutane to isobutylene, were made in this paper. A simplified model of the basic chemical reaction was constructed. The fields of the catalyst concentration, temperature fields and fields of dehydrogenation reaction product were obtained. The analysis of the circulation flows was carried out. The effect of slight structural modifications on the efficiency of the reaction product formation was carried out. The change of the circulating gas flow and catalyst due to the changes in apparatus construction was shown. The effect of fractional composition of the catalyst on the efficiency of the reaction product was demonstrated. It was obtained that formation of circulating flow is sensitive to the particle size selection.

## 2 PROBLEM FORMULATION

Chemical unit of the fluidized bed reactor intended for the dehydrogenation of isobutane is considered. Reactor has a cylindrical form and a height of the operating area is 11 meters and 5.2 meters in diameter. At the bottom of the reactor a feeder of gas feedstock is located. The load on the reactor unit feed is 20-40 tonnes per hour at a temperature of 550 ° C. Used gas feedstock is isobutane fraction of the following composition:

| Component | $H_2$ | i-$C_4H_{10}$ | n-$C_4H_{10}$ | i-$C_4H_8$ | n-$C_4H_8$ | $C_4H_6$ |
|---|---|---|---|---|---|---|
| Mass fraction,.% | 0.3-1.0 | 91-90 | 0.5-2.0 | 5.0-7.2 | 0.2-1.3 | 0.01-0.05 |

Table 1: Fractional composition of the gas feedstock.

There is aluminum chrome microspheroidal catalyst in the reactor with the particle diameters of 20-200 microns. A vertical pipe feeding regenerated catalyst at a temperature of 650 °C is along the axis of the reactor. The pipe length is of 7 m. The process of catalyst circulation from the reactor to the regenerator and back is continuous and the volume flow rate of the circulating catalyst is 60-100 tons per hour. Channel for discharging the catalyst into the regenerator is at the bottom part of the reactor. In the middle of the reactor in the working area angular falling type lattices with a free cross section about 30% are located. In the upper part of the reactor 6 pairs of cyclones are arranged for separation of exhaust gas from the reactor and returning the catalyst particles getting into the working zone of the reactor. Figure 1 demonstrates a scheme of meridian section of the reactor unit.

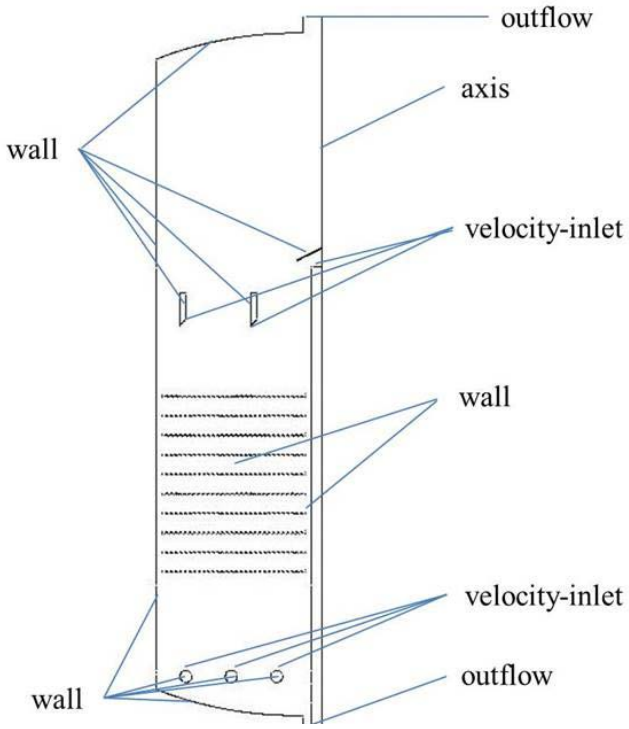



Figure 1: Scheme of meridian section of the reactor unit.

The main chemical reaction in the block is the dehydrogenation of isobutane, accompanied by the hydrogen evolution and heat absorption:

$$C_4H_{10} \leftrightarrow C_4H_8 + H_2 + \Delta H. \quad (1)$$

Dehydrogenation of $C_3$-$C_5$-paraffins is thermodynamically feasible only at temperatures 300-900 °C [15]. With non-localized energy supply to the hydrocarbon molecule the conversion will follow the path of cracking. Therefore, the dehydrogenation of paraffins can be carry out practically only in the presence of catalysts, that selectively increase the rate of dehydration by changing its activation energy, but not affectng the rate of reaction chain breaking. In

industrial processes the dehydrogenation of C3-C5 paraffins is carried out at temperatures of 520-620 °C and at a pressure close to atmospheric pressure or under vacuum [15]. At a temperature of 550-580 °C in the dehydrogenation of isobutane to aluminum chrome catalyst dehydrogenation reaction yield product (isobutene) is 35-50 % of mass fraction, cracking products is $\sum(C_1\text{-}C_3)$ = 3-7 % of mass fraction, isomerization (n-butylene ) = 0.2-0.5% of mass fraction, the oligomerization product yield (divinyl) is not more than 0.2% of mass fraction. The amount of coke formed on the catalyst is 0.2-2% of mass fraction. To remove high molecular hydrocarbon deposits use periodic catalyst regeneration by burning out the coke from the catalyst surface with fuel gas and air at a temperature of 650-670 °C.

## 3 MATHEMATICAL MODEL AND NUMERICAL SIMULATION

Considering chemical reactor unit consists of a large number of structural elements. The constructing of a three-dimensional mathematical model with all structural parts would lead to a great cost of calculation time. The cylindrical form of unit, the central pipe and circular arrangement of gas supply nozzles and cyclones allows the use of the model of axial symmetry. It is considered that an arbitrary meridional cross-section of the unit fully describes the processes taking place in the reactor. The geometry used in calculations corresponds to the figure 1. The axisymmetric model assumes that the reactor model is constructed by the geometry (shown in the figure) rotating around the axis of symmetry. The geometrical sizes of the model correspond to the geometric sizes of the test reactor unit.

### 3.1 Fluidized bed model

In the numerical simulation of a fluidized bed in a chemical reactor the continuous multiphase model was used, supplemented by kinetic theory of gases for the accounting the collisions of solid particles. At the same time several phases may be considered: gas, liquid or solid in any combination. For each of the phases holds the laws of mass conservation, momentum and energy.

The equation of mass conservation for the *i*-th phase is:

$$\frac{\partial \alpha_i \rho_i}{\partial t} + \nabla \cdot (\alpha_i \rho_i \vec{v}_i) = 0, \tag{2}$$

where for the *i*-th phase: $\alpha_i$ is volume fraction, $\rho_i$ is real density, $\vec{v}_i$ is velocity.

The equation of momentum conservation for the *i*-th phase can be expressed as:

$$\frac{\partial \alpha_i \rho_i \vec{v}_i}{\partial t} + \nabla \cdot (\alpha_i \rho_i \vec{v}_i \vec{v}_i) = -\alpha_i \nabla p + \nabla \cdot \overline{\overline{\tau}}_i + \alpha_i \rho_i \vec{g} + \sum_j \vec{R}_{ij}, \tag{3}$$

where $p$ is pressure, $\overline{\overline{\tau}}_i$ is stress tensor in the *i*-th phase, $\vec{R}_{ij}$ is the interaction force between the *i*-th and *j*-th phases. In equation (3) the stress tensor is written as:

$$\overline{\overline{\tau}}_i = \alpha_i \mu_i \left( \nabla \vec{v}_i + \nabla \vec{v}_i^T \right) + \alpha_i \left( \lambda_i - \frac{2}{3} \mu_i \right) \nabla \cdot \vec{v}_i \overline{\overline{I}}, \tag{4}$$

where $\mu_i$, $\lambda_i$ are shear and bulk viscosity, $\overline{\overline{I}}$ is unite tensor.

Interfacial interaction force is written as

$$\vec{R}_{ij} = K_{ij} (\vec{v}_i - \vec{v}_j), \tag{5}$$

where $K_{ij}$ is interfacial interaction coefficient.

The equation of energy conservation for the *i*-th phase is:

$$\frac{\partial \alpha_i \rho_i h_i}{\partial t} + \nabla \cdot (\alpha_i \rho_i \vec{v}_i h_i) = \alpha_i \frac{\partial p_i}{\partial t} + \overline{\overline{\tau}}_i : \vec{v}_i + \sum_j Q_{ij}, \tag{6}$$

where $h_i$ is enthalpy of the *i*-th phase, $Q_{ij}$ is heat transfer rate between *i*-th and *j*-th phases.

At movement of microspheroidal solid phase, consisting of granules, particles collide. As a result of the collision changes the kinetic energy of the granules. To account for the interaction of solids particles is used equation with the change of energy through a change in temperature of the particles [26]:

$$\frac{3}{2}\left[\frac{\partial \alpha_i \rho_i \theta_i}{\partial t} + \nabla \cdot (\alpha_i \rho_i \vec{v}_i \theta_i)\right] = \left(-p_i \overline{\overline{I}} + \overline{\overline{\tau}}_i\right) : \nabla \vec{v}_i + \nabla \cdot \left(k_{\theta i} \nabla \theta_i\right) - \gamma_{\theta i} + \phi_{ij}, \tag{7}$$

where $p_i$ is granules pressure of the *i*-th solid phase, $k_{\theta i}$ is diffusion coefficient of granules energy, $\theta_i$ is granules temperature of the *i*-th solid phase, $\gamma_{\theta i}$ is dissipation of energy because of particles collision, $\phi_{ij}$ is energy exchange between the *i*-th solid phase and the *j*-th liquid (gaseous) phase.

Written system of equations is not closed. To close the system use relationships obtained from experimental studies of the fluidized bed.

In fluidized bed models the force of interaction between phases is written in the form (5), where the coefficient of interfacial interaction $K_{ij}$ obtained in experimental studies, depending on types of the interacting phases. It is considered, that $K_{ij} = K_{ji}$. In the case of the interaction of two phases liquid or gaseous type the resistance is determined by the Schiller and Naumann model [17]. In the case of interaction between two phases, that are liquid (gaseous) phase (the *i*-th phase) and the granular solid phase (*j*-th phase) in (5) used Syamlal and O'Brien model [18]. In the case of interaction between two solid granular phases in equation (5) interfacial interaction coefficient is taken according Syamlal O'Brien model [19].

The term $\gamma_{\theta i}$ in equation (7) to the temperature of solid phase particles granules characterize the rate of energy dissipation in the *i*-th solid phase caused by collisions between phase particles. Energy dissipation due to the particles collision is written in the expression, obtained in Lun et al. model [20]. Coefficient of heat transfer intensity between *i*-th and *j*-th phase is determined for the interaction of the two liquid phases by expression of Ranz and Marshall model [21, 22]. In the case of granular solids for Nusselt number calculation at the heat exchange between the *i*-th solid phase and *j*-th liquid (gaseous) phase use an expression from the work of Gunn model [23]. For the heat exchange intensity balance relations $Q_{ij} = -Q_{ji}$ and $Q_{ii} = 0$ are performed.

The motion considered in the reactor at the accepted loads of gas feedstock and parameters of the used catalyst was turbulent. In the model allowed to calculate solutions used $k - \varepsilon$ turbulence model in which the motions of "secondary" granular solid phases were generated on the background of the turbulent motion of "primary" gas phase.

### 3.2 Chemical reaction model

Let us consider the basic chemical reaction only. It is described by the relation (1). The equation for the reaction product of can be written as:

$$\frac{\partial m_{prod}}{\partial t} = V_{cat} \cdot \rho_{feedstock} \cdot f(T_{feedstock}), \tag{8}$$

where $m_{prod}$ is mass of the dehydrogenation (isobutylene) reaction product, $V_{cat}$ is volume fraction of catalyst, $\rho_{prod}$ is the product (isobutylene) density, $f(T_{feedstock})$ is function that determines the conversion depending on the feedstock gas temperature.

The function $f(T_{feedstock})$ is based on the approximation of the experimental data and is:

$$f(T_{feedstock}) = 734.9314 - \frac{19702.791}{\sqrt{T_{feedstock}}}. \tag{9}$$

The dependence of the functions in the range of the temperature change from 550 to 650 °C is presented in Figure 2.

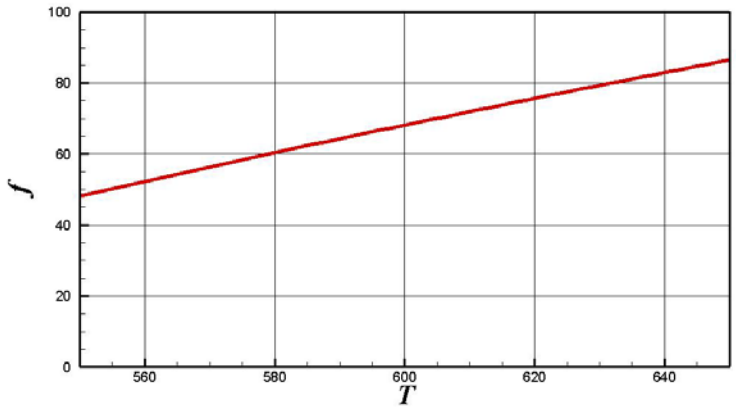



Figure 2: Function of conversion depending on the feedstock gas temperature.

### 3.3 CFD solver

Differential equations that describe the hydrodynamic and thermal processes in the field of computational model of the reactor were solved in CFD ANSYS Fluent package for axisymmetric unsteady flow scheme. The entire computational domain, shown in Figure 1, was divided into elements of a triangular or rectangular shape in different subregions of the computational domain, the dimensions of which were sufficient to determine the characteristic of the investigated phenomenon factors. For specified initial distribution of the catalyst in the reactor, after a while the flow entered a quasi-stationary mode. Hydrodynamic and thermal characteristic pictures of the reactor work calculated on this mode. Mass transfer due to the chemical reaction is implemented as a plug-in user-defined function (UDF). Time to quasi-stationary mode of one calculation depended on the conditions of the computational experiment and the average value of the order of 60 seconds of real time reactor's unit operation.

In each of the selected items are solved described above equations of the mathematical model of fluidized bed. Accordingly, the partition into a greater number of elements leads to a better accuracy of the calculations, but also leads to increase in the required computer resources. In this work, in calculations carried out the average number of finite elements was 500 thousand. Partitioning into finite elements near the gas feedstock supply models or arrays of nozzles built small enough to maximally correct reflect features of hydrodynamics and heat

and mass transfer relatively to the processes taking place in the operating chemical reactor. The average size of discrete element faces is 5 mm.

Boundary conditions were set on all elements of each of the constructed model of the reactor unit in accordance with the operating mechanism and used the solver. As the reactor model assumes rotationally symmetric, so on the axis of the constructed model was set conditions of axial symmetry. On all impermeable surfaces were conditions as walls. On models of gas feedstock supply nozzles and the outlet of the centrally located pipe of catalyst feed used velocity-inlet conditions (the value of the flow rate). Velocities were set according to the balance of the gas and catalyst flow rate. Temperatures of gas supplied and the catalyst were placed at 550 °C and 650 °C respectively. Outflow conditions (free output stream) were set in the area at the top of the reactor model for a gas outlet and at the bottom zone of the reactor to simulate the flow of catalyst into the regenerator. Also, through the part of free outlet stream at the top of the reactor there is a remove of fine catalyst particles. In the operating reactor particles from the top part return back into the working area over lattices with the help of cyclones. In constructed model on the output channels of cyclones were set velocity-inlet conditions in accordance with a flow rate of catalyst through the outlet channel in the upper part of the reactor.

## 4 RESULTS AND DISCUSSION

### 4.1 Results for base case of reactor

First were performed calculations for monodisperse catalyst. Particles size was selected of 100 microns. Operating time of the reactor unit is 90 seconds. Further calculation for two-fractional catalyst was carried out: a coarse fraction is 100 microns (80%), fine fraction is 50 microns (20%). Figure 3 shows the field of catalyst concentration and volume fraction of the reaction product of: ***a*** is for monodispersed catalyst, ***b*** is for polydisperse catalyst. The blue color corresponds to the absence of the phase under consideration.

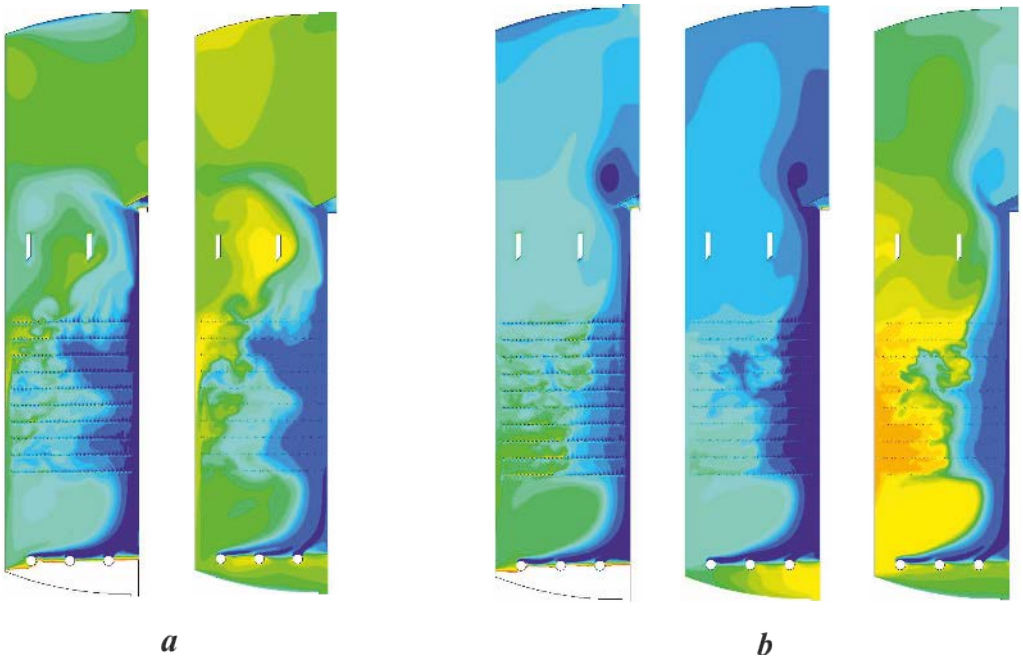



Figure 3: Calculation results.

It is evident that product formation corresponds to the catalyst concentration. In both cases there is an area with a low concentration of catalyst and product around the central tube. Detailed analysis has shown that we can see a downflow along pipe of incoming catalyst at a rate of about 2.5 m/s. In general, the presence of fine fraction contributes to a more intensive circulation and more uniform distribution of the catalyst in the unit. The volume of the reaction product located in unit in the case of monodisperse catalyst is 40% and for the case of polydisperse catalyst is 44%.

### 4.2 Reactor optimization variants

Two versions of a slight unit modification have been proposed to improve performance. The aim was to redirect flow downward along the tube of incoming hot catalyst into the central zone of the reactor. Following this purpose two variants of modifications were suggested.
a) The dispenser on pipe outlet, which should increase penetration of the catalyst into the heated reactor zone midway between the central pipe and the outer wall. Considered a deflector on the end of catalyst feed pipe.
b) Deflectors along the tube in the lattices zone provide catalyst redirection in the inter-lattice space. Number - between all the lattices. Slope - 45 degrees. Length - so that the edge of the deflector cover the edge of the lattice. Position - above the lattice, with a small gap between deflector and the lattice.

Figure 4 shows volume fraction of the reaction product fields for monodisperse and polydisperse two-fractional catalyst: ***a*** is the base of the reactor configuration; ***b*** is the presence of the deflector at the end of the catalyst feed pipe; ***c*** is the case with deflectors along the central tube.

Table 2 shows calculation results for volume fraction of the reaction product in the unit without modifications and with two variants of slight modifications.

| Calculation variant | Basic variant | | Modification 1 | | Modification 2 | |
|---|---|---|---|---|---|---|
| | mono-disperse | poly-disperse | mono-disperse | poly-disperse | mono-disperse | poly-disperse |
| Volume fraction, % | 40 | 44 | 49 | 42 | 36 | 41 |

Table 2: Volume fraction of the reaction product for different variants of the reactor constructions.

It is shown from Figure 4 and Table 2 that the presence of fine particles, zone of reaction product formation and the volume fraction are just about the same for all three cases. This is due to the fact that the presence of fine phase determines circulating flows in the unit, pushing and pulling the larger particles. Since the fine particles are moving fast on the carrier gas background phase, then they are less sensitive to small structural modifications.

A completely different picture is observed while considering in numerical simulation only one averaged fraction of coarse particles. The coarse particles have more inertia and respond to the geometrical features of the reactor unit. In original variant of the apparatus one global circulation flow on the whole volume of the unit is formed. The use of deflectors on the tube (modification 2) forms a main circulating flow in the upper part of the apparatus and a few smaller on the bottom. Such modification degrades the production at the bottom unit, and decreases overall efficiency. Throw the coarse particles in the middle zone of the reactor (modification 1) showed the best result because of breaking a single global circulation flow.

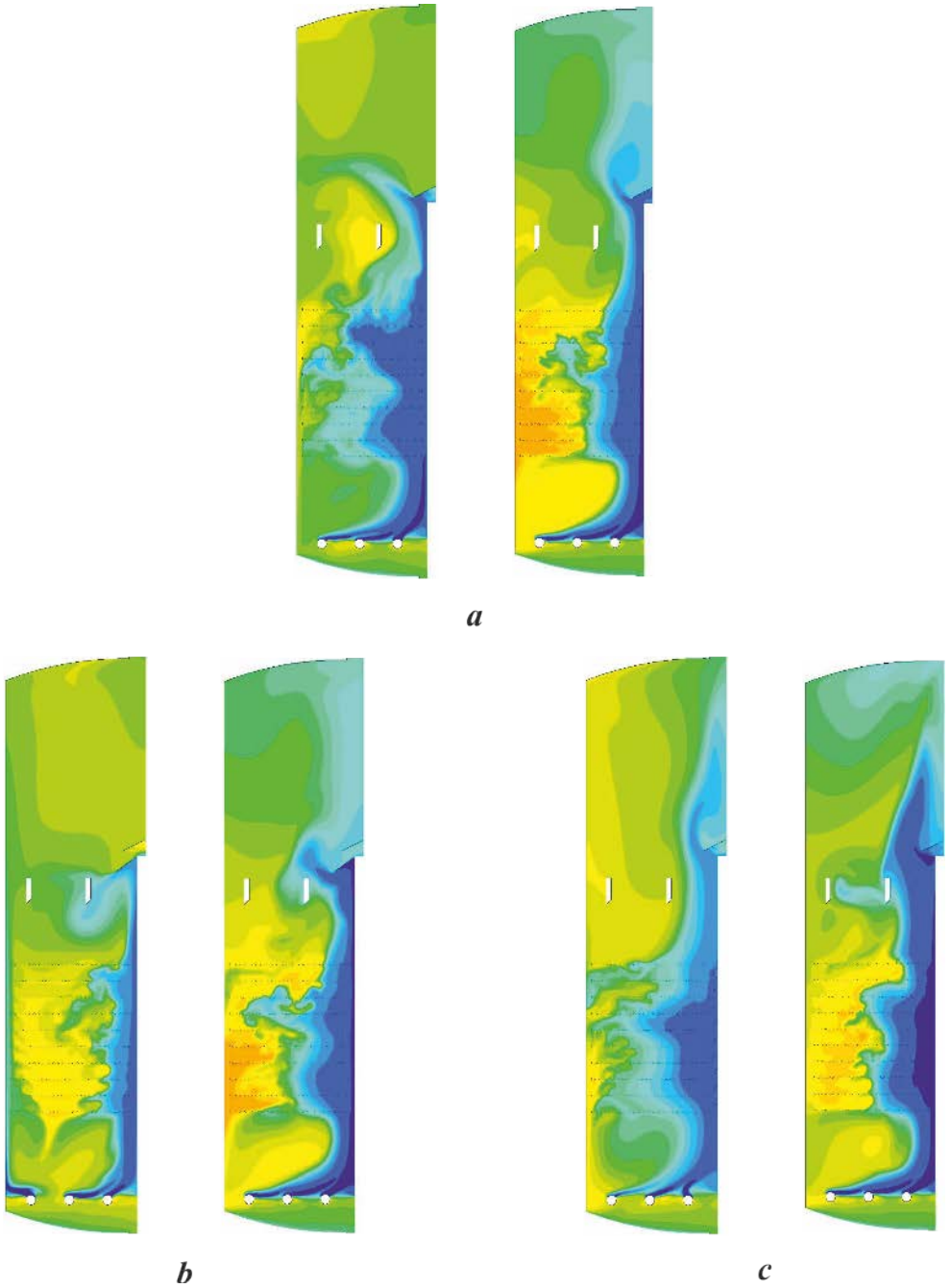



Figure 4: The results of calculations for units with modifications.

Calculations in this work were made for the averaged coarse fraction of catalyst of 100 microns, with an average speed of the cross section of the apparatus of 0.3 m / s. Note that the sensitivity of circulating flows formation of the reaction product does not depend only on the apparatus geometrical characteristics, but also on the fraction of the catalyst composition. The particle size of the main coarse fraction at specified gas flow rate will also determine the apparatus efficiency.

## 5 CONCLUSIONS

- A model of large-scale industrial fluidized bed reactor is constructed. The mathematical model is based on the Euler approach in describing discrete phase of particles motion as a continuous. Numerical implementation was carried out using Ansys Fluent software.
- A simplified chemical reaction model of isobutane to isobutylene dehydrogenation is constructed. The model is implemented and connected to a user-defined function (UDF).
- The fields of catalyst concentration, temperature field and the field of dehydrogenation reaction product were obtained. The analysis of circulation flows was carried out.
- The effect of slight structural modifications on the efficiency of product formation is investigated. The change of circulating gas flows and catalyst due to the change in apparatus construction is shown.
- The analysis of catalyst fractional composition effect on the efficiency of the reaction product was carried out. The sensitivity to the formation of circulating flow on the range of particle size is shown.

## REFERENCES

[1] J.F. Davidson, *Fluidization*. Academic Press, 1971.

[2] D. Kunii, O. Levenspiel, *Fluidization Engineering*. Butterworth-Heinemann, 1991.

[3] D. Gidaspow, *Multiphase Flow and Fluidization*. Academic Press, 1994.

[4] R. Jackson, *The Dynamics of Fluidized Particles*. Cambridge University Press, 2000.

[5] L. Gibilaro, *Fluidization Dynamics*. Butterworth-Heinemann, 2001.

[6] M.R. Gorliz, J.R. Grace, Augmentation of heat transfer by deflectors on circulating fluidized bed membrane walls. *International Journal of Heat and Mass Transfer*, **45**, 1149-1154, 2002.

[7] T. Samruamphianskun, P. Piumsomboon, B. Chalermsinsuwan, Effect of ring baffle configurations in a circulating fluidized bed riser using CFD simulation and experimental design analysis. *Chemical Engineering Journal*, **210**, 237–251, 2012.

[8] B. Chalermsinsuwana, T. Samruamphianskun, P. Piumsomboon, Effect of operating parameters inside circulating fluidized bed reactor riser with ring baffles using CFD simulation and experimental design analysis. *Chemical Engineering Research and Design*, **92**, 2479-2492, 2014.

[9] S. Piovano, G.L. Salierno, E. Montmany, M. D'Agostino, M. Maestri, M. Cassanello, Bed Expansion and Particle Classification in Liquid Fluidized Beds with Structured Internals. *Chemical Engeneering and Technology*, **38 (3)**, 423-430, 2015.

[10] R. Gujjula, N. Mangadoddy, Experimental Investigation of Hydrodynamics of Gas-Solid Flow in an Internally Circulating Fluidized Bed. *Canadian Journal of Chemical Engineering*, **93**, 1380-1391, 2015.

[11] S.A. Solov'ev, A.G. Egorov, A.A. Lamberov, S.R. Egorova, A.N. Kataev, Effect of the Design of a Feedstock Injection Device in a Fluidized_Bed Reactor on the Efficiency of

the Reaction, Using the Dehydrogenation of iso_Paraffins in a Fluidized Chromia–Alumina Catalyst Bed as an Example. *Catalysis in Industry*, **8 (1)**, 48-55, 2016.

[12] Y.-Q. Zhuanga, X.-M. Chena, Z.-H. Luob, J. Xiao, CFD–DEM modeling of gas–solid flow and catalytic MTO reactionin a fluidized bed reactor. *Computers and Chemical Engineering*, **60**, 1–16, 2014.

[13] W. Shuai, Z. Tianyu, L. Guodong, L. Huilin, S. Liyan, Multi scale heat transfer in fluidized bed reactors by Eulerian CFD modeling. *Fuel*, **139**, 646–651. 2015.

[14] M. Upadhyay, J.-H. Park, CFD simulation via conventional Two-Fluid Model of a circulating fluidized bed riser: Influence of models and model parameters on hydrodynamic behavior. *Powder Technology*, **272**, 260-268, 2015.

[15] M. Beccari, U. Romano, *Encyclopaedia of Hydrocarbons, Vol. 2*. ENI & Istituto della Enciclopedia Italiana G.Treccani, 2006.

[16] J. Ding, D. Gidaspow, A bubbling fluidization model using kinetic theory of granular flow. *AICHE Journal*, **36 (4)**, 523–538, 1990.

[17] L. Schiller, Z. Naumann, A Drag Coefficient Correlation. *Z. Ver. Deutsch. Ing*, **77**, 318-320, 1935.

[18] M. Syamlal and T. J. O'Brien, Computer Simulation of Bubbles in a Fluidized Bed. *AIChE Symp. Series*, **85**, 22–31, 1989.

[19] M. Syamlal, The Particle-Particle Drag Term in a Multiparticle Model of Fluidization. *National Technical Information Service, VA, DOE/MC/21353-2373, NTIS/DE87006500*, 1987.

[20] C.K.K. Lun, S.B. Savage, D.J. Jeffrey, Kinetic Theories for Granular Flow: Inelastic Particles in Couette Flow and Slightly Inelastic Particles in a General Flow Field. *J. Fluid Mech*, **140**, 223–256, 1984.

[21] W.E. Ranz, W.R. Marshall, Evaporation from Drops, Part I. *Chem. Eng. Prog.*, **48 (3)**, P. 141–146, 1952.

[22] W.E. Ranz, W.R. Marshall, Evaporation from Drops, Part II. *Chem. Eng. Prog.*, **48 (4)**, 173–180, 1952.

[23] D. J. Gunn, Transfer of Heat or Mass to Particles in Fixed and Fluidized Beds. *Int. J. Heat Mass Transfer*, **21**, 467–476, 1978.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SHOCK CAPTURING COMPUTATIONS WITH STABILIZED POWELL-SABIN ELEMENTS

**Giorgio Giorgiani[1], Hervé Guillard[2] and Boniface Nkonga[2]**

[1]Aix-Marseille Université
38 rue Joliot-Curie, La Jete, 13451 Marseille
e-mail: giorgio.giorgiani@univ-amu.fr

[2] INRIA Sophia Antipolis
2004 Route des Lucioles, 06902 Valbonne
e-mail: {herve.guillard,boniface.nkonga}@inria.fr

**Keywords:** Shock capturing, high-order, C1 approximations, splines, CFD, compressible flow.

**Abstract.** *In the last recent years, thanks to the increasing power of the computational machines, the interest in more and more accurate numerical schemes is growing. Methods based on high-order approximations are nowadays the common trend in the computational research community, in particular for CFD applications.*

*This work is focused on Powell-Sabin (PS) finite elements, a finite element method (FEM) based on PS splines. PS splines are piecewise quadratic polynomials with a global C1 continuity, defined on conforming triangulations. Despite its attractive characteristics, so far this scheme hasn't had the attention it deserves. PS splines are adapted to unstructured meshes and, contrary to classical tensor product B-splines, they are particularly suited for local refinement, a crucial aspect in the analysis of highly convective and anisotropic equations. The additional global smoothness of the C1 space has a beneficial stabilization effect in the treatment of advection-dominated equations and leads to a better capturing of thin layers. Finally, unlike most of other typology of high-order finite elements, the numerical unknowns in PS elements are located in the vertices of the triangulation, leading to an easy treatment of the parallel aspects.*

*Some geometrical issues related to PS elements are discussed here, in particular, the generation of the control triangles and the imposition of the boundary conditions. The PS FEM method is used to solve the compressible Euler equation in supersonic regime. A classical shock-capturing technique is used to reduce the oscillation around the discontinuity, while a variational multiscale formulation is used to introduce numerical diffusion in the streamwise direction. Some typical numerical examples are used to evaluate the performance of the PS discretization.*

## 1 INTRODUCTION

Numerical methods with high degrees of regularity are of interest in many physical applications, such as magnetohydrodynamics, shells analysis and vibrations. In particular, finite-element methods based on spline shape functions have raised a particular interest in the recent years. On one hand, $C^1$ spline methods allow to discretize higher-order derivatives (that is, derivatives of order superior to the second), and on the other hand, they provide accurate geometrical representation of the computational domain, see for example [1, 3] in the context of Isogeometric Analysis. Moreover, it has also been shown recently that the additional global smoothness of the spline interpolant introduces stability to the numerical solution for highly convective equations, see [1] and [7], and also for turbulence computations, see [2]. Hence, finite element methods based on spline shape functions can be advantageous also in the context of fluid-dynamic problems.

Nowadays, splines are extensively used in the graphical design industry to create smooth surfaces. The success of splines is basically due to the fact that they have very attractive and unique characteristics: they have a compact representation, they are able to represent curves and surfaces with an arbitrary level of regularity, simply increasing the polynomial degree of the basis, and they are efficient in terms of locality, that is, they can be modified locally by moving a control point, without perturbing the rest of the curve. Spline curves are usually represented as a linear combination of basis functions, called *B-splines*. The basis functions of a given order $p$ are defined with a recursive relation starting with piecewise constant basis functions (that is, $p = 0$),

$$B_{i,0}(\xi) = \begin{cases} 1 & \text{if } \xi_i \leq \xi < \xi_{i+1}, \\ 0 & \text{otherwise} \end{cases}$$

$$B_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} B_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} B_{i+1,p-1}(\xi),$$

where the points $\xi_i$ are the knots of the spline, $i = 1, 2, \ldots, n + p + 1$, and $n$ is the number of basis functions. Hence, the spline curve expressed as a combination of B-splines is

$$S(\xi) = \sum_{i=1}^{n} B_{i,p} \mathbf{P}_i, \tag{1}$$

where $\mathbf{P}_i$ are the control points of the spline.

Starting from the one-dimensional case (1), B-splines can be extended to higher dimensions through a tensor product representation. However, tensor product B-splines are restricted to structured rectangular meshes. The refinement procedure with tensor B-splines relies on the insertion of knots, leading hence to a global modification of the domain discretization. Thus, no local refinement is possible. Various solutions are proposed in literature to solve this limitation, for example, the introduction of *T-splines* in [3].

This work is focused on another approach, based on the definition of bivariate splines on irregular triangulations, that is, Powell-Sabin (PS) splines [5, 6, 7, 8]. PS splines are piecewise quadratic polynomial with $C^1$ continuity, defined on a unstructured triangulation of the domain. The major advantage with respect to tensor product B-splines is that PS splines allow a straightforward adaptive refinement of the mesh, which is a key ingredient in the simulation of anisotropic equations. Another advantage of PS elements is related to the fact that the unknowns are concentrated on the nodes of the triangulation. This means that there are no unknowns on

the element faces or in the interior of the elements. This is not the case for other $C^1$ interpolation techniques, such as, for example, the Clough-Tocher elements. The consequence is that the stencil is the same for each node of the mesh, which produces a linear system matrix with a regular block shape that facilitates the implementation aspect, enhances the parallel performance and improves the efficiency of the linear system solution.

In this work, a PS finite element method is presented for the solution of the two dimensional (2D) compressible Euler equation in supersonic regime. A SUPG stabilization technique is used to introduce numerical dissipation in the streamwise direction only, and an isotropic diffusion terms is added to avoid oscillation around the discontinuities.

The outline of the paper is as follows. Section 2 introduces the PS splines and their representation. Section 3 deals with the geometrical and mathematical tools to define the PS elements and in particular the generation of the shape functions. A detailed discussion is also given on the imposition of the boundary conditions. Finally, in Section 4 two classical numerical examples are presented.

## 2 POWELL-SABIN SPLINES

In this section is introduced the framework and the notation to define spline functions on a triangulation of a polygonal domain. Starting from the definition of bivariate polynomials on a triangle, the goal is to obtain an interpolant of a generic function with $C^1$ continuity. Considering a single triangle $\Omega_k$ of vertices $\mathbf{V}_i = (x_i, y_i)$, with $i = 1, 2, 3$, any bivariate polynomial $p(x, y)$ of degree $\leq 2$ in the space $\Pi_2 = \{\sum_{i=0}^{2} \sum_{j=0}^{2-i} a_{i,j} x^i y^j, a_{i,j} \in \mathbb{R}\}$ can be written in the Bernstein-Bézier representation

$$p(x, y) := p(\xi) = \sum_{\substack{i+j+k=2 \\ i,j,k \geq 0}} b_{i,j,k} B^2_{i,j,k}(\xi(x, y)), \tag{2}$$

where $\xi = (\xi_1, \xi_2, \xi_3)$ are the barycentric coordinates of a point $(x, y) \in \mathbb{R}^2$ and $B^2_{i,j,k}(\xi)$ are the Bernstein polynomials of degree 2 in $\Omega_k$, that is

$$B^2_{i,j,k} = \frac{2!}{i!j!k!} \xi_1^i \xi_2^j \xi_3^k. \tag{3}$$

The set of Bernstein polynomials $B^2_{i,j,k}(\xi)$ is a basis for the space of polynomials $\Pi_2$, see [4], hence the six coefficients $b_{i,j,k}$ uniquely define the second order polynomial $p(x, y)$ and they are called the Bézier ordinates with respect to the triangle $\Omega_k$[1].

Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain with boundary $\partial\Omega$. Let $\mathcal{T}$ be a conforming triangulation of $\Omega$ with vertices $\mathbf{V}_l$, $l = 1, \ldots, N_v$ and elements $\Omega_e$, $e = 1, \ldots, N_e$. Defining $\mathcal{S}_1^2$ as the linear space of piecewise quadratic polynomials on $\mathcal{T}$, the following interpolation problem is considered: given any set of triples $(f_l, f_{xl}, f_{yl})$, $l = 1, \ldots, Nv$, find $s(x, y) \in \mathcal{S}_1^2$ such that,

$$s(\mathbf{V}_l) = f_l, \quad \frac{\partial s}{\partial x}(\mathbf{V}_l) = f_{xl}, \quad \frac{\partial s}{\partial y}(\mathbf{V}_l) = f_{yl}, \quad l = 1, \ldots, Nv. \tag{4}$$

It is clear that such a problem has no solution in general: in fact, problem (4) requires the imposition of nine parameters to define the second order polynomial on each triangle, while only six coefficients are available, see equation (2).

[1]Note that (2) can be expanded as $p(\xi) = b_{2,0,0}\xi_1^2 + b_{0,2,0}\xi_2^2 + b_{0,0,2}\xi_3^2 + 2b_{1,1,0}\xi_1\xi_2 + 2b_{1,0,1}\xi_1\xi_3 + 2b_{0,1,1}\xi_2\xi_3$.

Thus, to obtain a solution to the interpolation problem (4), one possibility is to subdivide each triangle in sub-triangles and to define the interpolant in the refined triangulation. One of the solutions proposed by Powell and Sabin in [5] is based on the subdivision of each triangle in $\mathcal{T}$ into six smaller triangles (*PS6-split*). Hence, the conditions in (4) are imposed only on the vertices of the original triangulation, while in the other added nodes only $C_1$ continuity of the interpolating function is imposed. More details can be found in [5].

The so called Powell-Sabin refinement of $\mathcal{T}$ is denoted with $\mathcal{T}^*$ and is obtained dividing each triangle in $\mathcal{T}$ in six sub-triangles. The procedure to define $\mathcal{T}^*$ is described in [7] and reads as follows.

- Select a split point $\mathbf{C}_k$ inside each triangle $\Omega_k$ and connect it to the three vertices of $\Omega_k$.
- For each pair of triangles $\Omega_p$ and $\Omega_q$, with a common edge, connect the two points $\mathbf{C}_p$ and $\mathbf{C}_q$. If $\Omega_p$ is a boundary triangle, connect also $\mathbf{C}_p$ to an arbitrary point on the boundary edge.

The PS refinement defines also a set of points, called *PS points*, associated to each vertex. The PS points associated to the vertex $\mathbf{V}_l$ are defined as the midpoints of all the edges of the PS refinement containing $\mathbf{V}_l$, plus the point $\mathbf{V}_l$ itself. The PS points are fundamental for the definition of the shape functions as explained in the next section.

Having defined the PS refinement $\mathcal{T}^*$, the linear space of piecewise quadratic polynomials with $C^1$ continuity can be denoted as $S_2^1(\mathcal{T}^*)$. Each element $S_2^1(\mathcal{T}^*)$ is uniquely defined by its values and derivatives at the vertices of $\mathcal{T}$, thus the functional space $S_2^1(\mathcal{T}^*)$ has dimension $3N_v$.

## 3 POWELL-SABIN FINITE ELEMENTS

In this section PS splines are introduced in a finite-element framework to solve partial differential equations, defining shape functions belonging to the space $S_2^1(\mathcal{T}^*)$. Hence, it is necessary to express each element $s(x, y) \in S_2^1(\mathcal{T}^*)$ as a linear combination of PS spline basis functions, that is

$$s(x, y) = \sum_{l=1}^{N_v} \sum_{r=1}^{3} c_{l,r} B_l^{(r)}(x, y),$$

where the functions $B_l^{(r)}$ are called *Powell-Sabin B-splines* and $c_{l,r}$ are the coefficients of the representation.

Each B-spline $B_l^{(r)}(x, y)$ can be seen as the solution of the interpolation problem (4) with all $(f_i, f_{xi}, f_{yi}) = 0$ except for $(f_l, f_{xl}, f_{yl}) = (\alpha, \beta, \gamma) \neq 0$. The quantity $(\alpha, \beta, \gamma)$ is called a *triple* and represents the value of the function and the derivatives with respect the Cartesian axes in a given vertex. Then, it is easy to see that each basis $B_l^{(r)}$, $r = 1, 2, 3$, vanishes outside the molecule $M_l$ of vertex $\mathbf{V}_l$ defined as the union of all triangles $\Omega_e$ containing $\mathbf{V}_l$. This guarantees the compactness of the support for PS B-splines.

In principle, it is possible to define the PS B-splines simply choosing, for each vertex $\mathbf{V}_l \in \mathcal{T}$, three linearly independent vectors $(\alpha_{l,r}, \beta_{l,r}, \gamma_{l,r})$, $r = 1, 2, 3$, and solving the interpolation problem (4). This procedure generates three linearly independent basis functions for each vertex. However, in a finite element context, it is interesting to work with *normalized*

B-splines, that is, basis functions that form a partition of unity on $\Omega$, i.e.

$$B_l^{(r)}(x,y) \geq 0 \quad \forall x,y \in \Omega,$$
$$\sum_{l=1}^{N_v}\sum_{r=1}^{3} B_l^{(r)}(x,y) = 1 \quad \forall x,y \in \Omega,$$

which leads to the following constrains on the triples

$$\sum_{r=1}^{3} \alpha_{l,r} = 1, \quad \sum_{r=1}^{3} \beta_{l,r} = 0, \quad \sum_{r=1}^{3} \gamma_{l,r} = 0.$$

This property is also crucial to ensure the completeness of the basis.

The procedure proposed in [8] allows to obtain PS B-splines that form a partition of unity and with good characteristic from a numerical point of view. It is based on the fact that each set of three independent B-splines associated to a vertex $\mathbf{V}_l$ uniquely define a *control triangle* $T_l = (\mathbf{Q}_{l,1}, \mathbf{Q}_{l,2}, \mathbf{Q}_{l,3})$ for the PS spline, with vertices $\mathbf{Q}_{l,j} = (X_{l,j}, Y_{l,j})$, $j = 1,2,3$. The control triangle of a PS spline as a role similar to the control polygon for univariate splines in terms of control of the shape of the spline. In fact, in [8] it is shown that the control triangle and the coefficients $c_{l,j}$ define three control points $(X_{l,j}, Y_{l,j}, c_{l,j})$ defining a spatial plane that is always tangent to the spline surface at the vertex $\mathbf{V}_l$. Optimal shape functions from a numerical point of view are associated to a minimal area control triangle [8].

The points $\mathbf{Q}_{l,j}$ are related to the three triples and the vertex $\mathbf{V}_l$ via the linear system

$$\begin{bmatrix} \alpha_{l,1} & \alpha_{l,2} & \alpha_{l,3} \\ \beta_{l,1} & \beta_{l,2} & \beta_{l,3} \\ \gamma_{l,1} & \gamma_{l,2} & \gamma_{l,3} \end{bmatrix} \begin{bmatrix} X_{l,1} & Y_{l,1} & 1 \\ X_{l,2} & Y_{l,2} & 1 \\ X_{l,3} & Y_{l,3} & 1 \end{bmatrix} = \begin{bmatrix} x_l & y_l & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{5}$$

The positivity of the B-splines associated to the vertex $\mathbf{V}_l$ is guaranteed if all the barycentric coordinates with respect the control triangle $T_l$ of all the PS points associated to $\mathbf{V}_l$ are positive. This means geometrically that all the PS points associated to $\mathbf{V}_l$ are contained in the control triangle $T_l$, see Figure 1.

Hence, the procedure of computing the three shape functions for each vertex is:

- find a suitable control triangle for the vertex $\mathbf{V}_l$ that contains all the PS points associated to $\mathbf{V}_l$;
- compute the triple defining the three shape functions solving the linear system (5).

The control triangles for each vertex are constructed looking for the minimal area triangle that contains all the PS points. This lead to an optimization problem consisting in circumscribing a convex polygon with a minimal area triangle. An optimal algorithm is proposed by O'Rourke in [9] for solving such a problem, with a computational cost $\mathcal{O}(n)$, with $n$ the number of PS points. This is the algorithm used here in the numerical tests. In Figure 1 is depicted an example of control triangle, with the relative PS control points, obtained with the O'Rourke algorithm.

In order to understand the role of the control triangle in the definition of the basis functions, a comparison of the condition number of the linear system matrix as been carried out for three structured computational meshes for a square domain $[0,1]\times[0,1]$. In the first case, the minimal area triangle have been used to compute the basis. In the second case, the minimal area triangle

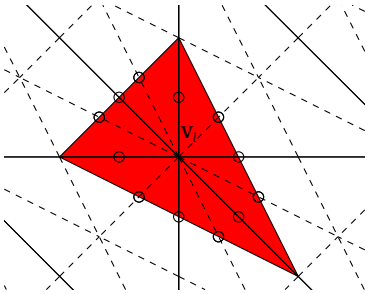



Figure 1: Powell-Sabin control triangle and relative Powell-Sabin points for the vertex $\mathbf{V}_l$.

Table 1: Condition number comparison for the stiffness matrix and the mass matrix obtained with optimal and non-optimal control triangles in a square domain: $Ne$ is the number of triangular elements of the mesh, $h$ is the element size.

| | | Cond. numb. of the stiffness matrix | | Cond. numb. of the mass matrix | |
|---|---|---|---|---|---|
| Ne | $h$ | optimal | non optimal | optimal | non optimal |
| 200 | 70.71E-03 | 83 | 389 | 171 | 4466 |
| 800 | 35.36E-03 | 326 | 453 | 163 | 4466 |
| 3200 | 17.68E-03 | 1305 | 1585 | 154 | 4469 |

has been expanded moving (to a distance equal to five times the original one) each vertex along the line joining the vertex with the node to which the triangle belongs. Table 1 shows the results: it is possible to notice how increasing the control triangle area degrades the condition number of mass matrix and the stiffness matrix.

## 4 STABILIZED APPROXIMATION OF THE EULER EQUATIONS WITH POWELL-SABIN ELEMENTS

The 2D inviscid Euler equations of gas dynamics are considered,

$$\begin{aligned} \frac{\partial \rho}{\partial t} + \boldsymbol{\nabla}\cdot(\rho \boldsymbol{u}) &= 0 \\ \frac{\partial \rho \boldsymbol{u}}{\partial t} + \boldsymbol{\nabla}\cdot(\rho \boldsymbol{u} \otimes \boldsymbol{u} + p\boldsymbol{\mathcal{I}}) &= \mathbf{0} \\ \frac{\partial \rho E}{\partial t} + \boldsymbol{\nabla}\cdot((\rho E + p)\boldsymbol{u}) &= 0 \end{aligned} \tag{6}$$

where $\rho$ is the density, $\boldsymbol{u}$ is the velocity vector, $E$ is the total energy, $\boldsymbol{\mathcal{I}}$ is the identity matrix and $p$ is the pressure, which is defined by the equation of state $p = (\gamma_g - 1)\rho\epsilon$, $\gamma_g = 1.4$ is the ratio of specific heats and $\epsilon = E - 1/2\|\boldsymbol{u}\|^2$ is the internal energy.

Problem (6) is recast in a conservative form,

$$\frac{\partial \boldsymbol{U}}{\partial t} + \boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{F}} = \mathbf{0}, \tag{7}$$

having defined the vector of conservative variables $\boldsymbol{U} = (\rho, \rho u, \rho v, \rho E)^T$ and the flux tensor $\boldsymbol{\mathcal{F}}$ with components, in Cartesian coordinates,

$$\boldsymbol{\mathcal{F}} = \left[\boldsymbol{F}_x\ \boldsymbol{F}_y\right], \quad \boldsymbol{F}_x = \left\{\begin{array}{c} \rho u \\ \rho u^2 + p \\ \rho uv \\ (\rho E + p)u \end{array}\right\}, \quad \boldsymbol{F}_y = \left\{\begin{array}{c} \rho v \\ \rho uv \\ \rho v^2 + p \\ (\rho E + p)v \end{array}\right\},$$

where $u$ and $v$ are the $x$ and $y$ components of the velocity vector.

A stabilized formulation based on PS splines is considered to solve problem (7). To simplify the presentation, in an abuse of notation, the same symbol is used for the numerical approximation, belonging to the finite dimensional space, and the exact solution $\boldsymbol{U}$. Hence, the stabilized finite element problem corresponding to (7) reads: find $\boldsymbol{U} \in [S_2^1(\mathcal{T}^*)]^4$ such that

$$\int_\Omega \boldsymbol{V} \cdot \frac{\partial \boldsymbol{U}}{\partial t}\, d\Omega - \int_\Omega \boldsymbol{\nabla V} : \boldsymbol{\mathcal{F}}\, d\Omega + \int_{\partial\Omega} \boldsymbol{\mathcal{F}V} \cdot \boldsymbol{n}\, d\Gamma + \mathcal{D}_{SUPG} + \mathcal{D}_{SC} = 0, \tag{8}$$

for all $\boldsymbol{V} \in [S_2^1(\mathcal{T}^*)]^4$. The terms $\mathcal{D}_{SUPG}$ and $\mathcal{D}_{SC}$ are the stabilization terms and they will be discussed later.

A fully implicit first-order time integration scheme is employed to approximate the time derivatives. The non-linear convective term is linearized using a Newton-Raphson procedure of the type

$$\int_\Omega \boldsymbol{\nabla V} : \boldsymbol{\mathcal{F}}^k\, d\Omega \approx \int_\Omega \boldsymbol{\nabla V} : \mathbb{A}^{k-1}\boldsymbol{U}^k\, d\Omega,$$

where $k$ stands for the current N-R iteration and the third order tensor $\mathbb{A}$ is the Jacobian of the flux,

$$\mathbb{A} = \frac{\partial \boldsymbol{\mathcal{F}}}{\partial \boldsymbol{U}},$$

and can be expressed as a second order tensor for each Cartesian direction, that is

$$\frac{\partial \boldsymbol{F}_x}{\partial \boldsymbol{U}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{\gamma_g-1}{2}\|\boldsymbol{u}\|^2 - u^2 & (3-\gamma_g)u & (1-\gamma_g)v & \gamma_g - 1 \\ -uv & v & u & 0 \\ \frac{\gamma_g-1}{2}u\|\boldsymbol{u}\|^2 - uH & -u^2(\gamma_g-1)+H & (1-\gamma_g)uv & \gamma_g u \end{pmatrix},$$

$$\frac{\partial \boldsymbol{F}_y}{\partial \boldsymbol{U}} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ \frac{\gamma_g-1}{2}\|\boldsymbol{u}\|^2 - v^2 & (1-\gamma_g)u & (3-\gamma_g)v & \gamma_g - 1 \\ \frac{\gamma_g-1}{2}v\|\boldsymbol{u}\|^2 - vH & (1-\gamma_g)uv & -v^2(\gamma_g-1)+H & \gamma_g v \end{pmatrix},$$

having defined the entalpy, $H = E + p/\rho$.

### 4.1 Imposition of the boundary conditions

Solid walls and free-stream boundaries are considered for defining the boundary conditions of system (6). Walls are considered as slip boundary conditions, that is, the normal component of the flux is set to zero on a wall. This kind of condition does not entail particular problems, and it is weakly imposed by the boundary integral in (8). On the contrary, inlet and outlet boundary conditions are imposed setting different components of the unknown vector at the boundary nodes, depending on the subsonic/supersonic regime at the boundary, see [15]. The problem of imposing boundary values deserves particular attention in the framework of PS finite elements. Since these kind of conditions are typically set on straight boundaries, the interest is focused on imposing boundary values with PS elements on polygonal boundaries.

A careful choice of the control triangle associated to the boundary node $\mathbf{V}_l$ simplify the task of imposing boundary values. Two different cases must be considered: the boundary node belongs to two edges forming an angle different to $\pi$, or equal to $\pi$. From equation (5), the

following relations are derived

$$\alpha_{l,1}\mathbf{Q}_{l,1} + \alpha_{l,2}\mathbf{Q}_{l,2} + \alpha_{l,3}\mathbf{Q}_{l,3} = \mathbf{V}_l, \tag{9}$$
$$\beta_{l,1}\mathbf{Q}_{l,1} + \beta_{l,2}\mathbf{Q}_{l,2} + \beta_{l,3}\mathbf{Q}_{l,3} = \mathbf{e}_x, \tag{10}$$
$$\gamma_{l,1}\mathbf{Q}_{l,1} + \gamma_{l,2}\mathbf{Q}_{l,2} + \gamma_{l,3}\mathbf{Q}_{l,3} = \mathbf{e}_y, \tag{11}$$

where $\mathbf{e}_x$ and $\mathbf{e}_y$ are respectively the unity vector in the $x$ and $y$ directions. Equation (9) states that $(\alpha_{l,1}, \alpha_{l,2}, \alpha_{l,3})$ are the barycentric coordinates of the vertex $\mathbf{V}_l$ with respect to the control triangle, while (10) and (11) mean that $(\beta_{l,1}, \beta_{l,3}, \beta_{l,3})$ and $(\gamma_{l,1}, \gamma_{l,2}, \gamma_{l,3})$ are respectively the barycentric coordinates of the vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ with respect to the control triangle.

In the case $\mathbf{V}_l$ forms an angle different from $\pi$, if the control triangle has two sides aligned with the two boundary edges as in Figure 2 (left), from equation (9) derives that only one shape function has value different from zero in $\mathbf{V}_l$. This allows to directly set the boundary value to the unknown associated to the non-zero shape function. Moreover, from (10) and (11) derives that the other two shape functions have zero derivative in the directions the two edges concurring in $\mathbf{V}_l$. In particular, using the notation depicted in Figure 2 (left) where triangle vertices are now called $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, it results

$$\begin{aligned} \alpha_{l,B} = \alpha_{l,C} = 0, (\Rightarrow \alpha_{l,A} = 1), \\ \begin{Bmatrix} \beta_{l,B} \\ \gamma_{l,B} \end{Bmatrix} \cdot \boldsymbol{t} = 0, \\ \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} \cdot \boldsymbol{r} = 0. \end{aligned}$$

Denoting with $U$ the component of $\boldsymbol{U}$ to be imposed, and $\hat{u}$ the $\mathcal{L}^2$ projection of the prescribed values on the space of univariate quadratic splines on $\partial\Omega$, the resulting equations relative to the unknowns of the vertex $\mathbf{V}_l$ are

$$\begin{aligned} U_{l,A} = \hat{u}(\mathbf{V}_l), \\ U_{l,A} \begin{Bmatrix} \beta_{l,A} \\ \gamma_{l,A} \end{Bmatrix} \cdot \boldsymbol{r} + U_{l,B} \begin{Bmatrix} \beta_{l,B} \\ \gamma_{l,B} \end{Bmatrix} \cdot \boldsymbol{r} = \boldsymbol{\nabla}\hat{u}(\mathbf{V}_l) \cdot \boldsymbol{r}, \\ U_{l,A} \begin{Bmatrix} \beta_{l,A} \\ \gamma_{l,A} \end{Bmatrix} \cdot \boldsymbol{t} + U_{l,C} \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} \cdot \boldsymbol{t} = \boldsymbol{\nabla}\hat{u}(\mathbf{V}_l) \cdot \boldsymbol{t}, \end{aligned}$$

that is

$$\begin{aligned} U_{l,A} &= \hat{u}(\mathbf{V}_l), \\ U_{l,B} &= \frac{\boldsymbol{\nabla}\hat{u}(\mathbf{V}_l) \cdot \boldsymbol{r} - \hat{u}(\mathbf{V}_l) \begin{Bmatrix} \beta_{l,A} \\ \gamma_{l,A} \end{Bmatrix} \cdot \boldsymbol{r}}{\begin{Bmatrix} \beta_{l,B} \\ \gamma_{l,B} \end{Bmatrix} \cdot \boldsymbol{r}}, \\ U_{l,C} &= \frac{\boldsymbol{\nabla}\hat{u}(\mathbf{V}_l) \cdot \boldsymbol{t} - \hat{u}(\mathbf{V}_l) \begin{Bmatrix} \beta_{l,A} \\ \gamma_{l,A} \end{Bmatrix} \cdot \boldsymbol{t}}{\begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} \cdot \boldsymbol{t}}, \end{aligned}$$

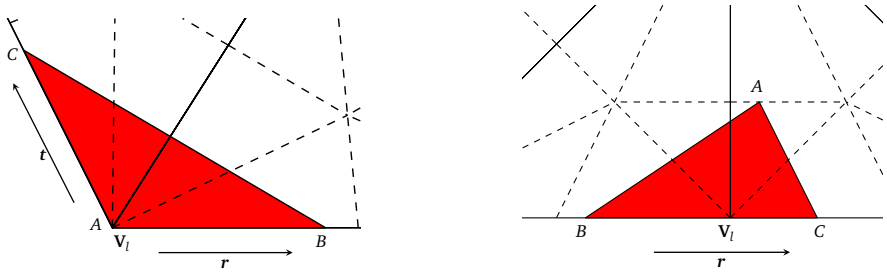



Figure 2: Control triangles for imposing boundary values: on an angle different from $\pi$ (left) and equal to $\pi$ (right).

where $U_{l,A}, U_{l,B}, U_{l,C}$ are the unknowns related to the shape functions with respectively triples $(\alpha_{l,A}, \beta_{l,A}, \gamma_{l,A})$, $(\alpha_{l,B}, \beta_{l,B}, \gamma_{l,B})$ and $(\alpha_{l,C}, \beta_{l,C}, \gamma_{l,C})$.

In the case $\mathbf{V}_l$ belongs to a plane angle, if the control triangle is aligned with the boundary as in Figure 2 (right), there is one shape function with zero value on $\mathbf{V}_l$, see equation (9). Hence, the unknown related to this shape function is not affected by the boundary condition. Moreover, again equations (10) and (11) guarantee that the same shape function has zero derivative in the direction parallel to the boundary. Thus, with reference to Figure 2 (right), results

$$\alpha_{l,A} = 0,\ \alpha_{l,B} \neq 0,\ \alpha_{l,C} \neq 0,$$
$$\begin{Bmatrix} \beta_{l,A} \\ \gamma_{l,A} \end{Bmatrix} \cdot \boldsymbol{r} = 0,$$

leading to the following relations for the unknowns in the vertex $\boldsymbol{V}_l$

$$U_{l,B}\alpha_{l,B} + U_{l,C}\alpha_{l,C} = \hat{u}(\mathbf{V}_l),$$
$$U_{l,B} \begin{Bmatrix} \beta_{l,B} \\ \gamma_{l,B} \end{Bmatrix} \cdot \boldsymbol{r} + U_{l,C} \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} \cdot \boldsymbol{r} = \boldsymbol{\nabla}\hat{u}(\mathbf{V}_l) \cdot \boldsymbol{r},$$

that is

$$U_{l,B} = \frac{\left(\hat{u}(\mathbf{V}_l) \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} - \alpha_{l,C}\boldsymbol{\nabla}\hat{u}(\mathbf{V}_l)\right) \cdot \boldsymbol{r}}{\left(\alpha_{l,B} \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} - \alpha_{l,C} \begin{Bmatrix} \beta_{l,B} \\ \gamma_{l,B} \end{Bmatrix}\right) \cdot \boldsymbol{r}},$$
$$U_{l,C} = \frac{\left(\hat{u}(\mathbf{V}_l) \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} - \alpha_{l,C}\boldsymbol{\nabla}\hat{u}(\mathbf{V}_l)\right) \cdot \boldsymbol{r}}{\left(\alpha_{l,B} \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix} - \alpha_{l,B} \begin{Bmatrix} \beta_{l,C} \\ \gamma_{l,C} \end{Bmatrix}\right) \cdot \boldsymbol{r}}.$$

### 4.2 Stabilization by artificial viscosity

Stabilized formulations are obtained adding extra terms in the Galerkin weak form, with the goal of reducing the instabilities with the introduction of artificial viscosity. Here, two stabilization terms are considered.

The first term is the streamline upwind Petrov-Galerkin (SUPG) term, used to produce a stable upwind discretization without introducing excessive numerical dissipation. The SUPG method was introduced by T. Hughes in [11] and it is one of the most established stabilized formulations in finite element flow computations. The SUPG method introduces a certain amount of artificial viscosity in the streamline direction only.

The second stabilization term is a shock capturing term, and it is used to prevent oscillations around the discontinuities arising in the solution in supersonic regimes. These spurious oscillations might lead to severe accuracy loss or stability problems. Differently from the SUPG stabilization term, the shock-capturing stabilization term introduce an isotropic artificial diffusion, but only in a sharp zone surrounding the discontinuities.

Since a first order time discretization is employed, a simplified version of the SUPG stabilization term is used

$$\mathcal{D}_{SUPG} = \sum_{e=1}^{Ne} \int_{\Omega_e} \mathbb{A}^T \boldsymbol{\nabla V} : \left(\tau_{\text{SUPG}} \mathbb{A} \boldsymbol{\nabla U}\right) d\Omega_e, \tag{12}$$

which does not take into account the complete residual of the original equation, but only the convective term.

The shock capturing stabilization term is

$$\mathcal{D}_{SC} = \sum_{e=1}^{Ne} \int_{\Omega_e} \boldsymbol{\nabla V} : \left(\tau_{\text{SC}} \boldsymbol{\nabla U}\right) d\Omega_e. \tag{13}$$

The terms $\tau_{\text{SUPG}}$ and $\tau_{\text{SC}}$ are stabilization matrices. Various options to compute these terms can be found in literature. In this work, for the SUPG stabilization term, two options have been retained: the first one is a simple diagonal matrix

$$\tau_{\text{SUPG}} \equiv \tau_{\text{SP}} = \tau \mathbf{I},$$

and a constant value $\tau$ is used in the whole mesh, $\tau = dt/2$, being $dt$ the time step. The second choice, denoted $\tau_{Tz}$, is inspired by the work of Tezduyar [13]

$$\tau_{Tz} = \left((1/2dt)^{(-2)} + \sum_{r=1}^{3}\sum_{q=1}^{3} \left|\boldsymbol{u} \cdot \boldsymbol{\nabla} B_r^{(q)}\right|\right)^{(-1/2)}.$$

For the shock-capturing operator, two choices described in [13, 12] are retained, and are described using four sensors. The first sensor is based on the relative density variation and it is defined, in each element, as

$$[\widetilde{\boldsymbol{\nabla}\rho}]_e = \sum_{r=1}^{3}\sum_{q=1}^{3} \left| \frac{\boldsymbol{\nabla}\rho \cdot \boldsymbol{\nabla} B_r^{(q)}}{|\boldsymbol{\nabla}\rho|} \right|.$$

The second sensor is based on the relative gradient according to the principle axes and is written

$$[\widetilde{\boldsymbol{\nabla U}}]_e = \left( \sum_{d=1}^{2} \left\| \frac{\partial \boldsymbol{U}}{\partial x_d} \right\|^2 \right)^{\frac{1}{2}},$$

while the third sensor scales the with the values of the unknowns

$$[\widetilde{\boldsymbol{U}}]_e = \|\boldsymbol{U}\|.$$

Finally the fourth sensor takes into account the local relative convective flux

$$[\widetilde{\mathbb{A}\boldsymbol{\nabla U}}]_e = \|\mathbb{A}\boldsymbol{\nabla U}\|.$$

Hence, the first shock capturing artificial viscosity is defined as a combination of the sensors with a managing parameter $\beta$ as follows

$$\tau_{\text{SC1}}(\beta) = [\widetilde{\boldsymbol{\nabla}\rho}]_e [\widetilde{\boldsymbol{\nabla}\boldsymbol{U}}]_e^{-(2-\beta)} [\mathbb{A}\widetilde{\boldsymbol{\nabla}\boldsymbol{U}}]_e^{-\beta},$$

and the second

$$\tau_{\text{SC2}}(\beta) = [\widetilde{\boldsymbol{\nabla}\rho}]_e [\widetilde{\boldsymbol{\nabla}\boldsymbol{U}}]_e^{-(2-\beta)} [\widetilde{\boldsymbol{U}}]_e^{(1-\beta)} [\mathbb{A}\widetilde{\boldsymbol{\nabla}\boldsymbol{U}}]_e^{-\beta}.$$

In [12], it is suggested to use $\beta = 1$ is for smooth shocks and $\beta = 2$ for strong shocks. A compromise between the two definitions is proposed,

$$\tau_{\text{SC}} = \frac{\tau_{\text{SC}}(\beta = 1) + \tau_{\text{SC}}(\beta = 2)}{2},$$

and it is the choice used in this work.

## 5 NUMERICAL TESTS

### 5.1 Numerical simulation of a supersonic flow past a forward facing step

This is classical test case for 2D numerical codes for supersonic flows and a detailed description can be found in [14]. It consists in a supersonic flow entering an infinite long wind tunnel with a step. The interaction of the supersonic flow with the step and the tunnel walls creates a typical patter of shock reflections.

The numerical set up of the simulation is as follows (see Figure 3): the computational domain has dimensions $3 \times 1$ length units, and the step is $0.2$ length units high and it is located at $0.6$ length units from the left boundary. The flow is entering from the left boundary with a uniform horizontal Mach $3$ velocity and pressure $p = 1$. The tunnel walls are non-slip boundaries, while the right boundary is a free exit. The simulation is started setting the variables at the inlet boundary values. No particular treatment is done for the corner of the step, which represent a singular point and it is known for introducing numerical errors, and no positivity correction was needed for the test cases here illustrated. However, with strong refinement in the corner, a negative pressure was developed in a point few elements after the corner and would require a special treatment. This point is still under investigation.

For the SUPG stabilization, the simple $\tau_{SM}$ is used. For the shock capturing term, the form $\tilde{\tau}_{\text{SC1}}$ is retained. The time step $dt$ is linearly incremented at the beginning of the simulation, until reaching a final CFL=1. In order to compare the results with [16], four computational meshes are employed with decreasing element size $h = 1/40, 1/80, 1/160, 1/320$. The results for a simulation time $t = 4$ are depicted in Figure 4 for the density variable. A correct shock reflection pattern is obtained and the results are globally similar to the reference [16]. While the Mach steam is almost completely disappeared in the finest mesh, as expected, there is no Kelvin-Helmholtz instabilities near the top of the domain. This could be due to the extra diffusion introduced by the triangular discretization, with respect the quadrangular used in [16], and it is still not fully understood.

In Figure 5 is also shown the shock profile on a horizontal line at $y = 0.1$, for the four meshes, with a comparison with the element size. It is possible to notice that there are no overshoots in the solution across the discontinuity and the shock is spread across about four element widths, which is acceptable.

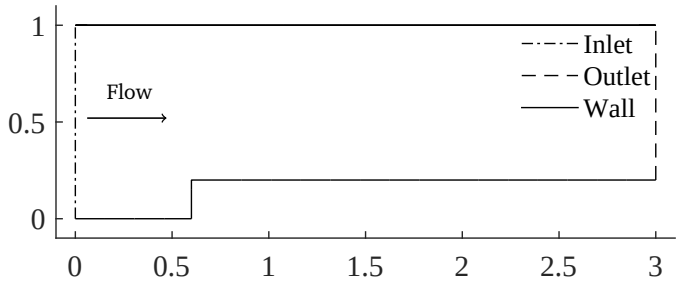



Figure 3: Forward facing step case: computational domain and boundary conditions.

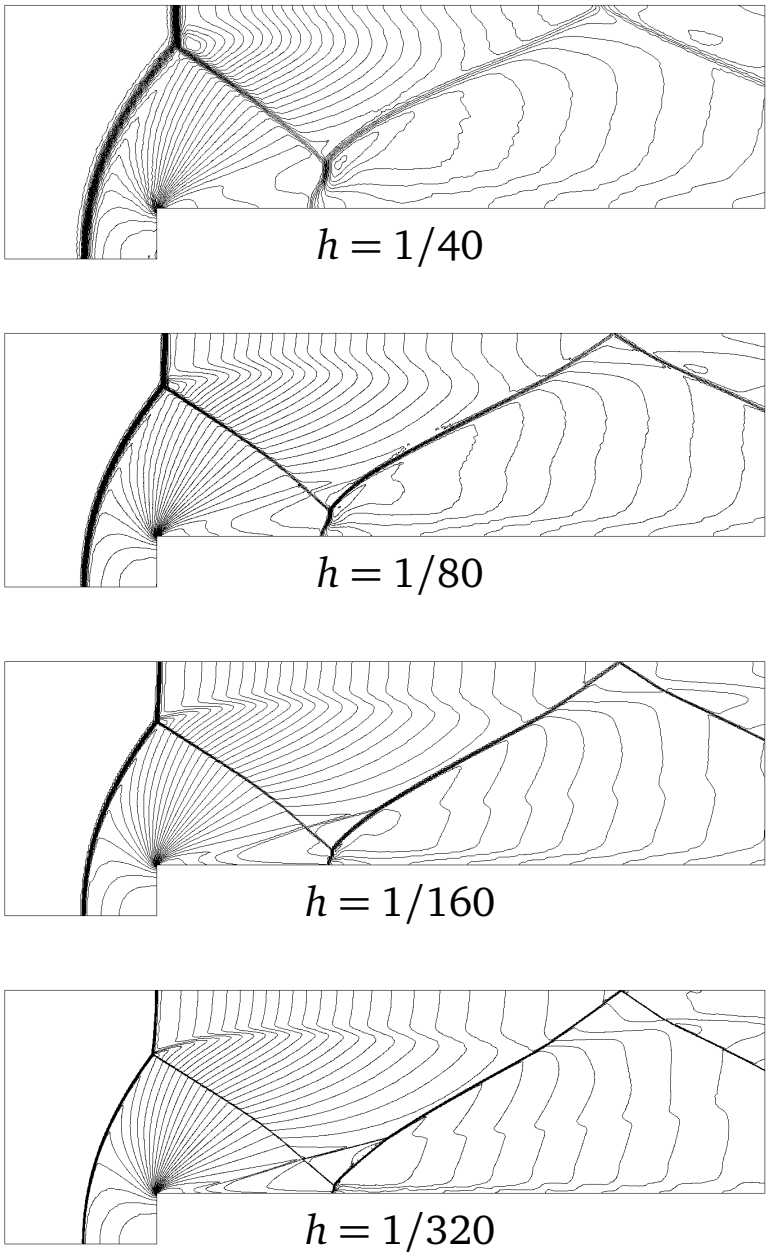



Figure 4: Forward facing step problem: density $\rho$, 30 equally spaced contour lines at $t = 4$.

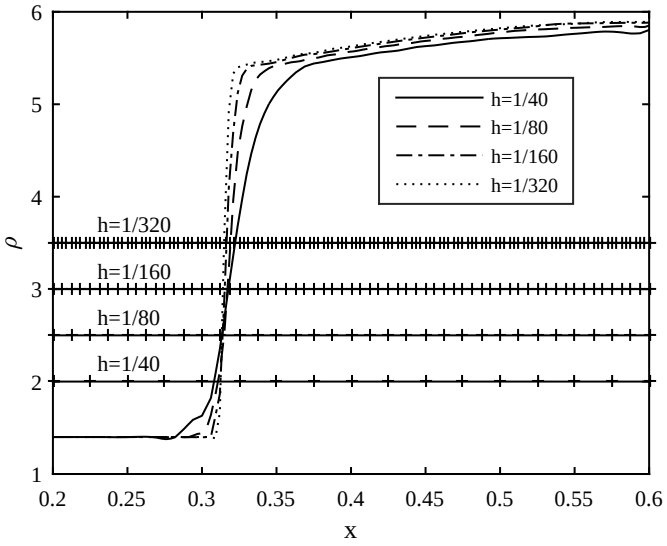



Figure 5: Forward facing step case: shock profile at a section $y = 0.1$ for the four meshes. The discretization is also displayed.

### 5.2 The Double Mach reflection test

The second test case considered is the Double Mach reflection test, also introduced in [14]. It represents the impinging of a strong shock wave, at Mach $10$, against an oblique wall, which produce a typical reflected shock with a curved shape, and a bubble of denser gas. The set up of the test is the following, see Figure 6 : the computational domain is a box of dimensions $3 \times 1$, and a wall boundary condition is set in the lower boundary starting from $x = 1/6$. The right boundary and the remaining of the lower boundary (from $x = 0$ to $x = 1/6$) is set to an outflow boundary, while the left and the top boundaries are inlet boundaries. In the upper boundary, a time variable condition is set, simulating the traveling of the shock wave to the right. In the initial condition, the shock wave creates an angle of $60^\circ$ from the $x$ axis and impinges the wall at $x = 1/6$. Post-shock conditions are set to the right of the wave, pre-shock to the left.

In order to avoid spurious oscillation of the solution in the first steps of the simulation, the fields are initialized with a smooth shock profile in the domain, with the discontinuity smeared over four elements. Moreover, the time step is increased with a cubic ramp until reaching a CFL=1 in $100$ time steps.

For the SUPG stabilization, similar results are obtained for $\tau_{SM}$ and $\tau_{Tz}$. The shock-capturing stabilization however resulted more critical in this case: the parameter $\tilde{\tau}_{\text{SC1}}$ used in the previous case resulted too diffusive, hence the form $\tilde{\tau}_{\text{SC2}}$ is used instead. As in [16], four meshes are considered for comparison purpose, with element size $h = 1/60, 1/120, 1/240, 1/480$. Again, a stable solution is obtained with a sharp shock resolution. However, the shock patterns inside the bubble is not obtained, and further investigation are needed. A positivity correction was used in this case: the value of the internal energy is artificially set to $10^{-5}$ in those Gauss points where a negative value is obtained.

## 6 CONCLUSIONS

In this paper is presented a finite element method based on Powell-Sabin (PS) splines, for the solution of the 2D compressible Euler equations in supersonic regime. The mathematical and geometrical tools to define the PS shape functions are presented, in particular, the subdivision

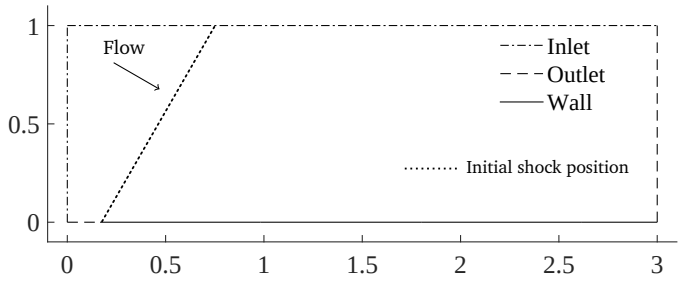



Figure 6: Double Mach reflection case: computational domain and boundary conditions.

of the elements that allows to define the PS splines as piecewise second order polynomials in each sub-triangle, the definition of the PS points and the control triangles.

A stabilized formulation is adopted, with SUPG and shock-capturing term. A simple isotropic artificial diffusion technique is adopted for the shock capturing term. Two classical examples are shown: the Woodward and Colella forward-facing step with Mach 3 flow and the Double Mach reflection test. Satisfactory results are obtained in terms of stabilization of the solution and resolution of the discontinuities. Further investigations are needed to understand the differences with the results with the reference solution and to set up an efficient positivity correction procedure.

## REFERENCES

[1] T. J. R. Hughes,J. A. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement, Comput. Methods Appl. Mech. Eng. 194 (39-41) (2005) 4135–4195.

[2] Y. Bazilevs, V.M. Calo, J. A. Cottrell, T. J. R. Hughes, A. Reali, G. Scovazzi, Variational multiscale residual-based turbulence modeling for large eddy simulation of incompressible flows, Comput. Methods Appl. Mech. Eng. 197 (1-4) (2007) 173–201.

[3] Y. Bazilevs, V.M. Calo, J. A. Cottrell, J. A. Evans, T. J. R. Hughes, S. Lipton and M. A. Scott and T. W. Sederberg, Isogeometric analysis using T-splines, Comput. Methods Appl. Mech. Eng. 199 (5-8) (2010) 229–263.

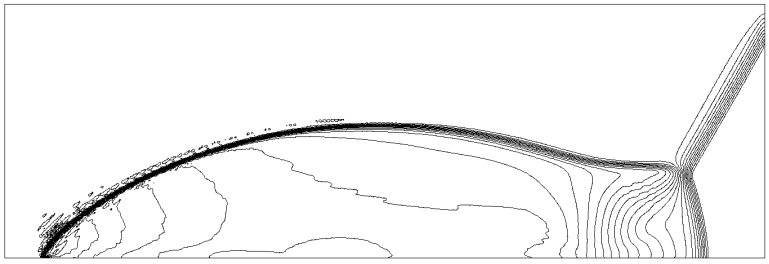
[4] M. J. Lai, L. L. Schumaker, Spline Functions on triangulations, Cambridge University Press, 2007.

[5] M. J. D. Powell, M. A. Sabin, Piecewise quadratic approximations on triangles, ACM Trans. Math. Softw. 3 (4) (1977) 316–325.

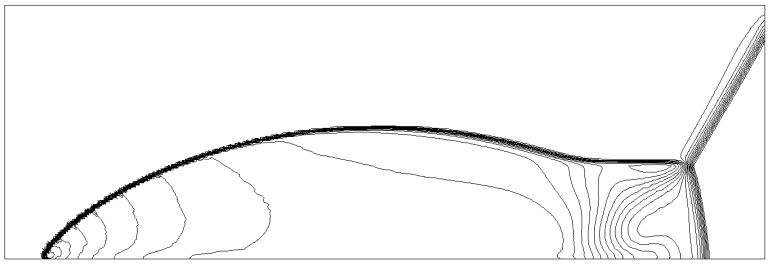
[6] N. Jaxon , X. Qian, Isogeometric analysis on triangulations, Computer-Aided Design 46 (2014) 45–57.

[7] H. Speleers, C. Manni, F. Pelosi, M. L. Sampoli, Isogeometric analysis with Powell-Sabin splines for advection-diffusion-reaction problems, Comput. Methods Appl. Mech. Eng., 221-222 (2012) 132–148.

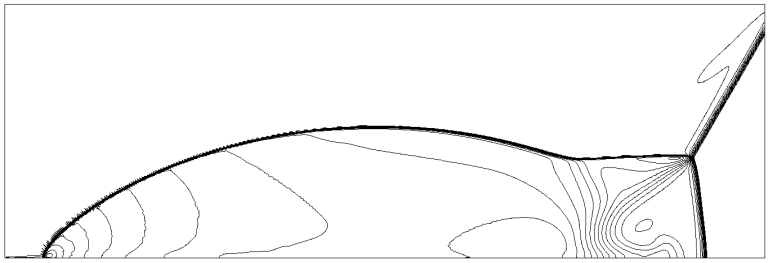
[8] P.Dierckx, On calculating normalized Powell-Sabin B-splines, Comput. Aided Geom. D., 15 (1) (1997) 61–78.



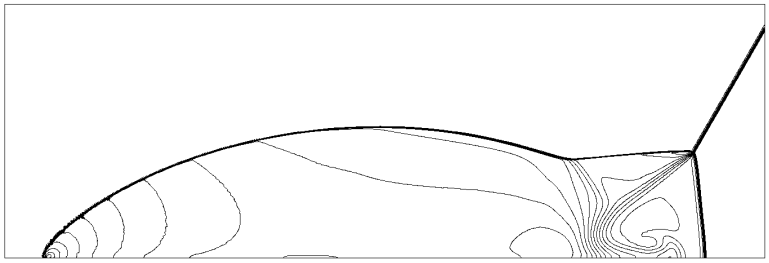
$h = 1/60$



$h = 1/120$



$h = 1/240$



$h = 1/480$

Figure 7: Double Mach reflection problem: density $\rho$, 30 equally spaced contour lines at $t = 0.2$.

[9] J. O'Rourke, A. Aggarwal, S. Maddila, M. Baldwin, An optimal algorithm for finding minimal enclosing triangles, J. Algorithm., 7 (2) (1986) 258–269.

[10] H. Speleers, P. Dierckx, S. Vandewalle, Powell-Sabin splines with boundary conditions for polygonal and non-polygonal domains, J. Comput. Appl. Math., 206 (1) (2007) 55–72.

[11] A. N. Brooks and T. J. R. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Eng., 32 (1990) 199–259.

[12] C. Wervaecke, H. Beaugendre, B. Nkonga, A fully coupled RANS Spalart-Allmaras SUPG formulation for turbulent compressible flows on stretched-unstructured grids, Comput. Methods Appl. Mech. Eng., 233-236 (2012) 109–122.

[13] T. E. Tezduyar, M. Senga, Stabilization and shock-capturing parameters in SUPG formulation of compressible flows, Comput. Methods Appl. Mech. Eng., 195 (2006) 1621–1632.

[14] P. Woodward, P. Colella, The numerical simulation of two-dimensional fluid flow with strong shocks, J. Comput. Phys., 54 (1) (1984) 115–173.

[15] J. Donea, A. Huerta, Finite element methods for flow problems, John Wiley & Sons, Chichester (UK), 2003.

[16] B. Cockburn, C. W. Shu, The Runge-Kutta discontinuous Galerkin method for conservation laws V, J. Comput. Phys., 141 (1998) 199–224.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL INVESTIGATION OF METASTABLE CONDENSING FLOWS WITH AN IMPLICIT UPWIND METHOD

**L. Azzini, T.P. van der Stelt, M. Pini**

Propulsion & Power group
Faculty of Aerospace engineering
Delft University of Technology
Kluyverweg, 1, Delft, Netherlands
contact e-mail: m.pini@tudelft.nl

**Keywords:** Two-phase flows, method of moments, nozzle expansion, metastable condensation, upwind flux

**Abstract.**

*This work proposes and assesses two numerical models for solving high-speed condensing flows in metastable conditions. Each model involves a set of governing equations (mass, momentum, and energy) for the mixture or the continuum phase, i.e. the vapor, and two additional transport equations to characterize the dispersed phase. Such relations are formulated through the so-called method of moments that allows to represent the wetness fraction and the number of droplets of the liquid.*

*The transport relations are discretized in space by means of a new coupled up-wind scheme. A segregated implicit time integration strategy is exploited to hasten the convergence of the full system to steady-state. The performance and accuracy of both models are thoroughly investigated on a reference quasi-1D problem and confronted against experimental data and more advanced two-phase flow models.*

*Results show that experimental observations are adequately predicted, especially concerning the droplets dimension. It is additionally inferred that the new upwind flux is beneficial to improve robustness of the underlying numerical methods. Finally, it is demonstrated that the continuum phase model outperforms the mixture one in terms of numerical stability and computational cost, thereby making it very promising for the extension to multi-dimensional problems.*

## NOMENCLATURE

| Acronyms | |
|---|---|
| EoS | Equation of State |
| ORC | Organic Rankine cycle |
| GDE | General Dynamic Equation |

| Symbols | |
|---|---|
| $A$ | Numerical flux jacobian |
| $a1$ | A matrix component |
| $a2$ | $\|A\|$ matrix component |
| $A_c$ | Cross sectional area |
| $b1$ | A matrix component |
| $b2$ | $\|A\|$ matrix component |
| $c1$ | A matrix component |
| $c2$ | $\|A\|$ matrix component |
| $e$ | Internal energy |
| $F$ | Numerical flux |
| $f$ | Distribution function |
| $G$ | Growth rate |
| $h$ | Enthalpy |
| $J$ | Nucleation rate |
| $L_{eig}$ | Left eigenvectors matrix |
| $M$ | Mach number |
| $m$ | Mass |
| $N$ | Droplet number per unit total mass |
| $P$ | Pressure |
| $Q$ | Source terms vector |
| $R$ | Droplet radius |
| $R_{eig}$ | Right eigenvectors matrix |
| $S$ | Source term (mass balance) |
| $T$ | Temperature |
| $t$ | Time |
| $U$ | Solution vector |
| $V$ | Generic vector (residual jacobian) |
| $v$ | Velocity |
| $x$ | Space coordinate |
| $Y$ | Liquid mass fraction |

| Greek letters | |
|---|---|
| $\alpha$ | Steam volume fraction |
| $\beta$ | Empirical parameter (in $J$) |
| $\gamma$ | Heat capacity ratio |
| $\Lambda$ | Eigenvalues matrix |
| $\mu_j$ | Moment of order -j |
| $\rho$ | Density |

| Subscripts | |
|---|---|
| * | Critical, correspondent to the Gibbs free energy maximum |
| 0 | Total property |
| average | Average properties on the domain (integral approximation) |
| c | Continuum phase |
| d | Dispersed phase |
| inlet | Domain inlet |
| m | Mixture |
| outlet | Domain outlet |

## 1 INTRODUCTION

Metastable condensation of transonic and supersonic flows is subject of abundant numerical and experimental research. The accurate characterization of homogeneous condensation, which occurs when liquid particles (nuclei) can durably form in a supersaturated high-speed vapor, is extremely relevant in a variety of industrial processes, whereby the release of latent heat produces unwanted fluid-dynamic penalties. Typical examples are homogeneous condensation in low-pressure steam turbines [1] or in converging-diverging nozzles [2].

The numerical computation of condensing flows can be approached in three different ways [3] commonly classified in i) Eulerian-Lagrangian (EL) methods, ii) Eulerian-Eulerian (EE) methods, and iii) method of moments (MM). The key difference among them lies in the treatment of the additional transport equations used to describe the second (dispersed) phase. Notably, the EL and EE approaches model the entire spectrum of droplet sizes, while the MM accounts only for the low-order statistical moments of the size distribution.

A first comparison amongst the three models was performed in [3]. The study pointed out that the EL method features the highest accuracy at the expense of high computational cost. Moreover, the extension to unsteady simulations and flows characterised by slip between phases is particularly challenging. The full EE method has a similar accuracy compared to EL. However, it is affected by severe numerical instabilities that may compromise the convergence rate. Results showed that the cost of simulations with the EE was 50% more demanding than with the EL.

The MM, originally introduced by Hill [4] and increasingly adopted in many research studies [5, 6, 7], exhibited substantial computational efficiency gain with respect to EL and EE methods, thus is deemed the most suitable model for multi-dimensional calculations in engineering applications. However, the accuracy obtained may be considerably lower, especially regarding the droplets dimension. [3] reported a discrepancy of about 20% on the average droplets radius compared to experimental data.

In the context of MM, two formulations can be devised to describe the motion of the two-phase mixture. The former, referred to as mixture formulation [8], expresses the mass, momentum, and energy balance in terms of mixture properties, whereas the latter, referred to as continuum phase formulation, specifies the conservation laws in terms of vapor phase. For this last case, no examples are found in the literature, and this work refers to the framework devised in [9] for the EE method.

The aim of this work is to compare the two formulations in order to establish an efficient numerical framework for the resolution of multi-dimensional condensing flows. To this purpose, a quasi 1-D model is employed. The transport equations are discretized in space by a novel upwind scheme while the whole system of conservation laws is integrated in time using an implicit method to enhance the numerical stability as well as the convergence to steady-state. The results are carried out on a widely popular nozzle test case in which non-equilibrium condensation takes place.

The paper is structured as follows: the first section describes the governing equations and the necessary closure relations. The second section reports the numerical resolution adopted. Finally, the last section reports the validation of the models and debates the results of the MM against reference solutions obtained with the EL and the EE.

## 2 GOVERNING EQUATIONS

This work considers quasi 1-D condensing flows. In principle, eight governing equations are needed to obtain the averaged properties of the liquid and vapor phase, namely their thermodynamic states, the velocities, the liquid mass fraction and the droplet radius.

For condensing flows with limited liquid mass fractions, the following assumptions proved to be adequate: i) the liquid and vapor are in mechanical equilibrium, ii) no slip between the two phases, and iii) the temperature of the dispersed phase is determined through a capillarity model. As a result, three conservation laws supplemented by two transport equations for the dispersed phase properties are sufficient to characterize the mixture.

### 2.1 Conservation laws

The mass, momentum, and energy balance for the continuum phase are written as

$$\frac{\partial \rho_c}{\partial t} + \frac{\partial \left(\rho_c v_c\right)}{\partial x} = S_c - \rho_c v_c \frac{\partial A_\mathrm{c}}{\partial x}\frac{1}{A_\mathrm{c}}, \tag{1}$$

$$\frac{\partial \rho_c v_c}{\partial t} + \frac{\partial \left(\rho_c v_c^2 + p_c\right)}{\partial x} = S_c v_c - \rho_c v_c^2 \frac{\partial A_\mathrm{c}}{\partial x}\frac{1}{A_\mathrm{c}}, \tag{2}$$

$$\frac{\partial (\rho_c e_{0,c})}{\partial t} + \frac{\partial \left(\rho_c h_{0,c}\right)}{\partial x} = S_c h_{0,d} - \rho_c h_{0,c} \frac{\partial A_\mathrm{c}}{\partial x}\frac{1}{A_\mathrm{c}}, \tag{3}$$

where $\rho_c$ is the phase density, $v_c$ the velocity, $P_c$ the pressure, $e_{o,c}$ the total internal energy, $h_{o,c}$ the total enthalpy, $A_c$ is the cross sectional area and $S_c$ is the source term accounting for mass exchange through the liquid-vapor interface, defined as

$$S_c = -\rho_m \frac{3Y}{R}\frac{\partial R}{\partial t}, \tag{4}$$

in which $R$ is the averaged droplet radius and $Y$ is the liquid mass fraction. The mixture equations can be retrieved by substituting in eq.(1) the mixture properties and neglecting $S_c$. For the sake of conciseness, mixture and the continuum phase approaches are referred to as model (a) and (b), respectively, hereafter.

Important differences arise between the two models: model (b) may suffer from more severe stiffness due to condensation source terms. On the contrary, in model (a) the thermodynamic non-equilibrium of the vapor phase renders the computation of the mixture properties iterative [8]. This procedure entails an extra computational burden and may affect the solver robustness. Eventually, note that the validity of model (b) is restricted to vapors containing negligible liquid volume fractions [9].

### 2.2 Method of Moments

As aforementioned, the method of moments replaces the full droplet size distribution with its low-order statistical moments. The generic moment of order -j is defined as

$$\mu_j = \int_0^\infty R^j f dr, \tag{5}$$

where $f$ is the distribution function, i.e. the radial derivative of the droplet number density, such that the total number of droplets $N$ is equal to

$$N = \int_0^\infty f dr. \tag{6}$$

As described in [5], the conservation law for each moment stems from the GDE (General Dynamic Equation) and eq.(5), obtaining

$$\frac{\partial}{\partial t}(\rho_m \mu_j) + \frac{\partial}{\partial x}(\rho_m \mu_j v_m) = j \cdot \int_0^\infty \rho_m R^{j-1} G f dr + \rho_m J(R_*) R_*^j, \tag{7}$$

where $G$ is the growth rate, $R_*$ the critical radius, and $J$ the nucleation rate.

These equations are usually solved through the method proposed in [4]. Nevertheless, though the model well correlates the experimental data for the continuum phase, as reported in [3], the method involves the resolution of four equations instead of the two theoretically required. Moreover, $G$ can be only expressed as linear combination of $R$ in the form

$$G = k_1 R + k_2, \tag{8}$$

where $k_1$ and $k_2$ are two parameters depending on the thermodynamic properties of the two phases. As outlined in [3],[7], the discrepancy between the experimental and numerical droplet averaged radius is larger than 20%.

For these reasons, this work adopts a 2-equations method. Thus, the transport equations for the moments of order -0 and -3 are finally written as

$$\frac{\partial}{\partial t}(\rho_m \mu_0) + \frac{\partial}{\partial x}(\rho_m \mu_0 v_m) + \rho_m \mu_0 v_m \frac{\partial A_c}{\partial x}\frac{1}{A_c} = \rho_m J(R_*), \tag{9}$$

$$\frac{\partial}{\partial t}(\rho_m \mu_3) + \frac{\partial}{\partial x}(\rho_m \mu_3 v_m) + \rho_m \mu_3 v_m \frac{\partial A_c}{\partial x}\frac{1}{A_c} = \rho_m J(R_*) R_*^3 + 3\rho_m R^2 G N, \tag{10}$$

where $\mu_0$ and $\mu_3$ are

$$\mu_0 = \int_0^\infty f dr = N, \tag{11}$$

$$\mu_3 = \int_0^\infty R^3 f dr = R^3 \int_0^\infty f dr = N R^3. \tag{12}$$

The remaining terms in eq. (9) and(10), namely the growth rate $G$, the nucleation rate $J$, the critical radius $R_*$, are taken from [7].

### 2.3 Thermophysical model

The thermodynamic model used for the continuum phase is the polytropic Van der Waals EoS (Equation of State), that allows to account for metastabilities in the two-phase region. The enthalpy as well as the saturation temperature and pressure are taken from [11], while the density and the capillarity model are as in [9]. Finally, the surface tension is given by [12], whereas viscosity and thermal conductivity are determined through the Chung's model [13].

The thermodynamic properties for the mixture can be retrieved following the procedure in [8] by using the quasi-Newton algorithm. Furthermore, the speed of sound of the mixture is estimated through the expression

$$c_m = \left[ \left(1 - \frac{\rho_m}{\rho_d} Y\right) \frac{1}{c_c^2} + \left(\frac{\rho_m}{\rho_d} Y\right) \frac{1}{c_d^2} \right]^{-\frac{1}{2}}. \tag{13}$$

Appendix A shows the procedure to obtain such relation.

## 3 NUMERICAL METHOD

The balance equations for either mixture or continuum phase are discretized using a cell-centered finite volume scheme of first order of accuracy [10]. The source terms are directly incorporated in the numerical flux without special treatment. On the contrary, the transport equations are discretized by using an upwind scheme specifically conceived in this study, which is further detailed in the following. The full set of governing equations is advanced in time by an implicit, segregated strategy. More precisely, at each iteration the mass, momentum, and energy equations are solved assuming frozen droplet properties.

### 3.1 Upwind scheme for transport equations

As can be observed, the last term of eq.(10) is a function of the radius $R$ of the dispersed phase. Since $R$ is in turn a combination of $\mu_0$ and $\mu_3$ through eq.(12) it can be readily incorporated in the left hand-side of eq.(10) by a simple algebraic manipulation as

$$\frac{\partial}{\partial t}(\rho_m \mu_3) + \frac{\partial}{\partial x}(\rho_m \mu_3 v_m) + \frac{\rho_m \mu_3 v_m}{A_c}\frac{\partial A_c}{\partial x} = \rho_m J(R_*) R_*^3 + \frac{\partial}{\partial x}\left(\int 3\rho_m R^2 G N dx\right). \quad (14)$$

The last integral is approximated as

$$\int 3 \cdot \rho_m R^2 G N dx = 3\left(\rho_m R^2 G N\right)_{average} \cdot (x - x_{inlet}) = \sum_{i=0}^{x_{out}} 3\left(\rho_m R^2 G N \frac{\Delta x}{x_{outlet} - x_{inlet}}\right)_i \cdot (x - x_{inlet}), \quad (15)$$

where $x_{inlet}$ and $x_{outlet}$ are the inlet and outlet abscissas of the domain. Concisely, the 2 transport equations can be written as

$$\frac{\partial}{\partial t}U + \frac{\partial}{\partial x}F = Q, \quad (16)$$

where

$$U = \begin{bmatrix} \rho_m \mu_0 \\ \rho_m \mu_3 \end{bmatrix} = \begin{bmatrix} \rho_m N \\ \rho_m N R^3 \end{bmatrix}, \quad (17)$$

$$Q = \begin{bmatrix} \rho_m J(R_*) \\ \rho_m J(R_*) R_*^3 \end{bmatrix} - U \cdot v_m \frac{\partial A_c}{\partial x}\frac{1}{A_c}, \quad (18)$$

$$F = \begin{bmatrix} \rho_m \mu_0 v_m \\ \rho_m \mu_3 v_m - \sum_{i=0}^{x_{out}} 3\left(\left(\rho_m \mu_0\right)^{\frac{1}{3}} \left(\rho_m \mu_3\right)^{\frac{2}{3}} G \cdot \frac{\Delta x}{x_{outlet} - x_{inlet}}\right)_i \cdot (x - x_{inlet}) \end{bmatrix}. \quad (19)$$

The final up-wind flux is given by

$$F_{\frac{i+j}{2}} = \frac{F_i + F_j}{2} - |A|_{\frac{i+j}{2}} \frac{U_j - U_i}{2}, \quad (20)$$

in which the physical flux $F$ is approximated for every cell -i as

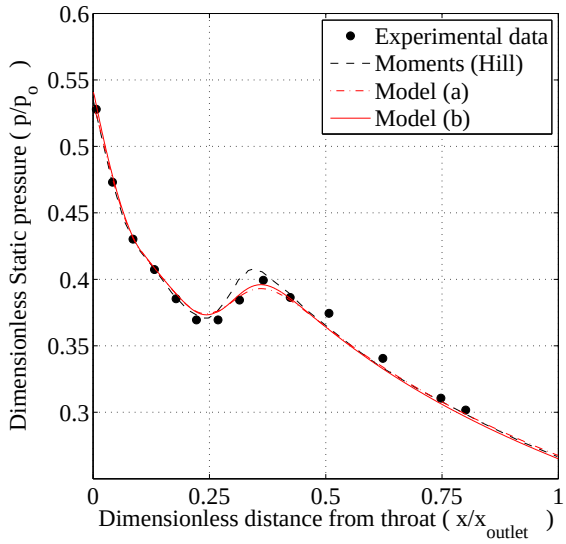
$$\sum_{j=1}^{i} 3\left(\rho_m R^2 G N \cdot \frac{\Delta x}{x_{outlet} - x_{inlet}}\right)_j \quad (21)$$

Appendix B reports the spectral decomposition of $A$ while Appendix C illustrates the derivation of the Jacobian.

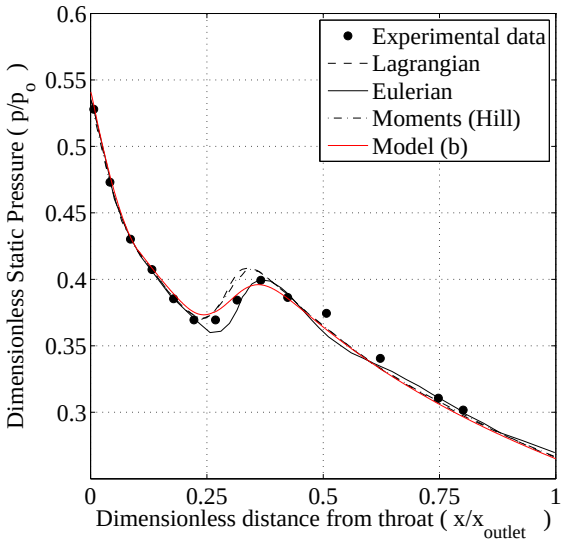
## 4 VALIDATION AND ASSESSMENT OF THE NUMERICAL MODELS

In this section, the results of model (a) and (b) are compared with the experimental data reported in [3] and with existing numerical solutions obtained with more sophisticated methods. The selected test case is the well-established Moore nozzle A, which is representative of supersonic wet-steam flows occurring in turbine flow passages. The shape of the profile is reported in Appendix D. All the calculations are performed on a 1000-cells grid.

Fig.1 shows the dimensionless pressure distribution along the nozzle for the present models, the Hill's method [5] and experimental observations. No significant differences are found, especially prior and after the onset of condensation.




(a) Comparison bewtween model (a), model (b) and Hill's method




(b) Comparison between model (b) and the three approaches reported in [3]

Figure 1: Pressure distribution along the nozzle, $x_{throat} = 0$

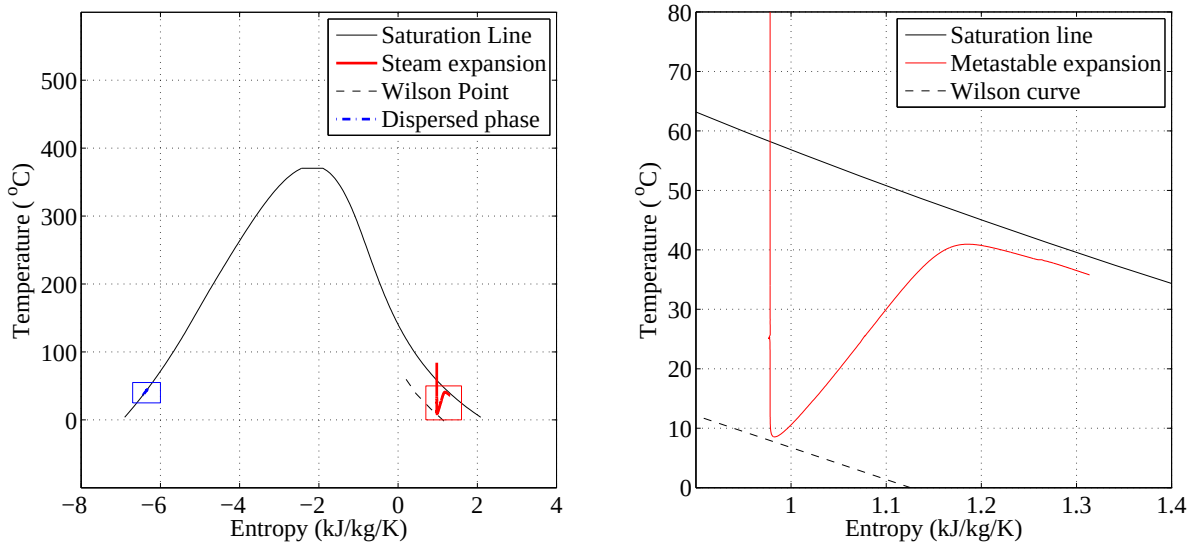
The pressure trends of both liquid and vapor are depicted in the *T-s* chart of Fig.2. The same diagram shows the experimentally derived Wilson points, extrapolated from [14] for the considered pressure range. For the sake of clarity, the same curves are also displayed in the *P-v* plane of Fig.3. When the steam subcooling is close to 50K (from Fig.2) and the continuum phase reaches the Wilson point, condensation starts rapidly. The sudden release of latent heat leads to a considerable static temperature rise of the continuum phase which is quickly brought back to thermodynamic equilibrium. The amount of heat released is proportional to the nucleation rate, see Fig.4, which shows a steep peak from 0 to $2 \cdot 10^{21} kg^{-1}s^{-1}$ in a narrow portion of the nozzle close to the throat.

A more in-depth physical explanation of the condensation phenomenon can be figured out by inspecting Eq.(22), valid for ideal compressible channel flow with heat addition [15]

$$\frac{dP}{P} = \frac{\gamma M^2}{1 - M^2} \left( \frac{dA_c}{A_c} - \left( 1 + \frac{\gamma - 1}{2} M^2 \right) \frac{dT_0}{T_0} \right), \tag{22}$$

in which $M$ is the mach number, $T_0$ is the total temperature, and $\gamma$ the specific heat ratio. In the early stage of condensation, the dominant term of Eq.(22) is the total temperature increase which causes a sudden static pressure rise, usually referred to as condensation shock wave.

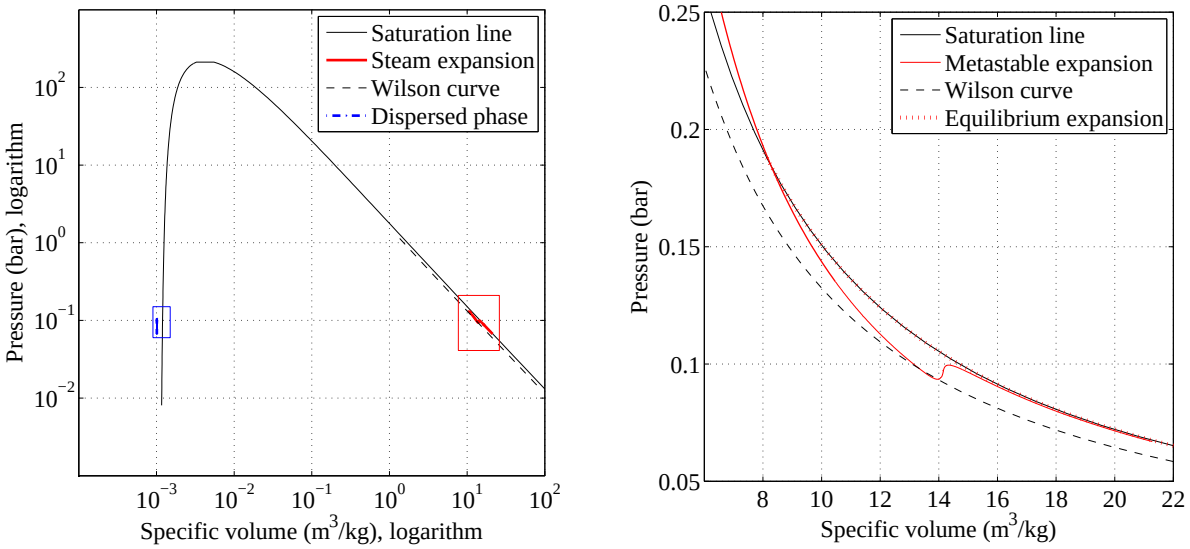
The flow then departs from metastable conditions to reach thermodynamic equilibrium. As the condensation proceeds, the nucleation rate $J$ falls down, therefore the release of latent heat and the total temperature variation reduce correspondingly. The flow motion is thus mainly governed by the area variation $dA_c$, which is positive for the given nozzle shape. Provided that the flow remains supersonic downstream of the shock, the expansion continues along the saturation line, as shown in both Fig.2 and 3.




(a) Liquid and vapor thermodynamic conditions

(b) Detail of the vapor expansion

Figure 2: Liquid and vapor expansion reported in the *T-s* diagram for model (a)




(a) Liquid and vapor thermodynamic conditions

(b) Detail of the vapor expansion

Figure 3: Liquid and vapor expansion reported in the *P-v* diagram for model (a)

As stated in [9], the liquid properties are barely dependent on the pressure in the specific thermodynamic range considered ($P < 0.25$ bar), see Fig.6. Furthermore, the temperature

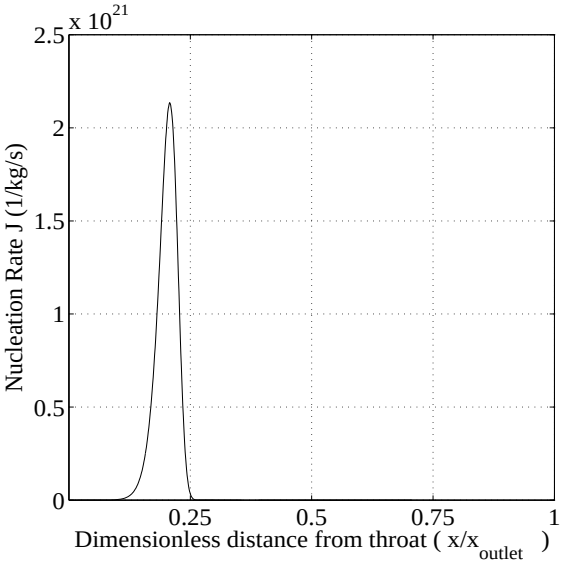



Figure 4: Nucleation rate along the nozzle, $x_{throat} = 0$

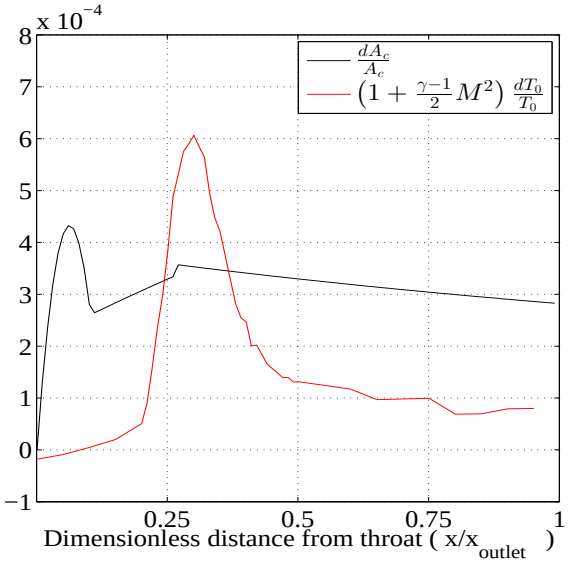



Figure 5: Main terms of eq.(22) along the nozzle, $x_{throat} = 0$

of the droplets $T_d$ is few degrees higher than the vapor, according to the capillarity model. Therefore, it can be inferred that the dispersed phase expansion is located close to the saturation line, with a slight temperature difference with respect to the steam, as proved by fig.2 and 3.

Fig.7 shows that the major deviations between the various models are located in the condensation region. In particular, the Lagrangian and the Hill's method overestimate the pressure local maximum, whereas the Eulerian simulation predicts a higher degree of subcooling. Note however that the pressure slope between the two pressure extrema predicted by the three models is very similar, suggesting a fairly close value of nucleation rate. On the other hand, model (a) and (b) tend to underestimate the pressure bump. The reason can be partly attributed to the inaccuracy of the thermodynamic model adopted for the calculations. As anticipated, this work makes use of the polytropic Van der Waals EoS, arguably less accurate than the model devised in [3].

The droplet averaged radius is recognized as remarkably challenging to detect for all two-phase computational models. Fig.8 depicts the droplet radius along the nozzle. Unexpectedly, the results of model (a) and (b) are in good agreement with the experimental data, with an deviation of 5% (0.095$\mu m$ diameter instead of 0.1$\mu m$ nominal) for model (b) and much lower for model (a). Conversely, the discrepancy for the Hill's method has been found higher than 20% (with a diameter of 0.077$\mu m$) and even worse for the Eulerian approach, which features deviations larger than 30%.

A numerical study was conducted to assess the convergence rate of model (a) and (b). The benchmark is the so-called single phase simulation, where it is supposed the vapor expanding in the nozzle without inception of condensation. Despite this simulation has no physical meaning, it is commonly utilized as reference for examining the convergence properties of the various methods. The simulations were carried out by discretizing the nozzle with 400 cells and employing the Euler explicit integration scheme with CFL number equal to 1. Convergence was achieved after reducing the residuals of all equations by five orders of magnitude. Table 1 reports the final results. Notwithstanding comparable accuracy, model (a) is three times more demanding than model (b). The rationale of this difference has been found in the iterative

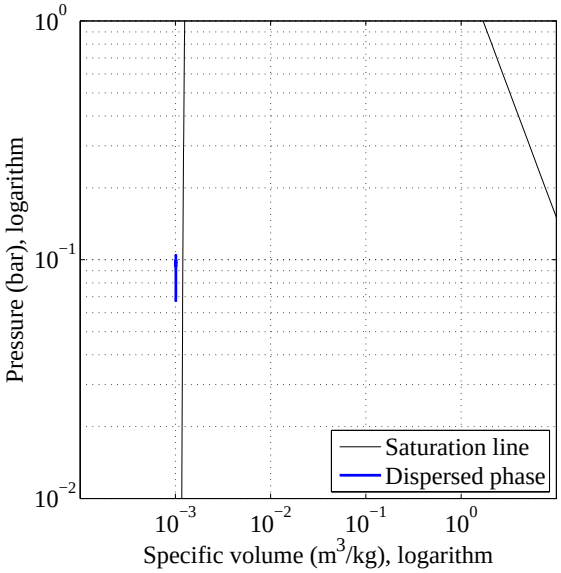



Figure 6: Detail of the liquid expansion from Fig. 3a, $x_{throat} = 0$

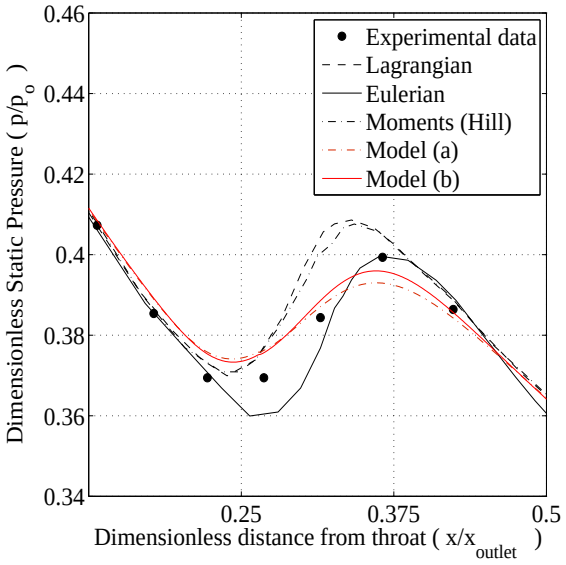
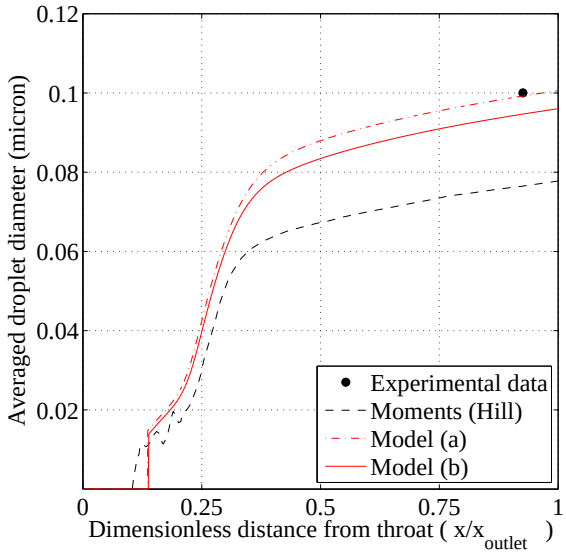
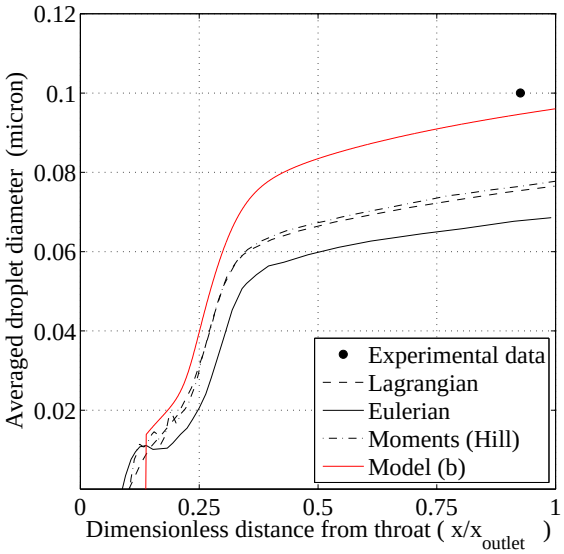



Figure 7: Detail of the pressure profile, comparison with [3], $x_{throat} = 0$




(a) Comparison between model (a), model (b) and Hill's method




(b) Comparison between model (b) and the three approaches reported in [3]

Figure 8: Averaged droplet diameter along the nozzle, $x_{throat} = 0$

algorithm, i.e. quasi-Newton method, adopted for computing the mixture properties.

A final analysis was carried out to investigate the numerical stability of both models. Several simulations were run on a mesh of 400-cells with the implicit method for progressively increasing CFL numbers. Table 2 lists the obtained results in terms of maximum allowable CFL number.

As expected, the maximum time step for the two-phase flow systems is significantly lower than for the benchmark. This is mainly due to the stiffness introduced by the onset of condensation which occurs at much faster time scale than wave propagation in the flow. On the other

| | Benchmark | Model (a) | Model (b) |
|---|---|---|---|
| Iterations | 5024 | 4700 | 4823 |
| Physical time (s) | 14 | 79 | 25 |

Table 1: Simulation time required for the benchmark (single-phase), model (a) and model (b), explicit time integration, CFL=1

| | Benchmark | Model (a) | Model (b) |
|---|---|---|---|
| Maximum CFL | 102 | 28 | 46 |

Table 2: Maximum allowable CFL for the benchmark (single-phase), model (a) and model (b). The computations are performance with implicit time integration and constant CFL

hand, the comparison shows that model (a) allows for a maximum CFL that is nearly half of the one of model (b).

## 5 CONCLUSIONS

In this study, two flow models based on the method of moments for quasi-1D condensing flows were compared to more advanced approaches based on the lagrangian, the full eulerian, and the Hill's formulation. The governing equations were formulated in terms of mixture or vapor phase (alternatively called continuum phase) properties while the method of moments was used to characterize the liquid wetness fraction and droplets number. A novel upwind flux was proposed to solve for the additional transport equations. Implicit time integration was adopted to enhance convergence to steady-state.

Both models were tested on a reference nozzle case for which experimental data exist. The results showed that both models are capable to predict the pressure distribution along the nozzle with accuracy similar to that of the more sophisticated flow models. Remarkably, the computed diameter of the droplets resulted very close to the experimental evidences.

Lately, a numerical study was conducted to assess the computational cost as well as the numerical stability of the two models. It was inferred that the introduction of the up-wind flux for the moments relations provided high robustness to the underlying numerical method. However, the analysis revealed that the mixture model is approximately three times more demanding than continuum phase model if solved by adopting the same CFL number and up to six times less efficient for maximum CFL number warranting stability. The main motivation was found to be associated with the iterative procedure used to retrieve the mixture properties.

Envisaged steps forward to the current work will include i) the investigation of a more efficient iterative algorithm, ii) the coupling with more complex equations of state iii) the validation of the models against condensing steam flows at high-pressure and ii) the extension to multi-dimensions to be pursued in the framework of an existing open-source CFD solver.

## REFERENCES

[1] J.B. Young, *Two-Dimensional, Non-Equilibrium, Wet-Steam Calculations for Nozzles and Turbine Cascades*, Journal of Turbomachinery 114, 569-579, 1992

[2] D.G. Elliot, E. Weinberg, *Acceleration of Liquids in Two-Phase Nozzles*, Technical report 32-987, National Aeronautics and Space Administration, 1968

[3] A.J. White, *A comparison of modelling methods for polydispersed wet-steam flow*, International journal for numerical methods in engineering 57, pp. 819834 (DOI: 10.1002/nme.705), 2003

[4] P.G. Hill, *Condensation of Water Vapour during Supersonic Expansion in Nozzles*, Journal of Fluid Mechanics, Vol. 25, part 3, pp. 593-620, 1966.

[5] F. Put, *Numerical Simulation of Condensation in Transonic Flows*, Thesis University of Twente, Enschede, 2003

[6] A.G. Gerber, A. Mousavi, *Application of quadrature method of moments to the polydispersed droplet spectrum in transonic steam flows with primary and secondary nucleation*, Applied Mathematical Modelling 31, pp. 15181533, 2007

[7] M. Giordano, S.J. Hercus, P. Cinnella, *Effects of modelling uncertainties in condensing wet-steam flows through supersonic nozzles*, V European Conference on Computational Fluid Dynamics ECCOMAS, Lisbon, Portugal,14-17 June 2010

[8] S. Dykas, W. Wróblewski, *Numerical modelling of steam condensing flow in low and high-pressure nozzles*, International Journal of Heat and Mass transfer 55, 6191-6199, 2012

[9] A.G. Gerber, M.J. Kermani, *A pressure based EulerianEulerian multi-phase model for non-equilibrium condensation in transonic steam flow*, International Journal of Heat and Mass Transfer 47, pp. 22172231, 2004

[10] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhliuser Verlag, Basel, Switzerland, 1992

[11] W. Wagner et al., *The IAPWS Industrial Formulation 1997 for the Thermodynamic Properties of Water and Steam*, ASME Journal of Engineering for Gas Turbines and Power 122, pp. 150182, 2000

[12] N.B. Vargaftik, B.N. Volkov, L.D. Voljak , *International Tables of the Surface Tension of Water*, Journal of chemical Physics Ref.Data12, 817, 1983

[13] B.E. Poling, J.M. Prausnitz, J.P. O'Connell , *The Properties of Gases and Liquids*, 5th edition, McGraw-Hill, 2001

[14] V.N. Gorbunov, Y.A. Ryzhov, U.G. Pirumov, *Non equilibrium Condensation in High-speed Gas Flows*, Gordon and Breach Science Publishers, 1989

[15] E.M. Greitzer, C.S. Tan, M.B. Graf, *Internal Flow, Concept and Applications*, Cambridge Engine Technology Series (No. 3), Cambridge University Press, 2004

## A Thermodynamic derivatives for a mixture

The speed of sound is defined as

$$c_m = \left(\frac{\partial P}{\partial \rho_m}\right)_{s_m}^{\frac{1}{2}}. \tag{23}$$

The introduction of the continuum phase volume title $\alpha_c$ and the use of the derivation properties allow to rewrite eq.(23) as

$$c_m = \left(\frac{\partial \rho_m}{\partial P}\right)_{s_m}^{-\frac{1}{2}} = \left(\frac{\partial \left(\alpha_c \rho_c + (1-\alpha_c)\,\rho_d\right)}{\partial P}\right)_{s_m}^{-\frac{1}{2}}, \tag{24}$$

thus,

$$c_m = \left[\alpha_c \left(\frac{\partial \rho_c}{\partial P}\right)_{s_m} + (1-\alpha_c)\left(\frac{\partial \rho_d}{\partial P}\right)_{s_m} + (\rho_c - \rho_d)\left(\frac{\partial \alpha_c}{\partial P}\right)_{s_m}\right]^{-\frac{1}{2}}, \tag{25}$$

Two simplifications are made: i) the derivative of $\alpha_c$ is neglected and ii) the two derivatives at $s_{mix}$ constant are approximated with the derivative of each phase, at constant $s_c$ and $s_d$ respectively. Furthermore, the volume title $\alpha_c$ is defined as

$$\alpha_c = \frac{V_c}{V_m} = 1 - \frac{V_d}{V_m} = 1 - \frac{\rho_m}{\rho_d} Y. \tag{26}$$

Therefore, the final expression for the speed of sound is

$$c_m = \left[\left(1 - \frac{\rho_m}{\rho_d} Y\right)\left(\frac{\partial \rho_c}{\partial P}\right)_{s_c} + \left(\frac{\rho_m}{\rho_d} Y\right)\left(\frac{\partial \rho_d}{\partial P}\right)_{s_d}\right]^{-\frac{1}{2}}, \tag{27}$$

thus,

$$c_m = \left[\left(1 - \frac{\rho_m}{\rho_d} Y\right)\frac{1}{c_c^2} + \left(\frac{\rho_m}{\rho_d} Y\right)\frac{1}{c_d^2}\right]^{-\frac{1}{2}}. \tag{28}$$

## B Numerical flux derivatives for moments equations

### B.1 Spectral decomposition

The flux jacobian $A$ is determined from eq.(19) as

$$A = \left(\frac{\partial F}{\partial U}\right) = \begin{bmatrix} a_1 & 0 \\ b_1 & c_1 \end{bmatrix}, \tag{29}$$

in which

$$a_1 = \frac{\partial F(1)}{\partial U(1)} = v_m, \tag{30}$$

$$b_1 = \frac{\partial F(2)}{\partial U(1)} = -G\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} \Delta x \left(\rho_m \mu_0\right)^{-\frac{2}{3}} \left(\rho_m \mu_3\right)^{\frac{2}{3}}, \tag{31}$$

$$c_1 = \frac{\partial F(2)}{\partial U(2)} = -2G\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} \Delta x \left(\rho_m \mu_0\right)^{\frac{1}{3}} \left(\rho_m \mu_3\right)^{-\frac{1}{3}} + v_m. \tag{32}$$

As $A$ is lower triangular, the eigenvalues matrix $\Lambda$ is equal to

$$\Lambda = \begin{bmatrix} v_m & 0 \\ 0 & -2G\frac{x-x_{inlet}}{x_{outlet}-x_{inlet}}\Delta x\left(\rho_m\mu_0\right)^{\frac{1}{3}}\left(\rho_m\mu_3\right)^{-\frac{1}{3}} + v_m \end{bmatrix}. \tag{33}$$

The right and left eigenvector matrices $R_{eig}$ and $L_{eig}$ are

$$R_{eig} = \begin{bmatrix} \frac{v_m-c_1}{\sqrt{b_1^2+(v_m-c_1)^2}} & 0 \\ \frac{b_1}{\sqrt{b_1^2+(v_m-c_1)^2}} & 1 \end{bmatrix} \tag{34}$$

and

$$L_{eig} = \frac{1}{Det(R_{eig})}\begin{bmatrix} 1 & 0 \\ -\frac{b_1}{\sqrt{b_1^2+(v_m-c_1)^2}} & \frac{v_m-c_1}{\sqrt{b_1^2+(v_m-c_1)^2}} \end{bmatrix}, \tag{35}$$

where

$$Det(R_{eig}) = \left(\frac{v_m - c_1}{\sqrt{b_1^2 + (v_m - c_1)^2}}\right). \tag{36}$$

Therefore, $|A|$ is written as

$$|A| = R_{eig}|\Lambda|L_{eig} = \begin{bmatrix} a_2 & 0 \\ b_2 & c_2 \end{bmatrix} = \begin{bmatrix} v_m & 0 \\ \frac{b_1(v_m-|c_1|)}{v_m-c_1} & |c_1| \end{bmatrix}, \tag{37}$$

### B.2 Boundary conditions

The number of boundary conditions that can be imposed is given by the eigenvalue analysis. From eq.(33), the first eigenvalue $\lambda_1$ is always positive. On the other hand, $\lambda_2$ may become negative. In this case, the condition on $\mu_3$ must be imposed at the domain outlet, but no information are known a priori to set this value.

It is worth pointing out that the expression for $\lambda_2$ in eq.(33) contains also an information related to the mesh, i.e. $\Delta x$, and not to the thermodynamics itself.

Furthermore, from the physics of the problem, it is arguably reasonable to impose the liquid phase fraction at the domain inlet, and to extrapolate at the domain outlet the value coming from inside.

## C Residual jacobian determination

Eq.(37) can be substituted in eq.(19) obtaining

$$F_{\frac{i+j}{2}} = \frac{F_i + F_j}{2} - \frac{1}{2}\begin{bmatrix} a_2 \cdot U_j(1) \\ b_2 \cdot U_j(1) + c_2 \cdot U_j(2) \end{bmatrix} + \frac{1}{2}\begin{bmatrix} a_2 \cdot U_i(1) \\ b_2 \cdot U_i(1) + c_2 \cdot U_i(2) \end{bmatrix} = \frac{F_i + F_j}{2} - \frac{1}{2}V_j + \frac{1}{2}V_i. \tag{38}$$

Therefore, the flux derivatives are

$$\frac{F_{\frac{i+j}{2}}}{U_i} = \frac{1}{2}\frac{\partial F_i}{\partial U_i} + \frac{1}{2}\frac{\partial F_j}{\partial U_i} - \frac{1}{2}\frac{\partial V_j}{\partial U_i} + \frac{1}{2}\frac{\partial V_i}{\partial U_i}, \tag{39}$$

$$\frac{F_{\frac{i+j}{2}}}{U_j} = \frac{1}{2}\frac{\partial F_i}{\partial U_j} + \frac{1}{2}\frac{\partial F_j}{\partial U_j} - \frac{1}{2}\frac{V_j}{U_j} + \frac{1}{2}\frac{V_i}{U_j}. \tag{40}$$

The physical flux derivative has already been shown in eq.(29), thus

$$\frac{\partial F_i}{\partial U_i}, \frac{\partial F_j}{\partial U_j} = A_i, A_j. \tag{41}$$

Despite of what happens for the conventional single phase, the physical flux $F_i$ depends also on $U_j$, due to the summatory in eq.(19). Therefore

$$\frac{\partial F_i}{\partial U_j}(1, 1:2) = 0, \tag{42}$$

$$\frac{\partial F_i}{\partial U_j}(2,1) = -\frac{x_i - x_{inlet}}{x_{outlet} - x_{inlet}} \left( G\Delta x \left(\rho_m \mu_0\right)^{-\frac{2}{3}} \left(\rho_m \mu_3\right)^{\frac{2}{3}} \right)_j, \tag{43}$$

$$\frac{\partial F_i}{\partial U_j}(2,2) = -2\frac{x_i - x_{inlet}}{x_{outlet} - x_{inlet}} \left( G\Delta x \left(\rho_m \mu_0\right)^{\frac{1}{3}} \left(\rho_m \mu_3\right)^{-\frac{1}{3}} \right)_j. \tag{44}$$

The last terms in eq.(39) are developed as

$$\frac{\partial V_j}{\partial U_i}(1, 1:2) = 0, \tag{45}$$

$$\frac{\partial V_j}{\partial U_i}(2,1) = U_j(1)\frac{\partial b_2}{\partial U_i(1)} + U_j(2)\frac{\partial c_2}{\partial U_i(1)} \tag{46}$$

$$\frac{\partial V_j}{\partial U_i}(2,2) = U_j(1)\frac{\partial b_2}{\partial U_i(2)} + U_j(2)\frac{\partial c_2}{\partial U_i(2)} \tag{47}$$

and

$$\frac{\partial V_i}{\partial U_i}(1,1) = a_2, \tag{48}$$

$$\frac{\partial V_i}{\partial U_i}(1,2) = 0, \tag{49}$$

$$\frac{\partial V_i}{\partial U_i}(2,1) = U_i(1)\frac{\partial b_2}{\partial U_i(1)} + b_2 + U_i(2)\frac{\partial c_2}{\partial U_i(1)}, \tag{50}$$

$$\frac{\partial V_i}{\partial U_i}(2,2) = U_i(1)\frac{\partial b_2}{\partial U_i(2)} + c_2 + U_i(2)\frac{\partial c_2}{\partial U_i(2)} \tag{51}$$

The same procedure for eq.(40) leads to

$$\frac{\partial V_i}{\partial U_j}(1,1) = 0, \tag{52}$$

$$\frac{\partial V_i}{\partial U_j}(1,2) = 0, \tag{53}$$

$$\frac{\partial V_i}{\partial U_j}(2,1) = U_i(1)\frac{\partial b_2}{\partial U_j(1)} + U_i(2)\frac{\partial c_2}{\partial U_j(1)} \tag{54}$$

$$\frac{\partial V_i}{\partial U_j}(2,2) = U_i(1)\frac{\partial b_2}{\partial U_j(2)} + U_i(2)\frac{\partial c_2}{\partial U_j(2)} \tag{55}$$

and to

$$\frac{\partial V_j}{\partial U_j}(1,1) = a_2, \tag{56}$$

$$\frac{\partial V_j}{\partial U_j}(1,2) = 0, \tag{57}$$

$$\frac{\partial V_j}{\partial U_j}(2,1) = U_j(1)\frac{\partial b_2}{\partial U_j(1)} + b_2 + U_j(2)\frac{\partial c_2}{\partial U_j(1)}, \tag{58}$$

$$\frac{\partial V_j}{\partial U_j}(2,2) = U_j(1)\frac{\partial b_2}{\partial U_j(2)} + c_2 + U_j(2)\frac{\partial c_2}{\partial U_j(2)} \tag{59}$$

Finally, all the quantities $a_2, b_2, c_2$ are evaluated in

$$U_{\frac{i+j}{2}} = 0.5 \cdot (U_i + U_j), \tag{60}$$

thus, the missing derivatives are

$$\frac{\partial a_2, b_2, c_2}{\partial U_i} = \frac{\partial a_2, b_2, c_2}{\partial U_j} = \frac{\partial a_2, b_2, c_2}{\partial U_{\frac{i+j}{2}}} \cdot \frac{\partial U_{\frac{i+j}{2}}}{\partial U_{i,j}} = \frac{1}{2}\frac{\partial a_2, b_2, c_2}{\partial U_{\frac{i+j}{2}}}, \tag{61}$$

$$\frac{\partial a_2}{\partial U_{\frac{i+j}{2}}} = 0 \tag{62}$$

$$\frac{\partial b_2}{\partial U_{\frac{i+j}{2}}} = \frac{1}{v_m - c_1}\left((v_m - |c_1|)\frac{\partial b_1}{\partial U_{\frac{i+j}{2}}} - b_1\frac{c_1}{\partial U_{\frac{i+j}{2}}}\frac{|c_1|}{c_1}\right) + \frac{b_1\left(v_m - |c_1|\right)}{\left(v_m - c_1\right)^2} \cdot \frac{\partial c_1}{\partial U_{\frac{i+j}{2}}}, \tag{63}$$

$$\frac{\partial c_2}{\partial U_{\frac{i+j}{2}}} = \frac{\partial c_1}{\partial U_{\frac{i+j}{2}}}\frac{|c_1|}{c_1}, \tag{64}$$

in which

$$\frac{\partial b_1}{\partial U_{\frac{i+j}{2}}(1)} = \frac{2}{3}\left(\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} G\Delta x \left(\rho_m \mu_0\right)^{-\frac{5}{3}} \left(\rho_m \mu_3\right)^{\frac{2}{3}}\right)_{\frac{i+j}{2}}, \tag{65}$$

$$\frac{\partial b_1}{\partial U_{\frac{i+j}{2}}(2)} = -\frac{2}{3}\left(\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} G\Delta x \left(\rho_m \mu_0\right)^{-\frac{2}{3}} \left(\rho_m \mu_3\right)^{-\frac{1}{3}}\right)_{\frac{i+j}{2}}, \tag{66}$$

$$\frac{\partial c_1}{\partial U_{\frac{i+j}{2}}(1)} = -\frac{2}{3}\left(\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} G\Delta x \left(\rho_m \mu_0\right)^{-\frac{2}{3}} \left(\rho_m \mu_3\right)^{-\frac{1}{3}}\right)_{\frac{i+j}{2}}, \tag{67}$$

$$\frac{\partial c_1}{\partial U_{\frac{i+j}{2}}(2)} = -\frac{2}{3}\left(\frac{x - x_{inlet}}{x_{outlet} - x_{inlet}} G\Delta x \left(\rho_m \mu_0\right)^{\frac{1}{3}} \left(\rho_m \mu_3\right)^{-\frac{4}{3}}\right)_{\frac{i+j}{2}}. \tag{68}$$

## D Nozzle geometry

The nozzle area $A_c$ is given by

$$A_c(x) = \begin{cases} a_0 + b_0 x + c_0 x^2 + d_0 x^3 & 0.0 < x < 0.041837 \\ a_1 + b_1 x + c_1 x^2 + d_1 x^3 & 0.041837 < x < 0.10612 \\ a_2 + b_2 x + c_2 x^2 + d_2 x^3 & 0.10612 < x < 0.4 \end{cases}, \tag{69}$$

in which $x$ is expressed in $m$ and $A_c$ in $m^2$. All the missing coefficient are reported in the table below [7] .

| | a | b | c | d |
|---|---|---|---|---|
| 0 | $1.52{\cdot}10^{-2}$ | $-5.1996{\cdot}10^{-4}$ | $6.7416{\cdot}10^{-1}$ | -8.7727 |
| 1 | $1.533{\cdot}10^{-2}$ | $8.0338{\cdot}10^{-3}$ | $2.6189{\cdot}10^{-2}$ | $7.3488{\cdot}10^{-3}$ |
| 2 | $1.4926{\cdot}10^{-2}$ | $1.4733{\cdot}10^{-2}$ | $5.4451{\cdot}10^{-5}$ | $-2.0589{\cdot}10^{-4}$ |

Table 3: Nozzle profile coefficients

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# LEVEL SET METHOD FOR SIMULATING THE INTERFACE KINEMATICS: APPLICATION OF A DISCONTINUOUS GALERKIN METHOD

**Roozbeh Mousavi[1], Florian Kummer[2], Martin Oberlack[2], Peter F. Pelz[1]**

[1]Technische Universität Darmstadt, Institut für Fluidsystemtechnik
L1|01 465, Otto-Berndt-Straße 2, 64287 Darmstadt, Germany
roozbeh.mousavi@fst.tu-darmstadt.de

[2]Technische Universität Darmstadt, Fachgebiet für Strömungsdynamik
L1|01 331, Otto-Berndt-Straße 2, 64287 Darmstadt, Germany

**Keywords:** Discontinuous Galerkin Method, Level Set Method, Re-Initialization Equation

**Abstract.** *A Discontinuous Galerkin (DG) Method was applied for simulating the kinematics of deforming interfaces. The level set method was used as an interface capturing method. The numerical implementations were performed in the context of the in-house code BoSSS developed at the Chair of Fluid Dynamics in Darmstadt University of Technology, see [1, 2]. As the higher-order spectral representation of the variables in the DG method, results in a precise solution to the level set advection equation, the interface kinematics could be accurately simulated without having to solve the level set re-initialization equation. But the solution exhibits an appropriate hp-convergence only if the gradient of the level set function does not have any singularity over the domain of computation. For instance, the signed-distance level set function of a circle has a singular gradient at the center of the circle. As the smoothed Heaviside and Delta functions used in the multiphase flow calculations, are commonly expressed in terms of the level set function, the level set function needs to remain signed distance in order to keep a uniform smoothing width. The signed distance property of the level set function can be recovered by solving the level set re-initialization equation. In order to obtain a solution with a monotonicity preserving behavior, a Godunovs scheme was applied for approximating the Hamiltonian of the re-initialization equation. Moreover, a notable stability improvement was achieved by adding an artificial diffusion along the characteristic lines. The solution exhibits an appropriate hp-convergence and almost no spurious movement of the interface was detected.*

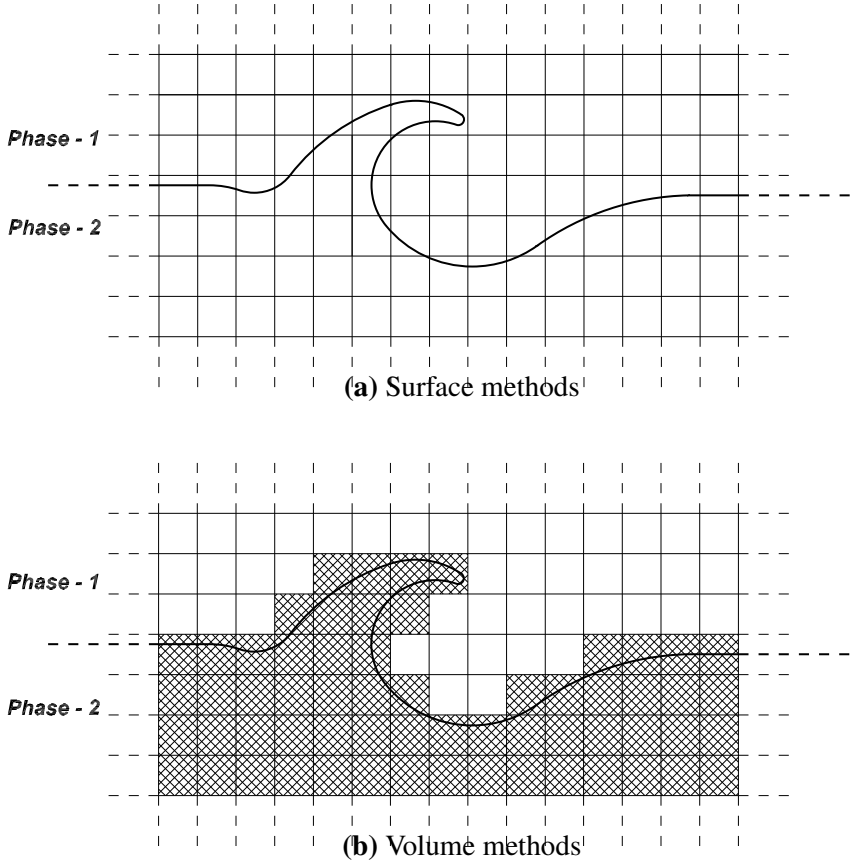
## 1 Introduction

The flows consisting of fluids with different properties, are termed as the multiphase flows. If the fluids are immiscible, they are separated by thin layers known as the interface. Interface is the region across which, the fluid properties as well as some of the flow variables are subjected to steep variations. In the context of the continuum mechanics, an interface is represented as a geometrical surface with zero thickness which may result in mathematical singularities, see [3]. A common approach to overcome the singularity problem is to make a diffuse interface assumption according to which, the fluid properties are supposed to vary smoothly across the interface, see [4].
A major issue in the numerical simulation of the multiphase flows is the manner of representing the interface and simulating its kinematics. Several methods with different levels of the accuracy and robustness have been proposed, among which, the level set method, see [5], is a robust one. The interface in the level set method is represented as the zero iso-value of a function which is called the level set function. The implicit representation of the interface provides an appropriate way for simulating the topological changes. Moreover, as the interface does not need to be reconstructed, this method is quite suitable when a precise calculation of the curvature is required. The accuracy of this method in simulating the interface kinematics is highly dependent on the preciseness of the numerical method applied for solving the corresponding advection equation, namely the level set advection equation. The discontinuous Galerkin (DG) method, see [6], is a modern technique providing a framework where a higher-order approximation can be implemented in a robust way. In this method, the variables are expressed in each cell in terms of an orthonormal basis polynomial space (OBPS). As in this method the in-cell variations are considered, an acceptable level of accuracy can be attained using a rather low spatial resolution. In this way, the total number of the numerical degrees of freedom is reduced although it is increased in each cell. The research presented in [7] is a well-known pioneer study on applying the DG method for solving the level set advection equation. As a consequence of the excellent accuracy they achieved, they claimed that the DG method is the best technique for solving the hyperbolic equations, such as the level set advection equation. But still a lack of Detailed $hp$-convergence analysis is notable in the literature. Although in the DG-based level set method the level set function does not need to be signed distance, this property is required for making a uniform diffusion thickness when a diffuse interface assumption is made, see [8, 9]. The signed distance property of a level set function can be maintained by solving an Eikonal equation termed as the level set re-initialization equation, see [10]. Application of the DG method for solving the re-initialization equation has been only considered in [9] yet. Although they have employed a set of the stabilization techniques successfully, a lack of performing a procedural error analysis is obvious in their publications. The present research is mainly motivated by the need of performing the mentioned lacks of the procedural error analyses concerning the solutions to the level set advection and re-initialization equations.

## 2 An Overview on the Numerical Techniques for Simulating the Interface Kinematics

The methods proposed for the numerical representation of the interface, can be classified into two categories, namely the surface methods and the volume methods, see [11]. The surface methods consider the interface itself as an object, either explicitly or implicitly. Whereas, the volume methods consider the phases at the either sides of the interface. Therefore, in the volume methods in order to obtain the normal vector to the interface, the interface needs to be reconstructed. As the interface curvature which is used to predict the surface tension effects, is

the divergence of the normal vector, any inaccuracy in calculating the normal vector intensively appears in the curvature. A common consequence of a non-precise prediction of the surface tension effects is the formation of a set of spurious vortical flows, see [12]. Hence, a surface method is often recommended when the surface tension effects is involved in the problem. The difference between these two types of methods is schematically demonstrated in figures 1a and 1b. The front tracking method introduced in [13], as well as the level set method introduced




**(a)** Surface methods

**(b)** Volume methods

**Figure 1** A schematic demonstration of the difference between the surface and volume methods for the numerical representation of an interface

in [14], are the common methods which can be classified under the category of the surface methods. The front tracking method explicitly represents the interface by a set of the connected massless particles. The particles are advected through the domain in a Lagrangian way together with a set of the conditions enforced on the interface. This method is very accurate if the interface is reconstructed by passing a spline of higher-degree over the particles. But the necessity of tracking a rather large number of particles makes this method expensive. Moreover, as the particles need to keep an optimum distance to each other, several particles are necessary to be added or removed during the simulation. In addition, a new reconnection procedure needs to be performed after any change in the configuration of the particles. Furthermore, for the simulation of the interface breakup or coalescence an ad-hoc procedure needs to be performed, see [15]. The level set method which is the subject of the present paper, was shortly introduced in section 1 and will be explained in detail subsequently. Although this method is highly robust, but there is no guaranty for the area/volume conservation if the numerical method applied to solve the

level set advection equation is not preciseness enough, see [5]. The particle level set method which was introduced in [16], implements a different way for maintaining the area/volume conservation. Each side of the interface in this method is assigned to a distinguished set of the massless particles which can be advected through the domain in a Lagrangian way. As the particles preserve the material characteristics in time, they can be used to reconstruct the interface in the regions where an area/volume loss (or gain) occurs.
The marker and cell method introduced in [17], as well as the volume of fluid method introduced in [18], are the common methods which can be classified under the category of the volume methods. The marker and cell method represents the phases on the either sides of the interface by a set of the massless particles which are advected through the domain in a Lagrangian way. The interface is then reconstructed in the multi-phase cells using the distribution density of the particles. This method is computationally very expensive due to requiring a large number of the particles. Moreover, needing to add additional particles for making an accurate simulation of the interface stretch, is an issue which reduces the robustness of the method. In the volume of fluid method, an indicator function is assigned to the phases at the either sides of the interface. The indicator function is commonly the volume fraction or the mass fraction of one of the phases. Therefore, it has a Heaviside distribution over the domain. The interface kinematics is simulated in this method by solving an advection equation for the indicator function. In the case of using a lower-order spatial discretization method, the Heaviside distribution of the indicator function is numerically smeared out. The region over which, this incorrect interface diffusion takes place, can be even developed by the velocity gradient in the direction normal to the interface. On the other hand, a higher-order numerical representation of the Heaviside distribution can lead to a numerical instability. Therefore, a lower-order method is used together with applying an interface reconstruction technique in order to prevent the development of the diffusion region. The common interface reconstruction techniques include reconstructing an interface of degree zero introduced in [18], and reconstructing an interface of degree $1$ introduced in [19]. Another common technique introduced in [20], is to compress the interface by adding a compression term to the advection equation. Although the implementation of the volume of fluid method is rather straightforward, the interface reconstruction techniques always reduce the accuracy of the curvature calculation. But the main advantage of this method is that the area/volume conservation is fulfilled. In [21], they used this property of the volume of fluid method, together with the ability of the level set method in a precise calculation of the curvature, for developing the idea of the coupled level set and volume of fluid method. The interface kinematics in this method is simulated by solving the volume of fluid advection equation. The normal vector which is used for a piecewise linear interface reconstruction, is obtained using the level set function. The updated interface is then used to re-initialize the level set function in order to maintain its signed-distance property.
Comparing the mentioned methods in terms of the accuracy, robustness and ease of the implementation, one can conclude that the classical level set method is an appropriate choice when a solver for a higher-order approximation of the solution to the level set advection equation is available, for instance, the code BoSSS where the DG method is applied.

## 3 Level Set Method for Modeling the Moving Interfaces

The interface in the level set method is represented implicitly as the zero-iso value of a function which is known as the level set function (Figure 2). The implicit representation provides the ability of handling any topological changes of the interface. As it is shown in the picture, although the level set function is defined over whole the domain, only its zero iso-value, sepa-

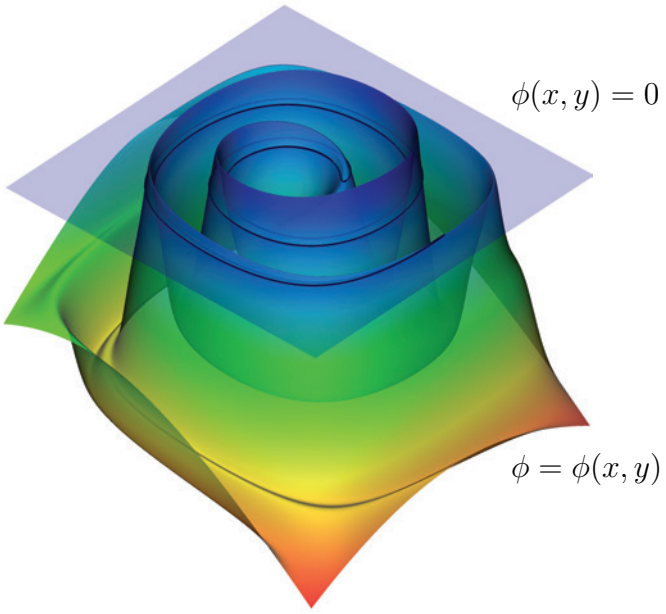



**Figure 2** Representing an interface as the zero iso-value of a level set function

rating the negative and positive regions of the function, is used for the interface representation. Therefore, any function that its zero iso-value represents the interface, can be used as the level set function. But the level set function is commonly designed to be a function that returns the signed-distance to the interface. The signed-distance level set function $\phi(\mathbf{x},t)$ of an interface $\mathcal{I}$ is defined as,

$$\phi(\mathbf{x},t)=\begin{cases}-d(\mathbf{x},t), & \phi<0,\\ 0, & \phi=0,\\ d(\mathbf{x},t), & \phi>0,\end{cases} \tag{1}$$

where $d(\mathbf{x},t)$ is the distance to the interface which is defined as,

$$d(\mathbf{x},t)=\min(|\mathbf{x}-\mathbf{x}_{\mathcal{I}}(t)|), \tag{2}$$

where $\mathbf{x}_{\mathcal{I}}(t)$ denotes the interface position. Parameterizing the interface with the surface coordinates ($\xi_{\mathcal{I}}$,$\eta_{\mathcal{I}}$), the interface position can be determined by,

$$\mathbf{x}_{\mathcal{I}}(t)=\mathbf{x}(\xi_{\mathcal{I}},\eta_{\mathcal{I}},t). \tag{3}$$

The value of the gradient of a signed-distance level set function is uniformly equal to 1. The advantage of using a signed-distance level set function is that this level set function can be used for constructing the distributions for which the distance to the interface is a parameter. For instance, a smoothed Heaviside function defined as,

$$\mathcal{H}_{\epsilon}(\phi)=\begin{cases}0, & \phi<-\epsilon,\\ \frac{1}{2}+\frac{\phi}{2\epsilon}+\frac{1}{2\pi}\sin\left(\frac{\pi\phi}{\epsilon}\right), & -\epsilon\leq\phi\leq\epsilon\\ 1, & \phi>\epsilon,\end{cases} \tag{4}$$

### 3.1 Level Set Advection Equation

Following [10], the evolution of an interface $\mathcal{I}$ can be determined by,

$$\frac{d\mathbf{x}_{\mathcal{I}}(t)}{dt}=\frac{\partial\mathbf{x}(\xi_{\mathcal{I}},\eta_{\mathcal{I}},t)}{dt}=\mathbf{u}_{\mathcal{I}}(\mathbf{x}(\xi_{\mathcal{I}},\eta_{\mathcal{I}},t),t), \tag{5}$$

where $\mathbf{u}_\mathcal{I}$ denotes the interface velocity. As $\phi(\mathbf{x}(\xi_\mathcal{I}, \eta_\mathcal{I}, t), t)$ is defined to be zero for all the time, one can write,

$$\begin{aligned}\frac{d\phi(\mathbf{x}(\xi_\mathcal{I}, \eta_\mathcal{I}, t), t)}{dt} &= \frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x_1}\frac{\partial x_1(\xi_\mathcal{I}, \eta_\mathcal{I}, t)}{\partial t} + \frac{\partial \phi}{dx_2}\frac{\partial x_2(\xi_\mathcal{I}, \eta_\mathcal{I}, t)}{\partial t} + \frac{\partial \phi}{\partial x_3}\frac{\partial x_3(\xi_\mathcal{I}, \eta_\mathcal{I}, t)}{\partial t} \\ &= \frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x_{1_\mathcal{I}}} u_{1_\mathcal{I}} + \frac{\partial \phi}{\partial x_{2_\mathcal{I}}} u_{2_\mathcal{I}} + \frac{\partial \phi}{\partial x_{3_\mathcal{I}}} u_{3_\mathcal{I}} = 0, \end{aligned} \tag{6}$$

where $x_i$ represents the components of $\mathbf{x}$, and $u_{i_\mathcal{I}}$ represents the components of $\mathbf{u}_\mathcal{I}$. Assuming the interface $\mathcal{I}$ to be a material surface, for which the velocity of the interface is equal to the velocity of the fluid particles located on the interface, one can write the equation describing the advection of the level set function $\phi$ as,

$$\frac{\partial \phi}{\partial t} + \mathbf{u} \cdot \boldsymbol{\nabla} \phi = 0, \tag{7}$$

where $\mathbf{u}$ is the velocity field.

### 3.2 Level Set Re-Initialization Equation

A signed-distance level set function keeps its signed-distance property if and only if the advection field meets a condition as,

$$\boldsymbol{\nabla} u_n \cdot \boldsymbol{\nabla} \phi = 0, \tag{8}$$

where $u_n$ is the component of the velocity field in the direction normal to the level set function, obtained as,

$$u_n = \mathbf{u} \cdot \boldsymbol{\nabla} \phi. \tag{9}$$

According to this condition, a signed-distance level set function remains signed-distance if $u_n$ does not have any spatial variation in the direction normal to the level set function, see [22].
If the advection field does not meet the condition (8), the signed distance property can be recovered by performing a re-initialization procedure that of course must not move the interface. It means that the re-initialization procedure is supposed to affect the level set function except its zero iso-value. The level set function can be re-initialized either by performing a geometrical technique such as the fast marching method introduced in [23], or operating the re-initialization over the field of level set function. The latter approach which was introduced in [24], is followed in the present research.
In order to derive the re-initialization equation instructively, one can start with considering the hyperbolic equation ()10) which describes the motion of the iso-values of a level set function in their normal directions,

$$\frac{\partial \phi}{\partial t} + (u_n \mathbf{N}) \cdot \boldsymbol{\nabla} \phi = 0, \tag{10}$$

where $\mathbf{N}$ represents the normal vector to each of the iso-values, see [5]. Since,

$$\mathbf{N} \cdot \boldsymbol{\nabla} \phi = \frac{\boldsymbol{\nabla} \phi}{|\boldsymbol{\nabla} \phi|} \cdot \boldsymbol{\nabla} \phi = \frac{|\boldsymbol{\nabla} \phi|^2}{|\boldsymbol{\nabla} \phi|} = |\boldsymbol{\nabla} \phi|, \tag{11}$$

equation (10) can be rewritten as,

$$\frac{\partial \phi}{\partial t} + u_n |\boldsymbol{\nabla} \phi| = 0. \tag{12}$$

By solving the equation (12) in a time interval $\Delta t$, the local value of $\phi$ increases by $(\Delta t)u_n$ times the value of its local gradient. In order to solve an Eikonal equation of the form $|\boldsymbol{\nabla}\phi| = 1$, one can follow a pseudo-time stepping approach and solve an equation of the form of the equation (12) with $u_n = 1$ and an additional source term 1 as,

$$\frac{\partial \phi}{\partial \tau} + |\boldsymbol{\nabla}\phi| = 1, \tag{13}$$

where $\tau$ is a pseudo-time, see [25]. By solving the equation (13) in a pseudo-time interval $\Delta\tau$, the local value of $\phi$ increases by the difference of the value of its local gradient and 1. It should be noted that the level set function has the negative sign on the opposite side of the interface. Therefore, equation (13) takes the following form in the region $\phi < 0$,

$$\frac{\partial \phi}{\partial \tau} - |\boldsymbol{\nabla}\phi| = -1, \tag{14}$$

The equations (13) and (14) together with the condition that the interface should not be affected by the re-initialization, can be then combined into the following compact form proposed in [10],

$$\frac{\partial \phi}{\partial \tau} + Sign(\phi^0)(|\boldsymbol{\nabla}\phi| - 1) = 0, \tag{15}$$

$$\phi(\mathbf{x}, 0) = \phi^0, \tag{16}$$

where $Sign(\phi^0)$ is a Signum function which is defined as,

$$Sign(\phi^0) = \begin{cases} -1 & \phi^0 < 0, \\ 0 & \phi^0 = 0, \\ 1 & \phi^0 > 0 \end{cases} \tag{17}$$

In [26], the re-initialization equation is rewritten in a more illuminating form as,

$$\frac{\partial \phi}{\partial \tau} + \mathbf{w} \cdot \boldsymbol{\nabla}\phi = Sign(\phi^0), \tag{18}$$

where $\mathbf{w}$ is the characteristic velocity of the hyperbolic equation (18) and defined as,

$$\mathbf{w} = Sign(\phi^0)\frac{\boldsymbol{\nabla}\phi}{|\boldsymbol{\nabla}\phi|}. \tag{19}$$

Inclusion of the Signum function in the definition of the characteristic velocity implies that the vector $\mathbf{w}$ points always outward the interface either within the region of $\phi^0 < 0$ or within the region of $\phi^0 > 0$. It means that the re-initialization of the level set function is started from the interface.

As it was mentioned before, the numerical representation of a jump may produce spurious spatial oscillations leading to a numerical instability. Accordingly, for solving the re-initialization equation (15), one needs to use a smoothed Signum function as an approximation to the exact Signum function (17). Although in the literature, see [24], commonly an infinitely smoothed Signum function is used as,

$$Sign_{\epsilon \to \infty}(\phi^0) = \frac{\phi^0}{\sqrt{(\phi^0)^2 + \epsilon^2}} \tag{20}$$

in the present research we used a finitely smoothed Signum function as,

$$Sign_{\epsilon}(\phi^0) = \begin{cases} -1 & \phi^0 < -\epsilon, \\ \dfrac{\phi}{\epsilon} + \dfrac{1}{\pi}\sin\left(\dfrac{\pi\phi^0}{\epsilon}\right) & -\epsilon \leq \phi^0 \leq \epsilon, \\ 1 & \epsilon < \phi^0, \end{cases} \tag{21}$$

in order to directly adjust the smoothing width. If the slope of the level set function is less or much less than 1, the smoothing width of the smoothed Signum function increases and consequently the speed of the characteristic lines is reduced. But if the slope is much higher than 1, the smoothed Signum function becomes too steep that may result in the numerical instability. In order to overcome this problem, it is proposed in [27] that an infinitely smoothed formulation in terms of the updated level set function $\phi$ instead of the initial level set function $\phi^0$ is used as,

$$Sign_{\epsilon\to\infty}(\phi) = \frac{\phi}{\sqrt{(\phi)^2 + (|\boldsymbol{\nabla}\phi|\epsilon)^2}}. \tag{22}$$

Multiplying $\epsilon$ by $|\boldsymbol{\nabla}\phi|$ in the formulation (22), modifies the smoothing width in order to prevent the numerical instability. As it was mentioned before, in the present research we prefer to use a smoothed Signum function with a finite width. Accordingly, the following formulation is constructed,

$$Sign_{\epsilon}(\phi) = \begin{cases} -1 & \phi < -\alpha\epsilon, \\ \dfrac{\phi}{\alpha\epsilon} + \dfrac{1}{\pi}\sin\left(\dfrac{\pi\phi}{\alpha\epsilon}\right) & -\alpha\epsilon \leq \phi \leq \alpha\epsilon, \\ 1 & \alpha\epsilon < \phi \end{cases} \tag{23}$$

where

$$\alpha = \begin{cases} 1, & |\boldsymbol{\nabla}\phi| \leq 1, \\ |\boldsymbol{\nabla}\phi|, & |\boldsymbol{\nabla}\phi| > 1, \end{cases}$$

## 4 Higher-Order Numerical Approximation

The numerical procedure of solving a scalar transport equation consists of two consecutive stages including the spatial discretization and temporal integration. The spatial discretization consists of representing the solution over a discrete domain and approximating the spatial differential terms converting the PDE to a system of temporal ODEs. The time evolution of the solution can be then obtained by performing a time integration. The present research is focused on the higher-order approximation to the spatial variations of the solution. Generally speaking, increasing the grid resolution is an essential way for improving the solution accuracy. However, the rate of convergence is in a direct relation to the order of the spatial discretization. In the case of applying a lower-order scheme, increasing the grid resolution leads to an error reduction that is relatively small comparing to the additional computational cost, see [28]. An instructive interpretation to this behavior can be made in the context of the Fourier analysis, see [29]: A higher-order spatial discretization, in principle gives the ability of resolving the modes of the solution which have higher wave numbers. Therefore, increasing the grid resolution results in an error reduction with a higher rate. Another consequence of using the higher-order schemes is reducing the numerical dissipation and dispersion errors, see [30]. These errors are quite determinant in making accurate solutions to the hyperbolic conservation laws. For instance,

considering a one-dimensional wave equation as,

$$\frac{\partial \phi}{\partial t} + a\frac{\partial \phi}{\partial x} = 0, \tag{24}$$

with $a$ as the wave speed, the dissipation error reduces the amplitude of the wave leading to the dissipation of the wave, and the dispersion error affects the speed of the wave and produces spurious oscillations. Therefore, if the solution to the equation (24) has high dissipation and dispersion errors, it results in simulating a dissipative wave moving with a wrong speed. Hence, a major advantage that a numerical method can have, is providing a context in which, a higher-order spatial discretization can be formulated properly.

### 4.1 Spatial Discretization

The spatial discretization methods are principally classified into three categories including the Finite Difference Method, the Finite Volume Method and the family of the spectral methods. The starting point in the procedure of a spatial discretization is converting the continuous domain to a discrete domain, over which the numerical solution is represented and the spatial differential terms are approximated.

A discrete domain in the finite difference method consists of a set of the nodes distributed in a structured way. Whereas in the Finite Volume Method as well as the spectral methods, the discrete domain is composed of a set of the sub-domains (cells) with arbitrary geometries which can be distributed in an unstructured way.

The solution in the Finite Difference Method is represented by its nodal values. Each nodal value together with the neighbour nodal values within a certain stencil, are associated with a polynomial distribution of a certain degree. Therefore, in the context of the Finite Difference Method, one can achieve a higher-order approximation by using a polynomial distributions of higher-degree over an expanded grid stencil. But having to expand the grid stencil in order to make a higher-order approximation, can be consider as a disadvantage of the Finite Difference Method.

The solution in the Finite Volume Method is represented in each cell uniformly as the cell-averaged value or with a linear variation limiting the order of approximation up to $2$. In this method, a PDE is converted to a system of temporal ODEs by integrating the PDE locally over each cell. Although there is a discontinuous variation of the solution at the border of the cells, each pair of the cells are connected via approximating the flux of the solution across the common border. As the numerical flux function is expressed in terms of the cell values, it can be approximated with a higher-order by expanding the stencil of the cells, such as the $3^{rd}$-order WENO (Weighted Essentially Non-Oscillatory) scheme proposed in [31] which is specifically designed to handle steep variations in the solution.

The solution in the spectral methods is represented as a certain composition of a spectrum of prescribed analytical functions. Therefore one can consider this type of methods as the mathematical spectralizers. For instance, the solution $\phi(\mathbf{x}, t)$ can be spectralized as,

$$\phi(\mathbf{x}, t) \approx \tilde{\phi}(\mathbf{x}, t) = \sum_{j=1}^{N_{DoF}} \hat{\phi}_j(t)\vartheta_j(\mathbf{x}), \tag{25}$$

where $\vartheta_j(\mathbf{x})$ represents the set of the analytical functions termed as the basis functions, $\hat{\phi}_j(t)$ represents the corresponding coefficients, and $N_{DoF}$ denotes the number of the numerical degrees of freedom within the domain, over which the solution is spectralized. As the basis

functions are prescribed, the solution is obtained by computing the unknown coefficients $\phi_j(t)$. It can be shown that the spectral methods yield exponential convergence rate, see [30]. The different spectral methods are distinguished based on the following characteristics:

- The domain, over which the solution is spectralized. In the spectral element methods such as the DG method, the spectralization is performed over each cell separately. Whereas in the the Finite Element Method as well as the Fourier Spectral Method, the spectralization is commonly made globally or over a wider stencil.
- The type of the basis functions and the domain within which, they are defined. In the Finite Element Method as well as the DG methods, the basis functions are defined within each cell. Whereas in the Fourier Spectral Method, they are defined over the whole domain. As stated in [32], the most suitable basis functions for the periodic problems are the trigonometric functions. On the other hand, the orthogonal polynomials are proven to be appropriate for the non-perodic problems.
- The third issue is the technique applied for determining the unknown coefficients $\hat{\phi}_n(t)$. For the non-periodic problems, the unknown coefficients are commonly determined applying the method of weighted residuals which is explained subsequently.

### 4.1.1 Method of Weighted Residuals

Considering the level set advection equation (7), as an approximate solution $\tilde{\phi}(\mathbf{x}, t)$ does not necessarily satisfy the equation, its substitution into the equation results in the appearance of a residual term as,

$$\frac{\partial \tilde{\phi}}{\partial t} + \boldsymbol{\nabla} \cdot \left( \tilde{\phi} \mathbf{u} \right) = \mathcal{R}(\tilde{\phi}). \tag{26}$$

The method of Weighted Residuals is based on seeking an approximate solution that satisfies a certain restriction imposed on the residual function. The restriction is placed by equating the Legendre inner product of the residual function and a test (weight) function to zero as,

$$\langle \chi_j(\mathbf{x}), \mathcal{R}(\mathbf{x}) \rangle = \int_\Omega \chi_j(\mathbf{x}) \mathcal{R}(\mathbf{x}) d\mathcal{V} = 0, \quad j = 1, \cdots, N_{DoF}. \tag{27}$$

Consequently, the equation (26) takes a form as,

$$\begin{aligned} \int_\Omega \chi_j(\mathbf{x}) \left[ \frac{\partial \tilde{\phi}}{\partial t} + \boldsymbol{\nabla} \cdot \left( \tilde{\phi} \mathbf{u} \right) \right] d\mathbf{x} &= \int_\Omega \chi_j(\mathbf{x}) \mathcal{R}(\tilde{\phi}) d\mathcal{V}, \\ j &= 1, \cdots, N_{DoF}, \end{aligned} \tag{28}$$

which is in fact a set of $N_{DoF}$ equations. The different Weighted Residuals methods are distinguished based on the test function they employ, see [30]. Among the different methods, the Galerkin Weighted Residual method uses the same set of the functions as the test functions, which is used as the basis functions. The Finite Element Method as well as the DG method follow the approach of the Galerkin Weighted Residual method.

## 4.2 Discontinuous Galerkin Method

### 4.2.1 Solution Representation

The DG method is classified in the category of the spectral element methods. The cell-wise representation of the solution results in the solution discontinuity across the borders of the cells, as it is shown in the figure 3. The amplitude of the discontinuities are adjusted by the grid resolution as well as the degree of the spectral representation. The procedure of applying the

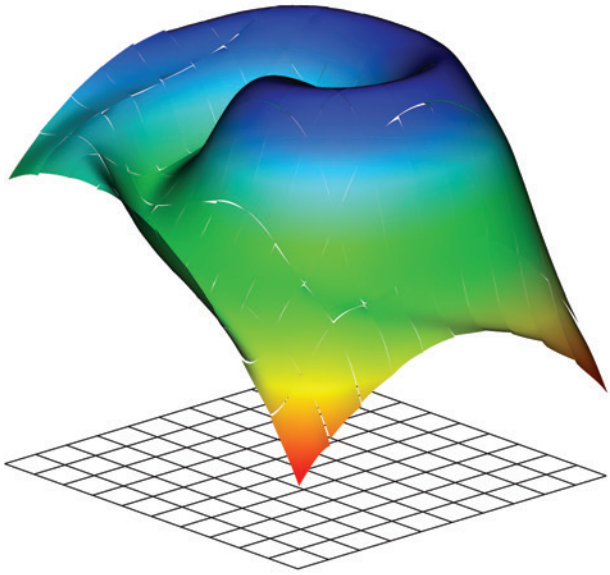


**Figure 3** A DG-based representation of a solution with observed cell-boundary discontinuities

DG method starts with approximating the physical domain $\Omega$ bounded by $\partial\Omega$ by a discrete domain $\Omega_h$ bounded by $\partial\Omega_h$ consisting of $N_C$ non-overlapping boundary conforming cells $\Omega_k^h$. The DG field $\phi_{hp,k}(\mathbf{x},t)$ which is defined as the DG-based representation of a solution $\phi(\mathbf{x},t)$, can be constructed over each cell $\Omega_k^h$ employing an orthogonal basis polynomials space $\vartheta_j(\mathbf{x})$ as,

$$\phi(\mathbf{x},t)|_{\Omega_k^h} \approx \phi_{hp}|_{\Omega_k^h}(\mathbf{x},t) = \sum_{j=1}^{N_P} \hat{\phi}(t)_{jk}\vartheta_{jk}(\mathbf{x}), \quad k = 1,\cdots,N_C, \tag{29}$$

where $N_P$ denotes the dimension of the orthogonal basis polynomials space. The letter $h$ in the subscript $hp$, indicates the finite spatial resolution, and the letter $p$ signifies the spectral representation. The dimension of the orthogonal basis polynomials space which is required for constructing a DG field of degree $p$, can be calculated as,

$$N_P = \frac{1}{D!}\prod_{1\leq l\leq D}(p+l), \tag{30}$$

where $D$ denotes the spatial dimension. It is more convenient to normalize the orthogonal basis polynomials space by the $\mathbb{L}^2$ norm of each of the polynomials, resulting an orthonormal basis polynomial space. The orthonormal basis polynomials space used in the present research is constructed over a set of $N_P$ monomials applying the Gram-Schmidt algorithm. For instance

following [37], the monomials space $\mathscr{P} := \{1, x, y, x^2, y^2, xy\}$ can be converted to an orthonormal polynomial space $\mathbb{P}$ as,

$$\Theta := \left\{ \frac{1}{2}, \frac{\sqrt{3}}{2}x, \frac{\sqrt{3}}{2}y, \sqrt{5}\left(\frac{3}{4}x^2 - \frac{1}{4}\right), \frac{3}{2}xy, \sqrt{5}\left(\frac{3}{4}y^2 - \frac{1}{4}\right) \right\}$$

In order to obtain the coefficients $\hat{\phi}(t)$ corresponding the DG-based representation of a solution $\phi(\mathbf{x}, t)$, the solution needs to be projected over the orthogonal basis polynomials space $\vartheta_{jk}(\mathbf{x})$ by performing an inner product as,

$$\begin{aligned} \hat{\phi}(t)_{jk} &= \langle \phi(\mathbf{x}, t), \vartheta_{jk}(\mathbf{x}) \rangle = \int_{\Omega_k^h} \phi(\mathbf{x}, t)\vartheta_{jk}(\mathbf{x}) d\mathcal{V}, \\ j &= 1, \cdots, N_P, \\ k &= 1, \cdots, N_C, \end{aligned} \tag{31}$$

The procedure of solving an equation in the modal DG method, which is used in the present study, is finalized by obtaining the coefficients $\hat{\phi}(t)$. Then the solution values can be calculated at any arbitrary point within the domain in the postprocessing stage.
As a result of the solution discontinuity across the borders of the cells, every point located on a cell border corresponds to a pair of the asymptotic values, namely the inner-cell value and the outer-cell value, denoted by $\phi_{hp}^-(\mathbf{x}^k)$ and $\phi_{hp}^+(\mathbf{x}^k)$, respectively. A conceptual representation of these values is given in figure 4 for a 1D problem. As it is indicated in this figure, the solution discontinuity is quantified by the jump operator $[\![\phi_{hp}]\!]_{x^k}$ which is defined as the difference between the inner- and the outer-cell values. The inner- and the outer-cell values are defined on the border $\partial\Omega_{h,k}$ of a cell $\Omega_{h,k}$ as,

$$\phi_{hp}^-(\mathbf{x}^k) := \lim_{\boldsymbol{\xi}^k \to \mathbf{x}^k} \phi_{hp}(\boldsymbol{\xi}^k), \quad \mathbf{x}^k \in \left(\Omega_k^h \backslash \partial\Omega_k^h\right), \tag{32}$$

$$\phi_{hp}^+(\mathbf{x}^k) := \lim_{\boldsymbol{\xi}^k \to \mathbf{x}^k} \phi_{hp}(\boldsymbol{\xi}^k), \qquad \mathbf{x}^k \notin \bar{\Omega}_k^h, \tag{33}$$

where

$$\mathbf{x}^k \in \left(\Omega_k^h \backslash \partial\Omega_k^h\right) := \{\mathbf{x}^k \in \Omega_k^h : \mathbf{x}^k \notin \partial\Omega_k^h\},$$

and

$$\mathbf{x}^k \notin \bar{\Omega}_k^h \equiv \mathbf{x}^k \notin \left(\Omega_k^h \bigcup \partial\Omega_k^h\right) := \{\mathbf{x}^k : \mathbf{x}^k \in \Omega_k^h \vee \mathbf{x}^k \in \partial\Omega_k^h\}.$$

### 4.2.2 Approximation of the Partial Differential Terms

The DG method employs the Galerkin Weighted Residual method for the numerical approximation of the partial differential terms. Considering the level set advection equation (7), this method can be implemented by multiplying the $N_P$ members of the orthonormal basis polynomial space $\vartheta_j$ as,

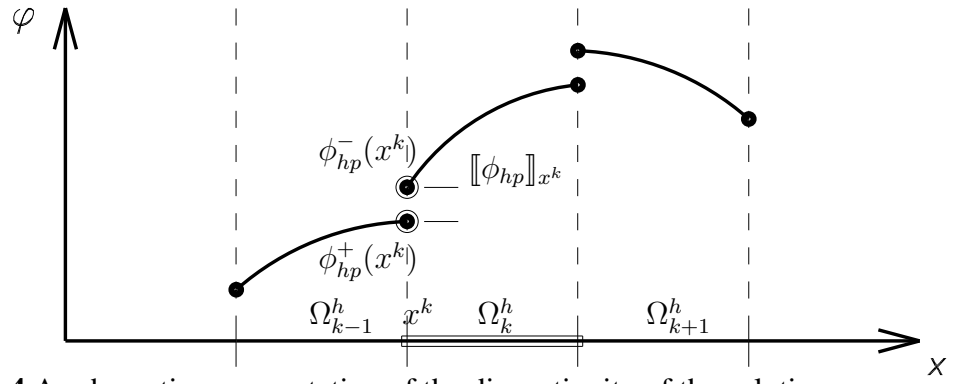



**Figure 4** A schematic representation of the discontinuity of the solution across each of the cell borders

$$\int_{\Omega^h_k} \vartheta_{jk}(\mathbf{x}) \left[ \frac{\partial \phi_{hp}}{\partial t} + \boldsymbol{\nabla} \cdot (\phi_{hp} \mathbf{u}_{hp}) \right] d\mathcal{V} = 0, \qquad (34)$$
$$j = 1, \cdots, N_P,$$
$$k = 1, \cdots, N_C,$$

leading to a matrix equation as,

$$\underbrace{\int_{\Omega^h_k} \vartheta_{jk}(\mathbf{x}) \left( \frac{\partial \phi_{hp}}{\partial t} \right) d\mathcal{V}}_{\textbf{TM}} + \overbrace{\int_{\Omega^h_k} \vartheta_{jk}(\mathbf{x}) (\boldsymbol{\nabla} \cdot \underbrace{(\phi_{hp} \mathbf{u}_{hp})}_{\text{Non-Linear Flux:}= \mathbf{f}} ) d\mathcal{V}}^{\textbf{AM}} = 0, \qquad (35)$$

where the matrices are of the size $1 \times N_P$. It is more convenient to consider the matrix corresponding to each of the terms separately as **TM** and **AM**, denoting the time and the advection matrices, respectively.

**Time Matrix (TM)** Substituting the DG-based representation of $\phi$ into this tensor yields,

$$\begin{aligned} \textbf{TM} &:= \int_{\Omega^h_k} \vartheta_{jk}(\mathbf{x}) \left( \frac{\partial}{\partial t} \left( \sum_{i=1}^{N_P} \hat{\phi}_{ik}(t) \vartheta_{ik}(\mathbf{x}) \right) \right) d\mathcal{V} \\ &= \sum_{i=1}^{N_P} \frac{\partial \hat{\phi}_{ik}(t)}{\partial t} \int_{\Omega^h_k} \vartheta_{jk}(\mathbf{x}) \vartheta_{ik}(\mathbf{x}) d\mathcal{V} \\ &= \sum_{i=1}^{N_P} \frac{d \hat{\phi}_{ik}(t)}{dt} \delta_{ji}. \end{aligned} \qquad (36)$$

where $\delta_{ji}$ is the Kronecker delta function defined as,

$$\delta_{ji} = \begin{cases} 1 & , \text{if } j = i \\ 0 & , j \neq i. \end{cases} \qquad (37)$$

**Advection Matrix (AM)** It should be noted that as the variable $\phi$ is the unknown variable and the velocity field $\mathbf{u}$ is prescribed, this advection term is not a non-linear term. Taking an integration by parts yields,

$$\mathbf{AM} \;:=\; \oint_{\partial\Omega_k^h} \vartheta_{jk}(\mathbf{x})\mathbf{n}_{\mathcal{S}} \cdot \mathbf{f} d\mathcal{S} - \int_{\Omega_k^h} \mathbf{f} \cdot \boldsymbol{\nabla}\vartheta_{jk}(\mathbf{x}) d\mathcal{V} \tag{38}$$

where $\mathbf{n}_{\mathcal{S}}$ is the normal vector to the cell border, directing outward of the cell. In the DG method, although the solution can be discontinuous across the cell borders, each common border of two adjacent cells must be assigned to a unique flux function $\mathbf{f}^*$ which is called the numerical flux. This explicit specification of the fluxes provides the conservative property of the DG method. The numerical flux is required to be substituted only into the first term of the formulation (38) as this term lives on the cell border where the adjacent cells are in connection. In order to construct a numerical flux at a cell border, the solution needs to be reconstructed there. The solution reconstruction in the DG method is commonly made by applying either a $1^{st}$-order upwind scheme or a $2^{nd}$-order central scheme using the inner- and the outer-cell values. An upwind numerical flux can be defined as,

$$\mathbf{f}_U^* = \begin{cases} \phi_{hp}^- \mathbf{u}_{hp}^-, & \dfrac{\mathbf{u}_{hp}^- \cdot \mathbf{n}_{\mathcal{S}} + \mathbf{u}_{hp}^+ \cdot \mathbf{n}_{\mathcal{S}}}{2} \geq 0, \\ \phi_{hp}^+ \mathbf{u}_{hp}^+, & \dfrac{\mathbf{u}_{hp}^- \cdot \mathbf{n}_{\mathcal{S}} + \mathbf{u}_{hp}^+ \cdot \mathbf{n}_{\mathcal{S}}}{2} < 0. \end{cases} \tag{39}$$

The upwind scheme is appropriate if the spatial differential term represents a directional phenomenon such as the advection. This is because this scheme is characterized by high dissipation and low dispersion errors. The central scheme is a proper choice for the terms representing the directionless phenomena such as the diffusion. This is because this scheme is characterized by low dissipation and high dispersion errors. Therefore, if it is applied for a directional problems, it does not dissipate the non-resolved modes of the solution and advects them through the domain with wrong speeds in wrong directions, leading to the formation of spurious spatial oscillations, see [38, 30]. The central numerical flux can be defined as,

$$\mathbf{f}_C^* = \frac{\mathbf{u}^- \phi_{hp}^- + \mathbf{u}^+ \phi_{hp}^-}{2} \tag{40}$$

Substituting the numerical flux $\mathbf{f}^*$ into the first term of the formulation (38), and substituting the DG-based representation of $\mathbf{f}$ into the second term, yields,

$$\begin{aligned} \mathbf{AM} \;&:=\; \oint_{\partial\Omega_k^h} \vartheta_{jk}(\mathbf{x})\mathbf{n}_{\mathcal{S}} \cdot \mathbf{f}^* d\mathcal{S} - \int_{\Omega_k^h} \left( \sum_{i=1}^{N_P} \hat{\mathbf{f}}_{ik}(t)\vartheta_{ik}(\mathbf{x}) \right) \cdot \boldsymbol{\nabla}\vartheta_{jk}(\mathbf{x}) d\mathcal{V} \\ &=\; \oint_{\partial\Omega_k^h} \vartheta_{jk}(\mathbf{x})\mathbf{n}_{\mathcal{S}} \cdot \mathbf{f}^* d\mathcal{S} - \left( \sum_{i=1}^{N_P} \hat{\mathbf{f}}_{ik}(t) \right) \cdot \int_{\Omega_k^h} \vartheta_{ik}(\mathbf{x}) \boldsymbol{\nabla}\vartheta_{jk}(\mathbf{x}) d\mathcal{V}. \end{aligned} \tag{41}$$

**Source Matrix (SM)** In the case of having an extra source term $s_\phi$ in the equation, implementing the Galerkin Weighted Residual method results in a source matrix **SM** as,

$$\mathbf{SM} := \int_{\Omega_k^h} \vartheta_{jk}\left(s_\phi\right) d\mathcal{V}. \tag{42}$$

Substituting the DG-based representation of $s_\phi$ as $s_\phi = \sum_{j=1}^{N_P} \hat{s}(t)_{jk}\vartheta_{jk}(\mathbf{x})$, yields,

$$\begin{aligned}\mathbf{SM} &:= \int_{\Omega_k^h} \vartheta_{jk}(\mathbf{x}) \left( \sum_{i=1}^{N_P} \hat{s}_{\phi,ik}(t)\vartheta_{ik}(\mathbf{x}) \right) d\mathcal{V} \\ &= \sum_{i=1}^{N_P} \hat{s}_{\phi,ik}(t) \int_{\Omega_k^h} \vartheta_{ik}(\mathbf{x})\vartheta_{jk}(\mathbf{x}) d\mathcal{V} \\ &= \sum_{i=1}^{N_P} \hat{s}_{\phi,ik}(t)\delta_{ji}. \end{aligned} \tag{43}$$

### 4.2.3 Gradient Calculation

The gradient of a variable $\phi$ in the DG method can be calculated in two different ways including,

- **Broken gradient:** In this method, considering the DG-based representation of the variable as $\phi = \sum_{j=1}^{N_P} \hat{\phi}(t)_{jk}\vartheta_{jk}(\mathbf{x})$, the gradient can be obtained by differentiating the prescribed orthonormal basis polynomial space $\vartheta_{jk}(\mathbf{x})$. Therefore, denoting the gradient by $\mathbf{G} = \boldsymbol{\nabla}\phi$, its DG-based representation can be obtained as,

$$\sum_{j=1}^{N_P} \hat{\mathbf{G}}_{jk}(t)\vartheta_{jk}(\mathbf{x}) = \sum_{j=1}^{N_P} \hat{\phi}_{jk}(t)\boldsymbol{\nabla}\vartheta_{jk}(\mathbf{x}) \tag{44}$$

- **By-flux gradient:** In this method, each component of the gradient vector is written in the form of a divergence as,

$$G_i = (\nabla_i)\phi \equiv \boldsymbol{\nabla} \cdot (\mathbf{e}_i\phi), \quad i = 1, 2, 3 \tag{45}$$

  where $\mathbf{e}_i$ represents the standard bases. In this way, each component of the gradient term takes the form of an advection term. Therefore, the same procedure described for the numerical approximation of an advection term can be followed, together with using the central scheme for determining the numerical flux. As it is shown in [39], this method leads to a more accurate calculation of the gradient than the former method.

## 4.3 Boundary Condition

As it was explained before, the numerical flux can be determined at the cell borders using a combination of the inner- and the outer-cell values $\phi_{hp}^+$ and $\phi_{hp}^-$. If the cell border is a part of the domain boundary $\partial\Omega$, the outer-cell value is an unknown. If a value $\phi_B$ is imposed on the boundary as a part of the problem definition (Dirichlet boundary condition), the unknown outer-cell value can be determined in two ways, including the direct method as,

$$\phi_{hp}^+ = \phi_B, \tag{46}$$

and the mirror method as,

$$\frac{\phi_{hp}^- + \phi_{hp}^+}{2} = \phi_B. \tag{47}$$

If the value $\phi_B$ is not specified, the unknown outer-cell value may be allowed to be equal to the inner-cell value, which means a zero gradient of $\phi$ normal to the domain boundary (Homogeneous Neumann boundary condition).

### 4.3.1 Numerical integration

In the procedures of implementing the spatial discretization methods as well as in the post-processing stages, one may require the numerical evaluation of definite integrals. The numerical integration methods are commonly referred in the literature as the quadrature rules. Among the different proposed methods, the Gaussian quadrature rules are highly efficient because of the minimal number of evaluations. In principle, a Gaussian quadrature rule consists of evaluating the integrand at a certain set of the integral points (quadrature points) and making a weighted summation of the calculated values, see [40]. For instance, a 1D Gaussian quadrature rule over a domain $[-1,1]$ can be written as,

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i), \tag{48}$$

where $\omega_i$ represents the prescribed weighting functions and $x_i$ represents the quadrature points which are distributed in a certain pattern within the domain of integration. The 1D version of a Gaussian quadrature rule can be simply extended to the corresponding multi-dimensional version by making the same distribution of the quadrature points over a multi-dimensional domain, and employing a multi-dimensional weighting function.
If the integrand has a Heaviside distribution, the accuracy of the integration can not be improved by increasing the order of quadrature rule. In this case, the accuracy can be improved by performing a multistage division of each cell and applying a lower-order quadrature rule over each sub-cell. This technique is commonly termed as the Brute Force integration.

### 4.3.2 Reference Cell

As a numerical grid consists of a set of the cells with various geometries and locations, it is impractical to construct an orthonormal basis polynomial space within each of them. Moreover, implementing a Gaussian quadrature rule for each of the cells is not a clever task. In order to overcome this problem, one can introduce a reference element in another domain characterized by the coordinates $(\xi_1, \xi_2)$ which are related to the coordinates $(x_1, x_2)$ using a transformation as,

$$\mathbf{x} = \mathbf{T}_k(\boldsymbol{\xi}) = \mathbf{M}_k\boldsymbol{\xi} + \mathbf{A}_k, \quad k = 1, \cdots, N_C \tag{49}$$

where $\mathbf{M}_k$ is a matrix which accounts for different types of linear deformation such as stretch, rotation and shear, see [1]. The vector $\mathbf{A}_k$ performs the translation. As the centroid of the reference cell is commonly located on $(\xi_1 = 0, \xi_2 = 0)$, the vector $\mathbf{A}_k$ gives the coordinates of the centroid of the cell $\Omega_k^h$. Figure 5 schematically shows a linear transformation between for instance, a triangular reference cell and a triangular cell $\Omega_k^h$. According to this transformation, a function $f(\mathbf{x})$ can be transferred to its corresponding function $g(\boldsymbol{\xi})$ as,

$$f(\mathbf{T}_k(\boldsymbol{\xi})) = \frac{1}{\sqrt{\det(\mathbf{M}_k)}} g(\boldsymbol{\xi}), \quad k = 1, \cdots, N_C \tag{50}$$

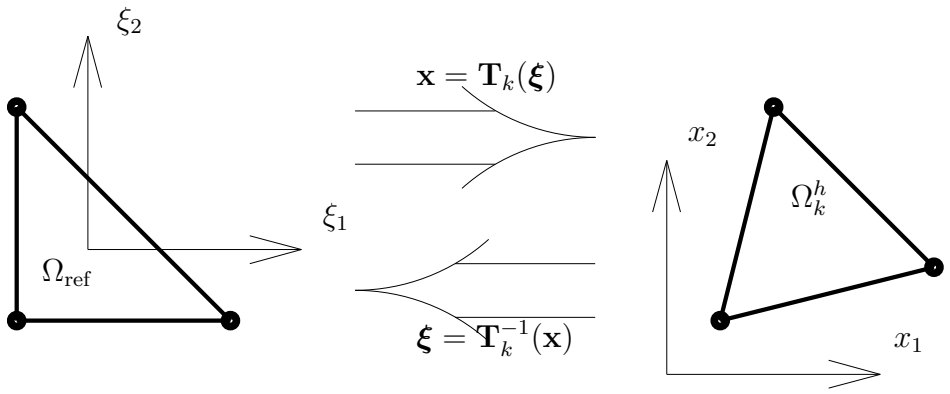



**Figure 5** A schematic representation of a linear transformation between the reference cell $\Omega_{ref}$ and a cell $\Omega_k^h$

This transformation can be applied to transfer, for instance, a solution $\phi_{hp}(\mathbf{x}, t)$ to $\phi_{ref}(\boldsymbol{\xi}, t)$, and the orthonormal basis polynomial space $\vartheta_{jk}(\mathbf{x})$ to $\vartheta_j(\boldsymbol{\xi})$, where $k = 1, \cdots, N_C$ and $j = 1, \cdots, N_P$. It should be noted that the values of the coefficients $\hat{\phi}_{jk}(t)$ stay the same in the transformation as they are only time dependent functions. Furthermore, an integral over a cell $\Omega_k^h$ can be transferred to an integral over the reference cell $\Omega_{ref}$ as,

$$\int_{\Omega_k^h} d\mathcal{V} = \det(\mathbf{M}_k) \int_{\Omega_{ref}} d\mathcal{V}_{ref}, \quad k = 1, \cdots, N_C \tag{51}$$

### 4.4 Time Integration

Considering a temporal ODE as,

$$\phi_t \equiv \frac{d\phi}{dt} = f(\phi, t), \tag{52}$$

the aim of a time integration method is to obtain a sequence of values $\tilde{\phi}^0, \tilde{\phi}^1, \cdots, \tilde{\phi}^n$ such that,

$$\begin{aligned} \tilde{\phi}^n &\approx \phi(t_n), \\ f^n &= f(\tilde{\phi}^n, t). \end{aligned} \tag{53}$$

There are two major families of the numerical time integration methods including the linear multi-step methods and the Runge-Kutta methods, see [33]. The former obtain obtains the $\tilde{\phi}^{n+1}$ either explicitly using the values of the later time steps, or implicitly using the values of the both later and previous time steps, in multiple steps, each of which includes a single stage. Whereas, the latter obtains it explicitly in a single step from $\tilde{\phi}^n$ to $\tilde{\phi}^{n+1}$, including multiple stages. The multi-stage evaluation makes the Runge-Kutta methods more stable than the linear multi-step methods. However, a possible consequence of using the higher-order schemes is the numerical instability. The Runge-Kutta methods with the total variation diminishing (TVD) property can prevent such a problem, see [34]. Considering the one-dimensional wave equation (24), the total variation of the variable $\phi$ can be calculated by,

$$TV = \int \left| \frac{\partial \phi}{\partial x} \right| dx. \tag{54}$$

The time integration of equation (24) is TVD if,

$$TV(\phi^{n+1}) \leq TV(\phi^n). \tag{55}$$

It can be proven that a TVD method is monotonicity preserving and also a monotone numerical scheme is TVD, see [35]. A numerical scheme is monotonicity preserving if the following properties are maintained in time, see [36]:

- No creation of any new extremum in the spatial distribution of the solution
- No decreasing values of the local minimums and no increasing values of the local maximums

It can be shown that the $1^{st}$-order forward (Euler) time integration method is a basic TVD Runge-Kutta method, see [34]. The Euler method can be written as,

$$\tilde{\phi}_t = \frac{\tilde{\phi}^{n+1} - \tilde{\phi}^n}{\Delta t}. \tag{56}$$

Therefore, each stage of a multi-stage higher-order TVD Runge-Kutta method can be constructed by applying an Euler method and combining the results with the initial data using a convex combination, see [5]. The convex combination is a linear combination with the positive coefficients, where the summation of the coefficients is equal to one. For instance following [34], the $3^{rd}$-order TVD Runge-Kutta method which is applied in the present study, can be constructed as,

$$\begin{aligned} \tilde{\phi}^1 &= \tilde{\phi}^n + \Delta t f^1 \\ \tilde{\phi}^2 &= \frac{3}{4}\tilde{\phi}^n + \frac{1}{4}\tilde{\phi}^1 + \frac{1}{4}\Delta t f^1 \\ \tilde{\phi}^{n+1} &= \frac{1}{3}\tilde{\phi}^n + \frac{2}{3}\tilde{\phi}^2 + \frac{2}{3}\Delta t f^2 \end{aligned} \tag{57}$$

### 4.5 Solving the Level Set Advection Equation

The spatial discretization of the advection term included in the level set advection equation (7) can be done using the DG method as explained in the section 4.2. The numerical flux is determined using a $1^{st}$-order upwind scheme. Moreover, a $3^{rd}$-order TVD Runge-Kutta scheme is used for the time integration as explained in the section 4.4.

### 4.6 Solving the Level Set Re-Initialization Equation

A stable solution to the re-initialization equation (15) can not be obtained by calculating the components of $\nabla\phi$ employing a central flux function and considering the term $Sign(\phi^0)(|\nabla\phi| - 1)$ as a source term. This approach does not necessarily result in a stable solution as it is not monotonicity preserving.
The behavior of the re-initialization equation with respect to a numerical scheme, can be better analyzed when it is considered as a Hamilton-Jacobi equation, see [5]. A general form of the Hamilton-Jacobi equation can be written for the level set function $\phi(\mathbf{x}, t)$ as,

$$\frac{\partial\phi}{\partial t} + H(\phi_x, \phi_y) = 0, \tag{58}$$

where $H$ is called the Hamiltonian, and $\phi_{x_i}$ represents $\partial\phi/\partial x_i$. The difficulties in solving the Hamilton-Jacobi equation arise from the fact that this equation develops solutions with singular derivatives even if the initial condition is smooth, see [41]. In order to overcome this problem, a generalized concept of the solution is proposed in [41] which is called the viscosity solution. This name refers to the method of vanishing viscosity which is used to prove the existence of

this type of solutions. The viscosity solution is a weak solution which does not need to be differentiable everywhere, while its existence, uniqueness and stability hold under certain assumptions. Therefore, the solution to every Hamilton-Jacobi equation needs to converge to a viscosity solution. It is proved in [42] that the class of monotone schemes results in the solutions which can meet this condition. Application of a variety of the monotone schemes, such as the Lax-Friedrichs scheme, the Roe-Fix scheme and the Godunov's scheme for solving the Hamilton-Jacobi equations are investigated in [43]. Concerning the re-initialization equation as an Hamilton-Jacobi equation, it is highly recommended in [5] to apply the Godunov's scheme. This approach is followed in the present research accordingly.

### 4.6.1 Godunov's Scheme

The Godunov's scheme was applied in [44] for solving the Hamilton-Jacobi equation. In order to describe this scheme, we consider the variables $G_i = \phi_{x_i}$ $(i = 1, 2)$ representing the components of $\boldsymbol{\nabla}\phi$ in 2D. Having the DG-based representation of $\phi$, the DG-based representation of each of the variables $G_i$ can be obtained using the "By-Flux" method as explained before. In order to use the "By-Flux" method, one needs to calculate the numerical fluxes $f_i^* = (\mathbf{e}_i\phi)^*$ $(i = 1, 2)$, either by using an upwind method or a downwind method as,

$$f_i^{*U} = \begin{cases} \phi^- \mathbf{e}_i, & \mathbf{e}_i \cdot \mathbf{n}_{\mathcal{S}} \geq 0 \\ \phi^+ \mathbf{e}_i, & \mathbf{e}_i \cdot \mathbf{n}_{\mathcal{S}} < 0 \end{cases} \tag{59}$$

$$f_i^{*D} = \begin{cases} \phi^- \mathbf{e}_i, & \mathbf{e}_i \cdot \mathbf{n}_{\mathcal{S}} < 0 \\ \phi^+ \mathbf{e}_i, & \mathbf{e}_i \cdot \mathbf{n}_{\mathcal{S}} \geq 0 \end{cases} \tag{60}$$

where $(i = 1, 2)$, and $\phi^-$ and $\phi^+$ denote the inner- and outer-cell values of $\phi$ at the border of a cell. In addition, $\mathbf{e}_i$ represents the standard bases and $\mathbf{n}_{\mathcal{S}}$ denotes the normal vector to the border of a cell. Following [45], the variables $G_i^U$ and $G_i^D$ $(i = 1, 2)$, are introduced using the numerical flux $f_i^{*U}$ and $f_i^{*D}$ respectively, for calculating $G_i$ . In the Godunov's scheme, the Hamiltonian is approximated by a numerical Hamiltonian $\tilde{H}$ as a function of $G_i^U$ and $G_i^D$, expressed as,

$$\begin{aligned} H(G_1, G_2) &\approx \tilde{H}(G_1^U, G_1^D, G_2^U, G_2^D) \\ &= \mathbf{ext}_x \mathbf{ext}_y \mathbf{H}(G_1, G_2), \end{aligned} \tag{61}$$

where

$$\mathbf{ext}_x H(G_1, G_2) = \begin{cases} \min(H(G_1^U, G_2), H(G_1^D, G_2)), & G_1^U \leq G_1^D \\ \max(H(G_1^U, G_2), H(G_1^D, G_2)), & G_1^U > G_1^D \end{cases}$$

$$\mathbf{ext}_y H(G_1, G_2) = \begin{cases} \min(H(G_1, G_2^U), H(G_1, G_2^D)), & G_2^U \leq G_2^D \\ \max(H(G_1, G_2^U), H(G_1, G_2^D)), & G_2^U > G_2^D \end{cases}$$

The numerical Hamiltonian (61) is then considered as a source term in the equation and the equation is solved by performing a time integration.

#### 4.6.2 Godunov's Scheme for Solving the Level Set Re-initialization Equation

The term $Sign(\phi^0)(|\boldsymbol{\nabla}\phi|-1)$ in the re-initialization equation corresponds to the Hamiltonian in the Hamilton-Jacobi equation. A compact formulation representing the application of the Godunov's scheme for solving the re-initialization equation is proposed in [27] which can be expressed in 2D as, $Sign(\phi^0)(|\boldsymbol{\nabla}\phi|-1)$

$$\approx \begin{cases} Sign(\phi^0)(\sqrt{\max[(G_1^{U+})^2,(G_1^{D-})^2]+\max[(G_2^{U+})^2,(G_2^{D-})^2]}-1), & Sign(\phi^0)\geq 0 \\ Sign(\phi^0)(\sqrt{\max[(G_1^{U-})^2,(G_1^{D+})^2]+\max[(G_2^{U-})^2,(G_2^{D+})^2]}-1), & Sign(\phi^0)<0 \end{cases} \tag{62}$$

where

$$\begin{aligned} G_i^{U+} &= max(G_i^U,0), \quad G_i^{U-} = min(G_i^U,0), \\ G_i^{D+} &= max(G_i^D,0), \quad G_i^{D-} = min(G_i^D,0), \end{aligned} \tag{63}$$

and $(i=1,2)$. A $3^{rd}$-order TVD Runge-Kutta time integration scheme is used for this equation as well.

## 5 Numerical Simulations and Results

This section is assigned to present a set of the test cases considered to investigate the application of the numerical schemes discussed in the previous section for solving the level set advection and re-initialization equations.

### 5.1 Error Calculation

The numerical results of the test cases considered in the present research are analyzed based on measuring three kinds of the errors including "volume/area loss", "interface $\mathbb{L}^1$-error" and "$\mathbb{L}^2$-error", which are briefly describes as follows:

**Volume/Area Loss** This error gives the difference between the volume/area occupied by the computed interface and the correct volume/area. Supposing that the region occupied by the interface corresponds to the negative part of the level set function, the volume/area of this region can be calculated as,

$$\Omega^h_{(\phi_{hp}<0)} = \int_{\Omega^h_{(\phi_{hp}<0)}} d\mathcal{V} \tag{64}$$

Therefore the volume/area loss can be calculated in percentage as,

$$\text{Area Loss} = 100 \times \frac{\Omega_{(\phi^{ref}<0)} - \Omega^h_{(\phi_{hp}<0)}}{\Omega_{(\phi^{ref}<0)}} \tag{65}$$

where $\phi^{ref}$ denotes a reference level set function which is compared to the calculated field of the level set function.

**Interface $\mathbb{L}^1$-Error** This error gives an $\mathbb{L}^1$ measure on the spurious movement of the interface. This error can be calculated as,

$$\mathbb{L}^1_{\mathcal{I}} = \frac{1}{\mathcal{L}} \int_{\Omega^h} |\mathcal{H}(\phi_{hp}) - \mathcal{H}(\phi^{ref})| d\mathcal{V} \tag{66}$$

where $\mathcal{L}$ is the interface circumference, $\phi^{ref}$ is the reference level set function and $\mathcal{H}(\phi)$ is the Heaviside function. The interface $\mathbb{L}^1$-Error as well as the area loss are calculated in the present research by taking a brute-force integration performing a five-stage cell division.

**$\mathbb{L}^2$ Error** This error gives an $\mathbb{L}^2$ measure on the accuracy of computing the field of a variable such as the level set function $\phi$. Having the reference field of the variable available as an analytic function, the $\mathbb{L}^2$ error is calculated as,

$$\mathbb{L}^2_{\phi_{hp}} = \|\phi_{hp} - \phi^{ref}\|_{\Omega^h} = \left( \int_{\Omega^h} [(\sum_{k=1}^{N_C} \sum_{j=1}^{N_P} \hat{\phi}_{jk} \vartheta_{jk}) - \phi^{ref}]^2 d\mathcal{V} \right)^{\frac{1}{2}} \tag{67}$$

where $\phi^{ref}$ is evaluated at the quadrature points in the procedure of the numerical integration.

### 5.2 Periodic Deformation of a Circle in a Swirl Flow

This section is assigned to verify the numerical solution to the level set advection equation (7) by simulating the periodic deformation of an eccentric circle in a prescribed velocity field corresponding to a time-dependent swirl flow. This test case which was originally considered in [46], is an appropriate benchmark for investigating the dissipative and dispersive errors of a numerical method applied for solving the level set advection equation.

**Problem Description** The domain of computation is a square with the lower-left corner located at $(0, 0)$ and the upper-right corner located at $(1, 1)$. The initial geometry of the interface is a circle with the radius $R = 0.15$, centered at ($x_c = 0.5$,$y_c = 0.75$). The signed-distance level set function of the interface is defined as,

$$\phi^0(\mathbf{x}) = \sqrt{(x - x_c)^2 + (y - y_c)^2} - R, \tag{68}$$

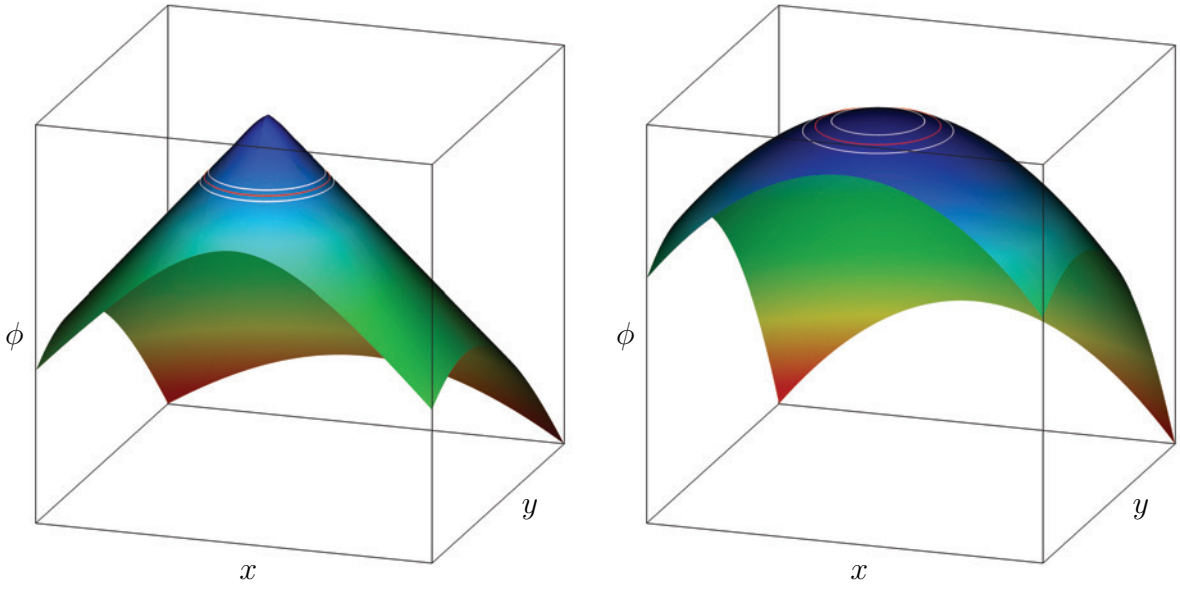
which corresponds to a singular gradient at the center of the circle. In order to investigate the effects of such a singularity, a non-signed-distance level set function is considered additionally, which does not correspond to a singular gradient. This function is expressed as,

$$\phi^0(\mathbf{x}) = (x - x_c)^2 + (y - y_c)^2 - R^2. \tag{69}$$

Figures 6 illustrate the difference between the initial signed-distance and non-signed-distance level set functions. As it is shown, the non-signed-distance level set function does not have a sharp apex. Therefore, this function can be properly projected to an orthonormal basis polynomial space of a certain degree. Figure 7 illustrates the difference between the broken gradients of the signed-distance and non-signed-distance level set functions after one time step. As it is shown in Figure 7a, the effect of the singularity is not limited to the center of the circle. In order to reduce the size of the singular region, one can either make a grid refinement, or use an orthonormal basis polynomial space of lower-degree.
The prescribed velocity field corresponding to the transient swirl flow, is defined as

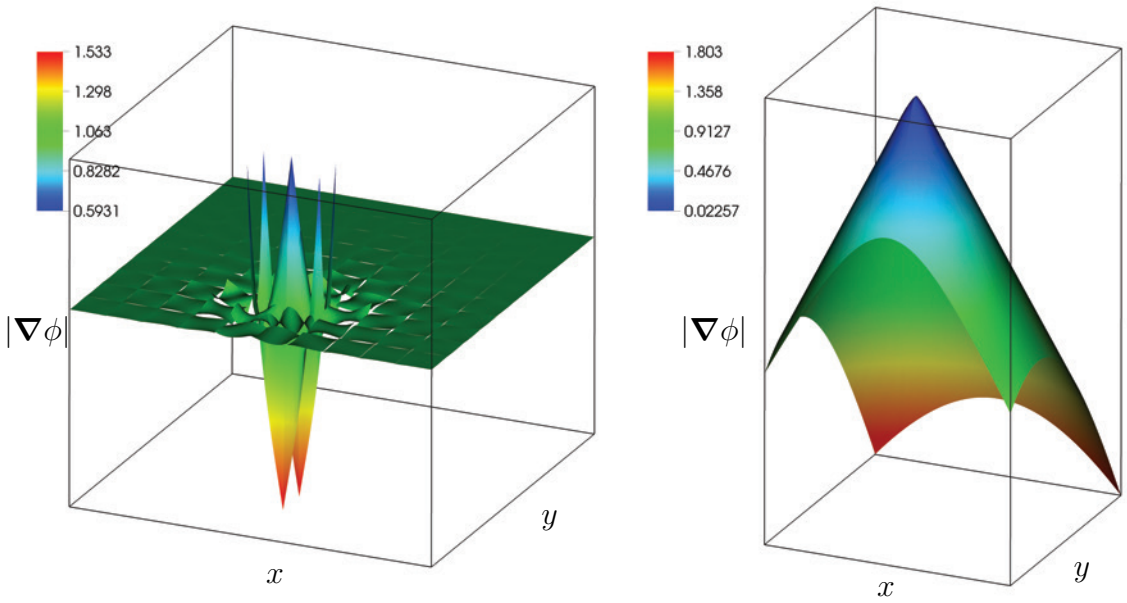
$$\begin{aligned} \mathbf{u}(\mathbf{x}) &= u_x(x, y)\mathbf{e_x} + u_y(x, y)\mathbf{e_y} \\ &= \sin(2\pi y) \sin^2(\pi x) \cos\left(\frac{\pi t}{T}\right) \mathbf{e_x} \\ &- \sin(2\pi x) \sin^2(\pi y) \cos\left(\frac{\pi t}{T}\right) \mathbf{e_y} \end{aligned} \tag{70}$$

where the term $\cos(\pi t/T)$ is multiplied for making a time periodicity in the interval $T$. The value of $T$ for this test case is $8$.




**(a)** Signed-distance **(b)** Non-signed-distance

**Figure 6** Initial signed-distance and non-signed-distance level set functions of a circle. The picture corresponds to $p = 3$ and $N_C = 10 \times 10$. The red curve represents $\phi = 0$ and the white curves represent $\phi = -0.01$ and $\phi = 0.01$.




**(a)** Signed-distance **(b)** Non-signed-distance

**Figure 7** Gradients of the signed-distance and non-signed-distance level set functions of a circle after one time step. The picture corresponds to $p = 3$ and $N_C = 10 \times 10$.

In order to improve the accuracy as well as the numerical stability, the velocity field is updated in each sub-stage of the $3^{rd}-$order TVD Runge-Kutta integration which is implemented in the present study. A vector plot of the velocity field is shown in figure 8.

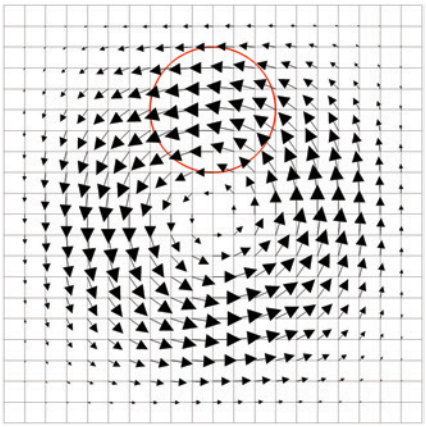
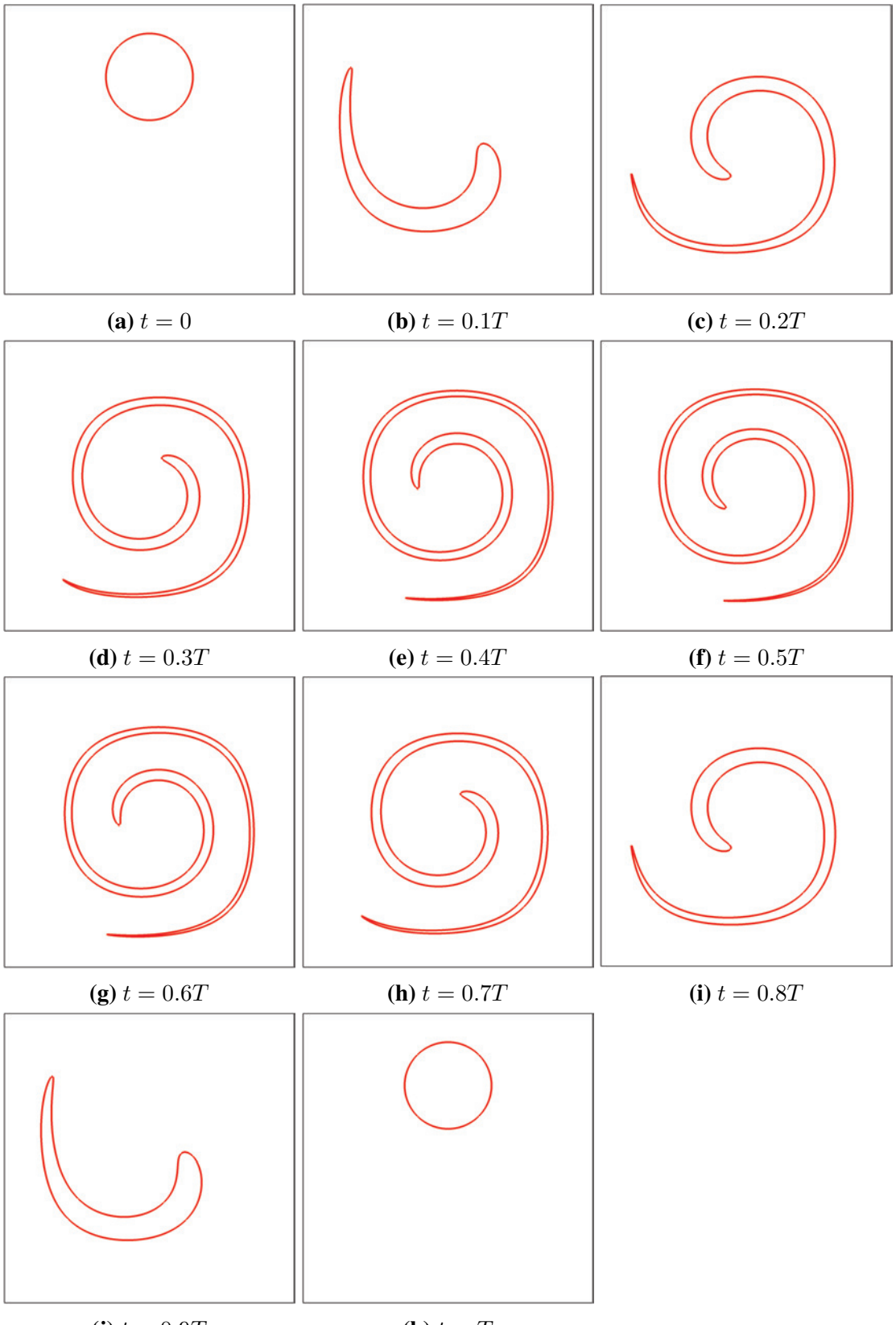


**Figure 8** The prescribed velocity vector field corresponding to a swirl flow

**Numerical Settings** This test case is assigned to perform two classes of the convergence studies including the $p$-convergence study and the $h$-convergence study, employing both the signed-distance and non-signed-distance level set functions. For the $p$-convergence study, the orthonormal basis polynomial space of degrees $p = 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10$ are used and the domain is discretized to a set of the quadrilateral cells with the resolution $N_C = 40 \times 40$. For the $h$-convergence study, the orthonormal basis polynomial space of degrees $p = 3,\ 4,\ 5,\ 6,\ 7$ are used and the domain is discretized to the sets of the quadrilateral cells with the resolutions $N_C = 10 \times 10,\ 20 \times 20,\ 40 \times 40,\ 80 \times 80$. The time step for this study is set to $0.000125$ corresponding to the CFL number of $1/p^2$. In addition to the convergence studies, a case with $p = 4$ and $N_C = 32 \times 32$ is considered in order to make a comparison with the available results reported in the literature. A homogeneous Neumann boundary condition is imposed on the entire boundary of the domain.

**Results** Figure 9 shows snapshots of the interface shape captured in one period of the deformation using $p = 7$, $N_C = 160 \times 160$ and an initially non-signed-distance level set function. The diagrams shown in the Figure 10 illustrate the p-Convergence studies on the area error, interface $\mathbb{L}^1$-error and level set $\mathbb{L}^2$-error at $t = T$, employing both the initially signed distance (denoted as SDLS) and initially non-signed distance (denoted as NSDLS) level set functions. A major fact illustrated in all of these diagrams is that the singularity dramatically reduces the p-convergence rate. Furthermore, as the interface $\mathbb{L}^1$-error measures a difference between the Heaviside functions but not a difference between the fields of the level set functions, such a large reduction in the p-convergence rate supports the argument that the negative effect of the singularity is not limited to the point of the singularity even at $t = T$ when the circle is reformed.



**(a)** $t = 0$ **(b)** $t = 0.1T$ **(c)** $t = 0.2T$

**(d)** $t = 0.3T$ **(e)** $t = 0.4T$ **(f)** $t = 0.5T$

**(g)** $t = 0.6T$ **(h)** $t = 0.7T$ **(i)** $t = 0.8T$

**(j)** $t = 0.9T$ **(k)** $t = T$

**Figure 9** Periodic deformation of an eccentric circle in a swirl flow. The picture corresponds to $p = 7$ and $N_C = 160 \times 160$. The level set function is initially non-signed-distance.

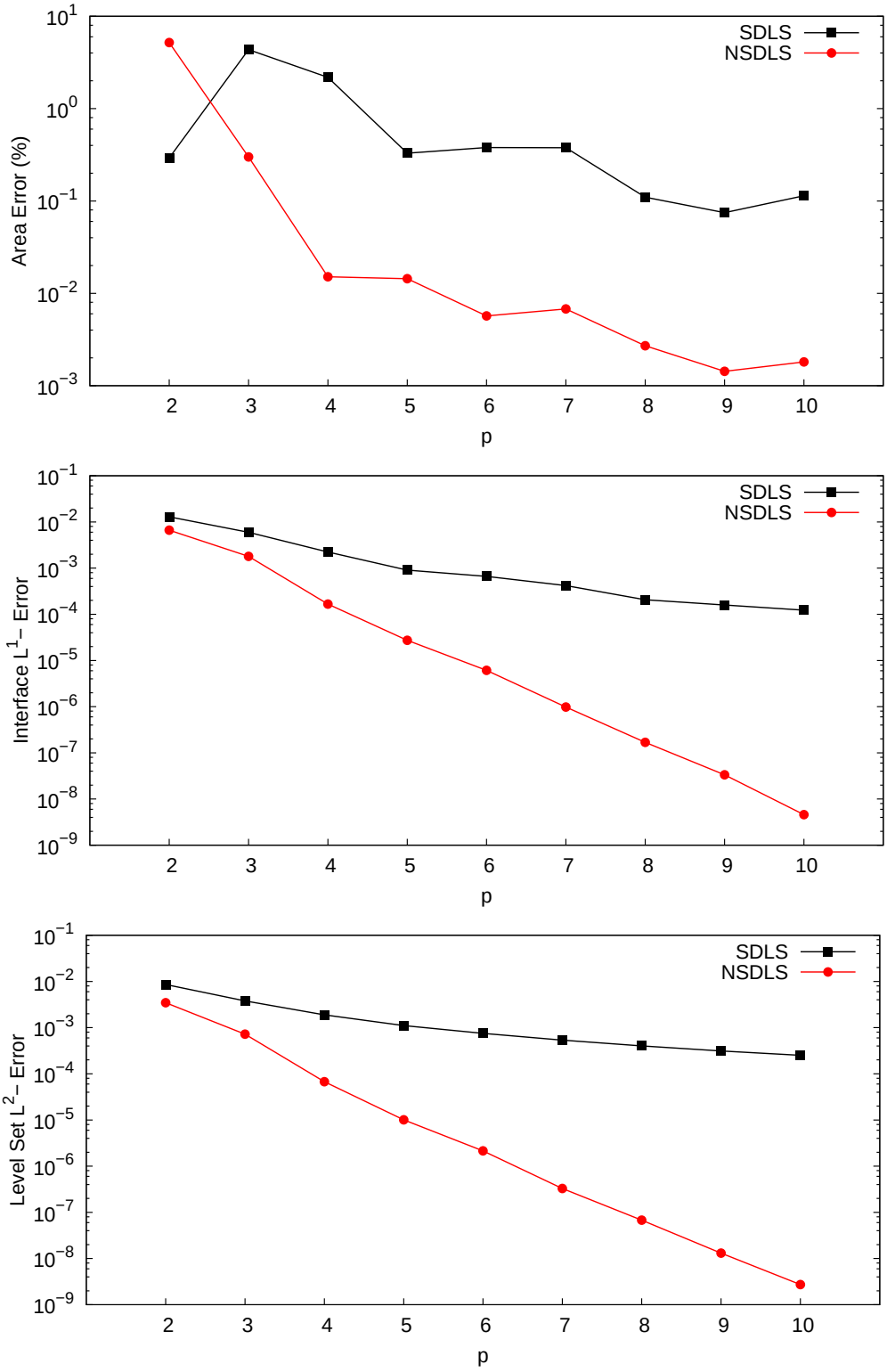
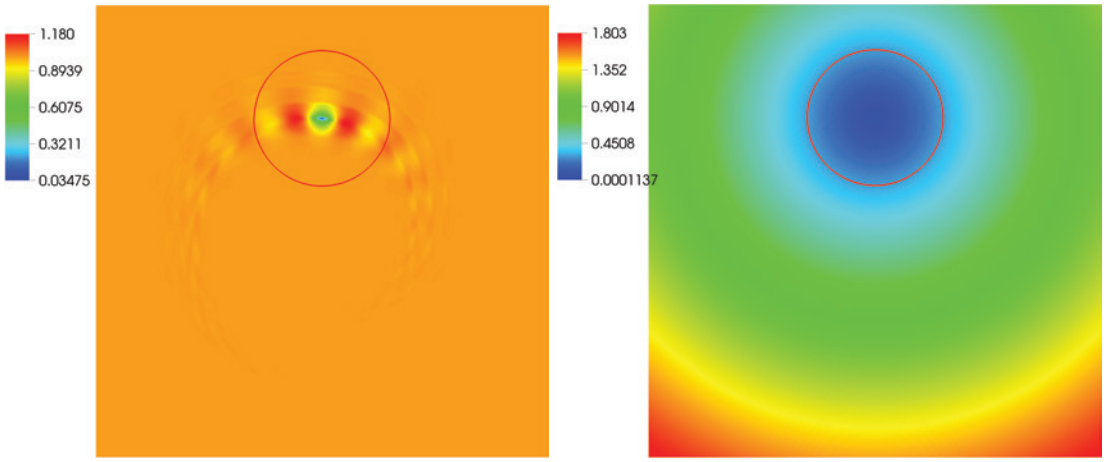



**Figure 10** p-Convergence study on the area error, the interface $\mathbb{L}^1$-error and the level set $\mathbb{L}^2$-error at $t = T$ for simulating the periodic deformation of a circle in a swirl flow. The grid resolution is $N_C = 40 \times 40$.

Figure 11 shows the difference between the gradients of the initially signed-distance and non-signed distance level set functions after one period of deformation. The left picture indicates the footprint of the singularity over the domain. Table 1 demonstrates an $h$-Convergence study




**(a)** $Signed - distance$ **(b)** $Non - signed - distance$

**Figure 11** Gradient of the initially signed-distance and non-signed-distance level set functions of a circle after one period of deformation in a swirl flow. The picture corresponds to $p = 10$ and $N_C = 40 \times 40$.

on the level set $\mathbb{L}^2$-error, using the initially signed-distance and non-signed-distance level set functions. This error is theoretically expected to behave as, see e.g. [6],

$$\mathbb{L}^2_\phi = C(p)(h)^{p+1} \tag{71}$$

where $h$ denotes the characteristics cell size, $p$ denotes the degree of the orthonormal basis polynomial space and $C$ denotes a coefficient which is dependent on $p$. According to this expression, increasing the grid resolution is theoretically expected to result in an exponential error reduction of order $p + 1$. The error reduction rate or the experimental error order (EEO) can be calculated for each error with respect to the error corresponding to the coarser grid, as

$$\begin{aligned} \mathbb{L}^2_{\phi,1} &= C(p)(h_1)^O, \\ \mathbb{L}^2_{\phi,2} &= C(p)(h_2)^O, \\ O &= \frac{\ln(\mathbb{L}^2_{\phi,2}/\mathbb{L}^2_{\phi,1})}{\ln(h_2/h_1)}, \end{aligned} \tag{72}$$

where $O$ is the $h$-convergence rate or EEO. The third column of the table illustrates that using an initially signed-distance level set function does not results in the expected $h$-convergence rate which is $p + 1$. Moreover, it is almost the same for all degrees of the orthonormal basis polynomial space. According to the fifth column of the table, employing an initially non-signed-distance level set function results in the expected $h$-convergence rate, but not for all of the grid resolutions. It means that using $N_C = 10 \times 10$ and $N_C = 20 \times 20$ is not suitable for this certain problem with the corresponding specifications. Diagram 12 is made by plotting the errors listed in table 1 versus the numerical degrees of freedom $N_{DoF}$. This diagram illustrates that using an orthonormal basis polynomial space of a higher-degree results in an accuracy which is more than the accuracy achieved by the one with a lower-degree but with the same $N_{DoF}$. This

**Table 1** $h$-Convergence study on the level set $\mathbb{L}^2$-error for simulating a periodic deformation of a circle in a swirl flow

| $p$ | $N_C$ | Signed Distance | | Non- Signed Distance | |
|---|---|---|---|---|---|
| | | Level Set $\mathbb{L}^2$-Error | EEO | Level Set $\mathbb{L}^2$-Error | EEO |
| 3 | $10 \times 10$ | $2.75E-02$ | $\cdots$ | $1.54E-02$ | $\cdots$ |
| | $20 \times 20$ | $1.07E-02$ | 1.36 | $4.99E-03$ | 1.63 |
| | $40 \times 40$ | $3.80E-03$ | 1.50 | $7.20E-04$ | 2.79 |
| | $80 \times 80$ | $1.03E-03$ | 1.89 | $1.39E-05$ | 5.69 |
| | $160 \times 160$ | $2.98E-04$ | 1.78 | $2.76E-07$ | 5.66 |
| 4 | $10 \times 10$ | $1.76E-02$ | $\cdots$ | $8.70E-03$ | $\cdots$ |
| | $20 \times 20$ | $7.13E-03$ | 1.31 | $2.54E-03$ | 1.78 |
| | $40 \times 40$ | $1.89E-03$ | 1.92 | $6.73E-05$ | 5.24 |
| | $80 \times 80$ | $5.23E-04$ | 1.85 | $7.62E-07$ | 6.47 |
| | $160 \times 160$ | $1.49E-04$ | 1.81 | $1.55E-08$ | 5.62 |
| 5 | $10 \times 10$ | $1.23E-02$ | $\cdots$ | $6.12E-03$ | $\cdots$ |
| | $20 \times 20$ | $4.45E-03$ | 1.47 | $9.64E-04$ | 2.67 |
| | $40 \times 40$ | $1.11E-03$ | 2.00 | $1.01E-05$ | 6.58 |
| | $80 \times 80$ | $3.12E-04$ | 1.83 | $8.32E-08$ | 6.92 |
| | $160 \times 160$ | $8.79E-05$ | 1.83 | $8.60E-10$ | 6.60 |
| 6 | $10 \times 10$ | $9.58E-03$ | $\cdots$ | $4.28E-03$ | $\cdots$ |
| | $20 \times 20$ | $2.94E-03$ | 1.70 | $2.61E-04$ | 4.04 |
| | $40 \times 40$ | $7.52E-04$ | 1.97 | $2.14E-06$ | 6.93 |
| | $80 \times 80$ | $2.08E-04$ | 1.85 | $8.91E-09$ | 7.91 |
| | $160 \times 160$ | $5.73E-05$ | 1.86 | $4.36E-11$ | 7.67 |
| 7 | $10 \times 10$ | $7.56E-03$ | $\cdots$ | $3.04E-03$ | $\cdots$ |
| | $20 \times 20$ | $2.02E-03$ | 1.90 | $6.50E-05$ | 5.55 |
| | $40 \times 40$ | $5.37E-04$ | 1.91 | $3.29E-07$ | 7.63 |
| | $80 \times 80$ | $1.47E-04$ | 1.87 | $8.22E-10$ | 8.64 |
| | $160 \times 160$ | $4.03E-05$ | 1.87 | $2.72E-12$ | 8.24 |

signifies the computational efficiency achieved by applying the higher-order methods. Table 2 makes a comparison between the results obtained in the present research and a number of the available results reported in the literature. As it is shown, the accuracy of the DG method is much higher than the FV method even if a higher-order WENO flux function is employed.

### 5.3 Periodic Deformation of a Slotted Disk in a Swirl Flow

This section is assigned to verify the solution to the level set advection equation (7) by simulating the periodic deformation of an eccentric slotted disk in a prescribed velocity field corresponding to a time-dependent swirling flow expressed as the equation 70.

**Problem Description** The domain of computation is a square with the lower-left corner located at $(0, 0)$ and the upper-right corner located at $(1, 1)$. The initial geometry of the interface

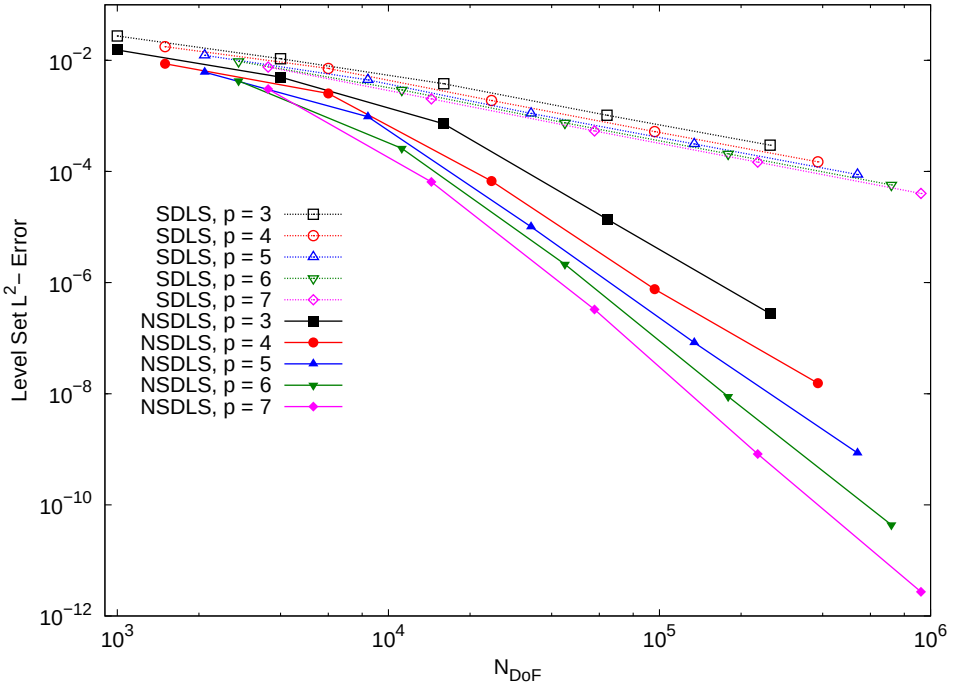



**Figure 12** Level set $\mathbb{L}^2$-error versus $N_{DoF}$ for simulating a periodic deformation of a circle in a swirl flow

**Table 2** Accuracy of the DG method comparing to a higher-order WENO Finite Volume method for simulating the periodic deformation of a circle in a swirl flow

| Method | $p$ | $N_C$ | $N_{DoF}$ | SDLS | Area Loss (%) | $\mathbb{L}^1$-error |
|---|---|---|---|---|---|---|
| DG [1] | 4 | $32 \times 32$ | 15360 | Yes | -3.81 | 4.38E-03 |
| DG [1] | 4 | $32 \times 32$ | 15360 | No | 0.1 | 7.73E-04 |
| DG [2] | 4 | $32 \times 32$ | 15360 | No | 0.71 | 5.5E-4 |
| $5^{th}$-Order WENO FV [3] | 0 | $128 \times 128$ | 16384 | Yes | -39.8 | 3.1E-2 |
| $5^{th}$-Order WENO FV [4] | 0 | $128 \times 128$ | 16384 | Yes | -0.71 | 1.4E-3 |

[1] Present. [2] DG method [7]. [3] Classical level set method [47]. [4] Particle level set method [47]

is a notched disk with the radius $R = 0.15$, the slot length $L_{Slot} = 0.25$ and the slot width $W_{Slot} = 0.5$, centered at $(x_c = .5, y_c = 0.75)$. The manually constructed signed-distance level set function of the interface is shown in the figure 13. The period of deformation is set to $T = 8$.

**Numerical Settings** The domain is discretized to a set of the quadrilateral cells with $N_C = 100 \times 100$, according to which, the slot passes 10 cells along its width. The degree of the orthonormal basis polynomial space is set to $p = 7$. The time step is set to $\Delta t = 0.0002$ corresponding to the CFL number of $1/p^2$. A homogeneous Neumann boundary condition is imposed on the entire boundary of the domain.

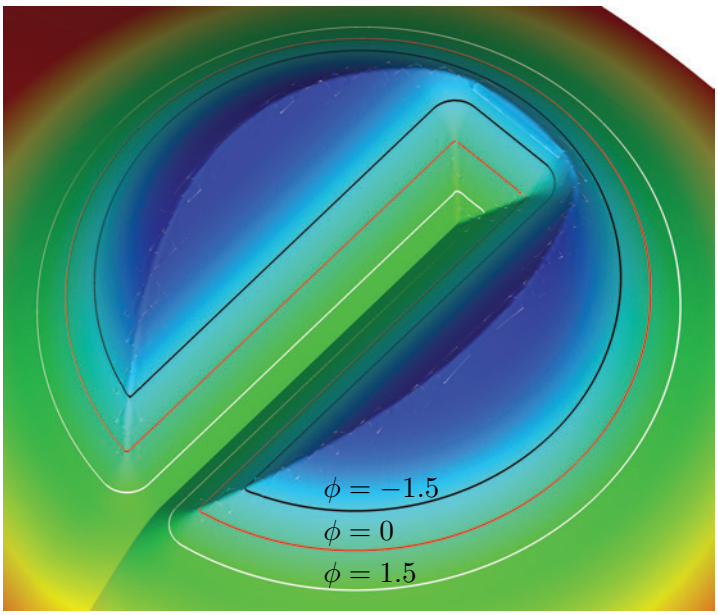



**Figure 13** The initial signed distance level set function of a slotted disk. The picture corresponds to $p = 7$ and $N_C = 100 \times 100$.

**Results** Figure 14 shows snapshots of the interface shapes captured in one period of deformation, using $p = 7$. As it is shown in the figure, despite the complexity in the geometry of the interface as well as the long period of deformation, the original shape of the interface is fairly regained at $t = T$. The area loss is $-1.85\%$ and the interface $\mathbb{L}^1$-error is $0.000974$.

### 5.4 Re-initializing the Level Set Function of a Circle

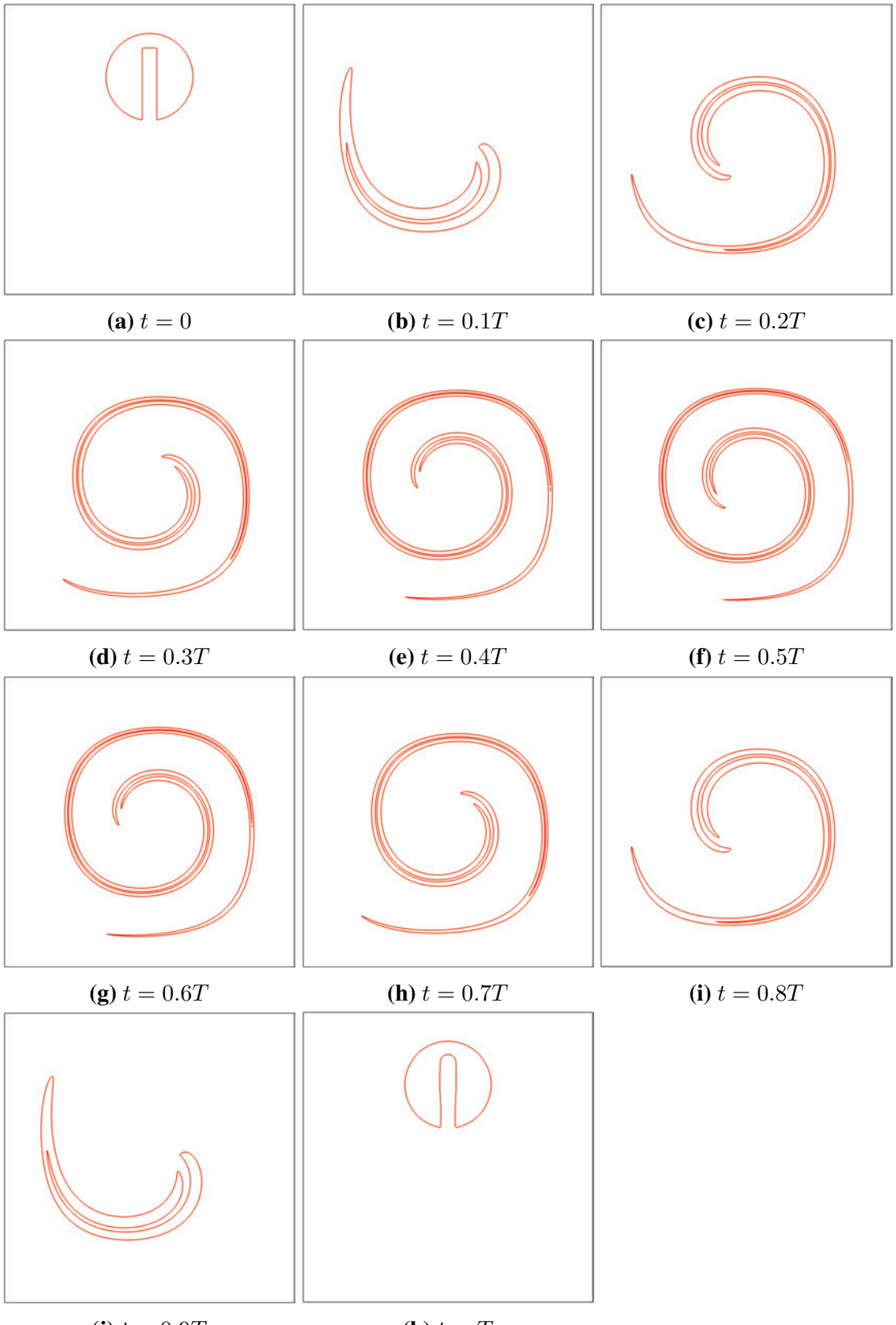
This section is assigned to verify the solution to the level set re-initialization equation (15) by re-initializing the level set function of a circle. As a result of the simple geometry of the interface, this test case clarifies the numerical challenges inherent in the application of the DG method for solving the re-initialization equation.

**Problem Description** The domain of computation is a square with the lower-left corner located at $(-0.5, -0.5)$ and the upper-right corner located at $(0.5, 0.5)$. The interface geometry is a circle with the radius $R = 0.25$, centered at $(x_c = 0, y_c = 0)$. The initial level set function of the interface is analytically expressed as,

$$\phi^0(\mathbf{x}) = 0.5 \cdot \left( \sqrt{(x - x_c)^2 + (y - y_c)^2} - R \right), \tag{73}$$

which corresponds to a gradient value or a slope of $0.5$. As it is discussed in the section 5.2, the gradient of this level set function is singular at the center of the circle.

**Numerical Setting** The domain is discretized to a set of the quadrilateral cells with $N_C = 20 \times 20$, according to which, the circle passes $10$ cells along its diameter. The degree of the orthonormal basis polynomial space is set to $p = 5$. The time step is set to $\Delta t = 0.002$ corresponding to the CFL number of $1/p^2$. The regularized Signum function 21 is used with the smoothing parameter $\epsilon = h$. As the signed distance property of a level set function needs to be recovered only in the region where the interface diffusion takes place, the re-initialization



**(a)** $t = 0$ **(b)** $t = 0.1T$ **(c)** $t = 0.2T$

**(d)** $t = 0.3T$ **(e)** $t = 0.4T$ **(f)** $t = 0.5T$

**(g)** $t = 0.6T$ **(h)** $t = 0.7T$ **(i)** $t = 0.8T$

**(j)** $t = 0.9T$ **(k)** $t = T$

**Figure 14** Periodic deformation of an eccentric slotted disk in a swirl flow. The picture corresponds to $p = 7$ and $N_C = 100 \times 100$.

equation is solved only in a narrow band around the interface in order to save the computational cost. In the direction normal to the interface, the narrow band includes a cell which is cut by the interface and one neighbor cell at each side. A homogeneous Neumann boundary condition is imposed on the boundary of the narrow band.

**Results** Figure 15 shows the level set function re-initialized in a narrow band around the interface. The red curve represents $\phi = 0$ and the white curves represent $\phi = -0.075$ and $\phi = 0.075$. As the re-initialization is not supposed to move the zero iso-value of the level set function (the interface), only one red curve is observed in the picture. But as the other iso-values can be moved, two iso-values corresponding to $\phi = -0.075$ and two iso-values corresponding to $\phi = -0.075$ are formed after the re-initialization. Figure 16 shows the gradient value of

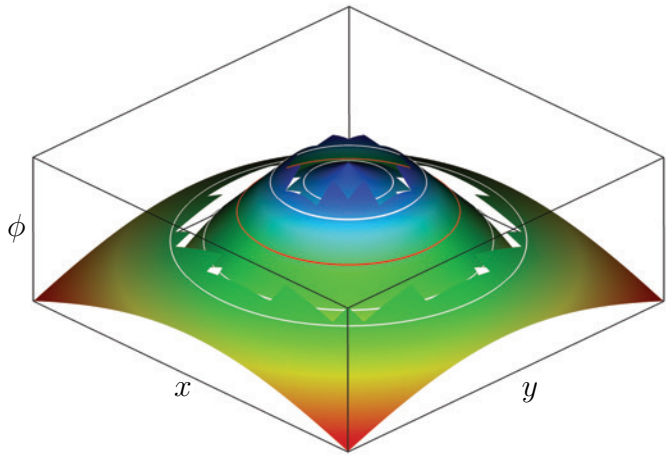



**Figure 15** The level set function of a circle with the initial slope of $0.5$ re-initialized in a narrow band around the interface. The picture corresponds to $p = 5$ and $N_C = 20 \times 20$.

the level set function after the re-initialization, implying spurious spatial oscillations in the field of the level set function. Diagram 17 shows the pseudo-time history of the $\mathbb{L}^2$-error of

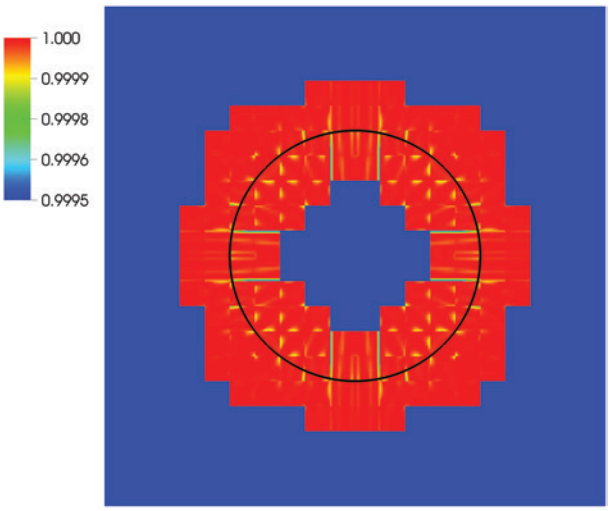



**Figure 16** Gradient value of the re-initialized level set function of a circle. The initial slope of the level set function is $0.5$. The picture corresponds to $p = 5$ and $N_C = 20 \times 20$.

$|\boldsymbol{\nabla}\phi|$ during the re-initialization. As it is shown, although the error decreases at the beginning with an appropriate rate, the solution starts to diverge after a certain pseudo-time. This is in connection with the spurious spatial oscillation observed in the figure 16. Following [9], the

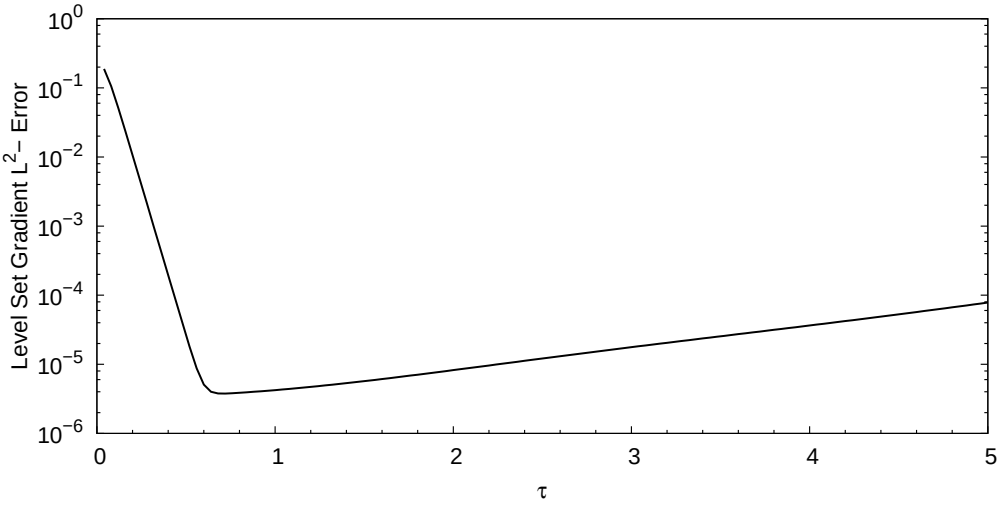



**Figure 17** $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$ during the re-initialization of the level set function of a circle. The initial slope of the level set function is $0.5$. The degree $p = 5$ is used with $N_C = 20 \times 20$.

solution stability can be improved by adding an artificial diffusion along the characteristic lines of the solution as,

$$\frac{\partial \phi}{\partial \tau} + Sign(\phi^0)(|\boldsymbol{\nabla}\phi| - 1) - \nu_{\mathbf{n}} \frac{\partial^2 \phi}{\partial \mathbf{n}^2} = 0 \tag{74}$$

where $\nu_{\mathbf{n}}$ is the coefficient of the artificial diffusion and $\mathbf{n}$ is the interface normal vector which is calculated as,

$$\mathbf{n} = \frac{\boldsymbol{\nabla}\phi}{|\boldsymbol{\nabla}\phi|}.$$

As

$$\frac{\partial \phi}{\partial \mathbf{n}} = \mathbf{n} \cdot \boldsymbol{\nabla}\phi,$$

the diffusion term can be calculated as,

$$\nu_{\mathbf{n}} \left( \frac{\partial^2 \phi}{\partial \mathbf{n}^2} \right) = \nu_{\mathbf{n}} (\mathbf{n} \cdot \boldsymbol{\nabla}(\mathbf{n} \cdot \boldsymbol{\nabla}\phi)).$$

As $|\boldsymbol{\nabla}\phi| = 1$, therefore,

$$\nu_{\mathbf{n}} (\mathbf{n} \cdot \boldsymbol{\nabla}(\mathbf{n} \cdot \boldsymbol{\nabla}\phi)) = 0,$$

which means that the diffusion term does not affect the solution in the steady state. Diagram 18 shows the effect of adding a diffusion term to the re-initialization equation. As it is shown, using the diffusion coefficients $\nu_{\mathbf{n}} = 0.0005,\ 0.001,\ 0.002$ has a notable stabilization effect. As adding a second-order term to the equation imposes more CFL restriction, a lower time step is required in order to use larger diffusion coefficients. Table 3 reports the effects of adding a diffusion term, on the area error, interface $\mathbb{L}^1$-error and the $\mathbb{L}^2$-error of the level set gradient after performing a re-initialization. The errors, indicate that the spurious movement of the interface as a result of the re-initialization, is quite small and it is further reduced by adding an artificial

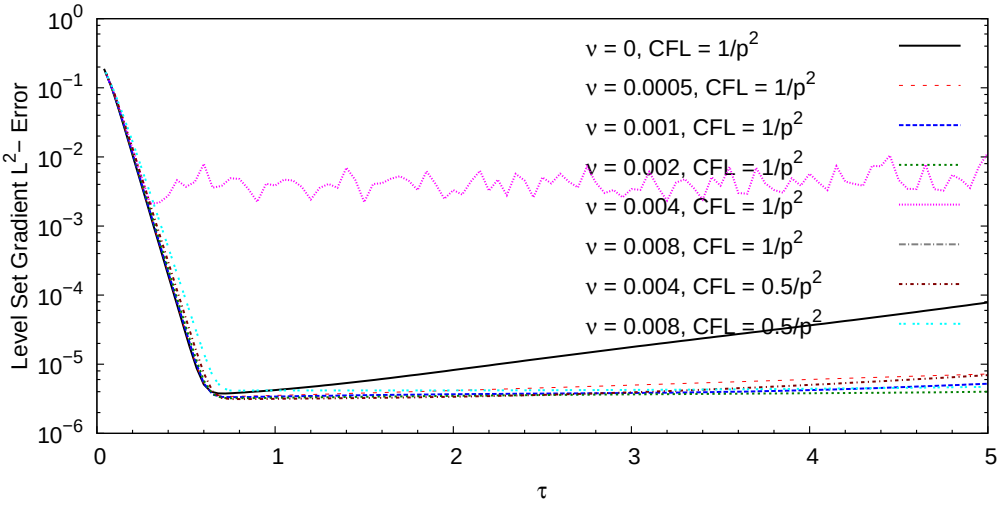



**Figure 18** Effect of the artificial diffusion on the $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$ during the re-initialization of the level set function of a circle. The initial slope of the level set function is $0.5$. The degree $p = 5$ is used with $N_C = 20 \times 20$.

**Table 3** Effects of adding a diffusion term to the re-initialization equation for re-initializing the level set function of a circle. The degree $p = 5$ is used with $N_C = 20 \times 20$.

| | $\nu = 0$ | $\nu = 0.002$ |
|---|---|---|
| Area Loss (%) | $-0.00136$ | $-0.00154$ |
| Interface $\mathbb{L}^1$-error | $4.0385E-06$ | $3.914E-07$ |
| LS Gradient $\mathbb{L}^2$-error | $2.616E-05$ | $4.762E-06$ |

diffusion to the re-initialization equation. Figure 19 shows the effects of the smoothing width $\epsilon$, the degree $p$ and the characteristic cell size $h$ as well as the artificial diffusion, on the pseudo-time history of the $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$. As it is shown, increasing the smoothing width of the Signum function improved the stability although it reduces the convergence rate. Moreover, it is shown that increasing the degree of the orthonormal polynomial space, improves the stability in addition to increasing the accuracy. Furthermore, it is shown that although increasing the grid resolution improves the stability, it imposes more restriction on the CFL number. Figure 20 shows the pseudo-time history of the $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$ to verify the conditional Signum function (23) for re-initializing the level set functions with very small or very large initial slopes. The diagrams indicate the proper performance of this Signum function.

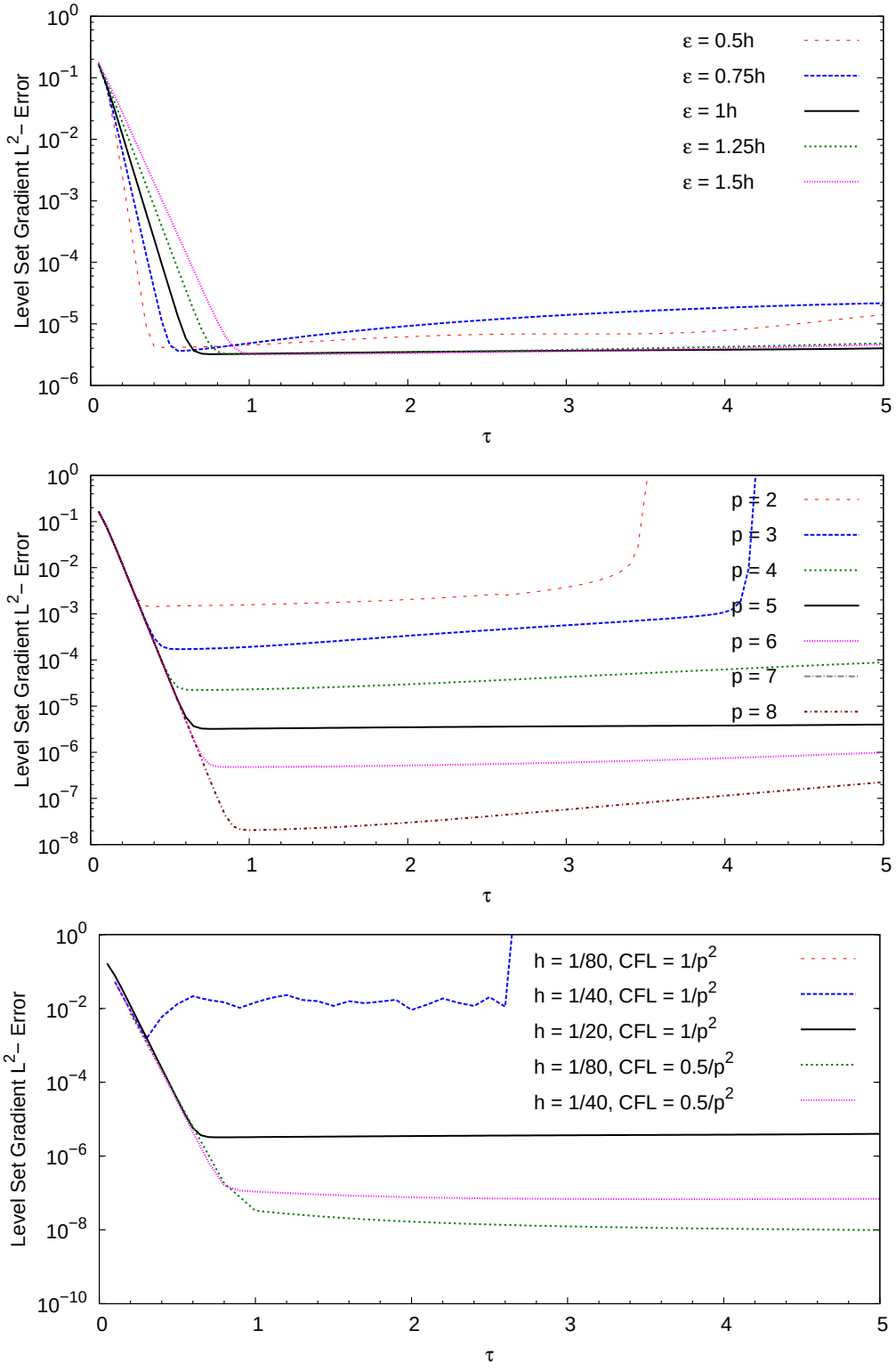



**Figure 19** Effects of $\epsilon$, $p$ and $h$ as well as the artificial diffusion, on the pseudo-time history of the $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$ during the re-initialization of the level set function of a circle. The initial slope of the level set function is $0.5$.

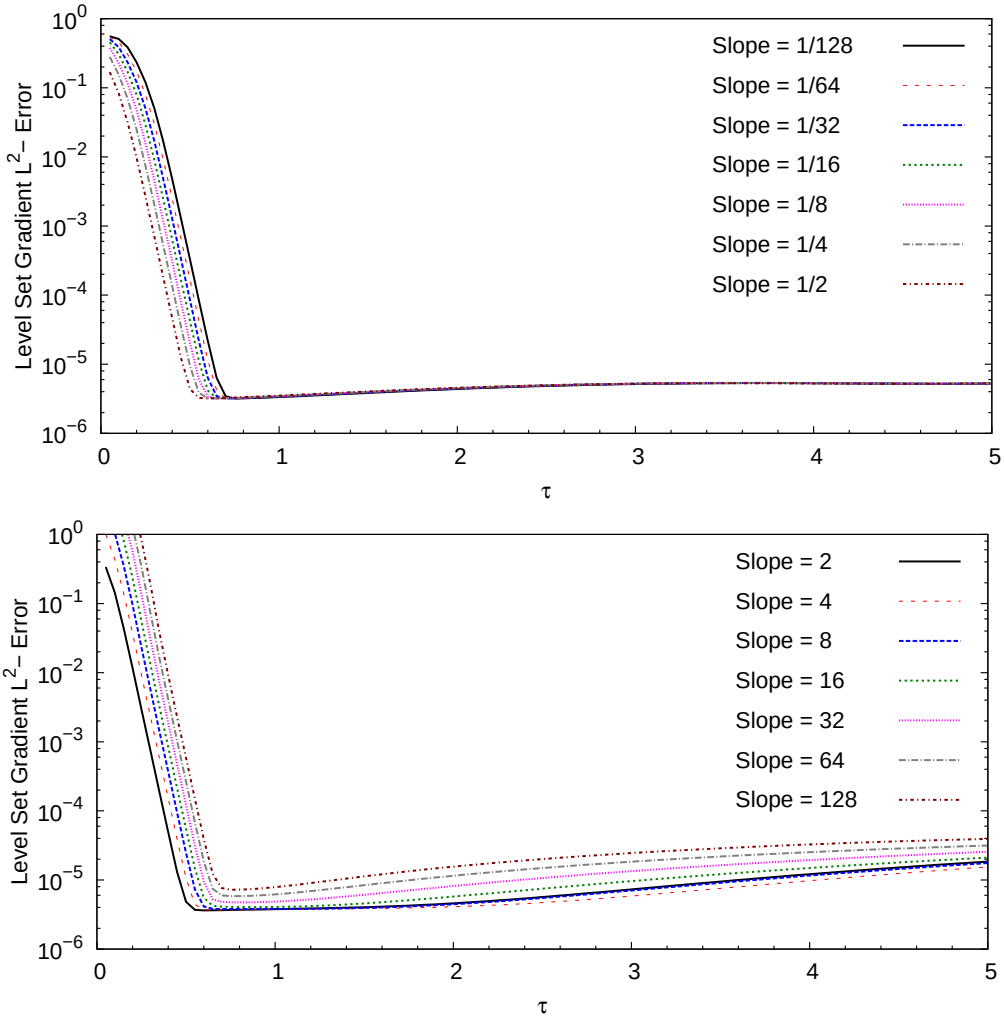



**Figure 20** pseudo-time history of the $\mathbb{L}^2$-error of $|\boldsymbol{\nabla}\phi|$ to verify the Signum function (23) for re-initializing the level set functions with very small or very large initial slopes

### 5.5 Re-initializing the Level Set Function of an Arc

This section is assigned to verify the solution to the level set re-initialization equation (15) by re-initializing the level set function of an arc. This test case is mainly aimed to investigate the $h$-convergence of the solution to the re-initialization equation. Such an investigation can not be done for the test case 5.4 because of the singularity in the gradient of the signed-distance level set function of a circle. On the other hand, if the re-initialization equation is solve over a narrow band around the interface in order to leave the point of singularity outside the domain of computation, the the narrow bands in different grid resolutions do not have the same thickness. Therefore, in the present test case an arc with the center located outside the domain of computation is considered and the re-initialization equation is solved over the entire domain.

**Problem Description** The domain of computation is a rectangle with the lower-left corner located at $(-0.5, -0.25)$ and the upper-right corner located at $(0.5, 0.25)$. The interface geometry is a circular arc with the radius $0.625$ centered at $(x_c = 0, y_c = 0.625)$. The initial level set function is analytically expressed by the equation 73 that corresponds to an initial slope of $0.5$.

**Numerical Setting** The domain is discretized to a set of the quadrilateral cells with $N_C = 10 \times 5,\ N_C = 20 \times 10,\ N_C = 40 \times 20$. The degree of the orthonormal basis polynomial space is set to $p = 5,\ 6,\ 7$. The time step is set to $\Delta t = 0.00025$. A regularized Signum function $Sign(\phi^0)$ is used with the smoothing parameter $\epsilon$ of $0.1$.

**Results** Figure 21 shows the level set function before and after the re-initialization. Table 4 lists the $\mathbb{L}^2$-errors of the gradient value of the level set function as well as the experimental error orders which are calculated using the expression (72). This table indicates an acceptable $h$-convergence of the solution.

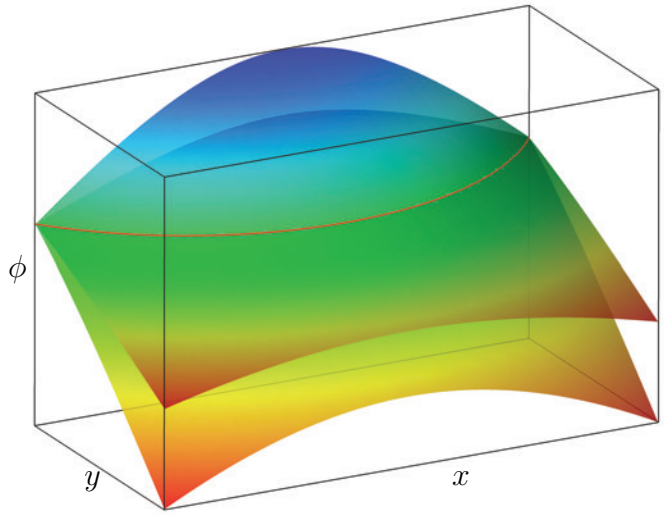



**Figure 21** Re-initialization of the level set function of an arc with the initial slope of $0.5$. The picture corresponds to $p = 5$ and $N_C = 10 \times 5$.

**Table 4** $h$-Convergence study on the $\mathbb{L}^2$-Error of $|\boldsymbol{\nabla}\phi|$ for re-initializing the level set function of an arc with the initial slop of $0.5$.

| $p$ | $N_C$ | $\mathbb{L}^2$-Error of $\lvert\boldsymbol{\nabla}\phi\rvert$ | EEO |
|---|---|---|---|
| | $10 \times 5$ | $3.10E-06$ | $\cdots$ |
| 5 | $20 \times 10$ | $3.08E-08$ | 6.65 |
| | $40 \times 20$ | $8.34E-10$ | 5.21 |
| | $10 \times 5$ | $5.90E-07$ | $\cdots$ |
| 6 | $20 \times 10$ | $2.53E-09$ | 7.87 |
| | $40 \times 20$ | $2.61E-11$ | 6.59 |
| | $10 \times 5$ | $5.61E-08$ | $\cdots$ |
| 7 | $20 \times 10$ | $3.58E-10$ | 7.29 |
| | $40 \times 20$ | $2.67E-12$ | 7.07 |

## 6 CONCLUSIONS

- A higher-order Modal Discontinuous Galerkin Method was successfully applied for solving the level set advection and re-initialization equations.
- As a result of the higher/order approximation, an accurate solution to the level set advection equation obtained without a necessity to solve the re-initialization equation.
- Concerning the solution to the level set advection equation, a singularity in the gradient of the level set function resulted in a large reduction of the the $hp$-convergence rate.
- Employing the Godunov's scheme in approximating the Hamiltonian of the level set re-initialization equation as well as adding a diffusion term, resulted in a stable solution to the equation.
- As a result of the higher/order approximation, an accurate solution to the re-initialization equation was found which does not move the interface spuriously.
- As a future work, filtering the solution of the re-initialization equation is expected to improve the long-term stability.

## 7 Acknowledgement

The present research was accomplished under a PhD grant from the Center of Smart Interfaces at Darmstadt University of Technology.

## REFERENCES

[1] F. Kummer, *The BoSSS Discontinuous Galerkin solver for incompressible fluid dynamics and an extension to singular equations*. PhD Dissertation, Technische Universität Darmstadt, 2012.

[2] R. Mousavi, *Level Set Method for Simulating the Dynamics of the Fluid-Fluid Interfaces: Application of a Discontinuous Galerkin Method*. PhD Dissertation, Technische Universität Darmstadt, 2014.

[3] Y. Shikhmurzaev, *Capillary Flows with Forming Interfaces*. Chapman and Hall-CRC, 2007.

[4] D. M. Anderson, G. B. McFadden, A. A. Wheeler, Diffuse-Interface Methods in Fluid Mechanics. *Annual Review of Fluid Mechanics*, **30**, 139–165, 1998.

[5] S. Osher, R. Fedkiw, *Level-set methods and dynamic implicit surfaces*. Springer-Verlag New York, Inc., 2003.

[6] D. A. D. Pietro, A. Ern, *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer-Verlag, 2011.

[7] E. Marchandise, J. F. Remacle, N. Chevaugeon, A quadrature-free discontinuous Galerkin method for the level set equation. *Journal of Computational Physics*, **212**, 338–357, 2006.

[8] S. Zahedi, K. Gustavsson, G. Kreiss, A conservative level set method for contact line dynamics. *Journal of Computational Physics*, **228**, 6361–6375, 2009.

[9] J. Grooss, J. S. Hesthaven, A level set discontinuous Galerkin method for free surface flows. *Journal of Computer Methods in Applied Mechanics and Engineering*, **195**, 3406–3429, 2006.

[10] M. Sussman, E. Fatemi, P. Smereka,S. Osher, An improved level set method for incompressible two-phase flows. *Computers and Fluids*, **27**, 663–680, 1998.

[11] J. H. Ferziger, M. Peric, *Computational Methods for Fluid Dynamics*. Springer-Verlag, 2002.

[12] B. Lafaurie, C. Nardone, R. Scardovelli, S. Zaleski, G. Zanetti, Modelling Merging and Fragmentation in Multiphase Flows with SURFER. *Journal of Computational Physics*, **113**, 134–147, 1994.

[13] J. Glimm, J. Grove, B. Lindquist, The bifurcation of tracked scalar waves. *SIAM Journal on Scientific and Statistical Computing*, **9**, 61–79, 1988.

[14] S. Osher, J. A. Sethian, Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, **79**, 12–49, 1988.

[15] S. O. Unverdi, G. Tryggvason, A Front-Tracking Method for Viscous, Incompressible, Multi-fluid Flows. *Journal of computational physics*, **100**, 25–37, 1992.

[16] D. Enright, R. Fedkiw, J. Ferziger, I. Mitchell, A Hybrid Particle Level Set Method for Improved Interface Capturing. *Journal of Computational Physics*, **183**, 83–116, 2002.

[17] F. H. Harlow, J. E. and Welch, Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of Fluids*, **8**, 2182–2189, 1965.

[18] C. W. Hirt, B. D. Nichols, Volume of fluid (VOF) method for the dynamics of free boundaries. *Journal of Computational Physics*, **39**, 201–225, 1981.

[19] D. L. Youngs, M. J. Morton, M. J. Baines, *Time-dependent multi-material flow with large fluid distortion, in: Numerical Methods for Fluid Dynamics*. Academic Press, New York, 1982.

[20] O. Ubbink, *Time-dependent multi-material flow with large fluid distortion, in: Numerical Methods for Fluid Dynamics*. PhD Dissertation, Imperial College of Science, Technology and Medicine, 1997.

[21] M. Sussman, E. G. Puckett, A Coupled Level Set and Volume-of-Fluid Method for Computing 3D and Axisymmetric Incompressible Two-Phase Flows. *Journal of Computational Physics*, **162**, 301–337, 2000.

[22] M. Raessi, J. Mostaghimi, M. Bussmann, Advecting normal vectors: A new method for calculating interface normals and curvatures when modeling two-phase flows. *Journal of Computational Physics*, **226**, 774–797, 2007.

[23] J. A. Sethian, A Fast Marching Level Set Method for Monotonically Advancing Fronts. *Proceedings of National Academy of Science*, **93**, 1591–1595, 1998.

[24] M. Sussman, P. Smereka, S. Osher, A level set approach for computing solutions to incompressible two-phase flow. *Journal of Computational Physics*, **114**, 146–159, 1994.

[25] E. Rouy, A. Tourin, A viscosity solutions approach to shape-form-shading. *SIAM journal on Numerical Analysis*, **29**, 867–884, 1992.

[26] M. Sussman, A. S. Almgren, J. B. Bell, P. Colella, L. H. Howell, M. L. Welcome, An Adaptive Level Set Approach for Incompressible Two-Phase Flows. *Journal of Computational Physics*, **148**, 81–124, 1999.

[27] D. Peng, B. Merriman, S. Osher, H. Zhao, M. Kang, A PDE-based fast local level set method. *Journal of Computational Physics*, **155**, 410–438, 1999.

[28] C. W. Shu, High-order Finite Difference and Finite Volume WENO Schemes and Discontinuous Galerkin Methods for CFD. *International Journal of Computational Fluid Dynamics*, **17**, 107–118, 2003.

[29] M. Ainsworth, Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. *International Journal of Computational Fluid Dynamics*, **198**, 106–130, 2004.

[30] G. E. Karniadakis, S. J. Sherwin, *Spectral/hp Element Methods for Computational Fluid Dynamics*. Oxford University Press, USA, 2005.

[31] X. D. Liu, S. Osher, T. Chan, Weighted Essentially Non-Oscillatory Schemes. *Journal of Computational Physics*, **126**, 202–212, 1996.

[32] B. Fornberg, *A practical guide to pseudospectral methods*. Cambridge University Press, 1998.

[33] L. N. Trefethen, *Finite difference and spectral methods for ordinary and partial differential equations*. Cornell University, 1996.

[34] S. Gottlieb, C. W. Shu, Total variation diminishing Runge-Kutta schemes. *Mathematics of Computation*, **67**, 73–85, 1998.

[35] A. Harten, High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, **49**, 357–393, 1983.

[36] C. Hirsch, *Numerical Computation of Internal and External Flows: Fundamentals of Computational Fluid Dynamics*. Butterworth-Heinemann, 2007.

[37] N. Emamy, *BoSSS for solving conservation laws*. Annual report, Technische Universität Darmstadt, Fachgebiet Strömungsdynamik, 2010.

[38] E. Marchandise, N. Chevaugeon, J. F. Remacle, Spatial and spectral superconvergence of discontinuous Galerkin method for hyperbolic problems. *Journal of Computational and Applied Mathematics*, **215**, 484–494, 2010.

[39] N. Emamy, *Numerical simulation of a droplet in a stationary electric field using DG*. PhD Dissertation, Technische Universität Darmstadt, 2014.

[40] R. Kress, *Numerical analysis*. Springer-Verlag New York, 1998.

[41] M. G. Crandall, P. L. Lions, Viscosity Solutions of Hamilton-Jacobi Equations. *Transactions of the American Mathematical Society*, **277**, 1–42, 1983.

[42] M. G. Crandall, P. L. Lions, Two Approximations of Solutions of Hamilton-Jacobi Equations. *Mathematics of Computation*, **43**, 1–19, 1984.

[43] S. Osher, C. W. Shu, High-Order Essentially Nonoscillatory Schemes for Hamilton-Jacobi Equations. *SIAM Journal on Numerical Analysis*, **28**, 907–922, 1991.

[44] M. Bardi, S. Osher, The Nonconvex Multidimensional Riemann Problem for Hamilton-Jacobi Equations. *SIAM Journal on Mathematical Analysis*, **22**, 344-351, 1991.

[45] J. Yan, S. Osher, A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations. *Journal of Computational Physics*, **230**, 232–244, 2011.

[46] R. J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis*, **33**, 627–665, 1996.

[47] D. Enright, F. Losasso, R. Fedkiw, A Fast and Accurate Semi-Lagrangian Particle Level Set Method. *Computers and Structures*, **83**, 479-490, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# LARGE EDDY SIMULATION OF TURBULENT COMPRESSIBLE FLOWS USING THE CHARACTERISTIC BASED SPLIT SCHEME AND MESH ADAPTATION

**Renato V. Linn**[1] **and Armando M. Awruch**[2]

[1]Federal University of Rio Grande do Sul
99 Osvaldo Aranha Ave. 4th floor, 90035-190, Porto Alegre, RS, Brazil
e-mail: renatolinn@ufrgs.br

[2] Graduate Program in Civil Engineering (PPGEC/UFRGS)
99 Osvaldo Aranha Ave. 3rd floor, 90035-190, Porto Alegre, RS, Brazil
e-mail: amawruch@ufrgs.br

**Keywords:** Computational Fluid Dynamics, Compressible Flow, Large Eddy Simulation, Mesh Adaptation

**Abstract.** *The simulation of high-speed turbulent compressible flows using a numerical method of Large Eddy Simulation (LES) combined with the Characteristic-Based Split scheme (CBS) and anisotropic mesh adaptation is presented in this work. The CBS scheme is a unified approach for Computational Fluid Dynamics (CFD) with capability of covering a wide range of flow speeds and types with good stability and accuracy compared with other numerical schemes of the same order [1]. Although LES of incompressible flows combined with the CBS scheme has already been successfully addressed, the compressible extension is not yet covered, been the main contribution of this work. The CBS scheme is employed in a Finite Element Method (FEM) context for space and time discretization using unstructured meshes with adaptation [2], allowing the representation of complex geometries with accuracy. The anisotropic mesh adaptation is performed with mesh refinement, mesh coarsening and edge swapping procedures. A compressible dynamic Smagorinsky model is employed for the compressible LES model. The developed code is used to investigate a complex turbulent transonic flow around a circular cylinder in a two-dimensional approach. Several complex flow features such as lamda-shock-waves, viscous interactions and Von Kármán vortex sheet effects are correctly captured by the mesh adaptation strategy and the computed aerodynamic coefficients are close to the experimental reported values.*

## 1 INTRODUCTION

The simulation of high-speed turbulent compressible flows at transonic and supersonic speed ranges has several difficulties. For aerodynamic problems, the accurate capturing of the complex three-dimensional shock waves and flow features developed is a challenging task which is hard to accomplished without making use of a mesh adaptation methodology. The adequate mesh to be used in the simulation depends on the desired resolution, the numerical algorithm employed, the flow characteristics and the geometry, thus being hard to be a priori determined even for an expert in CFD. Furthermore, the large mesh resolution required to correctly evaluate the smallest scales present in flows of high Reynold number, typically present in such type of problems, is prohibitive. Such constraints are alleviated by making use of a turbulence modelling strategy such as the Large Eddy Simulation (LES) combined with mesh adaptation.

Large Eddy Simulation (LES) directly calculates the large and energetic vortical structures in turbulent flows, while modelling the smaller-scales eddies. Therefore, compared to Reynolds-Averaged Navier-Stokes (RANS) models, the advantage of LES are significant, being RANS effective only for steady simulations of fluid flows. The LES model is combined with the Characteristic-Based split Scheme (CBS) [1] in this work to simulate turbulent compressible flows. The CBS scheme can handle with flows with a wide range of velocities. The capabilities of using the CBS scheme to solve compressible turbulent flows has not been investigated yet and it is the main objective of this work.

Mesh adaptation is combined with the turbulent flow solver in order to achieve a high resolution solutions together with low computational cost. The main procedures for mesh adaptation used in this work are mesh refinement, mesh coarsening and edge swapping. Mesh adaptation is evaluated anisotropically by making use of a metric-based methodology using the concept of Riemannian space to analyse the error as a tensorial quantity [2].

The developed algorithm is used to investigate a complex transonic turbulent flow around a circular cylinder in a two-dimensional approach. Even for a two-dimensional investigation, the obtained results are quite promising. The evaluated aerodynamic coefficients, Strouhal number and the complex flow features are in good agreements with experimental results.

## 2 GOVERNING EQUATIONS

For an arbitrary function $\mathcal{F}(x_i, t)$, the filtered variable $\bar{\mathcal{F}}(x_i, t)$ in physical space is represented by the convolution product [3]:

$$\bar{\mathcal{F}}(x_i, t) = \int_\Omega G(x_i - \xi_i, \Delta)\mathcal{F}(\xi_i, t)\, d\xi_i \tag{1}$$

where $x_i$ are the Cartesian coordinates $(i = 1, 2)$, $t$ is the time variable, $G$ is the filter kernel, $\Omega$ represents the flow domain and $\Delta$ is a measure of the filter width and is related to the computational mesh size. For compressible flows, the density weighted variable $\widetilde{\mathcal{F}}(x_i, t)$ is employed:

$$\widetilde{\mathcal{F}}(x_i, t) = \frac{\overline{\rho\mathcal{F}}}{\overline{\rho}} \tag{2}$$

and the resulting filtered compressible Navier-Stokes equations are:

$$\frac{\partial\overline{\rho}}{\partial t} + \frac{\partial}{\partial x_i}(\overline{\rho}\widetilde{u}_i) \tag{3}$$

$$\frac{\partial\left(\overline{\rho}\widetilde{u}_j\right)}{\partial t}+\frac{\partial}{\partial x_i}\left(\overline{\rho}\widetilde{u}_i\widetilde{u}_j\right)-\frac{\partial\widetilde{\tau}_{ij}}{\partial x_i}-\frac{\partial\tau_{ij}^{sgs}}{\partial x_i}+\frac{\partial\overline{p}}{\partial x_j}+\frac{\partial\left(\overline{\tau}_{ij}-\widetilde{\tau}_{ij}\right)}{\partial x_i}=0 \tag{4}$$

$$\frac{\partial\left(\overline{\rho}\widetilde{E}\right)}{\partial t}+\frac{\partial}{\partial x_i}\left(\overline{\rho}\widetilde{E}\widetilde{u}_i\right)-\frac{\partial}{\partial x_i}\left(\widetilde{k}\frac{\partial\widetilde{T}}{\partial x_i}\right)+\frac{\partial}{\partial x_i}\left(\overline{p}\widetilde{u}_i\right)-\frac{\partial}{\partial x_i}\left(\widetilde{\tau}_{ij}\widetilde{u}_j\right)-C_p\frac{\partial q_i^{sgs}}{\partial x_i}- \tag{5}$$
$$\frac{\partial}{\partial x_k}\left(\left[\overline{\tau}_{ik}-\widetilde{\tau}_{ik}\right]\widetilde{u}_i\right)-\frac{1}{2}\frac{\partial}{\partial x_i}\overline{\rho}\left(\widetilde{u_k u_k}u_i-\widetilde{u}_k\widetilde{u}_k\widetilde{u}_j-\tau_{kk}^{sgs}\widetilde{u}_i\right)+\frac{\partial}{\partial x_i}\left(\overline{k}\frac{\partial\overline{T}}{\partial x_i}-\widetilde{k}\frac{\partial\widetilde{T}}{\partial x_i}\right)=0$$

where $\overline{\rho}$ is the mean density, $\widetilde{u}_i$ are the filtered Cartesian components of the velocity, $\overline{p}$ is the mean pressure, $\widetilde{\tau}_{ij}$ and $\overline{\tau}_{ij}$ are the filtered and mean molecular viscous stress tensor, respectively, $\widetilde{E}$ is the filtered total energy, $\widetilde{T}$ is the filtered absolute temperature, $\widetilde{k}$ and $\overline{k}$ are the filtered and mean diffusion coefficient, respectively, $C_p$ is the specific heat coefficient at constant pressure, $\tau_{ij}^{sgs}$ is the sub-grid scale stress tensor and $q_i^{sgs}$ is the sub-grid turbulent heat flux. The filtered viscous stress tensor is approximated by:

$$\widetilde{\tau}_{ij}=\widetilde{\mu}\left[\left(\frac{\partial\widetilde{u}_i}{\partial x_j}+\frac{\partial\widetilde{u}_j}{\partial x_i}\right)-\frac{2}{3}\delta_{ij}\frac{\partial\widetilde{u}_k}{\partial x_k}\right] \tag{6}$$

where $\widetilde{\mu}$ is the molecular viscosity based on the Favre-filtered static temperature $\widetilde{T}$ and $\delta_{ij}$ is the Kronecker delta. The temperature dependence of the molecular velocity is evaluated through the Sutherland's relation [4]:

$$\widetilde{\mu}=\frac{1.45\widetilde{T}^{3/2}}{\widetilde{T}+110}\times 10^{-6} \tag{7}$$

where all variables are in the SI system and $\widetilde{T}$ is given in Kelvin. The closure of the conservative equations is done with the addition of the state equation of a perfect gas:

$$\overline{p}=\left(\gamma-1\right)\overline{\rho}\widetilde{e} \tag{8}$$

where $\gamma=C_p/C_v$ with $C_v$ is the specific heat coefficient at constant volume and $\widetilde{e}$ is the filtered internal energy.

The sub-grid scale stress tensor and the sub-grid turbulent heat flux are defined as:

$$\tau_{ij}^{sgs}=\overline{\rho}\left(\widetilde{u_i u_j}-\widetilde{u}_i\widetilde{u}_j\right) \tag{9}$$

$$q_i^{sgs}=\overline{\rho}\left(\widetilde{Tu_j}-\widetilde{T}\widetilde{u}_j\right) \tag{10}$$

The terms $\left(\overline{\tau}_{ij}-\widetilde{\tau}_{ij}\right)$, $\frac{\partial}{\partial x_i}\left(\overline{k}\frac{\partial\overline{T}}{\partial x_i}-\widetilde{k}\frac{\partial\widetilde{T}}{\partial x_i}\right)$ and $\frac{1}{2}\frac{\partial}{\partial x_i}\overline{\rho}\left(\widetilde{u_k u_k}u_i-\widetilde{u}_k\widetilde{u}_k\widetilde{u}_j-\tau_{kk}^{sgs}\widetilde{u}_i\right)$ are neglected due to its small contribution to the filtered conservation equations, following the results of [5, 6, 7, 8]. Finally, the total energy is approximated by [9, 10]:

$$\overline{\rho}\widetilde{E}=\overline{\rho}C_v\widetilde{T}+\frac{1}{2}\overline{\rho}\widetilde{u}_i\widetilde{u}_i+\frac{1}{2}\tau_{ii}^{sgs} \tag{11}$$

Based on the widely used sub-grid eddy viscosity model, the sub-grid stress tensor can be written as:

$$\tau_{ij}^{sgs} = 2\overline{\rho}\nu_{sgs}\left(\widetilde{S}_{ij} - \frac{1}{3}\widetilde{S}_{kk}\delta_{ij}\right) - \frac{1}{3}\tau_{kk}^{sgs}\delta_{ij} \tag{12}$$

where the rate of strain tensor $\widetilde{S}_{ij}$ is defined by:

$$\widetilde{S}_{ij} = \frac{1}{2}\left(\frac{\partial\widetilde{u}_i}{\partial x_j} + \frac{\partial\widetilde{u}_j}{\partial x_i}\right) \tag{13}$$

and the trace of sub-grid stress $\tau_{kk}^{sgs}$ tensor is modelled separately. The dynamic compressible Smagorinsky model is used in this work to evaluate the sub-grid viscosity $\nu_{sgs}$ and thus $\tau_{ij}^{sgs}$, $q_i^{sgs}$ and $\tau_{ii}^{sgs}$ [11].

The numerical solution of the flow equations is performed with the Characteristic-Based Split (CBS) scheme introduced by Zienkiewicz and Codina [12]. For the solution of the governing equations, the CBS algorithm uses a fractional step with a split. The four steps can be briefly described as [13]:

1. solve momentum equation without pressure terms,
2. calculate pressure solving a Poisson equation,
3. correct velocity components,
4. calculate additional scalar variables, such as temperature, from appropriate governing equations.

The time increment is assumed to be $\Delta t = t^{n+1} - t^n$ and the flux mass $U_i^{n+1} = \rho u_i$, evaluated at time $n+1$, is split into two terms:

$$U_i^{n+1} = U_i^n + \Delta U_i^* + \Delta U_i^{**} \tag{14}$$

A standard Galerkin finite element procedure is used for spatial discretization. Linear triangular and elements are employed in the present work. The spatial discretization of the variables is carried out as:

$$\begin{array}{llll} U_i = \mathbf{N}_u\widehat{\mathbf{U}}_i, & \Delta U_i = \mathbf{N}_u\Delta\widehat{\mathbf{U}}_i, & \Delta U_i^* = \mathbf{N}_u\Delta\widehat{\mathbf{U}}_i^*, & \Delta U_i^{**} = \mathbf{N}_u\Delta\widehat{\mathbf{U}}_i^{**} \\ u_i = \mathbf{N}_u\widehat{\mathbf{u}}_i, & p = \mathbf{N}_p\widehat{\mathbf{p}}, & \rho = \mathbf{N}_\rho\widehat{\rho} & \end{array} \tag{15}$$

where a hat superscript represents a nodal quantity and $\mathbf{N}_i$ are the shape functions such that:

$$\widehat{\mathbf{U}}_i = \left[U_i^1, U_i^2, \cdots, U_i^a, \cdots, U_i^m\right]^T \tag{16}$$

$$\mathbf{N}_j = \left[N_j^1, N_j^2, \cdots, N_j^a, \cdots, N_j^m\right] \tag{17}$$

with $a$ being a node (or variable), which varies from 1 to $m$, $i = 1, p$ and $j = u_i, p, \rho$.

The semi-discrete forms of the CBS equations are then weighted by $\mathbf{N}^T$ and integrated over the domain. The final equations for the four steps of the CBS scheme for the explicit formulation is summarized as:

- Step 1, obtain $\Delta\tilde{\mathbf{U}}_i^*$:

$$\mathbf{M}\Delta\widehat{\mathbf{U}}_i^* = \Delta t\Big[(\mathbf{C}-\mathbf{S})\left(\widehat{\mathbf{u}}_j\widehat{\mathbf{U}}_i\right) - (\mathbf{K}_\tau - \mathbf{T}_\tau)\,\mathcal{T}_{ij}\Big]^n - \frac{\Delta t^2}{2}\widehat{\mathbf{u}}_k\Big[(\mathbf{K}_u - \mathbf{T}_u)\left(\widehat{\mathbf{u}}_j\widehat{\mathbf{U}}_i\right) + (\mathbf{K}_p - \mathbf{T}_p)\,\widehat{\mathbf{p}}\Big]^n \tag{18}$$

- Step 2, obtain $\Delta\tilde{\rho}$:

$$\mathbf{M}\Delta\widehat{\rho} = \Delta t\Big[(\mathbf{D}-\mathbf{R}_u)\left(\widehat{\mathbf{U}}_i + \theta_1\Delta\widehat{\mathbf{U}}_i^*\right) - \Delta t\theta_1\mathbf{K}\widehat{\mathbf{p}}\Big]^n \tag{19}$$

- Step 3, obtain $\tilde{\mathbf{U}}_i$, thus establishing the values at $t^{n+1}$:

$$\mathbf{M}\widehat{\mathbf{U}}_i = \mathbf{M}\Delta\widehat{\mathbf{U}}_i^* + \Delta t[(\mathbf{D}-\mathbf{R}_u)\,\widehat{\mathbf{p}}]^n \tag{20}$$

- Step 4, solve energy equation to obtain the value $\Delta\tilde{\rho}\tilde{\mathbf{E}}$:

$$\mathbf{M}\Delta\widehat{\rho}\widehat{\mathbf{E}} = \Delta t\Big[(\mathbf{C}-\mathbf{S})\,\widehat{\mathbf{u}}_j\left(\widehat{\rho}\widehat{\mathbf{E}} + \widehat{\mathbf{p}}\right) - (\mathbf{T}_\tau - \mathbf{K}_\tau)\,(\mathcal{Q}_i + \mathcal{T}_{ij}\widehat{\mathbf{u}}_j)\Big]^n - \frac{\Delta t^2}{2}\Big[\widehat{\mathbf{u}}_k\,(\mathbf{K}_u - \mathbf{T}_u)\,\widehat{\mathbf{u}}_j\left(\widehat{\rho}\widehat{\mathbf{E}} + \widehat{\mathbf{p}}\right)\Big]^n \tag{21}$$

where $\mathcal{T}_{ij} = \tau_{ij} + \tau_{ij}^{sgs}$ is the total stress, $\mathcal{Q}_i = k\frac{\partial T}{\partial x_i} + q_i^{sgs}$ is the total heat flux and the vector and matrices are given by:

$$\mathbf{M} = \int_\Omega \mathbf{N}^T\mathbf{N}d\Omega \qquad \mathbf{R}_u = \int_\Omega \mathbf{N}^T\mathbf{N}n_i d\Omega \qquad \mathbf{S} = \int_\Gamma \mathbf{N}^T\mathbf{N}n_j d\Gamma \tag{22a}$$

$$\mathbf{C} = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_j}\mathbf{N}d\Omega \qquad \mathbf{D} = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_i}\mathbf{N}d\Omega \qquad \mathbf{K} = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_i}\frac{\partial\mathbf{N}}{\partial x_i}d\Omega \tag{22b}$$

$$\mathbf{K}_\tau = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_j}d\Omega \qquad \mathbf{K}_u = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_k}\frac{\partial\mathbf{N}}{\partial x_j}d\Omega \qquad \mathbf{K}_p = \int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_k}\frac{\partial\mathbf{N}}{\partial x_i}d\Omega \tag{22c}$$

$$\mathbf{T}_\tau = \int_\Gamma \mathbf{N}^T n_j d\Gamma \qquad \mathbf{T}_u = \int_\Gamma \mathbf{N}^T\frac{\partial\mathbf{N}}{\partial x_j}n_k d\Gamma \qquad \mathbf{T}_p = \int_\Gamma \mathbf{N}^T\frac{\partial\mathbf{N}}{\partial x_i}n_k d\Gamma \tag{22d}$$

with $\theta_1 = 0.5$ for a second-order accuracy (Crank-Nicolson scheme) in time for the velocity.

For transonic and supersonic speeds, an additional shock capturing dissipation is introduced to capture and smooth local oscillations in the vicinity of shocks. A method based on the Hessian of pressure is employed, modifying $\widehat{\mathbf{\Phi}}^{n+1} = \left\{\widehat{\rho}, \widehat{u}_i, \widehat{E}\right\}$ evaluated in time $n+1$ to $\widehat{\mathbf{\Phi}}_s^{n+1}$ by [14]:

$$\widehat{\mathbf{\Phi}}_s^{n+1} = \widehat{\mathbf{\Phi}}^{n+1} - \Delta t\mathbf{M}_L^{-1}C_e h^3\frac{|\mathbf{u}|+c}{p}\left|\frac{\partial^2 p}{\partial x_i^2}\right|_e^n\left(\int_\Omega \frac{\partial\mathbf{N}^T}{\partial x_i}\frac{\partial\mathbf{N}}{\partial x_i}d\Omega\right)\widehat{\mathbf{\Phi}}^n \tag{23}$$

where $h$ is the element size [14], $c$ is the local sound speed, $C_e$ is an user-informed constant and the subscript $e$ represents an element [1].

The following local time stepping is employed [13]:

$$\Delta t = \min\left(\frac{h}{c+|\mathbf{u}|}, \frac{h^2}{2\nu}\right)\theta_c \tag{24}$$

where $\theta_c$ is an user specified number between 0 and 1 used to not violate the Courant-Friedrichs-Lewy condition.

## 3 MESH ADAPTATION

Both the direction and the magnitude of the anisotropy are important and necessary information required for the construction of an anisotropic mesh adaptation procedure. These information can be evaluated through the use of the so called metric-based methods, where a Riemannian metric space is evaluated based on the interpolation error of the solution field. The most important features of the metric-based method employed in this work for two-dimensional problems is briefly described in this section. A more detailed overview of such methodology can be found in references [15, 2].

### 3.1 Metric Estimates

The length $\ell_{\mathcal{M}}(\mathbf{ab})$ of an edge $\mathbf{ab}$ can be continuously evaluated in a Riemannian metric space with a parametrization given by $\gamma(t) = \mathbf{a} + t\mathbf{ab}$ with $t \in [0,1]$ as:

$$\ell_{\mathcal{M}}(\mathbf{ab}) = \int_0^1 \|\gamma'(t)\|_{\mathcal{M}} dt = \int_0^1 \sqrt{\mathbf{ab}^T \mathcal{M}(\mathbf{a} + t\,\mathbf{ab})\,\mathbf{ab}}\, dt \tag{25}$$

where $\mathcal{M}$ is a $2 \times 2$ symmetric positive definite matrix called metric tensor, or just metric. The size $|T_K|_{\mathcal{M}}$ of an element $K$ evaluated in the Riemannian metric space is determined by:

$$|T_K|_{\mathcal{M}} = \int_K \sqrt{\det \mathcal{M}(\mathbf{x})} d\mathbf{x} \tag{26}$$

The anisotropic quality of an element $K$ can be monitored through a quality function $Q_{\mathcal{M}}$ that combines both sizing and orientation information [16]:

$$Q_{\mathcal{M}}(K) = \frac{\sum \ell^2_{\mathcal{M}}(A_K)}{|T_K|_{\mathcal{M}}} \tag{27}$$

where $A_K$ are the edges of an element $K$. In the previous equation, the numerator takes into account the sizing of the mesh. Decreasing the length of the edges evaluated in the Riemannian space also reduces $Q_{\mathcal{M}}$. The denominator is a measure of the orientation of the simplex in the Riemannian space, where a simplex oriented closer to the local field eigenvectors leads to a reduction of $Q_{\mathcal{M}}$. Thus, minimizing $Q_{\mathcal{M}}$ the quality of the element becomes maximized in an anisotropic sense.

From a discrete point of view, the metric field need to be interpolated to evaluate approximate length and volume in the Riemannian space. Considering a linear interpolation of the metric tensor, the integration of Eq. 25 is evaluated by [17]:

$$\ell_{\mathcal{M}}(\mathbf{ab}) \approx \frac{2}{3} \frac{\ell_0^2 + \ell_0 \ell_1 + \ell_1^2}{\ell_0 + \ell_1} \tag{28}$$

where $\ell_i(\mathbf{ab}) = \sqrt{\mathbf{ab}^T \mathcal{M}(\mathbf{x}_i)\mathbf{ab}}$ is the length of the edge in metric $\mathcal{M}(\mathbf{x}_i)$ with $i = 1, 2$ for the edge end-points $\mathbf{a}$ and $\mathbf{b}$, respectively. Similarly, the integral of Eq. 26 can be numerically approximated by:

$$|T_K|_{\mathcal{M}} \approx \sqrt{\det \frac{1}{3} \sum_{i=1}^{3} \mathcal{M}_i(\mathbf{x})}\, |T_K| \tag{29}$$

where $\mathcal{M}_i(\mathbf{x})$ is the metric at each of the $i$ vertices of the element $K$ (assumed as a triangle) with Euclidean oriented measure $|T_K|$.

The error $E_{\mathbf{L}^p}(\mathbf{ab})$ estimated for an edge $\mathbf{ab}$ is defined as the upper bound of the difference between a quadratic approximation and the evaluated linear interpolation measured in the $\mathbf{L}^p$ norm for a given continuous function $\phi$, which can be expressed as:

$$E_{\mathbf{L}^p}(\mathbf{ab}) \equiv \ell_{\mathcal{M}}(\mathbf{ab}) \tag{30}$$

with the metric $\mathcal{M}(\mathbf{x})$ defined as the normalized Hessian of the function $\phi$ measured in the $\mathbf{L}^p$ norm [18]:

$$\mathcal{M}(\mathbf{x}) = \mathcal{M}_{\mathbf{L}^p}(\mathbf{x}) = \left(\det|\mathbf{H}|\right)^{-\frac{1}{2p+d}} |\mathbf{H}| \tag{31}$$

where $\mathbf{H}$ is a symmetric matrix representing the Hessian of $\phi$. As $\mathbf{H}$ is a symmetric matrix, it can be decomposed into the product of orthonormal matrices $\mathcal{R}$ of associated eigenvectors and a diagonal matrix $\mathbf{\Lambda}$ of eigenvalues. To ensure that $\mathbf{H}$ is a positive-defined matrix and limited, $|\mathbf{H}| = \mathcal{R}\left|\tilde{\mathbf{\Lambda}}\right|\mathcal{R}^T$ is obtained as the Hessian matrix with normalized and limited eigenvalues $\tilde{\lambda}_i$ of $\tilde{\mathbf{\Lambda}} = \text{diag}\left(\tilde{\lambda}_i\right)$ such that [15]:

$$\tilde{\lambda}_i = \min\left(\max\left(\lambda_i, \frac{1}{h_{\max}^2}\right), \frac{1}{h_{\min}^2}\right) \tag{32}$$

with $h_{\max}$ and $h_{\min}$ as the maximum and minimum allowed edge size in the mesh and the Hessian matrix is evaluated by a double projection scheme using a weak formulation [19].

### 3.2 Metric Intersection

When more than one metric is simultaneously specified at one point, the resulted metric should cover the minimum error associated to each of these metrics. Therefore, a metric intersection procedure is used in this work. A common basis $\mathcal{P}$ of two metrics $\mathcal{M}_A$ and $\mathcal{M}_B$ is looked for, such that they are congruent to a diagonal matrix at its basis. The matrix $\mathcal{B} = {\mathcal{M}_A}^{-1}\mathcal{M}_B$ is introduced, such that $\mathcal{B}$ is diagonalizable with real eigenvalues. The normalized eigenvectors of $\mathcal{B}$ are $\mathbf{e}_i$ with $i = 1, 2$ and compose the common diagonalizable basis $\mathcal{P}$. The principal components of $\mathcal{M}_A$ and $\mathcal{M}_B$ projected in this basis can be obtained with [15]:

$$\mu_i = \mathbf{e}_i^T \mathcal{M}_A \mathbf{e}_i \quad \text{and} \quad \beta_i = \mathbf{e}_i^T \mathcal{M}_B \mathbf{e}_i \tag{33}$$

As $\mathcal{P}$ is also invertible, the metric intersection can be evaluated through:

$$\mathcal{M}_{A\cap B} = \mathcal{M}_A \cap M_B = \left(\mathcal{P}^{-1}\right)^T \text{diag}\left(\max\left(\mu_i, \beta_i\right)\right) \mathcal{P}^{-1} \tag{34}$$

For a third metric $\mathcal{M}_C$ to be intersected, the same procedure is performed, with $\mathcal{M}_{AB\cap C}$ employing $\mathcal{M}_{AB} = \mathcal{M}_{A\cap B}$ and so on for more metric intersections.

### 3.3 Mesh Refinement, Mesh Coarsening and Edge Swapping

The present mesh adaptation procedure uses the metric-based framework to evaluate the error associated to each edge of the mesh. When the error is above a specified value, the edge is split into two new edges. When the error is less than the desired, the edge is removed by an edge collapse procedure. Edge and face swapping are also used to improve the mesh quality.

When an edge $\mathbf{ab}$ is refined, a new node $\mathbf{c}$ is inserted generating two new edges: $\mathbf{ac}$ and $\mathbf{cb}$. A Riemannian centred is used, such that the new inserted node $\mathbf{c}$ verifies the following relation:

$$\ell_{\mathcal{M}}(\mathbf{ac}) = \ell_{\mathcal{M}}(\mathbf{cb}) \tag{35}$$

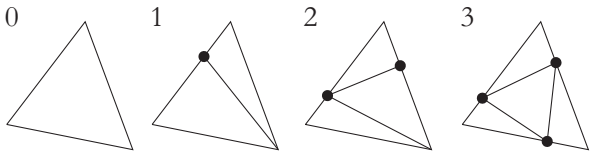



Figure 1: Edge-based anisotropic triangular subdivision scheme.

The spatial variation of the metric must be taken into account to solve the non-linear integral relation defined by Eq. 35. Moreover, using Eq. 25 and the parametrization $\mathbf{ac} = m\mathbf{ab}$ and $\mathbf{cb} = (1-m)\,\mathbf{ab}$ with $m \in [0,1]$, the Riemannian centred split defined in Eq. 35 can be evaluated by finding $m$ such that:

$$\int_0^1 \sqrt{\mathbf{ac}^T \mathcal{M}\left(\mathbf{a} + t\,\mathbf{ac}\right)\mathbf{ac}}\,dt = \int_0^1 \sqrt{\mathbf{cb}^T \mathcal{M}\left(\mathbf{c} + t\,\mathbf{cb}\right)\mathbf{cb}}\,dt \tag{36}$$

and considering a linear interpolation of the metric tensor along $\mathbf{ab}$, the value of $m$ is obtained as:

$$m = \frac{\ell_0^2 - 2^{-\frac{2}{3}}(\ell_0^3 + \ell_1^3)^{\frac{2}{3}}}{\ell_0^2 - \ell_1^2} \tag{37}$$

valid for all values of $\ell_0$ and $\ell_1$ such that $\ell_0 \neq \ell_1$. If $\ell_0 = \ell_1$, there is no size gradation and the Euclidean center is chosen for the split, with $m = 1/2$.

The refinement procedure must ensure that the topology of the mesh will not be violated, generating only valid elements. In order to satisfy this condition, the partitioning of an element is performed according to the number and position of the refined edges. All anisotropic partitioning cases are covered. For a mesh composed by triangular elements, four types of subdivision are possible, covering eight anisotropic cases (see Fig. 1). This type of edge subdivision is unique and can always be performed (existence and uniqueness conditions are satisfied), ensuring that the topology will not be violated.

An edge is coarsened by an edge collapse procedure. Let $\mathbf{ab}$ be an edge to be coarsened, with initial vertex $\mathbf{a}$ and final vertex $\mathbf{b}$. The vertices $\mathbf{a}$ and $\mathbf{b}$ are joined along the direction of the edge $\mathbf{ab}$ to a new point $\mathbf{c}$ between $\mathbf{a}$ and $\mathbf{b}$, collapsing all elements that share at least one edge with $\mathbf{ab}$. To hold the anisotropic information along the procedure, the choice of the collapsed point $\mathbf{c} = \mathbf{a} + n\mathbf{ab}$ is performed by finding $n \in [0,1]$ that minimizes the sum of all quality functions of the remaining elements affected by the collapse:

$$\min_{n\in[0,1]} \sum_{K\in B} Q_{\mathcal{M}}\left(K\right) \tag{38}$$

subjected to the following constraints:

$$|T_K| \;\forall\; K \in B > 0 \tag{39a}$$

$$\sum_{K\in A} |T_K| - \sum_{K\in B} |T_K| = 0 \tag{39b}$$

where $B$ is the remaining set of elements affected by the procedure after the collapse and $A$ the original set (see Fig. 2). A discrete approach is employed to solve Eq. 38 in this work. The continuous parametrization $n$ is replaced by $n_c \geq 2$ discrete positions $n_j \in [0,1]$:

$$n_j = \frac{(j-1)}{(n_c - 1)} \tag{40}$$

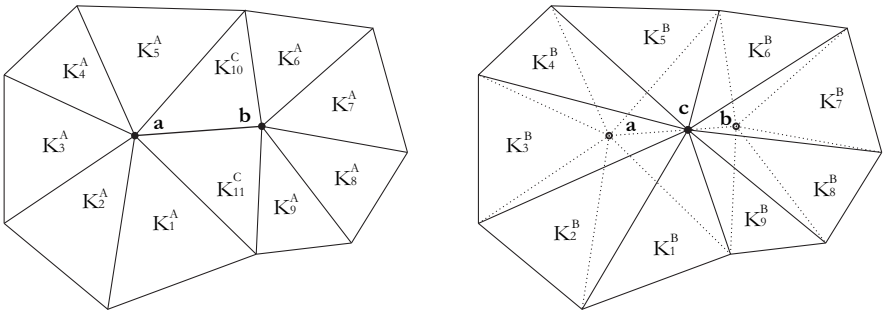



Figure 2: Edge collapse procedure.

with $j = 1, \ldots, n_c$. For each $n_j$ position, the summation of Eq. 38 is evaluated using Eq. 27 and considering a linear interpolation of the metric tensor between $\mathbf{a}$ and $\mathbf{b}$. Finally, the position $n_j$ which has the minimum value for this objective function and that does not violate the constraints of Eqs. 39 is chosen as the collapse point. This discrete approach is equivalent to choose between collapse the edge to one of the edge endpoints $\mathbf{a}$ or $\mathbf{b}$ if $n_c = 2$. For $n_c = 3$, the Euclidean midpoint is also considered as a possible position for the collapse and so on. In this work, $n_c = 10$ is adopted, allowing several possibilities for the choice of the position of point $\mathbf{c}$.

Edge swapping is performed in order to increase the mesh quality by alternating the edge connectivities. An edge $\mathbf{ab}$ that does not lye in the boundary of the domain shares two neighbour elements: $K_1$ and $K_2$. The internal edge of such elements can be swapped in order to obtain the elements $K_1'$ and $K_2'$ (see Fig. 3). The following condition is verified:

$$\max\left\{Q_{\mathcal{M}}\left(K'_1\right), Q_{\mathcal{M}}\left(K'_2\right)\right\} < \max\left\{Q_{\mathcal{M}}\left(K_1\right), Q_{\mathcal{M}}\left(K_2\right)\right\} \tag{41}$$

If such condition holds and no degenerated elements are obtained, then the swapping is performed, increasing the anisotropic quality of the mesh.

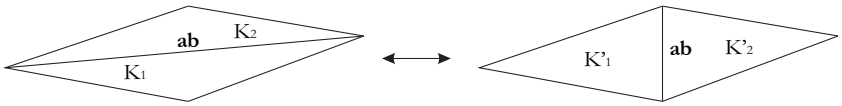



Figure 3: Edge swapping procedure.

### 3.4 Adaptation Algorithm

The overall algorithm implemented in this work performs the following sequence of local mesh modifications:

1. Refine edges with error above a given error threshold value $\eta_U$;
2. Edge swapping until convergence;
3. Coarse edges with error below a given error threshold value $\eta_L$;
4. Edge swapping until convergence.

which defines one adaptation iteration. An edge $\mathbf{ab}$ is refined if $E_{\mathbf{L}^p}(\mathbf{ab}) \geq \eta_U \varepsilon_{\mathbf{L}^p}$ and it is coarsened if $E_{\mathbf{L}^p}(\mathbf{ab}) \leq \eta_L \varepsilon_{\mathbf{L}^p}$, where $\varepsilon_{\mathbf{L}^p}$ is the target adaptation error measured in the

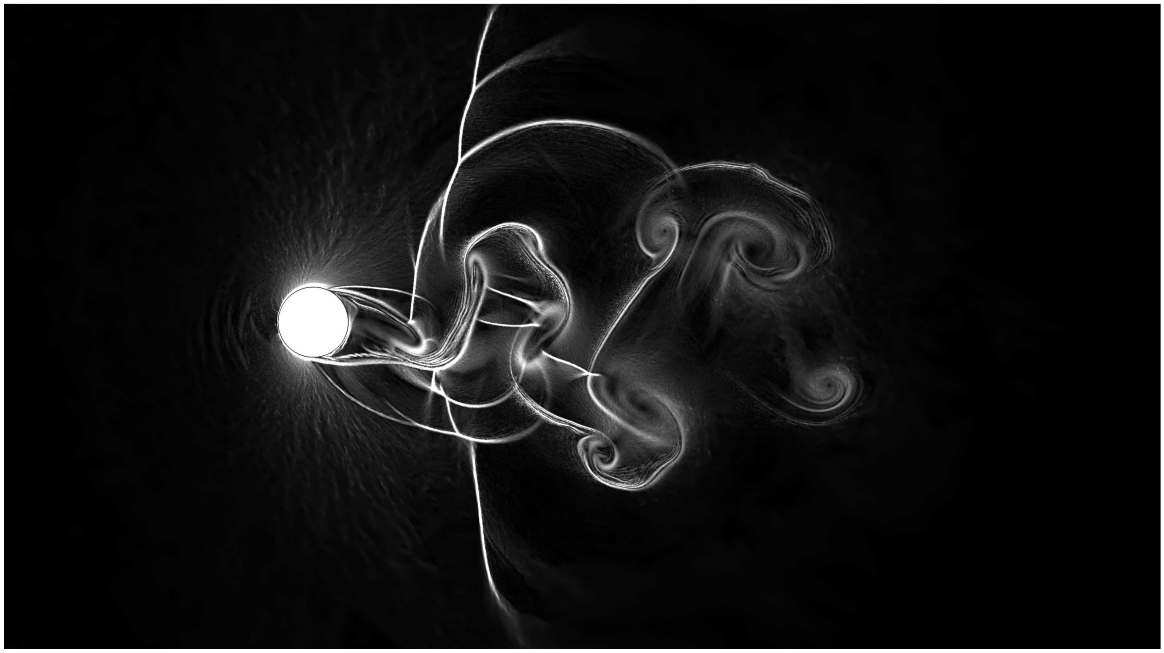
$\mathbf{L}^p$ norm. The upper and lower error threshold values $\eta_U$ and $\eta_L$ are assumed as $1.4$ and $0.6$, respectively [17]. As the investigated flows are compressible and turbulent, the continuous function $u$ chosen for the adaptation is the intersection of all conservative flow field variables, with $\phi = \rho \cap u_i \cap E$ for the error estimation in Eq. 31 and the adaptation is performed at each $t_a$ iterations of flow solver.

## 4 Test Case: Turbulent Transonic Flow Around a Circular Cylinder

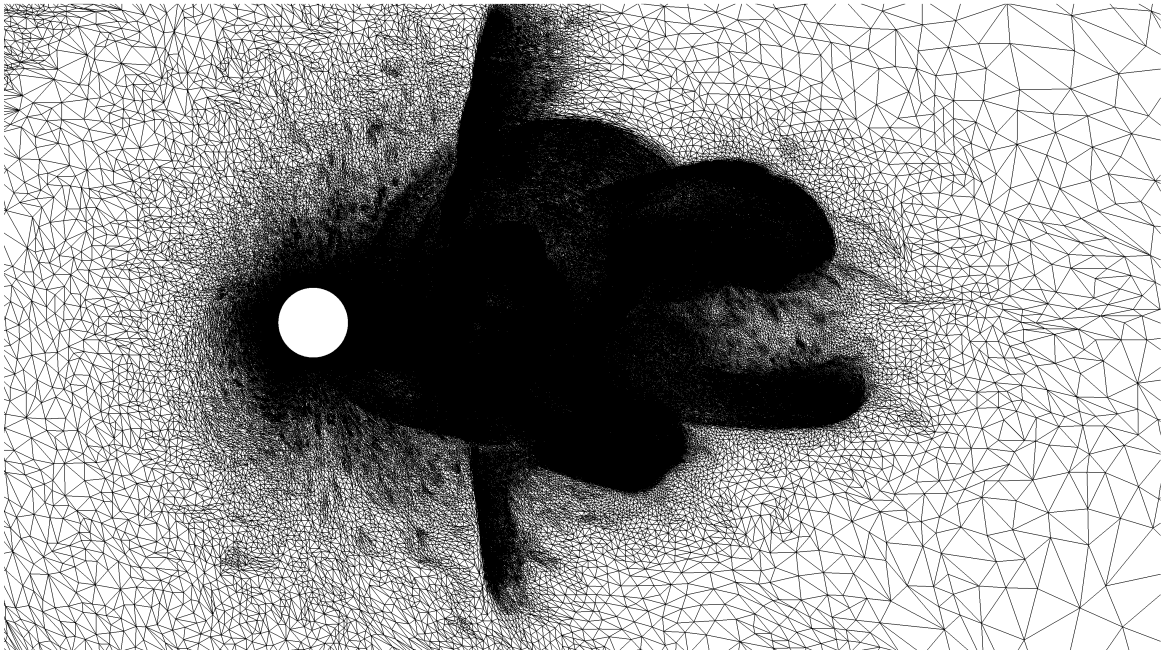
The studied test case consists of a turbulent transonic flow around a circular cylinder. The circular cylinder has a diameter of $D = 1$ with center located at coordinates $x = 0$, $y = 0$. The domain $\Omega$ is circular with radius $R = 25D$. The free stream Mach number employed is $M_\infty = 0.80$ and the free stream Reynolds number is $Re_\infty = 500.000$. The simulation is carried from the initial time set to zero to the final time $T = 50s$. Mesh adaptation is performed at each ten iterations of flow solver ($t_a = 10$) in order to control a target interpolation error of $\varepsilon_{\mathbf{L}^p} = 0.05$ measured in the $\mathbf{L}^2$ norm. The maximum and the minimum allowed edge sizes are $h_{\max} = 1D$ and $h_{\min} = 0.003D$, respectively.

Fig. 4 and Fig. 5 show the gradient magnitude of the mass density and the corresponding mesh at two different instants. In these figures, the complexity of the developed flow around the cylinder, involving complex viscous interactions with shock waves together with the high transient nature of such effects is evident. The separation point of the boundary layer is associated with the formation of lambda-shock-waves of different sizes and intensities. There is also the formation shock-waves linking and interacting with vortices rotating in opposite directions. Finally, there is the formation of a Von Kármán vortex sheet after a zone of intense interaction of viscous and shock-wave effects.

Fig. 6 and Fig. 7 show the unsteady lift $C_l$ and drag $C_d$ coefficients, respectively. The average drag coefficient obtained in the present work is $C_d = 1.61$ while the reference [20] experimental value is $C_d = 1.50$. The power spectra $P_l$ of the lift coefficient as function of the Strouhal number $St$ is shown in Fig. 8, where a pronounced pic close to the value of $St = 0.18$ is present. The experimental value for this pic response reported by [20] is also the exact same value. Finally, the power spectra $P_d$ of the drag coefficient as function of the Strouhal number $St$ is shown in Fig. 9. In this case, there is not only one pronounced pic.

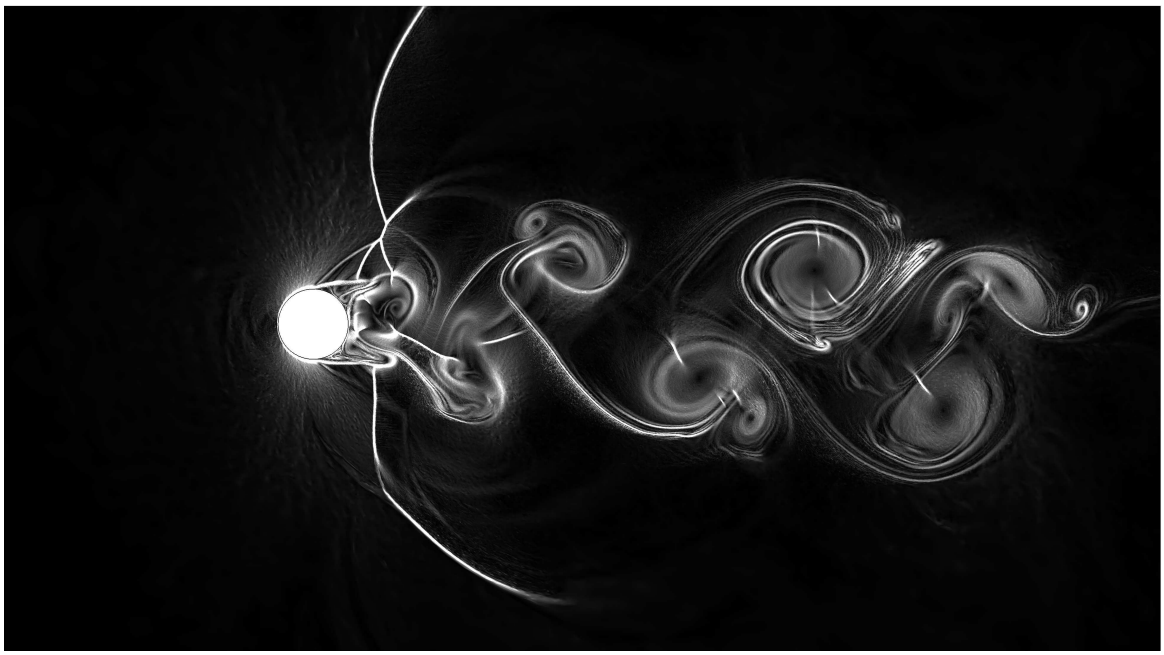


(a) Gradient magnitude of density. Scale: black (zero) to white (three).

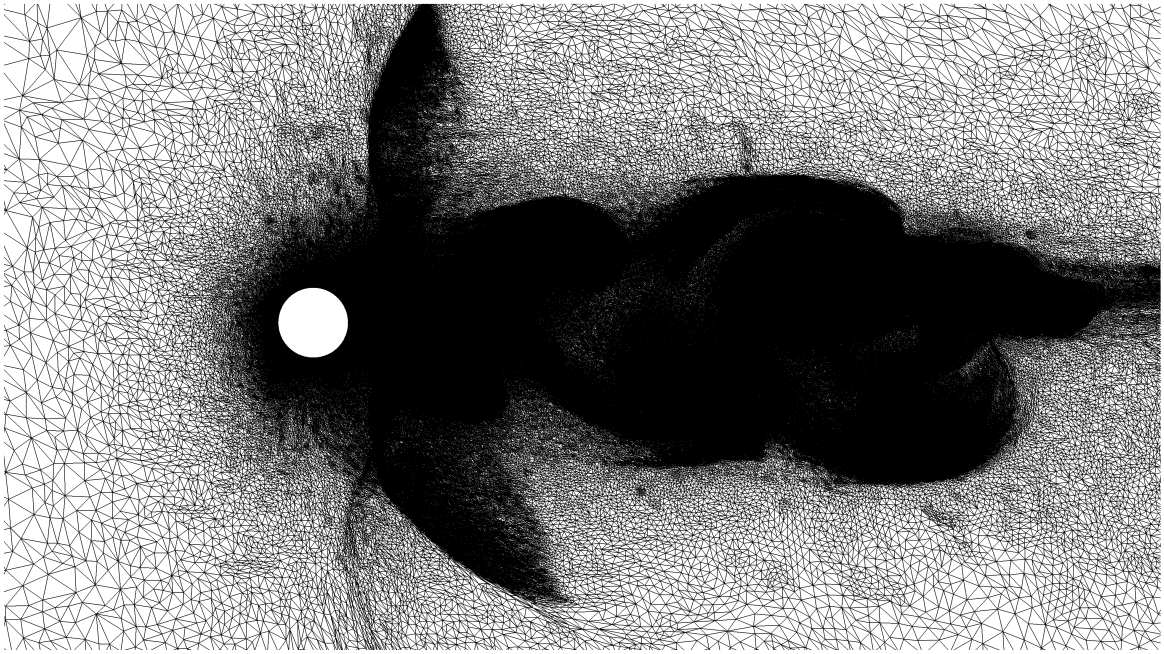


(b) Mesh

Figure 4: Flow field and mesh at time $t = 20s$.



(a) Gradient magnitude of density. Scale: black (zero) to white (three).



(b) Mesh

Figure 5: Flow field and mesh at time $t = 25s$.

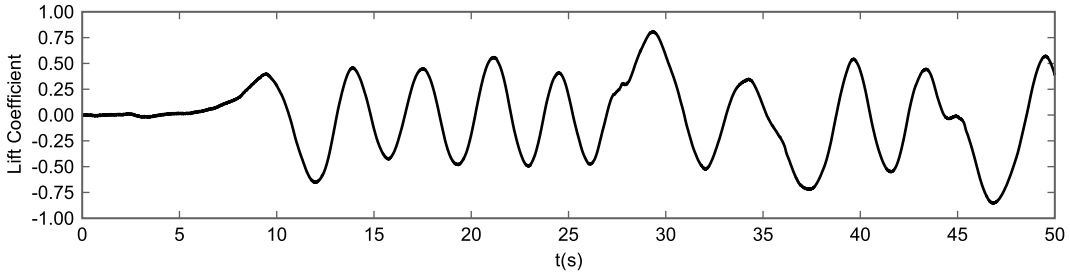



Figure 6: Unsteady lift coefficient.

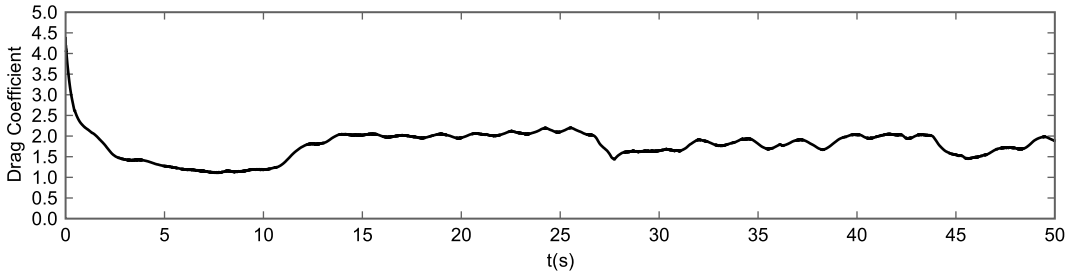



Figure 7: Unsteady drag coefficient.

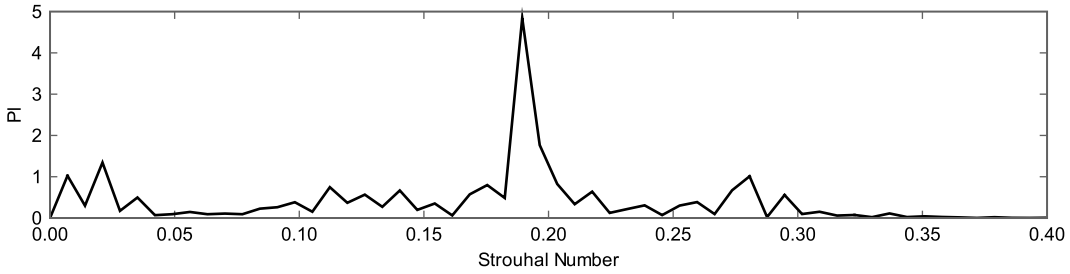



Figure 8: Power spectra of lift coefficient.

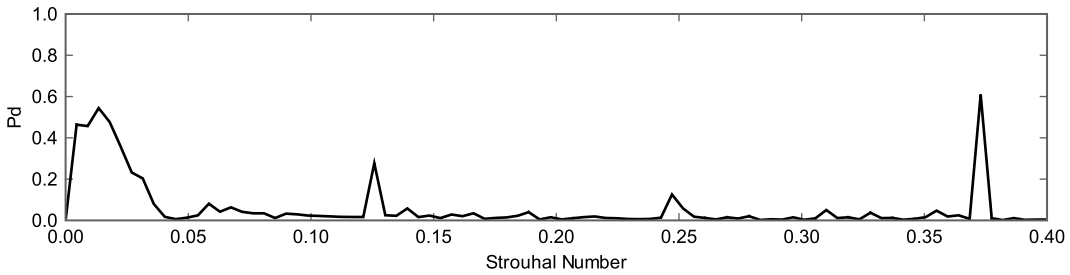



Figure 9: Power spectra of drag coefficient.

## 5 CONCLUSIONS

The coupling of the CBS algorithm with a dynamic Smagorinsky model for simulating turbulent compressible flows and anisotropic mesh adaptation is presented in this work. The developed code is used to investigate a complex turbulent transonic flow around a circular cylinder in a two-dimensional approach. The complex lamda-shock-waves, viscous interactions and Von Kármán vortex sheet effects are correctly captured by the mesh adaptation strategy, and the computed aerodynamic coefficients are in excellent agreement with the experimental reported values, even for a two-dimensional model. Although these preliminaries results are promising, an extension of the proposed methodology to a three-dimensional case is the next required step for validation of the computations.

**Acknowledgements**

The authors wish to thank to Propesq–UFRGS, CAPES and CNPq (Brazilian Research Committees) for their financial support and they thank also to CESUP (UFRGS Supercomputing Center) for its important contributions.

## REFERENCES

[1] P. Nithiarasu, R. Codina O.C. Zienkiewicz, The Characteristic-Based Split (CBS) scheme–a unifief approach to fluid dynamics. *International Journal for Numerical Methods in Engineering*, **66**,1514–1546, 2006.

[2] M. J. Castro-Diaz, F. Hiecht, B. Mohammadi, O. Pironneau, Anisotropic unstructured mesh adaptation for flow simulation. *International Journal for Numerical Methods in Fluids*, **25**,475–491, 1997.

[3] D. Knight, G. Zhou, N. Okong's, V. Shukla, Large Eddy Simulation of Compressible Flows Using Unstructured Grids. *International Journal of Computational Fluid Dynamics*, **13**,303–326, 2000.

[4] C. Hirsch, *Numerical Computation of Internal and External Flows, Vol 1, Fundamentals of Numerical Discretization*. Wiley, 1988.

[5] M.P. Martin, U. Piomelli, G. Candler, Sub-Grid Scale for Compressible Large Eddy Simulations. *Theoretical Computational Fluid Dynamics* **13**,361–376, 2000.

[6] S. Santhanam, S.K. Lele, J.H. Ferzinger, A Robust High-Order Compact Method for Large-Eddy Simulation. *Journal of Computational Physics* **191**,392–419, 2003.

[7] B. Vreman, B. Geurts, H. Kuerten, A Priori Tests of Large Eddy Simulation of the Compressible Plane Mixing Layer. *Journal of Engineering Mathematics*, **29**,299–327, 1995.

[8] B. Vreman, B. Geurts, H. Kuerten, Subgrid-Modeling in LES of Compressible Flow. *Applied Science Research* **54**,191–203, 1995.

[9] B. Kosovic, D.I. Pullin, R. Samtaney, Subgrid-Scale Modeling for Large Eddy Simulation of Compressible Turbulence. *Phys. Fluids*, **14**,1511–1522, 2002.

[10] T. Dubois, J.A. Domaradzki, A. Honein, The Subgrid Scale Estimation Model Applied to Large Eddy Simulations of Compressible Turbulence. *Physics of Fluids*, **14**,1781–1801, 2002.

[11] P. Moin, K. Squires, W. Cabot, S. Lee, A Dynamic Subgrid-Scale Model for Compressible Turbulence and Scalar Transport. *Physics of Fluids A* **11**,2746–2757, 1991.

[12] O.C. Zienkiewicz, R. Codina, A general algorithm for compressible and incompressible flow – part I: The split, characteristic-based scheme. *International Journal for Numerical Methods in Fluids*, **20**,869–885, 1995.

[13] O.C. Zienkiewicz, R.L. Taylor, P. Nithiarasu, *The Finite Element Method for Fluid Dynamics, 6th Edition*. Elseliver, 2005.

[14] P. Nithiarasu, O.C. Zienkiewicz, B.V.K. Satya-Sai, K. Morgan, M. Vázquez, Shock capturing viscosities for the general fluid mechanics algorithm. *International Journal for Numerical Methods in Fluids*, **28**,1325–1353, 1998.

[15] P. J. Frey, F. Alauzet, Anisotropic mesh adaptation for CFD computations. *Computer Methods in Applied Mechanics and Engineering*, **194**,5068–5082, 2005.

[16] C. Alexandra, V. Ducrot, Levelsets and anisotropic mesh adaptation. *Discrete and Continuous Dynamical Systems*, **30**,165–183, 2009.

[17] W. G. Habashi, J. Dompierre, Y. Bourgault, D. Ait-Ali-Yahia, M. Fortin, M. G. Vallet, Anisotropic Mesh Adaptation: Towards User-Independent, Mesh-Independent and Solver Independent CFD. Part I: General Principles. *International Journal for Numerical Methods in Fluids*, **32**,725–744, 2000.

[18] F. Alauzet, A changing-topology moving mesh technique for large displacements. *Engineering With Computers*, **30**,175–200, 2014.

[19] M. G. Vallet, C. -M. Manole, J. Dompierre, S. Dufour, F. Guibault, Numerical Comparison of some Hessian Recovery Techniques. *International Journal for Numerical Methods in Engineering*, **8**,987–1007, 2007.

[20] W.S. Murthy, C. Rose, From Drag, Skin Friction and Vortex Shedding Frequencies for Subsonic and Transonic Cross Flows on Circular Cylinder. *10th AIAA Fluid and Plasmodynamics Conference*, Albuquerque, NM, June 27-29, 1977.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF HYDROGEN JET INJECTION AND IGNITION IN SUPERSONIC FLOW

**Yulia V. Zakharova[1,2], Natalya N. Fedorova[1,2], Svetlana A. Valger[1]**

**Marat A. Goldfeld[2], Olga Vankova[2]**

[1] Novosibirsk State University of Architecture and Civil Engineering (Sibstrin)
Leningradskaya str., 113, 630008, Novosibirsk (Russia)
e-mail: swetla-ya@mail.ru

[2] Khristianovich Institute of Theoretical and Applied Mechanics SB RAS
Institutskaya st., 4/1 Novosibirsk (Russia)
{nfed, vankova, julsemen, gold}@itam.nsc.ru

**Keywords:** supersonic flows, mixing, combustion, numerical simulations, propulsion system.

**Abstract.** *In the present paper, the numerical and experimental study of the flows in combustion chamber has been performed. The experiments were conducted in the hot-shot wind tunnel. 3D numerical simulations were conducted by means of ANSYS FLUENT 16.0 on the RANS-based approach. The comparison and joint analysis have been performed for the calculated and experimental data. The gas-dynamic analysis has been carried out for channel flows with multiple hydrogen jet injection from walls. In the calculation, the influence of jet injection angle on the mixing process and structure of the supersonic flow in the channel is carried out. Preliminary 2D numerical results with hydrogen jet injection has shown that zone of combustion extend from the area of injection and cover the whole cavity. Further 3D studies were carried out with multiple hydrogen jet injection in a channel with backward-facing step. In this case the combustion layer extends over the bottom and top walls of the channel.*

## 1 INTRODUCTION

Scramjet (supersonic combustion ramjet) engine is one of the candidates for hypersonic flight propulsion system which will be used in wide range of flight Mach numbers from 4 to 12. Scramjet should be well operated in dual and scram modes depending on the flight Mach number [1]. In particular, scramjets are a promising technology that can enable efficient and flexible transport systems by removing the need to carry oxidizers and other propulsion limitations of conventional rocket engines. Supersonic combustion via the use of scramjet technology was successfully demonstrated for the first time worldwide by the HyShot II Program 2002 [2], followed by successful flights of the Hyper-X vehicles in 2004 at Mach number 6.8 and Mach number 9.6 [3]. Originally, many studies have been done for hydrogen-fueled concepts of supersonic combustion chamber, primarily for space transportation systems. In this case, the upper speed bound lies in the Mach number 12-16 range [4-5 and flow velocity at the entrance of combustor can achieve Mach number of 4-5. A lot of key issues have to be addressed for developing the dual mode ramjet technology. One of them consists in designing a combustion chamber able to operate in a very large Mach number range with enough high effectiveness.

The computer design tools are widely used in developing the optimum characteristics of the combustion chamber of different types [6], which must guarantee ignition and effective operation within the wide range of Mach numbers [7-8]. The main parameters to evaluate efficiency of the combustion chamber are the mixture degree, ignition condition, combustion efficiency and pressure losses. Results of experimental researches at different Mach numbers [9-10], however, have shown essential discrepancy between real characteristics of combustion chamber and anticipated values, in particular, in realization of self-ignition and achievement of high level of combustion efficiency.

The simulation of supersonic mixing and combustion, with a focus on the supersonic combustion ramjet (scramjet), poses many challenges. Even for non-reacting flows, the way to properly treat compressibility effects on turbulence at high Mach numbers is far from being resolved. Complex geometries posed additional difficulties. When combustion dynamics are added to the problem, uncertainty exists in the turbulence-chemistry interaction, shock wave-chemistry interaction, as well as the general lack of adequate knowledge of the effects of supersonic conditions on turbulence, reaction rates, and flame regimes. Review of these problems can be found in [11-12].

Backward facing step and cavity are widely used as effective flame holders at supersonic flow velocity in the channel. The results of numerical modelling and experimental investigations of high-enthalpy turbulent flows in the neighborhood of 90-degree BFS at the Mach numbers $M_\infty = 2-4$ are presented by authors in [13-14] under conditions of cold ($T_0$=300 K) and high-enthalpy ($T_0$=2500 K) external flows at $M_\infty = 2$, 3, and 4. It was found that temperature conditions can substantially affect the separation zone length and the flow structure behind the step. Different vortex structures of the separation zone behind the step were obtained, depending on the temperature factor. Nevertheless, it should be noted that these investigations were performed only for the external flow and for one step configuration.

The influence of step configuration on the structure of supersonic reacting flows in the channel under adiabatic and cold wall conditions was investigated in [15]. It is shown that ignition of mixtures essentially depends on the channel geometry and temperature conditions at the walls.

Results of modeling the interaction of a plane supersonic jet with a supersonic turbulent high-enthalpy flow in a channel are presented in [16]. Parametric studies show that an increase in the angle of inclination and the mass flow rate of the jet leads to an increase in the

depth of jet penetration into the flow, but more intense separated flows and shock waves are observed in this case.

The results of the numerical study on the cross-flow helium jet injection into a channel with BFS are presented in [17]. Significant 3D effects are revealed in the computations for the case of the round jet hole. It was shown that the main difference with the 2D case is the lower intensity of the interaction of the primary (air) and secondary flows due to the primary flow spreading around the jet.

The three common approaches for modeling turbulent flows usually use: and Reynolds-averaged Navier-Stokes equations (RANS) [18-19], direct numerical simulation (DNS) [20], large-eddy simulation (LES) [21-22]. It is well known that the RANS approach is the most computationally efficient and has a chance of completely modeling realistic aerospace systems. This is followed by LES, whereas DNS is still too costly for realistic engineering problems. However, the success of the RANS approach is dependent on type and complexity of flow in combustion chamber, and the procedure needs to be verified for every class of problem, making it non-universal [23]. Moreover, the approach is inherently steady and cannot deal with unsteady large-scale structures that determine the dynamics of many important flow problems. The main issue with LES, in comparison with RANS, is the computational cost.

There are a number of studies on hydrogen-fueled of scramjet, in which upper speed bound on flight Mach number was established in Mach 6-10 range. Mach number at the combustor entrance for this flight range should be 1.5-2.5. In spite of large amount of such investigations, the investigation of hydrogen combustion at high flow velocity is finding ever-increasing interest. Nevertheless, such researches were limited data for Mach numbers less than 3. The flow simulation in high-speed combustion chambers is one of important directions in these researches. The main goals of the present investigations were as follows:

- to get a new data on the flows in combustion channel;
- to investigate the influence of the injection angle, total temperature and Mach number on the self-ignition, flame propagation and combustion efficiency;
- to verify the numerical algorithm and code, including chemical schemes, on the channel flow experimental data.

## 2 COMBUSTOR MODEL AND EXPERIMENTAL SETUP

The combustor channel consists of rectangular channel with sudden expansion in the form of a BFS. The scheme of the model, position of injectors, and the basic dimensions of the combustor channel are presented in Figure 1. Model has two parts: multi-injector section and expanding section. In the first of part, fuel injection was carried out by means of sonic jets.

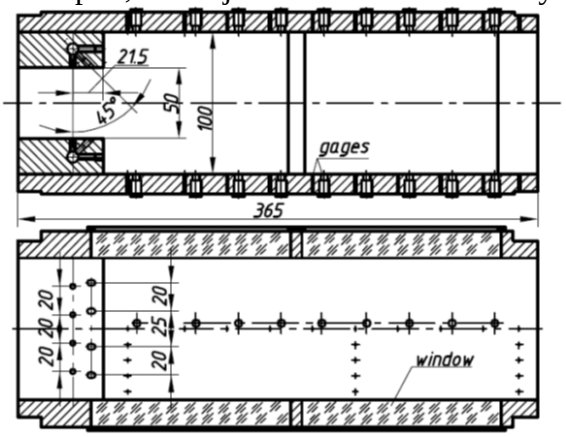



Figure 1: Model scheme of combustor and injector section.

The injecting holes have been developed as changeable elements and allowed to obtained different angles of fuel injection. Such a design enhances considerably the capabilities of the model. Variation of the fuel injection speed is possible by means of use different changeable fuel nozzles. Glass windows are placed along the whole inlet duct of the model to visualize the flow and to determine its structure (separation and attachment of the boundary layer, location and form of the flame). The system of fuel supply is intended for different types of fuel: hydrogen, kerosene and propane. Two separated systems can be used to supply the fuel into the main injectors and into the recirculation zone.

Test of combustor were performed at the hotshot wind tunnel IT-302M [24] with arc heating in the attached pipeline mode. Such mode of investigation allows an effective use of the advantages of the hotshot wind tunnel as a source of a high-enthalpy test gas (air). Photo of model and some element of the wind tunnel are shown in Fig. 2.

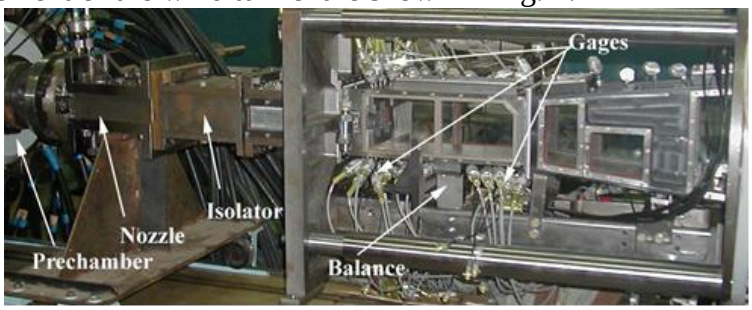



Figure 2: Elements of wind tunnel and model installation.

The choice of the initial values of air pressure in the discharge chamber and the voltage of capacitors allowed obtaining the required flow temperature. This approach ensures not only the necessary Mach number but also the required pressure and temperature at the combustor entrance. High parameters may be reached due to the absence of technological problems related to the temperature strength of the sharp edges of the model. In addition, an increase in the combustor dimensions allows one to extend significantly the capabilities of flow/flame diagnostics and to develop a modular principle of construction of the combustor for extending the range of parametric studies.

Peculiarity of this wind tunnel is the flow parameters which are falling during the operation time (100-150 ms). Therefore large numbers of runs were carried out with the pressure multiplicator for maintenance of constant equivalence ratio (ER) value. Models were tested at the following conditions at the duct entrance: Mach numbers $M_{en}$=2-4, total temperature $T_t$=2000-3000K, static pressure $P_{en}$=0.08-0.5 MPa, and fuel-air ER varied from 0.25 to 1.4.

During the tests, the parameters were measured as follows: the stagnation flow parameters in first and second prechambers; air and fuel flow rates; distributions of static pressure and heat flux in the model channel; Pitot pressure and temperature at model exit; base pressure distribution on backward facing step of injector device, forces acting on injector section. Large amount of measured stations allowed us to obtain detail distributions of static pressure and heat flux including transversal directions and base pressure. The video camera with frequency of about 800 frames per second was used for registration of flame images. The combustion efficiency was determined using an optical scheme of radiation registration in the ultraviolet range and one-dimensional calculation.

## 3 TESTING OF KINETIC SCHEMES

For numerical simulation of ignition of the hydrogen-air mixture it is necessary to use kinetic scheme that correctly predicts the ignition in supersonic external flow. Therefore, at

the first stage, several kinetic schemes were tested [25-27]. Calculations were provided in terms of experimental data [28] where hydrogen was injected into a supersonic flow through a tube of $d_1$ = 9.5 mm diameter. The tube was installed into a supersonic nozzle with diameter of the exit section d = 65.3 mm (Fig 3). The problem was solved as 2D axisymmetric. Computations were performed under the conditions presented in Table 1.

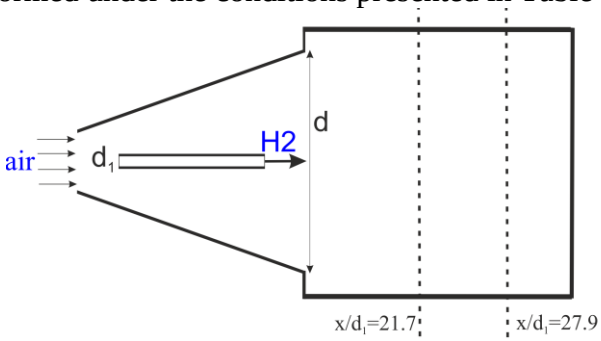



Figure 3: Calculation area for test problem.

| | Jet ($H_2$) | Free stream |
|---|---|---|
| Mach number, M | 2.00 | 1.90 |
| Static temperature, $T_{st}$, K | 251 | 1495 |
| Velocity, u, m/s | 2432 | 1510 |
| Static Pressure, $P_{st}$, MPa | 0.1 | 0.1 |
| Mass Fraction: | | |
| $H_2$ | 1.000 | 0 |
| $O_2$ | 0 | 0.241 |
| $N_2$ | 0 | 0.478 |
| $H_2O$ | 0 | 0.281 |

Table 1: Flow parameters for test problem.

Comparison of the computed and experimental data on $H_2O$ mass fraction in several cross-sections is provided in Fig. 4. The best agreement between the calculated and experimental data was obtained using the kinetic scheme with 19 direct and 19 reverse reactions [27]. Therefore this kinetic scheme was used for further calculations.

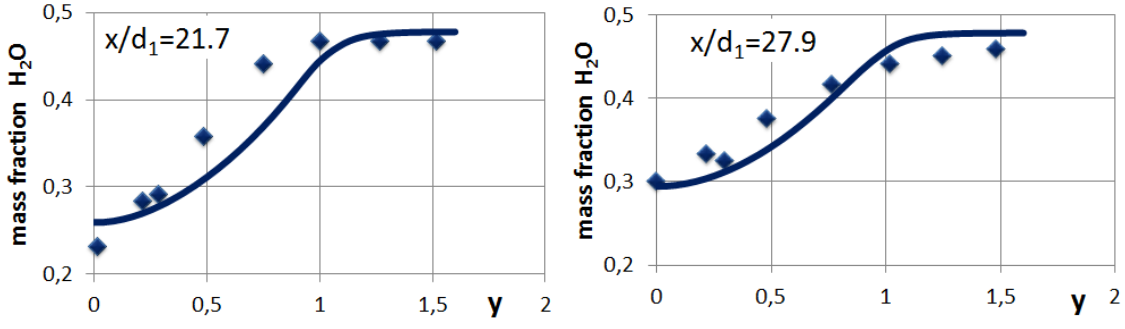



Figure 4: Comparison of experimental (symbols) and numerical (line) data on $H_2O$ mass fraction for test problem.

## 4 HYDROGEN JET INJECTION IN CHANNEL WITH CAVITY

Parametric computational studies of 2D supersonic air flow in a channel with a cavity were conducted. Geometry and flow parameters are conditioned by experiments provided in the hotshot wind tunnel IT-302M. Computational domain is a channel with a cavity located at the

bottom wall. The computational domain (Fig. 5) was bounded by the inlet cross section on the left, by the walls on the top and bottom, and by the outlet boundary on the right. At the inlet section, the turbulent flow parameters at Mach 2.8 were set accounting for boundary layer presence on the top and bottom walls. On the solid walls, the no-slip conditions and cold wall conditions $T_w$=300 K were assigned.

Hydrogen jet was injected from a slot of 2mm width located 50 mm prior the cavity front face (Fig.5). The investigation of injection angle influence on a flow in a channel was carried out in 2D approach. Computational results were obtained on the RANS-based approach closed by the k-$\omega$/SST model. The AUSM flux vector splitting scheme of second order was used for convective term approximation and an implicit temporal approximation is utilized for time integration. A structured grid with quadrilateral cells was used; the grid was refined toward the solid surfaces. The refinement factor was chosen in such a way that the dimensionless distance to the wall in the first computational node was $y_1^+$~1, and the laminar sublayer contained approximately ten computational nodes, which provided a fairly accurate resolution of the turbulent boundary layer in the near-wall region. Hydrogen was injected normally to the main flow. The detailed kinetic scheme [27] was used for modeling of combustion.

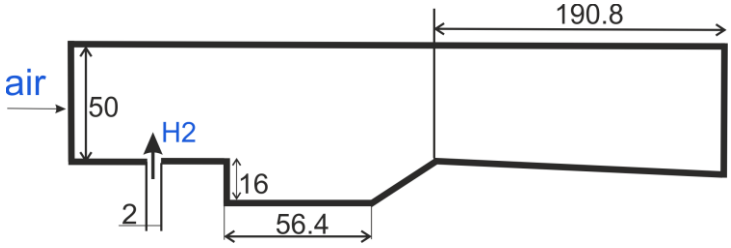



Figure 5: Channel with cavity.

The parameters of the examined flow were following: Mach number $M_\infty$ = 2.8, the total pressure $P_0 = 8.14\times10^5$ Pa and stagnation temperature $T_0$ = 1798 K. Hydrogen was injected at an angles $\alpha$ = 30°, 60° and 90° at sonic velocity with the following parameters: the total pressure $P_0 = 9\times10^5$ Pa and static temperature T = 300 K. The fields of main parameters allow estimating the influence of hydrogen jet injection angle on wave structure and combustion process in a channel. Hydrogen jet injection leads to reconstruction of the shock wave in a channel and forming an additional separation zones. The Mach number contours for the three injection angles of $\alpha$=30°, 60° and 90° are presented in Fig. 6.

The shock wave (1) that formed around the jet injection falls on the upper channel wall and then is reflected from it. Reflected shock wave (2) falls onto the mixing layer forming a subsonic region (3) in a vicinity of the rear cavity wall (Fig.6, a). The repeatedly reflected shock (3) comes to the upper wall and induces a local separation zone (4). Inside a cavity, a flow remains subsonic.

In case of 60° angle, the shock (1) is formed in front of the hydrogen jet. Furthermore, a local separation zone (3) is observed on the bottom wall before the hydrogen jet (Fig.6, b). Shock (1) falls on the upper wall and interacts with the shock (2) thereby a $\lambda$ - configuration is formed.

When hydrogen is injected at an angle of 90° the flow is decelerated to a subsonic speed, as evidenced by the absence of shocks in the channel after the cavity. The presence of local subsonic regions contributes to a better mixing.

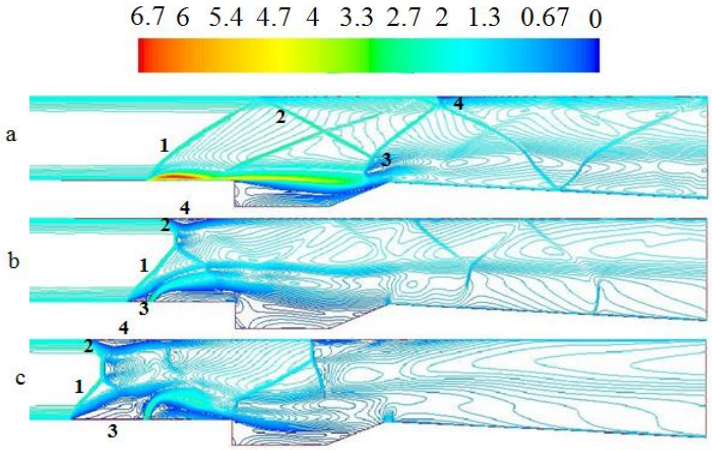



Figure 6: Contours of Mach number for different angles of hydrogen injection
a) 30°, b) 60°, c) 90°.

Further, the jet penetration depth and mixing rate of the primary and secondary flows can be estimated from the hydrogen mass fraction fields (Fig. 7). The mixture is ignited immediately behind the zone of hydrogen injection. Mixing and combustion layer extends over the cavity and along the bottom wall of the channel. When hydrogen was injected at an angle 90° the combustion process captures the whole cavity. Analysis of the $H_2$ and $O_2$ concentrations has shown that over the area of the combustion the mixture is lean that mean a lot of oxidant and lack of fuel. Inside a cavity the mixture is rich with a lack of oxidant.

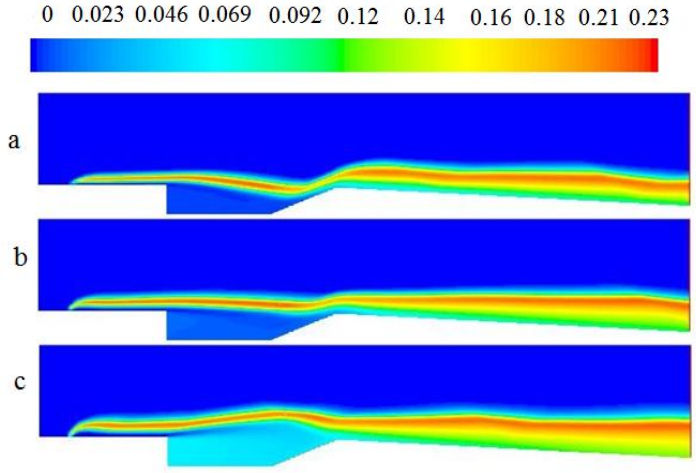



Figure 7: Mass fractions of $H_2O$ for different angles of hydrogen injection
a) 30°, b) 60°, c) 90°.

Modeling of multiple hydrogen jet injection in a channel with BFS was carried out in 3D approach. Channel geometry is matched with the experimental model (Fig.1). The calculation were conducted for the inlet parameters as follows: free stream Mach number $M_\infty = 4$, static pressure $P = 1.06\times10^5$ Pa and static temperature T = 600 K. Hydrogen was injected normally to the main flow at: static pressure $P = 25.8\times10^5$ Pa and static temperature T = 300 K. For modeling of combustion, the detailed kinetic scheme [27] was used.

The computed mass fraction of $H_2O$ in plane of symmetry is presented in Fig.8. It can be seen that the ignition of the hydrogen-air mixture takes place directly behind the step. Further, the flame propagates along the channel walls capturing the mixing layer. Near the exit boundary there is intense burning throughout the volume.

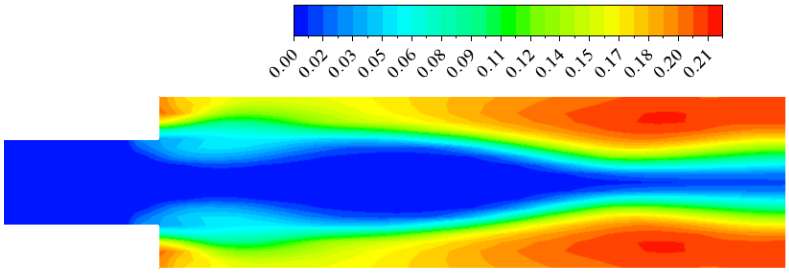



Figure 8: Mass fractions of $H_2O$ in 3D problem.

## 5 CONCLUSIONS

- 2D and 3D computational analysis has been performed for the problem of hydrogen jet injection into the supersonic flow in a channel with abrupt expansion.
- Several kinetic schemes were tested in terms of experimental data.
- The jet injection angle growth results in reorganization of the shock wave structure in a channel.
- In case of multiple injections the combustion layer extends over the bottom and top walls of the channel.

## 6 ACKNOWLEDGMENTS

The study was supported by Russian Science Foundation, Grant № 16-19-00010.

## REFERENCES

[1] Curran, E.T., Murthy, S.N.B. (Editors). *Scramjet Propulsion. Progress in Astronautics and Aeronautics*, **189**, 2001.

[2] R. R. Boyce, S. Gerard, and A. Paull, The HyShot Scramjet Flight Experiment—Flight Data and CFD Calculations Compared. *AIAA Paper 2003-7029*, 10, 2003.

[3] C. R. McClinton, X-43—Scramjet Power Breaks the Hypersonic Barrier: Dryden Lectureship in Research for 2006. *AIAA Paper 2006-1317*, 2006.

[4] Billig, F. Research on Supersonic Combustion. *J. of Propulsion and Power*. **9**(4): 499-514. 1993.

[5] Curran, E.T., Heiser, W.H., Pratt, D.T. Fluid Phenomena in Scramjet Combustion Systems. *Annual Review of Fluid Mechanics*. **28**, 1996.

[6] R. A. Baurle, Fuller,R. P., White, J. A., Chen,T. H., Gruber,M. R., and Nejad, A. S., An Investigation of Advanced Fuel Injection Schemes for Scramjet Combustion. *AIAA Paper 1998-0937*, 10, 1998.

[7] T. Cain and C. Walton, Review of Experiments on Ignition and Flameholding In Supersonic Flow, 38th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit, 7-10 July, Indianapolis, Indiana *AIAA 2002-3877*, 10, 2002.

[8] C. P. Goyne, C. G. Rodriguez†R. H. Krauss and J. C. McDaniel, C. R. McClinton, Experimental and Numerical Study of a Dual-Mode Scramjet Combustor , *Journal Of Propulsion and Power*, **22**, No. 3, May-June 2006.

[9] M.R. Gruber, J.M. Donbar, C.D. Carter, K.-Y. Hsu, Mixing and combustion studies using cavity-based flame holders in a supersonic flow, *Journal Propulsion and Power*, **20**, No. 5, 769–778, 2004.

[10] P.J. Ortwerth, A.B. Mathur, C. Segal, M.G. Owens, Combustion stability limits of hydrogen in a non-premixed, supersonic Flow, Proceedings of the 14th *International Symposium on Airbreathing Engines (ISABE)*, No. 1999-7137, 9, 1999.

[11] R. A. Baurle, Modeling of High Speed Reacting Flows: Established Practices and Future Challenges, *AIAA Paper 2004-0267*, 14, 2004.

[12] J. P. Drummond, Methods for Prediction of High-Speed Reacting Flows in Aerospace Propulsion, *AIAA Journal*, **52**, No. 3, 465-484, 2014.

[13] I.A. Bedarev, M.A. Goldfeld, Yu.V. Zakharova and N.N. Fedorova Investigation of temperature fields in supersonic flow behind a backward-facing step. *Thermophysics and Aeromechanics*, **16**, No. 3, 2009.

[14] Goldfeld M.A., Zakharova Yu.V., Fedorova N.N. A Numerical and Experimental Study of the High-Enthalpy High-Speed Cavity Flow. *Thermophysics Aeromechanics*, **19**, No. 6, 541 - 554. 2012.

[15] N. N. Fedorova, I. A. Bedarev, Yu. V. Zhakharova and M. A. Goldfeld Step configuration influence on combustion in premixed hydrogen-air supersonic flow. *ECCOMAS 2012*, September 10-14, Vienna, Austria, 2012.

[16] N. N. Fedorova, I. A. Fedorchenko, A. V. Fedorov Mathematical modeling of jet interaction with a high-enthalpy in an expanding channel, *Journal of Applied Mechanics and Technical Physics*, **54**, No. 2, 195–206, 2013.

[17] Irina A. Fedorchenko, Svetlana A. Valger, Dennis Daub Numerical study of helium jet injection into channel with sudden expansion, 2nd ECCOMAS Young Investigators Conference (YIC 2013), Bordeaux: France, 2 – 6 September 2013.

[18] Baurle, R. A., and Edwards, J. R., "Hybrid Reynolds-Averaged/Large-Eddy Simulations of a Coaxial Supersonic Free jet Experiment, *AIAA Journal*, **48**, No. 3, 551–571, 2010.

[19] V Candler, P.K. Subbareddy, J.M. Brock, Advances in Computational Fluid Dynamics Methods for Hypersonic Flows, *Journal of Spacecraft and Rockets*, **52**, No. 1, 17-28, 2015.

[20] Georgiadis, N., Yoder, D., Vyas, M., and Engblom, W., "Status of Turbulence Modeling for Hypersonic Propulsion Flowpaths, *AIAA Paper No*. 2011-5917, 2011.

[21] Peterson, D. M., and Candler, G. V., "Hybrid Reynolds-Averaged and Large-Eddy Simulation of Normal Injection into a Supersonic Crossflow," *AIAA Journal*, **26**, No. 3, 533–544, 2010.

[22] N. Georgiadis, D. Rizzetta, and C. Fureby, Large-Eddy Simulation: Current Capabilities, Recommended Practices, and Future Research, *AIAA Journal*, **48**, No. 8, 1772–1784, 2010.

[23] F. Ladeinde, X. Cai, and A. Alabi, A Critical Review of Scramjet Combustion Simulation, *47th AIAA Aerospace Sciences Meeting, AIAA Paper* 2009-127, 14, 2009.

[24] L.N. Pusyrev, M.I. Yaroslavtsev, Stabilization of Gas Parameters in Prechamber of Hypersonic Hot-Shot Wind Tunnel, *Izvestia Academii Nauk*, No.5, 86-93, (In Russian). 1990.

[25] Chul Park, Review of Chemical-Kinetic Problems of Future NASA Missions, I: Earth Entries// *Journal of Thermophysics and Heat Transfer*, **7**, No. 3, July-Sept. 1993.

[26] Foluso Ladeinde, A Critical Review of Scramjet Combustion Simulation (Invited). *AIAA Paper* 209-127, 13.

[27] Tien J. H., Stalker R. J., Release of Chemical Energy by Combustion in a Supersonic Mixing Layer of Hydrogen and Air// *Combustion and Flame.* No. 130, 329–348. 2002.

[28] John S. Evans, Charls J. Schexnayder Jr. Influence of Chemical Kinetics and Unmixedness on Burning in Supersonic Hydrogen Flames // *AIAA Jour*, **18**, No. 2, 188-193.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A 3D UNSTEADY NUMERICAL SIMULATION OF THE REFRIGERANT FLOW EVAPORATION IN A PLATE HEAT EXCHANGER

**Mirza Popovac, Gerwin Schmid and Michael Lauermann**

Austrian Institute of Technology
Giefinggasse 2, 1210 Vienna, Austria
mirza.popovac@ait.ac.at

**Keywords:** Multi-phase, pseudo-fluid, phase-change, CoolProp, OpenFoam, CFD.

**Abstract.** *This work presents a robust CFD approach for simulating the refrigerant flow evaporation within a plate heat exchanger of practical relevance with respect to the geometry and operating conditions. The purpose of this work is to assess the applicability of the numerical fluid flow simulations in predicting the thermal behavior and the fluid flow performance (including the phase change) of a plate heat exchanger with a realistic geometry and under realistic operating conditions. To that purpose, the basic structure of the CFD solver necessary for the numerical simulation of the occurring fluid flow phenomena is discussed here, together with the corresponding boundary conditions, fluid properties treatment, and related geometry issues. The results obtained for a specific plate heat exchanger are validated against experimental results found in available literature.*

## INTRODUCTION

One of the key components in various thermal machines are plate heat exchangers, which have the purpose to transfer the heat from a hot to a cold fluid. Typically it is expected that these heat exchangers exhibit the highest possible heat transfer coefficient with lowest possible hydraulic losses. The present paper focuses on a numerical method for analyzing the thermal behavior of a plate heat exchanger with refrigerant flow evaporation: while flowing through the plate channels of a heat exchanger, the refrigerant is evaporating and simultaneously cooling down the liquid (in this case the water) on the opposite side of the plate. The aim of the present work is to assess the applicability of this numerical approach, based on the computational fluid dynamics (CFD), to analyze the performance of plate heat exchangers of realistic geometries and under realistic operating conditions.

A great challenge in this work comes from the requirement for a realistic representation of the geometric features of the plate channels within a single layer of analyzed plate heat exchanger. On the one hand there is a huge variation in the characteristic length scales (high length to height ratio of the plate layer), and on the other there is an interlacing wavy geometry of the plate channels. Both of these aspects are setting high demands on the meshing, and as a result sufficiently high mesh quality cannot always be assured.

The most critical part of the present work, however, is the modeling of analyzed two-phase phenomena. Given the complexity of the desired application, in this paper the focus is on the robustness of the physical description of the problem under consideration, and the corresponding stability of the numerical simulation. Therefore, in the selected approach the two nonmiscible phases (refrigerant liquid and vapor phase) are treated as a single fluid with combined transport properties. The properties of individual phases are obtained from the *CoolProp* simulation platform for solving fluid equations of state and transport properties [1]. For thus defined pseudo-fluid, the governing equations for continuity, momentum and energy conservation are solved [2]. The phase-change process is described through the transfer function, which represents the volume fraction of a single phase. The assumption being introduced here is to express this function explicitly through the local temperature (comparing it with the saturation temperature), and implicitly include the related phase-change effects through the physical properties of the involved phases (e.g. the latent heat). The abovementioned approach significantly improves the numerical performance of the simulation method, in particular when the numerical mesh quality and the computational time requirements are concerned.

The numerical results, obtained using *OpenFOAM*® library suite [3], are validated against the available experimental results found in literature. The preliminary results using the adopted approach are showing that the basic flow pattern and the phase distribution within the plate heat exchanger channels can be identified. Comparing the results of the experiments and simulations, which are quantified through the integral values of momentum and heat transfer, the overall agreement is satisfactory.

## 1 HEAT EXCHANGER GEOMETRY

The creation of the plate heat exchanger geometry is very challenging task, and the meshing of this geometry is even more demanding. Nevertheless, this is the starting point for a numerical simulation, and it is a critical element when the quality of the simulation results is concerned. For a realistic representation of the geometric features of a single layer of the heat exchanger under investigation, *SolidWorks*™ package has been used to create a parametric model with characteristic width, length and chevron angle.

Shown in Figure 1 is the plate heat exchanger model of Huang et al. [4], which has been considered in the present analysis. Due to its highly complex geometry, only one layer of this heat exchanger has been included in the analysis: this layer consists of the channels through which the refrigerant is flowing, while the influence of the water on the secondary side has accounted for through the appropriate boundary condition. In this calculation the incoming and outgoing zones of the analyzed layer (yellow) were slightly modified with respect to the real geometry: there is no dedicated manifold distributor near the inlet or outlet of the heat exchanger, but rather the same geometry as throughout the whole plate (same channel size and chevron angle). Through the inlet (blue) and outlet (green) sections the operating conditions have been imposed. The remaining part (red) represents the walls through which the heat transfer with the secondary fluid (water) is taking place.

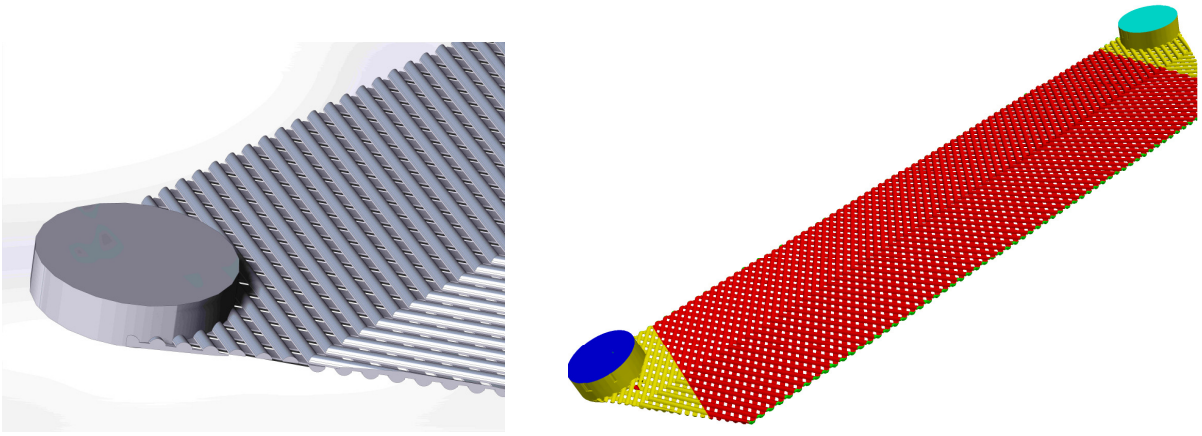


Figure 1: plate heat exchanger geometry detail (left), and the numerical mesh (right).

Using the geometry created as explained above, the numerical mesh has been created using *SnappyHexMesh* tool from the *OpenFoam* suite. For the given complexity of the simulated geometry, the only feasible meshing process is to use this automatic meshing approach, although its immanent feature is that they cannot assure the overall mesh quality (and the mesh quality will influence the stability and quality of the numerical simulation - especially in the case of the multi-phase flow simulations). One possibility of improving the mesh quality is to go for higher refinement levels (finer mesh), but that implies larger meshes (number of cells) and longer simulation times.

## 2 NUMERICAL FLUID FLOW SOLVER

For the present numerical analysis the open-source package *OpenFoam* has been used: it is based on the finite-volume method, and owing to its object-oriented structure it enables straightforward implementations of the necessary improvements (in this case, a realistic treatment of refrigerant properties, and the flow evaporation model). For the simulation of the refrigerant flow evaporation within a plate heat exchanger, the transient flow solver has been selected, based on the combined PISO-SIMPLE method.

The problem under consideration involves two nonmiscible phases of a single constituent: refrigerant liquid and vapor phase. In the adopted approach, these two phases are treated as a pseudo-fluid (with combined fluid properties), for which the time scales of the transport between the phases is much smaller than the integral time scale of the flow. This assumption

invokes the thermodynamic equilibrium requirement, in which case it is shown that the two-phase flow obeys the governing laws of a single-phase flow [5]. In other words, the flow is described with the equations for conservation of mass, momentum and enthalpy, as given below:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u_j}{\partial x_j} = 0 \tag{1}$$

$$\frac{\partial \rho u_i}{\partial t} + \frac{\partial \rho u_j u_i}{\partial x_j} = \frac{\partial}{\partial x_j}\left[(\mu + \mu_t)\frac{\partial u_i}{\partial x_j}\right] - \frac{\partial p}{\partial x_i} \tag{2}$$

$$\frac{\partial \rho h}{\partial t} + \frac{\partial \rho u_j h}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\frac{\lambda}{c_p} + \frac{\lambda_t}{c_p}\right)\frac{\partial h}{\partial x_j}\right] \tag{3}$$

The enthalpy approach is a conservative approach which is valid in a general phase-change case [6]. The enthalpy includes both the latent heat $H$ and the sensible heat (measured by the temperature $T$): $h = c_p T + H$, whereby the latent heat $H$ is being treated explicitly during the phase-change process. The combined properties of pseudo-fluid (density $\rho$, viscosity $\mu$, specific heat capacity $c_p$, thermal conductivity $\lambda$) are defined as a weight mixture of the individual fluid properties: $\phi = \phi_{vapor}\,\alpha_{vapor} + \phi_{liquid}\,\alpha_{liquid}$ ($\phi = \rho, c_p, \mu, \lambda$). The weighting function is represented by the volume fraction of a single phase $\alpha$, with $\alpha_{vapor} = 1 - \alpha_{liquid}$.

The effects of turbulence are introduced through the eddy viscosity $\mu_t$, here defined using the $k$-$\varepsilon$ model [7] with the realizability constrain imposed through the time scale $T_r$:

$$\mu_t = \rho c_\mu k T_r \tag{4}$$

$$\frac{\partial \rho k}{\partial t} + \frac{\partial \rho u_j k}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] + P_k - \rho\varepsilon \tag{5}$$

$$\frac{\partial \rho \varepsilon}{\partial t} + \frac{\partial \rho u_j \varepsilon}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma\varepsilon}\right)\frac{\partial \varepsilon}{\partial x_j}\right] + \frac{C_1 P_k - C_2 \rho\varepsilon}{T_r} \tag{6}$$

$$T_r = \max\left[\frac{k}{\varepsilon}, C_\tau\left(\frac{\nu}{\varepsilon}\right)^{1/2}\right] \tag{7}$$

## 2.1 Flow evaporation model

Given the complexity of the phase-change flow physics, its modeling represents a critical element in the numerical simulation of the plate heat exchanger with flow evaporation: it needs to be an appropriate phase-change model which will sufficiently capture the main flow phenomena, and yet be sufficiently fast and flexible in order to be suitable for an industrial application analyzed in this work. Furthermore, for the flow type under consideration here, there is the stability criterion (expressed through the Courant number) which correlates allowed spatial and temporal discretization characteristics (i.e. the time step and the mesh size). In the case of plate heat exchangers, however, the mesh quality is typically rather low. As a consequence, not only is the quality of the results reduced, but also the numerical simulation can get unstable or prohibitively long.

Aiming at the balance between the accuracy and applicability, a phenomenological approach for the phase-change model has been pursued, where the phase-change process is strictly controlled by the temperature. Namely, the phase kinetic is implemented by prescribing the phase fraction as the function of temperature: the pseudo-fluid is in its liquid phase if the local temperature is below the saturation temperature, and inversely it is in its vapor phase when the local temperature is above the saturation temperature. Corollary to this assumption is that the robustness of the model is significantly increased, while the main phase-change features are accounted for: the temperature is its dominant characteristic, and the saturation temperature is dynamically determined from the property table based on the local flow pressure. As the model parameter, the linear temperature range in which the change between the phases is taking place, can be extracted from the Clausius-Clapeyron relation [5].

### 2.2 Fluid properties

As the focus in this work is on the flows of refrigerants, in order to capture the totality of the fluid flow phenomena within a heat exchanger with flow evaporation, it is necessary to account for a realistic variation of the relevant fluid properties: density, viscosity, heat conductivity and specific heat capacity. To that purpose, the volumetric fields for fluid properties have been implemented into the simulation solver, and they are being fed from the external property tables by interpolating respective values based on the local flow conditions (pressure and temperature). These property tables are generated using simulation tools for fluid equations of state and transport properties, and in the present case the open-source software *CoolProp* has been used.

### 2.3 Boundary conditions

The inlet is defined by the fixed values (Dirichlet boundary condition type) for the velocity, temperature (enthalpy) and turbulent quantities, and the zero gradient (Neumann boundary condition type) for the pressure. The imposed velocity is obtained from the mass flux specified by the operating conditions, together with the related temperature, whereas the turbulent quantities at the inlet are obtained from the required turbulent intensity. The pressure at the outlet is defined by the fixed value specified by the operating conditions, and for the velocity, temperature and the turbulent quantities the zero gradient is specified.

The heat transfer through the plate walls is defined by the Robin type boundary condition: there the temperature of the refrigerant in the near-wall region within the plate channel is dependent on the heat transfer coefficient and the temperature of the secondary fluid at the opposite side of the plate wall. In this way the convective heat transfer from the secondary side (water) is accounted for, together with the 1D assumption for the conduction through the wall of the plate heat exchanger.

## 3 RESULTS

In order to assess the ability of the proposed simulation approach in predicting the thermal behavior and the fluid flow performance of a plate heat exchanger, the comparison between the numerical investigations and the experimental results has been performed. Selected for the comparison is the heat exchanger geometry of Huang et al. [4], with the chevron angle of $60^{o}$. For the quantification of the fluid flow and heat transfer characteristics, a series of numerical simulations has been performed for two refrigerants (R507a and R134a) at two operating conditions at approximately the same saturation temperature levels.

Summarized in the Table 1 are the boundary conditions applied for four different heat exchanger simulations. From the inlet mass flux, the imposed inlet velocity of the refrigerant has been recalculated for the high and low flow rates respectively. The imposed pressure at the outlet was selected such that for both refrigerants the saturation temperatures are approximately the same. On the thermal side of the secondary fluid, the inlet and outlet temperatures of the water are prescribed, and together with a specified mass flow rate the cooling rate is defined. This, in turn, gives the outlet temperature of the refrigerant, for the specified refrigerant inlet temperature.

| | | R507a high flow rate | R507a low flow rate | R134a high flow rate | R134a low flow rate |
|---|---|---|---|---|---|
| Refrigerant | $V_{in}$ $[m/s]$ | 0.03 | 0.02 | 0.03 | 0.02 |
| | $P_{in}$ $[kPa]$ | 756.3 | 749.1 | 324.3 | 315.6 |
| | $P_{out}$ $[kPa]$ | 742.4 | 742.8 | 306.7 | 307.2 |
| Water | $T_{in}$ $[^oC]$ | 16.3 | 16.3 | 15.2 | 14.5 |
| | $T_{out}$ $[^oC]$ | 12 | 12.6 | 11.2 | 11.5 |
| | $Q_{cool}$ $[kW]$ | 14.4 | 12.2 | 13.3 | 10.1 |

Table 1: operating conditions imposed as the boundary conditions.

The basic flow pattern can be seen in the Figure 2, which shows the velocity magnitude distribution of the refrigerant within the plate heat exchanger channels. The main conclusion from the obtained flow pattern is that good uniformity of the refrigerant flow has been achieved, which is an important ingredient for a good thermal behavior of the heat exchanger. The flow uniformity is the consequence of very complex surface structure, which distributes evenly the flow already in the entering section. However, such hydrodynamic behavior comes with the price, expressed in the overall pressure drop along the plate heat exchanger. The temperature distribution follows the main flow pattern: in the narrow parts of the plate channels (where two plates are interlacing) the velocity is increasing, carrying away the heat more effectively and hence locally reducing the refrigerant temperature. On the other hand, in larger channel pockets the refrigerant velocity is decreasing, and its temperature is increasing.

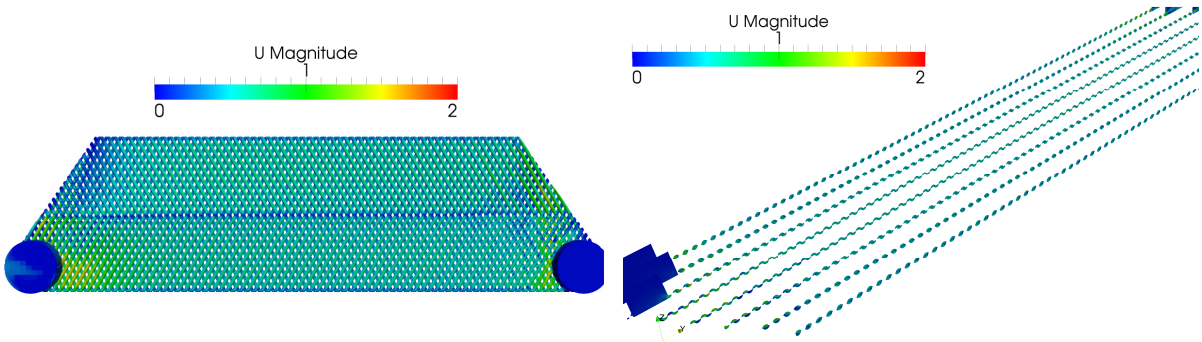



Figure 2: snapshot of the velocity distribution within the plate heat exchanger.

It is stipulated by the adopted phase-change model, that the phase distribution is dictated by the temperature distribution: the pockets of lower temperature evaporate the last, during the flow time. The overall characterization of the refrigerant flow evaporation, however, can be expressed through the domain integrated values of the phases. This has been shown in Figure 3, where the temporal evolution of the volume averaged liquid phase within the domain

(green line) indicates the characteristic time of analyzed plate heat exchanger needed for reaching its steady-state operation. The same conclusion can be drawn from the surface averaged temperature at the heat exchanger's outflow (red line).

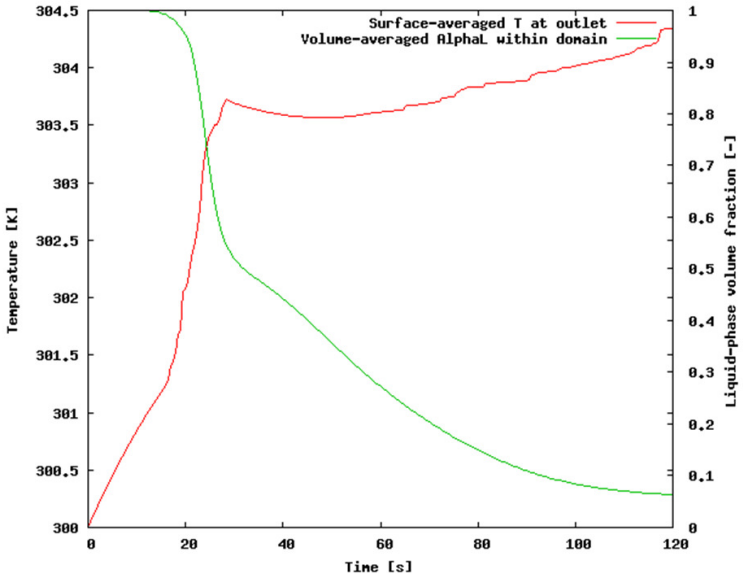



Figure 3: volume-averaged liquid phase, surface-averaged temperature.

From the data given in the Table 2 one can see that the prediction of the thermal characteristics is somewhat more difficult than the hydrodynamic characteristics. Namely, looking at the pressure drop, there is approximately 21% difference between the experimental and numerical results (in the average), whereas the difference in the heat transfer coefficient is approximately 27 % (in the average). Nevertheless, even for thus obtained results a satisfactory agreement can be considered, because in all investigated cases there is a correct trend when changing the operating conditions: this systematic discrepancy can be reduced by certain fine-tuning of the model, which is subject of the future work.

| | | R507a high flow rate | R507a low flow rate | R134a high flow rate | R134a low flow rate |
|---|---|---|---|---|---|
| Exp. | $dP$ $[kPa]$ | 10.3 | 8.8 | 12.5 | 9.4 |
| | $U$ $[kW/m^2K]$ | 1.46 | 1.64 | 1.41 | 1.23 |
| Sim. | $dP$ $[kPa]$ | 13.9 | 6.3 | 17.6 | 8.4 |
| | $U$ $[kW/m^2K]$ | 1.8 | 1.9 | 2.1 | 1.8 |

Table 2: experimental and simulation results for the pressure drop and the heat transfer coefficients, obtained for the analyzed refrigerants and operating conditions.

## 4 CONCLUSIONS

The aim of this paper was to assess the applicability of the computational fluid dynamics, for the analysis of thermal behavior of a plate heat exchanger with refrigerant flow evaporation. Given the complexity of the occurring phase-change phenomena on the one hand, and on

the other very demanding geometry which is considered, the focus here is on the robustness of the physical modeling and the stability of the numerical simulation.

Starting from the pseudo-fluid description of two-phase flows, the flow evaporation has been treated here in a single-phase flow manner. The transport properties of the pseudo-fluid have been defined as the weight mixture of the individual refrigerant properties (vapor and liquid), which are computed using *CoolProp* package. A standard transient *OpenFoam* solver has been extended to include the phase-change process, for which a robust physical description (based on a pre-defined function of the local temperature) has been deployed.

In order to validate the presented numerical approach, the comparison with the literature data has been performed. To that purpose, the geometry of the plate heat exchanger has been realistically represented, and the numerical analysis has been performed for the specified operating conditions. The simulation series has been performed for two different refrigerant and two operating conditions, and the results obtained from the numerical analysis compare reasonably well with the experimentally obtained values.

## ACKNOWLEDGEMENTS

The financial support for the work presented in this paper has been provided by FFG – Österreichische Forschungsförderungsgesellschaft mbH (project "*HighButane*", project number 843935).

## REFERENCES

[1] I.H. Bell, J. Wronski, S. Quoilin, V. Lemort (2014). Pure and pseudo-pure fluid thermophysical property evaluation and the open-source thermophysical property library CoolProp. Industrial Engineering Chemistry Research 53:2498–2508.

[2] M. Colarossi, N. Trask, D.P. Schmidt, M.J. Bergander (2012). Multidimensional modeling of condensing two-phase ejector flow. International Journal of Refrigeration 35:290-299.

[3] H.G. Weller, G. Tabora, H. Jasak, C. Fureby (1998). A tensorial approach to computational continuum mechanics using object-oriented techniques. Computers in Physics 12:620-631.

[4] J. Huang, T.J. Sheer, M. Bailey-McEwan (2012). Heat transfer and pressure drop in plate heat exchanger refrigerant evaporators. International Journal of Refrigeration 35:325-335.

[5] C. Kleinstreuer (2003). Two-Phase Flow: Theory and Applications. Taylor & Francis, New York, USA.

[6] W. Minkowycz (2009). Advances in numerical heat transfer. CRC Press, New York, USA.

[7] Jones, W. P. and Launder, B. E. (1972). The prediction of laminarization with a two-equation model of turbulence. Int. J. Heat Mass Transfer 15:301-314

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# RESEARCH OF UNSTEADY FLOW REGIMES IN CHANNEL OF HYPERSONIC INLET

**Natalia N Fedorova**[1, 2], **and Marat A. Goldfeld**[1]

[1] hristianovich Institute of Theoretical and Applied Mechanics SB RAS
4/1 Institutskaya, Novosibirsk, Russia,

[2]Novosibirsk State University of Architecture and Civil Engineering (Sibstrin)
113 Leningradskaya st., Novosibirsk, Russia
{nfed, gold}@itam.nsc.ru.

**Keywords:** Supersonic Flow, Channel, Numerical Simulation, Experimental Investigation, Wind Tunnel, Sock Wave. Boundary Layer Separation

**Abstract.** *In the paper, the numerical study of the flows in channels of variable cross section has been performed under the conditions of experiments carried out in blow down wind tunnel. The main goals of the present investigations were as follows: to get a new data on the flows in inlet channel at M=2-8 under conditions which close to start-unstart of inlet; to investigate the influence of the channel geometry, Mach and Reynolds number on the flow structure and inlet characteristics; to study the influence boundary layers on the channel flow structure and the inlet performances. The internal part of the channel has complicated configuration can be regulated over the length. The experimental investigations of the model have been carried out in the blow-down wind tunnel T-313 at Mach numbers from 2 to 6 and in the hot-shot wind tunnel IT-302M at Mach numbers from 5 to 8 in Reynolds number range from 8 to 56*10$^6$1/m. The comparison and joint analysis has been performed for the calculated and experimental data. The gas-dynamic patterns have been constructed for the complicated flows with multiplex interactions of the shock waves with the boundary layers developing on the compression surfaces. The parametric computations performed within the wide range of flow and geometric parameters allowed to carry out the experiments and provide a basis for explanation of flow features and the choice of optimum configurations.*

## 1 INTRODUCTION

Hypersonic airbreathing propulsion offers great potential for reliable and economical access to space as well as atmospheric flight. In particular, scramjets (supersonic combustion ramjets) are a promising technology that can enable efficient and flexible transport systems by removing the need to carry oxidizers and other propulsion limitations of conventional rocket engines. Supersonic combustion via the use of scramjet technology was successfully demonstrated for the first time worldwide by the HyShot II Program 2002 [1], followed by successful flights of the Hyper-X vehicles in 2004 at Mach number 6.8 and Mach number 9.6 [2].

One of basic elements of the scramjet is air inlet. Its influence on effectiveness of propulsion is dominant, especially at application of integral construction configurations. Feature of air inlet of the scramjet consists in fact that it should ensure operation of the engine in is unprecedented a wide range of flight conditions at high efficiency of all elements of the engine for reception of the maximum thrust. From this point of view, high level of mass flow rate and minimum total pressure losses must be ensured in a wide range of a flight velocities and altitudes [3, 4]. A lot of key issues have to be addressed for developing the dual mode ramjet technology. One of them consists in designing an inlet able to operate in a very large Mach number range, with enough pressure recovery while providing the necessary air mass-flow to the combustion chamber. The use of a sophisticated variable geometry simplifies the aerodynamic design of the inlet. But, in the same time, the mechanical feasibility of the inlet is more and more difficult task for implementation because of the very severe environment the inlet will encounter during the flight. At the contrary, a fixed geometry inlet will be much more feasible from the mechanical point of view but it will lead to the very difficult compromises at the choice of the design Mach number. At the too large contraction ratio, a very large spillage will reduce the installed thrust at low Mach number. If the contraction ratio will be too restricted, the general performances will be dramatically reduced at high Mach number [5].

Several studies have been carried out on optimizing inlet performance using internal inlet components. Amongst other parameters, inlet performance also depends on upstream flow. Different research groups have investigated the effect of major component properties on inlet characteristics, e.g., the modification of the inlet internal geometry [6] and configurations of side walls [7]. Variations of entrance configuration and inlet position reveal different effects on performances [8] and studies at the position of the inlet diffuser were carried out to avoid flow separation [9, 10]. Other studies included unstart criteria and inlet buzzing [11, 12], which must be prevented [13] to ensure the functionality of the ramjet and prevent destruction of the structure of inflowing and the inlet unstart [14, 15]. One technique is to control the inlet normal shock by adjusting internal parts of the inlet [16, 17]. All examples deal with inlet design and a given unchangeable main engine design [18, 19]. On the basis of still open issues from these studies, the influence of the upstream flow on inlet performance was identified as one of the key issues.

Reliable in-flight starting of the inlet is of critical importance for the successful operation of scramjet engines, particularly integral configurations with high-contraction inlets [20]. The present research is undertaken to examine the capabilities of various inlet starting methods based on two principles: boundary layer effects and variable geometries. If the overall Mach number range is limited (Mach 4-Mach 10), a completely fixed geometry can be defined for the scramjet inlet with relatively good possible performance in the whole flight envelope. At the contrary, an extension of the Mach number range needs to provide some variation of the geometry of the inlet channel in order to obtain acceptable performances. Indeed for higher Mach numbers, it will be necessary to use a high contraction ratio for the inlet in order to limit the supersonic Mach number at the entrance of the combustion chamber and to reduce diverg-

ing combustion chamber, as Mach number will increase. At the contrary, in order to extend the flight range to lower Mach numbers, the inlet contraction ratio must be reduced (with or without reduction of the capture area) to avoid a too large cowl spillage and the corresponding additive drag section of the combustion chamber must be open.

Therefore it is important to ensure the necessary mass flow rate and pressure recovery by means of adjustment of geometry and control of boundary layer bleed. Investigations of performances of adjustment inlet with or without boundary layer bleed were presented in works [21, 22].

Nevertheless these researches were limited to a narrow range of speeds of flight (Mach numbers) and did not assume an adjustment of cross-section of inlet channel depending on flight conditions. The computer design tools are widely used in developing the optimum geometry of the engines and their elements (inlet, combustion chamber, and nozzle) [23]. At the same time numerical simulation widely employs for determination of aerodynamic scramjet characteristics (thrust and drag). The flow simulation of high-speed inlets is one of important directions in these researches [24-26].

In the paper, the numerical study of the flows in channels of variable cross section has been performed under the conditions of experiments in wind tunnels. The comparison and joint analysis has been performed for the calculated and experimental data. The gas-dynamic analysis has been carried out for the complicated flows with multiple interactions of the shock waves with the boundary layers developing on the inlet surfaces.

The main goals of the present investigations were as follows:

- to get a new data on the flows in channel at M=2-8;
- to investigate the influence of the channel geometry, Mach and Reynolds number on the flow structure and inlet characteristics
- to verify the algorithm and code on the channel flow experimental data;
- to study the influence of the boundary layer before channel entrance on the channel flow structure and the inlet characteristics.

## 2 MODEL AND FACILITIES

The experimental model presents a channel with flat walls and 3D central body [4, 7]. The quasi-two-dimension configuration of the central body (nose part) allows one to simulate a boundary layer which is developed on the aircraft nose part and guarantees preliminary flow deceleration in two shock waves. The internal part of the channel has a complicated configuration with the adjusted area over the channel length. The variation of inlet geometry was performed by means of motion of the cowl. Several model configurations with different cowl position at different relative throat area have been considered (Fig. 1) for the corresponding Mach numbers to provide the inlet “start” at different test conditions.

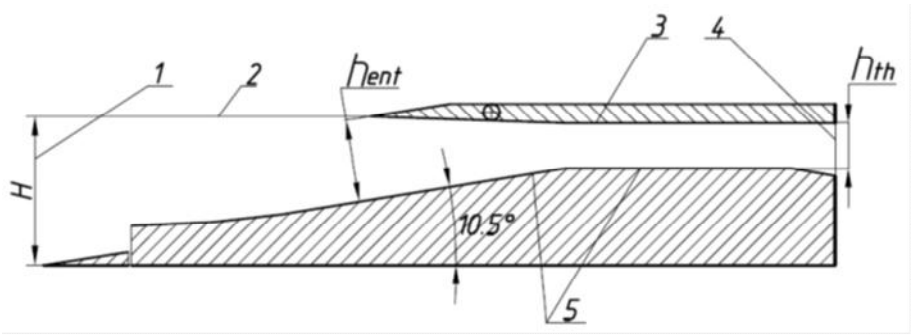



Figure 1: Scheme of model and calculation domain

Change of the channel configuration was accompanied by change of value of internal compression, which decreased at reduction of Mach number to provide inlet start in all range of Mach numbers. Change of internal contraction (relative height of a throat) is presented in Table 1 together with total contraction of captured air stream.

| M | 2 | 3 | 4 | 5 | 5.5 | 6 | 7 | 7.5 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| $H/h_{th}$ | 3.0 | 3.36 | 4.0 | 4.5 | 4.8 | 5.3 | 5.66 | 6.21 | 7.06 |
| $h_{ent}/h_{th}$ | 0.96 | 1.14 | 1.44 | 1.68 | 1.82 | 2.06 | 2.23 | 2.51 | 2.91 |

Table 1: Channel configurations

The bottom and top walls of the model were equipped with static pressure taps to measure the distribution of static pressure along the longitudinal axis. Additionally, in three cross-sections the installation of Pitot pressure probes has been provided to determine the local Mach numbers and total pressure recovery in the channel. The side walls of model and wind tunnels facility was equipped with transparent windows to provide Schlieren visualization of the internal flow structure. Glass windows are placed along the inlet throat of the model to visualize the flow to determine its structure (separation and attachment of the boundary layer, location and form of shock waves). During test, shadow, oil and sparkle visualizations were performed to analyze the flow structure at the entrance and in the duct of the inlet.

The experimental investigations of the model have been carried out in the blow-down wind tunnel T-313 at Mach numbers from 2 to 6 and in the hot-shot wind tunnel IT-302 at Mach numbers from 5 to 8. The blow-down wind tunnel T313 has been equipped profiled nozzles for the Mach number range from 2 to 6 and has test duration up to 600 s. Working section of wind-tunnel has a rectangular cross-section and located in a pressure chamber. The hot-shot wind tunnel IT 302M has been equipped profiled nozzles with exit diameter of 400 mm for the Mach number range from 5 to 8. Test duration can change from 100 to 200 ms depending on Mach number. Typical flow characteristics for both wind tunnels are presented in Table 2.

| Blow-down wind-tunnel T-313 | | | | | | Hot-shot wind-tunnel IT-302M (time range 10 - 70 ms) | | |
|---|---|---|---|---|---|---|---|---|
| M | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 |
| $P_t$, bar | 2.1 | 4.2 | 10.5 | 8.0 | 8.8 | 52÷31 | 330÷56 | 340÷124 |
| $T_t$, K | 288 | 280 | 2760 | 374 | 540 | 1780÷1500 | 2200÷1340 | 2000÷1600 |
| $Re*10^6/m$ | 24.8 | 36.2 | 54.4 | 16.9 | 9.4 | 16.8 ÷ 10.4 | 22.9÷6.8 | 14.9÷7.6 |

Table 2: Condition of the experiments in T-313 and in IT-302.

Peculiarity of IT-302M wind tunnel is decrease of flow parameter during operation time. That is why flow parameters for this tunnel are shown for typical duration of wind tunnel operation (from 10th ms to 70th ms) when regime of wind tunnel remains quasi-stationary.

## 3 MATHEMATICAL MODEL AND CALCULATION METHOD

The calculations have been performed on the basis of the full transient Favre-averaged Navier-Stokes equations and Wilcox two-equation turbulence model [27]. The original numerical algorithm [28] used earlier for the 2-D turbulent supersonic flow computations and showed a good potential to predict the properties of these flows. The same methodical ap-

proach was extended to a new class of internal supersonic flows. The temporal approximation presents a four-step finite-difference scheme of splitting according spatial variables. At each fractional step the finite-difference scheme was realized by scalar sweeps. The TVD-scheme of Flux Vector Splitting by van Leer of the third order of accuracy has been used for the approximation of convective terms. The viscous terms have been approximated with the central finite-difference relations of second order of accuracy.

The computation region was bounded by the bottom wall (central body) 5 from below and by upper wall 3 or free surface 2 from above. On the left and on the right the computational domain was bounded with input and output sections 1 and 4, respectively. In the input section 1, the free flow conditions or the profiles of all gas-dynamic and turbulent parameters obtained from the computation of the turbulent boundary layer on the plate have been assigned. The coincidence with experimental boundary layer integral parameters and skin friction coefficient has been provided in computations if such data were obtained in the experiments. The no-slip conditions for velocity have been specified on boundaries 4 and 5. Adiabatic condition for temperature was used in the computation for M=2 ÷6 corresponding blow-down wind tunnel and constant temperature ("cold wall") was specified for M=6 ÷8 computations in conditions of hot-shot wind tunnel. On the top boundary 2, so-called "simple wave" conditions were assigned in order to provide all the disturbances to way out from the computation region. The "soft" boundary conditions have been set on boundary 5. Special study was performed for determination of the temperature factor influence on the flow structure.

## 4 INLET FLOW SIMULATION

The computations were performed for the inlet configurations shown in Figure 1 under the conditions of experiments described in Table 1.

Fig. 2 presents the static pressure distributions along the central body and the cowl obtained in computations for Mach 2 to 6 carried out under the conditions in blow down wind tunnel. Data was made dimensionless by the value of pressure in the entrance section of the calculation region. The data obtained indicate a big lengthwise pressure non-uniformity and high level of pressure, which increase together with Mach number. At Mach numbers $M \geq 4$, a saw-toothed structure appears in the pressure distribution related to the alternation of shock waves and rarefaction waves inside the channel

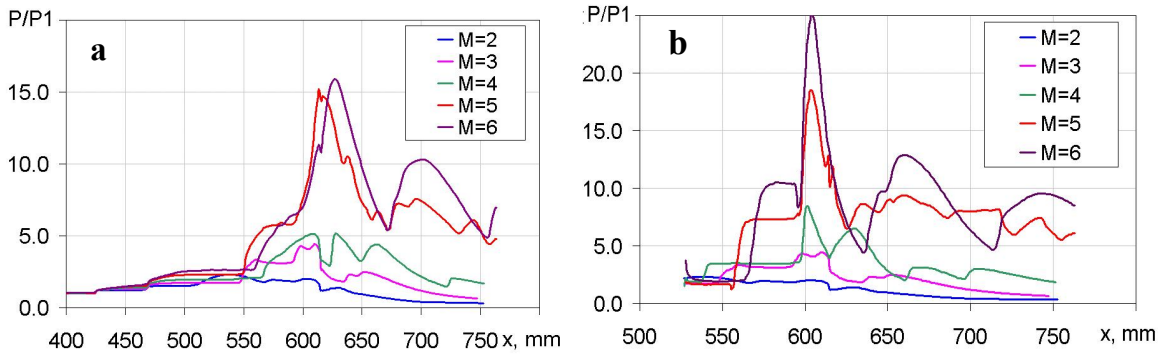



Figure 2: Computed pressure along central body (a) and cowl (b)

The computed distributions of relative total pressure $P0/P0_\infty$ are presented in Figure 3. Here pressure was averaged over the channel width along the computation grid lines. These data allow one to determine the total pressure losses in any cross section which is a good complement to the direct measurements in some assigned cross-sections. The joint analysis of

the calculated and experimental results has permitted one to conclude that investigated model configuration are satisfactory with respect to the total pressure recovery coefficient within the whole given range of Mach numbers.

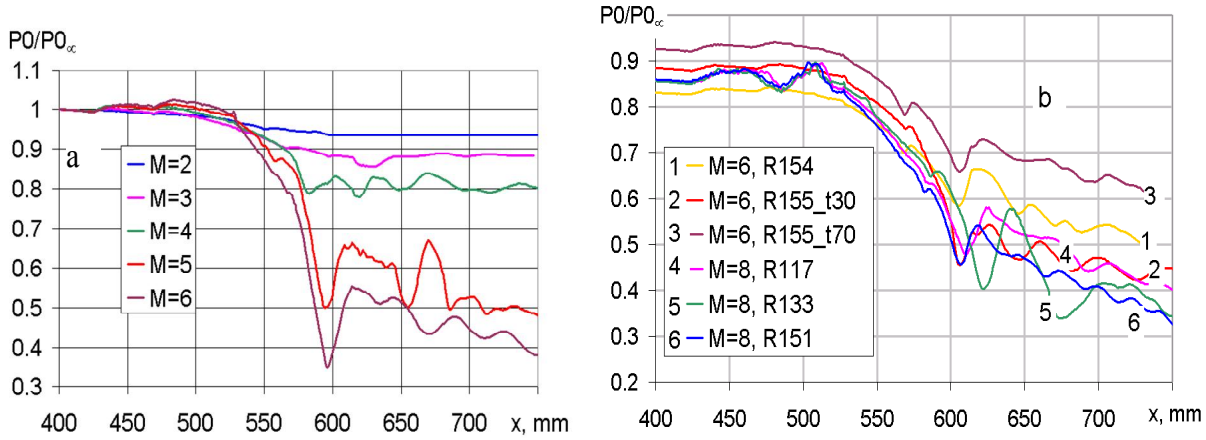



Figure 3: Averaged total pressure under blow down (a) and hot-shot (b) wind tunnel conditions

Comparison of results of calculation of static pressure distribution under test conditions in blow down wind tunnel is presented in Figure 4. This data shows good agreement of computed and experimental data on major part of the channel length of model. Appreciable quantitative difference is observed in the end of computational domain where experimental values can exceed results of calculation. Probably, this is effect of influence of subsonic diffuser, in which the throttling device should be installed. This effect becomes more significant at the rise of Mach number.

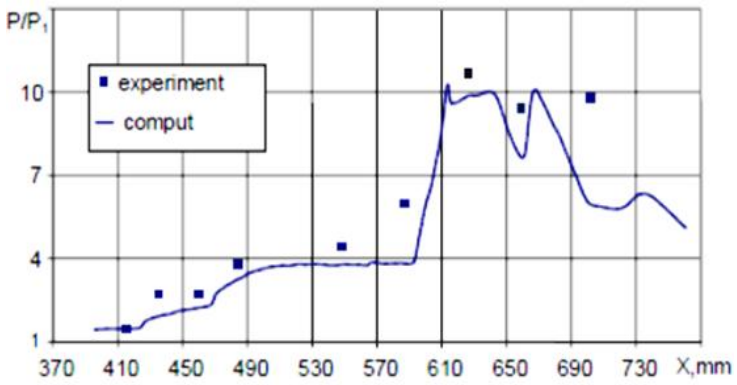



Figure 4: Central body pressure distribution at Mach number 6in blow down wind tunnel

The increment of shock wave intensity with Mach number rise results in the formation of an extensive separation region on the channel walls behind channel entrance. This region covers the channel section and causes the significant pressure rise. Determination of skin friction distribution on the walls allows detecting these regions and, correspondingly, area location of maximum pressure and heat flux. Results of calculation of skin friction distribution on the walls of a central body at Mach numbers 2-6 is presented in Figure 5. The presence of the boundary layer separation region arise at Mach number $M \geq 4$ when the skin friction coefficient become less than zero. One can see that separation zone increases at Mach number rise. Size of separation zone can reach 5 height of throat as it occurred at Mach num-

ber M=6. These results were confirmed by means of Schlieren visualization of flow in the model throat.

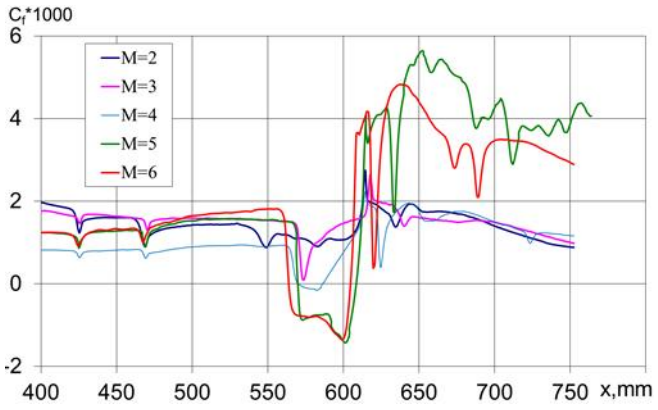



Figure 5: Computed skin friction along the central body for various Mach numbers

Three geometrical configurations were investigated at Mach number M=8 (Table 1) for three different relative internal contraction $h_{ent}/h_{th}$=3.11, 2.91 and 2.32 that corresponded experimental conditions of the runs 133, 151 and 117. Here $h_{ent}$ and $h_{th}$ is height of entrance in the channel and throat height, correspondingly. The static pressure distributions along the central body (a) and the cowl (b) are presented in Fig. 6. Comparison data obtained indicate qualitative agreement of the experimental and computational data.

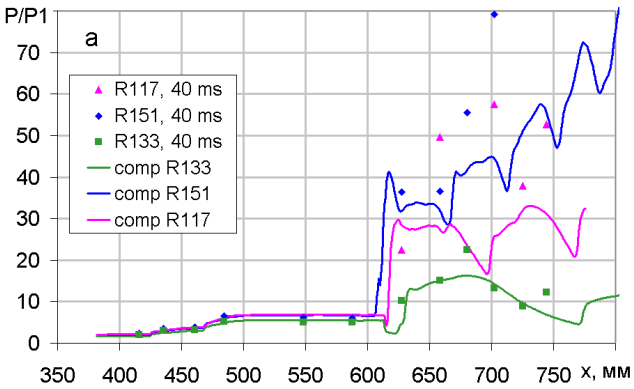



Figure 6, a: Static pressure distribution along central body at M=8

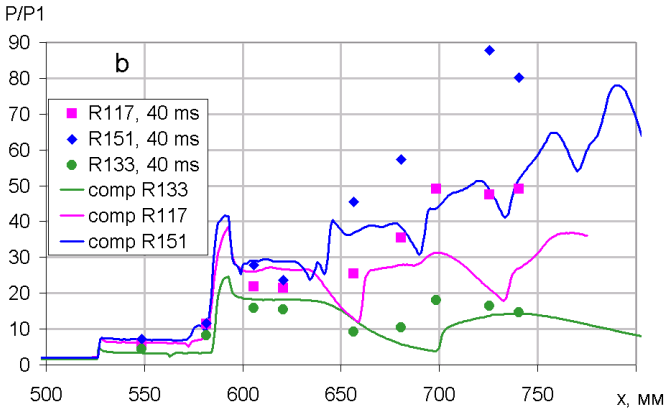



Figure 6, b: Static pressure distribution along cowl at M=8

The possible reason of disagreement between computed and experimental values may cause the fact that the cowl of the real model has the cavity and subsonic diffuser downstream behind the end of computation domain. Such geometrical configuration is source of formation of the massive flow separation. The separation zone moving upstream, leads to a significant static pressure rise inside the channel and, as a result, to increasing of the losses of total pressure. For computations under condition of hot-shot wind tunnel there is also a problem to define properly the incoming flow parameters, including turbulence level, and wall temperature. Computations were performed for constant conditions for fixed time, whereas they were changing in real flows.

The flow in inlet has very complex structure because of multiple shock waves and expansion fans and their interactions with boundary layer on channel walls, which is difficult-to-analyzed on the basis of only measurements and even visualization. Here mathematical modeling gives the instrumentation that helps one to construct the flow field picture. According to Schlieren visualization data, the M=4 flow structure is consistent with the calculated density contours flows, as it shown in Figure 7. The density contours are presented in the whole calculation region (a) and in the part of channel (c), which corresponds to the experimental flow visualization picture in model throat (b). Comparing Figure 7b to Figure 7c it is quite easy to identify the wave structures and corresponding flow regions on the experimental and calculation schemes. The gas-dynamic patterns of the flows constructed on the basis on a joint analysis of the computation and the visualization and measurements gives more deep understanding of the processes in the duct and helps one to design the efficient devices

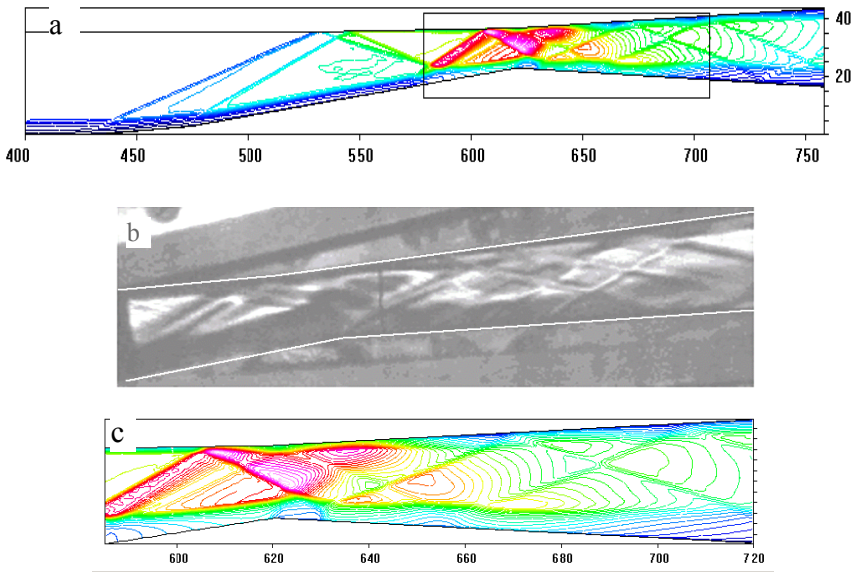



**Figure 7**: The computed density contours (a), (c) and experimental flow visualization (b) at M=4

In computations, the “start -unstart” of the channel was modeled depend on Mach number and corresponding internal contraction. When compression grows for example, in consequence of increase of angle of attack is growing, the separation zone on the central body and cowl becomes larger that covers the channel and brings about a significant pressure rise inside the channel. If cross section is deficient for realization of supersonic inflowing, separation zone is moving upstream along channel and its size is increasing. One can see the separation zone growth and movement of the separation zones and the shock waves system upstream in Figure 8. This process is accompanied by subsequent pressure rise up to moment of break of

inflowing and appearance of normal shock wave before channel entrance. These results conform to experimental data of pressure measurement and flow visualization in blow down wind tunnel.

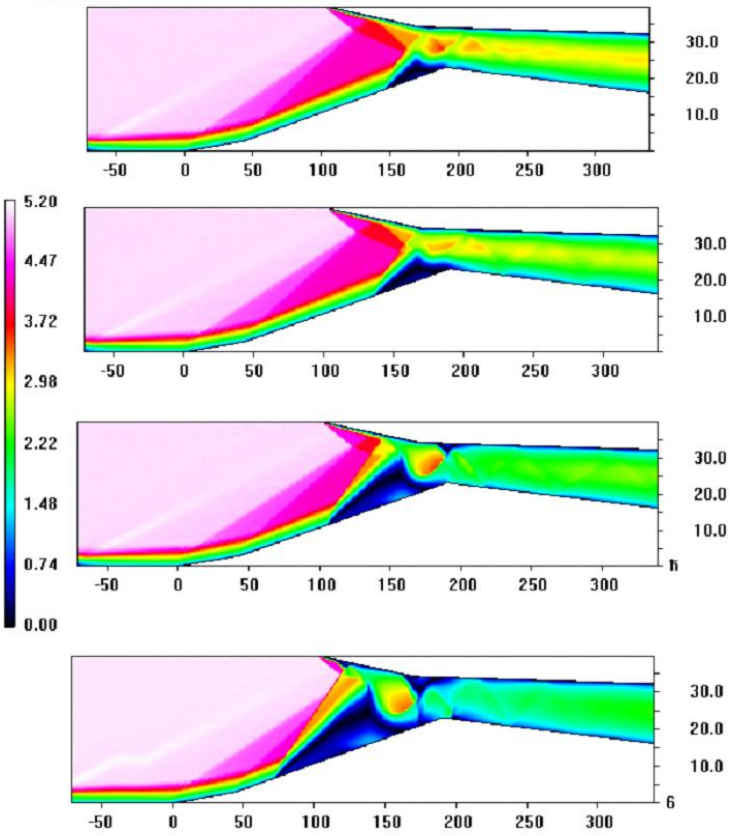



Figure8: Modelling the inlet "unstart" process at M=6

Thus, there is a transition from "start" to "unstart" state of the channel. Presented data show the presence of extensive subsonic areas, considerable decrease of flow velocity in a channel and increase of a boundary layer thickness.

The Figure 9 shows change of skin friction coefficient at process realization "unstart" the channel. It can be seen that the area of negative values of a skin friction increases in process of growth of the sizes of separation area and shifts upstream to a channel entrance. The data obtained allows defining extent of a separation region, which can reach till ten heights of the throat that corresponds approximately to half of length of the channel.

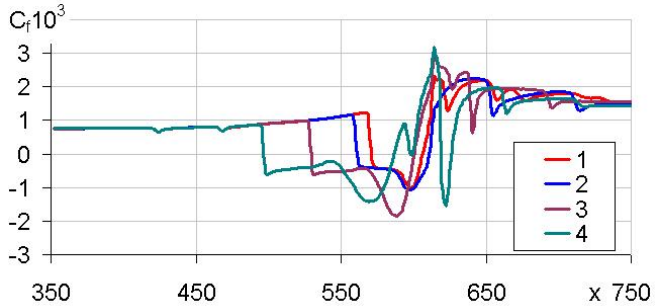



Figure 9: Skin friction coefficient distribution in the channel with separation zone.
Lines 1-4 corresponds pictures in Figure 8 from the upper to lower

## 5 CONCLUSIONS

Thus, the methods of physical experiment and computational modeling have been used in the work to study the properties of complex flows in the inlets with changeable geometrical configuration in a wide range of Mach number. The experimental findings present a basis for the mathematical model and calculation algorithm verification. At the same time, the parametric computations performed within the wide range of flow and geometric parameters help one to carry out the experiment and provide a basis for the choice of optimum configurations and explanation of flow features.

## ACKNOWLEDGEMENTS

This research was partially supported by the Russian Fund for Basic Research, project No. 16-08-00917.

## 6 REFERENCES

[1] R. R. Boyce, S. Gerard, and A. Paull, The HyShot Scramjet Flight Experiment—Flight Data and CFD Calculations Compared, *AIAA Paper* No. 2003-7029, Dec. 2003, 10p.

[2] C. R. McClinton, X-43—Scramjet Power Breaks the Hypersonic Barrier: Dryden Lectureship in Research for 2006, *AIAA Paper* 2006-1317, Jan. 2006, 12p.

[3] J. Seddon, and E. L. Goldsmith, Practical Intake Aerodynamic Design, *AIAA Education Series, AIAA, New York,* 1993, 398p.

[4] F. Falempin, and M. Goldfeld, Design and Experimental Evaluation of a Mach 2 – Mach 8 Inlet, *AIAA Paper No.* 2001-1890, Kyoto, 2001, 8p.

[5] E. T. Curran, W. H. Heiser, and D. T. Pratt, Fluid Phenomena in Scramjet Combustion Systems, *Annual Review of Fluid Mechanics,* **28**, Jan. 1996, pp. 323–360.

[6] D. M. Van Wie, Scramjet Inlets, *Scramjet Propulsion, Progress in Astronautics and Aeronautics, AIAA, Reston*, VA, Vol. **189**, Chap. 7, 2001, pp. 447-512.

[7] F. Falempin, M. Goldfeld, R. Nestoulia, A. Starov, Evaluation of a variable geometry inlet for dual mode ramjet, *AIAA Paper No.* 2002-5232, 2002, 8p.

[8] C.P. Wang, K.Y. Zhang, K.M. Cheng, Pressure Distribution Measurements in Scramjet Isolators under Asymmetric Supersonic Flow, *AIAA Paper No*. 2006-818, 2006. 10p.

[9] N. Bachchan, R. Hillier, Effects of Hypersonic Inlet Flow Non-Uniformities on Stabilising Isolator Shock Systems, *AIAA Paper No.* 2004-4716, Aug. 2004, 10p.

[10] A., Hamed, and J. S. Shang, Survey of Validation Data Base for Shockwave Boundary-Layer Interactions in Supersonic Inlets, *Journal of Propulsion and Power*, **7**, No. 4, 1991, pp. 617–625.

[11] K. Matsuo, Y. Miyazato, H.-D. Kim, Shock Train and Pseudo-Shock Phenomena in Internal Gas Flows, *Progress in Aerospace Sciences,* **35**, No. 1, 1999, pp. 33–100.

[12] C. Cox, C. Lewis, and R. Pap, Prediction of Unstart Phenomena in Hypersonic Aircraft, *AIAA Paper No.* 1995-6018, April 1995. 9p.

[13] P. J., Waltrup, and F. S. Billig, Structure of Shock Waves in Cylindrical Ducts, *AIAA Journal,* **11**, No. 10, 1973, pp. 1404–1408.

[14] B. U. Reinartz, C. D. Herrmann, and J. B. Ballmann, Aerodynamic Performance Analysis of a Hypersonic Inlet Isolator Using Computation and Experiment, *Journal of Propulsion and Power,* **19**, No. 5, 2003, pp. 868–875.

[15] J. L. Wagner, A. Valdivia, K. B. Yuceil, N. T. Clemens, D. S. Dolling, An Experimental Investigation of Supersonic Inlet Unstart, *37th AIAA Fluid Dynamics Conference and Exhibit, 2007 Miami, FL, AIAA Paper No.* 2007-4352, 2007, 11p.

[16] D. M. Van Wie, F. T. Kwok, R. F. Walsh, Starting Characteristics of Supersonic Inlets, *AIAA Paper No.* 96-2914, July, 1996, 9p.

[17] X. Veillard, R. B. Tahir, E. V. Timofeev, S. Mölder, Limiting Contractions for Starting Simple Ramp-Type Scramjet Intakes with Overboard Spillage. Journal of Propulsion and Power, **24**, No. 5, 2008, pp. 1042–1049.

[18] J. Haberle, A. Gulhan, Investigation of Two-Dimensional Scramjet Inlet Flowfield at Mach 7, Journal of Propulsion and Power, **24**, No. 3, 2008, pp. 446-459.

[19] S. Srikant, J.L. Wagner, M.R. Akella, N. Clemens, Unstart Detection in a Simplified-Geometry Hypersonic Inlet–Isolator Flow, Journal of Propulsion and Power, **26**, No. 5, 2010, pp. 1059-1071.

[20] M. T. Nair, N. Kumar, S. K. Saxena, Computational Analysis of Inlet Aerodynamics for a Hypersonic Research Vehicle, *Journal of Propulsion and Power,* **21**, No. 2, 2005, pp.286-291.

[21] D. Schulte, A. Henckels, R. Neubacher, Manipulation of Shock/Boundary-Layer Interactions in Hypersonic Inlets. *Journal of Propulsion and Power*, Vol. 17, No. 3, 2001, pp.411-417.

[22] D. Rozario, and Z. Zouaoui, Computational Fluid Dynamic Analysis of Scramjet Inlet, *5th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, AIAA Paper* No. 2007-30, 2007, 20p.

[23] G.V. Candler, P. K. Subbareddy, J.M. Brock, Advances in Computational Fluid Dynamics Methods for Hypersonic Flows. *Journal of Spacecraft and Rockets*, **52**, No. 1, 2015, pp.17-28.

[24] D. Knight, Automated optimal design of supersonic and subsonic diffusers using CFD, *European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS, Barcelona, September* 2000, 21 pp.

[25] C. Bourdeau, G. Carrier, D. Knight, and K. Rasheed, Three Dimensional Optimization of Supersonic Inlets, *AIAA Paper No.* 99-2108, 1999, 10p.

[26] S. Mölder, E. V. Timofeev, and R. B. Tahir, Flow Starting in High Compression Hypersonic Air Inlets by Mass Spillage, *AIAA Paper No*. 2004-4130, July 2004, 11p.

[27] D.C. Wilcox, *Turbulence Modeling for CFD. La Canada, California: DCW Industries Inc.* 1993. 460 p.

[28] A.V. Borisov, N.N. Fedorova, Numerical simulation of turbulent flows near the forward-facing steps, *Thermophisics and Aeromechanics*. **4**, No.1, 1996, pp. 69-83.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EXPLICIT DARCY'S LAW BOUNDARY CONDITION WITH COMBINED CONTINUUM AND DISCRETE MODEL FOR PRESSURE DRIVEN MEMBRANE APPLICATIONS

**Tomi I. Naukkarinen**[1], **and Teemu Turunen-Saaresti**[1]

[1] School of Energy Systems, Lappeenranta University of Technology (LUT)
Skinnarilankatu 34,53850 Lappeenranta, Finland
e-mail: tomi.naukkarinen, teemu.turunen-saaresti@lut.fi

**Keywords:** CFD-DEM, Darcy's law, Membranes, Dilute solute.

**Abstract.** *In this paper Darcy's law for dilute solutes, that is implemented explicitly to a boundary condition in computational fluid dynamics-discrete element method (CFD-DEM) solver, is presented for pressure driven membrane applications. The explicit treatment of the boundary condition with Darcy's law in segregated pressure based solver induces oscillations. To avoid oscillations the boundary condition is under relaxed by limiting change of the solvent flux between iteration cycles. With combination Darcy's law and CFD-DEM cake formation on a membrane and some of the hybrid processes can be modeled.*

## 1 INTRODUCTION

Pressure driven membrane process is widely used in civil engineering (e.g municipal wastewater and the desalination of sea water or brackish) and industry. The pressure driven processes suffer of fouling like other membrane processes. The fouling is a severe problem which increases maintenance cost, decreases productivity and in the some cases forces to shut down process. Cake formation, one type of the fouling, can be the problem at the membrane where it directly increases flow resistant, or at pretreatment phase, where solids are removed. In the former case, a cake effects indirectly to energy consumption and increases required pumping power.

Water quality regulations and the concern of trace contaminants imposes requirements to a water treatment technology and raises new ways to treat water e.g. hybrid ion exchange (IE) - pressure driven membrane processes as Al Abdulgader et al. [1] pointed out. In many hybrid IE-pressure driven membrane processes ion exchange and filtration is separated to different units while Kabay et al. [2] suggested hybrid process where ion exchange resins were immersed in feed and retentive. According Kabay et al. [2] in submerged system yield and separation efficiency is at high level. The submerged systems could be alternative method to traditional fixed bed columns.

Cake formation on a membrane and submerged hybrid processes impose additional requirements to modeling, particularly for computational fluid dynamics (CFD). Particle-particle collision and fluid-particle interaction forces have to be considered while flow regime changes from dilute to dense [3]. In addition, at small particle ($\sim \mu$m) scales cohesion forces have important role in settling and filtration [4]. Dong et al. [4] modeled successfully settling and filtration with DEM code but they assumed direction of fluid flow and used periodic boundary conditions. To gather information from systems, larger scale modelling e.g. fluidization or pneumatic convey is required [5].

Modeling of pressure driven membranes with CFD are common divided in two approached based on a boundary condition used for a membrane; dissolving and permeating boundary conditions [6]. Both of those approaches assumes boundary layer on membrane at permeate stream can be neglected which means constant pressure and concentration at the permeate stream. In addition, a membrane is modeled as boundary condition. In dissolving method there is no flux through a membrane and solute's boundary condition type is either is Dirichlet or Neumann. Use of the dissolving boundary condition can be justified by small modeling area where Reynolds number does not change. In the permeating boundary condition the solvent flux is defined by a equation similar to Darcy's law which is depend on the transmembrane pressure (TMP) and osmotic pressure [7]. Transmembrane pressure is pressure difference between feed and permeate streams. The pressure difference is the potential and the driving force for separation in the pressure driven membrane separation. The driving pressure is decreased by osmotic pressure. Osmotic pressure is pressure which is needed to prevent solvent flux across a semipermeate membrane from dilute solution to concentrate solution. Osmotic pressure rises from osmosis phenomenon. Osmotic pressure is proportional to solute concentration differences and solute type.

Drag, which is depended on the relative velocity difference and which is the one of the major

forces in the particle-fluid flows, has large impact on the pressure loss in the cake or resin bead layer. The drag force transfer momentum from the fluid to the particles and vise versa depend on the relative velocity difference. However, the pressure loss decreases TMP which decreases relative velocity difference and this interconnection effect is only possible to model with permeating boundary condition. Results of Dong et al. [4] support this assumption.

In this paper, Lagrangian-Euler approach, which is implemented to CFD-DEM program [8], is used with the explicit Darcy alike equation [6] to model particle fluid flow with the permeating membrane. The osmotic pressure is not included and only hydrodynamic model is taken account at this point. In addition, the segregated pressure-velocity coupling and PISO-loop [9] is used.

## 2 Numerical Methods and Case Setup

Permeating flux through the membrane is sensitive to TMP, which is solution of the pressure-velocity field, and at the same time the flux effects to velocity and pressure fields. This dependence with updating boundary values at the end of PISO-loop raise problems to achieve convergent solution.

To solve the convergence problem relaxation is added to the membrane boundary condition:

$$\begin{cases} \mathbf{J_{slv}} = \alpha \mathbf{J^o_{slv}} + (1-\alpha) L_p \Delta p_{tm} \hat{\mathbf{n}}, & \text{if } \|\mathbf{J_{slv}} - \mathbf{J^o_{slv}}\| \leq \beta \\ \mathbf{J_{slv}} = \mathbf{J^o_{slv}} + \beta \dfrac{\Delta p_{tm}}{\|\Delta p_{tm}\|} \hat{\mathbf{n}}, & \text{if } \|\mathbf{J_{slv}} - \mathbf{J^o_{slv}}\| > \beta \end{cases} \tag{1}$$

where $\mathbf{J_{slv}}$ is solvent flux, $L_p$ is hydraulic permeable coefficient, $\Delta p_{tm}$ is transmembrane pressure and superscript 0 is the mark of previous iteration. $\hat{\mathbf{n}}$ is surface normal. The equation 1 is sensitive to three parameters: $\alpha$, $\beta$ and $L_p$. Hydraulic permeable coefficient $L_p$ is the most important factor since it defines membrane sensitivity of TMP. $\alpha$ has relaxing effect to equation 1 and effects directly convergence rate. $\beta$ prevents diverging changes of flux between iteration rounds.

Within this paper pseudo-2D test computational domain is used: the flow is modeled in 2D and the particles are calculated with the 3D slice. The slice is used similar way in Ketterhagen et. al [10] to reduce computational demand and to take into account the three dimensional effects of particles. The computational domain is shown in figure 1. The flow with particles is introduced from the left, between the backward- and forward-facing step lies the membrane, where solvent permeates, and some of the particles are settled on the membrane. Rejection with some particles flows out of the domain at the right. A list of dimensions of test domain is shown in table 1. All test cases start with fully developed flow field without any particles. Solid mass fraction is fixed to 0.25 at the inlet until 300k of particles are fed in the flow domain. Constant and uniform velocity is used (0.5 m/s) at the inlet. Outlet pressure is set to be 109000 $Pa$. Constant viscosity is also assumed. Twenty pressure-velocity corrections per a time step are used to achieve convergent solution as a compromise between computational demand and accuracy.

In the test cases, boundary condition's ability to respond changing flow circumstances and the sensitivity of parameters $\alpha$ and $\beta$ are tested. In the first case the variety of $\alpha$ values and the fixed $\beta$ value are used. Table 2 includes the variables values of case I. In the case II limiter

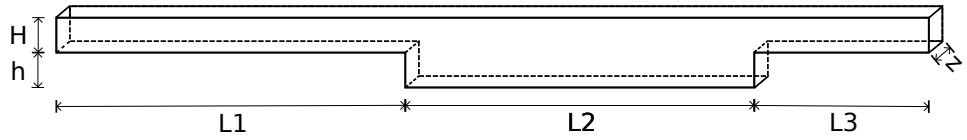



Figure 1: Outlay of computational domain

| Property | value | Unit |
|---|---|---|
| Geometry | | |
| H | 0.01 | m |
| h | 0.01 | m |
| L1 | 0.1 | m |
| L2 | 0.1 | m |
| L3 | 0.05 | m |
| z | 0.001 | m |

Table 1: Dimensions of the test cases

parameter $\beta$ is tested by increasing TMP by 50 $kPa$ during one fluid time step (0.000025 $s$). Three $\alpha$ values are chosen with six different $\beta$ values as can be seen from table 3. In the third case the sensitivity of relaxation factor $\alpha$ to the hydraulic permeable coefficient $L_p$, which is multiplied by $\gamma$ compared to the previous cases, is tested. Table 3 points out $\gamma$ values.

## 3 Results and Discussion

Averaged modeled solvent flux through the membrane is compared to the averaged result of equation 2. Basically the equation 2 is the same as the equation 1 without relaxation and the results of equation 2 are calculated with the modeled pressure on the membrane. Comparison is done by using absolute and relative errors. In the case I the averaged fluxes are calculated at time 0.3 $s$ which is about 0.05 seconds after particles hit membrane surface and induces pressure fluctuations. The flow field and the particle distribution are illustrated in figure 2 and the error results of the cases I are presented in table 4. $\alpha$ value 0.99 produces good results in the case I since the relative and absolute errors are minimal. $\alpha$ value 0.98 or lower produces over 50% relative errors, in addition absolute error that indicates that the flow direction is inward at the membrane. The case I results highlight sensitivity of the explicit boundary condition since only $\alpha$ value 0.99 produce reasonable results. These results indicates that strong relaxation is needed if pressure over membrane changes significantly, e.g filtration changes to backflush.

$$J_{slv,ave} = L_p \Delta p_{tm,m} \tag{2}$$

| Case I | $\alpha$[-] | $\beta$[-] | $L_p[m/(sPa)]$ |
|---|---|---|---|
| A | 0.99 | 0.005 | 2.77e-9 |
| B | 0.98 | 0.005 | 2.77e-9 |
| C | 0.97 | 0.005 | 2.77e-9 |

Table 2: Parameters of relaxation factors in case I

| | Case II | | | | | Case III | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | | | | | |
| | $\alpha$[-] | $\alpha$[-] | $\alpha$[-] | $\beta$[-] | $L_p[m/(sPa)]$ | $\alpha$[-] | $\beta$[-] | $\gamma$[-] |
| A | 0.99 | 0.9875 | 0.985 | $10^{-6}$ | 2.77e-9 | 0.99 | 0.005 | 0.8 |
| B | 0.99 | 0.9875 | 0.985 | $10^{-5}$ | 2.77e-9 | 0.99 | 0.005 | 0.9 |
| C | 0.99 | 0.9875 | 0.985 | $10^{-4}$ | 2.77e-9 | 0.99 | 0.005 | 1.0 |
| D | 0.99 | 0.9875 | 0.985 | $10^{-3}$ | 2.77e-9 | 0.99 | 0.005 | 1.1 |
| E | 0.99 | 0.9875 | 0.985 | $10^{-2}$ | 2.77e-9 | 0.99 | 0.005 | 1.2 |
| F | 0.99 | 0.9875 | 0.985 | $10^{-1}$ | 2.77e-9 | 0.99 | 0.005 | 1.3 |
| G | 0.99 | 0.9875 | 0.985 | $10^{-0}$ | 2.77e-9 | 0.99 | 0.005 | 1.4 |

Table 3: Parameters in the case II and III

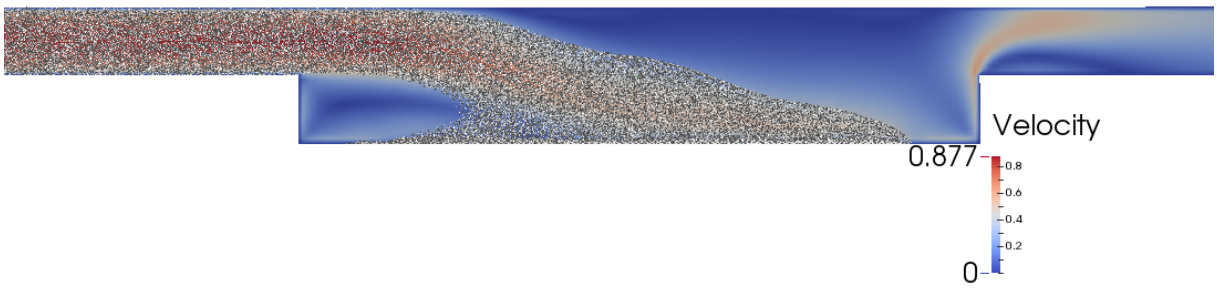



Figure 2: Particle distribution and flow field at 0.3 $s$ in the case I

The error results of case II suggest that large $\beta$ values are neglecting the limiting part of function in the equation 1 which leads to oscillations, on the other hand small $\beta$ value slows convergence. In addition, if large enough $\alpha$ value is chosen only small $\beta$ value can be disadvantage. The relative and absolute error results of the case II are gathered at the next time step after TMP changes and presented in table 5. In the case II the small values ($10^{-6} - 10^{-5}$) of $\beta$ slows convergence and errors stay at around 80 $\%$. In the caseII(1) $\alpha$ value 0.99 is not sensitive to bigger ($> 10^{-5}$) $\beta$ values so relative error stays at around 1 $\%$. In the case II(2) errors are first reducing while $\beta$ value increases, the minimum errors occurs when $\beta$ is $10^{-3}$. After reaching minimum relative error it is increasing again and 10 $\%$ relative error occurs when $\beta$ reaches value 1. If too small $\alpha$ value is chosen, as in the case II(3), the limiter prevents divergence but solution oscillates inside limiters boundaries.

| | Case I | |
|---|---|---|
| | rel.[%] | abs. [m/s] |
| A | 0 | 0.0000 |
| B | 92 | 0.2795 |
| C | 179 | -0.4006 |

Table 4: Relative and absolute errors in the case I

| | Case II | | | | | | Case III | |
|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | | |
| | rel.[%] | abs. [m/s] | rel.[%] | abs. [m/s] | rel.[%] | abs. [m/s] | rel.[%] | abs. [m/s] |
| A | 84 | 0.1360 | 84 | 0.1362 | 84 | 0.1359 | 0 | 0.0000 |
| B | 81 | 0.1120 | 81 | 0.1120 | 81 | 0.1116 | 0 | 0.0000 |
| C | 1 | 0.0003 | 1 | 0.0003 | 677 | -0.0236 | 0 | 0.0000 |
| D | 1 | 0.0001 | 0 | 0.0001 | 70 | 0.0636 | 0 | 0.0000 |
| E | 1 | 0.0001 | 1 | 0.0003 | 94 | 0.5560 | 0 | 0.0000 |
| F | 1 | 0.0001 | 1 | 0.0003 | 100 | -6.5918 | 1 | -0.0012 |
| G | 1 | 0.0001 | 10 | -0.0025 | 98 | 72.6242 | 319 | -0.1068 |

Table 5: Relative and absolute errors in the case II and III

The results, which can be found from table 5, of the case III clarify that there is dependency between $L_p$ and relaxation factor $\alpha$. Decreasing $\gamma$ value does not effect to errors but increasing $\gamma$ value increases both relative and absolute errors and induces oscillations. When $\gamma$ value exceed 1.3 the errors increase fast. This result suggest that there is minimum $\alpha$ value for every $L_p$ value which can still produce convergent solution. In the other words higher relaxation is demanded if $L_p$ increases.

The presented boundary condition is able to follow the pressure drop of packed particles as can be seen from figure 3. In the figure 3 the case I(A)'s averaged modeled flux and results of equation 2 are presented from 0 $s$ to 2 $s$. Between fluxes there are good agreement over hole sample time. In the beginning of the simulation flux is constant until particles start packing on the membrane. Packing particles induces pressure loss resulting in flux decreases until particles are fed in the domain and the hight of bed saturates.

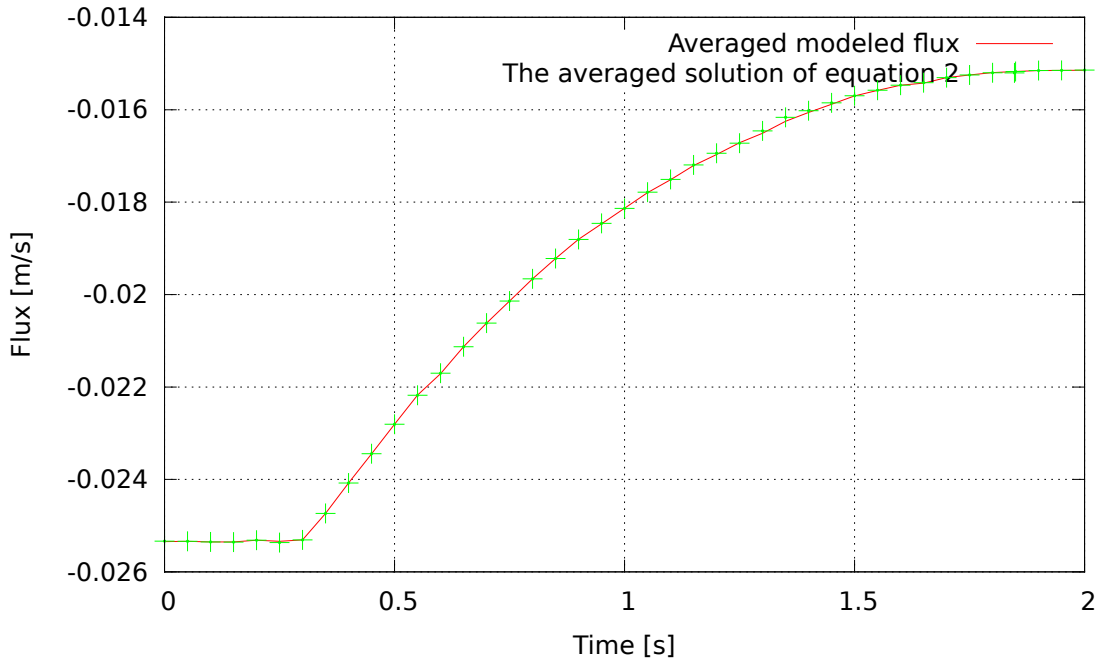



Figure 3: Averaged modeled flux and averaged flux of equation 2 over time in the case I(A)

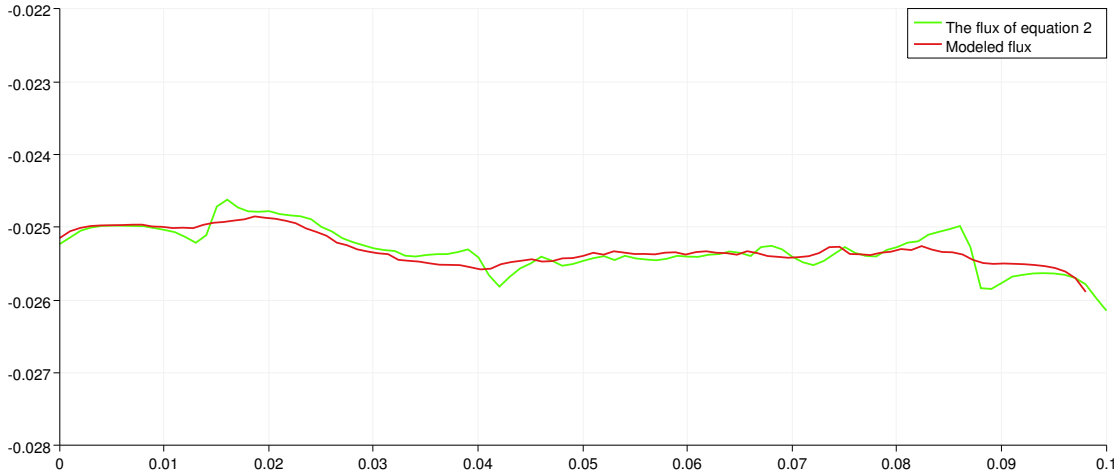



Figure 4: Flux of equation 2 and modeled flux through the membrane at 0.3 $s$ in the case I(A)

The averaged results of equation 2 and the averaged modeled fluxes are almost equal but point-to-point values on the membrane differ. This is revealed in the figure 4 where non-averaged fluxes of the Case I(A) are plotted at 0.3 $s$. In the figure 4 fluxes are mainly at good agreement but at 0.015, 0.04 and 0.09 $m$ fluxes differ. The spots (0.015 and 0.09 $m$) where modeled flux lags behind are the same spots where packed bed end in the figure 2. The lag proposes that more pressure-velocity corrections i.e inner iterations are needed and average agreement does not correlate with local unity.

The Darcy alike boundary condition has effect to particle distribution on the membrane as can be seen from figure 5. In the figure 5 the case I(A) is compared to constant flux case at 2 $s$. The constant flux is same as in the beginning of the case I(A) (-0.0254 $m/s$). The case I(A) is the upper picture in the figure 5 and the constant flux is lower picture in the figure 5. There are more particles in the constant flux case but in the case I(A) particles are more even distributed after the forward-facing step.

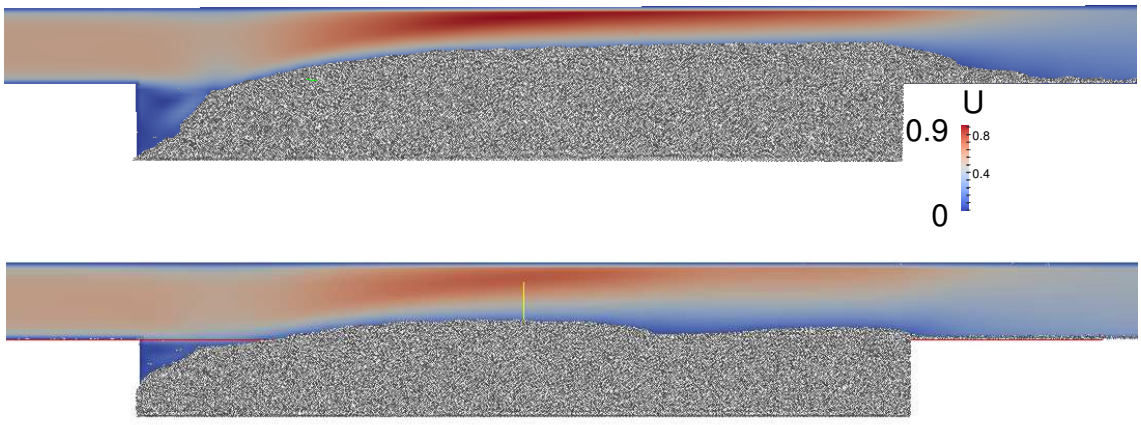



Figure 5: The flow field and particle distribution at 2 $s$ in the constant flux set up (upper) and in the case I(A) (lower).

## 4 CONCLUSIONS

In this paper explicit Darcy alike equation with relaxation factor and limiter was implemented to model permeate boundary condition. From the results can be seen the boundary condition is able to follow changing flow conditions. However, relaxation is needed and demand of relaxation depend on hydraulic permeable coefficient $L_p$ and pressure changes on a membrane. The used limiter can prevent divergence but small limiter value does not assure convergence which is why it can not be used alone.

## REFERENCES

[1] H. Al Abdulgader, N. Hilal, Hybrid ion exchange - pressure driven membrane processes in water treatment: A review. *Separation and Purification Technology*, **192**, 253–264, 2013.

[2] N. Kabay, I. Yilmaz, M. Bryjak, M. Yuksel, Removal of boron from aqueous solution by a hybrid ion exchange-membrane process. *Desalination*, **198**, 158–165, 2006.

[3] H.P. Zhu, Z.Y. Zhou, R.Y. Yang, A.B. Yu, Discrete particle simulation of particulate systems: Theoretical developments. *Chemical Engineering Science*, **62**, 3378–3396, 2007.

[4] K.J. Dong, R.P. Zou, A.B. Yu, G. Roach, DEM simulation of cake formation in sedimentation and filtration. *Minerals Engineering*, **22**, 921–930, 2009.

[5] H.P. Zhu, Z.Y. Zhou, R.Y. Yang, A.B. Yu, Discrete particle simulation of particulate systems: A review of major applications and findings. *Chemical Engineering Science*, **63**, 5728–5770, 2008.

[6] G.A. Fimbres-Weihs, D.E. Wiley, Review of 3D CFD modeling of flow and mass transfer in narrow spacer-filled channels in membranes modules. *Chemical Engineering and Processing*, **49**, 759–781, 2010.

[7] U. Merten, Flow relationships in reverse osmosis. *Industrial & Engineering Chemistry Fundamentals*, **2**, 229–232, 1963.

[8] C. Goniva, C. Kloss, N.G. Deen, J.A.M. Kuipers, S. Pirker, Influence of rolling friction on single spout fluidized bed simulation. *Particuology*, **10**, 582–591, 2012.

[9] R.I Issa, Solution of the implicitly discretised fluid flow equations by operator-splitting. *Journal of Computational Physics*, **62**, 40–65, 1986.

[10] W.R. Ketterhagen, Modeling granular segregation in flow from quasi-three-dimensional, wedge-shaped hoppers. *Powder Technology*, **179**, 126–143, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A MULTIGRID FORMULATION FOR FINITE DIFFERENCE METHODS ON SUMMATION–BY–PARTS FORM: AN INITIAL INVESTIGATION

**Andrea A. Ruggiu**[1], **Per Weinerfelt**[1,2], **Tomas Lundquist**[1] **and Jan Nordström**[1]

[1]Department of Mathematics, Computational Mathematics, Linköping University,
SE-581 83 Linköping, Sweden
e-mail: {andrea.ruggiu,per.weinerfelt,tomas.lundquist,jan.nordstrom}@liu.se

[2]Saab Aerospace,
SE-581 88 Linköping, Sweden
e-mail: per.weinerfelt@saab.se

**Keywords:** high order finite difference methods, summation–by–parts, multigrid, restriction and prolongation operators, improved convergence.

**Abstract.** *Several multigrid iteration schemes for high order finite difference methods are studied by comparing the effect of different interpolation operators. The usual choice of prolongation and restriction operators based on linear interpolation in combination with the Galerkin condition leads to coarse grid operators which are less accurate than their fine grid counterparts.*

*Moreover, these operators do not mimic the integration–by–parts property possessed by the original fine grid summation–by–part schemes and hence are intuitively less stable. In this paper, an alternative class of interpolation operators is considered to overcome these issues and improve the stability of the overall multigrid iteration scheme. As a pleasant side effect we find that also the efficiency of the iteration scheme is improved.*

## 1 INTRODUCTION

The multigrid method is a convergence acceleration technique based on a hierarchy of grids [1, 2]. The solution of a discrete problem on a fine grid is computed by solving a system of equations on the coarsest grid level. This technique is applied to problems in various branches of applied mathematics and engineering, such as the Helmholtz equation [3], magnetohydrodynamics [4] and computational fluid dynamics [5].

One of the main issues associated with this method is the construction of restriction and prolongation operators which transfer the information between grids. In this work we make use of a general relation between prolongation and restriction operators which was originally proposed in a different context [6]. These interpolation operators are consistent at the boundaries and guarantee that the accuracy of the original scheme is retained at the interior nodes on each grid level.

By using this result, we combine the multigrid method with an energy stable discretization for high order finite difference methods, based on the Summation–by–Parts (SBP) and Simultaneous–Approximation–Term (SAT) approach described in [7]. The improved stability of the new multigrid method is shown through numerical simulations of a hyperbolic problem in one dimension.

The rest of this paper is organized as follows: in Section 2 the main features of the two–level multigrid algorithm are presented. Section 3 introduces the SBP operators and the SAT penalty terms for high order finite difference discretizations. Section 4 deals with the construction of multigrid algorithms for SBP–SAT discretizations using the new class of interpolation operators. In Section 5 the new algorithm is compared to a multigrid scheme with conventional prolongation and restriction operators. Conclusions are drawn in Section 6.

## 2 THE MULTIGRID ALGORITHM

In this Section the multigrid approach is outlined by means of a two–level algorithm[1, 2, 8]. Let us consider the following steady–state problem

$$\begin{array}{rcl} Lu & = & f, \quad \text{in } \Omega, \\ Hu & = & g, \quad \text{on } \partial\Omega, \end{array} \tag{1}$$

where $L$ is a differential operator, $H$ is a boundary operator, $f$ and $g$ are given functions, and $\Omega$ is a domain. We assume that the boundary conditions are assigned in a way such that (1) is well–posed. The multigrid algorithm for solving (1) can be written in four steps:

1. Discretization;
2. Error smoothing;
3. Coarse–grid correction;
4. Fine–grid update.

### 2.1 Discretization

Construct a discretization for (1) on a grid $\Omega_1 \subset \Omega \cup \partial\Omega$ called *fine grid*. The resulting problem has the form

$$L_1\mathbf{u} = \mathbf{f}, \tag{2}$$

where $L_1$ is a discretization of the operator $L$ in (1) which also includes the boundary conditions defined by $H$. The vector $\mathbf{f}$ is a grid function which approximates $f$ on the nodes of $\Omega_1$, possibly augmented with boundary data, and $\mathbf{u}$ is an approximate solution to (1). Typically $\Omega_1$ has many nodes, and therefore it may be expensive to solve (2) directly.

The discrete operator $L_1$ is assumed to be invertible, since its continuous counterpart $L$ in (1), with the boundary conditions defined by $H$, leads to a well–posed problem. Even though we do not want to compute the inverse of $L_1$, the solution to (2) will be indicated by $L_1^{-1}\mathbf{f}$ in the rest of the paper.

### 2.2 Error smoothing

Define an error smoothing procedure. As an example, we may consider to march toward the solution to (2) from an initial guess $\mathbf{u}^{(0)}$ by solving

$$\begin{array}{rcll} \left(\frac{\mathrm{d}}{\mathrm{d}t} + L_1\right)\mathbf{w}\left(t\right) & = & \mathbf{f}, & t > 0, \\ \mathbf{w}\left(0\right) & = & \mathbf{u}^{(0)}, & \end{array} \tag{3}$$

for a given finite time $\Delta t > 0$ called *smoothing step*. The solution to (3) is

$$\mathbf{w}\left(\Delta t\right) = e^{-L_1 \Delta t}\mathbf{u}^{(0)} + \left(I - e^{-L_1 \Delta t}\right) L_1^{-1}\mathbf{f}, \tag{4}$$

where $I$ indicates the identity matrix on $\Omega_1$. Note that this technique yields a smoothing effect on the grid function $\mathbf{u}^{(0)}$ if all the eigenvalues of $L_1$ have positive real part. In particular, $\mathbf{w}\left(\Delta t\right) \to \mathbf{u} = L_1^{-1}\mathbf{f}$ as $\Delta t \to +\infty$.

By similarity, we may define a general smoothing technique for (2) as

$$\begin{array}{rcll} \mathbf{v}^k & = & S\mathbf{v}^{k-1} + \left(I - S\right) L_1^{-1}\mathbf{f}, & k = 1, \ldots, \nu, \\ \mathbf{v}^0 & = & \mathbf{u}^{(0)}, & \end{array} \tag{5}$$

where the exponential smoother $S_{exp} = e^{-L_1 \Delta t}$ yields a procedure based on time–marching as in (4). After $\nu$ steps, the iterative method (5) provides the grid function

$$\mathbf{v} = S^{\nu}\mathbf{u}^{(0)} + \left(I - S^{\nu}\right) L_1^{-1}\mathbf{f}. \tag{6}$$

This step of the algorithm is necessary since the high frequency modes on $\Omega_1$ may be unresolved on coarser grids (see Figure 1).

### 2.3 Coarse–grid correction

Consider $\Omega_2 \subsetneq \Omega_1$, called the *coarse grid*, and solve a correction problem on $\Omega_2$:

$$L_2\mathbf{d} = I_r\left(\mathbf{f} - L_1\mathbf{v}\right) \tag{7}$$

which is obtained from (2) and (4) by making use of

- a restriction operator $I_r : \Omega_1 \to \Omega_2$,
- a coarse–grid operator $L_2 : \Omega_2 \to \Omega_2$.

The coarse–grid operator can be built by using the Galerkin Condition [2]

$$L_2 = I_r L_1 I_p, \tag{8}$$

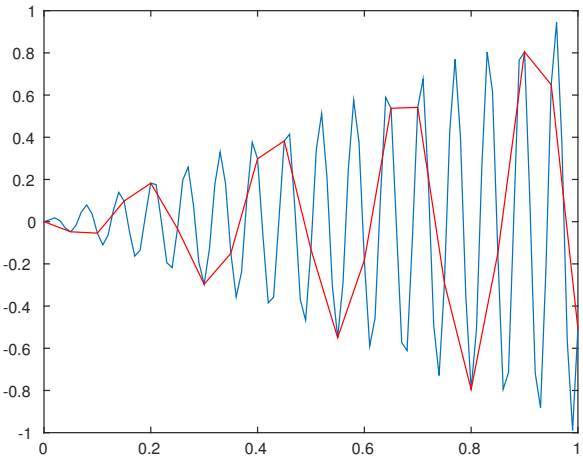



Figure 1: The function $x\sin(100x)$, $x \in [0,1]$, is represented on a fine (blue line) and on a coarse grid (red line). The fine grid is given by 101 equidistant points, while the coarse grid has got 21 nodes. The restricted grid function has a lower frequency than the original one.

where $I_p : \Omega_2 \to \Omega_1$ is a prolongation operator.

A common choice for $I_p$ in the one dimensional case is based on linear interpolation. Assuming for example that $\Omega_1 = \{x_j : x_j = jh, \quad j = 0, \ldots, N\}$ with $h = 1/N$, and that $\Omega_2$ consists of the odd nodes of $\Omega_1$, we may write

$$(I_p \mathbf{v})_m = \begin{cases} v_j, & m = 2j, \\ \frac{1}{2}\left(v_j + v_{j+1}\right), & m = 2j+1, \end{cases} \qquad j = 0, \ldots, N/2, \tag{9}$$

while the restriction operator is usually given in terms of the prolongation operator

$$I_r = \frac{1}{2} I_p^T. \tag{10}$$

### 2.4 Fine–grid update and multigrid iteration scheme

Update the fine grid solution $\mathbf{v}$ with an estimate of the correction $\mathbf{d}$ in (7), obtained through prolongation with $I_p$

$$\mathbf{u}^{(1)} = \mathbf{v} + I_p \mathbf{d}. \tag{11}$$

This relation, together with (6) and (7), provides an iterative method for solving (1):

$$\mathbf{u}^{(n+1)} = M\mathbf{u}^{(n)} + N\mathbf{f}, \tag{12}$$

where

$$M = CS^{\nu}, \qquad C = I - I_p L_2^{-1} I_r L_1 \qquad \text{and} \qquad N = (I - M)\, L_1^{-1}. \tag{13}$$

We will refer to $M$ as the multigrid iteration matrix and to $C$ as the coarse–grid correction operator.

The role of $M$ in the convergence of the iterative method (12) is central: by considering the errors $\mathbf{e}^{(n)} = \mathbf{u}^{(n)} - L^{-1}\mathbf{f}$, we get

$$\mathbf{e}^{(n+1)} = M\mathbf{e}^{(n)}, \tag{14}$$

which leads to convergence if the spectral radius of $M$ is less than 1.

### 2.5 The coarse–grid correction operator

In Section 2.3 we claimed that the Galerkin condition (8) may be used to build a coarse–grid operator $L_2$. This choice leads to some interesting theoretical properties about the coarse–grid correction operator $C$ in (12), (13). To start with, we prove

**Lemma 2.1.** *If the Galerkin condition* (8) *holds, then the coarse–grid correction operator $C$ in* (12)*,* (13) *is idempotent, i.e. $C^k = C$ for every $k \in \mathbb{N}$.*

*Proof.* By multiplying $C$ by itself we get

$$C^2 = \left(I - I_p L_2^{-1} I_r L_1\right)\left(I - I_p L_2^{-1} I_r L_1\right) = I - 2I_p L_2^{-1} I_r L_1 + I_p L_2^{-1} I_r L_1 I_p L_2^{-1} I_r L_1 = C,$$

since the Galerkin condition $L_2 = I_r L_1 I_p$ is fulfilled. □

The idempotency of $C$ implies that its eigenvalues are given by zeros and ones [9]. Moreover, it is possible to show

**Corollary 2.2.** *If the Galerkin condition* (8) *holds, the image of the prolongation operator $I_p$ is contained in the nullspace of $C$.*

*Proof.* Consider a vector $\mathbf{y} \in \mathrm{Im}\,(I_p)$, i.e. $\mathbf{y} = I_p\mathbf{z}$. When the coarse–grid correction operator acts on this vector, we find

$$C\mathbf{y} = \left(I - I_p L_2^{-1} I_r L_1\right) I_p\mathbf{z} = I_p\mathbf{z} - I_p L_2^{-1} I_r L_1 I_p\mathbf{z} = \mathbf{0},$$

since the Galerkin condition holds. □

The main consequence of Corollary 2.2 is that $C$ cancels all the grid functions which can be represented through the prolongation operator $I_p$. This implies that $C$ deals with smooth errors while the smoother $S$ damps the remaining high frequency modes.

## 3 THE SBP–SAT DISCRETIZATION

In this Section we introduce the Summation–by–Parts operators (SBP) and the Simultaneous–Approximation–Term (SAT) technique. The first ones mimic integration by parts, whereas the second one introduces penalty–like terms enforcing the boundary conditions weakly [7].

**Definition 3.1.** *We say that $D = P^{-1}Q$ is a $(p, q)$–accurate first derivative SBP operator, if $Q + Q^T = B = diag\,(-1, 0, \ldots, 0, 1)$, $P$ is a symmetric positive definite matrix, and the associated truncation errors are $\mathcal{O}\left(h^p\right)$ in the interior and $\mathcal{O}\left(h^q\right)$ at the boundaries.*

SBP operators based on diagonal norms $P$ are available for even orders $p = 2q$ in the interior, while the boundary closure is $q$th order accurate. Even though this is not the only possible choice, we will only consider these $(2q, q)$–accurate operators with diagonal $P$. For further details on the construction of SBP operators for the first derivative with $q \leq 4$, see [10].

The SBP finite difference operators together with a strong treatment of the boundary conditions only admits stability proofs for very simple problems. This was shown in [11], where SAT technique was proposed to complement the SBP schemes. By discretizing a well–posed Initial–Boundary–Value–Problem (IBVP) with both SBP operators and SAT penalty terms (the SBP–SAT approach), it is possible to prove that the corresponding semi–discrete problem is stable. As an example, we consider the advection equation in one dimension.

### 3.1 The advection problem

Consider

$$\begin{array}{rcll} u_t + u_x & = & 0, & 0 < x < 1, \quad t > 0, \\ u\left(x,0\right) & = & f\left(x\right), & 0 < x < 1, \\ u\left(0,t\right) & = & g\left(t\right), & t > 0, \end{array} \tag{15}$$

where both $f$ and $g$ are known data. The problem (15) is well–posed, since the analytical solution is a traveling wave, and the boundary condition is imposed at the inflow.

Let us consider an $(N+1)$–point uniform grid on $[0,1]$, given by $x_j = jh$ for $j = 0, \ldots, N$, where $h = 1/N$. We introduce the grid function $f_j = f\left(x_j\right)$ and to each grid point we also associate the approximate solution $u_j$. By applying the SBP–SAT discretization in space to (15), we get

$$\begin{array}{rcll} \mathbf{u}_t + P^{-1}Q\mathbf{u} & = & P^{-1}\sigma\left(u_0 - g\right)\mathbf{e}_0, & t > 0, \\ \mathbf{u}\left(0\right) & = & \mathbf{f}, & \end{array} \tag{16}$$

where $\mathbf{u} = (u_0, \ldots, u_N)$, $\mathbf{f} = (f_0, \ldots, f_N)$, $\sigma \in \mathbb{R}$ is a penalty parameter which can be tuned for stability, and $\mathbf{e}_0 = [1, 0, \ldots, 0]^T \in \mathbb{R}^{N+1}$. By defining the matrix $\widetilde{Q}_\sigma = Q - \sigma \mathbf{e}_0 \mathbf{e}_0^T$ and the vector $\mathbf{F} = -P^{-1}\sigma g \mathbf{e}_0$ we can rewrite (15) in a more compact form

$$\begin{array}{rcll} \mathbf{u}_t + P^{-1}\widetilde{Q}_\sigma\mathbf{u} & = & \mathbf{F}, & t > 0, \\ \mathbf{u}\left(0\right) & = & \mathbf{f}, & \end{array} \tag{17}$$

where $\widetilde{Q}_\sigma + \widetilde{Q}_\sigma^T = diag\left(-\left(1+2\sigma\right), 0, \ldots, 0, 1\right)$.

We want to find $\sigma$ such that the problem (17) is strongly stable, i.e. such that

$$\left\|\mathbf{u}\left(\cdot,t\right)\right\|^2 \leq K\left(t\right)\left(\left\|\mathbf{f}\right\|^2 + \max_{\tau\in[0,t]}\left\|g\left(t\right)\right\|^2\right) \tag{18}$$

in a suitable norm, with $K\left(t\right)$ bounded for any finite $t$ and independent of the data (for further details, see [7]). By multiplying the equation (17) by $\mathbf{u}^T P$ and adding the transpose we find

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}\right\|_P^2 = -u_N^2 + \left(1+2\sigma\right)u_0^2 - 2\sigma g u_0$$

where $\left\|\mathbf{u}\right\|_P = \sqrt{\mathbf{u}^T P \mathbf{u}}$. For $g = 0$ the method is stable for $\sigma \leq -\frac{1}{2}$. If $g \neq 0$ and assuming that $\sigma \neq -\frac{1}{2}$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}\right\|_P^2 = -\frac{\sigma^2}{1+2\sigma}g^2 - u_N^2 + \frac{\left[\left(1+2\sigma\right)u_0 - \sigma g\right]^2}{1+2\sigma},$$

which lead to an estimate of the form (18). Therefore, the semi–discrete problem (16) is strongly stable if $\sigma < -\frac{1}{2}$.

## 4 THE SBP–SAT–MG SCHEME

In the previous Section we have shown that operators which mimic integration by parts leads to stable semi–discrete problems. We conjecture that preserving this property on both fine and coarse levels should lead to an improved multigrid method.

Let us consider the following steady model problem

$$\begin{array}{rcll} u_x & = & f, & 0 < x < 1, \\ u\left(0\right) & = & g. & \end{array} \tag{19}$$

By applying the SBP–SAT method (see Section 3.1) we get

$$P^{-1}\widetilde{Q}_\sigma \mathbf{u} = \mathbf{F}, \tag{20}$$

where $\mathbf{F} = \mathbf{f} - P^{-1}\sigma g \mathbf{e}_0$ and $\widetilde{Q}_\sigma$ is defined above. Since our problem is written on the same form as (2), we can apply the multigrid scheme in Section 2 directly. The time–marching smoothing procedure (3) yields the semi–discrete problem (17), whose solution is stable. Following the remaining steps of the algorithm, we define a coarse grid correction problem using the Galerkin Condition (8). Unfortunately, the resulting coarse–grid operator does not satisfy the SBP property if standard restriction and prolongation operators like (9) and (10) are used.

By looking closely at the multigrid algorithm, we realize that the restriction operator chosen in (10) is consistent only in the interior. Hence, instead of using this relation, we use

$$I_r = P_2^{-1} I_p^T P \tag{21}$$

proposed in [6], where $P_2$ is a symmetric positive definite matrix of the same type as $P$ on the coarse grid. By satisfying (21) it is possible to build pairs of consistent and $p$th–order accurate prolongation and restriction operators according to the following definition of SBP–preserving interpolation operators given in [6].

**Definition 4.1.** *Let the row–vectors $\mathbf{x}_1^k$ and $\mathbf{x}_2^k$ be the projections of the monomials $x^k$ onto equidistant $1$–$D$ grids corresponding to a fine and coarse grid, respectively. We say that $I_r$ and $I_p$ are $2q$th–order accurate SBP–preserving interpolation operators if $I_r\mathbf{x}_1^k - \mathbf{x}_2^k$ and $I_p\mathbf{x}_2^k - \mathbf{x}_1^k$ vanish for $k = 0, \dots, 2q-1$ in the interior and for $k = 0, \dots, q-1$ at the boundaries.*

The SBP–preserving interpolation operators with minimal bandwidth for $q = 1$ are given by

$$I_p = \begin{bmatrix} 1 & & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & & \\ & 1 & & & & & \\ & \frac{1}{2} & \frac{1}{2} & & & & \\ & & \ddots & \ddots & & & \\ & & & & 1 & & \\ & & & & \frac{1}{2} & \frac{1}{2} & \\ & & & & & & 1 \end{bmatrix}, \qquad I_r = \frac{1}{2}\begin{bmatrix} 1 & 1 & & & & & \\ \frac{1}{2} & 1 & \frac{1}{2} & & & & \\ & \frac{1}{2} & 1 & \frac{1}{2} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{2} & 1 & \frac{1}{2} & \\ & & & & 1 & 1 & \end{bmatrix}. \tag{22}$$

Note that in this case $I_p$ coincides with the prolongation operator (9).

The SBP–preserving interpolators, together with the Galerkin Condition (8), lead to the following coarse grid operator

$$P_2^{-1}\widetilde{Q}_{2,\sigma} = I_r\left(P^{-1}\widetilde{Q}_\sigma\right) I_p = P_2^{-1} I_p^T P\left(P^{-1}\widetilde{Q}_\sigma\right) I_p = P_2^{-1}\left(I_p^T \widetilde{Q}_\sigma I_p\right),$$

where obviously $\widetilde{Q}_{2,\sigma} = I_p^T \widetilde{Q}_\sigma I_p$.

We claim

**Proposition 4.2.** *The interpolation operators satisfying* (21) *lead to a coarse grid operator $\widetilde{Q}_{2,\sigma}$ which satisfies the Summation–By–Parts property.*

The proof will be given in the full paper. Hence, through the Galerkin Condition (8) it is possible to build new SBP operators on the coarse grid. Moreover, these operators are consistent with (19) and can retain the accuracy of the original scheme at the interior nodes. We can show

**Proposition 4.3.** *Consider the $2qth$–order accurate SBP–preserving interpolation operators $I_p$ and $I_r$, with $q \geq 1$. If $D$ is a $(2q, q)$–first derivative SBP–operator in the fine grid, then $D_2 = I_r D I_p$ is a first derivative SBP–operator in the coarse grid. Moreover, it is $2qth$–order accurate in the interior nodes, while the accuracy on the boundary nodes is at least $q - 1$.*

The proof will be given in the full paper.

**Remark 4.4.** *For $q = 1$ the coarse grid operator $D_2$ has the same structure as $D$ and is first–order accurate at the boundaries.*

The operators for higher orders will be provided in the full paper.

## 5 NUMERICAL EXPERIMENTS

In this Section we perform numerical computations to test the properties of the new multigrid iteration scheme. Consider (19) as model problem with a manufactured solution given by

$$v(x) = e^{-x}\left(\cos(10\pi x) + \cos(2\pi x)\right), \quad 0 < x < 1.$$

To discretize (19) we use the formulation (20) on an (N+1)–point uniform grid on $[0, 1]$. The tests are carried out using a single–step smoother based on a modified Runge–Kutta scheme which was proposed in [12] for first order systems:

$$S = I - \Delta t L_1 + 0.6\Delta t^2 L_1^2 - 0.36\Delta t^3 L_1^3. \tag{23}$$

In the following tests, we consider that the multigrid iteration has reached convergence if $\|\mathbf{u}^{n+1} - \mathbf{u}^n\| < \varepsilon$, where $\varepsilon =$ tolerance.

### 5.1 IMPROVED CONVERGENCE

Consider a $(2, 1)$–order accurate discretization on a grid with $N = 500$ and a smoothing step $\Delta t = 5 \cdot 10^{-4}$. Figure 2 shows the spectral radius and the distribution of the eigenvalues of the multigrid iteration matrix $M$ in (13) by varying the interpolation operators. In particular, the SBP–preserving interpolation operators of 2nd order are tested against the classical choices (9) and (10). Note that even if in this case the prolongation operator is the same (see (22)), a minor change in the restriction operator affects the distribution of the eigenvalues of $M$.

To test the convergence properties, we have used both the tolerance $10^{-10}$ and $10^{-3}$. The results, shown in Table 1, suggest that the new class of operator is clearly superior.

| | SBP–preserving | Conventional |
|---|---|---|
| Number of iterations ($10^{-10}$) | 2372 | $> 10^6$ |
| Number of iterations ($10^{-3}$) | 1470 | 1541 |

Table 1: The number of iterations to convergence for the multigrid scheme with tolerance $10^{-10}$ and $10^{-3}$.

Another test with tolerance $\varepsilon = 10^{-10}$ and an increased smoothing step $\Delta t = 6.3 \cdot 10^{-4}$ makes the spectral radius for the SBP–preserving choice decrease, while the classical linear prolongation and restriction lead to instability. As a side effect, the number of iterations to convergence for the new interpolation operators drops to 1873. In Figure 3 the distributions of the eigenvalues are shown for this case. Note that the current smoothing step is close to the

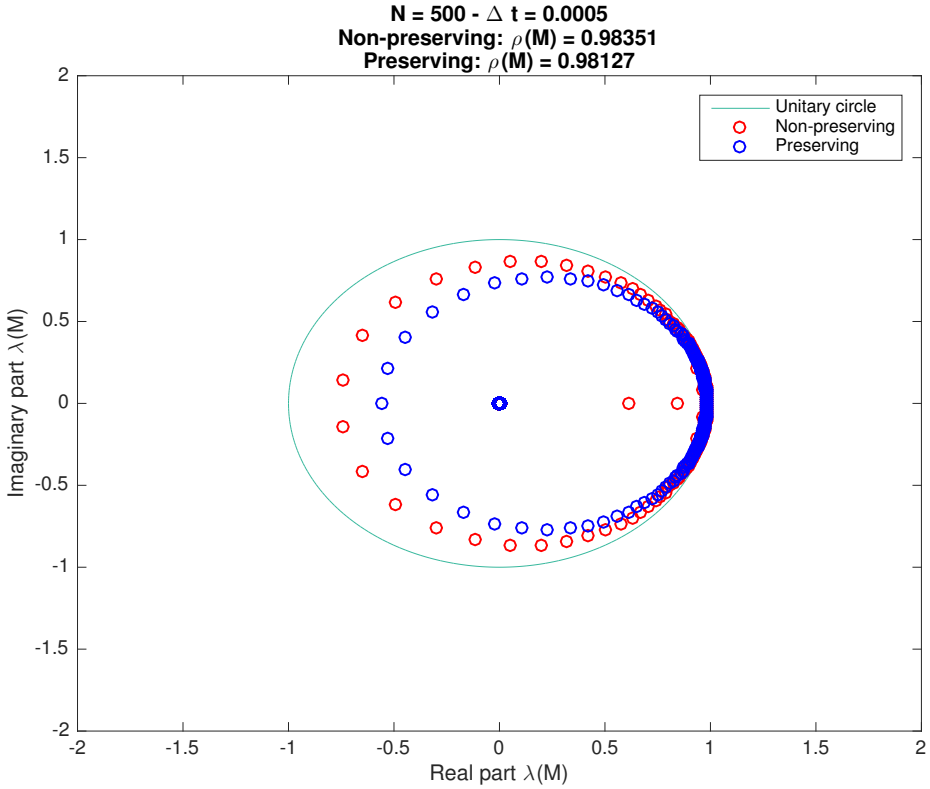



Figure 2: Eigenvalues of the multigrid iteration matrix for different choices of interpolation operators. The spectral radius of $M$ with SBP–preserving interpolators is less than the one with the conventional prolongation and restriction.

stability limit. This phenomenon can be evaluated by looking at the eigenvalues on the left side of the picture: increasing the time–step $\Delta t$, the magnitude of these eigenvalues becomes bigger.

These numerical experiments suggest that the SBP–preserving interpolation operators yield multigrid iteration schemes with significantly improved convergence properties.

## 5.2 HIGHER ORDER METHODS

The following tests are carried out by raising the order of accuracy of the fine grid first–derivative operator. The tolerance $\varepsilon = 10^{-10}$ was used.

| | Minimum iterations | $\Delta t$ | $\rho\,(M)$ | $\Delta t^*$ |
|---|---|---|---|---|
| $(4,2)$, SBP–preserving | 1435 | $5.8 \cdot 10^{-4}$ | 0.96882 | $5.9 \cdot 10^{-4}$ |
| $(4,2)$, Conventional | 6358 | $1.3 \cdot 10^{-4}$ | 0.99404 | $3.4 \cdot 10^{-4}$ |
| $(6,3)$, SBP–preserving | 1278 | $5.1 \cdot 10^{-4}$ | 0.96955 | $6.8 \cdot 10^{-4}$ |
| $(6,3)$, Conventional | 7030 | $9.8 \cdot 10^{-5}$ | 0.99485 | $2.6 \cdot 10^{-4}$ |
| $(8,4)$, SBP–preserving | 1856 | $6.8 \cdot 10^{-4}$ | 0.98502 | $7.2 \cdot 10^{-4}$ |
| $(8,4)$, Conventional | 10817 | $7.9 \cdot 10^{-5}$ | 0.99667 | $2.3 \cdot 10^{-4}$ |

Table 2: The minimum number of iterations and the corresponding smoothing step are presented for each order of accuracy. For the SBP–preserving choice the minimum amount of iterations is almost the same for any order. The spectral radius $\rho\,(M)$ and the time–step limit $\Delta t^*$ are also given for each scheme.

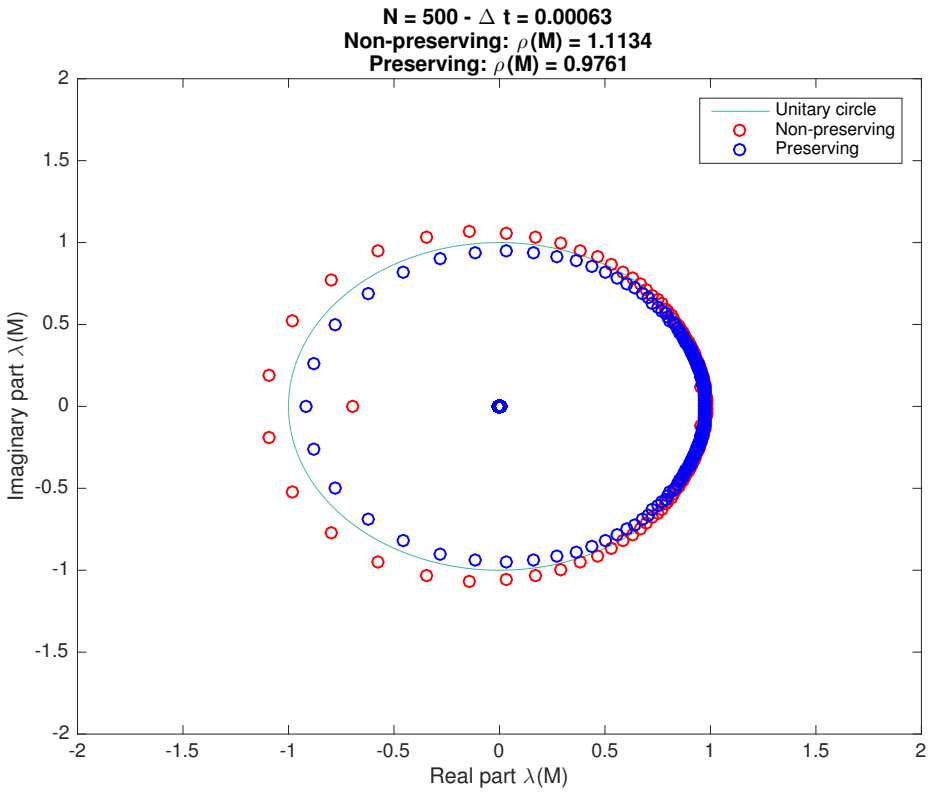



Figure 3: Increasing the smoothing step to $\Delta t = 6.3 \cdot 10^{-4}$ leads to an unstable multigrid iteration scheme, if the non–preserving choice is made.

In Table 2 the effects of the interpolation operators on the multigrid iteration scheme are shown. For each order of accuracy, the SBP–preserving interpolators lead to faster convergence. Moreover, the minimum amount of iterations does not change considerably for higher orders. Conversely, when the order of accuracy of the fine–grid operator is increased and the classical linear interpolation (9) and (10) is used, the number of iterations grows and the time–step limit $\Delta t^*$ decreases.

## 6 CONCLUSIONS

A new multigrid scheme for SBP–SAT discretizations has been proposed. So called SBP–preserving interpolation operators have been used to convey information between grids. When the Galerkin condition is considered, these prolongation and restriction operators lead to coarse–grid operators which satisfy the summation–by–part property.

Numerical experiments show that when compared to a multigrid scheme with the conventional choice of prolongation and restriction based on linear interpolation, the new scheme allows the use of larger smoothing steps and reaches convergence in fewer iterations. Hence, for every order of accuracy the SBP–preserving interpolation operators lead to multigrid iteration schemes which are more stable and have better convergence properties.

## REFERENCES

[1] U. Trottenberg, C.W. Oosterlee, A. Schuller, *Multigrid*, Academic Press, 2000.

[2] W. L. Briggs, V. E. Henson, S. F. McCormick, *A Multigrid Tutorial*, Second Edition, SIAM, 2000.

[3] C. Stolk, M. Ahmed, S. K. Bhowmik, *A Multigrid Method for the Helmholtz Equation with Optimized Coarse Grid Corrections*, SIAM Journal on Scientific Computing, Volume 36, Issue 6, 2014.

[4] J. H. Adler et al., *Monolithic Multigrid Methods for Two–Dimensional Resistive Magnetohydrodynamics*, SIAM Journal on Scientific Computing, Volume 38, pp. B1-B24, 2016.

[5] C. D. Santiago, C. H. Marchi, L. F. Souza, *Performance of geometric multigrid method for coupled two–dimensional systems in CFD*, Applied Mathematical Modelling, Volume 39, Issue 9, pp. 2602-2616, 2015.

[6] K. Mattson, M. H. Carpenter, *Stable and Accurate Interpolation Operators for High–Order Multi–Block Finite–Difference Methods*, SIAM Journal on Scientific Computing, Volume 32, Issue 4, 2010.

[7] M. Svärd, J. Nordström, *Review of Summation-By-Parts Schemes for Initial-Boundary-Value Problems*, Journal of Computational Physics, Volume 268, pp. 17-38, 2014.

[8] A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Berlin Heidelberg, pp.65-71, 2008.

[9] L. Hogben, *Handbook of Linear Algebra*, CRC Press, p. **2**-12, 2007.

[10] B. Strand, *Summation by Parts for Finite Difference Approximations for d/dx*, Journal of Computational Physics, Volume 110, pp. 47-67, 1994.

[11] M. H. Carpenter, D. Gottlieb, S. Abarbanel, *Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes*, Journal of Computational Physics, Volume 111 No. 2, pp. 220-236, 1994.

[12] B. Gustafsson, P. Lötstedt, *Proceedings of the Fourth Copper Mountain Conference on Multigrid Methods*, pp.181-233, SIAM, Philadelphia, 1989.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# IMPROVED DUAL TIME–STEPPING USING SECOND DERIVATIVES

**Jan Nordström and Andrea A. Ruggiu**

Department of Mathematics, Computational Mathematics, Linköping University,
SE-581 83 Linköping, Sweden
e-mail: {jan.nordstrom,andrea.ruggiu}@liu.se

**Keywords:** Dual–time stepping, Convergence acceleration, Steady–state solutions.

**Abstract.** *We present an improved and modified formulation of the dual time–stepping technique which makes use of two derivatives in pseudo–time. This new technique, based on the critically damped harmonic oscillator, retains the convergence properties to the stationary solution. Furthermore, when compared with the conventional time–marching, the dual time–stepping with two derivatives reduces the stiffness of the problem and requires fewer iterations for full convergence to steady–state.*

## 1 INTRODUCTION

The Dual Time–Stepping (DTS) technique has been first introduced in [1] for solving a large system of nonlinear equations obtained after an implicit time–discretization of the compressible Euler equations. In the original formulation, the DTS procedure consists in adding the derivative of the solution with respect to the so–called dual time and marching in time to reach the steady–state. Later, the same approach was successfully used in [2] for incompressible Euler and Navier–Stokes equations and it became commonly used in computational fluid dynamics. Other examples in which derivatives in fictitious time are introduced to solve systems of nonlinear equations include various engineering fields such as magnetohydrodynamics [3], simulations of launch environments [4] and electronics [5].

One drawback of the dual–time stepping technique is that the pseudo–time iterations must be fully converged in order to obtain an error estimate for time accuracy [6]. Moreover, if the dual time integration is carried out with an explicit scheme, the method may become unstable for pseudo–time steps exceeding the physical ones [7]. These two limitations may lead to a large number of dual–time iterations and hence to a computationally expensive method.

For these reasons, significant efforts have been made during the last decade to improve the performances of DTS. One strategy to accelerate the computations is to introduce a pre-conditioner multiplying the pseudo–time derivative [2, 8, 9]. Further improvements can be achieved by developing hybrid discretizations for the physical–time derivative. In [6, 10] the Alternating–Direction Implicit (ADI) scheme [11] is used in conjunction with the common BDF2. Another example is provided in [12], where the hybrid scheme is built with the Lower–Upper Symmetric–Gauss-Seidel (LU–SGS) method [13]. A further improvement of the dual time–marching is proposed in [14], where a local time–stepping approach is used.

The goal of this paper is to explore if we can accelerate the convergence of DTS by adding a second order pseudo–time derivative. The article is organized as follows: in Section 2, the DTS technique is presented and its convergence properties are shown. Section 3 describes a new class of dual time–marching procedures and introduces the second–derivative DTS. In Section 4, numerical simulations that corroborate the theoretical results are presented, while in Section 5 the main drawbacks of the scheme and alternative formulations are discussed. In Section 6 conclusions are drawn.

## 2 THE DUAL TIME–STEPPING TECHNIQUE

We start by illustrating how DTS is applied to a fully discretized hyperbolic problem.

### 2.1 The hyperbolic model problem

Consider the one–dimensional advection equation

$$u_t + a u_x = 0, \quad x \in \Omega, \quad t > 0, \tag{1}$$

where $a$ is a positive constant and $\Omega$ the spatial domain. Let $u_x \approx D\mathbf{u}$ be a general discretization of the spatial derivative, where $\mathbf{u}$ is the vector approximating the solution on a spatial grid. By applying Euler–backward scheme in time to (1) and indicating by $\Delta t$ the time–step, we get

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + aD\mathbf{u}^{n+1} = 0. \tag{2}$$

Here, $\mathbf{u}^{n+1}$ and $\mathbf{u}^n$ represent the approximated solution at the different times $t^{n+1} = (n+1)\Delta t$ and $t^n = n\Delta t$, respectively. The calculation of $\mathbf{u}^{n+1}$ by directly inverting the matrix

$(I/\Delta t + aD)$ in (2) may be excessively expensive. Instead, we apply the DTS technique by renaming $\mathbf{u}^{n+1}$ with $\mathbf{w}$ and adding the dual–time derivative $\mathbf{w}_\tau$ on the left hand–side of (2) obtaining

$$\mathbf{w}_\tau + \frac{\mathbf{w} - \mathbf{u}^n}{\Delta t} + aD\mathbf{w} = 0, \qquad \tau > 0. \tag{3}$$

If the solution $\mathbf{w}$ in (3) reaches steady–state, it will converge to $\mathbf{u}^{n+1}$ in (2). The scheme (3) can be rewritten in the following compact form

$$\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}, \qquad \tau > 0, \tag{4}$$

where $F = I/\Delta t + aD$, $\mathbf{R} = \mathbf{u}^n/\Delta t$ and $I$ is the identity matrix.

### 2.2 Nonlinear problems

Under mild restrictions, nonlinear differential problems can be related to linear formulations. As an example, consider a fully discretized problem using Euler–backward in time,

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + L\left(\mathbf{u}^{n+1}\right) = \mathbf{0}. \tag{5}$$

In (5), $L(\mathbf{u})$ is a nonlinear operator, typically coming from a nonlinear space approximation. Assuming small variations of the solution in time, a linearization of $L$ can be performed:

$$L\left(\mathbf{u}^{n+1}\right) = L\left(\mathbf{u}^n\right) + \frac{\partial L}{\partial \mathbf{u}}\Delta\mathbf{u} + O\left(\|\Delta\mathbf{u}\|^2\right), \tag{6}$$

where $\partial L/\partial\mathbf{u}$ is the Jacobian matrix of $L$ and $\Delta\mathbf{u} := \mathbf{u}^{n+1} - \mathbf{u}^n$. By substituting (6) into (5) we obtain the linear problem

$$\left(\frac{I}{\Delta t} + \frac{\partial L}{\partial \mathbf{u}}\right)\Delta\mathbf{u} \simeq -L\left(\mathbf{u}^n\right),$$

which can be solved using the dual–time stepping technique (4) by indicating with $F = I/\Delta t + \partial L/\partial\mathbf{u}$ and $\mathbf{R} = -L\left(\mathbf{u}^n\right)$. Hence, one can relate general problems to the linear setting, as long as the Jacobian matrix $\partial L/\partial\mathbf{u}$ is well–defined. The assumption of small variations in time can be fulfilled by considering sufficiently small time steps $\Delta t$.

### 2.3 Convergence

A general linear time–space discretization of a differential problem has the form

$$F\mathbf{u} = \mathbf{R}, \tag{7}$$

where $F$ is a nonsingular matrix and $\mathbf{R}$ is given and independent of $\mathbf{u}$. Moreover, we assume that $F$ is diagonalizable, i.e. $F = X\Lambda X^{-1}$ where $X$, $\Lambda$ are the matrices of the eigenvectors and eigenvalues of $F$, respectively.

By adding a dual–time derivative to the left hand–side of (7) we obtain (4), which converges in dual time to the solution of (7) since the following proposition holds.

**Proposition 2.1.** *Let all the eigenvalues of the diagonalizable $F$ have a positive real part. Then the dual–time dependent problem* (4) *converges to the solution of* (7).

*Proof.* Applying the eigendecomposition of $F$ to (4) yields

$$\mathbf{v}_\tau + \Lambda \mathbf{v} = X^{-1}\mathbf{R}, \qquad \tau > 0, \tag{8}$$

where $\mathbf{v} = X^{-1}\mathbf{w}$. By multiplying (8) with $e^{\Lambda\tau}$ from the left and integrating we find

$$\mathbf{v}\left(T\right) = e^{-\Lambda T}\mathbf{v}\left(0\right) + \Lambda^{-1}\left(I - e^{-\Lambda\tau}\right)X^{-1}\mathbf{R},$$

which converges as $T \to +\infty$ if all the eigenvalues of $F$ have positive real parts. Finally, the steady–state solution $\mathbf{w} = F^{-1}\mathbf{R}$ is recovered by multiplying $\mathbf{v}$ with $X$. □

**Remark 2.2.** *Since $F$ can be diagonalized, the eigenvalues contain all the information needed for convergence. In particular, the eigenvalue with the minimum real part determines the convergence rate in* (4)*.*

### 2.4 A note on preconditioning

To increase the convergence we may introduce a preconditioner $\Gamma$ which multiplies the first–derivative term in (4), yielding

$$\Gamma^{-1}\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}. \tag{9}$$

The optimal choice of $\Gamma$ in (9) depends on the specific problem, and will not be discussed in detail in this paper. We simply observe that the choice $\Gamma = cF^{-1}$, with $c > 0$, leads to a problem whose convergence does not depend on the eigenvalues of $F$, since (9) becomes

$$\mathbf{w}_\tau + c\mathbf{w} = cF^{-1}\mathbf{R}. \tag{10}$$

Note that, according to Proposition 2.1, this formulation is always convergent. On the other hand, even though the magnitude of $c$ can be chosen in order to get a fast convergence of (4), the formulation (10) requires the inverse of $F$, which we want to avoid.

### 2.5 Model Problem

The proof of Proposition 2.1 indicates that rather than considering the matrix–vector problem (7) at once, one may instead study the scalar model problem

$$w_\tau + \lambda w = r, \qquad \tau > 0, \tag{11}$$

defined by each row in (8) separately, with the corresponding steady–state solution

$$\lambda u = r, \qquad \lambda \in \mathbb{C} \setminus \{0\}. \tag{12}$$

## 3 THE SECOND–DERIVATIVE DTS TECHNIQUE

To possibly get an even faster decay to steady–state, we add two pseudo–time derivatives to the fully–discretized problem (7),

$$\mathbf{w}_{\tau\tau} + 2C\mathbf{w}_\tau + F\mathbf{w} = \mathbf{R}, \qquad \tau > 0, \tag{13}$$

where $C$ is an arbitrary matrix to be chosen in order to improve the convergence.

**Remark 3.1.** *A matrix parameter in front of the second derivative term in* (13) *would play the same role as $\Gamma^{-1}$ in* (9) *for the classical DTS formulation, and hence we consider* (13) *to be the most general DTS formulation involving two derivatives in dual time.*

We choose a diagonazable matrix $C = X\Theta X^{-1}$ in (13) with the same eigenvectors as $F$. This allows us to rewrite (13) as a system of independent ODEs of the form

$$w_{\tau\tau} + 2\theta w_\tau + \lambda w = r, \qquad \tau > 0, \tag{14}$$

where $\theta, \lambda$ are eigenvalues of $C$ and $F$, respectively. Note that the steady–state solution of (14) solves the model problem (12). Thus, the convergence properties of the classical and second–derivative DTS can be compared by studying the scalar equations (11) and (14).

The second–order ordinary differential equation (14) can be written as a system of first–order equations

$$\mathbf{z}_\tau + A\mathbf{z} = \mathbf{b}, \quad \text{where} \quad \mathbf{z} = \begin{bmatrix} w \\ w_\tau \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -1 \\ \lambda & 2\theta \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ r \end{bmatrix}. \tag{15}$$

By using the matrix exponential notation the solution to the system (15)

$$\mathbf{z}(\tau) = e^{-A\tau}\left(\mathbf{z}(0) - A^{-1}\mathbf{b}\right) + A^{-1}\mathbf{b} \tag{16}$$

converges to $A^{-1}\mathbf{b} = [u, 0]^T$, i.e. $w(\tau) \to u$, for any $\mathbf{z}(0)$ as $\tau \to +\infty$, if the eigenvalues of $A$ have positive real parts.

**Remark 3.2.** *The matrix exponential $e^{-A\tau}$ can be obtained by the Jordan form of $A = VJV^{-1}$, where $V$ is invertible and $J$ is a triangular matrix composed by Jordan blocks. In particular, $e^{-A\tau} = Ve^{-J\tau}V^{-1}$ and the eigenvalues of $A$ characterize the convergence of* (15)*. Note that for distinct eigenvalues, $J$ and $V$ are the matrices containing the eigenvalues and eigenvectors of $A$, respectively.*

### 3.1 Initial convergence analysis

A physical interpretation of (14) is given by the damped harmonic oscillator, if the coefficients $\theta$ and $\lambda$ are real. It is well–known that this system converges to steady–state if both $\theta$ and $\lambda$ in (14) are positive. Furthermore, the system approaches steady–state as quickly as possible, without oscillating, when it is critically damped, i.e. when $\theta = \sqrt{\lambda}$. In this section we will use these results as guidelines for the case with complex coefficients. Hence in the rest of the section, unless stated otherwise, we consider $\theta \in \mathbb{R}$, $\lambda \in \mathbb{R} \setminus \{0\}$.

From Proposition 2.1, the classical DTS in (11) converges to the steady–state solution as $T \to +\infty$ if $\lambda > 0$. For the second–derivative DTS (14) we prove

**Proposition 3.3.** *Let $\theta$ and $\lambda$ be real coefficients. The solution to the problem* (14) *converges to its steady–state solution as $\tau \to +\infty$ if $\theta$ and $\lambda$ are positive.*

*Proof.* The solution (16) converges to the steady–state solution if all the eigenvalues of $A$ have positive real parts. The characteristic polynomial of the matrix $A$ is $\mu^2 - 2\theta\mu + \lambda = 0$ and leads to the eigenvalues

$$\mu_{1,2} = \theta \pm \sqrt{\theta^2 - \lambda}. \tag{17}$$

If $\theta$ and $\lambda$ are positive, then both real parts of $\mu_{1,2}$ in (17) are positive and the solution converges. □

Next, our aim is to find conditions on $\theta$ to achieve a faster convergence than the classical Dual Time–Stepping technique in (11), i.e. we need

$$Re(\mu_{1,2}) \geq Re(\lambda), \tag{18}$$

where $\mu_1$ and $\mu_2$ are given by (17). Condition (18) gives rise to

**Proposition 3.4.** *The solution to the unsteady problem* (14) *converges to the steady–state faster than the solution to* (11) *as* $\tau \rightarrow +\infty$ *if*

$$0 < \lambda \leq 1 \qquad \textit{and} \qquad \lambda \leq \theta \leq \frac{1+\lambda}{2}. \tag{19}$$

*Proof.* By substituting (17) into (18) and observing that $\theta, \lambda \in \mathbb{R}$, we find

$$Re\left(\sqrt{\theta^2 - \lambda}\right) \geq \lambda - \theta, \tag{20a}$$

$$Re\left(\sqrt{\theta^2 - \lambda}\right) \leq \theta - \lambda. \tag{20b}$$

In order to prove the claim, we show that the values of $\theta$ outside the interval $[\lambda, (1+\lambda)/2]$ do not satisfy (20). If $\theta < \lambda$, then $Re\left(\sqrt{\theta^2 - \lambda}\right) > \theta - \lambda$ and consequently (20b) does not hold. Given that $\theta \geq \lambda$, the condition (20a) is always fulfilled. By squaring (20b), we get

$$\lambda\left[\lambda + (1 - 2\theta)\right] \geq 0,$$

which is satisfied for $\lambda > 0$ and $\theta \leq (1+\lambda)/2$. The remaining constraint of the claim, i.e. $\lambda \leq 1$, is necessary in order to guarantee that $\lambda \leq (1+\lambda)/2$. □

Proposition 3.4 provides a condition on the coefficient $\theta$ that leads to faster decay of (14) with respect to (11) for any fixed $\lambda \in \mathbb{R} \setminus \{0\}$. It is legitimate to ask if there exists an optimal choice of the free parameter $\theta$ among all the values which leads to improved convergence.

**Proposition 3.5.** *The choice* $\theta = \sqrt{\lambda}$ *provides the fastest decay for the second–derivative time–stepping formulation* (14)*.*

*Proof.* The eigenvalue of the matrix $A$ in (15) with the smallest real part determines the decay to the steady–state solution. According to (17), this eigenvalue has a real part given by

$$Re\left(\mu_1\right) = \begin{cases} \theta, & \text{if } \theta^2 < \lambda, \\ \theta - \sqrt{\theta^2 - \lambda}, & \text{if } \theta^2 \geq \lambda. \end{cases} \tag{21}$$

Since the real part of $\mu_1$ increases for $\theta$ less than $\sqrt{\lambda}$ and decreases for $\theta$ greater than $\sqrt{\lambda}$, we conclude that $\theta = \sqrt{\lambda}$ maximizes the real part of $\mu_1$. □

From (17), the optimal value of $\theta$ implies that the eigenvalues of $A$ in (15) are $\mu_1 = \mu_2 = \sqrt{\lambda}$. The new and optimal DTS formulation (14) can be rewritten as

$$w_{\tau\tau} + 2\sqrt{\lambda} w_\tau + \lambda w = r. \tag{22}$$

This formulation leads to convergence if $\lambda > 0$. Moreover, faster decay with respect to (11) is achieved if $0 < \lambda \leq 1$, since in this case $\sqrt{\lambda} \geq \lambda$.

Next, we consider directly the formulation (22) with $\lambda \in \mathbb{C} \setminus \{0\}$ and prove

**Proposition 3.6.** *The solution to the problem* (22) *converges to its steady–state solution as* $\tau \rightarrow +\infty$ *if, and only if,* $\lambda$ *is not a negative real number.*

*Proof.* The problem (22) can be written as the system of first–order equations (15) with $\theta = \sqrt{\lambda}$. The eigenvalues of $A$ are $\mu_1 = \mu_2 = \sqrt{\lambda}$ and lead to convergence if $Re\left(\mu_{1,2}\right) > 0$. The number $\sqrt{\lambda}$, interpreted as the principal square root of $\lambda$, has always a non–negative real part. If $\lambda$ is a negative real number, then $Re\left(\sqrt{\lambda}\right) = 0$ which implies non–convergence. □

In conclusion, the optimal choice in (13) is $C = F^{\frac{1}{2}} = X\Lambda^{\frac{1}{2}}X^{-1}$. In $\Lambda^{\frac{1}{2}}$ only the principal square roots are considered, i.e. the square roots with non–negative real parts.

### 3.2 The new DTS technique

Consider the new DTS technique applied to the original problem (7)

$$\mathbf{w}_{\tau\tau} + 2F^{\frac{1}{2}}\mathbf{w}_{\tau} + F\mathbf{w} = \mathbf{R}. \tag{23}$$

We can now prove

**Proposition 3.7.** *The decay to steady–state for the new DTS formulation* (23) *is determined by the square roots of the eigenvalues of* $F$.

*Proof.* The pseudo–time differential problem (23) can be written as a system of first–order equations

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{w}_{\tau} \end{bmatrix}_{\tau} + \begin{bmatrix} I & 0 \\ F^{\frac{1}{2}} & I \end{bmatrix}^{-1} \begin{bmatrix} F^{\frac{1}{2}} & -I \\ 0 & F^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} I & 0 \\ F^{\frac{1}{2}} & I \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_{\tau} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}. \tag{24}$$

Clearly, the convergence of the system is determined by the eigenvalues of $F^{\frac{1}{2}}$. □

The main consequence of Proposition 3.7 is that the new DTS formulation (23) converges to steady–state if the eigenvalues of $F$ are non–zero and do not lie on the negative real axis. If the eigenvalues of $F$ have positive real parts, then both the DTS formulations (4) and (23) are time convergent. In particular, the decay rates are determined by the eigenvalue with the smallest real part of $F$ and $F^{\frac{1}{2}}$, respectively.

**Remark 3.8.** *Note that the new DTS technique* (23) *can drive the solution to steady–state, when the classical one* (4) *fails to do that.*

**Remark 3.9.** *The square root of a number close to the imaginary axis has an output which is more distant from it. Similarly, if it is applied to a number with a large magnitude, the square root returns a number less distant from the origin. These two effects are illustrated in Figure 1.*

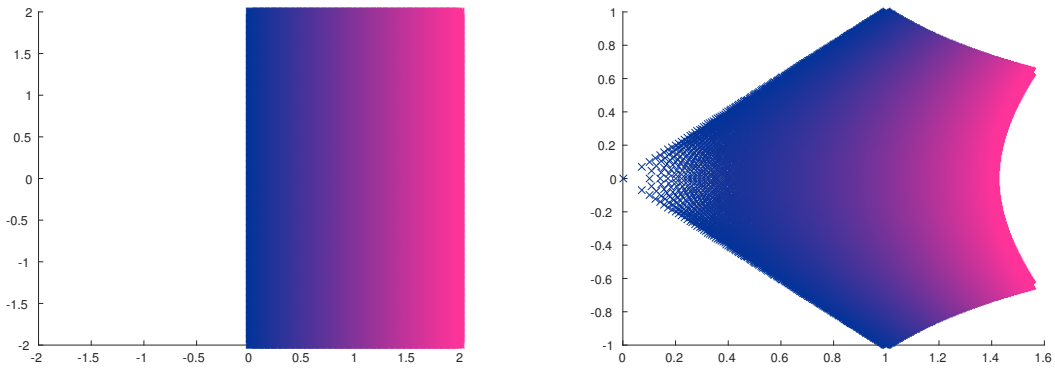



Figure 1: Complex numbers with nonnegative real parts (left figure) and their square root. Note that the distribution of points near the origin of the complex plane tends to rarefy.

As pointed out in Remark 3.9, if $F$ has eigenvalues close to the imaginary axis, the second–derivative DTS decays faster than the classical formulation. Another effect of the square root is that it narrows the spectrum of $F$ and enables the use of larger dual time–steps for an explicit time–integrator without causing stability issues.

## 4 NUMERICAL EXPERIMENTS

In this Section we perform numerical tests for both the classical (4) and the new DTS technique (23).

### 4.1 First–order ordinary differential equations

Consider the hyperbolic steady problem

$$\begin{array}{rcl} u_x & = & f, \quad 0 < x < 1, \\ u\left(0\right) & = & g, \end{array} \tag{25}$$

where $f\left(x\right) = 10\pi\cos\left(10\pi x\right)$ and $g = 1$. The solution to (25) is $u\left(x\right) = \sin\left(10\pi x\right) + 1$.

To discretize (25), we use an (N+1)–point uniform grid over $[0, 1]$, where $x_j = jh$, $j = 0, \ldots, N$ and $h = 1/N$. Let $\mathbf{f}$ be a grid function such that $f_j = f\left(x_j\right)$ and $\mathbf{u}$ the approximate solution to (25). By applying a Summation–by–Parts (SBP) discretization to (25) for the derivative and Simultaneous–Approximation–Terms (SAT) to impose the boundary condition (see the Appendix for details and [15] for references), we get

$$P^{-1}Q\mathbf{u} = \mathbf{f} - P^{-1}\mathbf{e}_0\left(u_0 - g\right), \tag{26}$$

where $\mathbf{e}_0 = \left[1, 0, \ldots, 0\right]^T \in \mathbb{R}^{(N+1)}$. Note that (26) has the form (7) with

$$F = P^{-1}\left(Q + E_0\right), \quad \mathbf{R} = \mathbf{f} + P^{-1}\mathbf{e}_0 g, \quad E_0 = \text{diag}\left(1, 0, \ldots, 0\right). \tag{27}$$

The penalty term in (26) makes the classical DTS technique (4) stable in the $P$–norm $\left\|\mathbf{w}\right\|_P = \sqrt{\mathbf{w}^T P\mathbf{w}}$ (see Appendix B). Also, the new DTS (23) applied to (26) gives rise to a stable scheme since Proposition 3.7 holds.

**Remark 4.1.** *The classical pseudo–time marching technique* (4) *is convergent since all the eigenvalues of* $F$ *have positive real parts. The new DTS formulation also converges since* $F^{\frac{1}{2}}$ *has only eigenvalues with positive real part, see [16] for details.*

We use a spatial increment $h = 0.01$ to represent the solution on $[0, 1]$ and the fourth order Runge–Kutta scheme as time–integrator. For both the schemes (4) and (23), we have used $\mathbf{w} = \left[1, \ldots, 1\right]^T$ as the initial guess. Let $\mathbf{w}^n$ be the solution to either (4) or (23) at the time $\tau^n = n\Delta\tau$. We consider the solution to be converged if $\left\|\mathbf{w}^n - \mathbf{u}\right\|_P < 10^{-6}$, where $\mathbf{u}$ is the solution to (26). The improved convergence can be seen directly by comparing the spectra of $F$ and $F^{\frac{1}{2}}$. From Figure 2, it is clear that the second–derivatives DTS has better convergence properties since the eigenvalue with minimum real part is further away from the imaginary axis. The minimum number of iterations to convergence for the classical DTS is 178, corresponding to $\Delta\tau = 0.0177$. Figure 3 shows that the new DTS (23) allows for the use of larger dual time–steps, since this formulation is less stiff than the classical one. The minimum number of iterations for the new DTS formulation (23) is 36, which is reached for $\Delta\tau = 0.199$. We can conclude that the new DTS formulation is approximately five times more efficient than the old one, for this problem.

### 4.2 A model of the time–dependent compressible Navier–Stokes equations

Next, we study both DTS approaches applied to the following system

$$\begin{array}{rcll} \mathbf{u}_t + A\mathbf{u}_x & = & \varepsilon B\mathbf{u}_{xx} + \mathbf{F}\left(x,t\right), & 0 < x < 1, \quad t > 0, \\ \mathbf{u}\left(x,0\right) & = & \mathbf{f}\left(x\right), & 0 < x < 1, \\ \left(u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x}\right)\left(0,t\right) & = & g_0\left(t\right), & t > 0, \\ \left(u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x}\right)\left(1,t\right) & = & g_1\left(t\right), & t > 0, \end{array} \tag{28}$$

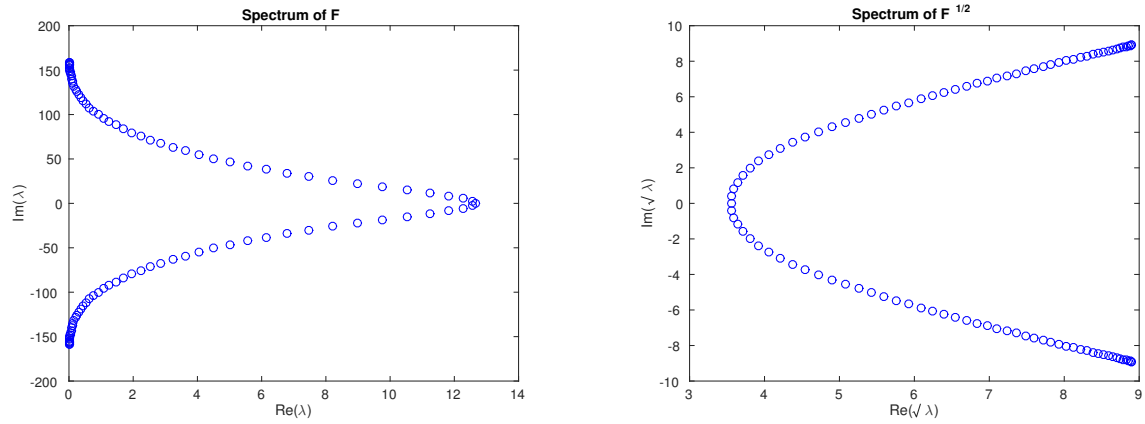



Figure 2: The spectrum of $F$ and of its square root for the first order problem.

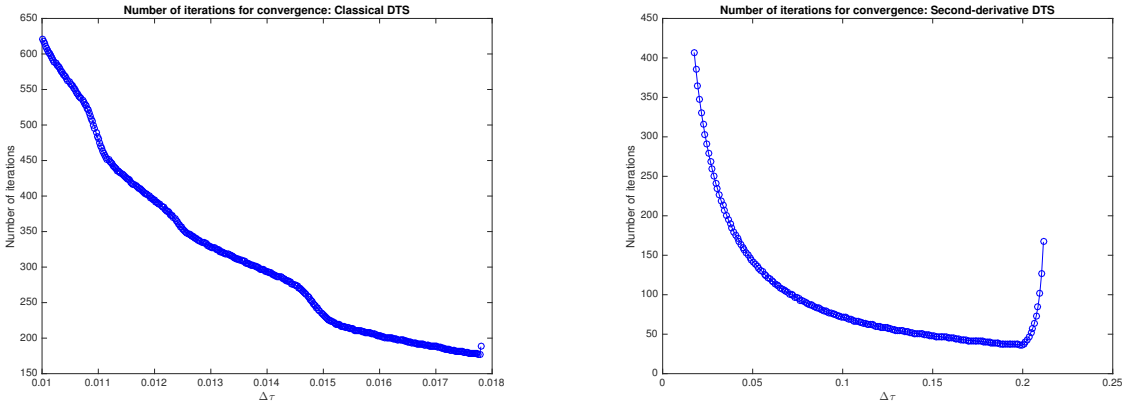



Figure 3: Number of iterations for the classical and new dual–time marching techniques. The new DTS is less stiff than the former formulation and a larger dual time–step can be chosen.

where $\mathbf{u}\left(x,t\right)=\left[u_1\left(x,t\right),u_2\left(x,t\right)\right]^T$, $\varepsilon=10^{-2}$. The matrices $A$ and $B$ are real and given by

$$A=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\qquad B=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}$$

while $\mathbf{F}\left(x,t\right)$, $\mathbf{f}\left(x\right)$, $g_0\left(t\right)$, $g_1\left(t\right)$ are given data.

The specific boundary conditions $\left(u_1+\sqrt{2}u_2-\varepsilon u_{2,x}\right)\left(0,t\right)=g_0\left(t\right)$ and $\left(u_1-\sqrt{2}u_2-\varepsilon u_{2,x}\right)\left(1,t\right)=g_1\left(t\right)$ applied to the linearized Navier–Stokes like system (28) makes the problem strongly well–posed, i.e. a unique solution to (28) exists and its norm is bounded by the boundary and initial data. Moreover, the corresponding semi–discrete problem in space is strongly stable, if the SBP–SAT approach is used. These theoretical results are shown in the Appendix.

Here we limit ourselves to the study of the fully–discrete problem

$$\frac{3\mathbf{v}^{n+1}-4\mathbf{v}^{n}+\mathbf{v}^{n-1}}{2\Delta t}+D\otimes A\mathbf{v}^{n+1}=\varepsilon D_2\otimes B\mathbf{v}^{n+1}+\widetilde{\mathbf{F}}^{n+1}+\mathbf{SAT},\tag{29}$$

with $\mathbf{v}^0=\widetilde{\mathbf{f}}$. The formulation (29) is obtained from (28) by discretizing in space with SBP–SAT and the 2nd order Backward Difference Formula (BDF2) in time. This two–step method

requires $\mathbf{v}^1$ as initial data, which is recovered using the same space discretization and Euler backward in time.

We consider a grid with $x_j = jh$, $j = 0, \ldots, N$ where $h = 1/N$ is the grid spacing, and the grid functions $\widetilde{\mathbf{f}}$, $\widetilde{\mathbf{F}}^n \in \mathbb{R}^{2(N+1)}$ which approximate $\mathbf{f}, \mathbf{F}(t^n)$ in the continuous problem (28). With each grid point we associate the approximate solution $\mathbf{v} \in \mathbb{R}^{2(N+1)}$, such that

$$v_{2j}^n \cong u_1(x_j, t^n), \; v_{2j+1}^n \cong u_2(x_j, t^n), \quad j = 0, \ldots, N.$$

In the fully–discrete problem (29), the symbol $\otimes$ denotes the Kronecker product defined by

$$A = \{a_{ij}\} \in \mathbb{R}^{m\times n}, \;\; B \in \mathbb{R}^{n\times p}, \qquad A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{m\times p}.$$

Moreover, $D$ and $D_2$ are SBP operators for the first and second derivatives and the vector $\mathbf{SAT}$ collects the penalty terms for the boundary conditions. The $\mathbf{SAT}$ term in (29) can be written as

$$\begin{aligned} \mathbf{SAT} = &- \left(P^{-1}E_0 \otimes \Sigma\right)\left[(I_{N+1} \otimes H_0)\,\mathbf{v}^{n+1} - \varepsilon\,(I_{N+1} \otimes G)\,D\mathbf{v}^{n+1} - \widetilde{\mathbf{g}}_0^{n+1}\right] \\ &+ \left(P^{-1}E_N \otimes \Sigma\right)\left[(I_{N+1} \otimes H_N)\,\mathbf{v}^{n+1} - \varepsilon\,(I_{N+1} \otimes G)\,D\mathbf{v}^{n+1} - \widetilde{\mathbf{g}}_N^{n+1}\right], \end{aligned} \tag{30}$$

where $E_0 = \text{diag}(1, 0, \ldots, 0)$, $E_N = \text{diag}(0, \ldots, 0, 1)$ and $I_M$ indicates the $M \times M$ identity matrix. Furthermore, we have used

$$\Sigma = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad H_0 = \begin{bmatrix} 1 & \sqrt{2} \\ 1 & \sqrt{2} \end{bmatrix}, \quad H_N = \begin{bmatrix} 1 & -\sqrt{2} \\ 1 & -\sqrt{2} \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \tag{31}$$

and $\widetilde{\mathbf{g}}_0^n = [g_0(t^n), g_0(t^n), 0, \ldots, 0]^T$, $\widetilde{\mathbf{g}}_N^n = [0, \ldots, 0, g_1(t^n), g_1(t^n)]^T$ are $2(N+1)$ vectors.

To solve the discrete problem (29) we can write the classical (4) and the new DTS formulation (23) by defining

$$\begin{aligned} F = &\frac{3}{2\Delta t} I_{2(N+1)} + D \otimes A - \varepsilon D_2 \otimes B \\ &+ \left(P^{-1}E_0 \otimes \Sigma\right)\left[(I_{N+1} \otimes H_0) - \varepsilon\,(I_{N+1} \otimes G)\,D\right] \\ &- \left(P^{-1}E_N \otimes \Sigma\right)\left[(I_{N+1} \otimes H_N) - \varepsilon\,(I_{N+1} \otimes G)\,D\right] \end{aligned} \tag{32}$$

and

$$\mathbf{R} = \frac{2\mathbf{v}^n}{\Delta t} - \frac{\mathbf{v}^{n-1}}{2\Delta t} + \left(P^{-1}E_0 \otimes \Sigma_0\right)\widetilde{\mathbf{g}}_0^{n+1} - \left(P^{-1}E_N \otimes \Sigma\right)\widetilde{\mathbf{g}}_N^{n+1} + \widetilde{\mathbf{F}}^{n+1}. \tag{33}$$

To obtain the computational results we have used the following manufactured solutions

$$u_1(x,t) = \cos(10\pi x - t), \qquad u_2(x,t) = \sin(10\pi x - t),$$

with a spatial increment $h = 0.01$ and a physical time–step $\Delta t = 0.1$. By choosing the fourth–order Runge–Kutta scheme as pseudo time–integrator, the optimal choice of dual time–step for the classical DTS (4) is $\Delta\tau = 0.002178$, see Figure 4.

With the stopping criterion $\|\mathbf{w}^n - \mathbf{u}\|_P < 10^{-6}$ this formulation reaches steady–state in 421 iterations. The new DTS (23) is less stiff than the classical time–marching technique (4), see Figure 5. Figure 6 shows the number of iterations needed for each dual–time step $\Delta\tau$. The optimal choice for the two–derivatives DTS is $\Delta\tau = 0.0722$ and it leads to convergence in 60 inner iterations. This implies that the new DTS is approximately seven times more efficient than the classical one.

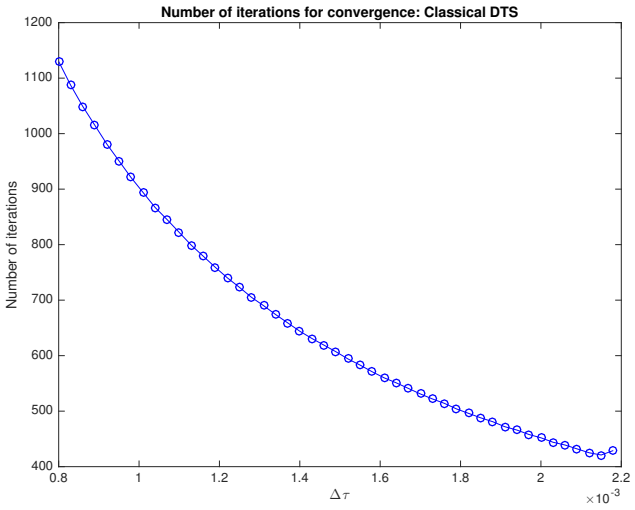



Figure 4: Number of iterations for convergence using the classical DTS.

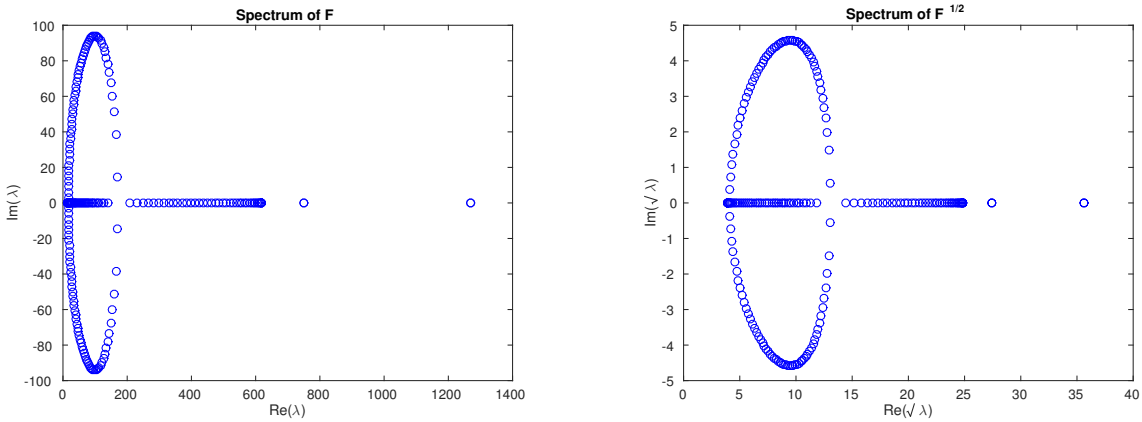



Figure 5: The spectrum of $F$ and of its square root for the linearized Navier–Stokes equations.

## 5 MAIN DRAWBACKS AND OPEN QUESTIONS

The previous numerical tests show that the new DTS formulation (23) has better convergence properties compared to the conventional time–marching technique (4). However, when we rewrite (23) in first–order form as in (15) we obtain a system which has twice as the dimensions of the one in (4). Moreover, the computation of the principal square root of $F$ may be excessively expensive to compute if the dimension of the system (7) is large. In Table 1, the computational times of both DTS techniques (4) and (23) are shown for the numerical experiment in Section 4.1. The last column provides the elapsed time for computing $F^{\frac{1}{2}}$.

Suppose that the square root of $F$ is given. Then from Table 1 we conclude that when the number of nodes increases, the second–derivative DTS (23) provides better results with respect to the classical technique (4). However, the computation of the square root becomes expensive. Therefore, we are interested in suboptimal formulations of (13) which do not involve fractional or negative powers of $F$.

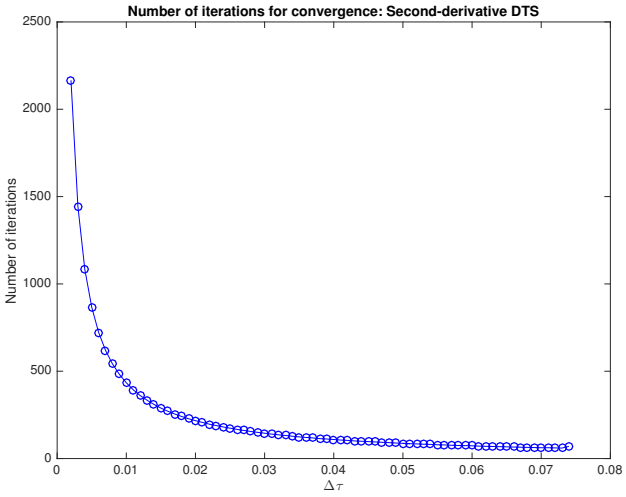



Figure 6: Number of iterations for convergence using the two–derivatives DTS.

| Nodes ($N$) | Classical DTS (sec) | Second–derivative DTS (sec) | Square root (sec) |
|---|---|---|---|
| 100 | 0.045337 | 0.107452 | 0.044975 |
| 1000 | 2.119714 | 1.820615 | 1.767909 |
| 2000 | 12.703043 | 8.927953 | 15.609013 |

Table 1: Execution times of the DTS schemes with optimal smoothing step for the hyperbolic steady problem (25). The elapsed time for the second–derivative DTS is indicated without the computation of the square root of $F$. The matrix $F^{\frac{1}{2}}$ is computed with the optimized routine **sqrtm** presented in [17].

### 5.1 Alternative formulations

Our goal is to provide provably convergent DTS schemes of the form (13), but avoid having to compute $F^{\frac{1}{2}}$. This system of second order differential equations can be written as a first–derivative formulation

$$\mathbf{z}_\tau + A\mathbf{z} = \mathbf{b}, \quad \text{where} \quad \mathbf{z} = \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -I \\ F & 2C \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}.$$

Let $K = K(F)$ be a polynomial in $F$. By choosing $C = (K^{-1}F + K)/2$, we can rotate the system as

$$\begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix}_\tau + \begin{bmatrix} I & 0 \\ K^{-1}F & I \end{bmatrix}^{-1} \begin{bmatrix} K^{-1}F & -I \\ 0 & K \end{bmatrix} \begin{bmatrix} I & 0 \\ K^{-1}F & I \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{w}_\tau \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{R} \end{bmatrix}. \tag{34}$$

Note that the optimal formulation with $C = K = F^{\frac{1}{2}}$ is represented in (34) and leads to (24).

There are two obvious alternatives for $K$. The first one is $K = \kappa I$ with $\kappa > 0$. This choice gives rise to a convergent formulation with a decay determined by $\kappa$ and $\lambda/\kappa$, where $\lambda$ is any eigenvalue of $F$. If $\kappa$ is big, the damping of the system is dominated by the scaled eigenvalues $\lambda/\kappa$. However, we would have the same behavior as that of a preconditioned classical DTS (9), with $\Gamma = \kappa I$. For small values of $\kappa$, every mode of the solution to (34) converges to steady–state uniformly. The second choice is $K = F$, which leads to the same damping as the classical DTS (4) if all the eigenvalues have real part less than one. Otherwise, the convergence

is dominated by the spurious eigenvalues 1. Therefore, these two choices do not lead to an improved formulation with respect to the classical DTS (4).

All other alternatives for $K$ that we have investigated lead to a matrix $C$ which involves inverse matrices or fractional powers of $F$. For this reason, we conclude that the choice $C = (K^{-1}F + K)/2$ in (13) leads to either inefficient or expensive DTS schemes. The existence of alternative formulations not affected by these two effects is still matter of research.

## 6 CONCLUSIONS

A new two–derivative dual–time stepping technique has been proposed. The new DTS technique has been analyzed and optimized theoretically. The formulation involves a parameter in front of the first derivative in dual time which can be chosen to obtain the highest possible decay rate.

We have compared the performances of the new formulation with the ones of the classical DTS. Our technique improves the decay rate with respect to the classical time–marching technique if the eigenvalues of the operator representing the system are near the imaginary axis. Furthermore, if the spectrum is not contained within the unitary circle, the new second–derivatives technique provides a system of equations which is less stiff than the DTS formulation.

Numerical computations for a first–order ordinary differential equation and a system modelling the linearized Navier–Stokes corroborate the theoretical results. The simulations reveal that the new formulation is more efficient than the standard one as the size of the problem increase, provided that the required matrix $F^{\frac{1}{2}}$ is given. However, if the computation of $F^{\frac{1}{2}}$ is required, the new DTS formulation is less efficient than the classical dual time–stepping technique.

## APPENDIX

## A SBP–SAT SPACE DISCRETIZATION

For the discretization in space of the differential problems we have used the Summation–By–Parts (SBP) operators in conjunction with the Simultaneous–Approximation–Terms (SAT) for the boundary treatment. The main feature of the first is to mimic the property of integration by parts, whereas the second are penalty–like terms that enforces the boundary conditions weakly.

**Definition A.1.** *$D = P^{-1}Q$ is a first–derivative SBP operator if $P$ is a symmetric positive definite matrix and $Q + Q^T = H = diag\,(-1, 0, \ldots, 0, 1)$.*

These operators can be built also for the second derivative [18].

**Definition A.2.** *$D_2 = P^{-1}\left(-S^T M + H\right) S$ is a second derivative SBP operator if $M$ is positive semidefinite and $S$ approximates the first derivative operator at the boundaries.*

As an example, choosing $S = P^{-1}Q$ in Definition A.2 leads to the so called wide version of $D_2$, i.e. $D_2 = D^2$. Both first– and second–derivative SBP operators can be built for even orders $2p$ at the interior, while at the boundary closure their accuracy is $p$. For further details on the costruction of the SBP operators for the first derivative with $p \leq 4$, see [19].

The SBP finite difference operators with a strong treatment of the boundary conditions only admits stability proofs for very simple problems. This result has been shown in [20], where SAT were proposed to enhance the SBP schemes. Discretizing a well–posed Initial–Boundary–Value–Problem (IBVP) in space with both SBP operators and the SAT penalty terms (SBP–SAT approach), it is possible to prove that the corresponding semidiscrete problem is stable. Further

theoretical details on the SBP–SAT discretization, well–posedness of an IBVP or the stability of its discretization are given in [15].

## B STABILITY OF THE FIRST NUMERICAL TEST

In this section we verify the stability of the classical DTS (4) applied to the discretized problem (26).

For each fixed $\tau > 0$ the dual–time marching technique can be rewritten as

$$\mathbf{w}_\tau = \mathbf{f} - P^{-1}\left(w_0 - g\right)\mathbf{e}_0 - P^{-1}Q\mathbf{w}.$$

The $P$–norm of the solution $\mathbf{w}$ is $\|\mathbf{w}\|_P = \sqrt{\mathbf{w}^T P \mathbf{w}}$. Thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\|\mathbf{w}\|_P^2 &= \mathbf{w}_\tau^T P\mathbf{w} + \mathbf{w}^T P\mathbf{w}_\tau = -\mathbf{w}^T\left(Q + Q^T\right)\mathbf{w} - 2\left(w_0 - g\right)w_0 \\ &= -w_N^2 - \left(w_0 - g\right)^2 + g^2 \leq g^2,\end{aligned}$$

where $w_0$ and $w_N$ are approximations for the solution at the boundaries. Since $g$ is a given data, the P–norm of the solution $\mathbf{w}$ is bounded in time. This implies strong stability of the classical DTS applied to the discretization (26). Equivalently, we have proven that $F$ in (27) has only eigenvalues with non–negative real part, since the energy of the solution to (23) is bounded for any $\tau > 0$.

## C WELL–POSEDNESS OF THE COMPRESSIBLE NAVIER–STOKES

Consider the model of the compressible Navier–Stokes equations (28). In Section 4.2 we claimed that the characteristic boundary conditions make the problem strongly well–posed in the Hadamard sense. To prove this statement we show that (28) admits a unique solution and that the norm of this solution is bounded by the given data $\mathbf{F}\left(x,t\right)$, $\mathbf{f}\left(x\right)$, $g_0\left(t\right)$ and $g_1\left(t\right)$.

We start by deriving the characteristic boundary conditions in (28). By premultiplying with $\mathbf{u}^T$ and integrating over $[0,1]$ we find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\mathbf{u}\left(\cdot,t\right)\right\|^2 + 2\varepsilon\int_0^1 \mathbf{u}_x^T B\mathbf{u}\mathrm{d}x &= \mathbf{u}^T\left(A\mathbf{u} - 2\varepsilon B\mathbf{u}_x\right)\left(0,t\right) \\ &\quad - \mathbf{u}^T\left(A\mathbf{u} - 2\varepsilon B\mathbf{u}_x\right)\left(1,t\right) + \int_0^1 \mathbf{u}^T\mathbf{F}\mathrm{d}x. \qquad (35)\end{aligned}$$

Furthermore, the boundary terms in (35) can be written as

$$\mathbf{u}^T\left(A\mathbf{u} - 2\varepsilon B\mathbf{u}_x\right) = \frac{1}{2\sqrt{2}}\left(u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x}\right)^2 - \frac{1}{2\sqrt{2}}\left(u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x}\right)^2.$$

and therefore the boundary conditions

$$\begin{aligned}\left(u_1 + \sqrt{2}u_2 - \varepsilon u_{2,x}\right)\left(0,t\right) &= g_0\left(t\right), \quad t > 0, \\ \left(u_1 - \sqrt{2}u_2 - \varepsilon u_{2,x}\right)\left(1,t\right) &= g_1\left(t\right), \quad t > 0\end{aligned} \qquad (36)$$

bound the boundary terms in (35). To prove the boundedness of the solution to (28) we use the Cauchy–Schwarz and Young inequalities with a constant $\eta > 0$ for the integral with the forcing term $\mathbf{F}$ in (35). Moreover, since the matrix $B$ is positive semidefinite,

$$\left\|\mathbf{u}\left(\cdot,T\right)\right\|^2 \leq e^{\eta T}\left\{\|\mathbf{f}\|^2 + \int_0^T e^{-\eta t}\left[\frac{1}{2\sqrt{2}}\left(g_0\left(t\right)^2 + g_1\left(t\right)^2\right) + \frac{1}{\eta}\left\|\mathbf{F}\left(\cdot,t\right)\right\|^2\right]\mathrm{d}t\right\} \qquad (37)$$

which proves that the solution to (28) is bounded. The estimate (37) together with the fact that we use the minimum number of boundary conditions implies both existence and uniqueness, and consequently well–posedness.

## D STABILITY OF THE COMPRESSIBLE NAVIER–STOKES

In this section we will prove that the discrete energy of the solution to

$$\mathbf{w}_t + D \otimes A\mathbf{w} = \varepsilon D_2 \otimes B\mathbf{w} + \widetilde{\mathbf{F}}(\tau) + \mathbf{SAT}, \quad t > 0, \tag{38}$$

is bounded. Without loss of generality we consider the homogeneous problem, i.e. $\widetilde{\mathbf{F}} = \widetilde{\mathbf{g}}_0 = \widetilde{\mathbf{g}}_N = 0$ and $D_2 = D^2$. Let $\mathbb{P} = P \otimes I_2$ and $\|\mathbf{w}\|_{\mathbb{P}} = \sqrt{\mathbf{w}^T \mathbb{P} \mathbf{w}}$. Thus

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{w}\|_{\mathbb{P}}^2 = 2\mathbf{w}^T\mathbb{P}\mathbf{SAT} - \mathbf{w}^T\left(\left(Q+Q^T\right)\otimes A\right)\mathbf{w} \\ + \varepsilon\left(D\mathbf{w}\right)^T\left(Q^T\otimes B\right)\mathbf{w} + \varepsilon\mathbf{w}^T\left(Q\otimes B\right)D\mathbf{w}.\end{aligned} \tag{39}$$

Making use of the facts that $Q + Q^T = E_N - E_0$ and $B$ is symmetric, we may write

$$\mathbf{w}^T\left(Q\otimes B\right)D\mathbf{w} = \mathbf{w}_N^T B\left(D\mathbf{w}\right)_N - \mathbf{w}_0^T B\left(D\mathbf{w}\right)_0 - \left(D\mathbf{w}\right)^T\left(P\otimes B\right)D\mathbf{w}, \tag{40}$$

$$\left(D\mathbf{w}\right)^T\left(Q^T\otimes B\right)\mathbf{w} = \mathbf{w}_N^T B\left(D\mathbf{w}\right)_N - \varepsilon\mathbf{w}_0^T B\left(D\mathbf{w}\right)_0 - \left(D\mathbf{w}\right)^T\left(P\otimes B\right)D\mathbf{w}, \tag{41}$$

where $\mathbf{w}_0$, $\mathbf{w}_N$, $\left(D\mathbf{w}\right)_0$, $\left(D\mathbf{w}\right)_N \in \mathbb{R}^2$ are numerical approximations of the solution and its derivative to the continuous problem (28) at the boundaries, respectively. By combining (39), (40), (41) and the expression of the vector $\mathbf{SAT}$ in (30), we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{w}\|_{\mathbb{P}}^2 = -2\varepsilon\left(D\mathbf{w}\right)^T\left(P\otimes B\right)D\mathbf{w} \\ + \mathbf{w}_0^T\left[\left(A-2\Sigma H_0\right)\mathbf{w}_0 - 2\varepsilon\left(B-\Sigma G\right)\left(D\mathbf{w}\right)_0\right] \\ - \mathbf{w}_N^T\left[\left(A-2\Sigma H_N\right)\mathbf{w}_N - 2\varepsilon\left(B-\Sigma G\right)\left(D\mathbf{w}\right)_N\right].\end{aligned} \tag{42}$$

The right hand–side of (42) can be seen as summation of three terms: the first one is non–positive, since $P \otimes B$ is positive semidefinite. The last two contributions are boundary terms, which can be expressed as

$$\begin{bmatrix}\mathbf{w}_i \\ \varepsilon\left(D\mathbf{w}\right)_i\end{bmatrix}^T\begin{bmatrix}A-2\Sigma H_i & -\left(B-\Sigma G\right) \\ -\left(B-\Sigma G\right) & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_i \\ \varepsilon\left(D\mathbf{w}\right)_i\end{bmatrix} = \mathbf{y}_i^T C_i \mathbf{y}_i, \quad i = \{0, N\}. \tag{43}$$

Inserting the expression of $H_0$, $H_N$, $G$ and $\Sigma$ in (31) into (43) leads to

$$C_0 = \begin{bmatrix}0 & 1 & 0 & 0 \\ -1 & -2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0\end{bmatrix}, \qquad C_N = \begin{bmatrix}0 & 1 & 0 & 0 \\ -1 & 2\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0\end{bmatrix}.$$

Since $C_0$ is negative semidefinite and $C_N$ positive semidefinite, the estimate (42) implies that the energy of the solution decreseas and proves the stability of the problem (38).

## REFERENCES


[1] A. Jameson, *Time dependent calculations using multigrid, with applications to unsteady flows past airfoils and wings*, AIAA paper 91-1596, AIAA 10th Computational Fluid Dynamics Conference, Honolulu, Hawaii, June 1991.

[2] A. Belov, L. Martinelli, A. Jameson, *A New Implicit Algorithm with Multigrid for Unsteady Incompressible Flow Calculations*, AIAA paper 95-0049, AIAA 33rd Aerospace Sciences Meeting, Reno, Nevada, January 1995.

[3] D. A. Knoll, P. R. McHugh, *Enhanced Nonlinear Iterative Techniques Applied to a Nonequilibrium Plasma Flow*, SIAM J. Scientific Computing 19, N. 1, 291-301, 1998.

[4] J. A. Housman, M. F. Barad, C. C. Kiris, *Space–Time Accuracy Assessment of CFD Simulations for the Launch Environment*, 29th AIAA Applied Aerodynamics Conference, 2011.

[5] T. Grasser, S. Selberherr, *Mixed–Mode Device Simulation*, Microelectronics Journal 31, pp.873–881, 2000.

[6] J. M. Hsu., A. Jameson, *An Implicit–Explicit Hybrid Scheme for Calculating Complex Unsteady Flows*, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, 2002-0714, January 2002.

[7] A. Arnone, M.-S. Liou, L.A. Povinelli, *Multigrid Time–Accurate Integration of Navier–Stokes Equations*, AIAA-93-3361, 1993.

[8] S. A. Pandya, S. Venkateswaran, T. H. Pulliam, *Implementation of Preconditioned Dual–Time Procedures in OVERFLOW*, AIAA Paper 2003-0072, 2003.

[9] B. T. Helenbrook, G. W. Cowles, *Preconditioning for dual-time-stepping simulations of the shallow water equations including Coriolis and bed friction effects*, Journal of Computational Physics, Volume 227, Issue 9, Pages 4425–4440, April 2008.

[10] C. Marongiu et al., *An Improvement of The Dual Time Stepping Technique For Unsteady RANS computations*, conference paper, European Conference for Aerospace Sciences (EUCASS), Moscow, 2005.

[11] D. W. Peaceman, H. H Rachford, Jr., *The Numerical Solution of Parabolic and Elliptic Differential Equations*, Journal of the Society for Industrial and Applied Mathematics, Vol. 3, No. 1, pp. 28–41, 1955.

[12] R. P. Dwight, *Time–Accurate Navier–Stokes Calculations with Approximately Factored Implicit Schemes*, Computational Fluid Dynamics 2004, Prooceedings of the third international conference on Computational Fluid Dynamics, ICCFD3, Toronto, 12-16 July 2004, Pages 211–217, 2006.

[13] S. Yoon, A. Jameson, *An LU–SSOR Scheme for the Euler and Navier–Stokes Equations*, AIAA Journal, 26, pp. 1025–1026, 1988.

[14] R. L. Bevan, P. Nithiarasu, *Accelerating incompressible flow calculations using a quasi–implicit scheme: local and dual time stepping approaches*, Computational Mechanics, 50(6), December 2012.

[15] M. Svärd, J. Nordström, *Review of Summation-By-Parts Schemes for Initial-Boundary-Value Problems*, Journal of Computational Physics, Volume 268, pp. 17-38, 2014.

[16] J. Nordström, T. Lundquist, *Summation–by–parts in time*, Journal of Computational Physics, Volume 251, pp. 487–499, 2013.

[17] N. J. Higham, *A new **sqrtm** for MATLAB*, Numerical Analysis Report No. 336, Manchester Centre for Computational Mathematics, Manchester, England, 1999.

[18] K. Mattson, J. Nordström, *Summation by Parts operators for finite difference approximations of second derivatives*, Journal of Computational Physics, Volume 199, pp. 503-540, 2004.

[19] B. Strand, *Summation by Parts for Finite Difference Approximations for d/dx*, Journal of Computational Physics, Volume 110, pp. 47–67, 1994.

[20] M. H. Carpenter, D. Gottlieb, S. Abarbanel, *Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes*, Journal of Computational Physics, Volume 111 No. 2, pp. 220-236, 1994.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A DIRECT SOLVER FOR THE ADVECTION-DIFFUSION EQUATION USING GREEN'S FUNCTIONS AND LOW-RANK APPROXIMATION

**Jonathan R. Bull**[1], **Sverker Holmgren**[1] **and Stefan Engblom**[1]

[1]Division of Scientific Computing
Department of Information Technology, Uppsala University
Polacksbacken, Uppsala, Sweden
e-mail: {jonathan.bull,sverker.holmgren,stefan.engblom}@it.uu.se

**Keywords:** advection diffusion, Green's function, semi-implicit, low-rank approximation

**Abstract.** *A new direct solution method for the advection-diffusion equation is presented. By employing a semi-implicit time discretisation, the equation is rewritten as a heat equation with source terms. The solution is obtained by discretely approximating the integral convolution of the associated Green's function with advective source terms. The heat equation has an exponentially decaying Green's function, allowing for a reduction of effort via low-rank matrix approximation. Simple low-rank approximations of the Green's function matrix are investigated as a precursor to using the Fast Multipole Method in higher dimensions. Results show that fast, stable and accurate computations can be achieved by this method. Low-rank approximation saves computational time at the expense of some accuracy. The new method is a template for developing fast, scalable preconditioners for advection-dominated problems including the unsteady Navier-Stokes equations.*

## 1 Introduction

One can express the solution to a PDE as an integral convolution of the associated Green's function with forcing and boundary terms. Upon discretising, the solution is the product of a discrete Green's function matrix with a vector right hand side; in general the matrix is full-rank. However, if the Green's function has a decaying kernel then the matrix can be well-represented by a low-rank approximation. This concept is sometimes used to precondition elliptic and parabolic equations arising from in a variety of applications [1, 2]. However, it has not been applied to problems featuring advection until now. The presence of advective terms in a PDE operator introduces directionality and non-locality: the effect of conditions far upstream is felt strongly at a particular location. Therefore, increasing separation does not imply decreasing influence, and the degree to which the matrix rank can be reduced is limited.

This paper presents a new direct solver for the advection-diffusion equation based on discrete Green's functions and low-rank matrix approximation. The problem of non-locality is dealt with by using a semi-implicit time discretisation to remove the advective term from the discrete system matrix. Specifically, the advective part is treated explicitly in time and the diffusive part implicitly in time, thereby transforming the equation into a steady heat equation with source terms at each timestep. The associated Green's function has the required decaying property, allowing the use of low-rank approximations.

Discretisation of the integral convolution can be achieved by a number of methods. In this paper, two simple midpoint-rule approximations are employed and their stability and accuracy assessed. Computational cost is then reduced by using several straightforward low-rank approximations of the discrete Green's function matrix. A series of numerical experiments in one dimension (1D) verify the analysis and demonstrate proof of concept. The method described in this paper is a template for more advanced strategies, in particular the use of the Fast Multipole Method (FMM) to compute optimal low-rank approximations in 2D and 3D bounded domains.

FMM is a hierarchical algorithm for computing matrix-vector products involving $N$ unknowns in $\mathcal{O}(N \log N)$ or even $\mathcal{O}(N)$ arithmetic operations to a given error tolerance [3]. It is commonly applied to N-body problems in astrophysics and molecular dynamics, and is beginning to be applied to solving elliptic PDEs arising in many areas of computational physics as well [4, 5, 6]. Recently, it has been applied as a preconditioner for elliptic PDEs in a technique known as Inverse FMM (IFMM) [7, 8]. In all existing applications, the governing physics are such that the mutual influence of one spatial location on another decays with separation distance. FMM exploits this quality by hierarchically compressing interactions between well-separated points into lower-dimensional spaces, reducing the number of interactions to be computed. The analogy with the multigrid (MG) method is strong: both are hierarchical and act mainly on low-frequency error components. It is envisioned that FMM could be competitive with MG for preconditioning the system of equations arising from unsteady 3D Navier-Stokes problems.

## 2 Method

### 2.1 Discretisation of advection-diffusion equation

We consider the advection-diffusion equation in one dimension (1D) on a finite domain $x \in V : [0, L]$ with boundary $S$

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0, \qquad x \in V, \tag{1a}$$

$$u(x, 0) = u_0(x), \qquad x \in V, \tag{1b}$$

$$u(x, t) = g(x, t), \qquad x \in S, \tag{1c}$$

where $u$ is a smooth scalar field, $a$ is the (positive) advection coefficient and $\nu$ is the diffusion coefficient. Dirichlet, Neumann and Robin boundary conditions can be specified for the problem and the domain $L$ may be periodic.

We subdivide the domain with $N$ uniform intervals of size $\Delta x = L/N$ and choose timestep $\Delta t$. Now we write (1) in semi-discrete form using matrix notation:

$$\frac{d\mathbf{u}}{dt} + A\mathbf{u} - D\mathbf{u} = \mathbf{b}, \tag{2}$$

where $\mathbf{u}$ is the solution vector, $A$ and $D$ are the discrete advection and diffusion operators and $\mathbf{b}$ is the vector containing boundary terms. We now discretise in time using a semi-implicit splitting scheme for the advective and diffusive terms:

$$\frac{d\mathbf{u}}{dt} - D\mathbf{u}^{n+1} \;=\; \mathbf{b} - A\mathbf{u}^n. \tag{3}$$

For simplicity, let $\mathbf{b} = 0$ and impose periodic boundary conditions. The spatial derivatives are discretised by finite differences with an upwind scheme for advection:

$$u_i^{n+1} - \frac{\nu\Delta t}{\Delta x^2}(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}) = u_i^n - \frac{a\Delta t}{\Delta x}(u_i^n - u_{i-1}^n). \tag{4}$$

The advective and diffusive CFL numbers are denoted $c_A = \frac{a\Delta t}{\Delta x}$ and $c_D = \frac{\nu\Delta t}{\Delta x^2}$ respectively. This is the simplest, first-order semi-implicit method for time-dependent PDEs; higher-order methods could also be used for greater accuracy [9].

### 2.2 Green's function

The semi-implicit discrete equation (3) is in the form of an initial-boundary value problem governed by the 1D heat equation:

$$\left(\frac{\partial}{\partial t} - \nu\frac{\partial^2}{\partial x^2}\right) u(x, t) = f(x, t), \qquad x \in V, \tag{5a}$$

$$u(x, 0) = u_0(x), \qquad x \in V, \tag{5b}$$

$$u(x, t) = g(x, t), \qquad x \in S, \tag{5c}$$

where in our case the source term $f(x, t) = -a\frac{\partial u^n}{\partial x}$ and $u^n$ denotes the solution at the previous timestep. The solution of (5) is expressed in terms of a Green's function $G(x - x', t - t')$ as the

sum of three integrals accounting for source terms, boundary and initial conditions [10]:

$$\begin{aligned} u(x,t) &= \int_0^T \int_V G(x-x',t-t')f(x',t')dx'dt' \\ &\quad + \int_V G(x-x',t-0)u_0(x')dx' \\ &\quad - \int_0^T \int_S \frac{\partial G}{\partial n} g(x',t')dS(x')dt', \quad x \in \mathbb{R}, t > t'. \end{aligned} \tag{6}$$

The Green's function satisfies

$$\left(\frac{\partial}{\partial t} - \nu \frac{\partial^2}{\partial x^2}\right) G(x-x',t-t') = \delta(x-x',t-t'), \tag{7}$$

where $\delta(x-x',t-t')$ is the Dirac delta function. The expression for Green's function in one dimension is [10]:

$$G(x-x',t-t') = (4\pi\nu(t-t'))^{-1/2} e^{-\frac{(|x-x'|)^2}{4\nu(t-t')}}. \tag{8}$$

Hereafter, only periodic (free-space) problems are considered so the third term is dropped. Let us treat the $n+1$th timestep as an initial-boundary value problem from time $t$ to $t+\Delta t$ with initial conditions $u(x,t) = u^n(x)$. The solution $u(t+\Delta t) = u^{n+1}(x)$ is given by

$$\begin{aligned} u^{n+1}(x) &= \int_t^{t+\Delta t} \int_V G(x-x',t-t')f(x',t')dx'dt' \\ &\quad + \int_V G(x-x',t+\Delta t-t)u^n(x')dx'. \end{aligned} \tag{9}$$

Taking a backwards-in-time approximation of the first integral, we obtain

$$\begin{aligned} u^{n+1}(x) &\approx \Delta t \int_V G(x-x',t+\Delta t-t)f(x',t)dx' \\ &\quad + \int_V G(x-x',t+\Delta t-t)u^n(x')dx' \\ &= \int_V (4\pi\nu\Delta t)^{-1/2} e^{-\frac{(|x-x'|)^2}{4\nu\Delta t}} [u^n + \Delta t f(x',t)]dx'. \end{aligned} \tag{10}$$

Note that the forcing term $u^n + \Delta t f(x',t)$ is identical to the right-hand side of the semi-implicit finite difference relation (4). The Green's function can be expressed in terms of a scaled wavenumber $\epsilon = (4\nu\Delta t)^{-1/2}$ and radial distance $r = |x-x'|$ as

$$G(r,\epsilon) = \frac{\epsilon}{\sqrt{\pi}} e^{-\epsilon^2 r^2}. \tag{11}$$

### 2.3 Integral approximation

The continuous integral (10) is approximated as a sum of integrations over each interval using the midpoint rule[1]. The mesh nodes are the points $x_i = \{0, \Delta x, \ldots, L-\Delta x, L\}$. Intervals can

[1]Caveat: the exponential function is not smooth at $x = x'$. In approximations that sample this point, the error may be dominated by this value.

be defined with the nodes either as their midpoints (*node-centred* scheme) or endpoints (*face-centred* scheme). Figure 1 (a) illustrates the node-centred scheme. In the case of non-uniform mesh spacing and in higher dimensions, it can be generalised as integration on a dual mesh. Using the node-centred scheme the solution at a node $x_i$ is:

$$\begin{aligned} u_i^{n+1} &= M_{ij}(u_j^n + \Delta t f_j) && (12)\\ &= \sum_{j=1}^{N} \frac{\epsilon\Delta x}{\sqrt{\pi}} e^{-(|j-i|\epsilon\Delta x)^2}[(1-c_A)u_j^n + c_A u_{j-1}^n]. && (13) \end{aligned}$$

The $[N \times N]$ matrix $M$ with entries $m_{ij} = \frac{\epsilon\Delta x}{\sqrt{\pi}} e^{-(|j-i|\epsilon\Delta x)^2}$ is symmetric positive-definite, diagonally dominant and circulant. Periodicity is imposed on $j$ such that if $j - i > N/2$ then $j = j - N$ and similarly, if $j - i < N/2$ then $j = j + N$.

The face-centred scheme is illustrated in Figure 1 (b). Using this scheme, the solution at a node $x_i$ is defined as:

$$\begin{aligned} u_i^{n+1} &= M_{ij}^F(u_j^n + \Delta t f_j) && (14)\\ &= \sum_{j=1}^{N} \frac{\epsilon\Delta x}{\sqrt{\pi}} e^{-((|j-i|+1/2)\epsilon\Delta x)^2}[(1-c_A)u_j^n + c_A u_{j-1}^n]. && (15) \end{aligned}$$

The salient difference between these two methods is that the latter does not include the peak of the Green's function. Although not critical for our Green's function, similar PDEs have Green's functions with singularities. In those cases, the face-centred scheme is preferable. Using either scheme, the error $\eta_i$ at a point $x_i$ is the sum of the errors in each interval:

$$\eta_i = \left| \int_L G(x,x')dx' - \sum_{j=1}^{N} \Delta x G(x_j, x_i) \right| \leq \sum_{j=1}^{N} \frac{\Delta x^3}{24} \left\| \frac{d^2 G}{dx^2} \right\|_{L_\infty(j)}. \quad (16)$$

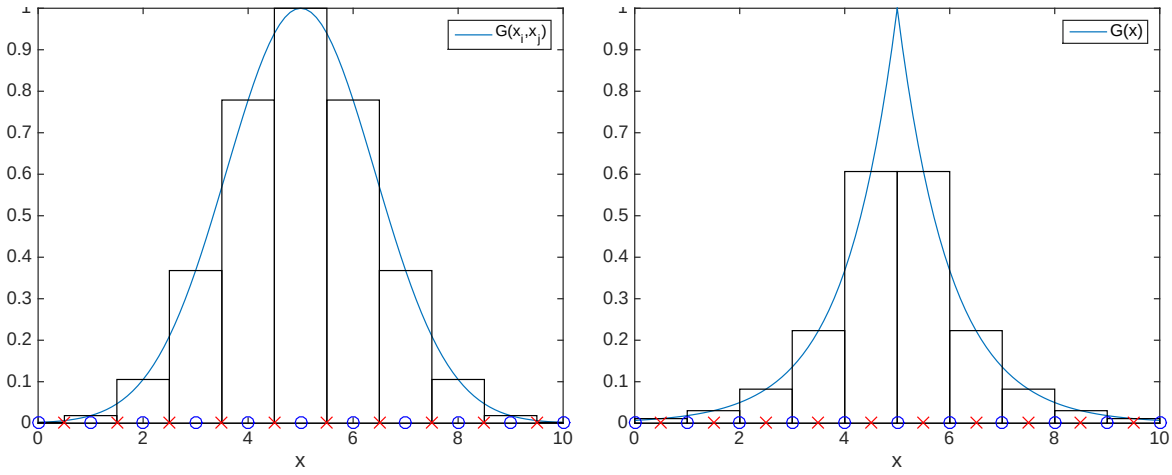
With the node-centred scheme the error is

$$\eta_i \leq \sum_{j=1}^{N} \frac{\Delta x^3}{24} \left\| -\frac{\epsilon}{2\sqrt{\pi}} e^{-(|j-i|\epsilon\Delta x)^2}(2(|j-i|\epsilon\Delta x)^2 - 1) \right\|_{L_\infty(j)} \leq \frac{L\epsilon\Delta x^2}{48\sqrt{\pi}}. \quad (17)$$

Therefore second-order convergence of the solution is expected. The face-centred scheme has the error

$$\begin{aligned} \eta_i &\leq \sum_{j=1}^{N} \frac{\Delta x^3}{24} \left\| -\frac{\epsilon}{2\sqrt{\pi}} e^{-(|j-i+1/2|\epsilon\Delta x)^2}(2(|j-i+1/2|\epsilon\Delta x)^2 - 1) \right\|_{L_\infty(j)}\\ &\leq \frac{L\epsilon\Delta x^2}{48\sqrt{\pi}} e^{-\epsilon\Delta x/2}(1-(\epsilon\Delta x)^2)\\ &= \frac{L\epsilon\Delta x^2}{48\sqrt{\pi}}(1-\epsilon\Delta x/2 + \mathcal{O}(\Delta x^2))(1-(\epsilon\Delta x)^2)\\ &= \frac{L\epsilon\Delta x^2}{48\sqrt{\pi}}, && (18) \end{aligned}$$

where $e^{-\epsilon\Delta x/2}$ has been expanded as a Taylor series about $\Delta x = 0$. Keeping the leading-order term only, the face-centred error (18) reduces to the same expression as for the node-centred scheme (17). Higher-order approximations may also be employed for the integral. Note that this is not the only source of error: the spatial and temporal discretisations also contribute.




(a) Integration on dual mesh (node-centred) (b) Integration on primary mesh (face-centred)

Figure 1: Approximations of Green's function with $N = 10$ centred at 5th interval. Primary mesh nodes shown in blue and midpoints in red.

## 2.4 Stability analysis

### 2.4.1 Node-Centred Scheme

Consider a uniform initial flow field $u(x) = C$: the solution is constant throughout time. Therefore a necessary, but not sufficient, condition for stability of the integral equation (12) (and also (14)) is that the row sum of matrix $M$ (resp. $M^F$) must not be greater than one. For the sake of accuracy the row sum must be as close to one from below as possible. Periodic boundaries and uniform $\Delta x$ are assumed. Thanks to the circulant property of $M$, the sum is identical for all rows $l$. The node-centred matrix row sum stability constraint is written:

$$\begin{aligned} \frac{\epsilon\Delta x}{\sqrt{\pi}} \sum_{j=1}^{N} e^{-(|j-l|\epsilon\Delta x)^2} &\leq 1 \\ \frac{\epsilon\Delta x}{\sqrt{\pi}} \left[ 1 + 2 \sum_{j=1}^{(N-1)/2} e^{-(j\epsilon\Delta x)^2} \right] &\leq 1. \end{aligned} \tag{19}$$

Now $\epsilon\Delta x = c_D^{-1/2}$ is substituted into (19) to obtain a function $f(c_D)$:

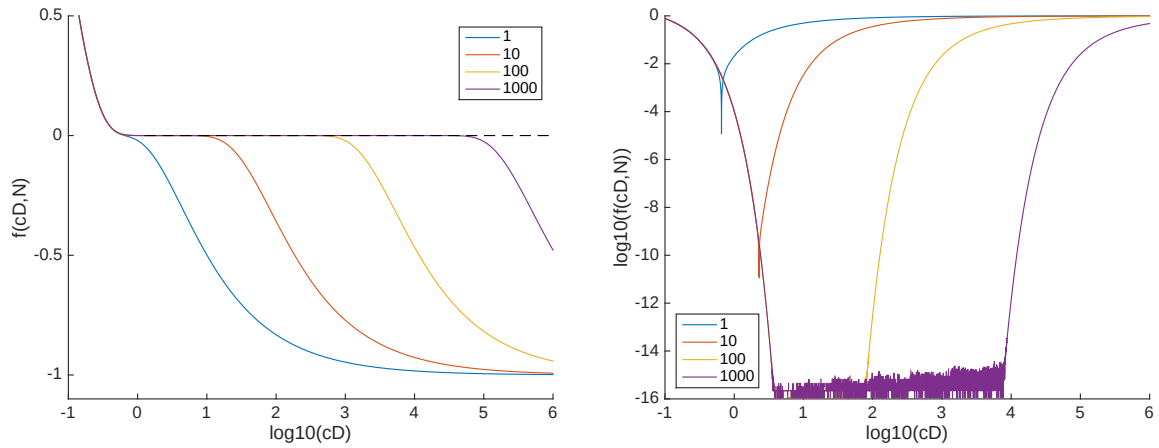
$$f(c_D) = \frac{1}{\sqrt{\pi}} c_D^{-1/2} \left[ 1 + 2 \underbrace{\sum_{j=1}^{(N-1)/2} (e^{1/c_D})^{-j^2}}_{S_N} \right] - 1 \leq 0. \tag{20}$$

The sum $S_N$ is convergent for $e^{1/c_D} > 1$, i.e. $c_D > 0$. However, no expression can be found for the limit of the sum as $N \to \infty$. Numerical calculations with large $N$ provide the following information:

1. as $c_D \to 0, S_N \to 0$,
2. as $c_D \to \infty, S_N \to N$,

3. at $c_D = 1, S_N = 0.3863$.

The stability function (20) is plotted (20) in Figure 2 on (a) semi-log and (b) log-log scales (real part) for several values of $N$. The function lies below the $c_D$ axis above a certain value, i.e. there is a *minimum* bound on $c_D$ for the inequality to be satisfied. The values of $c_D$ at which the function crosses the axis (spikes on the log-log plot) increases with $N$. In the simplest case of $N = 1$, the stability condition is $c_D \geq 0.66$. At $N = 10$, $c_D > 2.28$; at $N = 100$, $c_D > 3.73$, and at $N = 1000$, $c_D > 3.73$. The reason for the limited stability of the node-centred scheme is the presence of the central term at $j = i$ in (19). Accuracy also depends on $c_D$: there is a range of $c_D$ in which $f(c_D) \approx 0$ and $f(c_D) \leq 0$ with the zero-crossing defining the lower bound. The extent of the range increases with $N$. At larger $N$, the row sum equals one to machine precision across several orders of magnitude in $c_D$.




(a) Linear axes

(b) Re$(f(c_D, N))$ Log axes

Figure 2: Stability condition on node-centred scheme as a function of $c_D$ for $N = \{1, 10, 100, 1000\}$.

### 2.4.2 Face-Centred Scheme

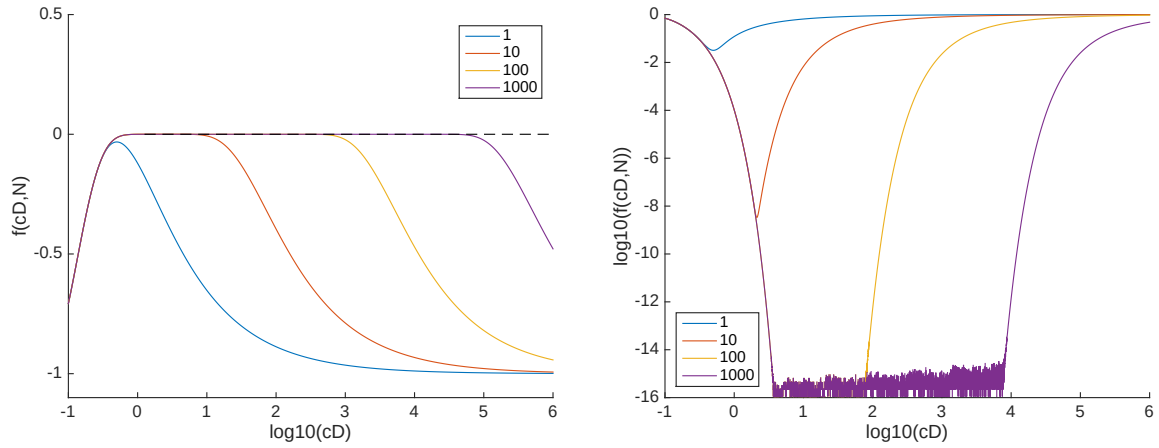
The matrix row sum using the face-centred scheme is:

$$\begin{aligned} \frac{\epsilon\Delta x}{\sqrt{\pi}} \sum_{j=1}^{N} e^{-((|j-l|+1/2)\epsilon\Delta x)^2} &\leq 1 \\ 2\frac{\epsilon\Delta x}{\sqrt{\pi}} \sum_{j=1}^{N/2} e^{-((j-1/2)\epsilon\Delta x)^2} &\leq 1. \end{aligned} \tag{21}$$

Again, we substitute $\epsilon\Delta x = c_D^{-1/2}$ into (21) to obtain:

$$f(c_D) = \frac{2}{\sqrt{\pi}} c_D^{-1/2} \sum_{j=1}^{N/2} (e^{1/c_D})^{-(j-1/2)^2} - 1 \leq 0. \tag{22}$$

The stability function is plotted in Figure 3. Now, the inequality is satisfied for all $c_D$. Nevertheless, the accuracy is still contingent upon $c_D$. As with the node-centred scheme, it is clear from Figure 3 (a) that the face-centred scheme is accurate over a finite region that increases rapidly with $N$. At $N = 1000$, machine-precision accuracy is achieved from $c_D = 3.75$ to $c_D \approx 10000$.




(a) Semi-log axes (b) Log-log axes

Figure 3: Stability condition on face-centred scheme as a function of $c_D$ for $N = \{1, 10, 100, 1000\}$.

## 2.5 von Neumann analysis

The above conditions are necessary but not sufficient for stability of the node- and face-centred schemes. Von Neumann's method is used to derive a second condition which, together with the first condition, is sufficient for stability of the method. We analyse the node-centred scheme first. The error term at time level $n+1$ and location $l$ is related to the error at the previous time level by:

$$\begin{aligned} e^{a(t+\Delta t)}e^{ik_m x} &= \frac{\epsilon\Delta x}{\sqrt{\pi}}\Bigg[[(1-c_A)e^{at}e^{ik_m x} + c_A e^{at}e^{ik_m(x-\Delta x)}] \\ &+ \sum_{j=1}^{(N-1)/2} e^{-(j\epsilon\Delta x)^2}[(1-c_A)e^{at}e^{ik_m(x-j\Delta x)} + c_A e^{at}e^{ik_m(x-(j-1)\Delta x)}] \\ &+ \sum_{j=1}^{(N-1)/2} e^{-(j\epsilon\Delta x)^2}[(1-c_A)e^{at}e^{ik_m(x+j\Delta x)} + c_A e^{at}e^{ik_m(x+(j-1)\Delta x)}]\Bigg], \end{aligned} \tag{23}$$

where the first term is the contribution from point $j = l$ and the next two terms are the contributions from points to the left and right of $l$ respectively. Upon dividing through by $e^{at}e^{ik_m x}$ and substituting the relations $e^{-(j+1)ik_m\Delta x} = e^{-jik_m\Delta x}e^{-ik_m\Delta x}$ and $e^{(j-1)ik_m\Delta x} = e^{jik_m\Delta x}e^{-ik_m\Delta x}$, we obtain

$$\begin{aligned} e^{a\Delta t} &= \frac{\epsilon\Delta x}{\sqrt{\pi}}\left[(1-c_A+c_A e^{-ik_m\Delta x})\left(1+\sum_{j=1}^{(N-1)/2} e^{-(j\epsilon\Delta x)^2}\left(e^{-jik_m\Delta x}+e^{jik_m\Delta x}\right)\right)\right] \\ &= (1-c_A+c_A e^{-ik_m\Delta x})\underbrace{\frac{\epsilon\Delta x}{\sqrt{\pi}}\left(1+2\sum_{j=1}^{(N-1)/2} e^{-(j\epsilon\Delta x)^2}\cos(jk_m\Delta x)\right)}_{M}. \end{aligned} \tag{24}$$

For stability it is required that $|e^{a\Delta t}| \leq 1$. For the most restrictive condition on $c_A$ term $M$ must be maximised. Using $\max(|\cos(jk_m\Delta x)|) = 1$, $M$ is reduced to the matrix row sum (19). Choosing $c_D$ such that the stability function in (20) equals zero to a sufficient level of precision,

we have $M \approx 1$. Thus (24) reduces to

$$\left|1 - c_A + c_A e^{-ik_m \Delta x}\right| \leq 1. \tag{25}$$

This represents a circle of radius $c_A$ and centre $1 - c_A$. Finally, we have the stability condition:

$$2c_A - 1 \leq 1 \quad \therefore c_A \leq 1. \tag{26}$$

This condition is typical of the upwind finite difference discretisation of the advection term and does not depend on the attributes of the Green's function method. An identical condition can also be derived for the face-centred scheme.

### 2.6 Low-rank approximation

In the two matrix methods defined above, entries far from the diagonal make a relatively small contribution to the solution so low-rank approximations to $M$ or $M^F$ can be justifiably used to reduce computational effort. The primary method of interest here is the fast multipole method (FMM), although this paper treats only the 1D case for which FMM is not applicable. After a generic description of FMM, two simple methods of constructing a low-rank matrix are presented. This work is a precursor step to using FMM for the 2D and 3D advection-diffusion equations.

**Fast multipole method:** The computational degrees of freedom are treated as a cloud of points (we do not specify the number of spatial dimensions $d$ to keep the description generic). A $k$-level tree mesh is constructed in which level one consists of one box containing all points and the $k$th level consists of $k^{2d}$ boxes. Each box at level $k$ contains a small number of points, say $p$ on average. A multipole expansion (polynomial series) of degree $m$ is employed at the box centroid to describe the equivalent source due to all point sources in the box.

The interactions between any two boxes on the same level are classified as strong or weak based on their radii, on the separation between their centroids and on a threshold value $\theta$. Strong interactions are computed directly between points. Weak interactions are computed by 'shifting' a multipole expansion from one box to the other. At levels 1 to $k-1$, a multipole expansion in a box is computed from multipole expansions inside its children. In this manner, only local interactions at level $k$ are computed directly and all other interactions are computed approximately via multipole expansions and shifts. This saves considerable effort especially for very large $N$.

In an abstract sense FMM is a method for hierarchical sparsification of a dense matrix. Strong interactions are clustered around the diagonal. Away from the diagonal, zeros are introduced by the multipole representation. The degree of sparsification depends on the user-defined parameters $p$, $m$ and $\theta$. Accuracy and computational cost can be fine-tuned via these parameters.

**Threshold value:** A minimum value is specified, below which the matrix entry is set to zero. The sparse matrix $M^{L1}$ is constructed according to

$$m_{ij}^{L1} = \begin{cases} m_{ij}, & m_{ij} > 1.0/N, \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

Likewise, the very sparse matrix $M^{L2}$ is constructed with the threshold value of $1.0/\sqrt{N}$.

**Specified bandwidth:** Let $P$ be the low-rank matrix bandwidth; in this case $P = \text{round}(N/8)$. Then the entries of sparse matrix $M^{L3}$ are given by

$$m_{ij}^{L3} = \begin{cases} m_{ij}, & i - P \leq j \leq i + P, \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

with the same periodic indexing as the full-rank matrix.

### 2.7 Comparison to iterative method

To compute a reference solution, an iterative method is employed based on the same implicit-explicit flux splitting:

$$Au_i^{n+1} = \left(1 - \nu\Delta t \frac{d^2}{dx^2}\right) u_i^{n+1} = (1 - c_A)u_i^n + c_A u_{i-1}^n, \tag{29}$$

where the matrix $A$ is

$$A = \begin{bmatrix} 1 + 2c_D & -c_D & \dots & -c_D \\ -c_D & 1 + 2c_D & -c_D & \\ & \ddots & \ddots & \ddots \\ & -c_D & 1 + 2c_D & -c_D \\ -c_D & \dots & -c_D & 1 + 2c_D \end{bmatrix}. \tag{30}$$

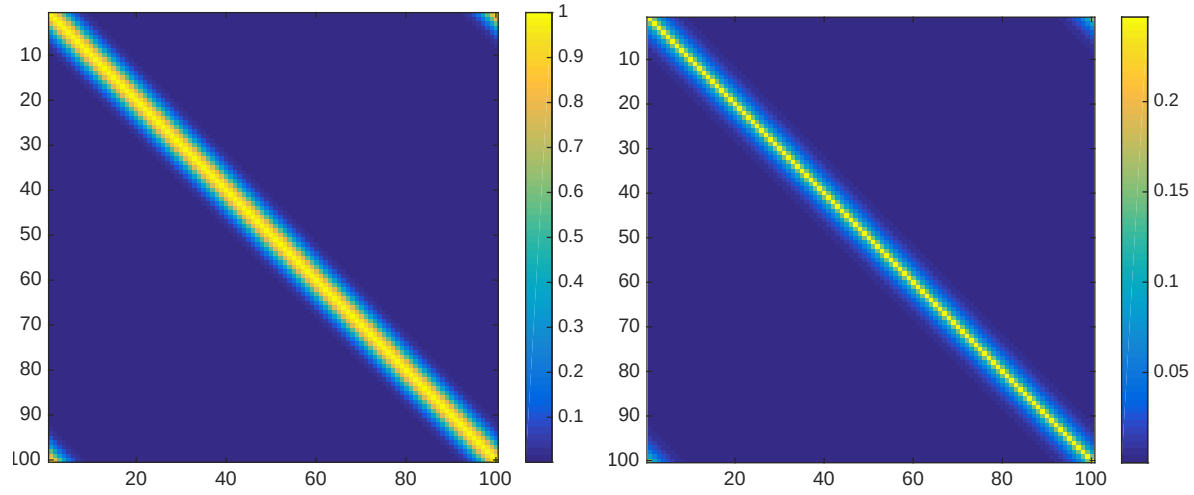
The Matlab `linsolve` function was used with `opts.SYM=true,opts.POSDEF=true` and all other options set to false. It is not a very efficient method but serves to provide a reference solution.

## 3 Numerical tests

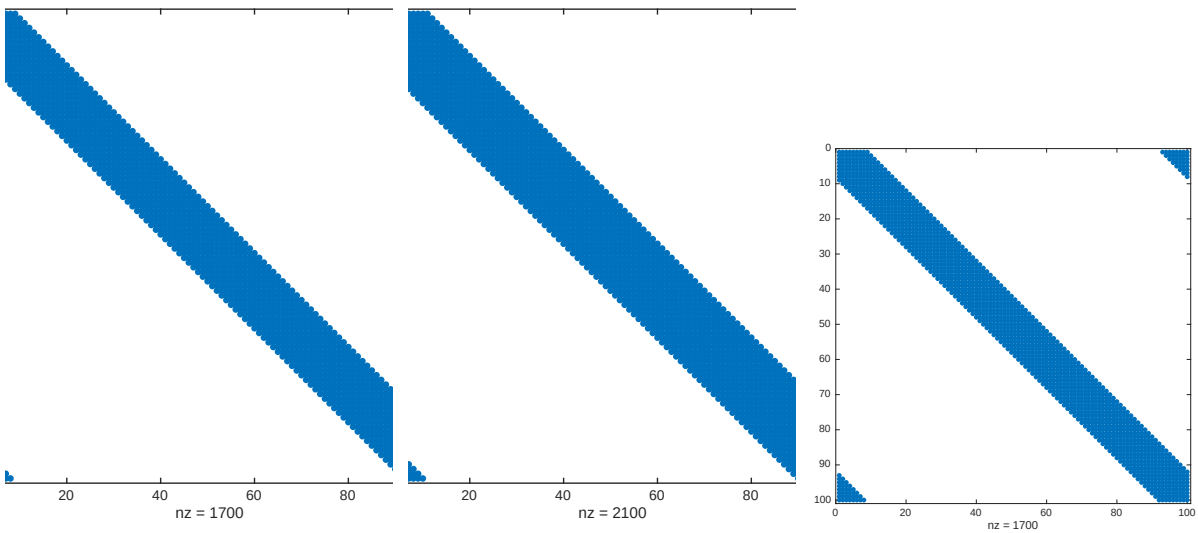
The full-rank and sparse matrix-vector products and the iterative method were implemented in Matlab to solve the 1D advection-diffusion equation with periodic boundary conditions. The advective and diffusive constants were set to $a = 1.0$ and $\nu = 0.1$ and simulations were run for 100 timesteps. Discrete optimisation of the diffusive CFL number was performed for the node- and face-centred schemes. The smallest values of $c_D$ that minimised the functions in Figures 2 and 3 were found for each value of $N$. These in turn dictated the timesteps, which were sufficiently small that the advective CFL condition (26) was also satisfied. Table 1 lists the optimised timesteps $\Delta t$ for each $N$ and each scheme. The values scale approximately with $N^{-2}$ and the node-centred scheme has an optimal timestep slightly larger than that of the face-centred scheme.

Figure 4 (a) shows the full-rank node-centred matrix $M$ coloured by magnitude. For comparison, Figure 4 (b) shows the inverse of matrix $A$ in (30). Figures (c,d,e) show the non-zeros of sparse matrices $M^{L1}$, $M^{L2}$ and $M^{L3}$. Table 2 shows the row sum and number of nonzeros in each matrix.

Figure 5 plots the solutions starting from (a) sinusoidal and (b) Gaussian initial conditions with $N$=100. Node-centred solutions are compared to the reference solution computed at nodes, and the face-centred scheme to the reference solution computed at interval midpoints. The $M^{L2}$ solution displays a large amplitude (dissipative) error, as expected from the fact that it has a row sum well below one. All the other node-centred solutions overlie the reference solution. The face-centred solution has a phase lag (dispersion error).




(a) $M$ (b) $A^{-1}$




(c) $M^{L1}$ (d) $M^{L2}$ (e) $M^{L3}$

Figure 4: Colour plot of (a) full-rank matrix, (b) inverse iterative matrix, (c,d,e) non-zeros of sparse matrices.

| $N$ | $\Delta t$ (node-centred) | $\Delta t$ (face-centred) |
|---|---|---|
| 100 | $3.85e-3$ | $3.75e-3$ |
| 200 | $9.64e-4$ | $9.37e-4$ |
| 400 | $2.41e-4$ | $2.34e-4$ |
| 800 | $6.02e-5$ | $5.86e-5$ |
| 1600 | $1.51e-5$ | $1.46e-5$ |
| 3200 | $3.76e-6$ | $3.66e-6$ |

Table 1: Optimal timesteps.

| matrix | row sum | NNZ |
|---|---|---|
| $A^{-1}$ | 1.00000 | 10000 |
| $M$ | 1.00000 | 10000 |
| $M^F$ | 1.00000 | 10000 |
| $M^{L1}$ | 0.99791 | 1700 |
| $M^{L2}$ | 0.95359 | 1100 |
| $M^{L3}$ | 0.99999 | 4900 |

Table 2: Matrix row sums and no. of nonzeros, $N$=100.

### 3.1 Error Convergence

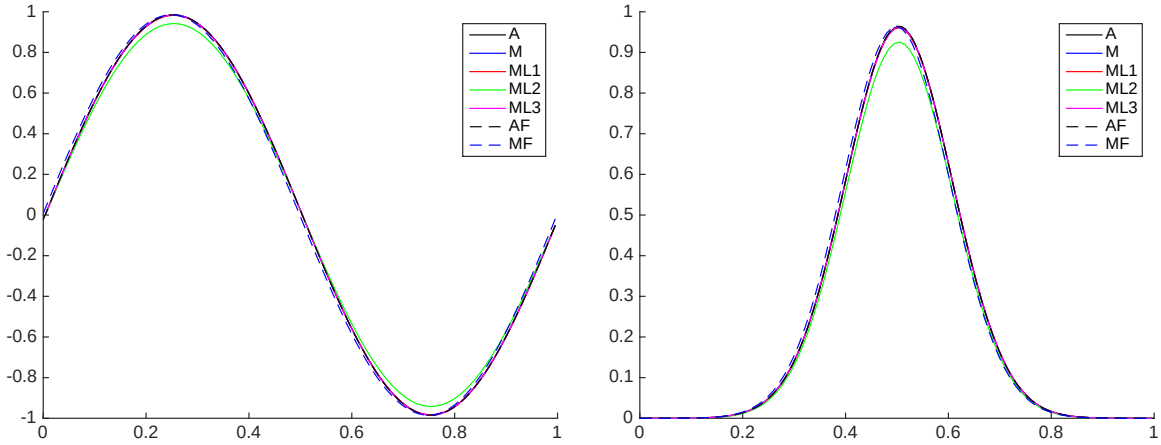
The $L_2$ norm of the solution errors with respect to the reference solutions (at the nodes or faces as appropriate) was calculated after one timestep, successively doubling $N$ from 100 up to 3200. Figure 6 (a) shows the error convergence on a log-log plot, starting from the sinusoidal initial condition. The full-rank node-centred matrix $M$ results in a convergence rate of approximately fifth order and approaches machine precision at large $N$. The $M^{L3}$ error converges at an identical rate and has the same magnitude: this low-rank approximation does not reduce accuracy. The $M^{L1}$ and $M^{L2}$ errors converge at between first and second order. The full-rank face-centred matrix $M^F$ error also converges at between first and second order. Figure 6 (b) shows the errors when starting from the Gaussian initial condition, with similar results to the (a).

### 3.2 CPU time

The time taken to compute one timestep with $N$=100 to $N$=3200 was found. The times to construct $A$, $M$, $M^{L1}$ etc. were not included. Figure 7 shows the CPU times. Similar results are obtained from both initial conditions. The reference solution is much slower than the direct solutions and the time increases faster than $N^2$. Direct solution times increase at about first order. The low-rank matrices constructed using the minimum-value rule, $M^{L1}$ and $M^{L2}$, are by far the fastest at large $N$. Matrix $M^{L3}$ with prescribed rank is as slow as the full-rank matrices at large $N$. Given that the error is as small as the full-rank scheme, this low-rank scheme clearly does not alter the matrix noticeably. Specifying a smaller bandwidth would have a more pronounced effect.

## 4 Conclusions

This paper presents a simple scheme for direct solution of the one-dimensional advection-diffusion equation in a periodic setting. By using a semi-implicit time discretisation the advective term becomes a forcing function and the left-hand side becomes the heat equation.




(a) sinusoidal initial condition (b) Gaussian initial condition

Figure 5: Computed solutions with (a) sinusoidal, (b) Gaussian initial conditions. —: reference (node), – –: reference (face), blue: $M$, red: $M^{L1}$, green: $M^{L2}$, magenta: $M^{L3}$, blue dashes: $M^F$.

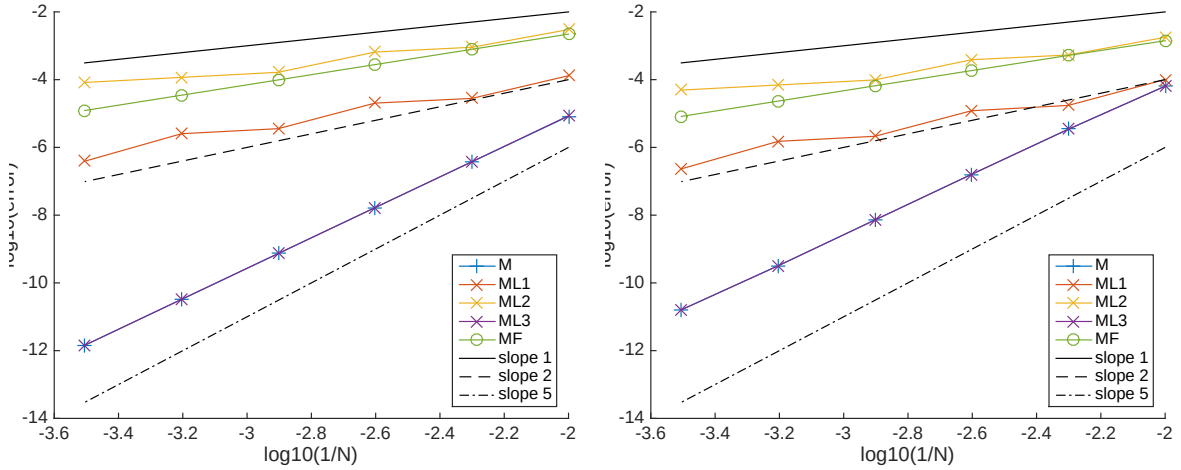
The associated Green's function has a decaying kernel, making it highly suitable for low-rank discrete approximation. Similar methods have been very successful in solving purely elliptic equations, but this is the first known application to a hyperbolic equation.

Several different full- and low-rank discrete approximations of the kernel were analysed and found to have differing stability limits and lead to differing error convergence rates. All obtained approximately linear scaling of CPU time with $N$. The node-centred full-rank scheme was not unconditionally stable but had the best accuracy and the error converged at fifth order. Almost machine-precision accuracy (with respect to the reference iterative solution) was attained by this scheme at $N = 3200$ in a single matrix-vector multiplication. The face-centred scheme, although unconditionally stable, was less accurate and had a much lower convergence rate. Specifically, the numerical wave speed was under-predicted although the amplitude was well-predicted. Low-rank node-centred approximations in which a threshold matrix entry was defined were shown to save considerable computational effort but were less accurate and had a lower error convergence rate. The balance of accuracy and cost may be problem-dependent and more work is needed to determine robust sparsification strategies.

The new method is a template for finding fast approximate solutions of advection-dominated problems. FMM is proposed as the ideal low-rank approximation algorithm for this application (although it is not suitable for 1D problems). FMM allows control over the balance of accuracy and speed, and scales very well to large numbers of parallel processes. In terms of solving PDEs, FMM is somewhat analogous to multigrid, in that low-frequency error components (i.e. distant interactions) are rapidly diminished. Indeed, FMM could become a competitor to multigrid as a solver/preconditioner for hyperbolic PDEs including the Navier-Stokes equations. Future work will extend the low-rank direct solution method to 2D and 3D problems on bounded domains.

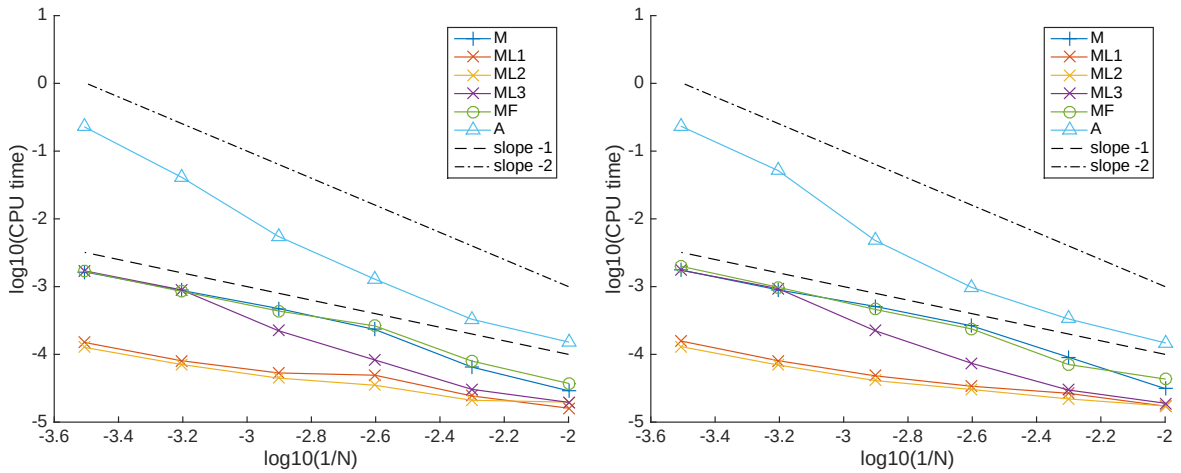
## REFERENCES

[1] M. Bebendorf, Hierarchical lu decomposition-based preconditioners for BEM, Computing 74 (3) (2005) 225–247.




(a) sinusoidal initial condition (b) Gaussian initial condition

Figure 6: Error convergence of direct solutions with respect to reference solution, (a) sinusoidal and (b) Gaussian initial conditions.




(a) sinusoidal initial condition (b) Gaussian initial condition

Figure 7: Time to solution.

[2] H. Branden, P. Sundqvist, An algorithm for computing fundamental solutions of difference operators, Numerical Algorithms 36 (4) (2004) 331–343.

[3] L. Greengard, V. Rokhlin, A fast algorithm for particle simulations, Journal of Computational Physics 73 (1987) 325–348.

[4] F. Ethridge, L. Greengard, A new fast-multipole accelerated Poisson solver in two dimensions, SIAM J. Sci. Comput. 23 (3) (2001) 741–760.

[5] L. Greengard, D. Gueyffier, P.-G. Martinsson, V. Rokhlin, Fast direct solvers for integral equations in complex three-dimensional domains, Acta Numerica 18 (2009) 243–275.

[6] R. Yokota, H. Ibeid, D. Keyes, Fast multipole method as a matrix-free hierarchical low-rank approximation, arXiv:1602.02244 [cs.NA] (2016).

[7] S. Ambikasaran, E. Darve, The inverse fast multipole method, arXiv:1407.1572 (2014).

[8] P. Coulier, H. Pouransari, E. Darve, The inverse fast multipole method: using a fast approximate direct solver as a preconditioner for dense linear systems, arXiv:1508.01835 (2015).

[9] U. Ascher, S. Ruuth, B. Wetton, Implicit-explicit methods for time-dependent partial differential equations, SIAM Journal on Numerical Analysis 32 (3) (1995) 797–823.

[10] P. Olver, introduction to partial differential equations, Springer, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN IMPROVED DISCRETISATION METHOD FOR BOUNDED CONVECTIVE SCHEMES ON UNSTRUCTURED CO-LOCATED GRIDS

**V. PRŽULJ**

Ricardo Software, Ricardo UK Ltd, Shoreham–by–Sea, West Sussex, BN43 5FG, UK
e-mail: vlado.przulj@ricardo.com, web page: www.ricardo.com/en-GB/What-we-do/Software

**Keywords:** CFD, Convection, Bounded Scheme, CBC/TVD Criteria, Finite–Volume, Co–located Grid, Unstructured Grid.

**Abstract.**

*In comparison to structured grids, a more refined discretisation of convective transport on unstructured grids is required to account for their complex topologies. In particular, and with reference to the given cell–face, the far upwind node is not readily available on unstructured grids. Also, the solution accuracy can be affected by numerical diffusion caused by skewness of the grid. In this paper, a novel method is introduced for the far upwind node reconstruction in the framework of the cell–centred finite–volume discretisation. The main idea is to form an upwind–biased stencil with a local axis that connects the central cell node and the geometric cell–face centre. Note that central and downwind cells share the given cell–face. Instead of the actual downwind node, its projection onto the local axis is used. In the previous methods, the local axis has been defined along the line connecting the central and downwind nodes. These methods can not mitigate the grid skewness error whereas by its definition the new method bypasses the issue of grid skewness. The present method is applied to several test cases, covering transport of a scalar step-profile, laminar vortex shedding from a circular cylinder and turbulent flows in two engine intake ports. Depending on the test–case, the high resolution convective schemes such as MINMOD, the modified SMART and bounded central differencing scheme are employed. The latter is selected for large eddy simulations of one of engine ports. The preceding schemes are formulated for non–uniform grids, and they obey the well known convective boundedness criteria. The used grid types also varied, including triangular, polyhedral and Cartesian cut–cell shapes. The convergence properties and accuracy of the new method are scrutinised against the previous one. For all cases, the bounded numerical solution is ensured, and the improved convergence rate of the new method is noted. The accuracy of results predicted by both methods is established through comparisons with either analytical or experimental and numerical results from literature.*

## 1 INTRODUCTION

Discretisation of fluid transport equations should ensure a convergent, accurate and physically bounded numerical solution. Closely related to the solution accuracy is the perplexing problem of boundedness. The difficulty is that proper bounds exist for any numerical problem but these bounds are not always easily perceived. The boundedness problem particularly affects the discretisation of convective transport on either structured or unstructured grids.

Calculation of cell–face variable values using linear interpolation between adjacent cell values, which is known as central differencing scheme (CDS), illustrates well the boundedness problem. If the numerical grid is not sufficiently refined this scheme can not suppress numerical oscillations accompanied by non–physical overshoots or undershoots, especially in regions with steep gradients of transported variables. This behaviour is predicted by Godunov's theorem [1] which states that the second and higher order linear schemes can not be monotonicity preserving. Only the first order accurate upwind differencing scheme (UDS) guarantees the solution boundedness whereby the cell–face value is taken to be that which prevails at the upwind cell. In order to maintain the bounded solution of the second and higher order convection schemes, these schemes should be designed as *non-linear* schemes, with provision of an appropriate amount of numerical diffusion.

The way in which the numerical diffusion is introduced, and its amount controlled, determines the accuracy, convergence properties and computational efficiency of the convective (numerical) scheme. A popular approach is to control the numerical diffusion by *boundedness criteria*. Two major criteria, total variation diminishing (TVD) [2, 3] and the convection boundedness criterion (CBC) [4], offer great flexibility in the construction and implementation of high resolution bounded schemes. These two criteria were formulated in terms of flux limiters and normalised variables, respectively. The review and comparison of TVD and CBC criteria, together with analysis of most higher–order bounded schemes can be found in [5].

In the framework of the cell-centred, finite-volume discretisation, each cell–face value is calculated with the help of a dimensionally split upwind–biased stencil, i.e. one–dimensional bounded scheme is applied along appropriately defined cell-face direction. On the structured, in general non–uniform and non–orthogonal grids, the upwind–biased stencil is defined along the local coordinates (grid lines) as shown in Figure 2 (left) for the east cell–face. The stencil involves the upwind $U$ and downwind $D$ cells (nodes) with respect to the cell $C$ which is itself upwind of the face $j$. The implementation of bounded schemes on multi–dimensional structured grids is relatively straightforward. Though it is very important to formulate these schemes in terms of non–uniform grid spacing where either normalised local coordinates [6, 7] or equivalent downwind and upwind geometric factors [8, p.48] can be used.

Considering unstructured grids, the implementation of the TVD/CBC schemes is neither straightforward nor unique task as the upwind node is not readily available. A sensible idea is to reconstruct the upwind node and its value. This can be done implicitly by projecting the central cell variable gradient onto the line connecting the central and downwind nodes as devised by [9, 10, 11]. Instead of the central cell gradient, the projection of the cell-face gradient in conjunction with the modified definition of the CBC normalised variable was employed in [12]. Both approaches are equivalent as they deliver the same definition of the normalised variables, and the actual position of the upwind node is not required. It was shown by [10, 11] that reconstructed upwind values should be bounded by maximum and minimum values of central node's neighbours. The effectiveness of upwind node reconstruction using formulation for the linear CBC/TVD schemes consistent with non-uniform grids was demonstrated in study

[11]. Alternative methods rely on the explicit definition of the upwind node, and require a geometric search to find an upwind cell closest to the defined upwind node. Then the upwind node value is the value at the upwind cell corrected with the help of the upwind cell gradient. In [13], the upwind node was defined in the same way as in the gradient projection method. Other authors, [14, 15], positioned the upwind node along the line normal to the cell-face. This line passes through the cell-face centre as well as through the nodes $C'$ and $D'$ which are projections of the $C$ and $D$ nodes onto the face normal, respectively. The recent multi–slope MUSCL method [16] specifies the upwind node along the line connecting the central node and cell-face centre. In essence, the upwind value is obtained by linear interpolation from the values available at the closest cell centres.

The majority of previous upwind node reconstruction methods calculated the bounded cell-face value at the cell–face interpolation point $j'$. This intersection point, defined by vector connecting the central and downwind nodes, did not always coincide with the face geometric centre $j$. Consequently, these methods, apart from [16], could not fully alleviate the grid skewness interpolation error. In the methods employing the line normal to the cell–face [14, 15], the interpolation point was the cell–face centre. However, the variable values and gradients at projected central and downwind nodes were taken to be those at the corresponding actual nodes. The skewness corrected TVD flux limiter, proposed recently by [17], was devised to correct the cell-face value implicitly. The limiter itself must be carefully bounded to satisfy the TVD constraints.

In this paper, a novel method, illustrated in Figure 2 (right), is proposed to mitigate the grid skewness error explicitly. For this, the upwind node $U$ is reconstructed along the cell–face direction defined by line connecting the central node $C$ and the face centre $j$. Instead of the actual downwind node $D$, its projection onto the cell–face direction, $D'$, is used. After an overview of the finite volume discretisation, bounded convective schemes are introduced with the help of normalised variable approach. Next, a new generalised implementation of convective schemes on unstructured grids is described. Then the results from four test cases are used to assess the performance of the proposed method.

## 2 OVERVIEW OF FINITE VOLUME DISCRETISATION

The discretisation of the governing equations is carried out by using the finite volume method to discretise spatial dimensions and finite differencing to discretise time. The solution domain is filled with an unstructured grid defined by edges of arbitrary polyhedra referred to as control volumes or cells. All flow variables are stored at geometric centres of the cells. For a non–moving control volume $V$, depicted in Figure 1, integration of fluid transport equations using second–order single–point quadrature operators leads to control volume balances of unsteady, convective, diffusive and source terms, [18, 19] :

$$\frac{d}{dt}\left(\rho V \phi\right)_P + \sum_{j=1}^{n_f} \left( C_j - D_j - \left(s_{\phi k}^A A_k\right)_j \right) = \left(s_\phi^V\right)_P V_P \,, \tag{1}$$

$$\text{with } C_j = \dot{m}_j \phi_j \,, D_j = \left(\Gamma_\phi \nabla\phi \cdot \vec{A}\right)_j$$

In the preceding equation, $\phi$ is a flow variable and $\Gamma_\phi$ represents its diffusion coefficient; time and density are denoted as $t$ and $\rho$, respectively. For a cell–face $j$, convective and diffusion fluxes are $C_j$ and $D_j$, respectively. The source terms $s_\phi^V$ and $s_\phi^A$ are related to the cell volume and cell face, respectively; $n_f$ is the number of cell-faces enclosing the cell. Mass fluxes $\dot{m}_j$ are

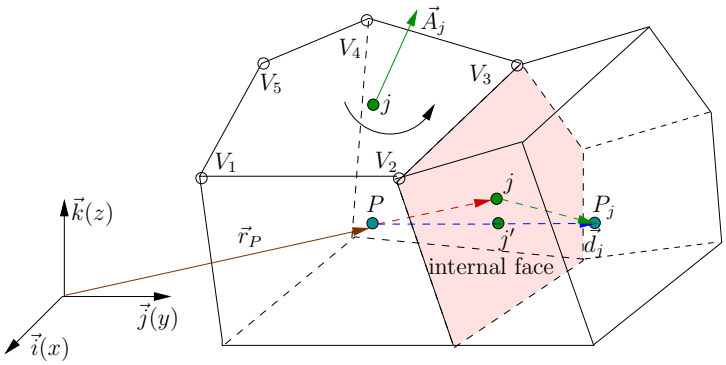



Figure 1: A polyhedral control volume (cell) $P$ as enclosed by its cell–faces $j$. A cell-face $j$, with the outward normal vector $\vec{A}_j$, is defined by a list of vertices $V_k$.

defined as:

$$\dot{m}_j = \rho_j \vec{U}_j^* \cdot \vec{A}_j \tag{2}$$

where the Rhie–Chow interpolation practice [20, 21] is used for the mass conserving face velocity $\vec{U}_j^*$. The mass fluxes must satisfy the integral mass balance over each cell:

$$\frac{d}{dt}\left(\rho V\right)_P + \sum_{j=1}^{n_f} \dot{m}_j = 0 \tag{3}$$

For time–dependent phenomena, the semi-discrete equation (1) can be written in a form that resembles the first order ordinary differential equation. Time 'marching' procedure is then used to advance the solution in time over a number of time steps with the time-step size $\Delta t$. The unsteady term is integrated over each time interval by either first order Euler or second order three time level (TTL) differencing scheme, see [21] . Both schemes are implicit and unconditionally stable.

The evaluation of cell–face values and gradients at both cell-centres and cell–faces is the essential ingredient of the discretisation procedure. Commonly used linear interpolation between the values at adjacent cells $P$ and $P_j$ (which amounts to the central differencing scheme, CDS) gives

$$\phi_j = f_j \phi_P + \left(1 - f_j\right) \phi_{P_j} \tag{4}$$

$$f_j = f_{j,P} = \frac{|\vec{r}_{P_j} - \vec{r}_j|}{|\vec{r}_{P_j} - \vec{r}_j| + |\vec{r}_j - \vec{r}_P|} \tag{5}$$

where $f_j$ is the cell–face interpolation factor. In addition to CDS, higher–order upwind schemes such as linear upwind (LUDS) [22] and quadratic upwind (QUICK) [23] (generalised by [24] as $\kappa$–schemes) are popular. As pure $\kappa$–schemes are not always appropriate for the convective transport, a simple and efficient approach is to blend them with a certain amount of the first order upwind (UDS) scheme as advocated by [21, 25]. In the case of CDS, the blending can be arranged as:

$$\phi_j = \phi_j^{UDS} + \gamma_\phi f_j^* \frac{\dot{m}_j}{|\dot{m}_j|}\left(\phi_{P_j} - \phi_P\right) \tag{6}$$

$$\phi_j^{UDS} = \begin{cases} \phi_{P_j} & \text{if} \quad \dot{m}_j < 0 \\ \phi_P & \text{if} \quad \dot{m}_j \geq 0 \end{cases}, \quad f_j^* = \begin{cases} f_j & \text{if} \quad \dot{m}_j < 0 \\ 1 - f_j & \text{if} \quad \dot{m}_j \geq 0 \end{cases} \tag{7}$$

where $f_j^*$ is now the *flow orientated* interpolation factor. The blending factor $\gamma_\phi$ depends on the grid resolution, and for sufficiently fine grids the values close to one can be used. The alternative to the blending approach is to adopt high resolution bounded schemes whose improved discretisation on the unstructured grids is the subject of this work.

For the cell gradients, a linear least-square procedure [26] is used in this study. Considering the cell–face gradient, the Gauss' formula $\nabla\phi \approx \oint \phi d\vec{A}/V$ can be applied to the control volume constructed around the face $j$. The derivation outlined in [19] leads to the following expression:

$$\nabla\phi_j \approx \overline{\nabla\phi}_j + \frac{\vec{A}_j}{\vec{A}_j \cdot \vec{d}_j}\left[\left(\phi_{P_j} - \phi_P\right) - \overline{\nabla\phi}_j \cdot \vec{d}_j\right] \;,\; \overline{\nabla\phi}_j = f_j\nabla\phi_P + (1-f_j)\nabla\phi_{P_j} \tag{8}$$

Further details regarding the discretisation of convective and diffusion fluxes can be found in [19, 27] .

The outcome of the discretisation of considered transport equation for $\phi$ is a linearised algebraic equation for each cell. The corresponding system of algebraic equations for each flow variable is under–relaxed implicitly [28] . The system is then arranged in a sparse matrix form $[\mathbf{a}]\phi = \mathbf{b}_\phi$, and solved by Bi–CGSTAB linear equation solver [29, 30] using incomplete lower–upper (ILU) pre-conditioners.

The non–linearity and coupling of the solved transport equations, including velocity–pressure coupling, is handled by generalised SIMPLE–based pressure correction method [31]. This method is suitable for both time marching and steady-state iterative solutions. Normalised residuals, defined as

$$R_\phi = \frac{\sum_{cells}|a_P\phi_P - b_\phi|}{\sum_{cells}|a_P\phi_P|} \;, R_{mass} = \frac{\sum_{cells}|\text{mass imbalance}|}{\sum_{cells}|\text{mass imbalance}|_{ref}} \tag{9}$$

serve as the convergence criterion for the SIMPLE–based iterative solution procedure. The reference mass imbalance in the preceding equation is the maximum imbalance in the first five iterations [32] . The solution is considered converged when all normalised residuals drop below a certain value, typically below $1. \times 10^{-5}$.

## 3 BOUNDED CONVECTIVE SCHEMES – NORMALISED VARIABLE APPROACH

High–resolution convective schemes require an upwind–biased stencil as defined along local coordinate $\xi$ in Figure 2 (left) for the structured grid. In this figure, the far upwind $U$, central $C$ and downwind $D$ nodes are defined with respect to the cell-face $j$ having positive mass flow rate $\dot{m}_j$. Obviously, labelling of the nodes depends on the mass flux direction. As shown in [11], the mathematical enforcement of monotonicity along local axis $\xi$ and interpolative boundedness (face value bounded by values at adjacent cells) is an easy way to derive both CBC [4] and TVD [3] boundedness criteria. These criteria, and bounded schemes based on them, can be formulated in terms of normalised variable approach introduced by [4, 33].

### 3.1 Normalised variable approach

In this approach, the convected variable $\phi$ is normalised as

$$\widetilde{\phi} = \frac{\phi - \phi_U}{\phi_D - \phi_U} \tag{10}$$

giving $\widetilde{\phi}_U = 0$ and $\widetilde{\phi}_D = 1$. The associated normalised variable diagrams (NVD), showing TVD (left) and CBC (right) regions, are presented in Figure 3. It follows that cell–face values

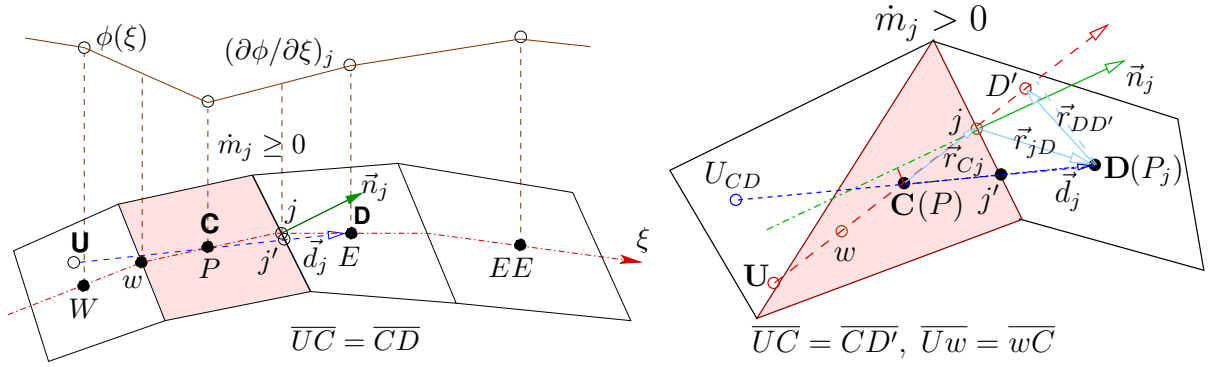



Figure 2: Upwind–biased stencil for structured (left) and unstructured (right) grids.

$\widetilde{\phi}_j$ should lie within the shaded areas (monotonic range $0 < \widetilde{\phi}_C < 1$), and on the line $\widetilde{\phi}_j = \widetilde{\phi}_C$ outside the monotonic range. Further, the convective schemes with linear characteristics such as $\kappa$–schemes (CDS, LUDS and QUICK) may violate boundedness criteria. Therefore bounded schemes have to be non–linear or piecewise linear schemes.

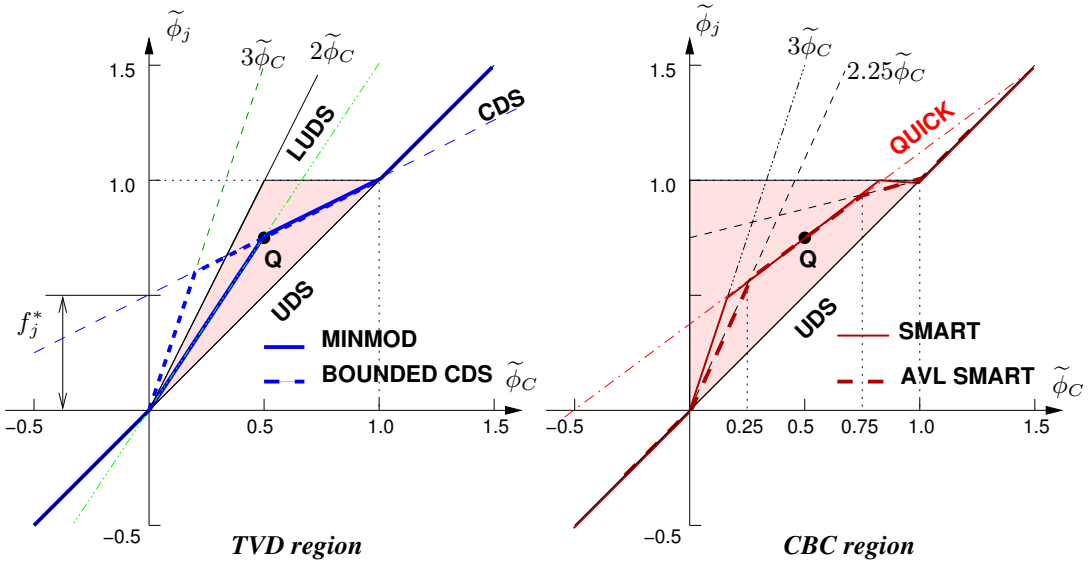



Figure 3: NVD approach. TVD (left) and CBC (right) regions and characteristics of well known convective schemes.

It is instructive to generalise linear schemes such as CDS, LUDS and QUICK on the structured non–uniform grids, Figure 2 (left), with the help of *flow orientated* interpolation factors as done in [8, p.48]:

$$\phi_j = \phi_j^{UDS} + \varphi_j \frac{\dot{m}_j}{|\dot{m}_j|}\left(\phi_{P_j} - \phi_P\right) = \phi_C + \varphi_j\left(\phi_D - \phi_C\right) \tag{11}$$

$$\varphi_j = \left(g_D - \alpha_j \frac{f_j^*}{f_w^*}\right) + \left(g_U + \alpha_j\right)\frac{\phi_C - \phi_U}{\phi_D - \phi_C} \tag{12}$$

$$g_D = \frac{f_j^{*2}\left(1 + f_w^*\right)}{f_j^* + f_w^*},\ g_U = \frac{f_w^{*2}\left(1 - f_j^*\right)}{f_j^* + f_w^*},\ f_w^* = \begin{cases} f_{j,E} & \text{if } \dot{m}_j < 0 \\ 1 - f_{j,W} & \text{if } \dot{m}_j \geq 0 \end{cases} \tag{13}$$

where the parameter $\alpha_j$ defines a family of the higher–order schemes. For example, the QUICK scheme is recovered for $\alpha_j = 0$. The quantity $\varphi_j$ plays a similar role as the flux limiter [3]. In

addition to geometric interpolation factors, the flux limiter is a function of the gradient ratio $r_j$:

$$r_j = \frac{\phi_C - \phi_U}{\phi_D - \phi_C} = \frac{\widetilde{\phi}_C}{1 - \widetilde{\phi}_C} \approx \frac{(\partial\phi/\partial\xi)_w}{(\partial\phi/\partial\xi)_j} \tag{14}$$

It is evident from Equation (12) that all NVD schemes $\widetilde{\phi}_j = f(\widetilde{\phi}_C)$ can be formulated in terms of flux limiters:

$$\widetilde{\phi}_j = \widetilde{\phi}_C + \varphi_j \left(1 - \widetilde{\phi}_C\right) \tag{15}$$

Note that Equation (11) for the cell-face value offers a straightforward implementation of NVD schemes.

### 3.2 Choice of bounded schemes

Characteristics of bounded schemes used in this study are shown in Figure 3. In comparison to the original SMART scheme [4], the AVL SMART [11] offers a good compromise between accuracy and convergence properties. Both schemes coincide with the third–order accurate QUICK scheme over a large part of the monotonic region. The TVD MINMOD scheme (Roe's limiter) is a combination of the second–order accurate LUDS and CDS schemes. It is more diffusive than the AVL SMART/SMART scheme, and therefore has better convergence properties.

With exception of implicit large eddy simulations (LES), the CDS scheme is an ideal choice for LES on the adequately resolved numerical grids, [34]. Unfortunately, achieving the good grid resolution is not always practical. To compensate for inadequate grid resolution, and maintain the numerical stability and boundedness, the bounded CDS (BCDS) scheme has been devised. As shown in Figure 3 (left), this composite scheme has a central differencing behaviour for $0.2 \leq \widetilde{\phi}_C \leq 1$.

The upwind node for bounded schemes on unstructured grids is usually placed at the same distance from the central node as the downwind node. For this situation, the flow orientated interpolation factor at the cell–face $w$ in Equation (13) is the same as at cell-face $j$, $f_w^* = f_j^*$. The corresponding flux limiters are then defined as:

$$\text{MINMOD:} \qquad \varphi_j = f_j^* \max\left[0, \min\left(r_j, 1\right)\right] \tag{16}$$

$$\text{Bounded CDS:} \qquad \varphi_j = f_j^* \max\left[0, \min\left(4r_j, 1\right)\right] \tag{17}$$

$$\text{AVL SMART:} \qquad \varphi_j = f_j^* \max\left\{0, \min\left[\left(1+\beta_1\right) r_j \,,\, \frac{1}{2}\left(\beta_1 + \beta_2 r_j\right) \,,\, 1+\beta_2\right]\right\} \tag{18}$$

$$\beta_1 = 1 + f_j^*, \;\; \beta_2 = 1 - f_j^*$$

## 4 GENERALISED IMPLEMENTATION OF BOUNDED SCHEMES

As outlined in the Introduction, original implementations of unstructured grid schemes employed an upwind–biased stencil whose axis was defined by a line connecting central $C$ and downwind $D$ nodes, Figure 2 (left). The acronym *UCD-method* will be used to refer to this type of upwind node reconstruction. In this case, the centred difference $(\phi_D - \phi_U)$ can be calculated from the gradient projection $\nabla\phi_C \cdot 2\vec{d}_{CD}$, which gives the upwind node value $\phi_U^*$ as:

$$\phi_U^* = \phi_D - \nabla\phi_C \cdot 2\vec{d}_{CD} \tag{19}$$

The preceding gradient projection implies a uniform stencil, that is virtual upwind node $U$ is at the same distance from $C$ as $C$ is from $D$. Also, gradient changes between upwind and downwind nodes are neglected.

The grids shown in Figure 2 are not only non–orthogonal but also skewed as the distance vector, connecting the central and downwind nodes, intersects the cell-face at an interpolation point $j'$ which does not coincide with the face geometric centre $j$. As a consequence, the UCD–method provides the cell–face value $\phi_{j\prime}$ at the interpolation point $j'$. Grid skewness is a common feature of unstructured grids. The associated error, analysed by [35], appears as a diffusion–like error. It can be easily corrected in an explicit manner [35, 36] as follows:

$$\phi_j = \phi_{j\prime} + \nabla\phi_j \cdot (\vec{r}_j - \vec{r}_{j\prime}) \tag{20}$$

The above correction, however, may produce an unbounded value at the cell-face centre.

A new method, illustrated in Figure 2 (right), adopts the line connecting the central cell $C$ and cell-face centre $j$ as a local coordinate for the upwind–biased stencil. This method, referred to as *UCF–method*, bypasses the grid skewness issue. However, a projection $D'$ of the downwind node $D$ onto the stencil axis as well as a variable value for this projected node are required. An orthogonal projection of $D$ appears to be a good choice as it minimises an interpolation error for the projected node value $\phi^*_{D'}$:

$$\phi^*_{D'} = \phi_D + \nabla\phi_D \cdot (\vec{r}_{D'} - \vec{r}_D) \tag{21}$$

Similar to the UCD–method, the upwind node $U$ and projected node $D'$ are equidistant to node $C$. The interpolation factor for the cell-face $j$ is calculated from:

$$f^*_{Cj} = \frac{\overline{Cj}}{\overline{CD'}} = \frac{\overline{Cj}}{\overline{Cj} + \overline{jD'}} = \frac{\vec{r}_{Cj} \cdot \vec{r}_{Cj}}{\vec{r}_{Cj} \cdot \vec{r}_{Cj} + |\vec{r}_{jD} \cdot \vec{r}_{Cj}|} \tag{22}$$

The position vector of the projected node is then:

$$\vec{r}_{D'} = \vec{r}_C + \frac{\vec{r}_{Cj}}{f^*_{Cj}} \tag{23}$$

Clearly, the position of projected node can be controlled by interpolation factor value. One can be tempted to define this value equal to $1/2$, i.e to design the upwind stencil as uniform. Numerical tests, however, revealed that biasing of interpolation factors towards $1/2$ as in:

$$f^*_j = \frac{1}{2}\left(f^*_{Cj} + \frac{1}{2}\right) \tag{24}$$

improves the convergence rate without the negative impact on the accuracy. Thus, the preceding definition of the flow orientated cell-face interpolation factor is used for flux limiters in Equations (16), (17) and (18), whereas the projected node $D'$ is placed according to Equation (23) using $f^*_j$ instead of $f^*_{Cj}$.

Another novelty of the current method is a calculation of the upwind node value using gradient projection at an imaginary cell–face $w$:

$$\phi_C - \phi^*_U = \frac{1}{f^*_j}\nabla\phi_w \cdot \vec{r}_{Cj} \Rightarrow \phi^*_U = \phi_C - \frac{1}{f^*_j}\nabla\phi_w \cdot \vec{r}_{Cj} \tag{25}$$

The required gradient at $w$ is reconstructed from the assumed linear distribution of gradients across the central cell (involving points $w$, $C$ and $j$):

$$\nabla\phi_w = \frac{1}{2f_j^*}\left[\left(1+2f_j^*\right)\nabla\phi_C - \nabla\phi_j\right] \tag{26}$$

where the variable gradient $\nabla\phi_j$ at the cell-face $j$ is given by Equation (8).

Both upwind node and projected downwind node values are bounded by values at neighbouring cells that surround the upwind and downwind node, respectively. This is done by checking $\phi_U^*$ and $\phi_{D'}^*$ against the minimum and maximum values of $\phi$ over the cells that share faces of $C$ and $D$, respectively:

$$\phi_U = max\left[\phi_{min}^{Cngb}, min\left(\phi_U^*, \phi_{max}^{Cngb}\right)\right],\ \phi_{D'} = max\left[\phi_{min}^{Dngb}, min\left(\phi_{D'}^*, \phi_{max}^{Dngb}\right)\right] \tag{27}$$

The convected variable $\phi$ is now normalised by replacing the downwind value $\phi_D$ in Equation (10) with the projected node value $\phi_{D'}$.

## 5 RESULTS AND DISCUSSION

The test cases used in this study include pure convection of scalar discontinuities, laminar vortex shedding from a circular cylinder at $Re = 105$, turbulent flow through an engine port and LES of the JSAE engine port. Unstructured grids made of triangular, polyhedral and Cartesian cut cells are employed. The convergence properties and accuracy of the new method are scrutinised for all test cases.

### 5.1 Convection of step profile

A steady–state convective transport of a passive scalar, $\nabla\cdot\rho\vec{U}\phi = 0$, is frequently used to test and validate convective schemes. Here, a step profile is convected over a square domain by the given velocity field $\vec{U} = (\sqrt{2}/2, \sqrt{2}/2)$. Figure 4 (left) shows the geometry and boundary conditions for this case. A mildly skewed triangular grid, having 25 cells per side of a square, is used, Figure 4 (right)).

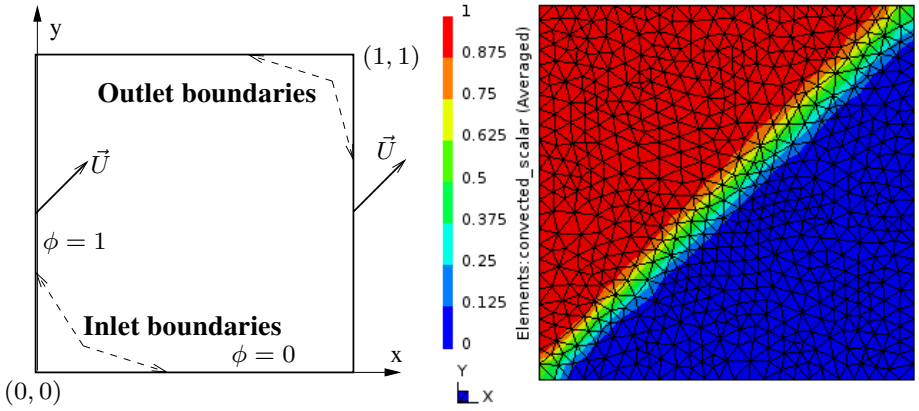



Figure 4: Step profile. Geometry with boundary conditions (left) and scalar distribution as computed by AVL SMART scheme (right).

The results were obtained by using double precision variables, and the convergence criterion based on the normalised residuals was $1\cdot 10^{-14}$. As Figure 5 indicates, the new UCF upwind node reconstruction method has a better convergence rate than the previous UCD method. In

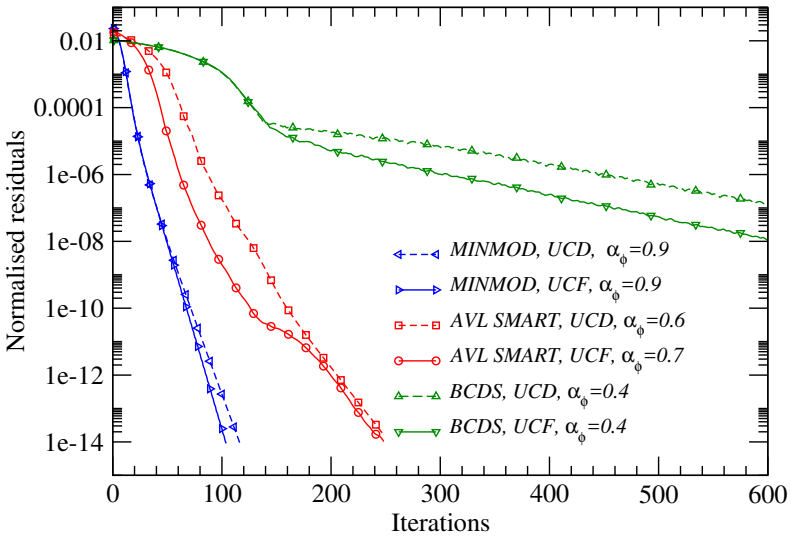



Figure 5: Step profile. Effect of upwind–node reconstruction methods on the convergence rate.

general, more accurate schemes require smaller under–relaxation factors $\alpha_\phi$ and more steady–state iterations to achieve the convergence. Though the bounded CDS has converged below the specified convergence criterion, its convergence properties in this case are not great.

The comparison of all schemes in terms of their ability to resolve the step profile is shown in Figure 6. Regardless of the upwind node reconstruction method, more accurate schemes

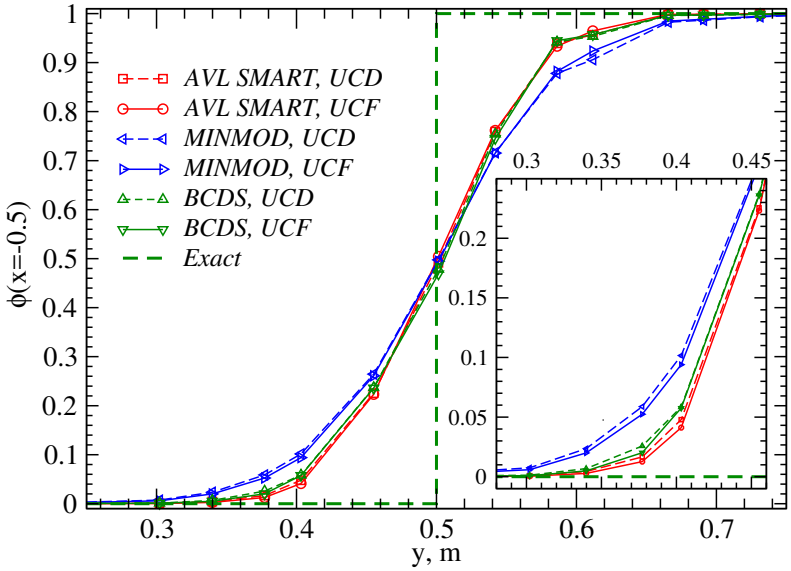



Figure 6: Step profile. Effect of upwind–node reconstruction methods on the scalar profile resolution at $x = 0.5$.

deliver better resolutions. The bounded CDS capability to resolve the step profile is quite close to AVL SMART, and significantly better than of MINMOD. Clearly, the proposed UCF method performs better than the UCD method for all tested schemes.

### 5.2 Laminar vortex shedding from a circular cylinder

The incompressible, unsteady flow around a two-dimensional circular cylinder is computed at Reynolds number $Re = \rho U_0 D/\mu = 105$, where $D = 1\ m$ is the cylinder diameter and

$U_0$ is the uniform inlet velocity. The flow is characterised by the presence of vortex shedding whose predicted strength is very sensitive to the employed convective scheme, [8]. A part of the solution domain, filled with triangles, is displayed in Figure 7. With the reference to the

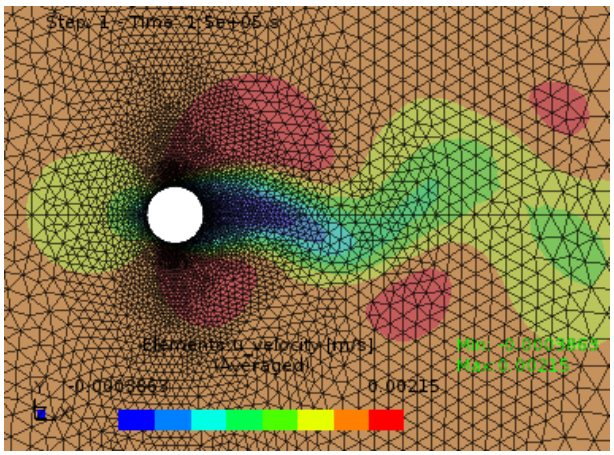



Figure 7: Circular cylinder. Triangular grid and vortex shedding pattern as visualised by distribution of $U$–velocity component at the given time.

cylinder centre, the inlet plane is placed at a distance $12D$, the outlet plane at $30D$ and the top and bottom symmetry planes at $12D$. A triangular grid, with 17,504 cells is employed. There are 160 cells along the cylinder wall, with approximate wall distances of $5.75 \times 10^{-3}\ m$.

Computations are performed using the third–order accurate AVL SMART scheme. Both first order Euler and second order TTL time schemes are used with a small time step size of $\Delta t = 0.5\ s$. This gives the non-dimensional time step size $\Delta t^* = U_0 \Delta t / D = 8.2 \cdot 10^{-4}$. The convergence criterion is set to $1. \times 10^{-6}$.

Figure 8 compares the time histories of the lift ($C_L$) and drag ($C_D$) coefficients (drag and lift forces normalised by $0.5\rho U_0^2 D$) as calculated by UCD and UCF upwind node reconstruction methods. Both methods deliver similar histories although the phase shift between two methods

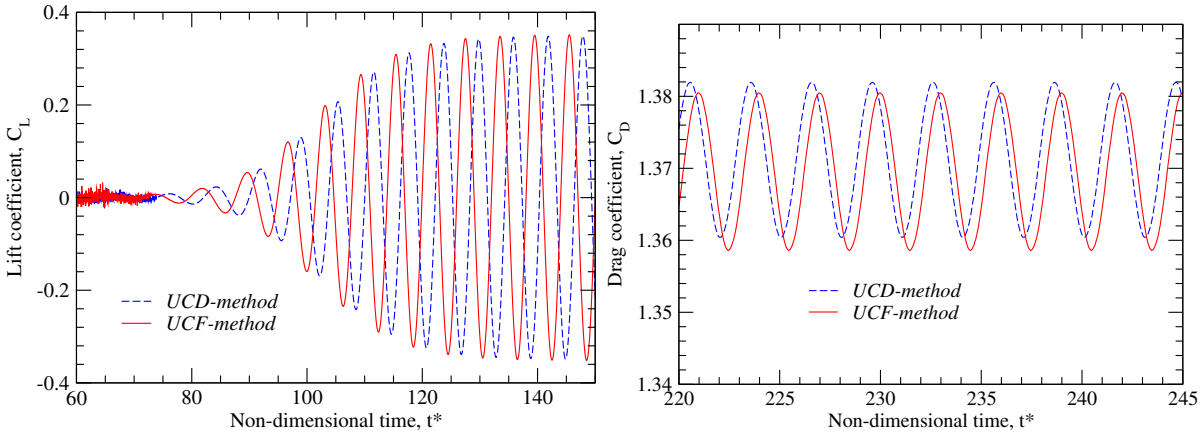



Figure 8: Vortex shedding from a circular cylinder at $Re = 105$. Time histories of the lift (left) and drag (right) coefficients as obtained by two upwind node reconstruction methods.

is evident. The integer flow parameters such as Strouhal number $St$, the time–averaged drag coefficient $\overline{C}_D$ and amplitudes of the lift coefficient $\widetilde{C}_L$ are compared in Table 1. For the same time–differencing scheme, the new UCF–method predicts lower values for the mean drag and

Table 1: Vortex shedding from a circular cylinder at $Re = 105$. Integral flow parameters as predicted by two upwind node reconstruction methods.

| Time scheme | Euler | | | Three time level | | |
|---|---|---|---|---|---|---|
| Flow parameter | $St$ | $\overline{C}_D$ | $\pm\widetilde{C}_L$ | $St$ | $\overline{C}_D$ | $\pm\widetilde{C}_L$ |
| UCD–method | 0.1660 | 1.371 | $\pm$0.350 | 0.1636 | 1.374 | $\pm$0.355 |
| UCF–method | 0.1665 | 1.370 | $\pm$0.352 | 0.1638 | 1.369 | $\pm$0.355 |

slightly higher values for the Strouhal number. The differences between the TTL and Euler scheme results are relatively small.

Note that predicted flow parameters are sensitive to the size of the solution domain. Therefore it is important to make a comparison with the results corresponding to the same or similar size of the solution domain. The correlations developed by [37], and based on the polynomial fitting of spectral element method results, are useful in this context as they take into account the solution domain extents. Thus the comparison of present results for a more accurate TTL scheme with the numerical correlations from [37] and the experiment of [38] is given in Table 2. It can be seen that new UCF–method results are in slightly better agreement with the

Table 2: Vortex shedding from a circular cylinder at $Re = 105$. Comparison of predicted integral parameters as obtained by TTL scheme with the results from literature. The values in brackets are relative differences with reference to values from [37, 38].

| Authors | $St$ | $\overline{C}_D$ | $\pm\widetilde{C}_L$ |
|---|---|---|---|
| Present, UCD | 0.1636 (-1.9 %) | 1.374 (0.4 %) | 0.355 (-0.8 %) |
| Present, UCF | 0.1638 (-1.7 %) | 1.369 (0.0 %) | 0.355 (-0.8 %) |
| Experiment, [38] | 0.1667 | - | - |
| Numerical, [37] | - | 1.369 | 0.358 |

results from literature than the UCD–method results.

### 5.3 Steady–state turbulent flow through an engine port

The turbulent flow trough a baseline intake port with the non–dimensional valve lift $L/D_i = 0.355$ ($D_i$ is the valve inner seat diameter) is computed. The MINMOD convective scheme is employed whereas turbulence is modelled using the time–scale bounded $k - \varepsilon$ model [39] in conjunction with enhanced wall functions [40].

The experimental port assembly is illustrated in Figure 9 (left). The air flow is supplied by an inlet plenum where the stagnation boundary conditions are maintained. The flow leaves a port cylinder via two outlet tubes where impulse torque meters are installed. The torque meters are placed close to outlet tubes ends. Note that $y$-axis starts at the port symmetry plane, and coincides with the outlet tubes axes. From the outlet tubes, the flow is discharged into atmosphere at the constant pressure. The flow experiments are conducted under adiabatic conditions (zero heat fluxes at all walls).

In the present simulations, the solution domain corresponds to the experimental rig flow geometry. However, the outlet tubes are truncated so that outflow pressure boundaries coincide with the starting positions of torque meters in the experimental rig. In this way the flow through the torque meter honeycomb is not modelled as porous media, i.e the presence of meters is ignored. The working fluid is air, and it is considered as an ideal compressible gas for the density predictions. The numerical grid, comprising $2,638,706$ internal polyhedral cells is used.

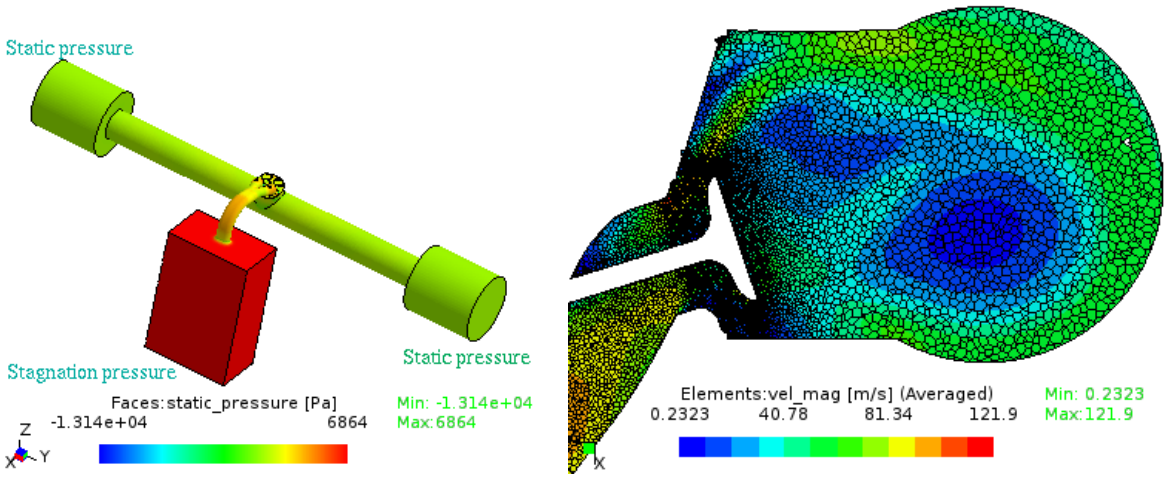



Figure 9: Intake port. Geometry (left) and velocity magnitude at the cross-section $y = 20\ mm$ (right).

Details of this grid around the valve, accompanied with the velocity magnitude distribution, are displayed in Figure 9 (right).

The convergence behaviour of the proposed UCF–method versus the UCD–method can be examined through the history of normalised residuals shown in Figure 10. With the UCF–

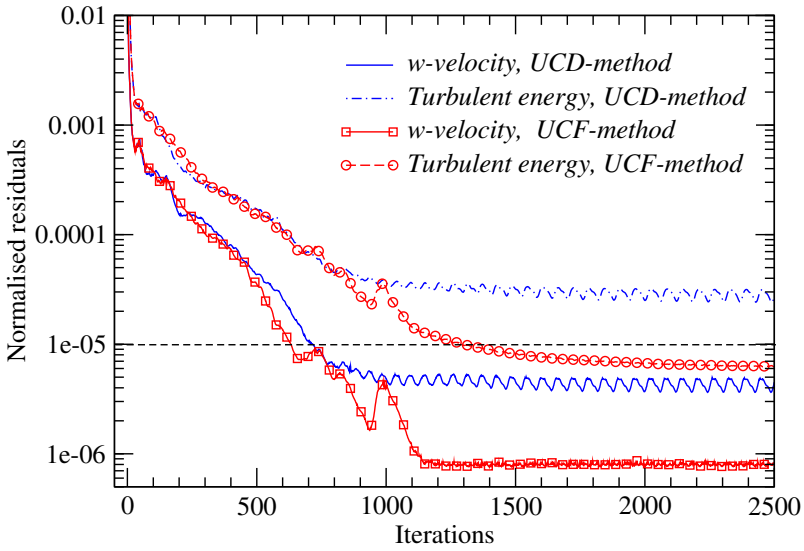



Figure 10: Intake port. Normalised residuals for $w$-velocity and turbulent kinetic energy corresponding to the two upwind node reconstruction methods.

method, the normalised residuals dropped below $1 \times 10^{-5}$ whereas this was not the case for the UCD–method. The calculated values of the inlet mass flow rate $\dot{m}_{in}$ and the torque $T$ (at the $y$–plane defined by position of a torque meter) are compared with the measured ones in Table 3. A good agreement between the predicted results with the measured ones is evident. Although the calculated torque value using the UCD–method appears quite close to the measured one, it should not be overlooked that the UCD–method has not fully converged.

Table 3: Intake port. Comparison of the predicted mass flow rate and torque values with the measured ones.

| Quantity | UCD–method | UCF–method | Measured |
|---|---|---|---|
| Mass flow rate $\dot{m}_{in}$, kg/s | 0.09841 | 0.09875 | 0.10236 |
| Relative difference, % | -3.86 | -3.53 | – |
| Torque $T \times 10^3$, Nm | 27.752 | 27.008 | 28.239 |
| Relative difference, % | -1.72 | -4.35 | – |

### 5.4 Large–eddy simulation of JSAE port

In this validation case, LES of JSAE engine port is performed using Smagorinsky–Lilly sub-grid scale model and bounded CDS scheme.

The port model was designed by Japanese Society of Automotive Engineers (JSAE) for their benchmark study, see [41]. Its geometry, shown in Figure 11 (left), has a simple $90^\circ$ axisymmetric design which provides the non–swirling flow around an engine valve at a lift of $7\,mm$ At the port inlet, the mass flow rate of $0.0254\ kg/s$ was prescribed, while the static pressure

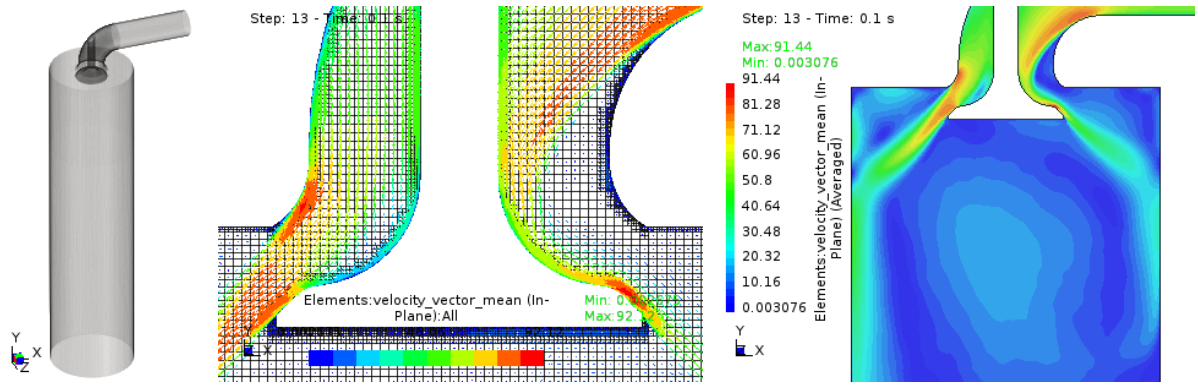



Figure 11: JSAE port. Geometry (left), details of Cartesian cut–cell grid overlaid with time–averaged velocity vectors (middle) and in–plane time–averaged velocity vector contours (right) around the valve in the symmetry plane $z = 0$ (right).

of $p_{out} = 101176\ Pa$ was specified at the outlet. The measured pressure drop, calculated as the difference between the total pressure at the inlet $p_{tot}$ and static pressure at the outlet was $\Delta p_{exp} = 3349\ Pa$.

The Cartesian cut–cell grid with 7.9 million cells and the maximum non–dimensional wall distance of $y^*_{max} = 51$ was considered suitable for the present LES. Unsteady runs were performed using the second order TTL time scheme with the time step size of $\Delta t = 1. \times 10^{-5}\ s$. Typically, three outer iterations per time step were required to reduce the normalised residual below the specified convergence criterion of $1. \times 10^{-5}$. Simulations were conducted over $10,000$ time steps ($0.1\ s$), and the last $5,000$ steps were used to obtain the time–averaged variables.

Figures 11 (middle, right) show velocity plots around the valve in the symmetry plane $z = 0$. On the right side of the valve and around the valve seat, a large recirculation region is present. The size of this region is influenced by the flow separation point on the port surface, above the valve seat. Thus this region determines the flow structure in the valve gap, which in turn affects the pressure drop across the port.

The sub–grid scale viscosity at the end of the simulation (as calculated by the UCF–method) is displayed in Figure 12 (left). On the right, time histories of the pressure drop $\Delta p(t) =$

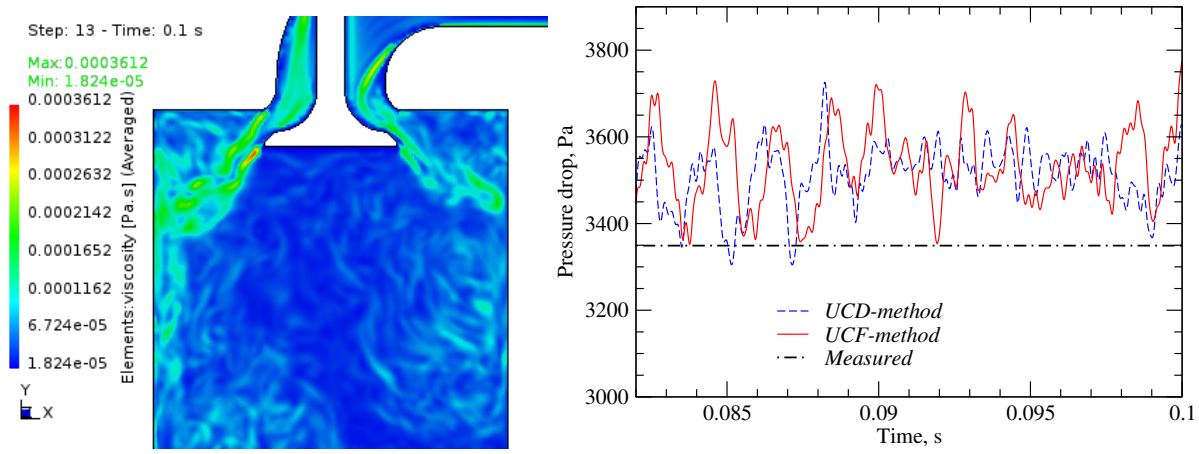



Figure 12: JSAE port. Sub–grid scale viscosity at the given time (left) and the effect of the upwind node reconstruction method on the pressure drop history (right).

$p_{tot} - p_{out}$ are plotted for both UCD and UCF upwind node reconstruction methods. The time averaged pressure drop values obtained by UCD and UCF methods are $\Delta p_{CD} = 3486\, Pa$ and $\Delta p_{CF} = 3527\, Pa$, respectively. With reference to the measured pressure drop $\Delta p_{exp} = 3349\ Pa$, these values over–predict the pressure drop for $4.1\%$ and $5.3\%$, respectively.

## 6 Conclusions

The implementation of bounded convective schemes on three–dimensional unstructured grids has been revisited. A central problem of the far upwind node reconstruction is generalised by defining an upwind–biased stencil along the local coordinate connecting the cell centre upstream of the cell–face and the cell-face centre. With this approach, a diffusion–like error introduced by grid skewness can be alleviated. The improved convergence properties and accuracy of the new method are demonstrated by computing four test cases. For each case, either analytical or available experimental/numerical results are used to validate the present method. In addition, the new method is scrutinised against the well established upwind node reconstruction method which by its definition can not mitigate the grid skewness error. Considering the performance of higher order schemes, the MINMOD scheme is well suited for the steady–state simulation whereas the modified SMART is recommended for unsteady simulations. In the case of LES, it is beneficial to employ the bounded central differencing scheme.

## References

[1] Roe, P.L., Characteristic–based schemes for the euler equations. *Annual Review of Fluid Mechanics* **18**:337–365 (1986).

[2] Harten, A., High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics* **49**:357–393 (1983).

[3] Sweby, P.K., High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM Journal Numer. Anal.* **21**:995–1011 (1984).

[4] Gaskell, P.H. and Lau, A.K., Curvature compensated convective transport: Smart, a new boundedness preserving transport algorithm. *International Journal for Numerical Methods in Fluids* **8**:617–641 (1988).

[5] Waterson, N.P. and Deconinck, H., Design principles for bounded higher–order convection schemes – a unified approach. *Journal of Computational Physics* **224**:182–207 (2007).

[6] Darwish, M.S. and Moukalled, F.H., Normalized variable and space formulation methodology for high–resolution schemes. *Numerical Heat Transfer B* **26**:79–96 (1994).

[7] Alves, M.A., Oliveira, P.J. and Pinho, F.T., A convergent and universally bounded interpolation scheme for the treatment of advection. *International Journal for Numerical Methods in Fluids* **41**:47–75 (2003).

[8] Przulj, V.P., *Computational Modelling of Vortex Shedding Flows*. Ph.D. Thesis, City University, London. http://openaccess.city.ac.uk/id/eprint/7565 (1998).

[9] Piperno, S. and Depeyre, S., Criteria for the design of limiters yielding efficient high resolution tvd schemes. *Computers & Fluids* **27**:183–197 (1998).

[10] Ubbink, O. and Issa, R.I., A method for capturing sharp fluid interfaces on arbitrary meshes. *Journal of Computational Physics* **153**:26–50 (1999).

[11] Przulj, V. and Basara, B., Bounded convection schemes for unstructured grids. *AIAA 2001–2593* (2001).

[12] Jasak, H., Weller, H.G. and Gosman, A.D., High resolution NVD differencing scheme for arbitrarily unstructured meshes. *International Journal for Numerical Methods in Fluids* **31**:431–449 (1999).

[13] Li, L.X., Liao, H.S. and Qi, L.J., An improved r-factor algorithm for TVD schemes. *International Journal of Heat and Mass Transfer* **51**:610–617 (2008).

[14] Hou, J., Simons, F. and Hinkelman, R., A new TVD method for advection simulation on 2d unstructured grids. *International Journal for Numerical Methods in Fluids* **71**:1260–1281 (2013).

[15] Zhang, D., Jiang, C., Liang, D. and Cheng, L., A review on TVD schemes and a refined flux–limiter for steady–state calculations. *Journal of Computational Physics* **302**:114–154 (2015).

[16] Touze, C.L., Murrone, A. and Guillard, H., Multislope MUSCL method for general unstructured meshes. *Journal of Computational Physics* **284**:389–418 (2015).

[17] Denner, F. and van Wachem, B.G.M., TVD differencing on three-dimensional unstructured meshes with monotonicity-preserving correction of mesh skewness. *Journal of Computational Physics* **298**:466–479 (2015).

[18] Demirdzic, I., Lilek, Z. and Peric, M., A collocated finite volume method for predicting flows at all speeds. *Int. Journal for Numerical Methods in Fluids* **16**:1029–1050 (1993).

[19] Przulj, V. and Basara, B., A SIMPLE–based control volume method for compressible flows on arbitrary grids. *AIAA 2002 –3289* (2002).

[20] Rhie, C.M. and Chow, W.L., Numerical study of the turbulent flow past an airfoil with trailing edge separation. *AIAA Journal* **21**:1525–1532 (1983).

[21] Ferziger, J. and Peric, M., *Computational Methods for Fluid Dynamics*. Springer, Berlin (1997).

[22] Warming, R.F. and Beam, R.M., Upwind second order difference schemes and applications in aerodynamic flows. *AIAA Journal* **14**:1241–1249 (1976).

[23] Leonard, B.P., A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Comp. Meth. Appl. Mech. Eng.* **19**:59–98 (1979).

[24] van Leer, B., Upwind difference methods for aerodynamic problems governed by the Euler equations. *Lectures in Applied Mathematics* **22**:327–336 (1985).

[25] Demirdzic, I. and Muzaferija, S., Numerical method for coupled fluid flow, heat transfer

and stress analysis using unstructured moving meshes with cells of arbitrary topology. *Comp. Methods Appl. Mech. Eng.* **125**:235–255 (1995).

[26] Muzaferija, S., *Adaptive Finite Volume Method for Flow Predictions Using Unstructured Meshes and Multigrid Approach*. PhD dissertation, University of London (1994).

[27] Przulj, V., Birkby, P. and Mason, P., Finite volume method for conjugate heat transfer in complex geometries using Cartesian cut–cell grids. In *CHT-08*, Marrakech, Morocco (2008).

[28] Patankar, S.V., *Numerical Heat Transfer and Fluid Flow*. McGraw–Hill, New York (1980).

[29] Van Der Vorst, H.A., Bi–CGSTAB: A fast and smoothly converging variant of Bi–CG for the solution of nonsymmetric linear systems. *SIAM J. Scientific Computing* **13**:631–644 (1992).

[30] Sleijpen, G. and Fokkema, D.R., Bicgstab($l$) for linear equations involving unsummetric matrices with complex spectrum. *Electronic Trans. Numer. Anal.* **1**:11–32 (1993).

[31] Przulj, V.P., Generalized SIMPLE–based pressure correction method for unstructured colocated grids. *AIAA Journal* **0**:1–12. Published online March 03, 2016, http://dx.doi.org/10.2514/1.J054505 (2016).

[32] Mathur, S.R. and Murthy, J.Y., A pressure based method for unstructured meshes. *Numerical Heat Transfer B* **31**:195–215 (1997).

[33] Leonard, B.P., Simple high–accuracy resolution program for convective modelling of discontinuities. *International Journal for Numerical Methods in Fluids* **8**:1291–1318 (1988).

[34] Georgiadis, N.J., Rizzetta, D.P. and Fureby, C., Large–Eddy Simulation: Current capabilities, recommended practices, and future research. *AIAA Journal* **48**(8):1772–1784 (2010).

[35] Jasak, H., *Error Analysis and Estimation in the Finite Volume Method with Applications to Fluid Flows*. Ph.D. Thesis, Imperial College, University of London,. (1996).

[36] Demirdzic, I. and Muzaferija, S., Numerical method for coupled fluid flow, heat transfer and stress analysis using unstructured moving meshes with cells of arbitrary topology. *Comp. Methods Appl. Mech. Eng.* **125**:235–255 (1995).

[37] Posdziech, O. and Grundmann, R., A systematic approach to the numerical calculation of fundamental quantities of the two–dimensional flow over a circular cylinder. *Journal of Fluids and Structures* **23**:479–499 (2007).

[38] Williamson, C.H., Oblique and parallel modes of vortex shedding in the wake of a circular cylinder at low reynolds numbers. *Journal of Fluid Mechanics* **206**. 579. (1989).

[39] Przulj, V., Penning, R. and Tiney, N., Conjugate heat transfer modelling of internal combustion engine structures and coolant flows. In *CHT-12*, Bath, England (2012).

[40] Przulj, V., Pragmatic wall treatment for RANS simulations on Cartesian cut–cell grids. In K. Hanjalic and Y. Nagano (eds.), *6th Int. Symp. Turbulence, Heat and Mass Transfer*, Rome, Italy (2009).

[41] Yamada, T., Comparisons of validation of engine intake port flow simulated by commercial CFD codes. *JSAE Paper* (20014699):20–26 (2001).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A LOW-MACH LOW-REYNOLDS PRECONDITIONING SCHEME WITH PARTICULAR ATTENTION ON VISCOUS TIME-STEPPING

**J. Fiedler and G. Ashcroft**

Institute of Propulsion Technology
German Aerospace Center address
Linder Höhe
51147 Cologne
e-mail: {jens.fiedler, graham.ashcroft}@dlr.de

**Keywords:** convergence acceleration, compressible flows, viscous flows, low Mach preconditioning

**Abstract.** *Industrial design and optimization processes rely increasingly on powerful and well-engineered CFD tools. Compressible solution methods, which perform very well at transonic and supersonic flow speeds, display a dramatic degradation of convergence as well as of solution quality as the incompressibility limit is approached (very low flow speeds). In many technical applications, especially in turbomachinery, the flow conditions vary strongly within the computational domain and with time. When the incompressibility limit is approached, a large disparity arises between the smallest and largest eigenvalues of the systems characteristic matrix. In order to overcome these problems low-Mach preconditioning methods have been devised to rescale the eigenvalues of the characteristic matrix of the system of governing equations and, hence, reduce the large inequality in the acoustic and convective flow speeds. Frequently, low flow speeds are observed in low Reynolds environments. As noted by several authors, often instability issues in these regions, such as cavities and boundary layers, arise by preconditioning. Often, this problem is due to an overestimation of the maximum allowable timestep size. In fact, in a low Reynolds regime, the influence of viscous effects on time marching schemes predominates. An important role in the determination of viscous time steps plays the von Neumann number (VNN), whereas the Courant Friedrichs Lewy criteria (CFL) influences the inviscid time step behaviour.*

*The objective of this work is the presentation of the theoretical background and results of a consistent Low-Mach preconditioning scheme based on a preconditioner proposed by Turkel, which has been extended to a wide range of Reynolds numbers. Furthermore, the interaction of low-Mach preconditioning, the von Neumann and Courant Friedrichs Lewy numbers on the convergence history and quality of results will be discussed. Its implementation is illustrated using DLR's in-house CFD code TRACE. To prove the robustness and correctness of the algorithm, we discuss a set of test cases like the lid-driven cavity at different Reynolds numbers influenced by CFL and VNN.*

## 1 INTRODUCTION

Complex technical systems, such as turbomachinery configurations, involve large variations of different flow topologies. Some regions contain very low flow speeds whilst others are decidedly compressible. Exemplary, in the main flow passage of a turbomachinery configuration relatively high flow speeds dominate. Locally small Mach numbers can be found in cavity and seals regions. Numerical simulations of turbomachinery components are mostly performed by density-based flow solvers. At low flow speeds, fully compressible solution methods converge extremely slow, due to the large disparity between the convective and acoustic speeds. In addition, a dramatic degradation of the solution quality can by observed.

In order to overcome these problems, low-Mach preconditioning operators have been developed to reduce the difference between the largest and smallest eigenvalues of the system's characteristic matrix [1]. Pre-multiplication of the time derivative of the Euler or Navier-Stokes system of equations by a suitable matrix rescales the various characteristic velocities. Preconditioning does not only reduce the stiffness of the system but also significantly increase the accuracy at low flow speeds. In the past, several preconditioning operators have been published, e.g. [10, 1, 12, 13]. A comprehensive overview of the low Mach preconditioning operators, developed during the last decades, can be found in the work of Depcik [2]. As noted by several authors low-Mach preconditioning can cause stability problems in viscous regimes [3, 6]. This is due to overestimation of the maximum allowable timestep size. In fact, in a low Reynolds regime the time steps are predominated by the viscous eigenvalues. This aspect of preconditioned flows has been discussed previously by Choi and Merkle [13] as well as Colin *et. al* [6]. Based on stability investigations the significant influence of viscosity on low Mach preconditioning has been proven. Based on a preconditioner proposed by Turkel [1], in this paper we discuss the influence of an additional limitation term of the preconditioning parameter $\beta^2$ on the numerical stability in a viscous environment, cf. [3, 6].

Moreover, the validity and robustness of the method is demonstrated using a classical lid-driven cavity configuration at different Reynolds, CFL and von Neumann numbers.

The improvements has been implemented in the CFD code TRACE, a fully implicit, parallel, hybrid, multi-block, Reynolds-Averaged Navier-Stokes flow solver specialised in the simulation of turbomachinery flows [5].

## 2 NUMERICAL METHOD

The non-dimensional time-dependent Navier-Stokes equations formulated in Cartesian coordinates read [4]

$$\frac{\partial \mathbf{Q}}{\partial t} + \frac{\partial \mathbf{f}}{\partial x} + \frac{\partial \mathbf{g}}{\partial y} + \frac{\partial \mathbf{h}}{\partial z} = \frac{1}{Re_a}\left(\frac{\partial \mathbf{f}_\nu}{\partial x} + \frac{\partial \mathbf{g}_\nu}{\partial y} + \frac{\partial \mathbf{h}_\nu}{\partial z}\right), \tag{1}$$

where $\mathbf{Q}$ is the solution vector of the conservative variables $(\rho, \rho u, \rho v, \rho w, \rho E)$ and $\mathbf{f}$, $\mathbf{g}$, $\mathbf{h}$ the inviscid fluxes given by, cf. [6]:

$$\mathbf{f} = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ \rho uH \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} \rho v \\ \rho vu \\ \rho v^2 + p \\ \rho vw \\ \rho vH \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} \rho w \\ \rho wu \\ \rho wv \\ \rho w^2 + p \\ \rho wH \end{pmatrix}, \tag{2}$$

where $\rho$, $u$, $v$, $w$ and $p$ are the density, velocities and pressure, respectively. The quantities $H$ and $E$ are the specific total enthalpy and energy. The viscous fluxes are given by

$$\mathbf{f}_\nu = \begin{pmatrix} 0 \\ \tau_{xx} \\ \tau_{xy} \\ \tau_{xz} \\ (\tau U)_x) \end{pmatrix}, \quad \mathbf{g}_\nu = \begin{pmatrix} 0 \\ \tau_{yx} \\ \tau_{yy} \\ \tau_{yz} \\ (\tau U)_y) \end{pmatrix}, \quad \mathbf{h}_\nu = \begin{pmatrix} 0 \\ \tau_{zx} \\ \tau_{zy} \\ \tau_{zz} \\ (\tau U)_z) \end{pmatrix}. \tag{3}$$

The components of the viscous stress tensor are referred to as $\tau_{ij}$ where $i, j = x, y, z$ . The vector $U$ contains the Cartesian velocity components. Source terms and heat flux contributions are not in the focus of the current work and have therefore been ignored. The fluid is considered as calorically perfect. Hence, the ratio of the specific heats is constant and set to 1.4. The acoustic Reynolds number is defined by

$$Re_a = \frac{\rho^* a^* L^*}{\mu^*}, \tag{4}$$

where $\rho^*$, $a^*$, $L^*$, $\mu^*$ are the (dimensional) reference density, speed of sound, length and viscosity. The system of equations (1) is spatially discretized using Roe's upwind approximation scheme [17]. The (preconditioned) inviscid fluxes at the cell faces are computed as

$$\mathbf{F}_{i+1/2} = \frac{1}{2}\left(\mathbf{F}_L + \mathbf{F}_R\right) - \frac{\partial \mathbf{Q}}{\partial \mathbf{U}} \mathbf{P}_U^{-1} |\mathbf{P}_U \mathbf{D}_U| \frac{\partial \mathbf{U}}{\partial \mathbf{Q}} \left(\mathbf{Q}_L - \mathbf{Q}_R\right), \tag{5}$$

where $\mathbf{Q}_L$ and $\mathbf{Q}_R$ are the states at the left and right side of the cell face. The term $\mathbf{P}_U^{-1}|\mathbf{P}_U\mathbf{D}_U|$, where $\mathbf{D}_U = \kappa \cdot \frac{\partial \mathbf{F}}{\partial \mathbf{U}}$, is known as the stabilisation term and leads after a matrix decomposition of the flux Jacobians, to $\mathbf{P}_U^{-1}\mathbf{M}_U|\mathbf{\Lambda}|\mathbf{M}_U^{-1}$. The primitive variable system $(\rho, u, v, w, p)$ is denoted as $U$. Based on a metric proposed by Hirsch [4] the preconditioned left and right eigenvector matrices are

$$\mathbf{M}_U^{-1} = \begin{pmatrix} \hat{\kappa}_x & 0 & \hat{\kappa}_z & -\hat{\kappa}_y & -\hat{\kappa}_x \frac{1}{a^2} \\ \hat{\kappa}_y & -\hat{\kappa}_z & 0 & \hat{\kappa}_x & -\hat{\kappa}_y \frac{1}{a^2} \\ \hat{\kappa}_z & \hat{\kappa}_y & -\hat{\kappa}_x & 0 & -\hat{\kappa}_z \frac{1}{a^2} \\ 0 & \hat{\kappa}_x & \hat{\kappa}_y & \hat{\kappa}_z & -\frac{\hat{\lambda}_1 - \hat{\lambda}_4}{\beta^2 a^2 \rho} \\ 0 & -\hat{\kappa}_x & -\hat{\kappa}_y & -\hat{\kappa}_z & \frac{\hat{\lambda}_1 - \hat{\lambda}_5}{\beta^2 a^2 \rho} \end{pmatrix} \tag{6}$$

$$\mathbf{M}_U = \begin{pmatrix} \hat{\kappa}_x & \hat{\kappa}_y & \hat{\kappa}_z & \frac{\rho\beta^2}{\hat{\lambda}_4 - \hat{\lambda}_5} & \frac{\rho\beta^2}{\hat{\lambda}_4 - \hat{\lambda}_5} \\ 0 & -\hat{\kappa}_z & \hat{\kappa}_y & \hat{\kappa}_x \frac{\hat{\lambda}_1 - \hat{\lambda}_5}{\hat{\lambda}_4 - \hat{\lambda}_5} & \hat{\kappa}_x \frac{\hat{\lambda}_1 - \hat{\lambda}_4}{\hat{\lambda}_4 - \hat{\lambda}_5} \\ \hat{\kappa}_z & 0 & -\hat{\kappa}_x & \hat{\kappa}_y \frac{\hat{\lambda}_1 - \hat{\lambda}_5}{\hat{\lambda}_4 - \hat{\lambda}_5} & \hat{\kappa}_y \frac{\hat{\lambda}_1 - \hat{\lambda}_4}{\hat{\lambda}_4 - \hat{\lambda}_5} \\ -\hat{\kappa}_y & \hat{\kappa}_x & 0 & \hat{\kappa}_z \frac{\hat{\lambda}_1 - \hat{\lambda}_5}{\hat{\lambda}_4 - \hat{\lambda}_5} & \hat{\kappa}_z \frac{\hat{\lambda}_1 - \hat{\lambda}_4}{\hat{\lambda}_4 - \hat{\lambda}_5} \\ 0 & 0 & 0 & \frac{\rho\beta^2 a^2}{\hat{\lambda}_4 - \hat{\lambda}_5} & \frac{\rho\beta^2 a^2}{\hat{\lambda}_4 - \hat{\lambda}_5}, \end{pmatrix} \tag{7}$$

where $\kappa_x$, $\kappa_y$ and $\kappa_z$ are the metric components. The preconditioning operator $\mathbf{P}_U$ and its

inverse $\mathbf{P}_U^{-1}$, proposed by Turkel [10] and formulated in primitive variables, reads

$$\mathbf{P}_U^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & -\frac{\beta^2-1}{a^2\beta^2} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \beta^{-2} \end{pmatrix} \qquad \mathbf{P}_U = \begin{pmatrix} 1 & 0 & 0 & 0 & \frac{\beta^2-1}{a^2} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \beta^2 \end{pmatrix} \tag{8}$$

The diagonal matrix $\mathbf{\Lambda}$ contains the (preconditioned) normalised eigenvalues which are given by

$$\hat{\lambda}_{1,2,3} = u\hat{\kappa}_x + v\hat{\kappa}_y + w\hat{\kappa}_z \tag{9}$$

$$\hat{\lambda}_{4,5} = \frac{1}{2}\left(1+\beta^2\right)\hat{\lambda}_1 \pm \frac{1}{2}\sqrt{\left(1-\beta^2\right)^2\hat{\lambda}_1^2 + 4\beta^2 a^2} \tag{10}$$

The components of the metric vector have been normalised by

$$\hat{\kappa}_i = \frac{\kappa_i}{\|\kappa\|} \tag{11}$$

which results in

$$\hat{\lambda}_i = \frac{\lambda_i}{\|\kappa\|} \tag{12}$$

The system of equations (1) is temporally discretised by a first order Euler backward discretisation scheme. Within this work steady state cases have been considered only and a pseudo-time step $\tau$ has been introduced to control the residual behaviour, cf. [8] The preconditioning operator modifies the pseudo-time $\tau$ and the residual $\mathbf{R}$. Therefore, the update vector $\triangle\mathbf{Q}$ at time step $m$ reads

$$\triangle\mathbf{Q} = -\mathbf{R}^m \left( \frac{\mathbf{P}_Q^{-1}}{\triangle\tau} + \left.\frac{\partial\mathbf{R}}{\partial\mathbf{Q}}\right|^m \right)^{-1} \tag{13}$$

The pseudo-time step $\triangle\tau_i$ considered for a cell $i$ is determined by applying the CFL condition

$$\triangle\tau_i = CFL\frac{\triangle h_i}{\lambda_{4,i} + \delta\frac{CFL}{VNN}\lambda_{\nu,i}} \tag{14}$$

where $\triangle h_i$ is a characteristic length of the cell volume, $\lambda_{4,i}$ is given by (10) and $\lambda_{v,i}$ is defined by [7, 14]

$$\lambda_{\nu,i} = \max\left(\frac{4}{3\rho}, \frac{\gamma}{\rho}\right)\left(\frac{\mu_L}{Pr_L} + \frac{\mu_T}{Pr_T}\right)\frac{1}{\triangle h_i}, \tag{15}$$

where $Pr_L$ and $Pr_T$ are the laminar and turbulent Prandtl numbers. The laminar and turbulent dynamic viscosity coefficients are denoted by $\mu_L$ and $\mu_T$, respectively. For purely inviscid consideration $\delta = 0$.

The preconditioning parameter $\beta^2$ is related to the local Mach number $M$ and is initially defined as [6]

$$\beta^2 = \min\left(\max\left(M^2, \beta^2_{min}\right), 1\right). \tag{16}$$

In order to avoid singularities in the preconditioning matrices $\beta^2_{min}$ should be set to some appropriate value. Moreover, the choice of $\beta^2_{min}$ strongly influences the stability of the preconditioned system. As stated by several authors [1, 15], $\beta^2_{min}$ should not be extremely small compared to the general flow velocity. Turkel [1] suggests to set $\beta^2_{min}$ to $K_1 M^2_{ref}$, where $K_1$ is a problem-dependent constant and $M_{ref}$ a reference Mach number, i.e. often the inflow Mach number. In the current investigations $\beta^2_{min} = 10^{-5}$. In the vicinity of stagnation points the local Mach number approaches zero. Darmofal and Siu [11] suggests to limit $\beta^2$ additionally by the local pressure gradient $K_2 \max_{faces}(|\triangle p|/\rho a^2)$, where $\triangle p = p_r - p_l$ and $p_r$ and $p_l$ are the pressures of the particular states and $K_2$ is a problem-dependent constant. Hence, $\beta^2$ reads

$$\beta^2 = \min\left(\max\left(M^2, \beta^2_{min}, K_2 \max_{faces}\left(\frac{|\triangle p|}{\rho a^2}\right)\right), 1\right). \tag{17}$$

Venkateswaran [3] suggests to limit $\beta^2$ at low $Re$ numbers using the diffusion velocity defined by

$$v_{vis} = \frac{\nu}{\triangle x}, \tag{18}$$

where $\nu$ is the kinematic viscosity related to the dynamic viscosity $\mu$ via $\nu = \mu/\rho$. Hence,

$$v_{vis} = \frac{\mu}{\rho \triangle x} \tag{19}$$

In order to determine a low Reynolds limit for $\beta^2$ the viscous Mach number is introduced by

$$M_{vis} = \frac{v_{vis}}{a}. \tag{20}$$

Within the current work, the results above have been combined to obtain

$$\beta^2 = \min\left(\max\left(M^2, \beta^2_{min}, K_2 \max_{faces}\left(\frac{|\triangle p|}{\rho a^2}\right), M^2_{vis}\right), 1\right). \tag{21}$$

The discretised algebraic system of equations has been solved using a symmetric Gauss-Seidel algorithm (SGS) and a predictor-corrector Gauss-Seidel scheme (PC-SGS). All calculations have been performed in double precision.

## 3 RESULTS

The lid-driven cavity configuration has been chosen to prove the correctness and accuracy of the algorithm. Often, this problem serves as a benchmark for the incompressible Navier-Stokes equations [18, 16]. Despite its simple setup, it allows a wide range of numerical properties and stability issues to be studied. Depending on the Reynolds number, the flows are characterised by multiple counter rotating recirculation regions in the corners of the cavity. The computational domain consists of a square where the upper wall moves at a constant horizontal velocity. All

walls are considered adiabatic. The Reynolds number considered within this configuration is defined by

$$Re_1 = \frac{\rho^* \, u^*_{Wall} \, L^*}{\mu^*}. \tag{22}$$

Numerical studies have been performed modifying the Reynolds number $Re_1$ by varying the reference length $L^*$. The latter scales the size of the configuration keeping $u^*_{Wall}$ constant and therefore the velocity distribution within the domain over the whole range of $Re_1$.

For $Re_1$ = 1000 grid independency studies have been carried out for different mesh sizes generated by successive refinement. Figures (5) and (6) show the v-velocity profiles along a horizontal line and the u-velocity profiles along a vertical line passing through the centre of the cavity. The results are in very good agreement with those obtained by Ghia *et. al.* [18]. Based on the mesh resolution results the computational grid is set to 81 x 81 cells without refinement at the walls.

The streamlines computed at Reynolds number $Re_1$ = 1 and $Re_1$ = 100 are shown in figures (1) and (3). Low-mach preconditioning does not show to have a significant influence on the results but provide a higher convergence rate at low CFL numbers, see figures (2) and (4). Calculations have been performed with CFL = 1 and CFL = 25. Both, the preconditioned and non-preconditioned solutions are in very good agreement with those published in the available literature, e.g [13].

In contrast to the non-preconditioned calculations obtained at $Re_1 = 1$ and $Re_1 = 100$ the quality of the results for $Re_1 = 1000$ presented in figure (7a) are not in accordance to those computed by low Mach preconditioning, see figure (7c). The solution quality has been considerably improved using a low Mach preconditioning technique. The influence of $M_{vis}$ on the numerical stability of a preconditioned calculation have been studied with SGS and PC-SGS. The solution quality diminishes drastically using a low Mach preconditioned SGS without additional modifications for low Reynolds environments as predicted by equation (21), figures (8) and (7b). The secondary recirculation patterns at the corners of the cavity are inadequately formed. The improvement in numerical stability through $M_{vis}$ on a low-Mach preconditioned predictor-corrector scheme is less important, but clearly recognisable, see figure (9). This emphasis the necessity to limit $\beta^2$ by $M_{vis}$ at low Reynolds numbers.

Low Mach preconditioning alleviates the strong disparity between the smallest and largest eigenvalues. Consequently, in low Reynolds regions it is inappropriate to use a purely inviscid scheme as the smaller the largest convective eigenvalue becomes the more important the largest viscous eigenvalue in a viscous regime becomes. Therefore, the convergence rate is no longer primarily determined by the low Mach preconditioner and CFL number but rather by the von Neumann number and largest viscous eigenvalue. Figures (10) and (11) present the convergence history of a lid-driven cavity calculation at $Re_1 = 1000$ at constant CFL and increasing VNN number up-to 10. Clearly, the viscous eigenvalue strongly alters the convergence rate but, based on the current configuration, has only a minor impact on the solution quality.

## 4 CONCLUSIONS

The numerical stability of a low-Mach preconditioned scheme is strongly influenced by a low-Reynolds modification of the preconditioning parameter $\beta$. In this work a consistent formulation of a preconditioning scheme applicable to a wide range of Reynolds and Mach numbers has been presented in the context of a fully-implicit three-dimensional compressible Navier-Stokes flow solver demonstrated. The validity has been proved using a classical lid-driven

cavity configuration. The influence of viscous timestep size has been studied and its impact on convergence rate shown. Moreover, for the configuration chosen in this work the viscous eigenvalues alter the computational result only slightly. In future work it is planned to investigate the viscous timestep behaviour depending on the ratio ($^{CFL}/_{VNN}$) and low-Mach preconditioning on different computational configurations.

## REFERENCES


[1] E. Turkel, Preconditioning techniques in computational fluid dynamics. *Annu. Rev. Fluid. Mech.*, **31**, 385–416, 1999

[2] Ch. D. Depcik, A perfect local preconditioner for the Navier-Stokes equations. *Master Thesis*, University of Michigan, 2002

[3] S. Venkateswaran and L. Merkle, Analysis of preconditioning methods for the Euler and Navier-Stokes equations. *von Kármán Institute for Fluid Dynamics, Lecture Series*, 1999

[4] Ch. Hirsch, *Numerical Computation of Internal and External Flows*. John Wiley & Sons, Ltd., 2007

[5] K. Becker, K. Heitkamp and E. Kügeler, Recent progress in a hybrid-grid CFD solver for turbomachinery flows. *Proceedings of the 5th European Congress on Computational Methods in Applied Science and Engineering (ECCOMAS 2010), Lisbon, Portugal, June 2010*

[6] Y. Colin, H. Deniau and J.-F. Boussuge, A robust low speed preconditioning formulation for viscous flow computations. *Computer & Fluids*, **47**, 1–15, 2011.

[7] R. C. Swanson, E. Turkel and J. A. White, An effective multigrid method for high-speed flows. *NASA CR-187602*, 1991.

[8] A. Jamenson, Time-dependent calculations using multigrid, with applications to unsteady flow past airfoils and wings. *AIAA-91-GT-1596*, 1991

[9] P. L. Roe, Approximative Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.*, **43**, 1981

[10] E. Turkel, Preconditioned Methods for Solving the Incompressible and Low-Speed Compressible Equations. *J. Comput. Phys.*, **72**, 1987

[11] D. L. Darmofal and K. Siu, A robust multigrid algorithm for the Euler equations with local preconditioning and semi-coarsing. *J. Comput. Phys.*, **151**, 1999

[12] J. M. Weiss and W. A. Smith, Preconditioning applied to variable and constant density flows. *AIAA Journal*, **33**

[13] Y.-H. Choi and C. L. Merkle, The application of preconditioning in viscous flows. *J. Comput. Phys.*, **105**, 1993

[14] S. Abarbanel and D. Gottlieb, Optimal time splitting for two- and three-dimensional Navier-Stokes equations with mixed derivatives. *J. Comput. Phys.*, **41**, 1981

[15] D. L. Darmofal and P. J. Schmid, The importance of eigenvectors for local preconditioners of the Euler equations. *J. Comput. Phys.*, **127**, 1996

[16] E. Erturk, Discussions on driven cavity flow. *Int. J. Numer. Meth. Fl.*, **60**, 275–294, 2008

[17] P. L. Roe, Approximative Riemann solvers, parameter vectors and difference schemes. *J. Comput. Phys.*, **43**, 357–372, 1981

[18] U. Ghia, K. N. Ghia and C. T. Shin, High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid methods. *J. Comput. Phys.*, **48**, 387–411, 1982

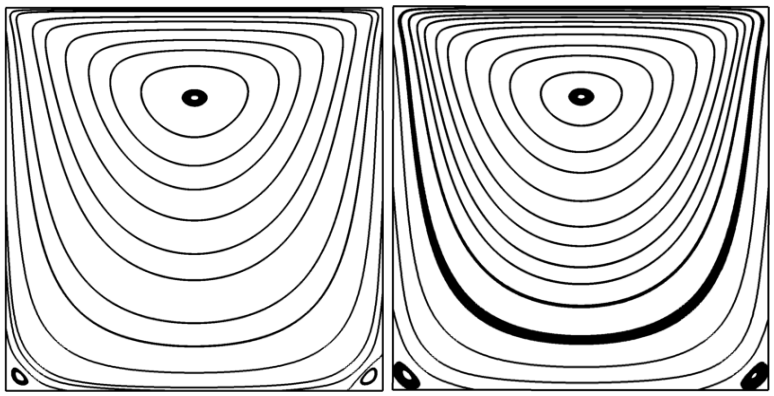


Figure 1: Streamlines in non-preconditioned (left) and preconditioned (right) calculations of a lid-driven cavity configuration for $\mathrm{Re}_1 = 1$, CFL = 25 and VNN = 0.

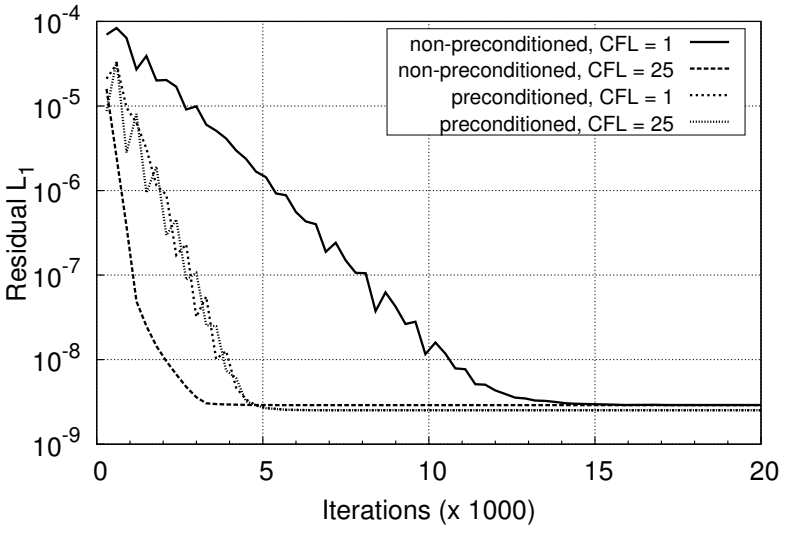



Figure 2: $L_1$ Residual for $\mathrm{Re}_1 = 1$ obtained with CFL = 1, 25 and VNN = 0.

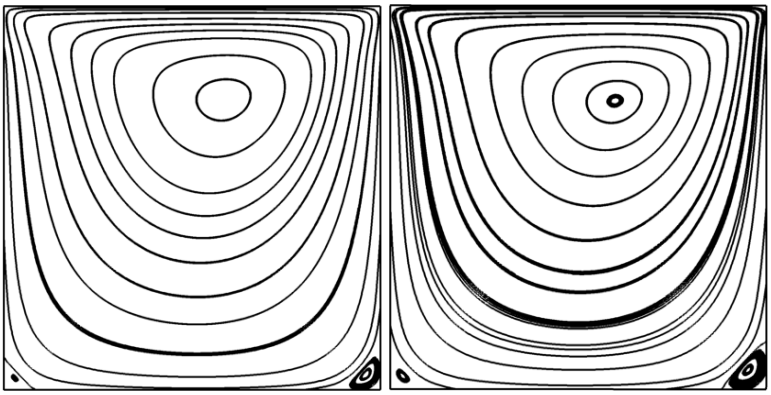


Figure 3: Streamlines in non-preconditioned (left) and preconditioned (right) calculations of a lid-driven cavity configuration for $\mathrm{Re}_1 = 100$, CFL = 25 and VNN = 0.

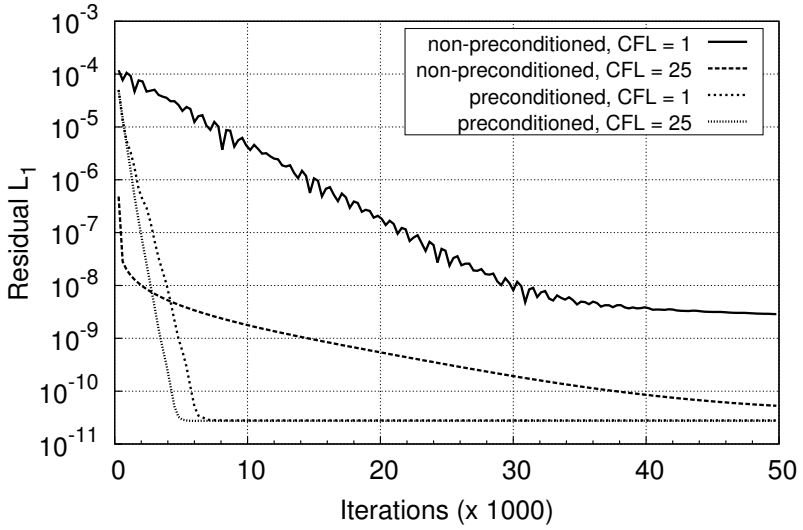



Figure 4: $L_1$ Residual for $\mathrm{Re}_1 = 100$ obtained with CFL = 1, 25 and VNN = 0.

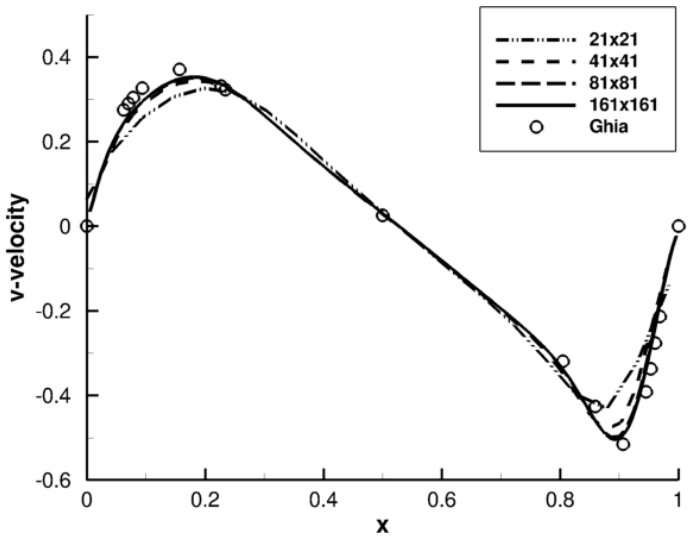



Figure 5: V-velocity profiles of a low-Mach preconditioned calculation along a horizontal line passing through the centre of the cavity at $Re_1 = 1000$, CFL = 25 and VNN = 0.

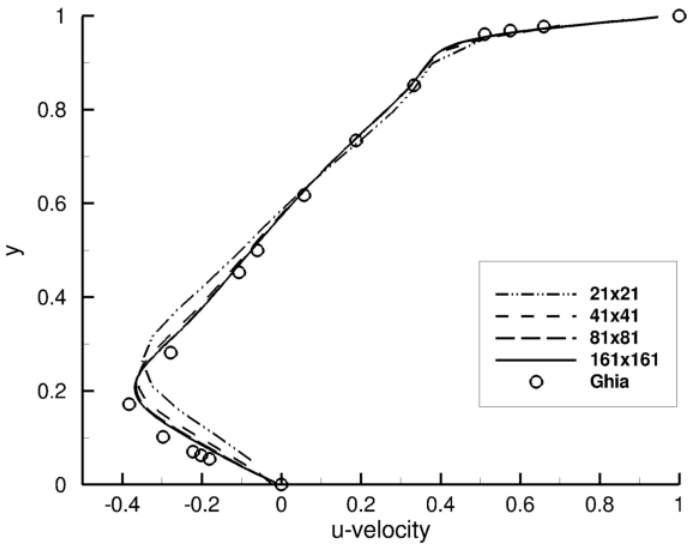
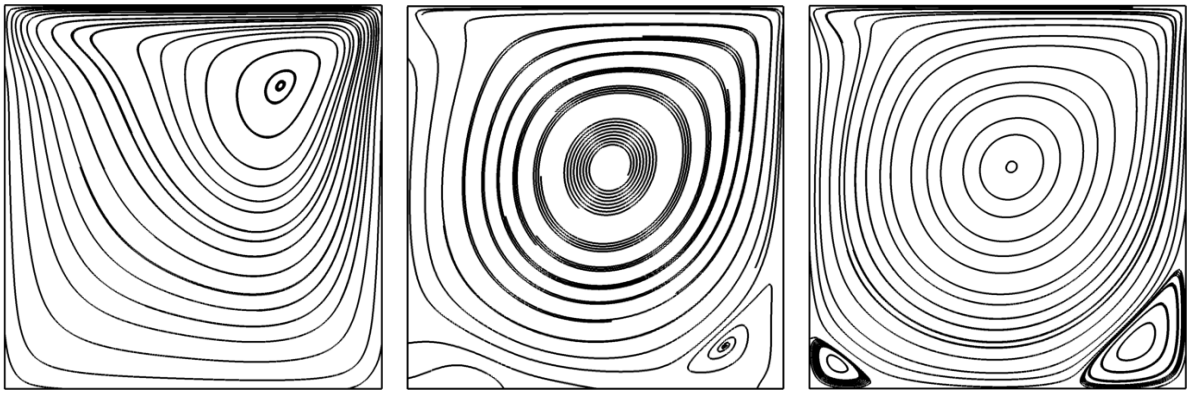



Figure 6: U-velocity profiles of a low-Mach preconditioned calculation along a vertical line passing through the centre of the cavity at $Re_1 = 1000$, CFL = 25 and VNN = 0.



(a) (b) (c)

Figure 7: Streamlines in a non-preconditioned (7a), a preconditioned calculation without (7b) and with low Reynolds modification (7c) of $\beta^2$ using a symmetric Gauss-Seidel solution scheme (SGS) for $\mathrm{Re}_1$ = 1000, CFL = 25 and VNN = 0.

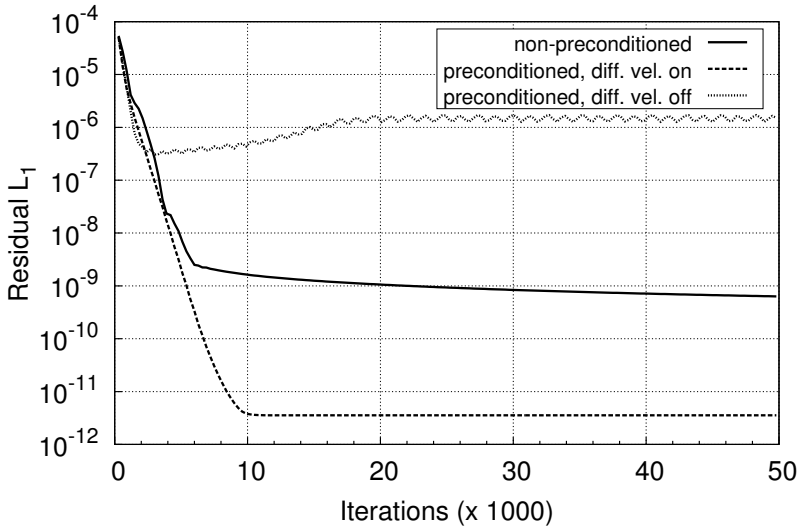



Figure 8: Influence of $M_{vis}$ on the numerical stability of a symmetric Gauss-Seidel solution scheme (SGS) for $\mathrm{Re}_1$ = 1000, CFL = 25 and VNN = 0.

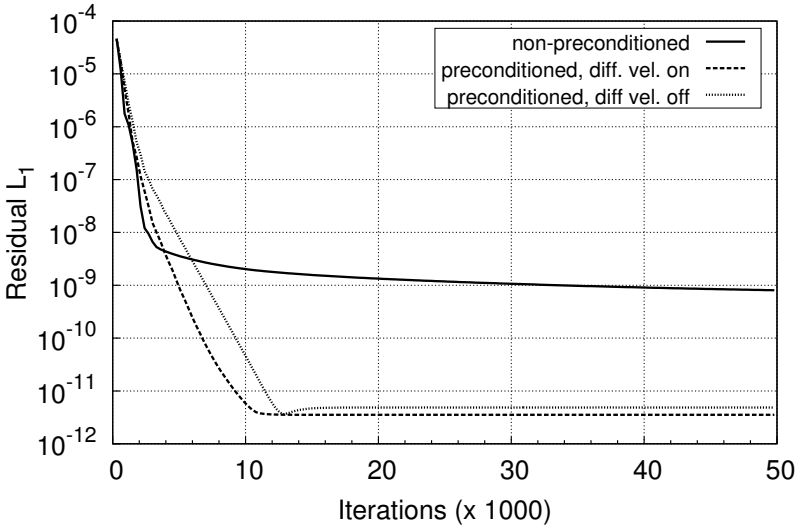



Figure 9: Influence of $M_{vis}$ on the numerical stability of a low Mach preconditioned predictor-corrector Gauss-Seidel method (PC-SGS) for $\mathrm{Re}_1$ = 1000, CFL = 25 and VNN = 0.

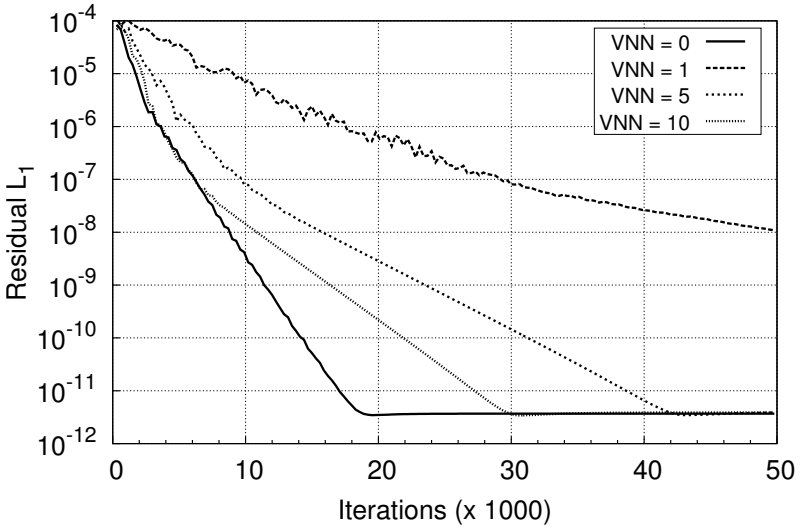



Figure 10: $L_1$ residual at increasing VNN of a low Mach preconditioned predictor-corrector Gauss-Seidel method (PC-SGS) at CFL = 1 and $Re_1 = 1000$.

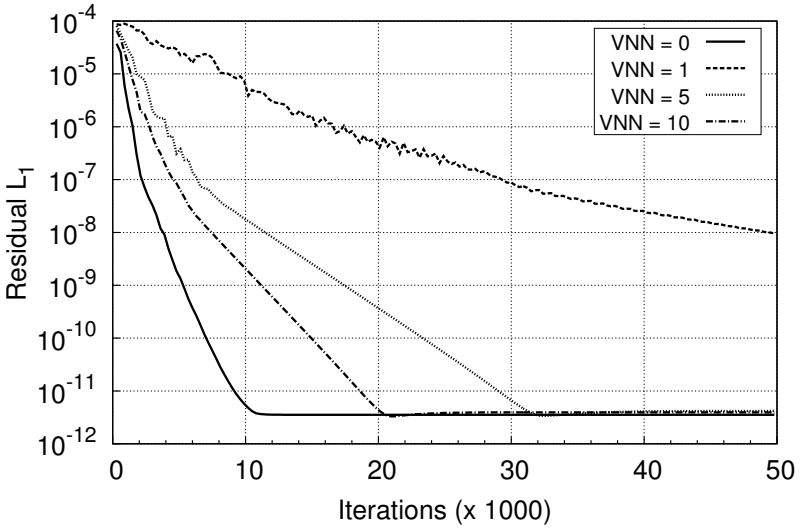



Figure 11: $L_1$ residual at increasing VNN of a low Mach preconditioned predictor-corrector Gauss-Seidel method (PC-SGS) at CFL = 10 and $Re_1 = 1000$.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# PERFORMANCE IMPROVEMENT OF FLOW COMPUTATIONS WITH AN OVERSET-GRID METHOD INCLUDING BODY MOTIONS USING A FULL MULTIGRID METHOD

**Kunihide Ohashi**[1], **Hiroshi Kobayashi**[2]

[1]National Maritime Research Institute
6-38-1 Shinkawa Mitaka Tokyo Japan
e-mail: k-ohashi@nmri.go.jp

[2]National Maritime Research Institute
6-38-1 Shinkawa Mitaka Tokyo Japan
e-mail: hiroshi@nmri.go.jp

**Keywords:** Overset-Grid Method, Full Multigrid Method, 3D RANS, Free Surface

**Abstract.** *A full multigrid method with an overset-grid method is developed to get a fast convergence. The weight values for an overstet-grid interpolation of flow variables can be determined on the fine grid level of multigrid, however, the weight value might be difficult to determine on coarser grid level. Therefore, the receptor cell value on coarser grid level is set and kept by the interpolation and correction of the multigrid method. Then, the flow variables on the coarse grid is interpolated to the fine grid as initial flows of the full multigrid method. The present method is applied to steady and unsteady simulations. The elapased time of steady simulations with and without free surafce is reduced about 35% of the time without the full multigrid method. The present method also succeeds to reduce the computational time remarkably on the unsteady simulation.*

# 1 INTRODUCTION

An overset-grid method can treat complex geometries by assembling the computational grid around each geometry and generating the overset information to combine the all grids. Additionally, the overset-grid method has a lot of flexibility in design processes, various appended shapes can be computed without generating the all computational grid just by replacing the computational grid around an appendage and generating the overset information.

The overset relation is generally composed on the certain division number of grid points, and performance improvement method to reduce a computational time plays important role with increasing the number of grid points.

The full multigrid method(FMG) is efficient to get fast convergence. A flow computation is started from the coarsest grid on the multigrid level, then the flow variables are interpolated to the finer grid as the initial flow. The interpolation cycle is continued until the finest grid level. The weight values for the overset-grid interpolation of flow variables can be determined on the finest grid of multigrid level, however, the weight value might be difficult to determine on coarser grid level. Therefore, the flow variables of the coarse grid are interpolated to the finest grid, then values are updated by using the overset interpolation on the finest grid. After the interpolation, the flow variables on the finest grid level are transformed to the coarser grid level and maintained the values. The computation on the coarse grid level is continued with using the overset interpolated values. Then, the flow variables on the coarse grid are interpolated to finer grid level as the initial flow of the full multigrid method.

The present method is applied to steady simulations w/ and w/o a free surface with several appendages to a ship hull. Then, the method is applied to an unsteady simulation including a body motion. Through the comparisons of results with and without FMG, the effectiveness of the present method is examined.

# 2 COMPUTATIONAL METHOD

## 2.1 Base solver

An in-house structured CFD solver[1] is employed. The governing equation is 3D RANS equation for incompressible flows. Artificial compressibility approach is used for the velocity-pressure coupling. Spatial discretization is based on a finite-volume method. A cell centered layout is adopted in which flow variables are defined at the centroid of each cell and a control volume is a cell itself. Inviscid fluxes are evaluated by the third-order upwind scheme based on the flux-difference splitting of Roe. The evaluation of viscous fluxes is second-order accurate. For unsteady flow simulations, a dual time stepping approach is used in order to recover incompressibility at each time step. It is consisted from the second order two-step backward scheme for the physical time stepping and the first order Euler implicit scheme for the pseudo time. The linear equation system is solved by the symmetric Gauss-Seidel (SGS) method.

For free surface treatment, an interface capturing method with a single phase level set approach is employed.

Body motions are taken into account by a moving grid technique. Grid velocities are contained in the inviscid terms to satisfy the geometrical conservation law. The grid velocities are derived from the volume where an each cell face sweeps. The boundary condition on a body is given as the velocities of the body motion.

### 2.2 Overset-grid method

The weight values for the overset-grid interpolation are determined by an in-house system[2]. The detail of the system can be found on [2], the summary is described.

1. The priority of the computational grid is set.
2. The cells of a lower priority grid and inside a body is identified (called as in-wall cell in here).
3. Receptors cells which the flow variables have to be interpolated from donor cells are defined. Two cells on a higher priority grid and facing to the outer boundary are set as receptor cells to satisfy the third order discretization of NS solver. Additionally, two cells neighborhood of in-wall cells, the cells of a lower priority grid and inside the domain of a higher priority grid are also set as the receptor cell.
4. The weight values for the overset interpolation are determined by solving the inverse problem based on Ferguson spline interpolation.

Flow variables of the receptor cell are updated when the boundary condition is set. The forces and moments are integrated on the higher priority grid to eliminate the lapped region on body surfaces. At first, the cell face of the lower priority grid is divided into small pieces. Secondly, the small piece is projected to the cell face of the higher priority grid by using the normal vector of the higher priority face. Then the 2D solid angle is computed and the small piece is decided in or out of the higher priority face. Once the small piece is in the higher priority face, the area ratio of the piece is set to zero. Finally, the area ratio is integrated on the lower priority face, then we have the ratio to integrate the forces and moments on lower priority face.

### 2.3 Full multigrid method

The multigrid method can obtain a fast convergence by reducing the residual which has the large wave length on a coarse grid. The agglomerate type is applied to the present method. A coarse grid is defined two cells in each $(i, j, k)$ direction, thus one cell on the coarse grid is constructed by eight cells on the fine grid.

The solution $\boldsymbol{q}_l$ on the fine grid level($l$) is transferred to the coarse grid level($l+1$) as follows.

$$\boldsymbol{q}_{l+1} = T^l_{l+1}\boldsymbol{q}_l \tag{1}$$

where $T^l_{l+1}$ is defined as

$$T^l_{l+1}\boldsymbol{q}_l = (\Sigma \boldsymbol{q}_l V_l)/V_{l+1} \tag{2}$$

The equation to be solved on the coarse grid level is expressed as

$$\frac{\partial \boldsymbol{q}_{l+1} V_{l+1}}{\partial \tau} = -\boldsymbol{R}_{l+1}(\boldsymbol{q}_{l+1}) - \boldsymbol{P}_{l+1} \tag{3}$$

where $\boldsymbol{R_{l+1}}$ is the residual on the coarse grid, and the forcing term $\boldsymbol{P}_{l+1}$ is given as follows.

$$\boldsymbol{P}_{l+1} = Q^{l+1}_l \boldsymbol{R}_l(\boldsymbol{q}_l) - \boldsymbol{R}_{l+1}(\boldsymbol{q}_{l+1}) \tag{4}$$

where $Q_l^{l+1}$ is transfer operator for the residual and is defined by the sum of eight cells on the fine grid.

$$Q_l^{l+1}\boldsymbol{R}_l(\boldsymbol{q}_l) = \Sigma(\boldsymbol{R}_l) \tag{5}$$

The solution on the coarse grid is expressed as $\boldsymbol{q}_{l+1}^+$, then the correction to the fine grid level is defined as

$$\boldsymbol{q}_l^+ = \boldsymbol{q}_l + I_{l+1}^l(\boldsymbol{q}_{l+1}^+ - \boldsymbol{q}_{l+1}) \tag{6}$$

where $I_{l+1}^l$ is an interpolation operator for a correction.

The computation starts from the coarsest grid on FMG, then the the solution is interpolated to the finer grid as the initial flows, and FMG achieves the fast convergence. An overset relation can be determined on the finest grid, however the relation may be difficult to construct on a coarse grid level. An additional method is required to apply FMG for the overset-grid. While one solving the equation on a coarse grid level, the solution interpolation and transfer of multi-grid method are made in a specific computational step. The receptor cell value is updated on the finest grid level, then transferred to coarser grid level. The present method can take account the receptor value on coarser grid level. The influence of a specific computational step(SCS) is also investigated in the present research.

The cell flag to be solved or unsolved on a coarse grid level is determined by using the cell flag on a fine grid level. An unsolved cell flag is included in the cell on a coarse grid, the cell is flagged as an unsolved cell(Fig.1). And furthermore, a cell flag on the coarser grid level is determined in the same way. Once the receptor cell value on the finest grid is included in coarser grid level, the correction to the finer grid from a coarse grid is treated as zero.

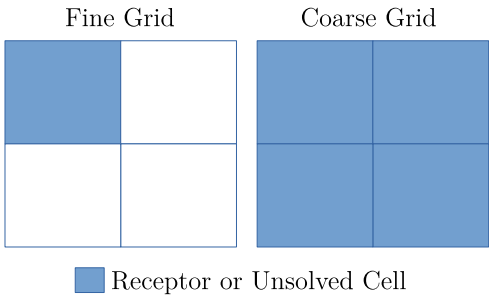



Figure 1: Flagging of cells with multigrid level

## 3 COMPUTED RESULTS

### 3.1 Steady double model flow

The FMG and overset-grid methods are applied to the flows around a main hull with several appendages. Reynolds number is $3.0 \times 10^6$ and EASM[3] is applied to compute Reynolds stresses. Table 1 shows the division number of computational grids in each direction. The grids are arranged with the priority of the overset interpolation. As appendages, two bilge keels, two aft hull fins, one duct which is situated just before a propeller, and a rudder are arranged with refined rectangular grids which resolve the flows around the aft fins and the rudder. The computational grid are consisted from 10 grids with about 3.0 millions cells.

Table 1: Division number of computational grid

| Grid | IM×JM×KM |
|---|---|
| Bilge keel×2 | 33×65×57 |
| Aft fin×2 | 49×65×41 |
| Rect. fin1,2 | 53×33×65 |
| Duct | 61×61×37 |
| Rudder | 101×81×73 |
| Rect. | 101×81×73 |
| Hull | 113×121×77 |

Table 2: Number of solved cell

| Grid | Fine | | Medium | | Coarse | |
|---|---|---|---|---|---|---|
| | total | solved | total | solved | total | solved |
| Bilge keel×2 | 229376 | 69632 | 28672 | 8704 | 3584 | 896 |
| Aft fin×2 | 245760 | 116736 | 30720 | 14592 | 3840 | 1536 |
| Rect. fin1 | 212992 | 65901 | 26624 | 7878 | 3328 | 1 |
| Rect. fin2 | 212992 | 66044 | 26624 | 7889 | 3328 | 0 |
| Duct | 129600 | 124936 | 16200 | 15436 | 2025 | 1440 |
| Rudder | 576000 | 297152 | 72000 | 36996 | 9000 | 3894 |
| Rect. | 576000 | 269380 | 72000 | 31412 | 9000 | 242 |
| Hull | 1021440 | 908720 | 127680 | 111726 | 15960 | 11526 |

Three levels multigrid method is applied. Table 2 shows the total and solved cell number with the multigrid level. The cell number to be solved becomes zero on the refined rectangular grid at the coarse grid level, hence the full multigrid starts with the medium grid.

Fig.2 shows the time history of L2-norm of the pressure with changing the SCS from 5 to 40. Fig.3, 4 show the comparisons of the total resistance coefficient and the elapsed time. The total resistance coefficient is non-dimensionalized by $\frac{1}{2}\rho U_0^2 L_0^2$. The spikes of L2-norm are existed when the interpolation and the correction of the multigrid are made with SCS intervals. The difference of L2-norm between SCS=5 and 10 is small, while L2-norm becomes larger after 301 time steps where the computational grid changes to the fine grid with using SCS=20 and 40. The total resistance coefficient shows small difference with the SCS change until 300 time steps, while larger fluctuation with SCS=20 and 40 appears after 301 time steps. From the above results, SCS is determined as 10 from hereafter.

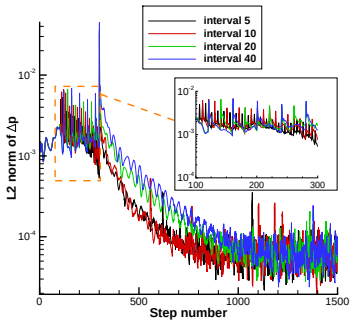



Figure 2: Time history of L2 norm($\Delta p$)

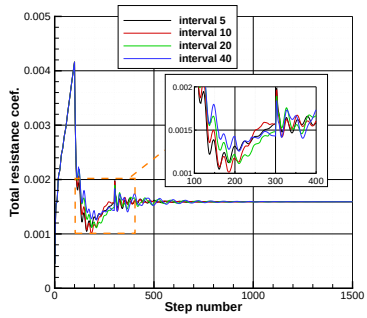



Figure 3: Convergence history of total resistance coefficient

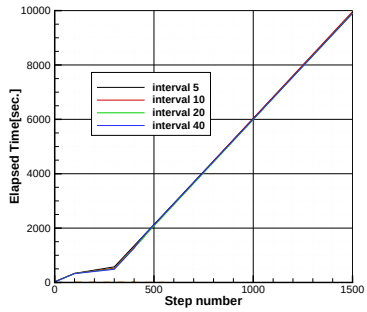



Figure 4: Comparison of elapsed time

From Fig.5 to Fig.7 show the comparisons of the results with and without FMG. The time history of L2-norm of FMG has the spike on 301 time steps when the computational grid shifts from the medium level to the fine level(Fig.5). FMG takes few times until 300 time steps with the computation on the medium grid, then the slope of FMG becomes same as the result without FMG(Fig.7). The steps from 500 to 800 where the total resistance coefficient almost converges are magnified in Fig.6, and the elapsed time at 650 steps where the total resistance coefficient converges within $\pm 0.5\%$ from the averaged value is compared. FMG takes 3271 seconds, while the multigrid(without FMG) takes 5118 seconds, thus FMG succeeds to reduce the computational time about 37% in this case.

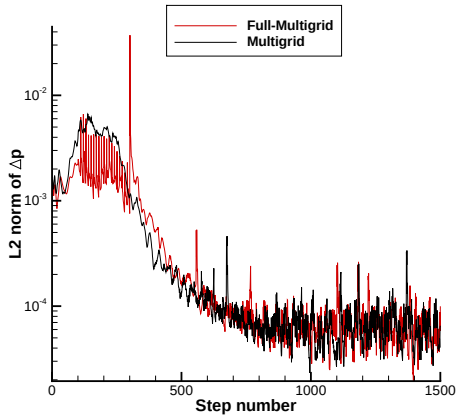



Figure 5: Comparison of L2 norm($\Delta p$)

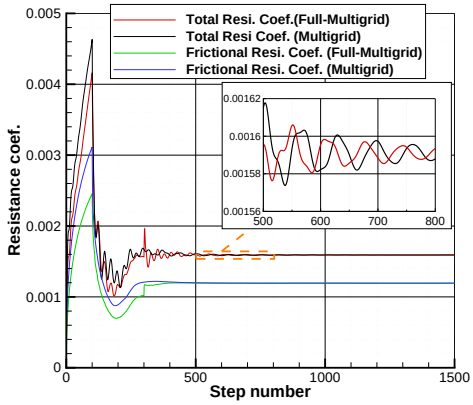



Figure 6: Convergence history of resistance coefficient

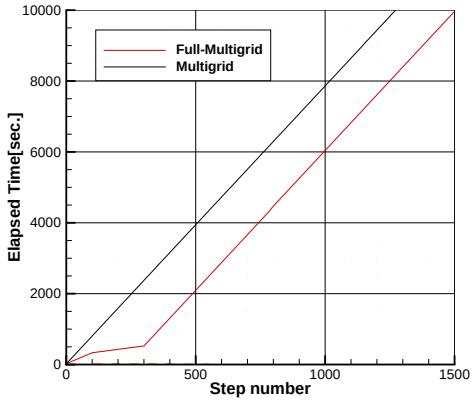



Figure 7: Comparison of elapsed time

### 3.2 Steady free surface flow

The FMG method is applied to the free surface flow with the overset-grid. The computational grid is consisted from 5.07 million cells with the hull grid, the rudder grid and two rectangular grids. Reynolds number is $1.074 \times 10^7$, Froude number is 0.261 and $k - \omega SST$[4] model is applied. Table 3 shows the division number of computational grids in each direction.

Table 3: Division number of computational grid

| Grid | IM×JM×KM |
|---|---|
| Rudder | 113×121×77 |
| Rect. Rud. | 49 × 49 × 45 |
| Hull | 113×121×77 |
| Rect. | 113×121×77 |

Table 4 shows the total and solved cell number with the multigrid level. The cell number to be solved remains on the coarse grid level, hence FMG starts from the coarse grid level in this case. The computation on the coarse grid level continues until 500 steps, then the computation

Table 4: Number of solved cell

| Grid | Fine | | Medium | | Coarse | |
|---|---|---|---|---|---|---|
| | total | solved | total | solved | total | solved |
| Rudder | 135168 | 126710 | 16896 | 15836 | 2112 | 1584 |
| Rect. Rud. | 101376 | 70660 | 12672 | 8606 | 1584 | 116 |
| Hull | 3276800 | 3149958 | 409600 | 392792 | 51200 | 43944 |
| Rect. | 1554432 | 1329118 | 194304 | 164464 | 24288 | 18160 |

continues on the medium grid, finally the computation on the fine grid level starts from 1001 steps.

Fig.8 shows the comparison of L2-norm with and without FMG. The time history of L2-norm of FMG has the spikes at 501 and 1001 time steps when the computational grid shifts from the coarse and medium grid levels to finer grid level. The total resistance coefficient also shows small diffraction when the grid level moves to finer grid level(Fig.9), then converges rapidly. The elapsed time of FMG is relatively small until the grid level reaches the fine grid, then the slope becomes same as the result without FMG(Fig.10). The steps from 1600 to 2300 where the total resistance coefficient almost converges are magnified in Fig.9. The elapsed time at 2200 steps with FMG where the total resistance coefficient converges is compared with the elapsed time without FMG at 1800 steps. FMG takes 34142 seconds, while the multigrid(without FMG) takes 52624 seconds, thus FMG succeeds to reduce the computational time about 35%.

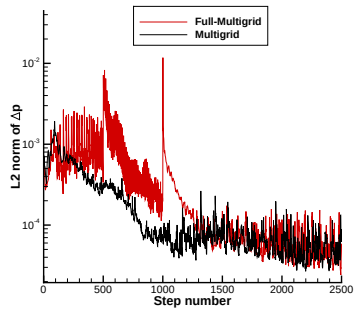



Figure 8: Comparison of L2 norm($\Delta p$)

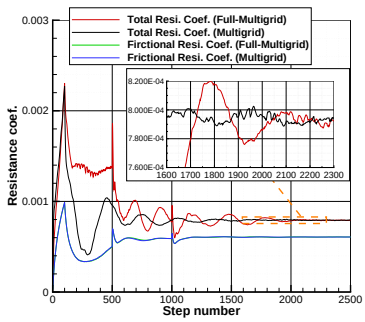



Figure 9: Convergence history of total resistance coefficient

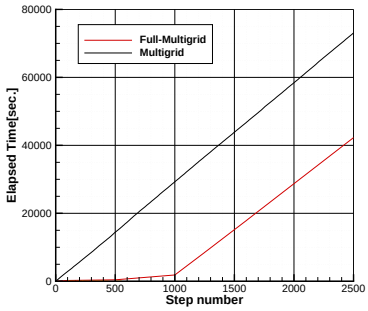



Figure 10: Comparison of elapsed time

### 3.3 Unsteady simulation with body motion

The FMG method is applied to the unsteady flow computation with a prescribed rotational motion. The computation of a marine propeller with a fixed revolution number is selected. All grids rotate around the x-axis with the revolution number. Reynolds number is $4.649 \times 10^5$ and $k - \omega SST$[4] model is applied. The non-dimensional revolution number is 2.0. Table 5 shows the division number of computational grids in each direction. The computational grid is consisted from 2.09 million cells with the the five blade grids and the cylinder grid on the background.

Table 5: Division number of computational grid

| Grid | IM×JM×KM |
|---|---|
| Blade | 49×65×33 |
| Cylinder | 117×145×97 |

Table 6 shows the total and solved cell number with the multigrid level. The cell number to be solved remains on the coarse grid level, hence FMG starts from the coarse grid level. The computation on the coarse grid level continues until 500 steps, then the computation continues on the medium grid level, finally the computation on the fine grid level starts from 1001 steps.

Fig.11 shows the comparison of L2-norm with and without FMG. The non-dimensionalized physical time step with FMG is $3.0 \times 10^{-4}$, while the time step without FMG is $5.0 \times 10^{-5}$, thus the time step with FMG takes larger on the present unsteady simulation. L2-norm with FMG is slightly reduced with the computational steps, while the gradient of L2-norm without FMG is comparatively small. The resistance coefficient with FMG converges rapidly, on the other hand, the resistance coefficient without FMG remains the fluctuation(Fig.12). Fig.13 shows the comparison of the elapsed time with and without FMG. Due to the difference of the physical time step, the direct comparison is difficult but the elapsed time with FMG is remarkably reduced.

Fig.14 shows the pressure distribution on the body surfaces and the velocity distribution with vectors just behind the blades. The back side of the blade surfaces becomes negative pressure values due to the angle of attack which is caused by the inflow to the blades. The axial velocity is accelerated by the blades, hence the variables are smoothly interpolated and computed in each computational grid. Thus it can be said the present overset grid-method and FMG work well.

Table 6: Number of solved cell

| Grid | Fine | | Medium | | Coarse | |
|---|---|---|---|---|---|---|
| | total | solved | total | solved | total | solved |
| Blade1 | 98304 | 92160 | 12288 | 11520 | 1536 | 1152 |
| Blade2 | 98304 | 92160 | 12288 | 11520 | 1536 | 1152 |
| Blade3 | 98304 | 92160 | 12288 | 11520 | 1536 | 1152 |
| Blade4 | 98304 | 92160 | 12288 | 11520 | 1536 | 1152 |
| Blade5 | 98304 | 92160 | 12288 | 11520 | 1536 | 1152 |
| Cylinder | 1603584 | 1422870 | 200448 | 174881 | 25056 | 18866 |

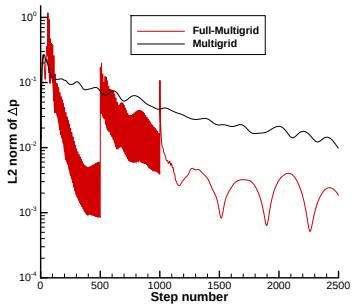



Figure 11: Comparison of L2 norm($\Delta p$)

## 4 CONCLUSIONS

- The full multigrid method to the overset-grid method has been developed.
- Present method succeeds to reduce the computational time about 35% in the steady computations with and without the free surface.
- Present method also can be applied to the unsteady simulation, and the elapsed time of the computation is remarkably reduced.

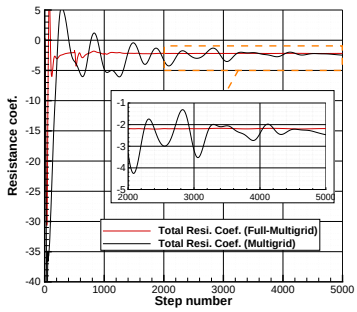



Figure 12: Convergence history of total resistance coefficient

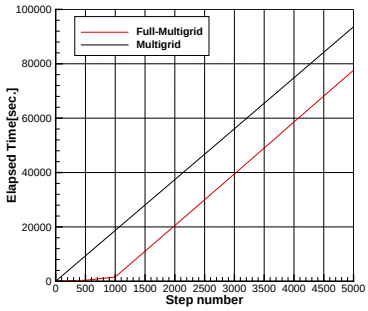



Figure 13: Comparison of elapsed time

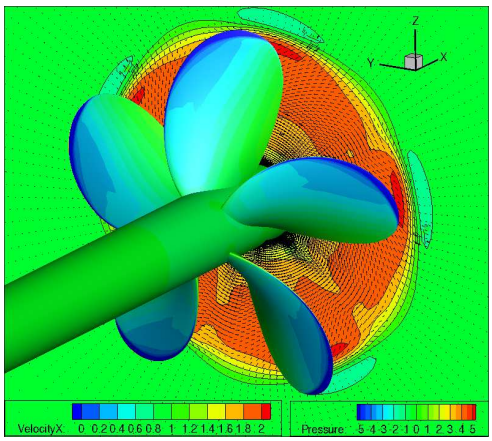



Figure 14: Pressure distribution on body surfaces and flow behind the blades

## REFERENCES

[1] K., Ohashi, T., Hino, N., Hirata, H., Kobayashi, Development of NS solver with a structured overset grid method *The 28th Computational Fluid Dynamics Symposium*, Japan, 2014.

[2] H., Kobayashi, Y., Kodama, Developing spline based overset grid assembling approach and application to unsteady flow around a moving body, *MARINE2015*, Roma, Italy, June 15-17, 2015.

[3] C.L., Rumsey, T.B., Gatski, Recent Turbulence Advances Applied to Multielement Airfoil Computations, *Journal of Aircraft*, vol.38(5), pp.904-910, 2001.

[4] F.R., Menter, Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications, *AIAA Journal*, vol.32(8), pp.1598-1605, 1994.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SOLVING LINEAR SYSTEMS WITH MULTIPLE RIGHT-HAND SIDES WITH GMRES: AN APPLICATION TO AIRCRAFT DESIGN

**Aloïs Bissuel[1], Grégoire Allaire[2], Laurent Daumas[1], Frédéric Chalot[1] and Michel Mallet[1]**

[1] Dassault Aviation
78, quai Marcel Dassault 92552 Saint-Cloud CEDEX 300 France
e-mail: {alois.bissuel, laurent.daumas, frederic.chalot, michel.mallet}@dassault-aviation.com

[2] CMAP
UMR 7641 École Polytechnique CNRS
Route de Saclay, 91128 Palaiseau CEDEX France
e-mail: allaire@cmap.polytechnique.fr

**Keywords:** Linearized Navier-Stokes, GMRes, Block GMRes, Finite Elements, HPC

**Abstract.** *To create efficient new aerodynamic designs or predict the onset of flutter, the linearised Navier-Stokes equations might be used. In some cases, many right-hand sides must be solved keeping the same matrix. In this paper, techniques which enable to solve several right-hand sides at the same time, such as Block* GMR*es, or reuse pieces of information computed in the previous solves, such as Krylov space recycling, are investigated. They will be tested on both simple and industrial test cases.*

# 1 INTRODUCTION

To create efficient new designs, aircraft manufacturers may resort to automatic optimisations tools relying on linearised Navier-Stokes solver [3]. Flutter analysis is another application where one can use time-domain linearisation of the Navier-Stokes equations [2]. In both cases, one may have to solve the same linear system with different right-hand sides, be it one for each geometric variable (direct method) or for each cost function (adjoint method) in the case of optimization, or one for each structural eigenmode in the case of flutter analysis. Due to their size and sparsity, these linear systems are solved using iterative methods such as GMRes [14].

Extensions of the GMRes method to multiple right-hand sides were consequently investigated. They can be grouped in two main classes. The first one is based on reusing information from the first solve to speed-up convergence of the subsequent systems. Using a projector, spectral information from the first solve can spectrally deflate at lower cost a GMRes method [10]. It is expected to attain near asymptotic convergence right from the beginning, removing transient plateaus. Another method could be to project all the right hand-sides on the Krylov space generated during the GMRes solve of the first vector to make these converge at least partially, lowering the number of iterations necessary for their complete solving.

The second class of extension of GMRes to several right-hand sides is to solve them all at once, using block Krylov spaces [5]. To improve the convergence when using small Krylov spaces with restarts, spectral deflation was implemented [9]. Decreased computation time is expected to come from both an accelerated convergence and a reduced parallel communication time due to block communications.

These methods will be tested on 2D and 3D examples, ranging from one hundred thousand to thirty million unknowns, on parallel architectures. Both convergence and timing results will be presented.

# 2 AETHER

## 2.1 Compressible Navier-Stokes code

AETHER is a compressible Navier-Stokes code developed in-house by Dassault Aviation. It is a finite elements solver on unstructured mesh, stabilized by the SUPG method [7]. To enable better properties of the implicit matrices, the compressible Navier-Stokes equations are written in entropic variables, so that they become symmetric. Several turbulence models are available (Spalart-Allmaras, k-$\epsilon$, k-$\omega$, K-KL, *etc*).The non-linear code is fully implicit, using GMRes to solve the Newton iterations. The code is made parallel using MPI as message passing library.

## 2.2 Linearised Navier-Stokes equations

For aeroelasticity problems, a simple way to understand the linearised Navier-Stokes equations is to consider the non-linear ones in a reduced form:

$$E(\mathbf{V}, \mathbf{x}) = 0 \tag{1}$$

Here, $\mathbf{V}$ is the physical state of the systems (*i.e.* the flow variables), and $\mathbf{x}$ corresponds to coordinates of the aeroplane. If the surface of the aircraft is slightly modified by a small displacement $\delta\mathbf{x}$, for instance a twist, then the flow variables will change accordingly by an increment $\delta\mathbf{V}$:

$$E\left(\mathbf{V} + \delta\mathbf{V}, \mathbf{x} + \delta\mathbf{x}\right) = 0 \tag{2}$$

A linear expansion of this equation, remembering that the base flow state verifies (1), yields

$$\frac{\partial E}{\partial \mathbf{V}}\delta\mathbf{V} = -\frac{\partial E}{\partial \mathbf{x}}\delta\mathbf{x} \tag{3}$$

Thus, if a small known displacement $\delta\mathbf{x}$ is applied to the aircraft, solving this linear system enables to know the variation of the flow variables $\delta\mathbf{V}$. If a harmonic movement needs to be applied to the structure, complex numbers are used. For details, see [2]. When doing aerodynamic optimisation, either a direct approach (comparable to the derivation in real number made before) or a adjoint one can be used, as explained in [3].

## 3 BLOCK GMRES

When solving for several right-hand sides, it might be assumed that searching iteratively the solution in several directions at the same time could be advantageous. Herein lies the main idea of the block GMRes method.

### 3.1 Standard block GMRes

Block GMRes is simply a generalisation of the standard iterative GMRes method [14] for vector with several columns. We want to solve the following linear problem :

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \text{with} \quad \mathbf{A} \in \mathbb{R}^{N\times N}, \quad \mathbf{b} \in \mathbb{R}^{N\times s}, \quad \mathbf{x} \in \mathbb{R}^{N\times s} \tag{4}$$

Here, $N$ is the dimension of the linear system, and $s$ is the number of right-hand side we want to solve simultaneously. Krylov spaces, which are necessary to GMRes, can be extended to a block definition.

$$\mathcal{B}_n(\mathbf{A}, \mathbf{y}) = \text{block span}\left(\mathbf{y}, \mathbf{A}\mathbf{y}, \ldots, \mathbf{A}^{n-1}\mathbf{y}\right) \tag{5}$$

Block span means that the linear combinations of the individual columns of the block vector are taken into account, i.e.

$$\forall \mathbf{u} \in \mathcal{B}_n(\mathbf{A}, \mathbf{y}), \exists\, (\boldsymbol{\gamma}_k) \in \mathbb{R}^{s\times s} \quad \mathbf{u} = \sum_{k=0}^{n-1} \mathbf{A}^k \mathbf{y} \boldsymbol{\gamma}_k$$

Starting from an initial guess $\mathbf{x}_0$, a block Krylov method for solving the linear system (4) finds an approximate solution $\mathbf{x}_n$ such that

$$\mathbf{x}_n - \mathbf{x}_0 \in \mathcal{B}_n(\mathbf{A}, \mathbf{r}_0) \tag{6}$$

with $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ being the initial residual.

The algorithm of the block GMRes is presented in [5]. Its implementation is summarised in algorithm 1. Compared to standard GMRes, there is little conceptual difference. Only the normalisations are transformed into QR-factorisations and the Hessenberg matrix is defined by blocks.

The Hessenberg matrix $\overline{\mathbf{H}}_n = \left(\widetilde{\boldsymbol{\eta}}_{i,j}\right) \in \mathbb{R}^{n+1\times n}$ is upper triangular with one extra lower diagonal. To solve the least-square problem with only a simple triangular solve as in line 17 of algorithm 1, the matrix $\overline{\mathbf{H}}_n$ needs to be reduced to a purely upper triangular matrix. A complete QR decomposition would be both too costly and impractical to update, as $\overline{\mathbf{H}}_n$ is created column-by-column. As introduced in [6], Householder transformations can be put to use efficiently for this task. These transformations are applied to blocks of dimension $2s \times s$ to reduce them to

**Algorithm 1** Non-restarted Block GMREs with $m$ Krylov vectors

$\mathbf{v}_0\boldsymbol{\rho}_0 = QR(\mathbf{r}_0)$ ▷ QR factorisation of the initial residual
$\mathbf{q} \leftarrow \mathbf{e_1}\boldsymbol{\rho}_0$ ▷ Least-square problem right-hand side
**for** $n = 1,m$ **do**
  $\tilde{\mathbf{v}} \leftarrow \mathbf{A}\mathbf{v}_{n-1}$
  **for** $k = 0,n-1$ **do** ▷ Arnoldi orthonormalisation of the next vector
    $\widetilde{\boldsymbol{\eta}}_{k,n-1} \leftarrow \langle \mathbf{v}_k, \tilde{\mathbf{v}} \rangle$
    $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} - \mathbf{v}_k\widetilde{\boldsymbol{\eta}}_{k,n-1}$
  **end for**
  $\mathbf{v}_n\widetilde{\boldsymbol{\eta}}_{n,n-1} = QR(\tilde{\mathbf{v}})$
  Apply previous Householder transformations to $\left(\widetilde{\boldsymbol{\eta}}_{k,n}\right), k = 0, \ldots, n-2$
  Create and apply Householder transformation $\mathbf{Q}_n$ on $\begin{pmatrix} \widetilde{\boldsymbol{\eta}}_{n-1,n-1} \\ \widetilde{\boldsymbol{\eta}}_{n,n-1} \end{pmatrix}$
  Apply Householder transformation $\mathbf{Q}_n$ to $\mathbf{q}$
  Compute residual : norm of the last block of $\mathbf{q}$
  **If** residual lower than $\varepsilon$ **exit loop**
**end for**
Apply Householder transformations backwards to $\mathbf{q}$
$\mathbf{k}_n \leftarrow \mathbf{H}_n^{-1}\mathbf{q}$ ▷ Triangular solve
$\mathbf{x}_n \leftarrow \mathbf{x}_0 + \mathbf{V}_{n+1}\mathbf{k}_n$ ▷ New approximate solution
$\mathbf{r}_n \leftarrow \mathbf{V}_{n+1}\mathbf{k}_n$ ▷ New residual

upper triangular form. When a column is added to $\overline{\mathbf{H}}_n$, all previous Householder transformations need to be applied so as to get the modified $\boldsymbol{\eta}_{i,n}$. This transformed column is added as the last column of $\mathbf{H}_n$ which is upper triangular.

Restarting the block GMREs algorithm is simply a matter of restarting it with the new residual.

### 3.2 Spectral deflation of block GMREs

When solving a linear problem with the GMREs method, restarts are a necessity for memory reasons and computational time, as each new Krylov vector must be orthogonalised against all previous vectors. When restarted, Krylov methods have a known tendency to stall convergence. To alleviate these convergence problems, Morgan introduced in [8] a deflation of the smallest eigenvector of the matrix $\mathbf{A}$, which are hard to find during the iterations. Some numerical tricks are introduced by Röllin in [13]. This theory can be adapted to the block case.

The deflated restart is done in several steps as explained in algorithm 2. First of all, the harmonic Ritz vectors are computed, solving a standard eigenvalue problem and not a generalized one (see [13] for details). The $k$ lowest ones are stored and orthogonalized. To force the new iterations to explore directions yet unsearched, the residual of the previous iteration is added to this basis. The new Krylov basis and Hessenberg matrix are also created. Finally, the residual in the new Krylov basis has to be computed. The idea for this critical step came from [12] : $\mathbf{q}$ is simply filled with the orthonomalising coefficients created during the orthogonalisation of the residual against the harmonic Ritz vectors.

Deflation of the right-hand sides, which is a way to eliminate linear dependence between the right-hand sides [5], was not used for several reasons. First of all, it is hard to correctly implement it jointly with spectral deflation. At restarts with spectral deflation, the new starting Krylov space and Hessenberg matrix must follow the Arnoldi relation. Moreover, the residual is a

**Algorithm 2** Deflated restart of the block GMRes algorithm with $k$ harmonic Ritz vectors

1: Compute the $k$ first harmonic Ritz vectors $\mathbf{p}_i$ of $\mathbf{H}_n$
2: Orthonormalise $\mathbf{P}_k = (\mathbf{p}_i)_{i=1,k}$
3: Append $\mathbf{q}_n$ to $\mathbf{P}_k$ which becomes $\mathbf{P}_{k+1}$
4: Orthonormalise the last column of $\mathbf{P}_{k+1}$ against the first columns
5: Store the orthonormalisation coefficients in $\mathbf{q}$
6: $\mathbf{V}_{k+1}^{new} \leftarrow \mathbf{V}_{n+1}\mathbf{P}_{k+1}$
7: $\overline{\mathbf{H}}_k^{new} \leftarrow \mathbf{P}_{k+1}^T\overline{\mathbf{H}}_n\mathbf{P}_k$

linear combination of the harmonic Ritz vectors [13]. Deflation of the right-hand sides can delete and add (previously deleted) right-hand sides, and to enforce the aforementioned Arnoldi relation is difficult. To circumvent this problem, one could use the GCRO-DR algorithm [11], which will be introduced shortly below. The second reason is that a deflation of the right-hand sides was implemented (but not along spectral deflation), and no linear (near-)dependency could be found. As in aeroelasticity applications, the right-hand sides are quite similar, and the solutions have some sort of commonality between them, deflation was hoped to be quite powerful, but ultimately it was not found to be of any effect in our problems.

## 4 KRYLOV SPACE RECYCLING

Instead of solving all the right-hand sides at the same time, the information created during the solving of one right-hand side can be reused to solve all the other ones. Several methods can be devised, depending on which piece of information is to be recycled. When a single right-hand side is solved using the GMRes method, a Krylov space is generated, which in itself might be interesting to recycle. Moreover, if deflation is used, the smallest harmonic Ritz vectors are also computed. They are hard to find, and their approximation might take several restarts to be correct, so they might be of higher value than a single Krylov space.

### 4.1 Harmonic Ritz vector reuse

A first way to effectively reuse the harmonic Ritz vectors created during a previous solve with a deflated GMRes method is GCRO-DR [11]. It consists in a GMRes loop nested within a GCR (Gradient Conjuguate Residual). The necessary orthogonality constraint for GCR is enforced through a carefully chosen projector, akin to deflation projectors. See [4] for an interesting comparison between deflation projectors and deflation as treated in section 3.2. GCRO-DR was put to use in [15] for linearised Navier-Stokes computational studies with some success. It was chosen not to implement this algorithm due to its complexity and its likeness to deflated GMRes (see [11]).

Instead, a simpler method was tried. As explained in [10], a projection onto the small harmonic Ritz vectors at each restart of a standard GMRes (*i.e.* non deflated) could be sufficient to attain near-asymptotic convergence right from the start, and also exhibit good deflation properties at restart (no significant change of the convergence rate). For the projection and the GMRes method to interweave properly, the projection has to be carefully crafted. The projection, dubbed MinRes in [10], is explained in algorithm 3.

The MinRes projection should be used in the following way. A first right-hand side is solved using deflated GMRes. At the last restart, the Krylov space and the Hessenberg matrix created by the deflation and corresponding to the smallest harmonic Ritz vectors are saved. The other right-hand sides are then solved : before each restart of a standard GMRes cycle, the residual and

**Algorithm 3** MinRes projection

**Require:** $\mathbf{x}_0$ approximate solution, $\mathbf{r}_0$ residual, $\mathbf{V}_{k+1}$ and $\overline{\mathbf{H}}_k$ verifying Arnoldi relation

1: $\mathbf{c} \leftarrow \mathbf{V}_{k+1}^T \mathbf{r}_0$
2: Solve $\min \|\mathbf{c} - \overline{\mathbf{H}}_k \mathbf{d}\|$ ▷ Simple triangular solve after upper-triangularising $\overline{\mathbf{H}}_k$
3: New solution is $\mathbf{x}_k \leftarrow \mathbf{x}_0 + \mathbf{V}_k \mathbf{d}$
4: New residual is $\mathbf{r}_k \leftarrow \mathbf{r}_0 - \mathbf{V}_{k+1} \overline{\mathbf{H}}_k \mathbf{d}$

the approximate solution are updated with the MinRes projection, using the saved Krylov space and Hessenberg matrix.

Another variant which was tested is to solve all the right-hand sides at the same time, one of them being solved with a deflated GMRes, and all the other being projected on the same Krylov space using the Minres projection. The idea is to leverage the similarities between the right-hand sides, and solving one may lower to a large extent the other residuals. Unfortunately, this technique proved unsuccessful in reducing with more than one order of magnitude the residuals of the other right-hand sides.

## 5 TEST CASES

These two techniques for solving linear systems with multiple right-hand sides will be tested on a two dimensional test-case, and a three-dimensional one, the latter being more representative of industrial problems.

### 5.1 2D test case: RAE2822

The first test case is a classic 2D wing profile. An illustration of the mesh used is shown in figure 1. This unstructured mesh has 35,000 nodes. A simulation on this test case has 140,000 unknowns.

This test case was used at a Mach number of 0.734 and an angle of attack of 3°. The base flow around which the linearised calculations are made does not exhibit any shocks or flow separation. Four movements of the profile were used for the purpose of creating four right-hand sides. The deformations corresponds in that order to a thickening, a pitching, a vertical displacement of the profile and an aileron deflection.

#### 5.1.1 Block GMRes

To test the efficiency of the block GMRes method, several combinations of the number of right-hand sides and of Krylov space were tried. The figure 2 indicates the results.

The test runs are sorted according to the order of the block Krylov space, which is equivalently the number of iterates with the matrix $\mathbf{A}$ . The parameter "p" indicates the number of right-hand sides solved at the same time. The dotting pattern is identical when solving the same number of right-hand sides. The smaller the dots, the lower the number of right-hand sides. The abscissa was chosen to be the number of matrix block-vector product (instead of matrix vector product). This enables to compare easily a block GMRes run with a standard one. For instance, if a solve with four right-hand sides takes less iterations than a solve with a single right-hand side (and that the four right-hand sides take roughly the same number of standard GMRes iterations to be solved), then solve the four of them at the same time is advantageous in terms of matrix-vector products.

The first comment to be made is that as a general rule, when a higher number of right-hand

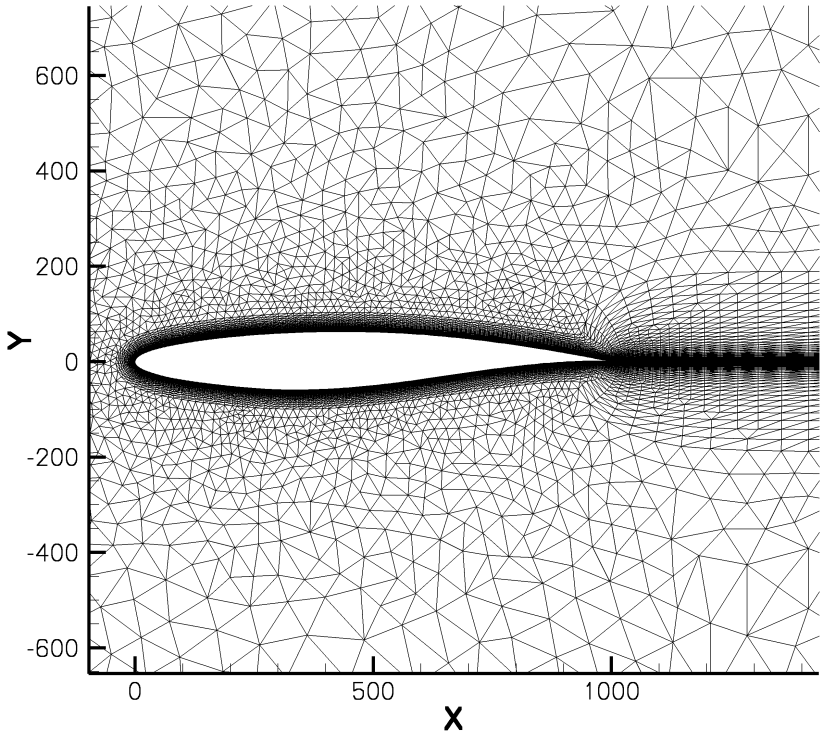



Figure 1: Mesh used for the 2D test case RAE2822

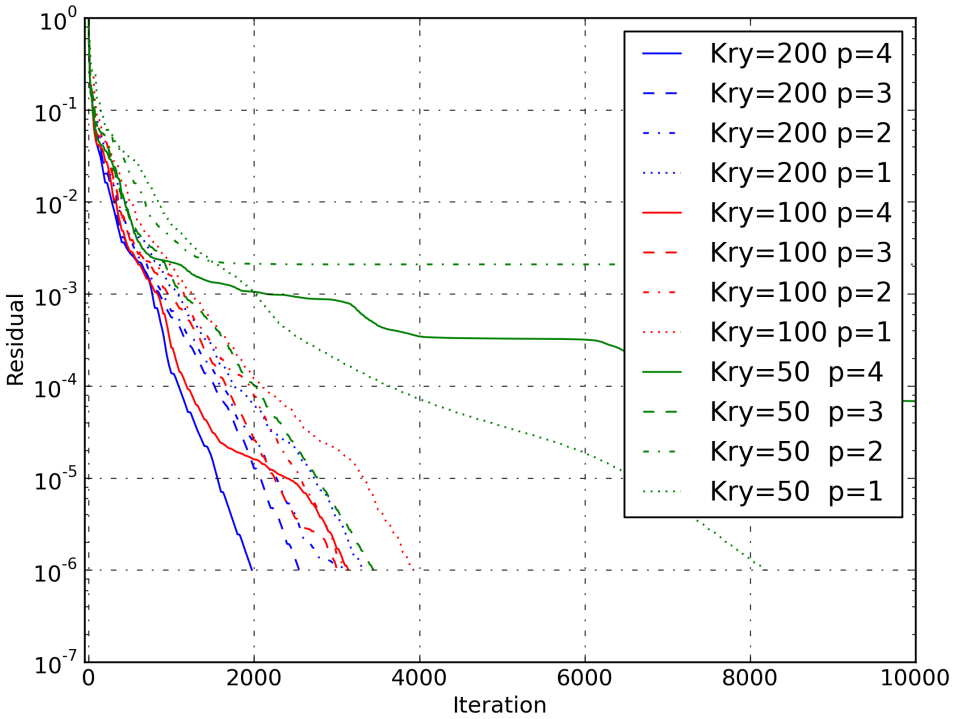



Figure 2: Convergence of the right-hand sides with block GMRes. In abscissa is the number of matrix block-vector products. “kry” is the order of the block Krylov space

| Krylov | 200 | 200 | 200 | 200 | 100 | 100 | 100 | 100 |
|---|---|---|---|---|---|---|---|---|
| p | 4 | 3 | 2 | 1 | 4 | 3 | 2 | 1 |
| Time (s) | 356 | 244 | 135 | 35 | 272 | 149 | 61 | 25 |
| Time/p (s) | 89 | 81 | 68 | 35 | 68 | 50 | 30 | 25 |
| Iterations | 1974 | 2546 | 3174 | 3315 | 3140 | 3021 | 3092 | 3931 |
| Iterations/s | 5.5 | 10.4 | 23.4 | 95.2 | 11.6 | 20.2 | 51.1 | 155.2 |
| Iterations*p/s = matvec/s | 22.2 | 31.3 | 46.9 | 95.2 | 46.2 | 60.7 | 102.2 | 155.2 |

| Krylov | 50 | 50 | 50 | 50 |
|---|---|---|---|---|
| p | 4 | 3 | 2 | 1 |
| Time (s) | — | 80 | — | 33 |
| Time/p (s) | — | 27 | — | 33 |
| Iterations | — | 3444 | — | 8185 |
| Iterations/s | — | 43.0 | — | 245.3 |
| Iterations*p/s = matvec/s | — | 128.9 | — | 245.3 |

Table 1: 2D test case: timing of block GMRes solve. Hyphen indicates a lack of convergence

sides are computed at the same time, the convergence is achieved in less iterations. This is verified for the solves with a block Krylov space of 200. It would have also been the case for the block Krylov space of 100, if it were not for the solve with four right-hand sides, whose convergence exhibits a strange slowing around 2000 iterations. Also, when the order of the block Krylov space is reduced, the robustness of the solve is decreased.For the Krylov space of order 50, with four and two right-hand sides, the convergence stalled for unexplained reasons.

Some timing details are provided in table 1. From these results, it can be inferred that block GMRes is slower than standard GMRes. The small convergence acceleration is not sufficient to overcome the the extra work of mainly orthonormalisation due to the larger number of vectors in memory.

To make a fairer comparison between block GMRes and its single right-hand side counterpart, table 2 shows a comparison this time grouped by equal memory footprint of the Krylov space. Here the block GMRes method is faster than the standard GMRes. First of all, the block GMRes needs less iterations to converge. Also, due to the blocking of the parallel communications and of the scalar products in the orthonormalisations, the matrix vector product is faster in block GMRes. Both of these effects combine to give a significant timing edge to the blocked Krylov method. This comparison is not realistic of industrial applications because of the large number of vectors used. Deflated GMRes has enabled Dassault Aviation to use quite small Krylov spaces without loss of robustness. Its blocked version is not yet robust enough.

| | 800 vectors | | | 400 vectors | | |
|---|---|---|---|---|---|---|
| Kry | 200 | 400 | 800 | 100 | 200 | 400 |
| p | 4 | 2 | 1 | 4 | 2 | 1 |
| Time (s) | 356 | 208 | 129 | 272 | 135 | 177 |
| Time/p (s) | 89 | 104 | 129 | 68 | 68 | 177 |
| Iterations | 1974 | 2336 | 2819 | 3140 | 3174 | 3023 |
| Iterations/s | 5.5 | 11.2 | 21.8 | 11.6 | 23.4 | 17.1 |
| Iterations*p/s = matvec/s | 22.2 | 22.5 | 21.8 | 46.2 | 46.9 | 17.1 |

Table 2: 2D test case: timing with equal memory footprint

#### 5.1.2 Krylov space recycling

Krylov space recycling was tested. The first right-hand side was solved using a standard deflated GMRes, so as to get the lowest harmonic Ritz vectors of the matrix. The convergence curve is shown on figure 3.

In this test case, the Krylov space size was of 200 vectors. The number of harmonic Ritz vectors which were kept was 60. The target residual was a normalized $10^{-5}$, meaning that the residual had to be 5 orders of magnitude lower for the computation to be stopped. From the second right-hand side onward, the system was solved using non-deflated GMRes with Minres projection. For the first three orders of magnitude, it is competitive with deflated GMRes, but then the convergence stalls (except for the second right-hand side for which it happens one order of magnitude lower). For clarity, the graph was trimmed at 12 000 iterations.

### 5.2 3D test case

The 3D test case is a composed of a dummy half-fuselage with a generic swept wing. The profile has a 10% thick symmetric airfoil section (NACA 64A010 profile). The wing has a 1m span. It has been tested previously in CFD and in wind tunnel [2]. A view of the half-model is shown on figure 4. The flow conditions are a Mach number of 0.88 and an angle of attack of 0°. The computational mesh totals more than 6 million nodes, which makes more than 30 million unknowns to be solved. This test case is deemed to be representative of industrial applications.

The movements corresponding to the right-hand sides are test polynomials on the wing of the coordinates $x$, $y$ and $z$ of increasing degree. Seven of them are shown on figure 5. They are a blend of a pitching, heaving, bending, twisting movement of the wing. They seem quite similar. It is thus expected that solving one may considerably help solving the others.

#### 5.2.1 Block GMRes

The block GMRes solver was tested on the 3D case. Ten right-hand sides were solved at the same time. Due to memory restrictions, we were forced to use a block Krylov space of order 50. The convergence curve is shown on figure 6. The convergence for the overall residual is presented with a bold line, and the residual of the single vectors are plotted with thin lines. After the 2000th iteration, the convergence is slowing down, and is completely stalled at the end of the computation. The residual failed to be lowered by 5 orders of magnitude, which is by experience

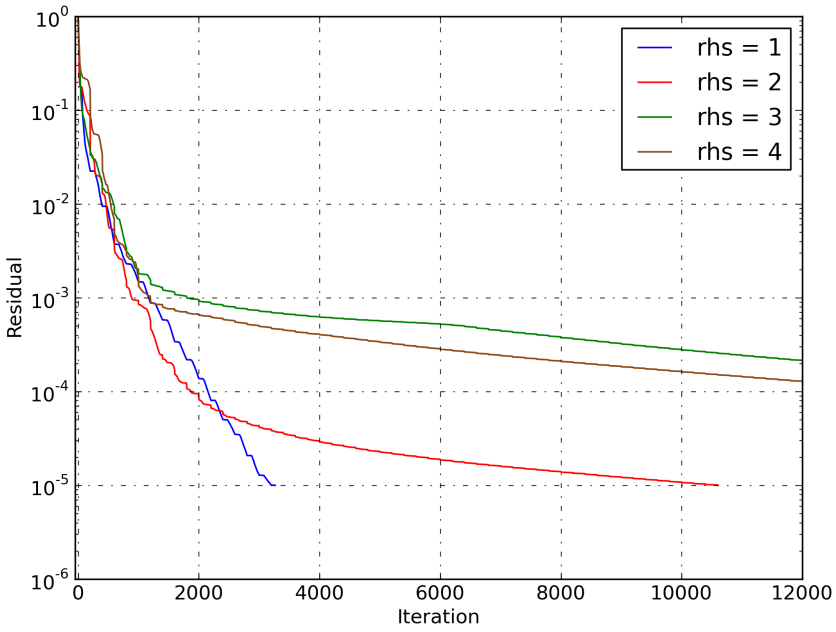



Figure 3: 2D test case: convergence of the Krylov space recycling method

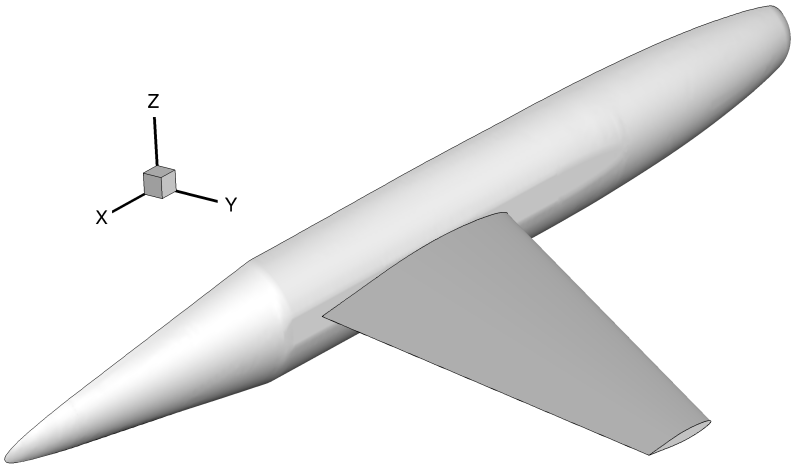



Figure 4: General view of the 3D test case

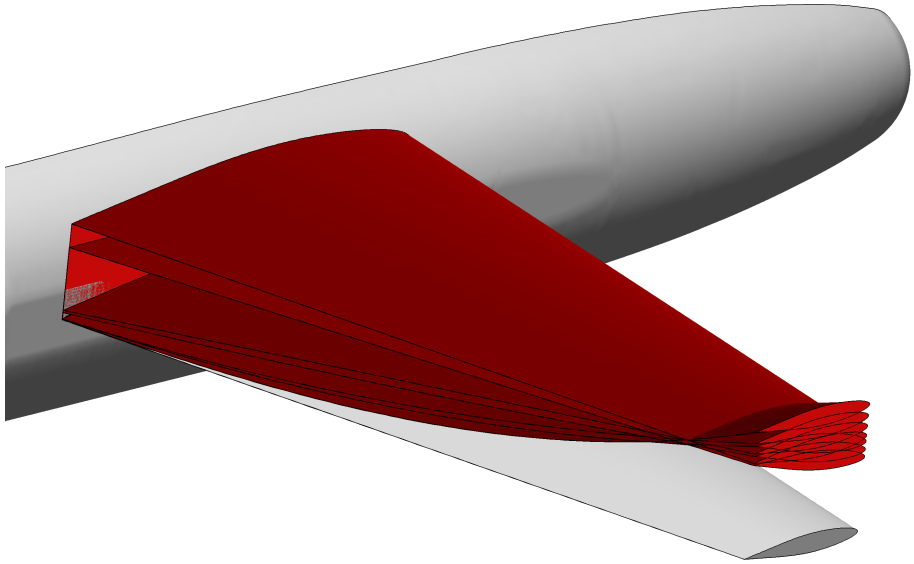


Figure 5: Seven movements corresponding to test polynomials on the wing

a requirement to get a good solution. By 4000 iterations, a standard GMRes computation would have already converged to that precision. As the convergence curve is not conclusive, no timings are shown.

### 5.2.2 Krylov space recycling

Recycling the harmonic Ritz vectors proved more interesting on the 3D case than on the 2D one. Figure 7 shows the convergence of the GMRes solver coupled with the Minres projection. The Krylov space was composed of 200 vectors, and 40 harmonic Ritz vectors were kept and used for the Minres projection. It is interesting to see that all the right-hand sides have the same asymptotic convergence rate. Nevertheless, this rate is smaller than the one achieved with the deflated GMRes solve (first right-hand side). Finally, the asymptotic convergence was not reached rapidly. The fact that no plateaus are found in the GMRes solve shows that this test case might not be the one where this method could shine [10]. To conclude with, the convergence penalty is far larger than the overhead time imposed by the deflation in standard GMRes.

## 6 CONCLUSION

In this paper, extensions of the GMRes algorithm to simultaneously solve several right-hand sides were implemented and tested on simple and industrial test cases. Block GMRes, the natural extension of GMRes to several right-hand sides, can accelerate matrix vector products on parallel computing architectures. Its lack of robustness on small Krylov space does put it at a disadvantage compared to standard GMRes. To compare favourably, it needs too large a Krylov space to be fast enough for industrial use. Krylov space recycling, which is a simpler

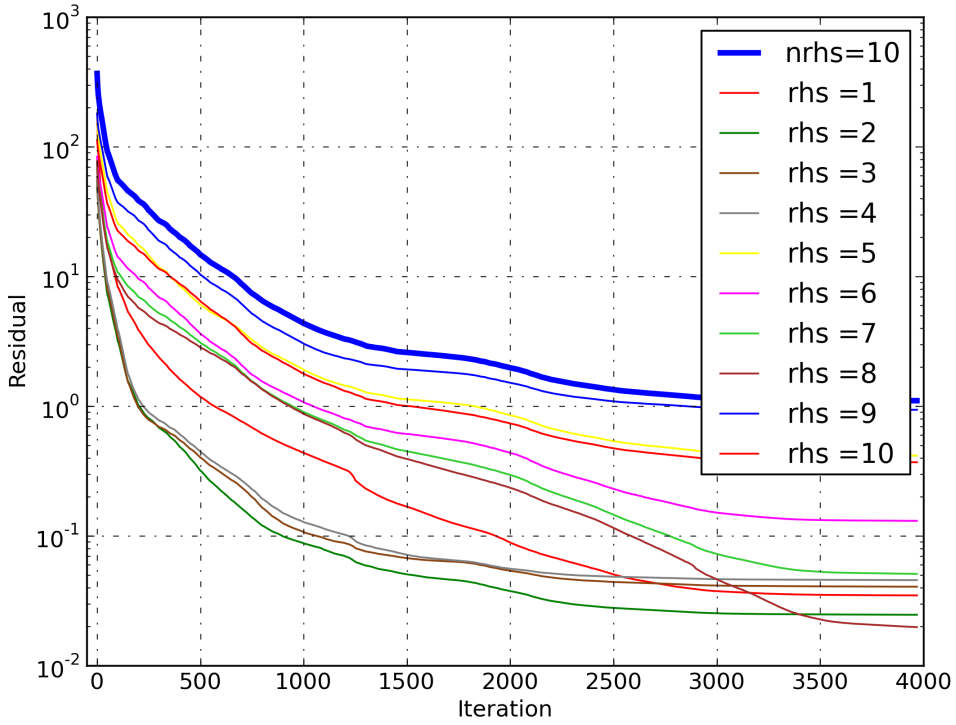



Figure 6: 3D test case: convergence of the block GMRes solver. Bold line is the total residual.

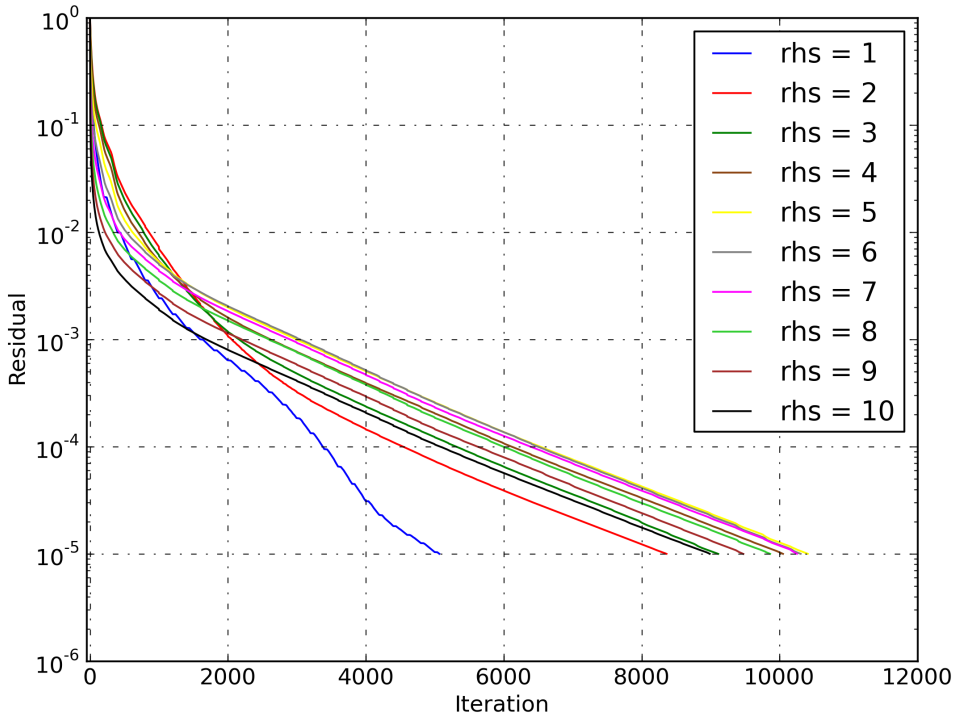



Figure 7: 3D test case: Krylov recycling. First right-hand side is solved with deflated GMRes

way of solving several right-hand sides at the same time, was disappointing on the simpler test case, where it stalled the convergence. On the industrial test case, though all right hand sides converged, the asymptotic convergence rate was not good enough and not reached soon enough for this method to be competitive.

This study reversely highlighted the performance of the deflated GMRes method for solving large linear systems, as various extensions of this method proved to be not as fast. Thus, future works will be devoted to accelerating the solves without changing the GMRes method. A study is currently underway to look at two-level preconditionners, which have shown to be promising to reduce the impact of domain decomposition on convergence in CFD computations[1].

## ACKNOWLEDGEMENT

This work is supported by a CIFRE fellowship funded in part by ANRT (National Association for Research and Technology) on behalf of the French Ministry of Higher Education and Research, and in part by Dassault Aviation.

## REFERENCES

[1] H. Alcin, B. Koobus *et al.*, *Efficiency and scalability of a two-level Schwarz algorithm for incompressible and compressible flows*, Int. J. Number. Meth. Fluids, 2012

[2] L. Daumas, F. Chalot, N. Forestier, Z. Johan, *Industrial use of linearized CFD tools for aeroelastic problems*, IFASD, 2009

[3] L. Daumas, Q. Dinh, S. Kleinveld, G. Rogé, *Automatic shape optimization using parametric CAD applied to sonic boom reduction*,3rd AIAA Multidisciplinary Design Optimization Specialist Conference, pp. 23-26. 2007.

[4] A. Gaul, M. Gutknecht, J. Liesen, R. Nabben, *A Framework for Deflated and Augmented Krylov Subspace Methods*, arXiv:1206.1506v2, Feb. 2013

[5] M. H. Gutknecht, *Block Krylov Spaces Methods for Linear Systems With Multiple Right-Hand Sides: An Introduction*, Feb. 2006

[6] M. Gutknecht, T. Schmelzer, *Updating the QR decomposition of block tridiagonal and block Hessenber matrices generated by block Krylov space methods*, Research Report No. 2005-10, 2005

[7] M. Mallet, *A Finite Element Method for Computational Fluid Dynamics*, PhD Thesis, Stanford University, 1985

[8] R. B. Morgan, *A restarted GMRES method augmented with eigenvectors*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1154–1171.

[9] R. B. Morgan, *Restarted Block GMRes With Deflation Of Eigenvalues*, Applied Numerical Mathematics, 54.2 (2005): 222-236

[10] R. B. Morgan, W. Wilcox, *Deflated Iterative Methods for Linear Equations with Multiple Right- Hand Sides*, arXiv 0707.0505 2007

[11] M. L. Parks, E. de Sturler *et al.*, *Recycling Krylov subspaces for sequences of linear systems*, SIAM J. Sci. Comput., Vol. 28, No. 5, pp. 1651–1674

[12] X. Pinel, M. Montagnac, *Block Krylov Methods to Solve Adjoint Problems in Aerodynamic Design Optimization*, AIAA Journal, Vol. 51, No. 9, Sept 2013

[13] S. Röllin, W. Fichtner, *Improving the accuracy of GMRes with deflated restarting*, SIAM J. Sci. Comput. Vol. 30, No. 1, pp. 232–245

[14] Y. Saad, M. Schultz, *GMRes: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems*, SIAM J. Sci. Stat. Comput., Vol. 7, No. 3, July 1986

[15] S. Xu, S. Timme, K. J. Badcock, *Krylov Subspace Recycling for Linearised Aerodynamics Analysis using DLR-TAU*, IFASD 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPARISON OF DIFFERENT SPATIAL/ANGULAR AGGLOMERATION MULTIGRID SCHEMES FOR RADIATIVE HEAT TRANSFER COMPUTATIONS

**Georgios N. Lygidakis**[1] **and Ioannis K. Nikolos**[2]

[1,2] Technical University of Crete, School of Production Engineering and Management
University Campus, Chania, GR-73100, Greece
[1]e-mail: glygidakis@isc.tuc.gr
[2]e-mail: jnikolo@dpem.tuc.gr

**Keywords:** Radiative Heat Transfer, Finite-Volume Method, Multigrid Scheme, Spatial/Angular Agglomeration, FAS, FMG.

**Abstract.** *During the past decade the 3D unstructured grids have become an important tool for radiative heat transfer simulations, extending their applications to even more complex enclosures. Nevertheless, the corresponding solvers appear to be inferior in terms of efficiency, compared to those for structured meshes. One remedy to this shortcoming appears to be the agglomeration multigrid method, based on the solution of the numerical problem on successively coarser spatial and angular resolutions, derived from the initial finest ones through the fusion of their neighbouring control volumes and control angles respectively. Considering this state, the enhancement of an in-house academic solver with different spatial/angular agglomeration multigrid schemes to accelerate the finite-volume method for the prediction of radiative heat transfer, is reported in this study. The incorporated multigrid methods are based on the relaxation of radiative transfer equation with the FAS approach, considering though different types of sequentially coarser spatial and angular resolutions, as well as different V-cycle types. More specifically, a nested, a uniform and an alternate scheme were developed, while they were examined in conjunction with the V(1,0), V(1,1), V(2,0) and V(2,1) V-cycles types. To further accelerate the numerical solution, a combined FMG-FAS strategy was included, according to which the whole procedure begins from the coarsest discretization (spatial and angular) and as the number of iterations is increased the FAS extends to the finer resolutions, up to the initial finest one. The proposed numerical schemes were validated against a benchmark test case, considering radiative heat transfer through a strongly scattering medium in a cubic enclosure with highly reflecting surfaces. The obtained results reveal the superiority of the nested scheme along with the V(2,0)-cycle type strategy, while they highlight the significant contribution of the angular extension of the multigrid technique.*

## 1 INTRODUCTION

During the last decade the three-dimensional unstructured grids have become an important tool for radiative heat transfer computations, extending their applications to more complex geometries, like combustion chambers. Although unstructured meshes offer the largest possible flexibility in the treatment of complicated enclosures, along with the minimum user interaction for their generation/adaptation, the corresponding solvers appear to be inferior in terms of efficiency, compared to those for structured grids [1]. One remedy to this shortcoming is the multigrid method, originally developed by Brandt [2] to increase the convergence rate of the numerical solution of elliptic problems, e.g. of incompressible fluid flow problems in CFD (Computational Fluid Dynamics). Since then it has been applied though to various types of computational simulations, e.g., compressible fluid flow [3-5], radiative heat transfer [6-8], etc. Its main concept depends on the construction of coarser resolutions, in order the low-frequency errors at the finest discretization to become high-frequency ones at the coarser ones, and as such to be damped in a more efficient way [1, 3, 6]. Besides multigrid method's significant contribution in test cases with unstructured grids, it was additionally revealed to be a valuable tool for simulations involving higher-order accurate spatial schemes [6, 7, 9-11] or higher-order governing equations [12]. Various types of the multigrid methodology have been developed during the past years, whose differences are identified mainly on the way the coarser discretizations are generated, as well as on the relation associating the successive grid resolutions [1, 6].

The agglomeration multigrid method, originally proposed by Lallemand [13], appears to be one of the most widely implemented schemes. It considers a sequence of coarser spatial resolutions with polyhedral elements, derived through the fusion of neighboring control volumes of their finer levels in an arbitrary way (isotropic agglomeration) [6]. Nevertheless, the aforementioned agglomeration strategy has been identified to lead to reduced performance in test cases involving hybrid grids with highly stretched elements on boundary surfaces [14]. A semi-coarsening or directional coarsening agglomeration method was proposed by Mavriplis [14] to mitigate this drawback. According to this technique the nodes of prismatic and hexahedral elements are treated separately, i.e., their control cells are merged only if they are aligned with the normal to the boundary direction [1]; for the rest nodes the aforementioned isotropic strategy is applied. As a result, the grid anisotropy, caused by the utilization of prismatic or hexahedral elements, is moderated [15]. Another popular approach is the full-coarsening directional agglomeration [1, 4, 16, 17], according to which the boundary control cells are fused initially, while a line-agglomeration step is then performed, for merging the prismatic or hexahedral control volumes along implicit lines, directly above the corresponding already agglomerated boundary ones. As a result, a deeper reduction of DoFs (Degrees of Freedom) is succeeded and consequently further acceleration is gained, without though the produced discretizations to differentiate significantly from the initial topology. Specifically for radiative heat transfer simulations, which consider the solution of RTE (Radiative Transfer Equation), for each control cell and each control angle, an angular extension of the aforementioned agglomeration strategy has been proposed by the authors, either coupled with the spatial one or not [6, 7]. According to this methodology, angularly coarser resolutions are also generated (besides the spatially coarser ones), resulting in further improvement of the computational performance of the iterative methodology, via a nested spatial/angular agglomeration multigrid scheme [6, 7].

As far as the relation associating each two successive resolutions is concerned the FAS (Full Approximation Scheme) approach is revealed to be the most widely applied one [1, 3]. It considers the relaxation of the governing PDEs (Partial Differential Equations) only at the ini-

tial finest level, while at the coarser ones an approximate version of the same PDEs is solved [3], accomplishing gradually a V- or W-cycle scheme. The computed variables and flux balances are restricted (in terms of smoothing) from the finer to the coarser mesh, while the corresponding calculated corrections are prolonged (in terms of interpolation) from the coarser to the finer one [1, 3, 4, 6]. In case of multigrid accelerated radiative heat transfer problems, considering additionally angularly coarser resolutions, the approximation of the angularly coarser PDEs is performed by employing the FAS scheme much in the same way to the spatial ones [6, 7]. In order to gain additional acceleration, a combined FMG-FAS (Full Multigrid-Full Approximation Scheme) approach can be implemented [1, 4, 15]. According to this method the FAS V-cycle is incorporated in the FMG process; the whole procedure begins from the coarsest grid and, as the number of iterations is increased, the finer FAS levels are added up to the initial finest one [1, 4, 15].

In this work the development and comparison of different spatial/angular agglomeration multigrid schemes for the acceleration of FVM radiative heat transfer computations, is reported. It is based upon a previous work of the authors [6, 7], incorporating though further enhancements, namely different sequences of spatial and angular coarser resolutions, different V-cycle types, a full-coarsening directional agglomeration strategy and a combined FMG-FAS approach. More precisely, the nested, uniform, and alternate schemes are compared. According to the first nested one (reported in [6, 7]) the angular V-cycle is accomplished at each step of the spatial one. If the uniform scheme is selected each coarser resolution is constructed by simultaneously coarsening its finer one spatially and angularly, while in case of the alternate technique each coarser level is obtained from the finer one by coarsening it either in spatial or angular dimension. As far as the different V-cycle types are concerned the V(1,0), V(1,1), V(2,0) and V(2,1) are assessed; the first number in parentheses denotes the number of relaxations performed prior to restriction, while the second one denotes the corresponding number of relaxations after the prolongation [6]. For simulations involving hybrid grids, in order increased accuracy to be achieved in regions with prismatic elements [18], a full-coarsening directional agglomeration strategy is additionally developed [1, 4, 5]. It succeeds greater reduction of DoFs and consequently greater acceleration, comparing to this obtained with isotropic agglomeration methodology. Finally, a combined FMG-FAS approach is developed and tested against the only-FAS technique [1, 4, 15]; it considers the division of the whole procedure in two stages, the preliminary and the main one [1, 4, 15]. It begins from the coarsest resolution (preliminary stage) while, as the number of iterations is increased, the finer FAS levels are added up to the initial finest one (main stage) [1, 4, 15]. The proposed numerical schemes are evaluated against a benchmark test case, considering radiative heat transfer through a strongly scattering medium in a cubic enclosure with highly reflecting surfaces [19]. According to the produced results the nested spatial/angular scheme along with the V(2,0)-cycle type appears to be the preferred choice for such simulations. Furthermore, they highlight the significant contribution of the angular extension of the multigrid technique.

## 2 RADIATIVE HEAT TRANSFER COMPUTATION

The radiative intensity $I_p$ of a node $p$ at position $r$ and time $t$ along a path $s$ through an absorbing, emitting and scattering gray medium is obtained by the time-dependent RTE as follows [10, 11, 20, 21]

$$\frac{1}{c}\frac{dI_p(\vec{r},\hat{s})}{dt}+\frac{dI_p(\vec{r},\hat{s})}{ds}=-(k_\alpha+\sigma_s)I_p(\vec{r},\hat{s})+k_\alpha I_b(\vec{r})+\frac{\sigma_s}{4\pi}\int_{4\pi} I_p'(\vec{r},\hat{s}_i')\Phi(\hat{s},\hat{s}_i')d\omega \quad (1)$$

where $c$ is the propagation speed of radiation, while $k_\alpha$ and $\sigma_s$ are the absorption and scattering coefficient respectively [10, 11]. The LHS terms express the radiative intensity change rate per ray direction and time, while the RHS expresses its attenuation by absorption and scattering processes (first term), and its augmentation by blackbody energy (second term) and scattering phenomenon (third term) [6, 10]. Despite the presence of the temporal term, the aforementioned expression can be employed for both steady-state and transient problems [6, 10, 11, 20].

The finite-volume method is applied for the discretization of the computational field and consequently for the solution of the RTE [6, 12]; therefore, the computational domain has to be discretized both spatially and angularly. A node-centered median dual control volume method is employed for the spatial discretization, according to which the construction of the control cell of a node is achieved by connecting the lines defined by the edge midpoints, the barycenters of faces and the barycenters of elements sharing this node [3, 18]. Angular discretization is succeeded by dividing the directional domain, represented by a *sphere* in three dimensions, into a discrete number of solid control angles with lines of constant longitude and latitude [18, 21]; an equal division strategy is followed in this work, hence, the $4\pi$ steradians derive $N_\theta$x$N_\varphi$ control angles [6, 10, 18]. Thus, equation (1) is integrated over the control volume of each node $p$ and each solid control angle $\Delta\Omega^{mn}$ deriving the following formulation [10]

$$\Delta I_p^{mn} \frac{V_p \Delta\Omega^{mn}}{c\Delta t} + \sum_i I_i^{mn} D_{ci}^{mn} \Delta A_i = \left[ -\left(k_\alpha + \sigma_s\right) I_p^{mn} + S_R^{mn} \right] V_p \, \Delta\Omega^{mn} \tag{2}$$

where $V_p$ is the volume of the examined control cell. The source term $S_R^{mn}$, which includes the contribution of the blackbody energy and the scattering phenomenon from other solid control angles, is expressed as [10]

$$S_R^{mn} = k_\alpha I_b + \frac{\sigma_s}{4\pi} \int_{4\pi} I_p^{m'n'} \, \Phi\left(m'n', mn\right) \, d\omega \tag{3}$$

where $\Phi$ denotes the Scattering Phase Function. If isotropic scattering is assumed, a constant value is assigned to it [18, 22]; alternatively, medium's anisotropic attitude has to be modelled, e.g. using Legendre polynomials (as in this study) [18, 22]. Finally, $D_{ci}^{mn}$ is the directional weight, defined by the examined control angle $\Delta\Omega^{mn}$ and the unit normal vector of the examined control volume surface, while $I_i^{mn}$ is the corresponding radiative intensity [10, 11]. For its definition the step scheme is adopted in this work, according to which the radiative intensity at a downstream face is set equal to that of the upstream node; this is expressed for two neighboring nodes $p$ and $q$ as [6, 10]

$$I_i^{mn} D_{ci}^{mn} = I_p^{mn} D_{ci,out}^{mn} + I_q^{mn} D_{ci,in}^{mn} \tag{4}$$

where $D_{ci,out}^{mn}$ denotes the directional weight going outwards the examined control volume (of node $p$), while $D_{ci,in}^{mn}$ the corresponding weight coming into the same cell [6]. Considering this scheme, an edge-based data structure is used to reduce the computational load [3]. Furthermore, in order to alleviate the overhang problem, derived by the combination of unstructured grids with the angular division of the computational domain, the pixelation method is applied transforming appropriately the aforementioned weights [10, 11]. In that way the effect of overlapped directional weights is considerably reduced and the accuracy of the final solution is increased.

For the same accuracy reasons, a higher-order accurate spatial scheme is implemented reducing the effect of false scattering, derived by the spatial discretization of the computational field. Particularly, a second-order scheme is used in this work, based on the well-established

in CFD MUSCL (Monotone Upstream Scheme for Conservation Laws) methodology [23]. Therefore, the values of radiative intensity at the RHS of equation (4) are reconstructed prior to the implementation of the step scheme; the required gradients are obtained with the Green-Gauss linear representation method [10]. Moreover, in order to control the aforementioned reconstructed values, especially at boundary areas entailing high intensity gradients, the scheme is coupled with the Van Albada-Van Leer [23] or the Min-mod [24] slope limiters [10, 11].

Finally, the appropriate contributions of the boundary conditions are added to the fluxes of the corresponding nodes; they are computed in an implicit way applying the step scheme between the examined node and a *ghost* node outside the computational field [10, 11]. Opaque and diffusive boundary surfaces, as well as mirroring ones are assumed in this study [10, 11].

Since the required flux balance has been obtained for each node and each solid control angle, equation (2) is transformed as [10, 11]

$$\Delta I_p^{mn} \frac{V_p \Delta\Omega^{mn}}{c\Delta t} = R_p^{mn} \tag{5}$$

where $R_p^{mn}$ is the fluxes sum, while $\Delta t$ is the pseudo-time step defined via a local time-stepping technique [3, 10, 11]. For the iterative relaxation of equation (5) an explicit second-order temporal accurate scheme, using a four-stage Runge-Kutta method (RK(4)) [25], is employed. Besides the local time-steeping technique and the edge-based data structures, further acceleration of the solution process is obtained with parallel processing, based on the domain decomposition approach and the MPI (Message Passing Interface) library functions [18, 26, 27].

## 3 THE SPATIAL/ANGULAR AGGLOMERATION MULTIGRID SCHEME

### 3.1 Spatial/angular agglomeration strategy

The first issue to be defined for the multigrid accelerated solution of RTE is the agglomeration strategy, i.e., the methodology which has to be applied for the generation of the sequence of the coarser resolutions. Similar procedures are followed for the construction of both the coarser spatial and angular discretizations. They are performed on a topology-preserving framework at each partition, in which the initial grid is divided for parallel processing [1, 6]; they are confined though by predefined limitations, ensuring consistency of the restriction and prolongation processes [1, 6].

Concerning the spatial agglomeration strategy, a similar to the advancing front technique is implemented, as the whole procedure begins from the solid wall boundary nodes, while it extends gradually to the internal ones [1, 6]; the starting point is justified by the fact that the proposed spatial algorithm was initially developed for CFD simulations [1, 4]. If no such boundary nodes are present to the examined sub-grid, the process begins from the *core* nodes at the overlapping regions (among the examined partition and its adjacent ones) [10, 11, 18]. As mentioned above, predefined rules limit though this procedure, e.g., a boundary node can be merged only with another boundary node of the same surface type, an internal node can be fused only with its neighboring also internal nodes, etc. [1, 6]. Taking into account those limitations the isotropic agglomeration procedure begins with the construction of the so-called *seed* list, including the solid wall boundary nodes. A loop is performed over them, examining their eligibility for fusion with their adjacent non-agglomerated yet ones; if no constraint is identified they are merged with their neighboring ones creating *supernodes*, while if a limitation exists they are simply transferred to the next multigrid level as *singleton supernodes* [1,

6]. A new list is then constructed including the nodes touched by the agglomeration front, which are actually the adjacent ones to the already examined and agglomerated nodes, while at next the previous step of examination and fusion is repeated. The procedure is assumed accomplished when all the *core* nodes have been examined. The *ghost* nodes are not included to the main agglomeration process, but they are fused according to the merging path of their corresponding *core* nodes at the adjacent partitions; as a result, virtual *ghost supernodes* are produced [1, 6]. The procedure continues with the construction of the corresponding *superedges*, connecting the derived *supernodes* and representing the interfaces of their control cells. In case an even coarser grid is required, the whole process is repeated. Further details for the applied isotropic spatial agglomeration procedure can be found in [1] and [6].

As mentioned in Introduction, the isotropic agglomeration strategy appears to be less effective in test cases involving hybrid grids, with highly stretched elements (prisms or hexahedra) at boundary surfaces [1, 4, 5, 14-17]; the latter are employed in order increased accuracy to be obtained at those areas [18]. Thus, a full-coarsening directional agglomeration procedure was also incorporated to the proposed algorithm, which is based on the methodology of Nishikawa and Diskin [5]. Further modifications have been included though; the agglomeration process isn't limited across the sub-domains' internal boundaries, allowing virtual *ghost supernodes* to be generated, while an additional limitation prohibits the fusion process of the control cells at prismatic layers in order the topology of the initial grid to be preserved more accurately [1]. Similarly to isotropic agglomeration, the directional one begins with the construction of the initial *seed* list, containing though the boundary nodes of only the prismatic elements [1, 5]. Since they are examined and merged in *supernodes*, a new list is created, filled with the nodes of the next prismatic layer; the latter are fused according to the agglomeration path of their corresponding boundary nodes, constructing the so-called implicit lines [1, 5]. The previous steps are repeated until all the prismatic nodes have been examined for fusion; the isotropic method is applied then for the rest nodes (tetrahedral and pyramidical) [1].

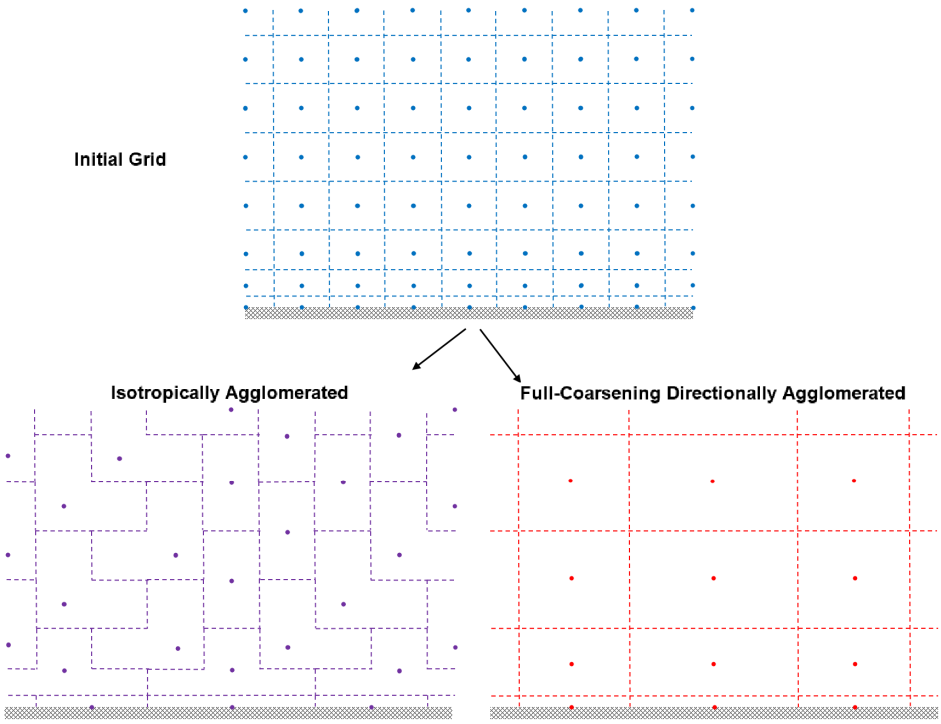



Figure 1: Isotropic and full-coarsening directional agglomeration of a 2D quadrilateral grid.

Figure 1 illustrates a schematic example of the previously described directional procedure (right), compared with this of the corresponding isotropic one (left). It is obvious that directional agglomeration preserves more accurately the topology of the initial grid; on the other hand, the isotropic approach derived a mesh with a topology modified compared to the initial one, due to the arbitrary polyhedral control cells it entailed. Further details for the aforementioned full-coarsening directional agglomeration procedure can be found in [1].

As far as the angular agglomeration is concerned, a similar to the spatial strategy is followed [6]. For the construction of each coarser resolution, every two successive control angles in both azimuthal and polar directions are fused, deriving new solid control *superangles*. Considering the equal separation strategy, followed for the division of the initial directional domain, each coarser level includes the one quarter of the number of solid control angles of its finer one [6]. The procedure is limited though by only one predefined constraint (unlike the spatial process), defining that only the control angles belonging to the same quadrant of the directional *sphere* can be fused together [6]. In that way consistency during the restriction and prolongation processes is ensured [6]. A more detailed description of the angular agglomeration strategy can be found in [6].

### 3.2 The spatial/angular agglomeration multigrid FAS approach

As mentioned in Introduction, this study aims to compare different combined spatial/angular agglomeration multigrid schemes for the acceleration of FVM radiative heat transfer computations. Particularly, three such schemes have been developed, namely a uniform, an alternate and a nested scheme. Their differences focus on the sequence of the spatial and angular levels, employed during the V-cycle of the FAS method. The procedures of the generation of the successively coarser resolutions (spatial and angular) as well as of the solution strategy (implementation of the second-order spatial accurate scheme and fix-up method, evaluation of directional weights and pixelation coefficients, etc.) are performed in the same way for all the developed schemes [6].

As far as the uniform spatial/angular agglomeration multigrid scheme is concerned, each coarser resolution is constructed by simultaneously coarsening its finer one spatially and angularly. According to the FAS approach, the solution procedure begins with the relaxation of equation (5) with the Runge-Kutta method at the initial finest resolution, while at next the values of the nodal fluxes and radiative intensity are restricted to the next coarser level as follows [6]:

$$R_{P,restricted}^{MN} = \left(I_R\right)_{h,mn}^{H,MN} R_p^{mn} = \sum R_p^{mn}$$
$$I_{P,restricted}^{MN} = \left(I_I\right)_{h,mn}^{H,MN} I_p^{mn} = \frac{\sum I_p^{mn} \cdot V_p \cdot \Delta\Omega^{mn}}{V_P \cdot \Delta\Omega^{MN}} \tag{6}$$

Actually, a simple summation operator is applied for flux balances, considering both the included control cells and solid control angles. The corresponding operator for the values of radiative intensity defines a spatial-angular averaging process instead. Since the restriction process is accomplished, a similar to equation (5) relation is relaxed for this level (*H,MN*); its RHS is substituted though by the following value

$$R_{P,FAS}^{MN} = R_P^{MN}\left(I_P^{MN}\right) + \underbrace{\left[R_{P,restricted}^{MN} - R_P^{MN}\left(I_{P,restricted}^{MN}\right)\right]}_{A_H^{MN}} \tag{7}$$

where the first term denotes the fluxes of the agglomerated *supernode P* computed at this level (*H,MN*) similarly to the initial finest one using the previously constructed *superedges* [1, 6]. The forcing function $A_H^{MN}$ includes the restricted fluxes to the *supernode P* and the fluxes of the same node, computed with the restricted values of radiative intensity from the finer resolution. It is quite obvious that for the first internal iteration of the Runge-Kutta scheme, the RHS term equals the restricted from the finer level flux balance, confirming in that way the approximation character of the FAS approach [1, 3, 6]. The aforementioned procedure of relaxation and restriction are repeated up to the coarsest generated resolution, while at next the obtained corrections are prolonged to the next finer level applying a simple point-injection scheme as [1, 3, 6]

$$\left(I_p^{mn}\right)^{l+} = \left(I_p^{mn}\right)^{l} + \Delta I_p^{mn} = \left(I_p^{mn}\right)^{l} + \left(I_I\right)_{H,MN}^{h,mn} \Delta I_P^{MN} = \left(I_p^{mn}\right)^{l} + \left(I_P^{MN} - I_{P,restricted}^{MN}\right) \tag{8}$$

The prolongation procedure is repeated similarly up to initial finest level, at which point the FAS V-cycle gets accomplished [1]. Figure 2 depicts a schematic representation of the aforementioned uniform method with a V(1,0)-cycle type and two spatial-angular coarser levels.

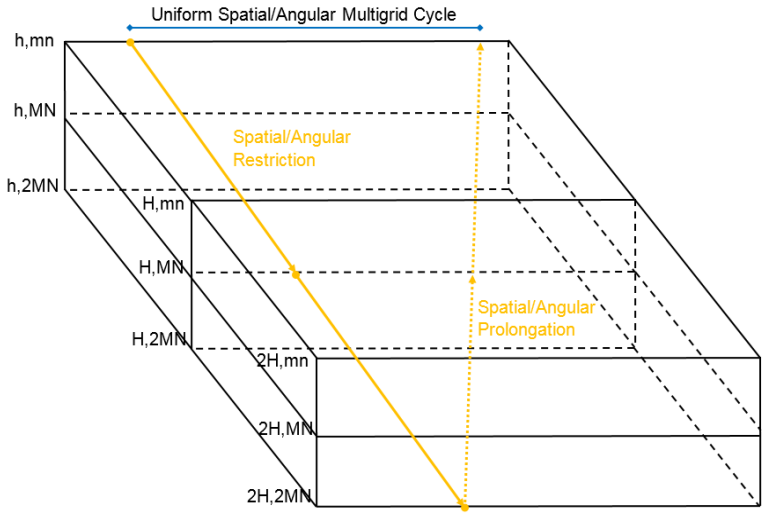



Figure 2: Uniform spatial/angular agglomeration multigrid cycle.

It should be highlighted that an equal number of coarser spatial and angular discretizations has to be used for the implementation of this approach. However, in most test cases, more spatial resolutions than angular ones can be generated [7]; to alleviate this shortcoming the coarsest generated angular discretization can be employed for the next coarser spatial levels too.

The alternate spatial/angular agglomeration multigrid scheme defines instead an alternation of spatial and angular coarsening at the sequence of FAS levels; as mentioned in Introduction, each coarser level is obtained from the finer one by coarsening it either in spatial or angular dimension. The whole procedure begins again with the solution of equation (5) at the initial finest level (spatially and angularly), while at next the nodal fluxes and radiative intensities are restricted to the next spatially coarser grid; no angular coarsening has been performed to this level. Since the solution is obtained at this discretization, the derived fluxes and radiative intensities are restricted to the next coarser level, provided with angular coarsening only. The aforementioned steps are repeated alternatively up to the coarsest resolution (spatially and angularly), while at next the corrections are prolonged accordingly, i.e., alternatively, up to the

initial spatially and angularly finest level. The operators, used for the aforementioned processes, are the same applied for the only-spatial or the only-angular agglomeration multigrid schemes [6]. Similarly to the uniform approach an equal number of coarser spatial and angular discretizations has to be used for the implementation of this scheme. Figure 3 includes a schematic representation of the previously described alternate method with a V(1,0)-cycle type and two spatial/angular coarser levels.

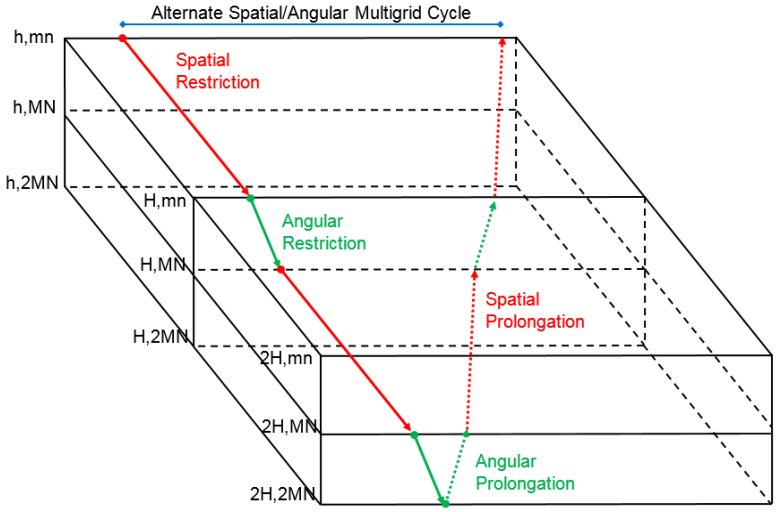



Figure 3: Alternate spatial/angular agglomeration multigrid cycle.

Besides the aforementioned schemes, a nested spatial/angular agglomeration multigrid scheme is assessed in this study, originally proposed in [6]. According to this approach, a complete angular FAS V-cycle is executed in each level of the spatial FAS V-cycle; the spatial and angular, restriction and prolongation, operators are applied similarly to the only-spatial and only-angular multigrid schemes [6]. Figure 4 illustrates its schematic representation, considering a V(1,0)-cycle type and two spatial/angular coarser levels. More details for the implementation of the nested spatial/angular agglomeration multigrid method can be found in [6].

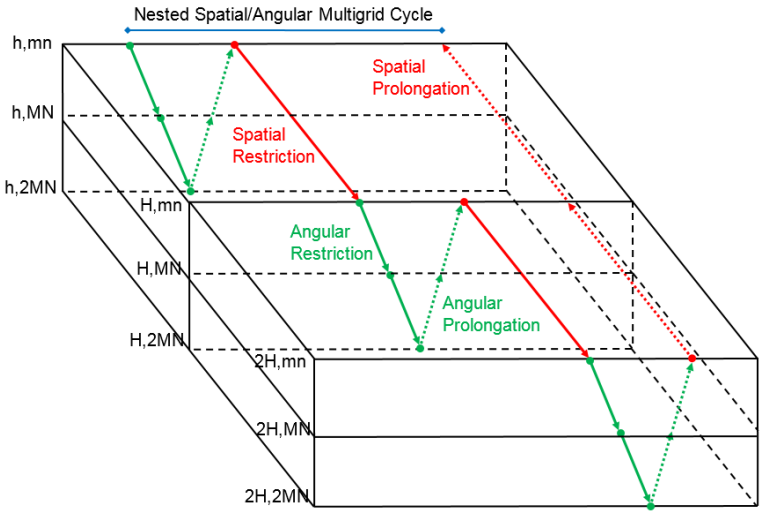



Figure 4: Nested spatial/angular agglomeration multigrid cycle.

### 3.3 FAS V-cycle types

Different V-cycle types are assessed in this study as well, namely the V(1,0), V(1,1), V(2,0) and V(2,1). The first number in parentheses denotes the number of relaxations performed prior to restriction, while the second denotes the corresponding number after the prolongation [1, 3]. Depending mainly on the implemented iterative procedure, as well as on the encountered problem's nature, various V-cycle types have been reported in the open literature as the most appropriate for them [1, 3, 5, 14-17]. For example, if an explicit iterative scheme is applied for an inviscid compressible fluid flow simulation a V(1,0)-cycle is usually preferred, while in case of a turbulent flow problem along with an implicit method, a V(2,1) appears to be the commonly preferred choice [1, 3]. As such, the four aforementioned V-cycle types have been incorporated in the proposed solver to study their effect on the code's computational efficiency.

### 3.4 The combined FMG-FAS agglomeration multigrid approach

A combined FMG-FAS agglomeration multigrid approach was also developed to further accelerate the solution procedure of the radiative heat transfer problems [1, 3, 4, 15]. Independently of the employed FAS scheme (only-spatial, only-angular, uniform, alternate or nested), it considers the division of the whole procedure in two stages, the preliminary and the main one [1, 4, 15]. As such, the solution process begins from the coarsest resolution and as the number of iterations/cycles is increased the FAS is extended to the next finer level, using the derived solution by the coarsest discretization as an initial guess [1]. In that way a cheaper initial condition is obtained, compared to the usually used uniform one [3]; for the interpolation of the solution the same single point-injection scheme, used for the prolongation process, is applied. The FAS solution process is continued, including only the two coarsest resolutions, while as the number of iterations/cycles is increased again, the previous steps are repeated adding successively more FAS levels up to the initial finest discretization, at which point the preliminary stage ends and the main one begins [1]. For simplicity reasons, in combined spatial/angular agglomeration multigrid schemes, the spatial and angular resolutions are extended simultaneously to their finer ones during the preliminary stage (similarly to the procedure followed for the uniform approach). Finally, the number of multigrid cycles/iterations performed during the preliminary stage may differ depending on the implemented multigrid scheme (only-spatial, only-angular, uniform, alternate or nested).

## 4 VALIDATION RESULTS

For the evaluation of the efficiency improvement, contributed by the different combinations of the proposed multigrid schemes (i.e., by employing the uniform, alternate, or nested scheme, different V-cycle types, and the combined FMG-FAS method), a benchmark test case was encountered, considering radiative heat transfer in a cubic enclosure with edge length equal to unity [7, 10, 19]. A schematic representation of the cubic geometry is depicted in Figure 5. The included cold medium is assumed purely scattering ($\sigma_s=1m^{-1}$, $k_\alpha=0m^{-1}$, $T_m=0K$). The wall surfaces are considered similarly cold ($T_w=0K$), except for the bottom face at which a constant heating energy is implemented ($E=1W/m^2$), while for the bases normal to the y-direction mirroring boundary conditions are imposed [7, 10, 19]. The computational domain is represented by a hybrid grid, composed of *22,500* nodes, *65,480* tetrahedra and *19,460* prisms; for angular discretization the directional domain is divided in *16* azimuthal and *8* polar angles. Parallel processing was performed on a workstation with an AMD FX$^{(tm)}$-8350 eight-core processor at *4.00* GHz, while the initial mesh was divided in two sub-domains. In order to alleviate the overhang problem the pixelation method was implemented, while false

scattering effect was reduced with the incorporated second-order spatial accurate scheme, coupled with the Min-mod limiter [10, 11].

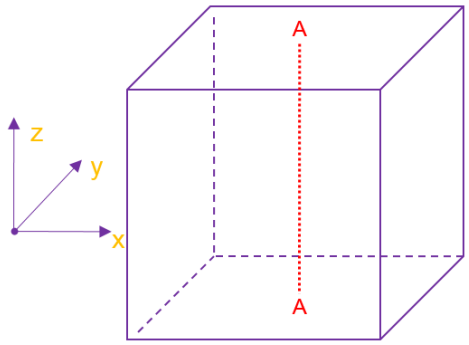



Figure 5: Geometry of the cubic enclosure.

Prior of validating the efficiency improvement, the accuracy of the proposed solver was assessed. The medium is assumed to exhibit anisotropic scattering behavior, which is modelled with the F2 scattering phase function, based on the Legendre polynomials [18, 19]. Two simulations were encountered for the evaluation of solver's accuracy, in which different values of wall emissivity $\varepsilon_w$ were assumed, namely *1.0* and *0.1*. In Figure 6 the extracted distributions of dimensionless incident radiative heat flux along the A-A line are illustrated, compared to the corresponding ones reported by Kim and Lee [19]. It is obvious that a very good agreement has been achieved. The second test case (with wall emissivity equal to *0.1*) required much more computation time, compared to the first one, due to the increased radiation exchange between the discretized solid control angles [7]. Therefore, it was selected in order the efficiency evaluation simulations to be conducted with.

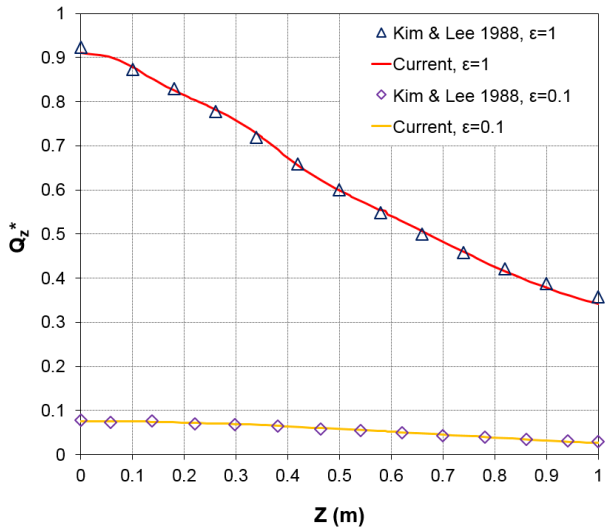



Figure 6: Distribution of the dimensionless incident radiative heat flux along the line A-A for two different values of wall emissivity ($\varepsilon_w$=*1.0* and $\varepsilon_w$=*0.1*).

For the validation of the computational performance's improvement the radiative intensity residual has to be computed at each multigrid cycle *k* [7]

$$residual = \frac{\sum_{p=1}^{N_p}\sum_{m=1}^{N_\theta}\sum_{n=1}^{N_\varphi}\left|I_p^{mn,k+1} - I_p^{mn,k}\right|}{N_P N_\theta N_\varphi} \tag{9}$$

where $N_p$ denotes the number of nodes, while $N_\theta$ and $N_\varphi$ the number of polar and azimuthal control angles, respectively, all of them regarding the initial finest resolution (spatially and angularly). Besides the contribution of the combined spatial/angular agglomeration multigrid schemes, developed in this study, this of the only-spatial and only-angular approaches was validated too. The notation SxAy, originally proposed in [7], is used in this work as well, where x denotes the number of spatial multigrid levels, while y the number of angular ones.

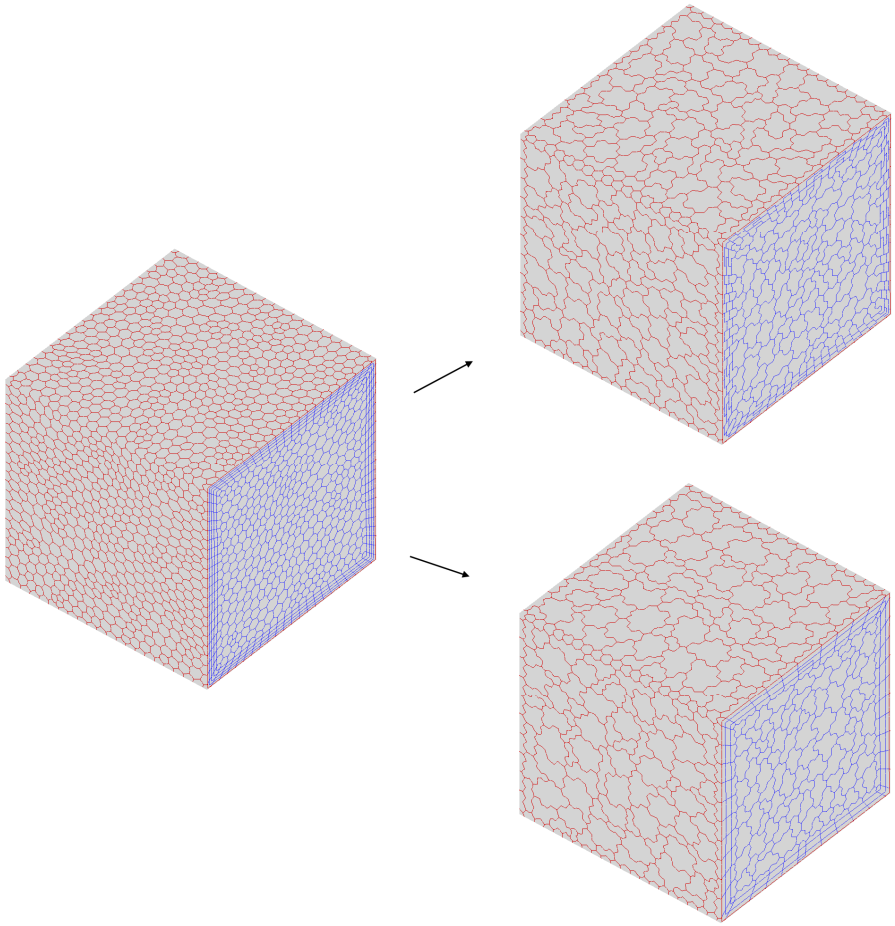


Figure 7: Initial and coarser via isotropic (top) and directional (bottom) fusion surface control-volume grids.

At the first stage of the evaluation, the incorporated full-coarsening directional agglomeration strategy was assessed. Thus, three successively coarser grids were generated, implementing both the isotropic and the directional agglomeration methodology. Figure 7 illustrates the initial and the first-level coarser surface control-volume grids, extracted with both approaches (isotropic-top, directional-bottom). It is obvious that the second agglomeration type preserved more accurately the topology of the initial grid, while it lead to a more extended reduction of DoFs. Contrary to that, the isotropic agglomeration produced arbitrary polyhedral control volumes even in the prismatic region, differentiating significantly the topology of the extracted grid from the initial one. The effect of this differentiation is demonstrated in Figure 8, in-

cluding the radiative intensity convergence history per number of iterations and wall-clock time for two only-spatial S4A1 V(1,0)-cycle simulations, implementing isotropic and directional agglomeration respectively. No significant difference can be identified at the iterations/cycle curve, but the situation is changing at the time curve; particularly, isotropic strategy produced a temporal speed-up coefficient equal to *~2.35* for the final residual *1.0E-10*, while directional one equal to *~2.51*. Despite a reduction of three or four orders of magnitude is usually more than enough for such test cases, speed-up coefficients have been computed for a final residual value equal to *1.0E-10*, in order the contribution of the proposed multigrid method to be more clearly demonstrated.

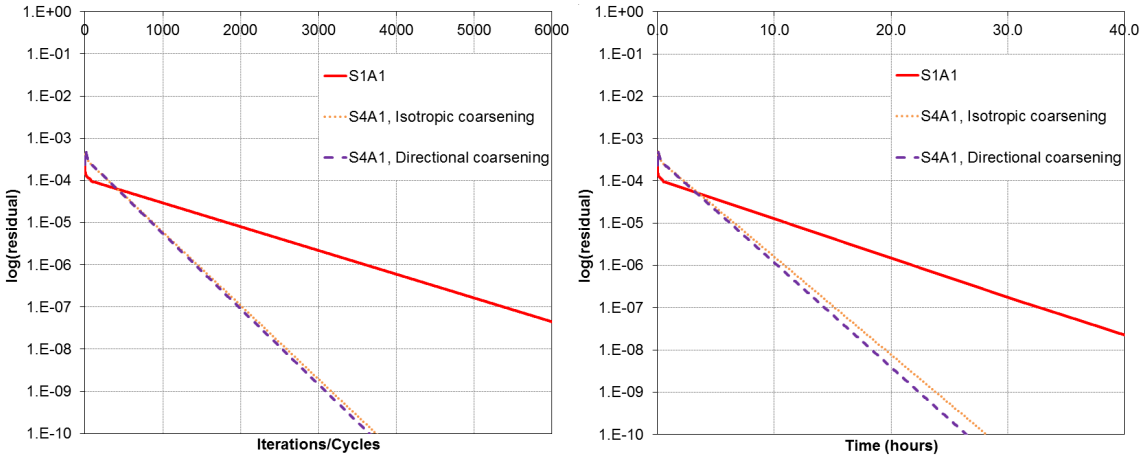



Figure 8: Radiative intensity convergence history per number of iterations and wall-clock time for the only-spatial agglomeration multigrid scheme, implementing isotropic and directional agglomeration approach.

The evaluation of the incorporated multigrid schemes was continued with the only-spatial and only-angular ones implementing though different V-cycle types, namely, V(1,0), V(1,1), V(2,0) and V(2,1). For the implementation of the only-angular as well as of the combined spatial/angular schemes, angular coarsening was performed analogously; the initial finest 16-8 discretization was reduced successively to 8-4 and 4-2. Figures 9 and 10 contain the radiative intensity convergence history per number of iterations and wall-clock time of the aforementioned simulations.

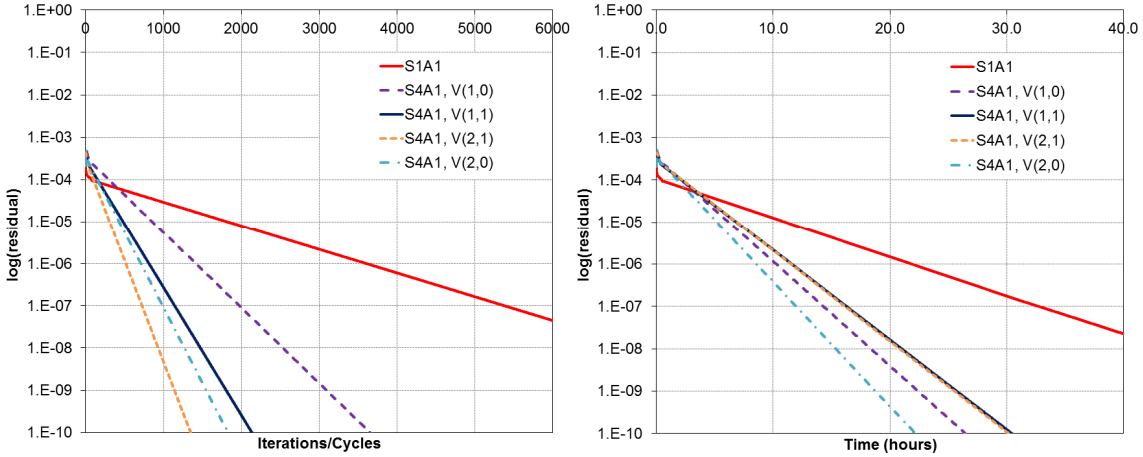



Figure 9: Radiative intensity convergence history per number of iterations and wall-clock time for the only-spatial agglomeration multigrid scheme employing different V-cycle types.

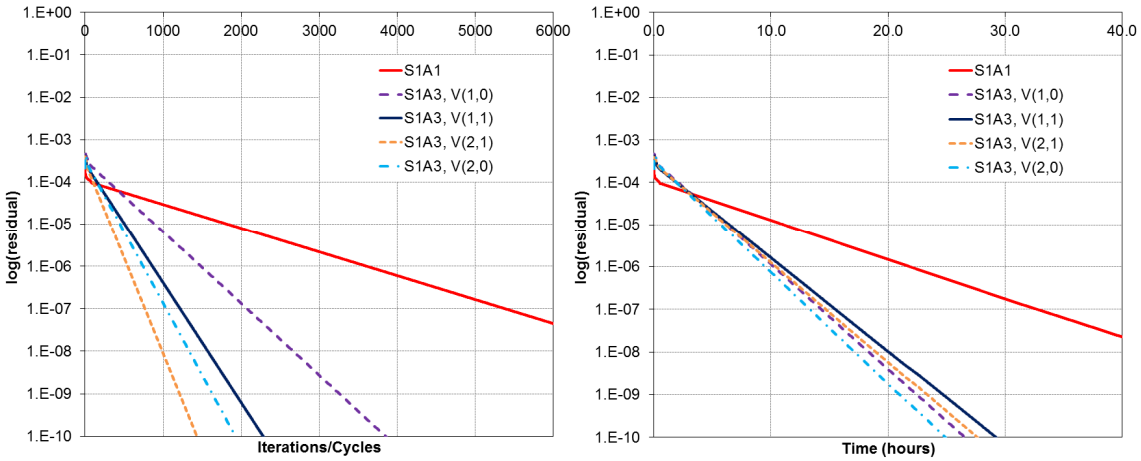



Figure 10: Radiative intensity convergence history per number of iterations and wall-clock time for the only-angular agglomeration multigrid scheme employing different V-cycle types.

The V(2,0)-cycle type seems to be the most efficient one, succeeding a time-acceleration rate equal to *~3.00* and *~2.67* with the only-spatial and the only-angular multigrid schemes respectively. The V(1,0) is revealed as the next most efficient type, achieving temporal speed-up coefficients equal to *~2.51* and *~2.49* with the spatial and angular approach respectively. Despite the fact that the only-spatial scheme appears to be more effective than the only-angular one in this test case, the speed-up coefficients (extracted with the V(1,0)-cycle type) of the only-angular scheme highlight the significance of the proposed angular extension to the multigrid technique.

The assessment is continued with simulations implementing the uniform, alternate and nested spatial/angular methodologies, along with the aforementioned V-cycle types. For the fourth FAS level of the uniform and alternate methods (besides the third one), the coarsest angular resolution (4-2) was used as well. Figures 11 to 13 illustrate the radiative intensity convergence history per number of iterations and wall-clock time of the aforementioned simulations. The nested scheme along with the V(2,0)-cycle type appears to derive the most efficient simulation, succeeding a temporal acceleration coefficient equal to *~4.21*. The V(2,0)-cycle alternate multigrid approach is revealed as the next most efficient type, with a corresponding coefficient equal to *~3.45*. As expected, the aforementioned combined spatial/angular multigrid schemes achieved better temporal rates than the only-spatial or only-angular approaches. Unlike them, the uniform scheme doesn't seem to have a significant effect in the computational performance of the proposed algorithm; actually, it produced worse speed-up coefficients, even compared with the only-spatial or only-angular multigrid methods. This is attributed to the extreme coarsening, entailed by the simultaneous spatial and angular reduction of the corresponding DoFs. Once more the V(2,0)-cycle scheme provided better acceleration coefficients, comparing to the rest V-cycle types, independently of the implemented FAS scheme (uniform, alternate or nested). Furthermore, the simulation involving the nested scheme with the V(1,0)-cycle type failed to converge; this failure originates possibly from the insufficient relaxation of RTE prior prolonging the radiative intensity corrections from the angular to the spatial V-cycle. As far as the rest two V-cycle types are concerned, it seems that the relaxation after the prolongation process does not contribute to the computational performance improvement; on the other hand, it appears to add just extra computation time. The previous state is confirmed from the fact that V(1,1)-cycle type is revealed even slower from the V(2,1).

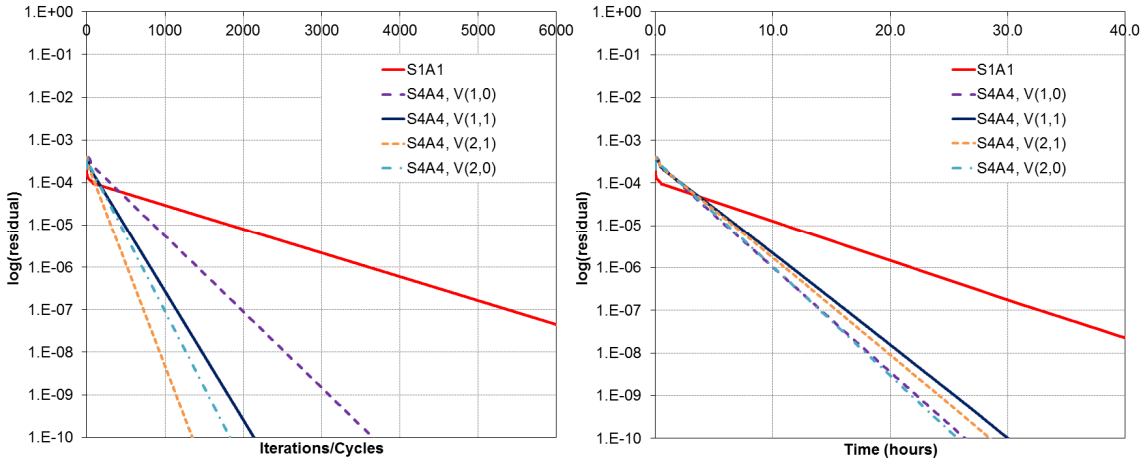



Figure 11: Radiative intensity convergence history per number of iterations and wall-clock time for the uniform spatial/angular agglomeration multigrid scheme employing different V-cycle types.

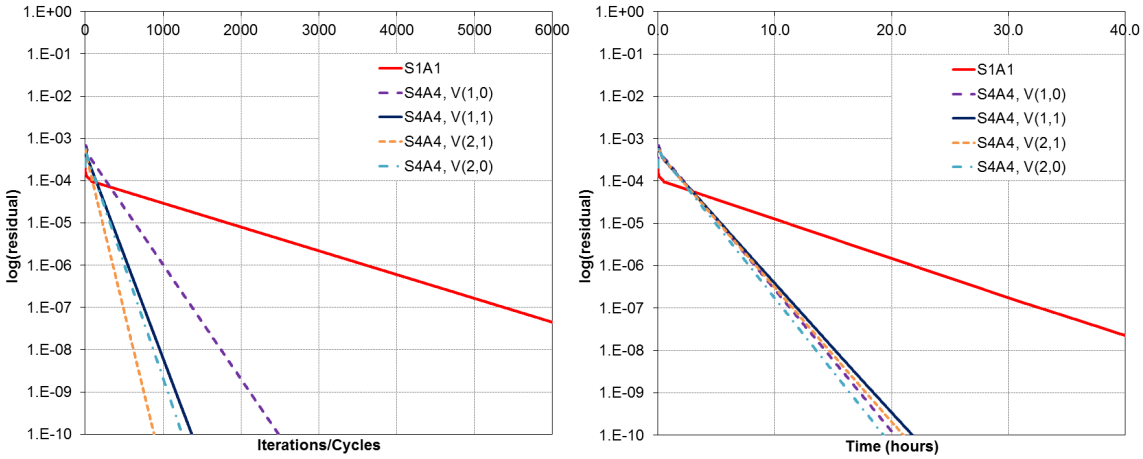



Figure 12: Radiative intensity convergence history per number of iterations and wall-clock time for the alternate spatial/angular agglomeration multigrid scheme employing different V-cycle types.

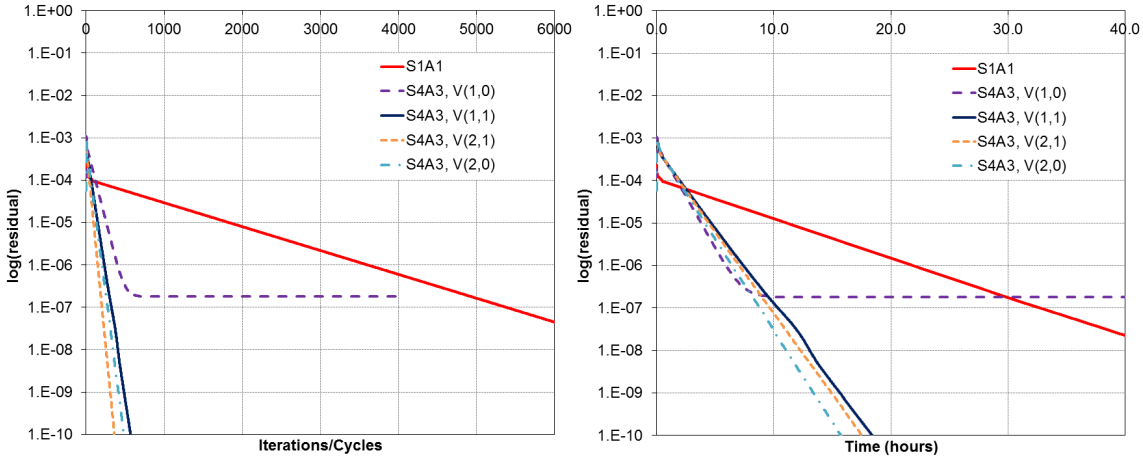



Figure 13: Radiative intensity convergence history per number of iterations and wall-clock time for the nested spatial/angular agglomeration multigrid scheme employing different V-cycle types.

The contribution of the combined FMG-FAS approach was then evaluated for all the incorporated multigrid schemes (only-spatial, only-angular, uniform, alternate and nested). Figures 14 includes the radiative intensity convergence history per number of iterations and wall-clock time of the aforementioned simulations. Different numbers of iterations/cycles were tested for each of the implemented multigrid schemes during the preliminary stage; the most efficient results for each approach are presented in the aforementioned figure. The nested spatial/angular and only-angular methods appear to be the most efficient ones, achieving a temporal speed-up coefficient equal to *~4.81*. The next more effective scheme is revealed to be the alternate one, while the only-spatial one provided the worst results; considering the latter observation, the significant contribution of the angular extension of the multigrid technique is once more demonstrated.

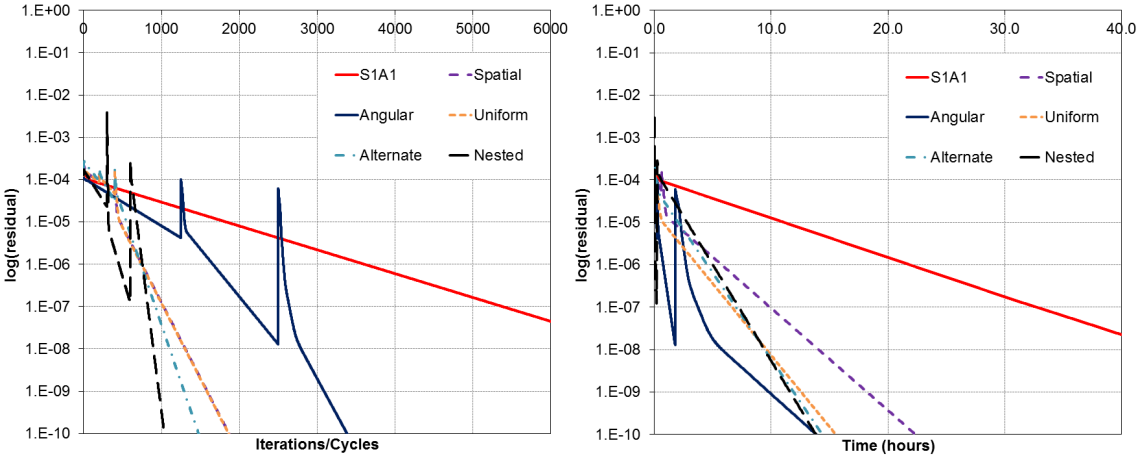



Figure 14: Radiative intensity convergence history per number of iterations and wall-clock time for different combined FMG-FAS agglomeration multigrid schemes.

## 5 CONCLUSIONS

In this work the development and comparison of different spatial/angular agglomeration multigrid schemes for the acceleration of FVM radiative heat transfer computations, was reported. It was based upon a previous study of the authors [6, 7], incorporating though further enhancements, namely different sequences of spatial and angular coarser resolutions, different V-cycle types, a full-coarsening directional agglomeration strategy and a combined FMG-FAS approach. Based on the results presented in the previous Section, the following conclusions can be extracted: a) Full-coarsening directional agglomeration should be the preferred choice in test cases using hybrid unstructured grids. Unlike the isotropic method, the directional one leads to a deeper reduction of DoFs, while simultaneously it preserves more accurately the topology of the initial grid; as a result a greater efficiency improvement is achieved. b) The V(2,0)-cycle type appears to be superior, compared to the rest tested ones (V(1,0), V(1,1) and V(2,1)), independently of the implemented FAS scheme (only-spatial, only-angular, uniform, alternate or nested). Furthermore, the relaxation after the prolongation process, defined by the latter two types (V(1,1) and V(2,1)), seems not to contribute at all to the computational acceleration. c) The nested spatial/angular agglomeration multigrid approach is revealed to be much more efficient than the rest ones, while the next most effective appears to be the alternate scheme. d) Considering that the aforementioned schemes derived higher acceleration compared to the only-spatial one, as well as the impressive improvement entailed by the only-angular approach, the significant contribution of the angular extension of the mul-

tigrid technique is confirmed. e) Independently of the implemented multigrid scheme (only-spatial, only-angular, uniform, alternate or nested), additional acceleration can be achieved with the implementation of the combined FMG-FAS approach.

## REFERENCES


[1] G.N. Lygidakis, S.S. Sarakinos, I.K. Nikolos, Comparison of different agglomeration multigrid schemes for compressible and incompressible flow simulations, *Advances in Engineering Software*, in press, 2016.

[2] A. Brandt, Multigrid techniques with applications to fluid dynamics: 1984 guide, *VKI Lecture Series*, 1-176, 1984.

[3] J. Blazek, *Computational Fluid Dynamics: Principles and Applications*, Kidlington, Elsevier Science, 2001.

[4] G.N. Lygidakis, I.K. Nikolos, Numerical analysis of flow over the NASA common research model (CRM) using the academic CFD code Galatea, *ASME Journal of Fluids Engineering*, vol. **137**, 071103, 2015.

[5] H. Nishikawa, B. Diskin, Development and application of parallel agglomerated multigrid methods for complex geometries, *20th AIAA Computational Fluid Dynamics Conference,* Honolulu, Hawaii, USA, June 27-30, 2011, AIAA 2011-3232.

[6] G.N. Lygidakis, I.K. Nikolos, Using a parallel spatial/angular agglomeration multigrid scheme to accelerate the FVM radiative heat transfer computation – part I: Methodology, *Numerical Heat Transfer, Part B: Fundamentals*, **66**, 471-497, 2014.

[7] G.N. Lygidakis, I.K. Nikolos, Using a parallel spatial/angular agglomeration multigrid scheme to accelerate the FVM radiative heat transfer computation – part II: Numerical results, *Numerical Heat Transfer, Part B: Fundamentals*, **66**, 498-525, 2014.

[8] G. Kim, S. Kim, Y. Kim, Parallelized unstructured-grid finite volume method for modeling radiative heat transfer, *Journal of Mechanical Science and Technology*, **19**, 1006-1017, 2005.

[9] R. Capdevila, C.D. Perez-Segarra, A. Oliva, Development and comparison of different spatial numerical schemes for the radiative transfer equation resolution using three-dimensional unstructured meshes, *Journal of Quantitative Spectroscopy and Radiative Transfe*r, **111**, 264-273, 2010.

[10] G.N. Lygidakis, I.K. Nikolos, Using a high-order spatial/temporal scheme and grid adaptation with a finite-volume method for radiative heat transfer, *Numerical Heat Transfer, Part B: Fundamentals*, **64**, 89-117, 2013.

[11] G.N. Lygidakis, I.K. Nikolos, Improving the accuracy of a finite-volume method for computing radiative heat transfer in three-dimensional unstructured meshes, *3rd South-East European Conference on Computational Mechanics, ECCOMAS-IACM Special Interest Conference*, Kos, Greece, June 12-14, 2013, 599-620.

[12] P. Hassanzadeh, G.D. Raithby, Finite-volume solution of the second order radiative transfer equation: Accuracy and solution cost, *Numerical Heat Transfer, Part B: Fundamentals*, **53**, 374-382, 2008.

[13] M.H. Lallemand, *Schemas decentres multigrilles pour la resolution des equations d' Euler en elements finis*, PhD Thesis, Universite de Provence, France, 1988.

[14] D.J. Mavriplis, Directional coarsening and smoothing for anisotropic Navier-Stokes problems, *Electronic Transactions in Numerical Analysis*, **6**, 182-197, 1997.

[15] N.K. Lambropoulos, D.G. Koubogiannis, K.C. Giannakoglou, Acceleration of a Navier-Stokes equation solver for unstructured grids using agglomeration multigrid and parallel processing, *Computer Methods in Applied Mechanics and Engineering*, **193**, 781-803, 2004.

[16] H. Nishikawa, B. Diskin, J.L. Thomas, Recent advances in agglomerated multigrid, *15th AIAA Aerospace Sciences Meeting,* Grapevine, Texas, USA, January 7-10, 2013, AIAA 2013-0863.

[17] V. Hannemann, Structured multigrid agglomeration on a data structure for unstructured meshes, *International Journal for Numerical Methods in Fluids*, **40**, 361-368, 2002.

[18] G.N. Lygidakis, I.K. Nikolos, Using the finite-volume method and hybrid unstructured meshes to compute radiative heat transfer in 3-D geometries, *Numerical Heat Transfer, Part B: Fundamentals*, **62**, 289-314, 2012.

[19] T.K. Kim, H. Lee, Effect of anisotropic scattering on radiative heat transfer in two-dimensional rectangular enclosures, *International Journal of Heat and Mass Transfer*, **31**, 1711-1721, 1988.

[20] B. Hunter, Z. Guo, Comparison of the discrete-ordinates method and the finite-volume method for steady-state and ultrafast radiative transfer analysis in cylindrical coordinates, *Numerical Heat Transfer, Part B: Fundamentals*, **59**, 339-359, 2011.

[21] G.D. Raithby, Discussion of the finite volume method for radiation, and its application using 3D unstructured meshes, *Numerical Heat Transfer, Part B: Fundamentals*, **35**, 389-405, 1999.

[22] H. Jimbo, R. Liming, T. Heping, Effect of anisotropic scattering on radiative heat transfer in two-dimensional rectangular media, *Journal of Quantitative Spectroscopy and Radiative Transfe*r, **78**, 151-161, 2003.

[23] G.D. Van Albada, B. Van Leer, W.W. Roberts, A comparative study of computational methods in cosmic gas dynamics, *Journal of Astrophysics and Astronomy*, **108**, 46-84, 1982.

[24] P.K. Sweby, High resolution schemes using flux limiters for hyperbolic conservation laws, *SIAM Journal for Numerical Analysis*, **21**, 995-1011, 1984.

[25] M.H. Lallemand, *Etude de schemas Runge-Kutta a 4 pas pour la resolution multigrille des equations d' Euler 2D*, Raport de Recherche, INRIA, 1988.

[26] V. Venkatakrishnan, Implicit schemes and parallel computing in unstructured grid CFD, *26th Computational Fluid Dynamics Lecture Series Program, VKI*, Rhode Saint-Genese, Belgium, March 13-17, 1995.

[27] S. Lanteri, Parallel solutions of compressible flows using overlapping and nonoverlapping mesh partitioning strategies, *Parallel Computing*, **22**, 943–968, 1996.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# LINEARISED FREQUENCY DOMAIN GUST ANALYSIS OF LARGE CIVIL AIRCRAFT

**P. Bekemeyer[1], R. Thormann[2] and S. Timme[3]**

[1]Ph.D. Student
University of Liverpool
e-mail: philipp.bekemeyer@liverpool.ac.uk

[2]Research Associate and [3]Lecturer
University of Liverpool
e-mail: {reik.thormann,sebastian.timme}@liverpool.ac.uk

**Keywords:** Large Civil Aircraft, Gust Response, Frequency Domain, Computational Fluid Dynamics, Reynolds-averaged Navier-Stokes Equations

**Abstract.** *Predicting loads due to atmospheric turbulence is crucial in the aircraft design and certification process. Efficient methods are important regarding the large number of simulations needed to cover the parameter space of e.g. Mach number, flight altitude and gust shape. Since commercial aircraft operate at transonic flight speeds, applied methods should consider aerodynamic nonlinearities such as shocks and boundary layer separation. Based on a recently presented method using frequency domain computational fluid dynamics for gust interaction, an extension towards industry-relevant three-dimensional cases is proposed. Results are shown for a large civil aircraft solving the Reynolds-averaged Navier-Stokes equations. Complex-valued surface pressures from time-linearised simulations are compared to nonlinear unsteady time-marching investigations. Responses to 1-cos gusts are obtained to discuss time histories of load factors and worst case surface pressure distributions.*

## 1 INTRODUCTION

Various dynamic response simulations need to be performed in order to achieve aircraft certification. Fast and reliable prediction tools are required to analyse several different gust parameters and flight conditions. Furthermore, aerodynamic nonlinearities, e.g. recompression shocks and shock induced boundary layer separation, should be considered to accurately predict loads. Current industrial practice is based on potential flow equations, mostly the doublet lattice method (DLM) [1], which predicts unsteady aerodynamic loads uncoupled from the steady flowfield. Gust loads are included using a frequency domain sampling and projecting surface forces on structural modes. While this offers fast predictions throughout the flight envelope, aerodynamic nonlinearities are neglected. At transonic flow conditions in particular, DLM is not predicting loads conservatively and thus DLM correction methods or more accurate tools need to be applied.

Recently computational fluid dynamics (CFD) has become more interesting for loads prediction, offering accurate results also at aforementioned nonlinear conditions. While unsteady time-marching gust simulations are technically possible also for large configurations [2], they are currently not feasible in an industrial environment due to overwhelming computational cost. However, linear frequency domain based methods, also known as time-linearised methods, have shown large runtime improvements while maintaining the accuracy of the underlying nonlinear CFD model [3]. More commonly used for forced-motion simulations, an extension towards gust excitation has been presented recently for the Reynolds-averaged Navier-Stokes (RANS) equations [4]. Results are shown for an aerofoil at subsonic and transonic flight conditions. Aerodynamic responses to 1-cos gusts are obtained by superposing several frequency domain results at discrete frequencies deploying a real-valued weighting function.

First applications for time-linearised methods were demonstrated in the field of turbomachinery [5, 6, 7]. Assuming harmonic blade motion of small amplitude, the Euler equations are linearised around a nonlinear steady-state solution. Solving then for the first harmonic is offering huge time savings compared to a time-marching approach [8, 9]. Considering external flows, initial results for an aerofoil are presented in [10]. Forced motion responses are shown for an aerofoil, wing and an aircraft in [11], while a delta wing under small harmonic oscillations of elastic modes as well as control surfaces is discussed in [12]. Compared to solving the unsteady nonlinear Euler equations in a time-marching approach, a significant speed-up is reported throughout, independent of the techniques applied to solve the resulting linear systems of equations.

Considering the RANS equations, initial work has again been published for turbomachinery applications [13]. Stall flutter including large separation regions inside a blade cascade is investigated with good agreement to experimental data. For external flows, analysing forced-motion responses, time-saving factors between one and two orders of magnitude have been demonstrated for aerofoils and wings [14, 15]. Regarding a full civil aircraft at cruise conditions speed-up of nearly two orders of magnitude has been reported [16].

In this paper an application of the linear frequency domain method for gust solving the RANS equations is demonstrated for a full aircraft configuration including fuselage, wing, tail and nacelles. Unsteady results are presented for a transonic Mach number after a steady elastic trimming process. Aerodynamic responses to sinusoidal gusts are compared to their time domain counterparts while discussing accuracy and efficiency of both methods. Finally, responses to several 1-cos gusts, based on international certification requirements, are evaluated in a very efficient way by exploiting the consolidated frequency domain technique.

## 2 METHODS

The nonlinear governing equation in semi-discrete form is written as

$$\dot{\mathbf{w}} = \mathbf{R}(\mathbf{w}, \mathbf{v}_\mathrm{g}) \tag{1}$$

where $\mathbf{w}$ is the state-space vector of unknowns, $\mathbf{R}$ denotes the nonlinear residual corresponding to the unknowns and $\mathbf{v}_\mathrm{g}$ describes external disturbances due to gusts.

The difference between an equilibrium solution $\mathbf{w}_0$ and the instantaneous state-space vector $\mathbf{w}$ is introduced as

$$\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_0 \tag{2}$$

and accordingly for external disturbances. A first order Taylor expansion is used to express the residual in Eq. (1) around the equilibrium point assuming small pertubations

$$\Delta\dot{\mathbf{w}} = \mathbf{R}(\mathbf{w}_0, \mathbf{v}_{\mathrm{g}0}) + \frac{\partial \mathbf{R}}{\partial \mathbf{w}}\Delta\mathbf{w} + \frac{\partial \mathbf{R}}{\partial \mathbf{v}_\mathrm{g}}\Delta\mathbf{v}_\mathrm{g} \tag{3}$$

where $A = \frac{\partial \mathbf{R}}{\partial \mathbf{w}}$ denotes the Jacobian matrix. The first term on the right-hand side of the latter equation is by definition zero and equivalent to the nonlinear steady flow solution that accounts for aerodynamic nonlinearities.

Subsequently, the system is transferred into frequency domain assuming the disturbance vector $\Delta\mathbf{w}$ and external excitation vector $\Delta\mathbf{v}_\mathrm{g}$ change harmonically in time. After rearranging, Eq. (3) becomes

$$\left(A - i\omega I\right)\hat{\mathbf{w}} = -\Xi(\omega)\frac{\partial \mathbf{R}}{\partial \mathbf{v}_\mathrm{g}}\hat{\mathbf{v}}_\mathrm{g} \tag{4}$$

with $\hat{\mathbf{w}}$ and $\hat{\mathbf{v}}_\mathrm{g}$ as complex-valued Fourier coefficients.

The response to an arbitrary time domain signal is obtained by superposing frequency domain responses with the complex-valued weights denoted $\Xi(\omega)$ [1]. In Fig. 1 the time and frequency domain representations for three different excitation types are shown. While the sinusoidal excitation only needs one single frequency domain solve, the two other signals affect an infinite range of frequencies. Comparing the 1-cos and pulse excitation it can be seen that the pulse omits zero amplitudes, making it possible to excite all frequencies of interest with a single unsteady time-marching simulation. Obtained frequency domain results are then used to reconstruct the time history of the loads by applying an incomplete inverse Fourier transform. Furthermore, it is also possible to get the time history of the complete flowfield.

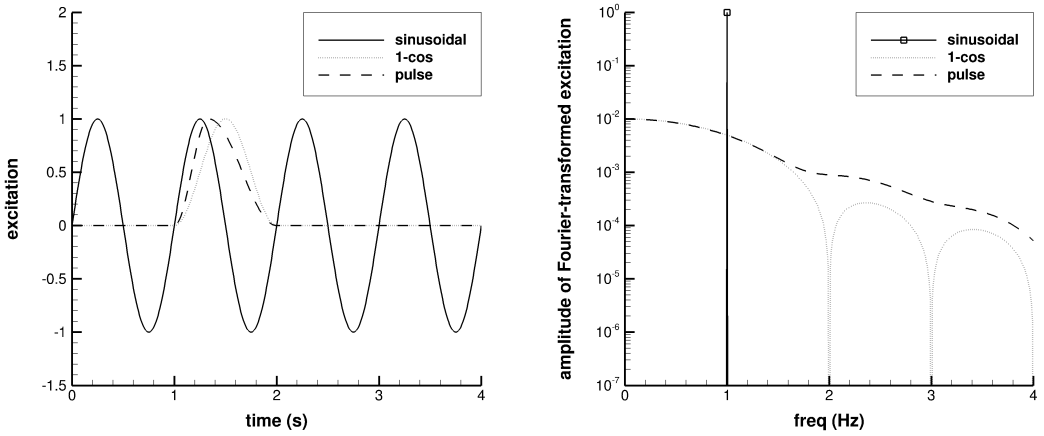
The right-hand side in Eq. (4) can be modified by applying the chain rule. Neglecting the weighting factor $\Xi(\omega)$ it becomes

$$\frac{\partial \mathbf{R}}{\partial \mathbf{v}_\mathrm{g}}\hat{\mathbf{v}}_\mathrm{g} = \frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}}\frac{\partial \dot{\mathbf{x}}}{\partial \mathbf{v}_\mathrm{g}}\hat{\mathbf{v}}_\mathrm{g} \tag{5}$$

where $\dot{\mathbf{x}}$ denotes artificial mesh velocities applied to model the gust during the CFD calculation using the field velocity approach [17]. Since the relation between gust disturbance $\mathbf{v}_\mathrm{g}$ and artificial mesh velocity $\dot{\mathbf{x}}$ is simply

$$\dot{\mathbf{x}} = -\mathbf{v}_\mathrm{g} \tag{6}$$

[1] Compared to previously presented work in [4] this paper applies complex-valued weighting functions. While resulting in a slightly different analytical gust expression, this offers a further improvement in terms of computational efficiency since samples at discrete frequencies can be re-used to reconstruct an arbitrary time signal.




(a) Time domain (b) Frequency domain

Figure 1: Excitation representation in time and frequency domain

Eq. (5) becomes

$$\frac{\partial \mathbf{R}}{\partial \mathbf{v}_\mathrm{g}} \hat{\mathbf{v}}_\mathrm{g} = -\frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}} \hat{\mathbf{v}}_\mathrm{g} \tag{7}$$

The derivative in Eq. (7) is computed using a finite difference evaluation, applied around the equilibrium point, with a known gust shape vector $\hat{\mathbf{v}}_\mathrm{g}$. Thus Eq. (7) becomes

$$\frac{\partial \mathbf{R}}{\partial \mathbf{v}_\mathrm{g}} \hat{\mathbf{v}}_\mathrm{g} = \frac{\mathbf{R}(+\varepsilon \mathbf{v}_\mathrm{g}) - \mathbf{R}(-\varepsilon \mathbf{v}_\mathrm{g})}{2\varepsilon} \tag{8}$$

with $\varepsilon$ denoting the finite difference step size. Two additional residual solves are necessary to construct the right-hand side before solving the linear system, while forming and storing the matrix $\frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}}$ explicitly can be avoided.

Furthermore, an analytical description for the gust vector is introduced as

$$\hat{\mathbf{v}}_\mathrm{g}(\mathbf{x}, \omega) = v_{gz} e^{i\boldsymbol{\varphi}(\mathbf{x},\omega)} \tag{9}$$

where $v_{gz}$ and $\boldsymbol{\varphi}(\mathbf{x}, \omega)$ are the constant gust amplitude in z-direction and the phase shift at every mesh point, respectively. The phase shift can either be obtained from a Fourier analysis of the time domain signal or more elegantly using the analytical expression

$$\boldsymbol{\varphi}(\mathbf{x}, \omega) = (\mathbf{x} + x_0)\, \omega \tag{10}$$

where $x_0$ represents the distance between gust and aircraft. The relation between the angular frequency $\omega$ and gust length of interest $L_g$ is simply

$$\omega = \frac{2\pi U_\infty}{L_g} \tag{11}$$

with $U_\infty$ as freestream velocity. Similar expressions can be derived for horizontal gusts by considering the corresponding amplitude and modifying the phase shift accordingly.

Results are produced using the DLR-TAU code [18] solving the RANS equations in conjunction with the Spalart-Allmaras turbulence model [19]. Fluxes are discretised applying the

Jameson-Schmidt-Turkel scalar artificial dissipation scheme [20]. Steady-state solutions are obtained utilising the backward Euler method with LU-SGS iteration [21]. For steady-state as well as unsteady simulations a 2v multigrid scheme is employed. Nonlinear time-marching solutions are computed using dual time-stepping of second order accuracy considering 128 steps per period. In addition, a Cauchy convergence criterion with an abort value of $10^{-8}$ for the relative error on drag coefficient is used to speed-up nonlinear time domain simulations. Gusts are introduced by adding an artificial mesh velocity during each time step which is prescribed according to the investigated gust shape [17]. Convergence for all linearised systems is achieved with a preconditioned generalised conjugate residual solver with deflated restarting [22]. For preconditioning a block incomplete lower upper factorisation of the approximate Jacobian matrix A with zero level of fill-in is applied [23]. The number of Krylov vectors necessary to solve each linear system is chosen based on engineering judgement and the guidelines published in [22], resulting in 100 vectors of which 20 are part of the deflated restarting process.

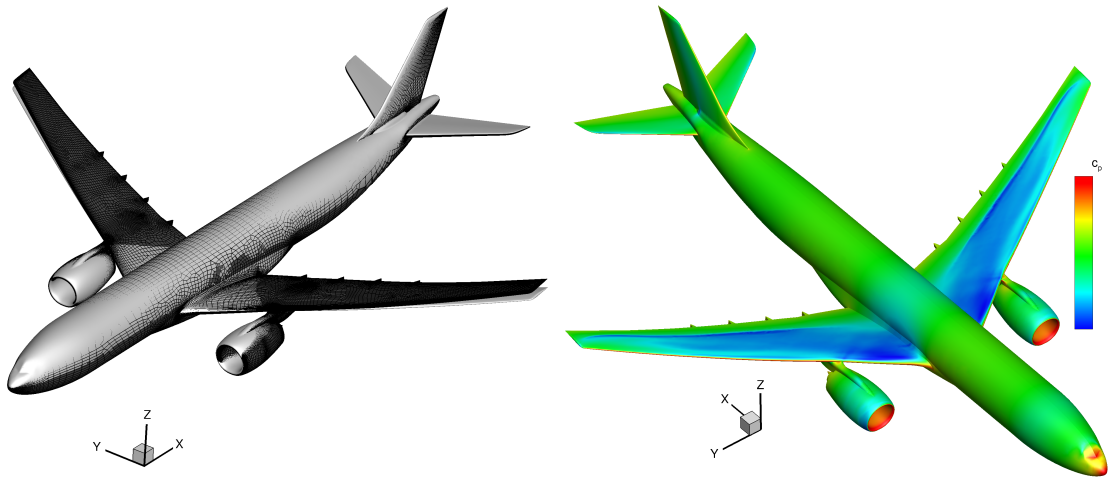
## 3 RESULTS

Results considering a large civil aircraft encountering either sinusoidal or 1-cos gusts are presented. First, convergence behaviour of the linear system is analysed and the influence of the finite difference step size $\varepsilon$ is assessed. Then, several frequency domain gust responses at different frequencies are compared to corresponding nonlinear time domain simulations to outline the capability of the method for complex three-dimensional geometries. Therefore, complex surface pressure distributions are evaluated as well as representative sections on the wing and elevator. Finally, different 1-cos gust signals, chosen according to international certification requirements, are discussed. Besides time histories for the load factor $n_z$, also instantaneous surface pressure distributions corresponding to maximum load factor are reconstructed.

The investigated test case is a civil aircraft with a wingspan of approximately 60 m including elevator and fin. The computational mesh comprises nearly 8 million points of which 130,000 are on the surface. During the steady-state simulation at transonic flight conditions an elastic trimming is performed so that the lift balances the weight and zero pitching moment occurs. Elastic effects are captured using the first 94 structural modes while rigid-body modes are neglected. The trimming process is based on Broyden's method and adjusts angle of attack and elevator deflection iteratively until the desired coefficients are reached. Within each iterative trimming step, surface loads are calculated and the elastic deformation is updated accordingly. Subsequently, the density residual is driven to converge seven order of magnitude. The resulting surface shape in comparison to the undeformed aircraft is shown in Fig. 2(a). The most amplified structural modes are first wing bending and first wing twist, causing a decrease in sectional lift towards the wing tip. The corresponding surface pressure distribution is displayed in Fig. 2(b) with a strong shock visible along the wing span at roughly 70% chord length. On the elevator a suction area around the leading edge, but no shock formation, is present.

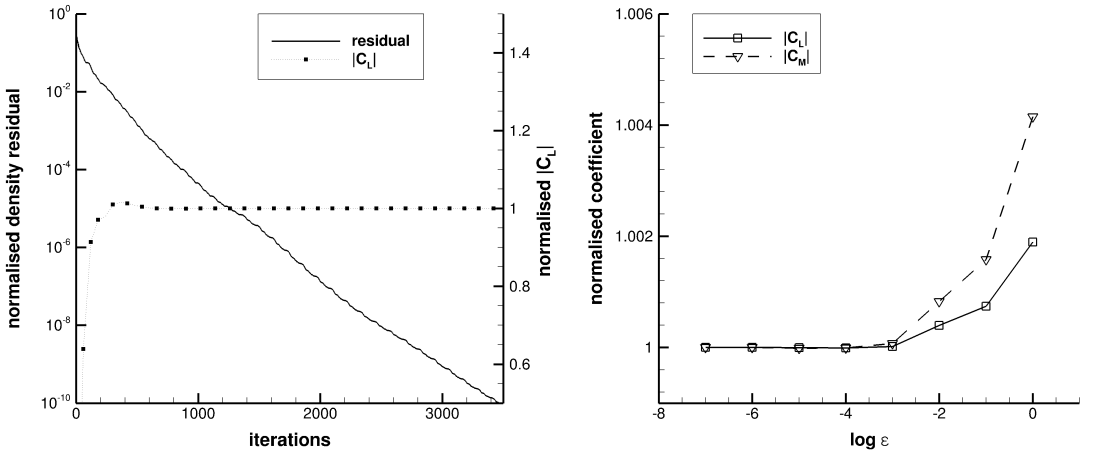
### 3.1 Investigation of Numerical Accuracy and Computational Cost

The convergence behaviour of the frequency domain solve is investigated first. A gust length of $L_g = 100$ m and an amplitude of $0.01\%$ of the freestream velocity is chosen to ensure a dynamically linear response. While the linearity criterion can be relaxed as shown in Sec. 3.2, it is selected here to demonstrate the exact reproduction between time and frequency domain results. The convergence of the density residual, shown in Fig. 3(a), together with the magnitude of the lift coefficient normalised by its final value, confirms that even for this complex geometry




(a) Deformed and undeformed surface (b) Steady surface pressure distribution

Figure 2: Deformed surface and steady-state surface pressure coefficient for civil aircraft



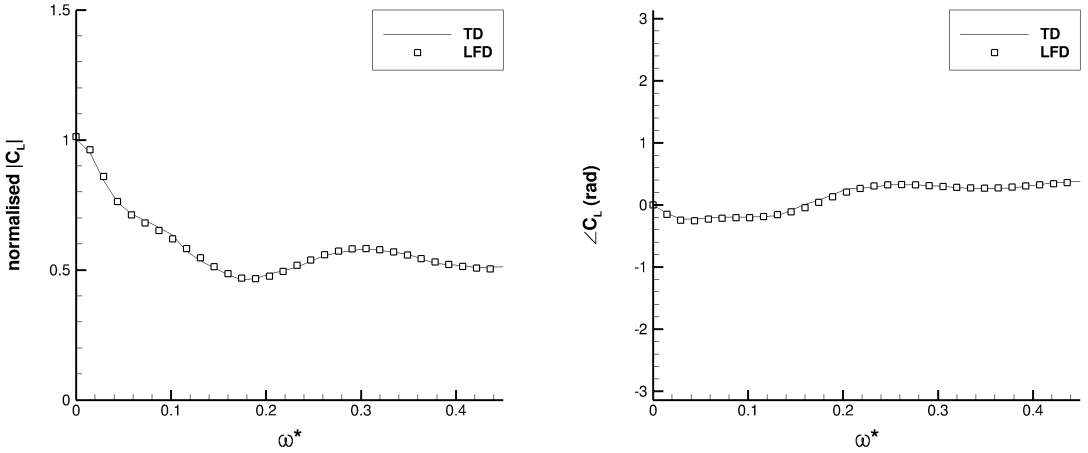

(a) Density residual and $|C_L|$ over iterations (b) Influence of finite difference step size $\varepsilon$

Figure 3: Numerical investigation of frequency domain gust approach

it is possible to converge to machine precision. The lift coefficient remains unchanged once the residual has converged four orders of magnitude. If only integrated loads are of interest, this offers time savings of additional 50% compared to aiming for full convergence. Subsequently, the abort criteria of six orders of magnitude based on the density residual is used, resulting in converged solutions for integrated loads as well as surface pressures.

The influence of the finite difference step size $\varepsilon$ for forming the right-hand side in Eq. (8) is analysed next, investigating the same test case. The magnitudes of lift and pitching moment, presented in Fig. 3(b), are both normalised so that they converge towards one for small $\varepsilon$. Below $\varepsilon = 10^{-3}$ both coefficients are independent of the step size, while larger step sizes cause magnitudes to increase with a higher impact on the moment. A step size of $10^{-4}$ is then used throughout, ensuring results are independent of the step size.

The proposed method is validated at several frequencies by comparing frequency response functions of lift coefficient between time domain (TD) and linear frequency domain (LFD).



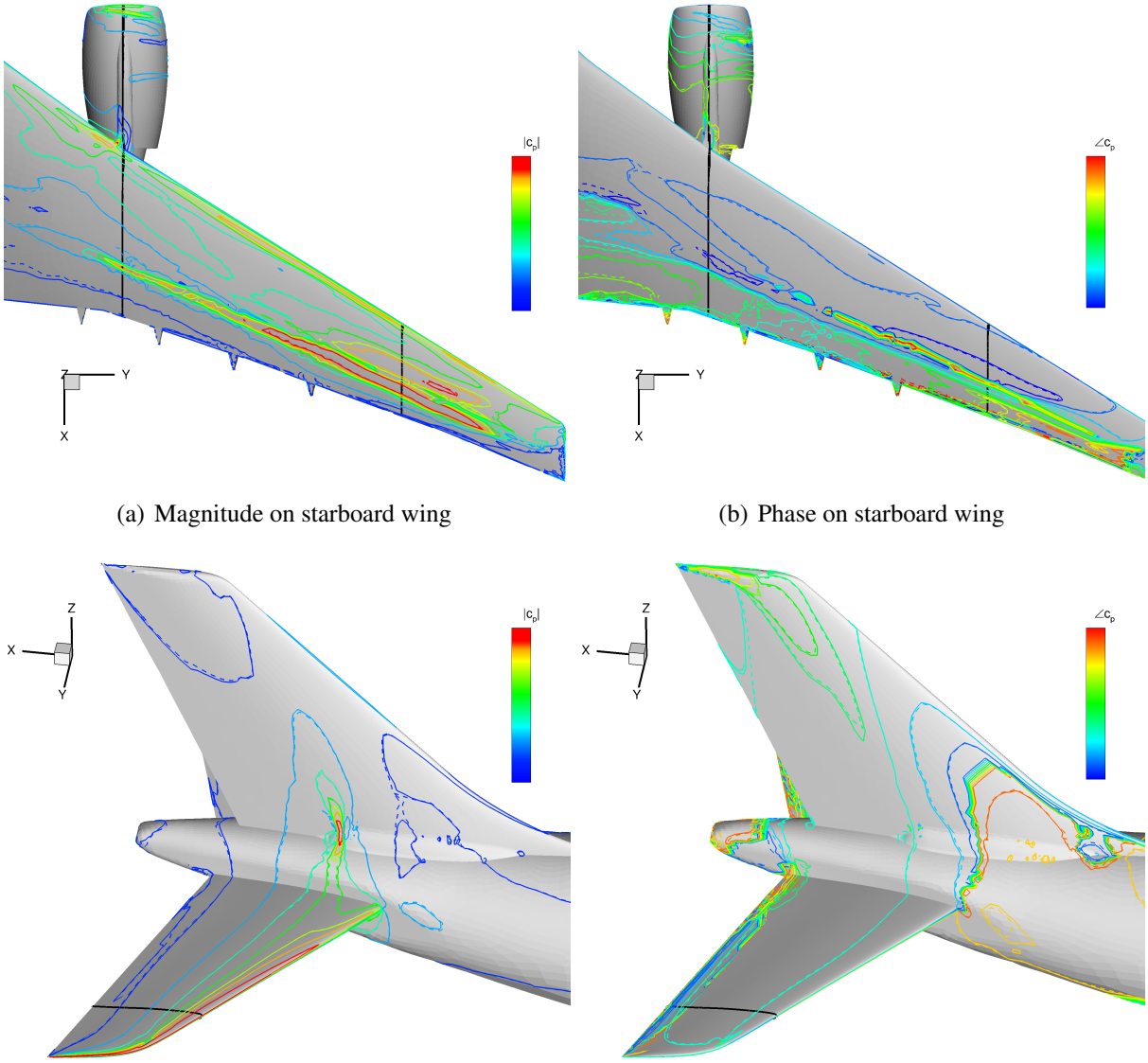

(a) Magnitude of lift frequency response function (b) Phase of lift frequency response function

Figure 4: Complex-valued frequency response functions of lift for time and frequency domain

Instead of producing time domain solutions separately for each frequency of interest, a pulse signal is used to excite all frequencies at once with one unsteady simulation. Then, a Fourier transform of the unsteady lift coefficient is performed and resulting complex-valued Fourier coefficients are weighted by the Fourier transform of the input signal. The amplitude of the excitation during the unsteady simulation is set to $0.01\%$ of the freestream velocity, ensuring a linear dynamic response. Good agreement between time and frequency domain is observed for the magnitude normalised using the quasi steady result and phase of lift coefficient as can be seen in Figs. 4(a) and 4(b).

For a further validation complex-valued surface pressure distributions are compared between both methods for a gust length of $L_g = 100$ corresponding to a reduced frequency of about $0.073$. While the frequency domain results are readily available after the linear system is solved, time domain solutions are generated in a similar fashion like the transfer function of lift coefficient. During the time-marching simulation the instantaneous surface pressure distributions are stored and then used in a Fourier transform. Results for the starboard wing are visualised in Figs. 5(a) and 5(b) for magnitude and phase with solid and dashed lines denoting time and frequency domain solutions, respectively. Good agreement between both methods at all locations, even at severe flow topologies, is obtained. The highest amplified region is around the shock at 70% chord length on the upper surface again with no differences between the simulations. Also around the wing-pylon junction, causing complex flow topologies due to vortices, excellent agreement is observed. Some minor discrepancies arise between the wing-fuselage junction in magnitude as well as phase. In Figs. 5(c) and 5(d) results are compared for the elevator and fin with similar good agreement. Highest magnitudes are located around the leading edge caused by the suction area since no shock formation is present. The phase is nearly constant on the elevator but shows large gradients near the trailing edge and during the transition from elevator leading edge to fuselage.

Values are extracted from the wing and elevator surface to compare time and frequency domain results in more detail, the locations of which are indicated by black lines in Fig. 5. Pressure magnitudes are scaled by the maximum value of the corresponding section while the x-axis is normalised by the local chord length. The first location is at 75% semi wingspan and




(a) Magnitude on starboard wing

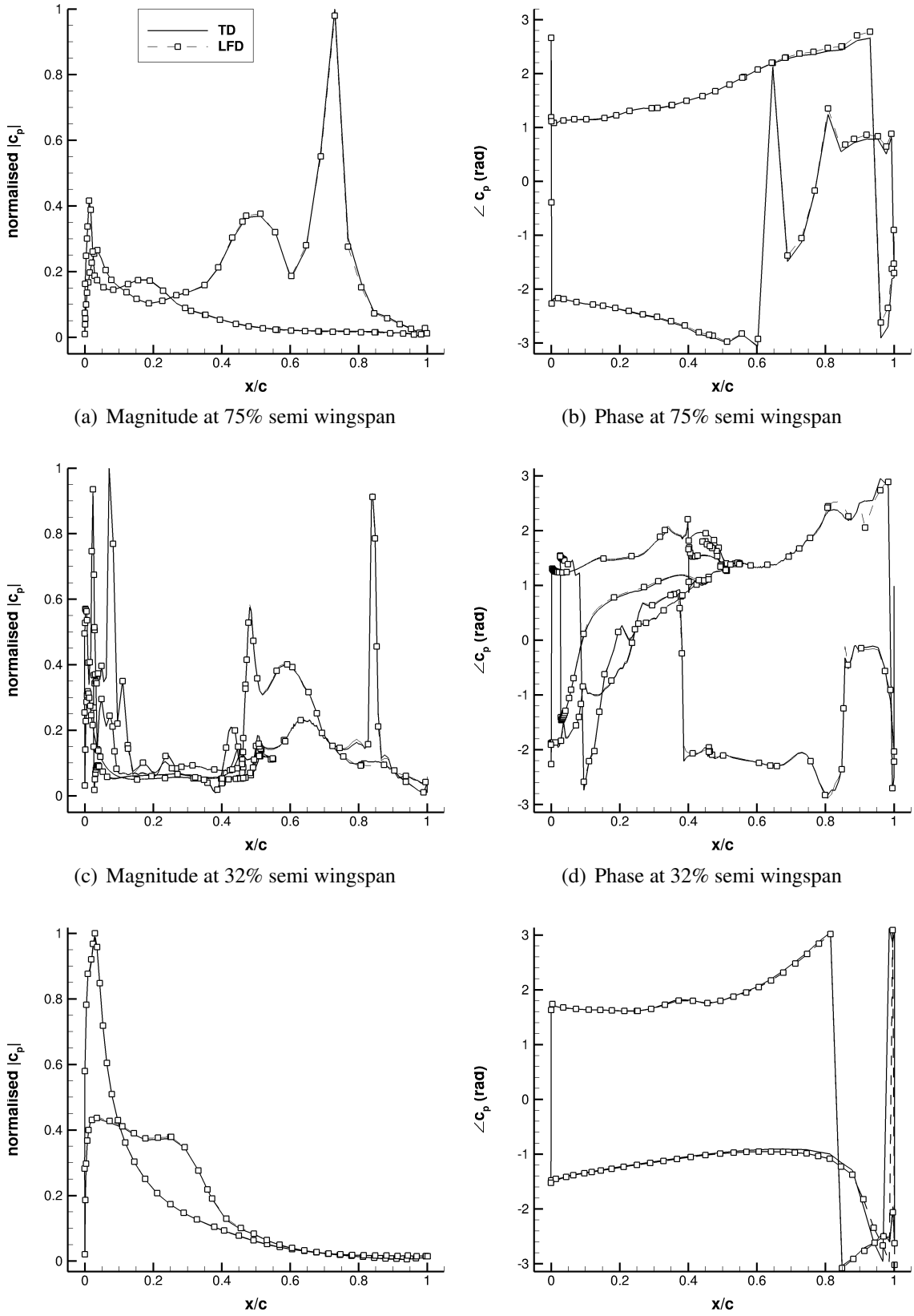
(b) Phase on starboard wing

(c) Magnitude on tail

(d) Phase on tail

Figure 5: Complex-valued surface pressure with solid and dashed lines as time and frequency domain solution, respectively. Black lines show location of extracted sections

contains complex-valued pressures around the strong shock on the upper surface. Results for magnitude and phase are shown in Figs. 6(a) and 6(b) with good agreement. A small offset at the shock location is observed for phase with values being slightly overpredicted by the frequency domain method. For the second section at 32% semi wingspan magnitude and phase are compared in Figs. 6(c) and 6(d), respectively. Even in such complex flow situation due to the junction of wing, pylon and nacelle both methods are in excellent agreement. As in the previous section, similar behaviour can be seen for phase with small differences on the wing around the shock location. The third slice at 75% semi elevator span is displayed in Figs. 6(e) and 6(f). In contrast to surface pressures on the main wing, no shock formation is present, resulting in maximum values around the suction area instead. As before, a nearly perfect agreement for magnitude between both methods is found, while minor differences for phase occur around the trailing edge since the time-marching simulation is predicting a slightly steeper gradient.




(a) Magnitude at 75% semi wingspan

(b) Phase at 75% semi wingspan

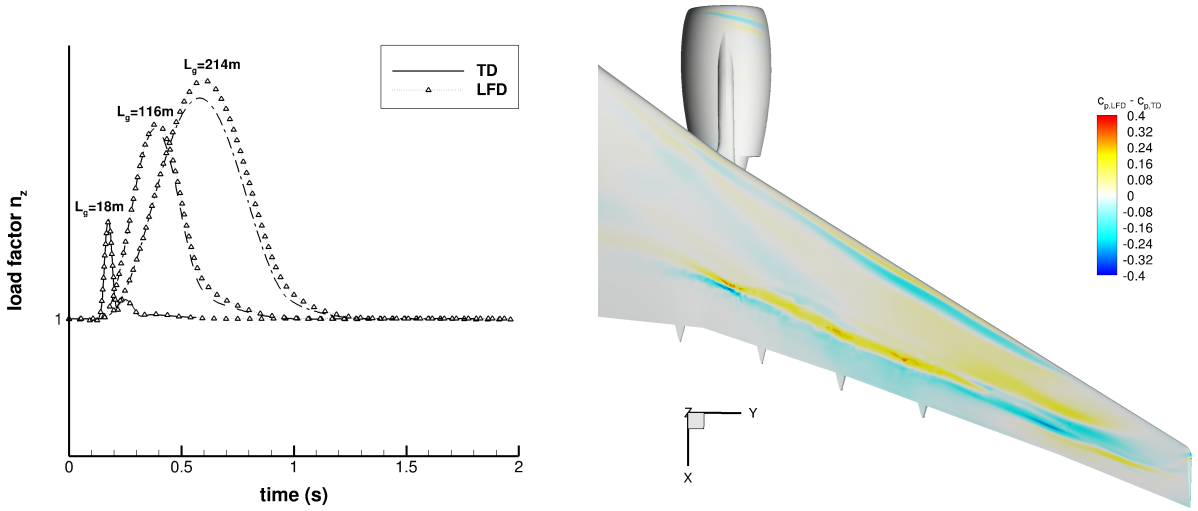
(c) Magnitude at 32% semi wingspan

(d) Phase at 32% semi wingspan

(e) Magnitude at 75% semi elevatorspan

(f) Phase at 75% semi elevatorspan

Figure 6: Complex-valued pressure distributions at wing and elevator sections




(a) Load factor time history

(b) Absolute surface pressure difference

Figure 7: Time responses to several 1-cos gusts and absolute pressure differences for $L_g = 116$ m

In terms of computational cost, obtaining a single frequency domain response is about two orders of magnitude faster than the corresponding time domain solution. However, the memory required to solve the linear system increases compared to a time-marching approach. While an unsteady RANS simulation needs to be performed for each set of parameters separately, complex-valued frequency domain results can be used in a weighted incomplete inverse Fourier transform to reconstruct time responses of integrated values as well as the whole flowfield in a rapid manner. For gust excitation it is advisable to precompute around 15 discrete frequencies to investigate an arbitrary time domain signal e.g. 1-cos gusts. The resulting computational cost is still one order of magnitude below a single unsteady investigation while offering results for all dynamic gust responses at the defined flight condition.

### 3.2 Application According to Certification Requirements

After the investigation of numerical accuracy and computational efficiency presented above, the frequency domain gust approach is used to predict dynamic responses to 1-cos gusts in an efficient way. The linear system is solved at 15 reduced frequencies between 0 and 0.6. Responses to 1-cos gusts are obtained by applying a complex-valued weighting function as discussed in Sec. 2. The chosen gust lengths are $L_g = 18$ m, 116 m and 214 m with amplitudes as defined by the European Aviation Safety Agency in CS 25.341 [24]. Load factors are reconstructed and visualised in Fig. 7(a) for all three gust lengths. For the shortest gust length of $L_g = 18$ m excellent agreement between both methods is observed. With increasing gust length also the gust amplitude is growing, causing dynamically nonlinear responses near the maximum load factor. Since the frequency domain approach assumes a dynamically linear response, reconstructed load factors are slightly overpredicted for gust lengths of $L_g = 116$ m and 214 m. The absolute surface pressure difference for $L_g = 116$ m at maximum load factor is displayed in Fig. 7(b). The highest error occurs around the shock foot since nonlinear shock motion and a nonlinear amplitude decrease arises during the time-marching simulation. Minor differences around the leading edge at the stagnation line are also caused from the same amplitude mechanism.

## 4 CONCLUSIONS

This paper outlines a method for efficiently computing the aerodynamic response to gust encounter for industry-relevant test cases. The Reynolds-averaged Navier-Stokes equations are linearised to obtain responses to sinusoidal gust excitation, a method previously presented for aerofoils. Arbitrary time domain signals such as 1-cos gusts can be simulated using a complex-valued weighting in combination with superposition of responses at discrete frequencies. Computational cost is reduced by more than two orders of magnitude compared to equivalent unsteady nonlinear time-marching simulations.

The method is applied to a large civil aircraft including flow through engines, elevator and fin at transonic cruise conditions. The steady-state 1g flight shape is generated through an iterative elastic trimming procedure. Responses at various frequencies are compared to a pulse excitation, validating the method at all frequencies of industrial interest. Complex surface pressure distributions are discussed with good agreement on the wing and tail globally as well as in specific wing and elevator sections. Finally, characteristic 1-cos gusts are reconstructed and results are matched to time-marching simulations. Although results are produced for a very complex geometry including pylon and nacelles, an excellent agreement between the time and frequency domain methods is obtained. This includes the global lift coefficient as well as surface pressure distributions, thus demonstrating the maturity of the method.

Work to extend the method to generate a reduced-order model for the three-dimensional test case is currently in progress. Also, it is desirable to incorporate nonlinear aerodynamic effects to increase the accuracy of predicted loads further at transonic conditions. In addition, introducing structural and rigid-body motions in the response simulation is desirable to account for elastic surface deformation and changes in the flight path during the aircraft gust interaction.

## ACKNOWLEDGEMENTS

The research leading to these results was co-funded by Innovate UK, the UK's innovation agency, as part of the Enhanced Fidelity Transonic Wing project.

## REFERENCES

[1] Albano, E. and Rodden, W. P., "A doublet lattice method for calculating lift distribution on oscillating surfaces in subsonic flow," *AIAA Journal*, Vol. 2, No. 7, 1969, pp. 279–285.

[2] Reimer, L., Ritter, M., Heinrich, R., and Krüger, W., "CFD-based Gust Load Analysis for a Free-flying Flexible Passenger Aircraft in Comparison to a DLM-based Approach," *22nd AIAA Computational Fluid Dynamics Conference*, 2015, AIAA 2015-2455.

[3] Thormann, R., Nitzsche, J., and Widhalm, M., "Time-linearized Simulation of Unsteady Transonic Flows with Shock-Induced Separation," *European Congress on Computational Methods in Applied Sciences and Engineering*, 2012, ECCOMAS 2012.

[4] Bekemeyer, P. and Timme, S., "Reduced Order Gust Response Simulation using Computational Fluid Dynamics," *57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2016, AIAA 2016-1485.

[5] Whitehead, D. S. and Grant, J. R., "Force and Moment Coefficients of High Deflection Cascades," *2nd International Symposium on Aeroelasticity in Turbomachinery*, 1981.

[6] Verdon, J. M. and Caspar, J. R., "Development of a Linear Unsteady Aerodynamic Analysis for Finite-Deflection Subsonic Cascades," *AIAA Journal*, Vol. 20, No. 9, 1982, pp. 1259–1267.

[7] Verdon, J. M. and Caspar, J. R., "A Linearized Unsteady Aerodynamics Analysis for Transonic Cascades," *Journal of Fluid Mechanics*, Vol. 149, 1984, pp. 403–429.

[8] Hall, K. C. and Clark, W. S., "Linearized Euler Predictions of Unsteady Aerodynamic Loads in Cascades," *AIAA Journal*, Vol. 31, No. 3, 1993, pp. 540–550.

[9] Hall, K. C., Clark, W. S., and Lorence, C. B., "A Linearized Euler Analysis of Unsteady Transonic Flows in Turbomachinery," *Journal of Turbomachinery*, Vol. 116, No. 3, 1994, pp. 477–488.

[10] Mortchéléwicz, G. D., "Application des Équations d'Euler Linéarisées á la Prévision du Flottement," *85th Meeting of the AGARD Structures and Materials Panel*, 1997, AGARD Report 822: Numerical Unsteady Aerodynamic and Aeroelastic Simulation.

[11] Widhalm, M., Dwight, R. P., Thormann, R., and Hübner, A. R., "Efficient Computation of Dynamic Stability Data with a Linearized Frequency Domain Solver," *5th European Conference on Computational Fluid Dynamics*, 2010, ECCOMAS CFD 2010.

[12] Weishäupl, C. and Laschka, B., "Small Disturbance Euler Simulations for Delta Wing Unsteady Flows due to Harmonic Oscillations," *Journal of Aircraft*, Vol. 41, No. 4, 2004, pp. 782–789.

[13] Clark, W. S. and Hall, K. C., "A Time-Linearized Analysis of Stall Flutter," *Journal of Turbomachinery*, Vol. 122, No. 3, 2000, pp. 467–476.

[14] Pechloff, A. and Laschka, B., "Small Disturbance Navier-Stokes Computations for Low-Aspect-Ratio Wing Pitching Oscillations," *Journal of Aircraft*, Vol. 47, No. 3, 2010, pp. 737–753.

[15] Thormann, R. and Widhalm, M., "Linear-Frequency-Domain Predictions of Dynamic-Response Data for Viscous Transonic Flows," *AIAA Journal*, Vol. 51, No. 11, 2013, pp. 2540–2557.

[16] Thormann, R. and Widhalm, M., "Forced Motion Simulation using a Linear Frequency Domain Solver for a Generic Transport Aircraft," *International Forum on Aeroelasticity and Structural Dynamics (IFASD)*, 2013, IFASD 2013-17A.

[17] Parameswaran, V. and Baeder, J. D., "Indicial aerodynamics in compressible flow-direct computational fluid dynamic calculations," *Journal of Aircraft*, Vol. 34, No. 1, 1997, pp. 131–133.

[18] Schwamborn, D., Gerhold, T., and Heinrich, R., "The DLR TAU-Code: Recent Applications in Research and Industry," *European Conference on Computational Fluid Dynamics*, 2006, ECCOMAS CFD 2006.

[19] Spalart, P. R. and Allmaras, S. R., "A One-Equation Turbulence Model for Aerodynamic Flows," *Recherche Aerospatiale*, Vol. 1, 1994, pp. 5–21.

[20] Jameson, A., Schmidt, W., and Turkel, E., "Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes," *AIAA Journal*, 1981, pp. 1981–1259.

[21] Dwight, R., "An Implicit LU-SGS Scheme for Finite-Volume Discretizations of the Navier-Stokes Equations on Hybrid Grids," *DLR-FB-2005-05*, 2006.

[22] Xu, S., Timme, S., and Badcock, K. J., "Krylov Subspace Recycling for Linearised Aerodynamics Analysis using DLR-TAU," *International Forum on Aeroelasticity and Structural Dynamics (IFASD)*, 2015, IFASD-2015-186.

[23] Saad, Y., *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2nd ed., 2003.

[24] European Aviation Safety Agency, "Certification Specifications for Large Aeroplanes (CS-25)," 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# CONSISTENT NON-REFLECTING BOUNDARY CONDITIONS FOR BOTH STEADY AND UNSTEADY FLOW SIMULATIONS IN TURBOMACHINERY APPLICATIONS

**Daniel Schlüß, Christian Frey, Graham Ashcroft**

German Aerospace Center (DLR)
Institute of Propulsion Technology
Linder Höhe, 51147 Cologne, Germany
email: daniel.schluess@dlr.de, christian.frey@dlr.de, graham.ashcroft@dlr.de

**Keywords:** Non-Reflecting Boundary Conditions, Turbomachinery, Steady and Unsteady CFD

**Abstract.** *This paper presents the implementation of a non-reflecting boundary condition for steady and unsteady turbomachinery flow computations. Here, the truncation of the computational domain can lead to spurious numerical reflections due to artificial open boundary surfaces. To face this issue, Giles introduces a popular set of non-reflecting boundary conditions for turbomachinery applications. Whereas the steady formulation is exact within the linearisation approach, Giles suggests an approximate boundary condition for unsteady simulations. Resulting from this approximation, the unsteady boundary conditions are not perfectly non-reflecting. Thus, steady and time averaged unsteady flow solutions do not necessarily coincide, even if the flow field contains no unsteadiness.*

*We suggest a single boundary condition formulation suitable for steady and unsteady simulations. This approach applies a modal decomposition and, thus, undesired incoming modes can easily be ruled out. Steady modes can be handled as in the steady case. Apart from the assumptions made by starting from the linearised, two-dimensional Euler equations, this approach does not require any further approximation, and, therefore, the time-averaged unsteady and the steady solutions coincide in the limit of steady flows.*

*We apply and assess the presented boundary condition in steady and unsteady computations of two turbomachinery test cases.*

## 1 INTRODUCTION

Computational fluid dynamics (CFD) based upon the Reynolds-averaged Navier-Stokes equations (RANS) has become a vital tool for both design of and research on turbomachinery applications in the aerospace, energy and automotive industries. Steady computations remain the backbone of industrial design processes especially due to their extensive utilization in shape optimization methods. Unsteady turbomachinery flow simulations, on the other hand, attain increasing importance for understanding blade row interactions which is needed to achieve higher aerodynamic loads and closer blade row spacings to meet the desire for more efficient and lighter engines. Moreover, these design trends require additional activities in fluid dynamics related disciplines like aeroelasticity and aeroacoustics, which again heavily rely on unsteady CFD [1, 2].

The quality of flow simulations depends, beside the choice of an appropriate closure model for the RANS equations and the properties of the numerical scheme, directly on the boundary conditions imposed. Straightforward far-field or simple 1D characteristic boundary conditions can lead to spurious, numerical reflections deteriorating the flow solution inside. Thus, many aerodynamic problems, such as the flow around an airfoil or wing, are commonly simulated in very large computational domains to minimise the impact of the far-field boundary conditions on the region of interest.

In turbomachinery flows, however, this is usually not possible. Firstly, instead of the whole machine, only single components or certain sets of stages or blade rows are considered which leads to artificial, open boundaries rather close to the blades. Additionally in steady turbomachinery CFD, adjacent blade rows in their respective relative frames of reference are coupled via so-called mixing planes [3]. These mixing planes pose boundary conditions to the adjacent domains which strengthens the need for non-reflecting boundary conditions (NRBC) in turbomachinery applications.

Numerical reflections, that emerge when applying less advanced boundary conditions, can disturb the flow field in the neighbourhood of a blade leading to an incorrect surface pressure distribution and hence aerodynamic work or losses. We will give an example for this later in the application section. In particular, reflecting boundary conditions can deteriorate the prediction of aeroelastic or aeroacoustic phenomena as shown by Kersken et al. [4] and Ashcroft and Schulz [5].

The mathematical theory of non-reflecting boundary conditions is presented in a review paper by Hidgon [6]. Based upon this theory, Giles introduces a set of NRBC for 2D turbomachinery flows [7, 8]. Giles' approach assumes wavelike perturbations around an average flow state. Thus, the linearised Euler equations are considered neglecting viscous effects. Transforming these perturbations into the spectral, i.e. time and wavenumber, domain and exploiting an eigenvector analysis, one obtains a set of modal perturbations and their respective direction of propagation. This decomposition enables the definition of a flow state at the boundaries that does not induce undesired incoming perturbations. The straightforward implementation of these NRBC requires the frequencies and wave numbers of all perturbations at the boundary to be known. Due to the Fourier transform applied here, these boundary conditions are non-local in time and space. Therefore, Giles suggests an approximate boundary condition for unsteady computations derived from theoretical work of Engquist and Majda [9]. This boundary condition is derived from a Taylor series expansion about the ideal one-dimensional boundary condition and, thus, is local in time and space apart from the fact that it requires averaged quantities. A more detailed description of the implementation of these boundary conditions is given

in [10]. Saxer and Giles present an extension of the non-local, steady NRBC to 3D flows [11].

However, resulting from the approximation, Giles states the unsteady boundary conditions may produce significant unphysical reflections for outgoing modes with large circumferential wavenumbers [8]. Accordingly, steady solutions obtained utilizing the exact steady NRBC and time averaged unsteady flow solutions may differ even for naturally steady flows. Hagstrom proposes a higher order approach for the approximation of unsteady NRBC [12]. This method is local in time but employs a Fourier transform in space. Henninger et al. [13] apply it to acoustic and aeroelastic turbomachinery test cases and demonstrate improved non-reflecting properties compared to Giles' boundary condition.

To overcome the inconsistency of steady and unsteady NRBC and ensure their comparability in steady and time-accurate turbomachinery flow simulations, the authors propose a boundary condition formulation suitable for both. This approach applies a modal decomposition in time and circumferential direction. Thus, the zeroth harmonic in time can be handled by a steady boundary condition. Disturbances of higher harmonics are treated independently. Accordingly, undesired incoming modes can easily be ruled out. Apart from the assumptions made by starting from the linearised, two-dimensional Euler equations, this approach does not require any further approximation, and, therefore, the time-averaged unsteady and the steady solution coincide in the limit of steady flows.

In this paper, we firstly outline the theory behind NRBC and the approximation made by Giles. Subsequently, we present the implementation of an exact method and apply it to steady and unsteady turbomachinery computations. In order to validate our implementation of this boundary condition in DLR's 3D RANS solver for internal and turbomachinery flows, TRACE, we give a comparison to characteristic and 2D non-reflecting steady boundary conditions and approximate unsteady 2D NRBC. Thereby, the favourable behaviour of the exact method is demonstrated.

## 2 NON-REFLECTING BOUNDARY CONDITIONS

Before presenting our implementation of an exact boundary condition, we want to give a brief overview of one way to construct non-reflecting boundary conditions in the context of turbomachinery CFD as they are implemented in TRACE. A more elaborate description including the mathematical derivation and well-posedness analysis is given in Giles' technical report [7] and implementation details can be found in his UNSFLO report [10].

### 2.1 General approach

For many turbomachinery flows we can assume that variations in pitchwise direction predominate variations in spanwise direction [11]. Thus, we can adapt Giles' originally two-dimensional formulation for three-dimensional flows as a reasonable approximation. Here, the original technique is applied in blade-to-blade planes, i.e. the coupling of planes of constant relative channel height is neglected.

Let $q = (\varrho, u, v, w, p)$ be the vector of primitive variables with density $\varrho$, pressure $p$ and velocity components $u$,$v$ and $w$ aligned such that $u$ is normal to the boundary but in flow direction whilst $v$ and $w$ point along the boundary in pitchwise and spanwise direction, respectively. For the construction of non-reflecting boundary conditions we further consider sufficiently small perturbations around a mean flow state. So we can linearised the three-dimensional Euler equa-

tions in primitive form neglecting changes in spanwise direction:

$$\frac{\partial q}{\partial t} + A\frac{\partial q}{\partial x} + B\frac{\partial q}{\partial y} = 0 \tag{1}$$

with

$$A = \begin{pmatrix} \overline{u} & \overline{\varrho} & 0 & 0 & 0 \\ 0 & \overline{u} & 0 & 0 & 1/\overline{\varrho} \\ 0 & 0 & \overline{u} & 0 & 0 \\ 0 & 0 & 0 & \overline{u} & 0 \\ 0 & \gamma\overline{p} & 0 & 0 & \overline{u} \end{pmatrix} \text{ and } B = \begin{pmatrix} \overline{v} & 0 & \overline{\varrho} & 0 & 0 \\ 0 & \overline{v} & 0 & 0 & 0 \\ 0 & 0 & \overline{v} & 0 & 1/\overline{\varrho} \\ 0 & 0 & 0 & \overline{v} & 0 \\ 0 & 0 & \gamma\overline{p} & 0 & \overline{v} \end{pmatrix} \tag{2}$$

Overlined variables denote mean flow conditions and $\gamma$ represents the ideal gas heat capacity ratio. In the following, we only consider wave-like perturbations of the form

$$q = \mathrm{Re}\left(\hat{q}\, e^{\mathrm{i}(kx+my-\omega t)}\right) \tag{3}$$

where $x$ is again normal to the boundary in flow direction and $y$ is aligned with the boundary in pitchwise direction. Note that the latter is the circumferential direction in rotational turbomachinery flows. Then $k$ and $m$ denote wave numbers along $x$ and $y$, respectively, with angular frequency $\omega$. Though this approach seems rather restrictive, we can in fact assemble any flow state by superposition of these fundamental perturbations within the linearised theory. Substituting (3) into equation (1), we obtain:

$$\left(-\omega I + kA + mB\right)\hat{q} = 0 \tag{4}$$

With nontrivial solution $\hat{q}$, this yields the dispersion relation

$$\det\left(-\omega I + kA + mB\right) = 0 \tag{5}$$

or rearranged

$$\det\left(-\omega A^{-1} + kI + mA^{-1}B\right) = 0. \tag{6}$$

Assuming we know $\omega$ and $m$, equation (6) can be interpreted as a characteristic polynomial to the eigenvalue problem

$$\left(-\omega A^{-1} + mA^{-1}B\right) r = -kr \tag{7}$$

with eigenvalues $-k$ and right eigenvectors $r$. The 3D linearised Euler equations yield a characteristic polynomial of degree five. Accordingly, there are five eigenvalues $-k_i$ regarding their possible multiplicity. Analogously, we introduce a set of left eigenvectors $l_i$ satisfying

$$l_i\left(-\omega A^{-1} + mA^{-1}B\right) = -k_i l_i. \tag{8}$$

Note that the left eigenvectors are row vectors whereas the right eigenvectors are column vectors.

In order to derive non-reflecting boundary conditions from these mathematical considerations, it is helpful to clarify the physical interpretation of the left and right eigenvectors. As mentioned earlier, we can decompose any solution $q$ into a sum of single Fourier modes $\hat{q}$, each with distinct values of $\omega$ and $m$. Each of these modes is again a set of five fundamental waves with same $\omega$ and $m$ but (possibly) different $k$. Since the wave numbers normal to the boundary, $k_i$, are the negative eigenvalues of $(-\omega A^{-1} + mA^{-1}B)$, they allow us to distinguish whether the corresponding waves propagate into or out of the computational domain.

The change in primitive variables induced by such a wave is expressed by the respective right eigenvector $r_i$. In other words, the right eigenvectors form a set of linearly independent basis vectors and every perturbation from a mean state can be decomposed into a linear combination of right eigenvectors with weights $\alpha_i$ :

$$q = \mathrm{Re}\left(\left[\sum_{i=1}^{5} \alpha_i r_i e^{\mathrm{i}k_i x}\right] e^{\mathrm{i}(my-\omega t)}\right) \tag{9}$$

For different eigenvalues $k_i \neq k_j$, each left eigenvector $l_i$ is orthogonal to $r_j$, i.e. $l_i r_j = 0$. For eigenvalues of multiplicity larger than one, the corresponding eigenvectors can be constructed such that $l_i r_j = 0$ for $i \neq j$. Because of this perpendicularity relation, the left eigenvectors help us to determine the share of their corresponding right eigenvector in any arbitrary perturbation, i.e. $\alpha_i = l_i \hat{q}$. So the pivotal idea for the construction of non-reflecting boundary conditions is to decompose the flow state at a boundary and rule out the incoming waves. This can be done by requiring for any combination of $\omega$ and $m$

$$l_i \hat{q} = 0 \tag{10}$$

for each $l_i$ belonging to an incoming wave .

Applying the above to the three-dimensional linearised Euler equations, we obtain the following wave numbers:

$$k_{1,2,3} = \frac{\omega - m\overline{v}}{\overline{u}} \tag{11}$$

$$k_4 = \frac{(\omega - m\overline{v})\,(\overline{a}\Psi - \overline{u})}{\overline{a}^2 - \overline{u}^2} \tag{12}$$

$$k_5 = -\frac{(\omega - m\overline{v})\,(\overline{a}\Psi + \overline{u})}{\overline{a}^2 - \overline{u}^2} \tag{13}$$

with speed of sound $a$ and

$$\Psi = \begin{cases} \sqrt{\Delta} & \text{if } \Delta > 0, \\ -\mathrm{i}\ \mathrm{sign}\,(\omega - m\overline{v})\,\sqrt{-\Delta} & \text{if } \Delta < 0 \end{cases} \tag{14}$$

and

$$\Delta = 1 - \frac{(\overline{a}^2 - \overline{u}^2)\,m^2}{(\omega - m\overline{v})^2}. \tag{15}$$

The wave numbers $k_{1,2,3}$ are real and the group velocity normal to the boundary $\frac{\partial \omega}{\partial k} = u$ is positive. Thus, the respective perturbations $r_{1,2,3}$ propagate convectively in flow direction. They are incoming waves at an inflow boundary and outgoing ones at an outflow. The calculation of $k_4$ and $k_5$ requires the distinction of two cases. If $\Delta > 0$ then $\Psi$ is real. For flows that are subsonic normal to the boundary it can be shown that one wave is incoming and one is outgoing. Taking the positive branch of the root yields the wave corresponding to $k_4$ also propagates in flow direction whereas the other one travels in the opposite direction. If $\Delta < 0$, $\Psi$ becomes complex. We choose the sign of $\Psi$ such that $\mathrm{Im}(k_4)$ is real and, according to equation (3), the corresponding wave propagates downstream, which is consistent with the case $\Delta > 0$. The wave associated with $k_5$ then again travels upstream. In the particular case of

$\Delta = 0$, acoustic resonance occurs, which involves additional challenges for the construction of boundary conditions [4, 14]. But this case is not considered in the present work.

For flows that are supersonic normal to the boundary, there is no upstream running wave. Yet, there are only very few axially supersonic turbomachinery applications and, secondly, supersonic inflow and outflow boundary conditions are rather straightforward and can be found in many textbooks on CFD (e.g. [15]). Hence, this paper does not consider boundary conditions for normally supersonic flows.

To write down the eigenvectors and formulate boundary conditions based upon them, it is convenient to define

$$\lambda = \frac{m}{\omega}. \tag{16}$$

The matrix of right eigenvectors $R = (r_1\ r_2\ r_3\ r_4\ r_5)$ is defined by:

$$R(\lambda) = \begin{pmatrix} -\overline{\varrho} & 0 & 0 & \frac{\overline{\varrho}(1-(1-\overline{v}\lambda)Ma_x\Psi)}{2(1-Ma_x)} & \frac{\overline{\varrho}(1+(1-\overline{v}\lambda)Ma_x\Psi)}{2(1+Ma_x)} \\ 0 & \overline{a}\,\overline{u}\lambda & 0 & \frac{\overline{a}(1-\overline{v}\lambda)(\Psi-Ma_x)}{2(1-Ma_x)} & \frac{-\overline{a}(1-\overline{v}\lambda)(\Psi+Ma_x)}{2(1+Ma_x)} \\ 0 & \overline{a}\,(1-\overline{v}\lambda) & 0 & \frac{\overline{a}^2\left(1-Ma_x^2\right)\lambda}{2(1-Ma_x)} & \frac{\overline{a}^2\left(1-Ma_x^2\right)\lambda}{2(1+Ma_x)} \\ 0 & 0 & \overline{a} & 0 & 0 \\ 0 & 0 & 0 & \frac{\overline{\varrho}\,\overline{a}^2(1-(1-\overline{v}\lambda)Ma_x\Psi)}{2(1-Ma_x)} & \frac{\overline{\varrho}\,\overline{a}^2(1+(1-\overline{v}\lambda)Ma_x\Psi)}{2(1+Ma_x)} \end{pmatrix} \tag{17}$$

with $Ma_x$ denoting the boundary normal Mach number. From this, the significance of each right eigenvalue becomes apparent. As $r_1$ only affects the density, it constitutes an entropy perturbation. The second and third eigenvectors represent vorticity disturbances in the blade-to-blade plane and in spanwise direction, respectively. Since $\frac{\omega - m\overline{v}}{\overline{u}}$ is a triple eigenvalue of the dispersion relation (6), the determination of $r_1$, $r_2$ and $r_3$ is not unique. However, choosing them this way yields a physically vivid set of eigenvectors and their required orthogonality is evident at once. The two remaining eigenvectors correspond to upstream and downstream running acoustic perturbations, i.e. isentropic, irrotational pressure waves.

The left eigenvector matrix can be derived likewise. Another possibility to obtain the left eigenvector matrix is to invert the right eigenvector matrix:

$$L = \begin{pmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \end{pmatrix} = R^{-1} = \begin{pmatrix} \frac{-1}{\overline{\varrho}} & 0 & 0 & 0 & \frac{1}{\overline{\varrho}\,\overline{a}^2} \\ 0 & \frac{-\overline{u}\lambda}{\overline{a}} & \frac{1-\overline{v}\lambda}{\overline{a}} & 0 & \frac{-\lambda}{\overline{\varrho}\,\overline{a}} \\ 0 & 0 & 0 & \frac{1}{\overline{a}} & 0 \\ 0 & \frac{1-\overline{v}\lambda}{a} & \frac{\overline{u}\lambda}{\overline{a}} & 0 & \frac{(1-\overline{v}\lambda)\Psi}{\overline{\varrho}\,\overline{a}^2} \\ 0 & -\frac{1-\overline{v}\lambda}{a} & -\frac{\overline{u}\lambda}{\overline{a}} & 0 & \frac{(1-\overline{v}\lambda)\Psi}{\overline{\varrho}\,\overline{a}^2} \end{pmatrix} \tag{18}$$

Due to the fact that there is one outgoing wave at an inflow and four outgoing waves at an outflow, we need to extrapolate outgoing perturbations from the interior. To do so, we define characteristic variables $c = (c_1\ c_2\ c_3\ c_4\ c_5)^T$ such that these characteristics coincide with the weights $\alpha_i$ in equation (9) for plane waves running perpendicularly to the boundary. Note this is the case if $m = 0$ or equivalently $\lambda = 0$. Hence, the forward and backward transformations are given by

$$c = L_{1d}\, q \tag{19}$$

with

$$L_{1d} = L(0) = \begin{pmatrix} \frac{-1}{\overline{\varrho}} & 0 & 0 & 0 & \frac{1}{\overline{\varrho}\,\overline{a}^2} \\ 0 & 0 & \frac{1}{\overline{a}} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\overline{a}} & 0 \\ 0 & \frac{1}{a} & 0 & 0 & \frac{1}{\overline{\varrho}\,\overline{a}^2} \\ 0 & -\frac{1}{a} & 0 & 0 & \frac{1}{\overline{\varrho}\,\overline{a}^2} \end{pmatrix} \tag{20}$$

and

$$q = R_{1d}\, c \tag{21}$$

with

$$R_{1d} = R(0) = \begin{pmatrix} -\overline{\varrho} & 0 & 0 & \frac{\overline{\varrho}}{2} & \frac{\overline{\varrho}}{2} \\ 0 & 0 & 0 & \frac{\overline{a}}{2} & -\frac{\overline{a}}{2} \\ 0 & \overline{a} & 0 & 0 & 0 \\ 0 & 0 & \overline{a} & 0 & 0 \\ 0 & 0 & 0 & \frac{\overline{\varrho}\,\overline{a}^2}{2} & \frac{\overline{\varrho}\,\overline{a}^2}{2} \end{pmatrix}. \tag{22}$$

## 2.2 Steady boundary conditions

The boundary condition generally consists of two steps, i.e. mean flow conditions and circumferential perturbations are treated separately. As changes in the mean flow represent plane waves at the boundary, we can directly express them by changes in the 1D characteristics. In the following, we write down inflow and outflow boundary conditions by means of characteristics in a compact form, so equations for in- and outflow boundaries are only given separately where necessary. For this purpose, we separate the left and right eigenvector matrices and their corresponding characteristics depending on the direction of propagation of their associated waves. In normally subsonic flows, the first four waves propagate downstream and the fifth one runs upstream. Accordingly, we write for an inflow

$$\left(\begin{array}{c} L^{in} \\ \hline L^{out} \end{array}\right) = \left(\begin{array}{c} l_1 \\ l_2 \\ l_3 \\ l_4 \\ \hline l_5 \end{array}\right) \tag{23}$$

$$\left(\begin{array}{c} c^{in} \\ \hline c^{out} \end{array}\right) = \left(\begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \\ \hline c_5 \end{array}\right) \tag{24}$$

$$\left(R^{in}\middle|R^{out}\right) = \left(r_1\ r_2\ r_3\ r_4\middle|\, r_5\right) \tag{25}$$

and for an outflow

$$\left(\begin{array}{c} L^{out} \\ \hline L^{in} \end{array}\right) = \left(\begin{array}{c} l_1 \\ l_2 \\ l_3 \\ l_4 \\ \hline l_5 \end{array}\right) \tag{26}$$

$$\left(\begin{array}{c} c^{out} \\ \hline c^{in} \end{array}\right) = \left(\begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \\ \hline c_5 \end{array}\right) \tag{27}$$

$$\left(R^{out}\middle|R^{in}\right) = \left(r_1\ r_2\ r_3\ r_4\middle|\, r_5\right). \tag{28}$$

To meet the boundary values specified by the user, we define a residual vector for both inflow and outflow boundaries

$$\mathfrak{R}_{bd} = \begin{cases} \begin{pmatrix} \overline{p}(\overline{s} - s_{bd}) \\ \overline{\varrho}\,\overline{a}\,(\overline{v} - \overline{u}\tan(\alpha_{circ,bd})) \\ \overline{\varrho}\,\overline{a}\,(\overline{w} - \overline{u}\tan(\alpha_{rad,bd})) \\ \overline{\varrho}(\overline{h_t} - h_{t,bd}) \end{pmatrix} & \text{for inflow boundaries} \\ \\ \left(\overline{p} - p_{bd}\right) & \text{for outflow boundaries} \end{cases} \tag{29}$$

Here, the subscript $bd$ denotes user specified boundary values. Since for turbomachinery flows, the user commonly specifies stagnation pressure and stagnation temperature at an inflow boundary, the associated specific entropy $s$ and specific stagnation enthalpy $h_t$ need to be calculated from the former. The angles $\alpha_{circ}$ and $\alpha_{rad}$ represent the angles between the velocity vector and the boundary normal in pitchwise and spanwise direction. Mixing planes can be covered by defining analogously

$$\mathfrak{R}_{bd} = L^{in}_{1d}\,\delta\overline{q}_{mp} \tag{30}$$

with $\delta\overline{q}_{mp}$ being the difference between the flow states at each side of the mixing plane in the same frame of reference.

To meet the boundary values specified by the user or mixing plane, the required change of characteristics is obtained via a Newton-Raphson step:

$$\mathfrak{R}_{bd} + \frac{\partial\mathfrak{R}}{\partial q}\frac{\partial q}{\partial c^{in}}\delta\overline{c^{in}} = 0 \tag{31}$$

Note that $\frac{\partial q}{\partial c^{in}}$ is the backward transformation of the incoming characteristics, $R^{in}_{1d}$. The derivation of the residual Jacobian $\frac{\partial\mathfrak{R}}{\partial q}$ can be found in [10]. For the adaption of Giles' original boundary condition to a cell-centred solver, where the boundary condition is applied between two pseudo-time updates, we also need to modify the update of the mean charactersistics at the faces due the pseudo-time change of the outgoing characteristics in the interior [16]. We define another residual

$$\mathfrak{R}_i = L^{out}_{1d}\,(\overline{q_f} - \overline{q_i}) \tag{32}$$

where subscripts $i$ and $f$ denote inner cell values and face values, respectively. Note that the inner values have already been updated by the pseudo-time solver whereas face values are still about to be updated by the boundary condition. Finally, the update of mean characteristics reads

$$\delta\overline{c} = -L_{1d}\left[R^{in}_{1d}\left(\frac{\partial\mathfrak{R}}{\partial q}R^{in}_{1d}\right)^{-1}\mathfrak{R}_{bd} + R^{out}_{1d}\mathfrak{R}_i\right] \tag{33}$$

To apply the actual non-reflecting boundary condition, we perform a Fourier decomposition of the disturbances about the mean flow along the boundary in the pitchwise direction. The spatial transformation for any quantity $\phi$ reads

$$\hat{\phi}_i = \frac{1}{P}\int_0^P \widetilde{\phi}\, e^{-\mathrm{i}m_i\, y}\, dy. \tag{34}$$

with wave numbers $m_i = \frac{2\pi i+\theta}{P} \neq 0$, pitch $P$, inter-blade phase angle $\theta$ and local deviation from the average state $\widetilde{\phi} = \phi - \overline{\phi}$. Note that $\theta$ vanishes in steady computations. Accordingly, the backward transformation is given by

$$\widetilde{\phi} = Re\left(\sum_{i=-\infty}^{\infty}\hat{\phi}_i\, e^{\mathrm{i}m_i\, y}\right). \tag{35}$$

In order to attain non-reflecting properties at the boundary, equation (10) must be satisfied for every mode with $m_i \neq 0$ and $\omega = 0$. We can scale the second to fifth left eigenvectors by $\omega$ and then set $\omega$ to zero. Consequently, the left eigenvector matrix for steady boundary conditions reduces to

$$L_s = \begin{pmatrix} -\frac{1}{\overline{\varrho}} & 0 & 0 & 0 & \frac{1}{\overline{\varrho}\,\overline{a}^2} \\ 0 & -\frac{\overline{u}}{\overline{a}^2} & -\frac{\overline{v}}{\overline{a}^2} & 0 & \frac{-1}{\overline{\varrho}\,\overline{a}^2} \\ 0 & 0 & 0 & \frac{1}{\overline{a}} & 0 \\ 0 & -\frac{\overline{v}}{\overline{a}^2} & \frac{\overline{u}}{\overline{a}^2} & 0 & \frac{\beta}{\overline{\varrho}\,\overline{a}^3} \\ 0 & \frac{\overline{v}}{\overline{a}^2} & -\frac{\overline{u}}{\overline{a}^2} & 0 & \frac{\beta}{\overline{\varrho}\,\overline{a}^3} \end{pmatrix} \tag{36}$$

with

$$\beta = \begin{cases} \mathrm{i}\,\mathrm{sign}(m)\sqrt{\overline{a}^2 - (\overline{u}^2+\overline{v}^2)} & \text{for } \overline{u}^2+\overline{v}^2 < \overline{a}, \\ -\mathrm{sign}(\overline{v})\sqrt{(\overline{u}^2+\overline{v}^2) - \overline{a}} & \text{for } \overline{u}^2+\overline{v}^2 > \overline{a}. \end{cases} \tag{37}$$

We can express the actual boundary condition, equation (10), in terms of characteristics which is helpful to write the overall boundary update as a sum of mean characteristic changes and local changes. Note that the outgoing characteristics have to be extrapolated from the interior.

$$L^{in}_s\hat{q} = L^{in}_s\left(R^{in}_{1d}\,L^{in}_{1d}\,\hat{q}_f + R^{out}_{1d}\,L^{out}_{1d}\,\hat{q}_i\right) = L^{in}_s\left(R^{in}_{1d}\,\hat{c}^{in}_f + R^{out}_{1d}\,\hat{c}^{out}_i\right) = 0 \tag{38}$$

The above equation yields ideal values for the incoming characteristics as a function of the outgoing ones. To meet the ideal values, the required changes in the characteristic at an inflow boundary is given by

$$\begin{pmatrix}\delta\hat{c}_1\\ \delta\hat{c}_2\\ \delta\hat{c}_3\\ \delta\hat{c}_4\\ \delta\hat{c}_5\end{pmatrix} = \begin{pmatrix} -\hat{c}_{f,1} \\ -\frac{\beta+\overline{v}}{\overline{a}+\overline{u}}\hat{c}_{i,5} - \hat{c}_{i,2} \\ -\hat{c}_{f,3} \\ \left(\frac{\beta+\overline{v}}{\overline{a}+\overline{u}}\right)^2\hat{c}_{i,5} - \hat{c}_{f,4} \\ \hat{c}_{i,5} - \hat{c}_{f,5} \end{pmatrix}. \tag{39}$$

Accordingly, the outflow boundary condition reads

$$\begin{pmatrix} \delta\hat{c}_1 \\ \delta\hat{c}_2 \\ \delta\hat{c}_3 \\ \delta\hat{c}_4 \\ \delta\hat{c}_5 \end{pmatrix} = \begin{pmatrix} \hat{c}_{i,1} - \hat{c}_{f,1} \\ \hat{c}_{i,2} - \hat{c}_{f,2} \\ \hat{c}_{i,3} - \hat{c}_{f,3} \\ \hat{c}_{i,4} - \hat{c}_{f,4} \\ \frac{2\overline{u}}{\beta-\overline{v}}\hat{c}_{i,2} - \frac{\beta+\overline{v}}{\beta-\overline{v}}\hat{c}_{i,4} - \hat{c}_{f,5} \end{pmatrix}. \tag{40}$$

Subsequently, these characteristic changes can be transformed from the wavenumber domain back into the physical domain according to equation (35) yielding a local change of the characteristics $\delta\tilde{c}$.

For supersonic ($\overline{u}^2+\overline{v}^2 > \overline{a}^2$), but normally subsonic ($\overline{u} < \overline{a}$) flows, $\beta$ and $L_s$ are independent of $m$ according to equations (36) and (37). Hence, no Fourier transformation is required and equations (39) and (40) can be applied directly at each face substituting any $\hat{c}$ by $\widetilde{c}$. Note that the steady, supersonic boundary conditions become local boundary conditions in this case apart from their dependence on averaged quantities.

Prescribing no incoming perturbations means that entropy and enthalpy are uniform along the inflow boundary only within the linearised theory, but second-order perturbations may still occur. Thus, we replace the condition for $\delta\tilde{c}_1$ and $\delta\tilde{c}_4$ in the physical domain and instead enforce uniform entropy and enthalpy by locally driving the following residual

$$\begin{pmatrix} \widetilde{\mathfrak{R}_1} \\ \widetilde{\mathfrak{R}_2} \end{pmatrix} = \begin{pmatrix} \overline{p}\,\widetilde{s} \\ \overline{\varrho}\widetilde{h_t} \end{pmatrix} \tag{41}$$

to zero. This can be done by means of a Newton-Raphson step

$$\begin{pmatrix} \widetilde{\mathfrak{R}_1} \\ \widetilde{\mathfrak{R}_2} \end{pmatrix} + \begin{pmatrix} \frac{\gamma R}{\gamma-1}\overline{p} & 0 & 0 \\ \frac{\gamma}{\gamma-1}\overline{p} & \overline{\varrho}\,\overline{a}\,\overline{v} & \frac{\overline{\varrho}}{2}\left(\overline{a}^2 + \overline{a}\,\overline{u}\right) \end{pmatrix} \begin{pmatrix} \delta\tilde{c}_1 \\ \delta\tilde{c}_2 \\ \delta\tilde{c}_4 \end{pmatrix} = 0 \tag{42}$$

where $R$ denotes the specific gas constant. The resulting condition for the update of $\widetilde{c_1}$ and $\widetilde{c_4}$ reads

$$\begin{pmatrix} \delta\widetilde{c_1} \\ \delta\widetilde{c_4} \end{pmatrix} = \begin{pmatrix} -\frac{\gamma-1}{\gamma R}\widetilde{s} \\ \frac{-2}{\overline{a}(\overline{a}+\overline{u})}\left(\widetilde{h_t} + \frac{\widetilde{s}}{\gamma R}\overline{a}^2 + \overline{a}\,\overline{v}\,\delta\widetilde{c_2}\right) \end{pmatrix}. \tag{43}$$

We further have to determine the integral change of characteristics due to circumferential perturbations because the mean flow boundary conditions already ensure that averaged flow quantities match the boundary values specified by the user or the mixing plane. Accordingly, we have to subtract the integral change afterwards in order to guarantee the mean change of characteristics is not affected by the treatment of perturbations. For this purpose we introduce

$$\overline{\delta\widetilde{c}} = \frac{1}{P}\int_0^P \delta\widetilde{c}\,dy. \tag{44}$$

Finally, the overall update of a boundary face can be written as

$$\delta q = \left[\sigma R_{1d}^{in}\left(\delta\overline{c}^{in} + \delta\widetilde{c}^{in} - \overline{\delta\widetilde{c}^{in}}\right) + R_{1d}^{out}\left(\delta\overline{c}^{out} + \delta\widetilde{c}^{out} - \overline{\delta\widetilde{c}^{out}}\right)\right]. \tag{45}$$

The relaxation factor $\sigma$, which needs to be chosen suitably in the range of $0$ to $1$, is necessary to retain the well-posedness of the mathematical problem [7]. For the application in a cell-centred solver, the updated face values have to be extrapolated appropriately to the ghost cells.

### 2.3 Approximate unsteady boundary conditions

To overcome the disadvantage of the straightforward, exact unsteady NRBC being spatially and temporally non-local, Giles proposes an approximate, local unsteady boundary condition [7]. This approach has been introduced by Engquist and Majda for general wave equations [9], but Giles applied it to two-dimensional linearised Euler equations in general and turbomachinery flows in particular.

As in steady boundary conditions, the averaged flow field and perturbations are handled separately. The mean flow boundary condition is identical to the steady mean flow boundary condition except that averaged quantities in the residual (29) are additionally averaged in time rather than only circumferentially. Unsteady turbomachinery flows are (within the framework of RANS) considered to be periodic and, therefore, temporal averaging means averaging a quantity over one period in this context.

The actual approximate boundary condition, however, handles perturbations from the mean state. The condition for a (hypothetically) perfectly non-reflecting boundary treatment has been discussed in section 2.1. But the universal application of equations (10) and (18) requires the decomposition of the boundary flow field into the Fourier domain. To avoid this costly transformation, the central concept is to express the exact unsteady boundary condition by use of a Taylor series expansion about the one-dimensional characteristic boundary condition. Recall from equation (16) that $\lambda \rightarrow 0$ and from equations (15) and (14) that $\Psi \rightarrow 1$ in the one-dimensional case due to vanishing circumferential wave numbers $m$. Then, the second order approximation of the left eigenvector matrix $L$ reads

$$L_a = L|_{\lambda=0} + \lambda \left.\frac{\partial L}{\partial \lambda}\right|_{\lambda=0} = L_{1d} + \frac{m}{\omega} \left.\frac{\partial L}{\partial \lambda}\right|_{\lambda=0}. \tag{46}$$

Now the boundary condition $L_a^{in}\,\widetilde{q} = 0$ can be rearranged starting with multiplying by $\omega$. Comparing equations (1) and (4), we can infer that we can replace $\omega$ by $\mathrm{i}\frac{\partial}{\partial t}$ and $m$ by $-\mathrm{i}\frac{\partial}{\partial y}$ yielding

$$L_{1d}^{in}\frac{\partial \widetilde{q}}{\partial t} - \left.\frac{\partial L^{in}}{\partial \lambda}\right|_{\lambda=0} \frac{\partial \widetilde{q}}{\partial y} = 0. \tag{47}$$

Giles' shows his original formulation may become ill-posed at inflows [7]. To face this ill-posedness, he proposes a modification by giving up the perfect orthogonality of $l_4$ and $r_5$ and thereby the condition for perfect non-reflecting behaviour. A multiple of $(\lambda\, l_{2,1d})$ is added to $l_{a,4}$ such that $l_{a,4}\, r_5$ is minimized under the constraint of the inflow boundary condition being well-posed. Expressing $\widetilde{q}$ in terms of characteristics, we can again extrapolate the outgoing characteristic $\widetilde{c}_5$ from the interior and obtain the following inflow boundary condition

$$\frac{\partial}{\partial t}\begin{pmatrix}\widetilde{c}_1\\ \widetilde{c}_2\\ \widetilde{c}_4\end{pmatrix} + \begin{pmatrix}\overline{v} & 0 & 0\\ 0 & \overline{v} & \frac{1}{2}(\overline{a}+u)\\ 0 & \frac{1}{2}(\overline{a}-\overline{u}) & \overline{v}\end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix}\widetilde{c}_1\\ \widetilde{c}_2\\ \widetilde{c}_4\end{pmatrix} + \begin{pmatrix}0\\ \frac{1}{2}(\overline{a}-\overline{u})\\ 0\end{pmatrix}\frac{\partial \widetilde{c}_5}{\partial y} = 0. \tag{48}$$

The outflow boundary condition is already well-posed in its original formulation and, thus, remains unchanged. It reads

$$\frac{\partial \widetilde{c}_5}{\partial t} + \overline{v}\frac{\partial \widetilde{c}_5}{\partial t} + \begin{pmatrix}0 & \frac{1}{2}(\overline{a}+\overline{u}) & 0\end{pmatrix}\frac{\partial}{\partial y}\begin{pmatrix}\widetilde{c}_1\\ \widetilde{c}_2\\ \widetilde{c}_4\end{pmatrix} = 0. \tag{49}$$

As the characteristic $c_3$, representing vorticity perturbations in spanwise direction, is decoupled in this quasi-three-dimensional approach, this characteristic is treated as a convective perturbation and, hence, extrapolated appropriately.

The appropriate numerical method to solve these differential equations at the boundaries depends on the solution methods for the flow equations in the interior. The solution algorithm applied in TRACE is presented in [5]. The primitive variables at the boundary and in the ghost cells are reconstructed from the change of characteristics in the same manner as in the steady boundary conditions.

Due to the chosen approach, to express the left eigenvector matrix by a Taylor series expansion of order two in $\lambda$, this boundary is only perfectly non-reflecting for planar waves impinging normally upon the boundary. In general, the boundary conditions become more reflecting with increasing $\lambda$. Accordingly, flows with large circumferential wave numbers and flows that are dominated by low frequency perturbations can induce spurious reflections at the boundary. The non-reflecting properties of the inflow boundary condition are further diminished by the modification to gain well-posedness. Details on reflections coefficients can be found in [7] and applications to two-dimensional pressure-waves in uniform mean flow assessing the reflection properties are presented in [17].

A first order approximation of these boundary conditions is also implemented in TRACE. In the application section of this paper, both boundary conditions are applied and compared to the exact approach. Note that the first order approximation can also be regarded as a characteristic, one-dimensional boundary condition.

### 2.4 Exact NRBC

It can be difficult to distinguish between the impact of approximate, unsteady boundary conditions and actual unsteady effects when comparing unsteady and steady flow simulations. The authors rank direct comparability of steady and unsteady computations as highly desirable in order to asses unsteady flow phenomena in turbomachinery design and research. Therefore, we do not consider Fourier transforms in time and space to be necessarily circumvented due to their computational effort. The fundamental concept of the exact boundary conditions is applying condition (10) to any incoming mode of the spatially and temporally decomposed boundary flow field. For this purpose, we first determine the temporal Fourier decomposition at the boundary according to the approach He proposed in the context of his phase lag method [18]. For the implementation in TRACE, the reader is referred to [19]. The choice of considered harmonics is analogous to the phase lag set of harmonics. From here on, we distinguish temporally Fourier transformed quantities, denoted by subscript $\omega$, and circumferentially Fourier transformed quantities, denoted by subscript $m$. For simplicity, this has been omitted up to this point. The temporal Fourier coefficients $\hat{q}_\omega$ are again expressed in terms of characteristic variables $\hat{c}_\omega = L_{1d}\hat{q}_\omega$. For each frequency in the Fourier domain, the flow can also be Fourier transformed along the boundary according to equation (34) and the respective outgoing characteristics are again extrapolated from the interior characteristics $\hat{c}^{out}_{(\omega,m),i}$ . The exact non-reflecting boundary condition for every mode, i.e. for every combination of $\omega$ and $m$, reads

$$L^{in}\hat{q}_{(\omega,m),f} = L^{in}\left(R^{in}_{1d}\,\hat{c}^{in}_{(\omega,m),target} + R^{out}_{1d}\,\hat{c}^{out}_{(\omega,m),i}\right) = 0. \tag{50}$$

$L$ no longer needs to be approximated because $\omega$ and $m$ are known in this context. The condition for updating the incoming modal characteristics is

$$\delta\hat{c}^{in}_{(\omega,m)} = \left[-\left(L^{in}R^{in}_{1d}\right)^{-1} L^{in}R^{out}_{1d}\,\hat{c}_{(\omega,m),i}\right] - \hat{c}^{in}_{(\omega,m),f}. \tag{51}$$

The outgoing characteristic is modified according to

$$\delta\hat{c}^{out}_{(\omega,m)} = \hat{c}^{out}_{(\omega,m),i} - \hat{c}^{out}_{(\omega,m),f} \tag{52}$$

The mean flow $\hat{q}_{0,0}$ is treated in the same manner as in the steady or approximate unsteady boundary condition according to equation (33).

Transforming $\delta\hat{c}_{(\omega,m)}$ back into the physical domain, we obtain face-wise, harmonic characteristics $\delta\hat{c}_{\omega}(y)$. The change of primitive variables at each face is relaxed again:

$$\delta\hat{q}_{\omega} = \sigma R_{1d}\delta\hat{c}_{\omega} \tag{53}$$

Subsequently, the boundary states can be extrapolated to the ghost cells in the frequency domain. Finally, the current ghost cell state is reconstructed by means of an approximate, inverse Fourier transform, again exploiting the phase lag functionality in TRACE [19].

Due to its universal approach, this boundary condition is suitable and consistent for both steady and unsteady simulations. Beyond, this approach is perfectly consistent with the non-reflecting boundary condition for frequency domain methods, presented by Frey [20, 21], which is rather favourable to compare unsteady time-domain computations and frequency domain computations. Chassaing and Gerolymos [17] demonstrate the advantageous reflection properties towards Giles' approximate boundary conditions for certain acoustic waves in uniform mean flow. Yet, they observe considerably slower convergence. This is considered by the authors of this paper to be due to the issue of temporal Fourier coefficients evolving rather slowly over possibly many periods, which is a known handicap of the phase lag approach.

## 3 APPLICATION

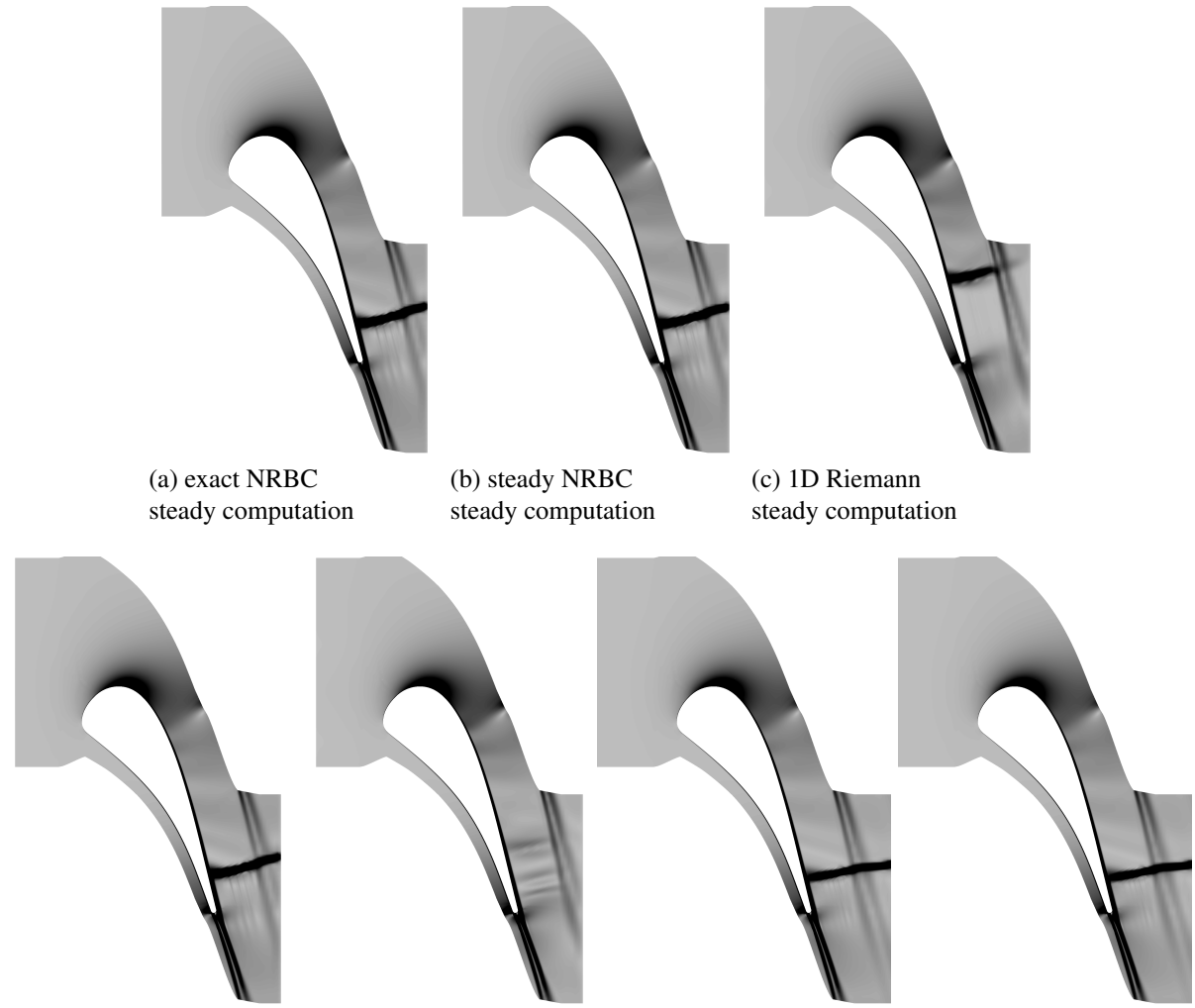
To validate our implementation of the exact boundary conditions and asses their non-reflecting properties, we show their application to two turbomachinery test cases. The behaviour of the exact NRBC is compared to steady NRBC, first and second order approximate NRBC and simple, one-dimensional, steady Riemann boundary conditions. For the description of Riemann boundary conditions, the reader is referred to [15].

When discussing boundary conditions for turbomachinery flows, the appropriate definition of a mean state is not trivial. This issue has been excluded from the theory section to stick to the actual subject of non-reflecting boundary conditions. In the following applications, however, all mean flow states are obtained by flux averaging [10] along the pitch and additionally in time for unsteady flows.

All computations are performed on structured grids using DLR's finite-volume 3D RANS solver TRACE [22, 23] applying Roe's upwind scheme [24] extended to second order accuracy through van Leer's MUSCL extrapolation [25] and an appropriate limiter function for convective fluxes. Viscous fluxes are computed based on gradients obtained by a central finite difference scheme in combination with Wilcox's linear eddy viscosity turbulence model [26]. For steady simulations a first order implicit pseudo-time marching technique is employed. The same pseudo-time method is used for subiterations of single time steps in unsteady computations.

### 3.1 VKI LS89

The first test case is a two-dimensional, linear turbine cascade designed and investigated at the von Karman Institute for Fluid Dynamics. The airfoil is typical of high pressure turbine nozzle guide vanes. A detailed description can be found in [27].



(a) exact NRBC steady computation

(b) steady NRBC steady computation

(c) 1D Riemann steady computation

(d) exact NRBC unsteady computation

(e) 2nd order approx. BC unsteady computation

(f) exact NRBC steady computation long domain

(g) steady NRBC steady computation long domain

Figure 1: Pseudo-Schlieren images obtained by plotting density gradient magnitude (black corresponds to large gradients)

All calculations are carried out for a supercritical operating point with 415 K stagnation temperature and 147500 Pa stagnation pressure at the inlet and 78000 Pa static pressure at the outlet. The inflow is purely axial. Steady computations are conducted applying one-dimensional Riemann boundary conditions, steady NRBC and exact NRBC.

Unsteady computations apply the second order approximate boundary conditions and the exact NRBC. The base frequency is set to 6100 Hz. The underlying assumption is that a downstream rotor blade with the same pitch faces purely axial flow in its relative frame of reference at this blade passing frequency. Therefore, this frequency is estimated to have a realistic order of magnitude for turbomachinery flows. One period is resolved by 64 time steps using a second order Euler backward scheme. For every physical time step, 25 pseudo-time iterations are performed. Though we perform unsteady computations, both solutions converge to a steady state.

Inlet and outlet planes are located at 50 % axial chord length away from the leading and

training edge. The computational grid comprises 21016 cells. Furthermore, we conduct steady computations with axial spacing between boundaries and the blade of 300 % axial chord length applying steady and exact NRBC to produce reference results. The extended grid comprises 36668 cells.

In figure 1, the magnitude of the density gradients is plotted for all computations in order to mimic the Schlieren flow visualization technique. The flow around the airfoil is supercritical, characterised by a shock close to suction side trailing edge impinging upon the exit boundary of the short computational domain. The flow field predicted employing the steady (Fig. 1b) and exact NRBC (Fig. 1a) in steady computations as well as the exact NRBC in an unsteady computation (Fig. 1d) agree qualitatively with the reference results obtained from steady computations in the long domain applying both steady (Fig. 1g) and exact NRBC (Fig. 1f). The shock position, in particular, is well captured.

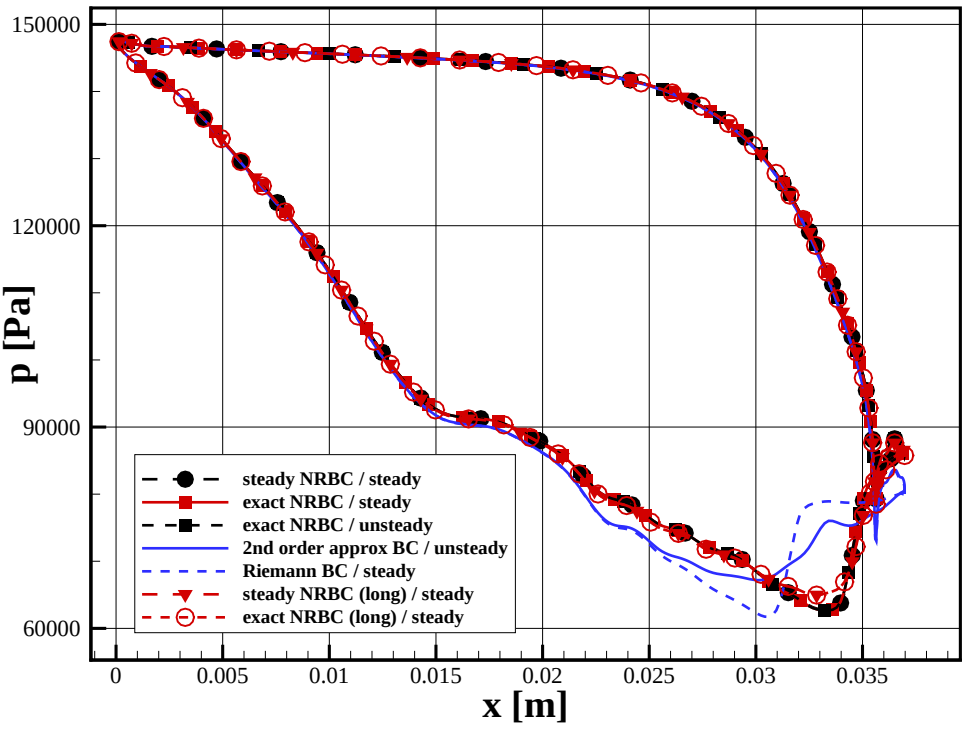



Figure 2: Blade pressure distribution

However, the solution obtained by Riemann boundary conditions (Fig. 1c) shifts the shock slightly upstream and diminishes its intensity. In the flow field resulting from the unsteady computation employing the second order approximate boundary condition (Fig. 1e), a distinct shock is not visible. This behaviour becomes more apparent in figure 2. Blade pressure distributions of all simulations are compared. The figure shows that the approximate boundary condition "smears out" the shock leading to a different suction side pressure distribution and, thus, aerodynamical load. Similarly, the Riemann boundary condition leads to a comparable error in the pressure distribution. The shift and weakening of the shock is evident here as well. Accordingly, an accurate prediction of integral forces, flow turning, performance and losses cannot be expected for this type of flow from the approximate boundary condition and the Riemann boundary condition.

The pressure distributions of all other computations coincide almost perfectly. Yet, in the very vicinity of the shock, all short domain computations deviate slightly from the reference results of the long domain. But among themselves, the exact and steady NRBC short domain results agree very well.

### 3.2 Transonic compressor rig

The second test case is a single-stage, transonic compressor rig of the Darmstadt University of Technology [28]. Steady and unsteady computations are performed for the aerodynamic design point. At a rotational speed of 20000 revolutions per minute the compressor achieves a stagnation pressure ratio of about 1.5 for a massflow of 16.3 kg/s. In order to conduct unsteady computations efficiently, the stator is scaled to 32 blades per row, which enables simulations having two stator blades and one rotor blade (16 blades per row) with equal pitch. The computational grid contains 254728 cells.

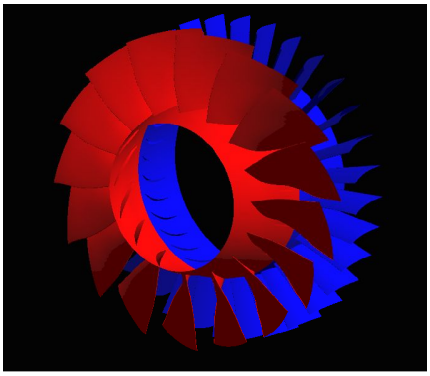


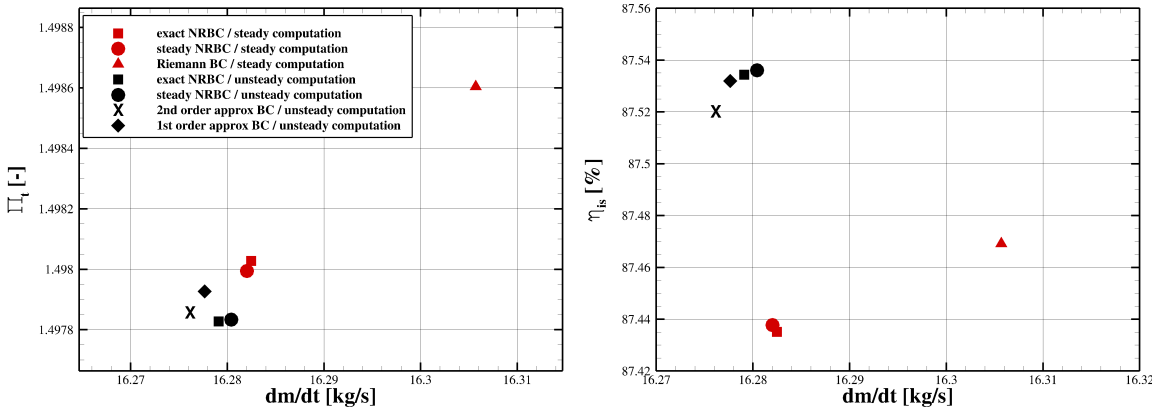
Figure 3: Computational model of the Darmstadt transonic compressor rig

Steady computations employ the mixing plane approach for blade row coupling [3], which is conservative due to its formulation based upon flux averaged quantities. Several computations are conducted applying one-dimensional Riemann boundary conditions, steady and exact NRBC. Unsteady computations utilize a conservative, non-matching interface [29] and a third order implicit Runge-Kutta time discretization scheme with 64 time steps per segment passing period. At each time step, 25 pseudo-time iterations are performed. First and second order approximate boundary conditions are applied as well as steady and exact NRBC.

Figure 4 shows the impact of boundary conditions on integral flow quantities in the compressor. The one-dimensional Riemann boundary conditions lead to an increase in the massflow of approximately 0.16 % compared to all other computations and 0.03 percentage points increase in isentropic efficiency towards steady computations employing steady and exact NRBC. This deviation is not huge, but significant in the context of turbomachinery performance analysis. All other solutions agree well regarding massflow and stagnation pressure ratio.

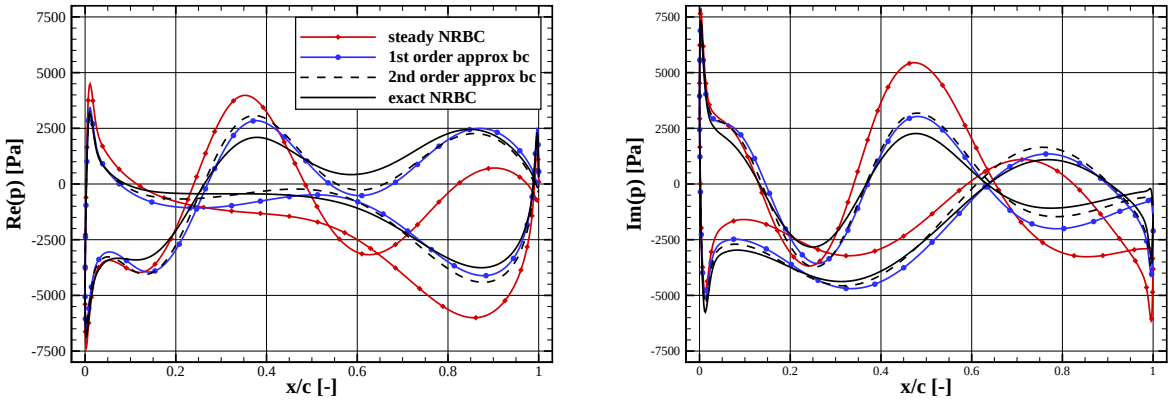
In particular, the steady and the exact NRBC predict abutting operation points and efficiencies. The shift in efficiency from steady to unsteady calculations can be explained from the inherently different mechanisms of loss production in steady and unsteady computations. Since any non-uniformity is mixed out in the mixing plane, this approach is known to overestimate losses in turbomachinery flows [30].

Considering the unsteady calculations, the influence of the chosen boundary condition on the operating point and performance is rather small. This might lead to the conclusion that steady boundary conditions can properly be applied in unsteady computations.




(a) Stagnation pressure ratio over massflow (b) Isentropic efficiency over massflow

Figure 4: Operating points for steady and unsteady computations employing different boundary conditions




(a) Real part (b) Imaginery part

Figure 5: Unsteady blade pressure distribution along the 80 % relative mass flow streamsurface (complex Fourier coefficient of first harmonic)

However, even in the case of similar integral values, like massflow or efficiency, the flow field may still differ noticeably. The unsteady blade pressure distribution of the stator is depicted in figure 5. The plot shows the complex amplitude of pressure associated with the first harmonic (i.e. the segment passing frequency) along a stream surface that is defined such that 80 % of the overall massflow run through the passage below this stream surface. Due to artificial reflections arising when steady boundary conditions are applied in unsteady flows, the steady NRBC lead to a significantly different unsteady pressure distribution. The Exact NRBC and approximate NRBC yield comparable, but not perfectly matching unsteady pressure distributions.

Chassaing and Gerolymos demonstrate the good non-reflecting properties of the exact NRBC applied to unsteady flows with uniform underlying mean flow [17] whereas the derivation of the steady NRBC depends on the assumption of a steady flow field. Thus, it can be assumed that the exact NRBC provide a better prediction of the unsteady pressure field than the steady boundary conditions do. Since unsteady pressure fluctuations play a key role for aeroelastic and aeroacoustic phenomena, the inaccurate prediction of the unsteady pressure field can be

detrimental especially for these disciplines. Thus, the steady NRBC is not suitable for unsteady flows.

## 4 CONCLUSIONS

We have implemented an exact formulation of non-reflecting boundary conditions in TRACE, which is neither limited to steady flows nor suffers from the loss in accuracy raising from the approximation of the left eigenvectors. Thus, the presented exact NRBC represents a single method for both steady and unsteady turbomachinery flow simulations. Moreover, this method is consistent with the natural formulation of NRBC in frequency domain methods. However, the limitations arising from the linearisation of the quasi two-dimensional Euler equations still hold.

We have demonstrated the strong non-reflecting properties of the exact NRBC in two test cases. The steady flow in a two-dimensional supercritical, highly loaded turbine cascade is captured well. We observe good agreement with the non-reflecting steady NRBC as well as with reference solutions obtained on a larger domain. Moreover, the solution coincide when switching to unsteady simulations. This behaviour is advantageous compared to the combination of steady and approximate, unsteady boundary conditions since the flow field predicted employing the approximate unsteady NRBC varies significantly from the steady solution obtained employing steady NRBC. Note that no unsteadiness is observed in the unsteady solutions. So the change cannot be explained by unsteady effects.

The application to a transonic compressor stage showed that the exact method is capable to predict the operating point and performance in good agreement with the steady boundary condition. In contrast, less advanced one-dimensional Riemann boundary conditions diminish the solution quality in both test cases and are, therefore, considered inferior in turbomachinery flows.

We further demonstrated, that steady NRBC applied in unsteady computations do not necessarily predict significantly different integral quantities such as performance data or the operating point. However, in time resolving simulations, the steady NRBC lead to reflected waves that propagate through the domain and, possibly, pollute the simulation of unsteady flow phenomena.

In summary, the exact NRBC perform very satisfactory in both test cases. However, the lagged convergence properties due to the challenging determination of temporal Fourier coefficients may limit the range of practical applicability. Thus, investigations on this issue are needed to further improve convergence speed and robustness.

## ACKNOWLEGEMENT

Financial support by MTU Aero Engines (co-sponsorship of the first author) is gratefully acknowledged.

## REFERENCES

[1] J. D. Denton and W. N. Dawes, “Computational fluid dynamics for turbomachinery design,” *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 213, no. 2, pp. 107–124, 1998.

[2] W. Dawes, "Turbomachinery computational fluid dynamics: asymptotes and paradigm shifts," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1859, pp. 2553–2585, 2007.

[3] J. Denton and U. Singh, "Time marching methods for turbomachinery flow calculations," VKI Lecture Series 1979-7, von Karman Institute., 1979.

[4] H.-P. Kersken, G. Ashcroft, C. Frey, N. Wolfrum, and D. Korte, "Nonreflecting boundary conditions for aeroelastic analysis in time and frequency domain 3D RANS solvers," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[5] G. Ashcroft and J. Schulz, "Numerical modelling of wake-jet interaction with application to active noise control in turbomachinery," in *Proceedings of Aeroacoustics Conferences*, American Institute of Aeronautics and Astronautics, 2004.

[6] R. L. Higdon, "Initial-boundary value problems for linear hyperbolic systems," *SIAM Rev.*, vol. 28, pp. 177–217, June 1986.

[7] M. B. Giles, "Non-reflecting boundary conditions for the Euler equations," tech. rep., MIT Dept. of Aero. and Astr., 1988. CFDL Report 88-1.

[8] M. B. Giles, "Nonreflecting boundary conditions for Euler calculations," *AIAA Journal*, vol. 28, no. 12, pp. 2050–2058, 1990.

[9] B. Engquist and A. Majda, "Absorbing boundary conditions for the numerical simulation of waves," *Math. Comp.*, vol. 31, pp. 629–651, May 1977.

[10] M. Giles, "UNSFLO: A numerical method for the calculation of unsteady flow in turbomachinery," tech. rep., Gas Turbine Laboratory Report GTL 205, MIT Dept. of Aero. and Astro., 1991.

[11] A. P. Saxer and M. B. Giles, "Quasi-three-dimensional nonreflecting boundary conditions for Euler equations calculations.," *Journal of Propulsion and Power*, vol. 9, no. 2, pp. 263–271, 1993.

[12] T. Hagstrom, *Computational Wave Propagation*, ch. On High-Order Radiation Boundary Conditions, pp. 1–21. New York, NY: Springer New York, 1997.

[13] S. Henninger, P. Jeschke, G. Ashcroft, and E. Kügeler, "Time-domain implementation of higher-order non-reflecting boundary conditions for turbomachinery applications," in *Proceedings of ASME Turbo Expo 2013*, 2015.

[14] C. Frey and H.-P. Kersken, "On the regularisation of non-reflecting boundary conditions near acoustic resonance," in *Accepted for: Proceedings of ECCOMAS Congress 2016*, 2016.

[15] C. B. Laney, *Computational gasdynamics*. Cambridge University Press, 1998.

[16] S. Robens, C. Frey, P. Jeschke, E. Kügeler, A. Bosco, and T. Breuer, "Adaption of Giles non-local non-reflecting boundary conditions for a cell-centered solver for turbomachinery applications," in *Proceedings of ASME Turbo Expo 2013*, 2013.

[17] J. Chassaing and G. Gerolymos, "Time-domain implementation of nonreflecting boundary-conditions for the nonlinear euler equations," *Applied Mathematical Modelling*, vol. 31, no. 10, pp. 2172 – 2188, 2007.

[18] L. He, "Method of simulating unsteady turbomachinery flows with multiple perturbations," *AIAA Journal*, vol. 30, pp. 2730–2735, Nov. 1992.

[19] R. Schnell and D. Nürnberger, "Investigation of the tonal acoustic field of a transonic fanstage by time-domain CFD-calculation with arbitrary blade counts," in *Proceedings of ASME Turbo Expo 2004*, 2004.

[20] C. Frey, G. Ashcroft, H.-P. Kersken, and C. Voigt, "A harmonic balance technique for multistage turbomachinery applications," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[21] C. Frey, G. Ashcroft, and H.-P. Kersken, "Simulations of unsteady blade row interactions using linear and non-linear frequency domain methods," in *Proceedings of ASME Turbo Expo 2015*, 2015.

[22] K. Becker, K. Heitkamp, and E. Kügeler, "Recent progress in a hybrid-grid CFD solver for turbomachinery flows," in *Proceedings Fifth European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, 2010.

[23] K. Becker and G. Ashcroft, "A comparative study of gradient reconstruction methods for unstructured meshes with application to turbomachinery flows," in *52nd AIAA Aerospace Sciences Meeting*, (National Harbor, MD, USA), Jan. 2014.

[24] P. L. Roe, "Approximate Riemann solvers, parameter vectors, and difference schemes," *Journal of Computational Physics*, vol. 43, no. 2, pp. 357–372, 1981.

[25] B. van Leer, "Towards the ultimate conservative difference scheme. V. a second-order sequel to Godunov's method," *Journal of Computational Physics*, vol. 32, no. 1, pp. 101–136, 1979.

[26] D. C. Wilcox, "Reassessment of the scale-determining equation for advanced turbulence models," *AIAA Journal*, vol. 26, pp. 1299–1310, Nov. 1988.

[27] T. Arts and M. L. De Rouvroit, "Aero-thermal performance of a two dimensional highly loaded transonic turbine nozzle guide vane: A test case for inviscid and viscous flow computations," in *Proceedings of ASME Turbo Expo 1990*, 1990.

[28] G. Schulze, C. Blaha, D. Hennecke, and J. Henne, "The performance of a new axial single stage transonic compressor," in *Proceedings of the 12th International Symposium on Air Breathing Engines (ISABE)*, 1995.

[29] H. Yang, D. Nürnberger, and H.-P. Kersken, "Towards excellence in turbomachinery computational fluid dynamics," *Journal of Turbomachinery*, vol. 2006, no. 128, pp. 390–402, 2006.

[30] G. Fritsch and M. B. Giles, "An asymptotic analysis of mixing loss," *Journal of turbomachinery*, vol. 117, p. 367, 1995.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# REDUCED ORDER MODELLING OF GUST ANALYSIS USING COMPUTATIONAL FLUID DYNAMICS

**R. Thormann[1], P. Bekemeyer[2] and S. Timme[3]**

[1]Research Associate
University of Liverpool
e-mail: reik.thormann@liverpool.ac.uk

[2]Ph.D. Student and [3]Lecturer
University of Liverpool
e-mail: {philipp.bekemeyer, sebastian.timme}@liverpool.ac.uk

**Keywords:** Reduced Order Modelling, Gust Response, Computational Fluid Dynamics, Reynolds-averaged Navier-Stokes Equations

**Abstract.** *The simulation of gust responses is a crucial task in the design and certification process of a new aircraft. Moreover, if high accuracy is desired, the computational cost can be overwhelming. Linear frequency-domain methods have previously shown significant reduction in computational cost for motion-induced aerodynamics as well as for gust excitations. Time-domain signals are reconstructed by a superposition of responses at several discrete frequencies. Rather than using the frequency-domain method directly, a reduced order model is constructed projecting the time-depending linearised Reynolds-averaged Navier-Stokes equations on a basis obtained by proper orthogonal decomposition. The resulting small-sized ordinary differential equations for the modal coefficients are integrated in time. Results are presented for a two-dimensional NACA 0012 aerofoil covering sub- and transonic conditions including a case with shock-induced separation. Responses due to 1-cos as well as to sharp-edged gusts are compared between the reduced order model and its non-linear full order counterpart showing time histories of the lift coefficient and worst case surface pressure coefficients.*

## 1 Introduction

Gust load analysis is a key task during design and certification of new aircraft. Simulations have to be performed for a huge number of parameter combinations varying e.g. Mach number, altitude, load factor and gust length. The industrial process currently relies on linear potential methods, such as doublet lattice, which are commonly corrected by quasi-steady data either obtained from wind-tunnel experiments or computational fluid dynamics (CFD) simulations. Since corrections are commonly introduced only at zero frequency, deviations occur at higher frequencies important for shorter gust lengths and unsteady transonic effects – such as resonance behaviour and inverse shock motion – and cannot be captured accurately [1]. Additionally, this approach requires a mapping between the surface representations of the different methods which can become complicated for industrial cases.

In the past few years CFD aerodynamics alone have been used to investigate gust encounter. Examples from simple aerofoils to civil aircraft are available [2, 3, 4]. However, solving the non-linear Reynolds-averaged Navier-Stokes (RANS) equations in the time domain is too time-consuming to cover the complete flight envelope. Reduced order modelling is one alternative to overcome high computational cost. Several approaches are possible to achieve a reduced order model (ROM) for gust interactions including autoregressive methods [3] and eigenvalue realisation [5]. Another model reduction technique is based on proper orthogonal decomposition (POD) [6], which was, with respect to fluid dynamics, first introduced to model coherent structures in turbulent flow fields [7]. Snapshots are computed with the full order method covering the parameter space of interest and a small eigenvalue problem is solved subsequently to obtain a reduced linear basis. The idea of POD using frequency domain sample data was first introduced for the investigation of a simple twelve-degrees-of-freedom mass-spring-damper system combined with an incompressible three-dimensional vortex lattice method [8].

Linearised frequency domain (LFD) methods have also proven to retain the RANS accuracy at significantly reduced computational cost for forced motion simulations [9]. The RANS equations are linearised around a steady state and then solved in the frequency-domain. Moreover, this method was combined with the POD technique for a pitch-plunge aerofoil [10]. An extension to three-dimensional cases, generating snapshots for all structural modes of interest, has also been presented [11]. Recently, this approach was extended towards gust responses and a ROM is presented [12]. The LFD method is first used to generate snapshots of sinusoidal gusts while varying the reduced frequency. Afterwards, a reduced basis is computed using the POD approach and the linearised RANS equations are projected onto the subspace resulting in a small linear system per gust frequency.

In this paper, the approach is modified to solve the projected equations in the time domain to simulate non-periodic excitation, such as 1-cos and sharp-edged gusts more efficiently. Moreover, a time-domain ROM offers the additional advantage to couple the system with a controller for gust load allevation. Viscous results are shown for the NACA0012 aerofoil including non-linearities such as transonic shock waves and boundary layer separation. Responses are compared between the reduced order model and its non-linear full order counterpart showing time histories of the lift coefficient and worst case surface pressure coefficients.

## 2 Methods

The RANS equations in semi-discrete form read:

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t}(t) = \mathbf{R}(\mathbf{W}(t), \mathbf{v}_g(t)) \tag{1}$$

with $\mathbf{W}$ denoting the vector of fluid unknowns, $\mathbf{v}_g$ the external excitation due to a gust encounter and $\mathbf{R}$ the non-linear residual function. Separating the variables into a steady and a time-dependent part:

$$\mathbf{W}(t) = \overline{\mathbf{W}} + \widetilde{\mathbf{W}}(t)$$

and expanding eq. (1) around this steady state in a Taylor series yields:

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}}{\mathrm{d}t}(t) = \mathbf{R}(\overline{\mathbf{W}}, 0) + \frac{\partial \mathbf{R}}{\partial \mathbf{W}}\,\widetilde{\mathbf{W}}(t) + \frac{\partial \mathbf{R}}{\partial \mathbf{v}_g}\,\widetilde{\mathbf{v}}_g(t) + \text{H.O.T.}$$

Since $\overline{\mathbf{W}}$ is the steady state, $\mathbf{R}(\overline{\mathbf{W}}, 0)$ is very small and can be omitted. The linear equation for the perturbations is obtained by neglecting all non-linear terms:

$$\frac{\mathrm{d}\widetilde{\mathbf{W}}}{\mathrm{d}t}(t) = \frac{\partial \mathbf{R}}{\partial \mathbf{W}}\,\widetilde{\mathbf{W}}(t) + \frac{\partial \mathbf{R}}{\partial \mathbf{v}_g}\,\widetilde{\mathbf{v}}_g(t) \tag{2}$$

Assuming a harmonic excitation at frequency $\omega$, eq. (2) can be transferred into the frequency domain, yielding the governing equations for the LFD method:

$$\left[\frac{\partial \mathbf{R}}{\partial \mathbf{W}} - j\omega\underline{\mathbf{I}}\right]\widehat{\mathbf{W}} = -\frac{\partial \mathbf{R}}{\partial \mathbf{v}_g}\,\widehat{\mathbf{v}}_g \tag{3}$$

Thus a huge, but sparse system of linear equations is obtained relating the Fourier coefficient of the gust excitation $\widehat{\mathbf{v}}_g$ to the Fourier coefficient of the fluid unknowns $\widehat{\widetilde{\mathbf{W}}}$. The right-hand side vector is computed applying the chain rule:

$$\frac{\partial \mathbf{R}}{\partial \mathbf{v}_g}\,\widehat{\mathbf{v}}_g = \frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}}\,\frac{\partial \dot{\mathbf{x}}}{\partial \mathbf{v}_g}\,\widehat{\mathbf{v}}_g$$

where $\dot{\mathbf{x}}$ is the vector of mesh velocities since the gust is simulated applying the field velocity method [13]. The equation can be further simplified by the simple relation $\dot{\mathbf{x}} = -\widehat{\mathbf{v}}_g$. As for the LFD method for forced motion, the fluid Jacobian matrix $\frac{\partial \mathbf{R}}{\partial \mathbf{W}}$ is computed analytically, while the influence of the grid velocities in term $\frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}}$ is computed using central finite differences. The complex-valued vector $\widehat{\mathbf{v}}_g$ can be calculated by the analytical expression

$$\widehat{\mathbf{v}}_g(\mathbf{x}, \omega) = v_{gz} e^{j\omega(\mathbf{x}-\mathbf{x}_0)}$$

with $v_{gz}$ denoting the gust amplitude, $\mathbf{x}$ the grid coordinate and $\mathbf{x}_0$ a reference point. The frequency of excitation is defined by

$$\omega = U_\infty \frac{2\pi}{L_g}$$

with $L_g$ as the gust length and $U_\infty$ as the freestream velocity.

### 2.1 Proper Orthogonal Decomposition

Snapshots are generated at discrete frequencies with sinusoidal excitation solving the linearised frequency-domain system given in eq. (3). Solutions $\widehat{\mathbf{W}}_k$ are stored as columns in the snapshot matrix $\underline{\mathbf{S}}$. The POD basis $\underline{\mathbf{\Phi}}$ is obtained as a linear combination of snapshots

$$\underline{\mathbf{\Phi}} = \underline{\mathbf{S}}\boldsymbol{\nu}_k \tag{4}$$

where the vector $\boldsymbol{\nu}_k$ is scaled so that vectors in $\underline{\mathbf{\Phi}}$ are unit length. The eigenvalue problem

$$\underline{\mathbf{S}}^H\underline{\mathbf{S}}\boldsymbol{\nu}_k = \lambda_k\boldsymbol{\nu}_k, \tag{5}$$

with $\underline{\mathbf{S}}^H\underline{\mathbf{S}}$ a symmetric, positive definite matrix, is solved to ensure the best possible approximation in eq. (4). The relative information content a certain mode $\lambda_k$ contributes to the system, also often referred to as energy, is given by

$$r_k = \lambda_k \left( \sum_i \lambda_i \right)^{-1} \tag{6}$$

and can be used to decrease the number of modes further by only considering those with a high relative information content.

In order to use the POD basis also for a time-domain ROM the POD basis can either be enhanced by its complex-conjugate or alternatively the complex-conjugate snapshots are included before calculating the POD basis. The time-dependent expression for the perturbations in eq. (2) can be reduced to an ordinary differential equation with the number of degrees of freedom corresponding to the POD subspace:

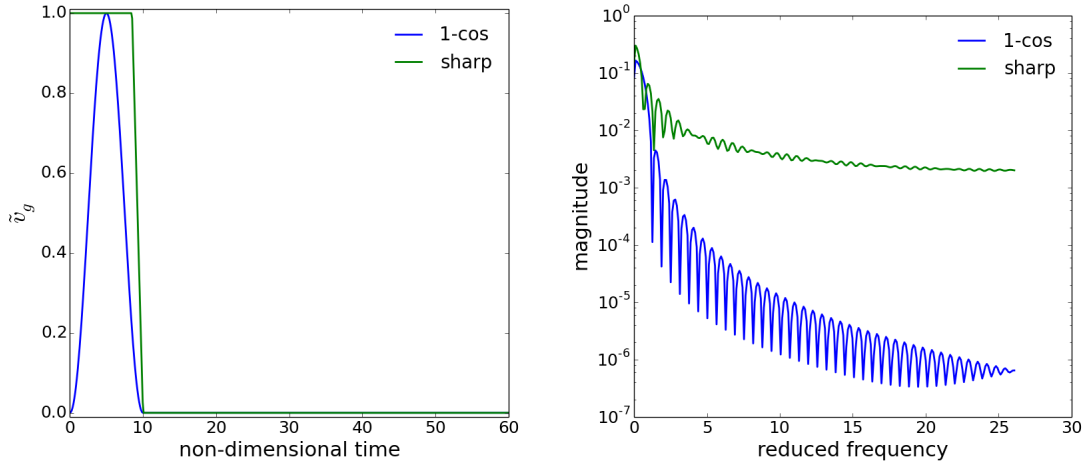
$$\frac{\mathrm{d}\widetilde{\mathbf{q}}}{\mathrm{d}t}(t) = \underline{\mathbf{\Phi}}^H \frac{\partial \mathbf{R}}{\partial \mathbf{W}} \underline{\mathbf{\Phi}}\widetilde{\mathbf{q}}(t) + \underline{\mathbf{\Phi}}^H \frac{\partial \mathbf{R}}{\partial \dot{\mathbf{x}}} \widetilde{\mathbf{v}}_g(t). \tag{7}$$

with $\widetilde{\mathbf{q}}$ denoting the POD or modal coefficients used to reconstruct the flow solution as

$$\widetilde{\mathbf{W}} = \underline{\mathbf{\Phi}}\,\widetilde{\mathbf{q}}.$$

A backward Euler scheme combined with Newton's method is used to integrate eq. (7) in time. The small-sized differential equation is stiff and an explicit time integration scheme was thus unstable for suitable time step sizes.

The same basis can be used to project the LFD system in eq. (3) onto the POD subspace as shown in [12]. However, there are two advantages of the time-domain ROM over the frequency-domain method. First, maximum loads cannot be computed within the frequency-domain and an additional inverse Fourier transformation has to be performed to reconstruct the flow field. Secondly, the magnitudes at high frequencies for a 1-cos gust excitation are negligibly small as seen in Figure 1(b). Therefore, the responses of a limited number of sinusoidal gust can be computed with a frequency-domain ROM and superposed according to their participation on a 1-cos gust. High freqencies cannot be neglected if a sharp-edged gust is considered on the other hand. A frequency-domain ROM must compute the response for all Fourier coefficients reducing significantly the efficiency of the method compared to its time-domain couterpart.




(a) time-domain (b) frequency-domain

Figure 1: Time- and frequency-domain signals of 1-cos ($L_g = 10\,\mathrm{m}$) and sharp-edged gust

### 2.2 Computational Fluid Dynamics Solver

The generation of snapshots to obtain the POD basis as well as the computation of the full-order reference solution is done with an in-house, semi-meshless Navier-Stokes flow solver [14, 15] coupled with the Spalart-Allmaras turbulence model [16]. Convective fluxes are discretised using upwind schemes, specifically the Osher solver for the mean flow equations [17]. A weighted least squares procedure calculates the gradients of the flow variables, required for viscous fluxes as well as source terms in the turbulence model. The steady-state solution is obtained applying a fully implicit backward Euler method with local time-stepping, while additionally a second order dual-time stepping is utilised in unsteady time-domain simulations. Linear equations arising from the implicit time integration and from the full-order LFD system are solved throughout using a restarted generalised conjugate residual method preconditioned with an incomplete lower-upper factorisation [18].

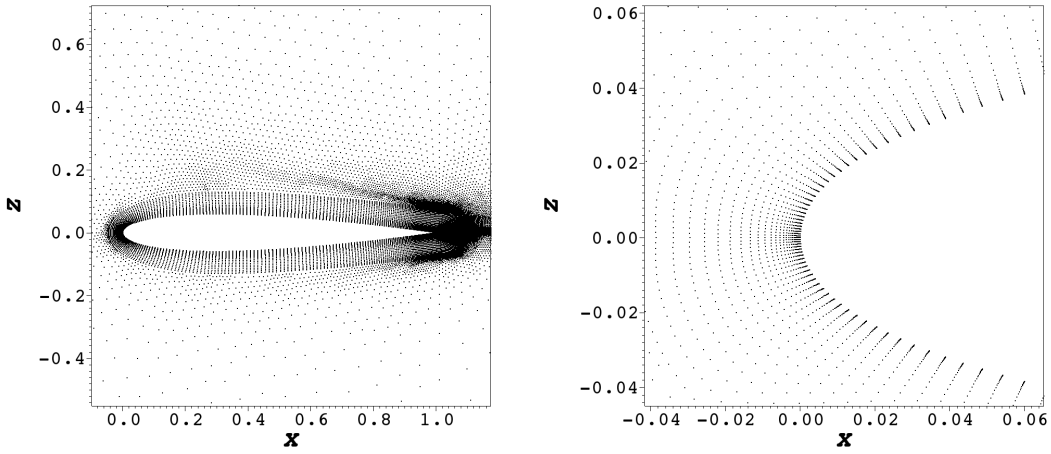
## 3 Results

Results are presented for the NACA 0012 aerofoil using a computational domain discretised with about $30 \times 10^3$ points. The point distribution has a structured layer near the wall to ensure a sufficient boundary layer resolution, seen in Figure 2, while the farfield distance is set to 50 chord lengths. Comparisons are shown for three different flow conditions to demonstrate the capability of the ROM for a variety of problems including sub- (S1) and transonic (T1) cases and an additional case exhibiting a shock-induced flow separation (T2). Comparisons of the full order model with a frequency-domain ROM are presented for S1 and T1 in [12]. The Reynolds number based on the chord length is 10 million, while the remaining flow parameters are summarised in Table 1.

The steady pressure coefficients on the surface are shown in Figure 3(a). A recompression shock can be observed for the transonic test cases T1 and T2. The shock position of T2 with a higher angle of attack is more upstream than in case T1. Since the shock position in attached flow moves downstream with increasing angle of attack, this reverse shock motion indicates flow separation. In Figure 3(b) the corresponding non-dimensional x-

Table 1: Main flow parameters of three considered test cases

| Case | Mach number | Angle of attack [deg.] |
|---|---|---|
| S1 | 0.3 | 0.0 |
| T1 | 0.8 | 0.0 |
| T2 | 0.8 | 0.0 |




(a) computational domain (b) computational domain, zoomed-in

Figure 2: Computational domain of NACA 0012

velocity component is visualised. Besides the large supersonic zone, a region with negative velocities, as expected, can be observed behind the shock.

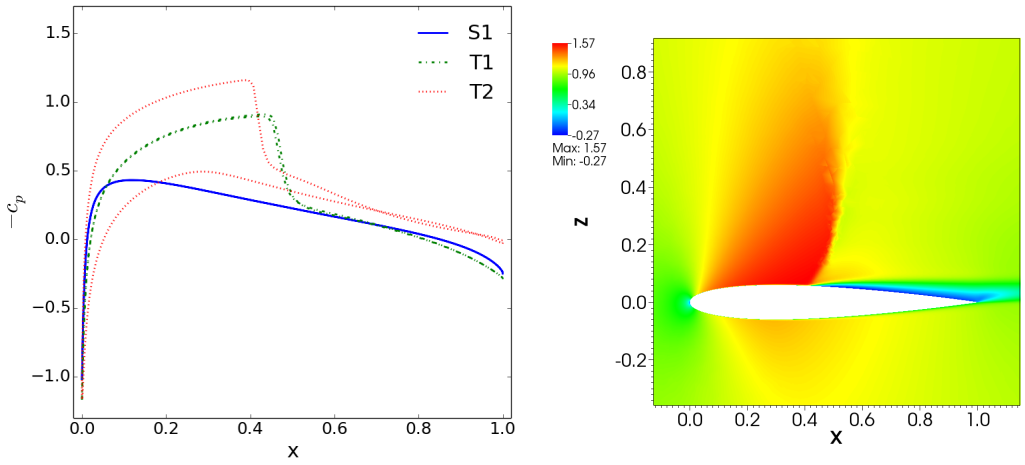
The ROM is generated using 40 snapshots of sinusoidal gust excitations computed at reduced frequencies in the interval $(0, 2\pi]$. The frequencies are distributed following the formula given in [12]:

$$\omega_k = \omega_0 + e^{-k\Delta\omega} \tag{8}$$

with suggested parameters $\omega_0 = 0.0$ and $\Delta\omega = 0.25\pi$. The responses to these sinusoidal gusts are computed with the LFD solver. The relative information content of the POD modes, i.e. the influence of each mode, and the dominant mode for case T2 are shown in Figure 4.

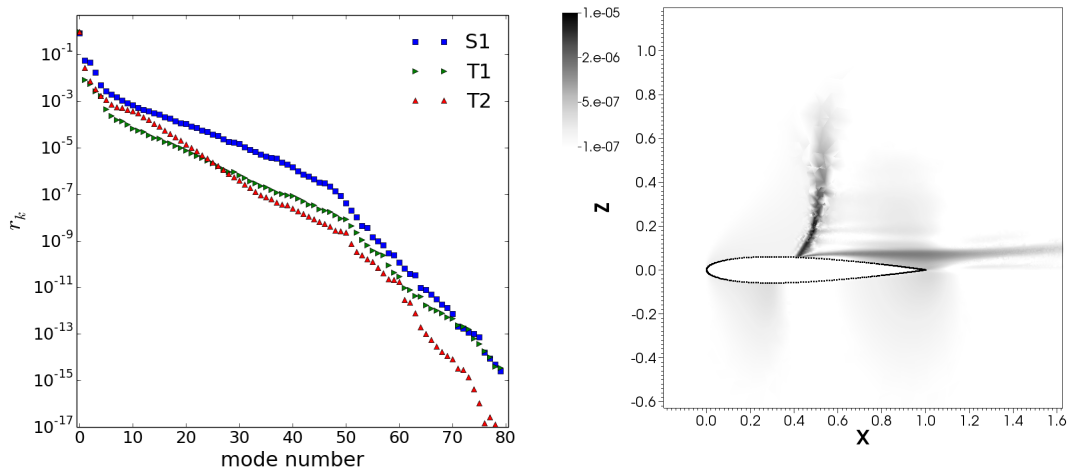
A drop of two orders of magnitude within the first 5 modes can be observed in Figure 4(a) for all test cases. A linear decrease until about the 50th mode can then be found for both attached-flow cases with a larger gradient in transonic flow. It is interesting to note, that the slope for the detached-flow case T2 shows a bump between the 10th and 15th mode and a linear decrease thereafter. This reflects a richer physical content close to the buffet onset. Since the ordinate is presented in logarithmic scale the linear decrease indicates the exponential decay of the relative information content. A criteria of $\sum_k r_k = 0.9999$ is used for the POD reduction throughout resulting in 29 basis vectors for case S1, 16 for T1 and 18 for T2.

For the detached-flow test case T2, the pressure's magnitude of the first POD mode, i.e. the mode with highest information content of $r_k = 0.955$, is displayed in Figure 4(b). In addition to the shock region, which is also present for the transonic attached-flow




(a) surface pressure coefficients (b) x-velocity component for T2

Figure 3: Steady surface pressure coefficients for all cases and non-dimensional x-velocity for case T2



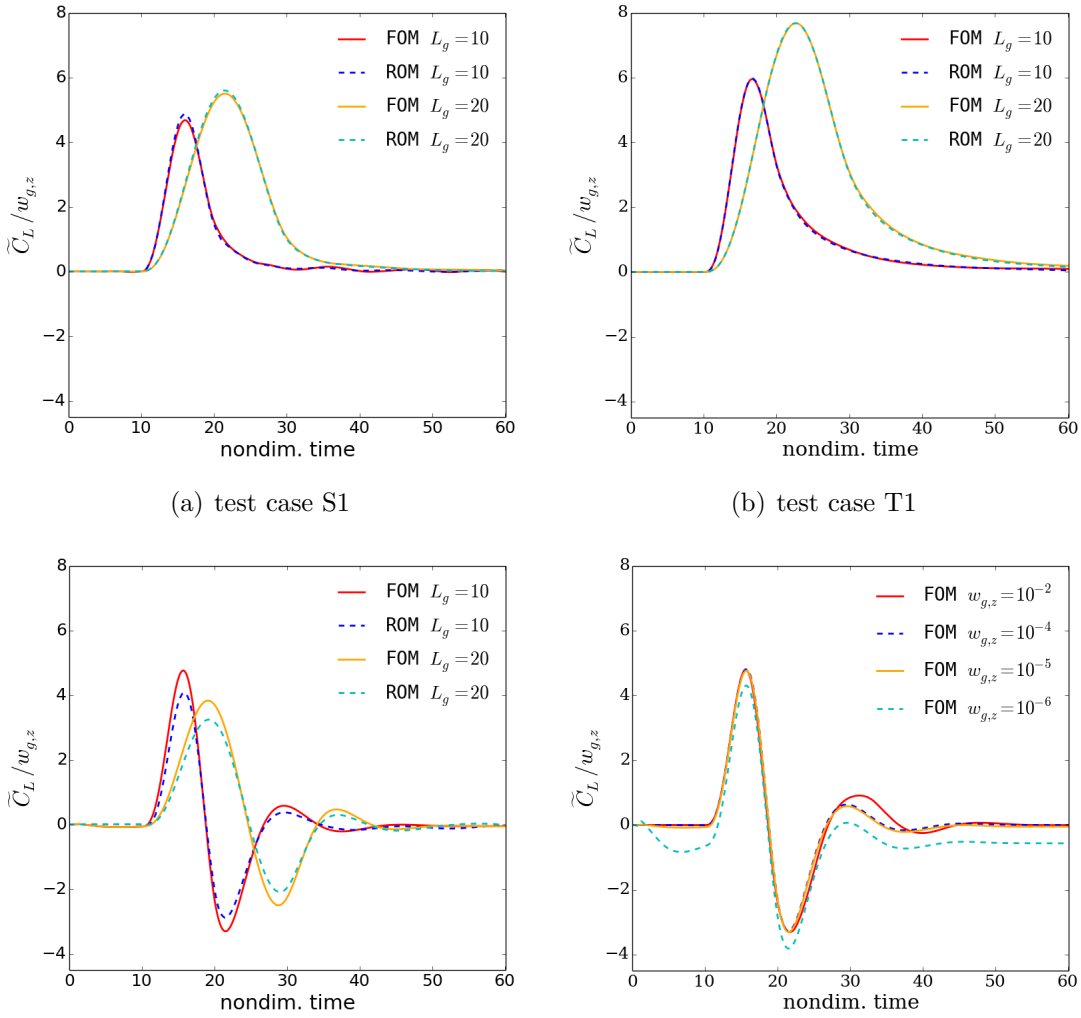

(a) relative information content (b) isobars of first POD mode for T2

Figure 4: Information content of POD modes and isobars of first POD mode for test case T2

case [12], high magnitudes can be observed around the free shear layer. If the shock is moving, the detached-flow topology changes as well explaining the similar magnitudes in both regions.

### 3.1 Responses due to 1-cos gusts

Comparisons of the lift response due to different 1-cos gusts are presented in Figure 5. The gust lengths are $L_g = 10\,\mathrm{m}$ and $L_g = 20\,\mathrm{m}$ with a reference point 10 m upstream of the aerofoil. The full order reference solutions (FOM) are computed using the non-linear time-domain solver with a non-dimensional time step of 0.02. For both cases with attached flow the non-dimensional gust amplitude $w_{g,z} = v_{g,z}/U_\infty = 10^{-2}$ is chosen throughout. Overall, good agreement can be found between the time-domain ROM and the full-order reference. For the subsonic case in Figure 5(a) the maximum is slightly overpredicted by the ROM because of amplitude non-linearities. The influence of the gust amplitude on




(a) test case S1

(b) test case T1

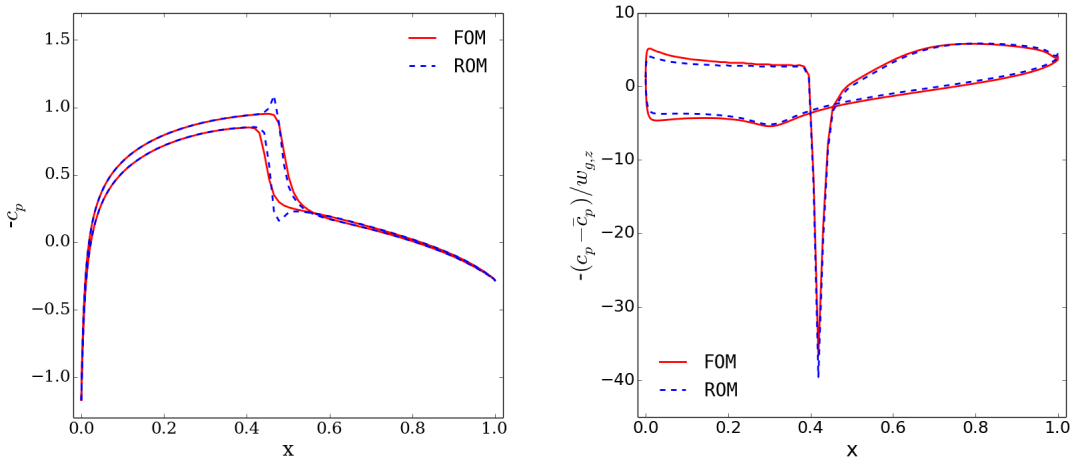
(c) test case T2

(d) test case T2 at $L_g = 10\,\mathrm{m}$

Figure 5: Change in lift coefficient due to different 1-cos gusts

the lift coefficient is presented in [12] showing a decrease in amplitude of the non-linear solution compared to a time-linearised solver. The agreement between both methods is better for the attached-flow transonic case T1. The maximum response in lift coefficient and the decay afterwards agree well.

The lift response changes significantly for case T2 with shock-induced separated flow as seen in Figure 5(c) and 5(d), exhibiting a distinct minimum following the global maximum. This indicates a large influence of an additional frequency, which is known from resonance peaks in frequency response functions close to buffet onset [19, 20]. Detached-flow cases are more sensitive concerning the excitation amplitude and its effect is thus analysed in Figure 5(d). While the global maximum in lift response is almost independent of the gust amplitude, the subsequent oscillations show a clear influence which diminishes for non-dimensional amplitudes $w_{g,z} \leq 10^{-4}$. At very small amplitudes with $w_{g,z} \leq 10^{-6}$, the solution of the RANS solver becomes inaccurate because of limited machine precision. In Figure 5(c) the results of the linear ROM are compared to non-linear responses with a gust amplitude of $w_{g,z} = 10^{-5}$. The reduced formulation is capable to reproduce the complex progression of the lift coefficient qualitatively, while underestimating its extrema.




(a) test case T1 (b) test case T2

Figure 6: Surface pressure coefficients at maximum lift response for $L_g = 10\,\mathrm{m}$

In Figure 6, the instantaneous pressure coefficients are presented at the time step of the maximum lift response for both transonic test cases. For the attached-flow case T1 an amplitude was chosen yielding visual differences to the steady-state solution in Figure 3(a). Although amplitude effects can be observed at the shock location, the ROM prediction agrees well with the corresponding full order solution. The values around the shock on both surfaces are overpredicted by the linear ROM, caused by the superposition of the steady-state solution and the POD modes with rather high amplitudes. A smaller gust amplitude is simulated for the detached-flow case T2 and a plot of instantaneous pressure coefficients would simply show the steady state. Thus, the normalised difference to the steady surface pressure coefficient

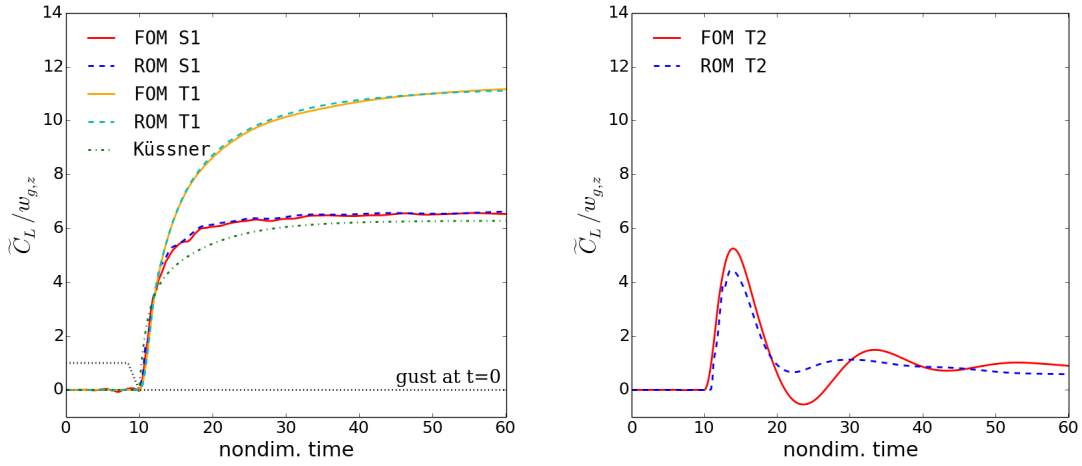
$$\widetilde{c}_p = \frac{c_p(t) - \bar{c}_p}{w_{g,z}}$$

is displayed in Figure 6(b) instead. Good agreement between both methods is obtained in the region of separated flow, while the ROM predicts a slightly larger shock peak and smaller pressure differences around the leading edge.

### 3.2 Responses due to sharp-edged gust

After simulating responses due to 1-cos gusts, the capabilities of the time-domain ROM are demonstrated for a gust shape not easily represented by a Fourier series with a finite number of coefficients. Contrary to a 1-cos gust, a sharp-edged gust excites also high frequencies, as shown in Figure 1(b). The Fourier series cannot be truncated which contradicts the application of a ROM in frequency-domain.

Responses to a smoothed sharp-edged gust with a non-dimensional gust amplitude of $10^{-3}$ are computed for the three test cases. The step in gust velocity was smoothed with a linear increase over 1.5 chord length, shown as dotted black line in Figure 7(a), to improve convergence of the full order model. Results for the same gust shape are then calculated by the ROM using the same POD basis as for the 1-cos responses. For both attached-flow cases, the response in lift coefficients shows qualitatively the same progression as the Küssner function [21]. Excellent agreement can be found between results computed




(a) attached-flow cases S1 and T1

(b) detached-flow case T2

Figure 7: Change in lift coefficient due to smoothed sharp-edged gust with $w_{g,z} = 10^{-3}$

by the ROM and the full order model. The response of the subsonic case obtained by the non-linear flow solver shows minor oscillations which are filtered by the ROM. The corresponding lift coefficient for the detached-flow case T2, shown in Figure 7(b), differs significantly exhibiting a maximum and damped oscillations thereafter. This distinct behaviour can be explained by looking at frequency responses due to pitch motion as presented in [19, 20] for constant Mach number and increasing angle of attack.

At attached-flow conditions the lift response is similar to predictions by linear potential theory such as Theodorson [22] starting with a large value at zero frequency and decreasing monotonically thereafter. However, close to shock buffet onset the quasi-steady derivative is decreasing significantly, which is the reason for the reduced lift gain after the sharp-edged gust passed the aerofoil. Moreover, the frequency response function exhibits a local maximum at the buffet frequency where an eigenvalue of the fluid Jacobian matrix is weakly damped causing the oscillations in Figure 7(b). The ROM predicts a maximum at the same time step, but at a smaller value and the lift coefficient oscillates at much smaller amplitude.

## 4 Conclusion

The paper describes a time-domain reduced order model based on proper orthogonal decomposition for gust response analysis. The reduced order model is trained by responses due to sinusoidal gusts computed with a linear frequency-domain solver. The linearised Reynolds-averaged Navier-Stokes equations are projected onto the subspace yielding a small-sized ordinary differential equation for the modal coefficients, subsequently used to reconstruct the physical solution from the projection basis. Different disturbances such as 1-cos or sharp-edged gusts can then be predicted at low cost.

Results are presented for the NACA 0012 aerofoil at three different flow conditions covering sub- and transonic flow without boundary layer separation and a shock-induced, separated flow. Lift responses of 1-cos gusts with two different wave lengths show an excellent agreement between the reduced order model and the non-linear, time-domain full order reference solution for the attached-flow cases. Only a minor difference can be observed in the subsonic case with the linear reduced order model overpredicting the max-

imal lift response due to amplitude non-linearities in the full order model. A comparable good agreement is obtained for the instantaneous surface pressures showing only small differences around the shock location for the transonic case. The lift responses due to a sharp edged gust give a progression similar to the Küssner function for both attached-flow cases demonstrating a good agreement between the considered methods. The lift response for the test case with shock-induced separation however changes significantly showing additional oscillations for both gust shapes due to the proximity to an inherent flow instability. The reduced order model can predict these features qualitatively, while underestimating amplitudes.

Future work will extend the presented methods to include additional Taylor coefficients to account for amplitude non-linearities. Moreover, the reduced order model will also be coupled with a controller for gust load alleviation.

## 5 Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 636053.

## REFERENCES

[1] Dimitrov, D. and Thormann, R., "DLM-Correction Methods for Aerodynamic Gust Response Prediction," *International Forum on Aeroelasticity and Structural Dynamics*, 2013, IFASD 2013-24C.

[2] Wales, C., Jones, D., and Gaitonde, A., "Prescribed Velocity Method for Simulation of Aerofoil Gust Responses," *Journal of Aircraft*, Vol. 52, No. 1, 2015, pp. 64–76.

[3] Raveh, D. E., "CFD-Based Models of Aerodynamic Gust Response," *Journal of Aircraft*, Vol. 44, No. 3, 2007, pp. 888–897.

[4] Reimer, L., Ritter, M., Heinrich, R., and Krüger, W., "CFD-based Gust Load Analysis for a Free-flying Flexible Passenger Aircraft in Comparison to a DLM-based Approach," *22nd AIAA Computational Fluid Dynamics Conference*, 2015, AIAA 2015-2455.

[5] Wales, C., Gaitonde, A., and Jones, D., "Reduced order modelling for aeroelastic aerofoil response to a gust," *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, 2013, AIAA 2013-0790.

[6] Dowell, E. H., Hall, K. C., Thomas, J. P., Kielb, R. E., Spiker, M. A., Li, A., and Denegri, Jr., C. M., "Reduced Order Models in Unsteady Aerodynamic Models, Aeroelasticity and Molecular Dynamics," *ICAS - 26th Congress of International Council of the Aeronautical Sciences*, 2008, ICAS 2008-0.1.

[7] Lumley, J. L., *The Structures of Inhomogeneous Turbulent Flow*, Atmospheric Turbulence and Radio Wave Propagation, 1965.

[8] Kim, T., "Frequency-Domain Karhunen-Loève Method and Its Application to Linear Dynamic Systems," *AIAA Journal*, Vol. 36, No. 11, 1998, pp. 2117–2123.

[9] Thormann, R. and Widhalm, M., "Linear-Frequency-Domain Predictions of Dynamic-Response Data for Viscous Transonic Flows," *AIAA Journal*, Vol. 51, No. 11, 2013, pp. 2540–2557.

[10] Hall, K. C., Thomas, J. P., and Dowell, E. H., "Proper Orthogonal Decomposition Technique for Transonic Unsteady Aerodynamic Flows," *AIAA Journal*, Vol. 38, No. 10, 2000, pp. 1853–1862.

[11] Thomas, J. P., Dowell, E. H., and Hall, K. C., "Three-Dimensional Transonic Aeroelasticity Using Proper Orthogonal Decomposition-Base Reduced-Order Models," *Journal of Aircraft*, Vol. 40, No. 3, 2003, pp. 544–551.

[12] Bekemeyer, P. and Timme, S., "Reduced Order Gust Response Simulation using Computational Fluid Dynamics," *57th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, AIAA SciTech*, 2016, AIAA 2016-1485.

[13] Parameswaran, V. and Baeder, J. D., "Indicial aerodynamics in compressible flow-direct computational fluid dynamic calculations," *Journal of Aircraft*, Vol. 34, No. 1, 1997, pp. 131–133.

[14] Kennett, D. J., Timme, S., Angulo, J., and Badcock, K. J., "An Implicit Meshless Method for Application in Computational Fluid Dynamics," *International Journal for Numerical Methods in Fluids*, Vol. 71, No. 8, 2013, pp. 1007–1028.

[15] Kennett, D. J., Timme, S., Angulo, J., and Badcock, K. J., "Semi-Meshless Stencil Selection for Anisotropic Point Distributions," *International Journal of Computational Fluid Dynamics*, Vol. 26, No. 9-10, 2012, pp. 463–487.

[16] Spalart, P. R. and Allmaras, S. R., "A One-Equation Turbulence Model for Aerodynamic Flows," *Recherche Aerospatiale*, Vol. 1, 1994, pp. 5–21.

[17] Osher, S. and Solomon, F., "Upwind Difference Schemes for Hyperbolic Systems of Conservation Laws," *Mathematics of Computation*, Vol. 38, 1982, pp. 339–374.

[18] Saad, Y., *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2nd ed., 2003.

[19] Nitzsche, J., "A Numerical Study on Aerodynamic Resonance in Transonic Separated Flow," *International Forum on Aeroelasticity and Structural Dynamics*, June 2009, IFASD-2009-126.

[20] Thormann, R., Nitzsche, J., and Widhalm, M., "Time-linearized Simulation of Unsteady Transonic Flows with Shock-Induced Separation," *European Congress on Computational Methods in Applied Sciences and Engineering*, 2012, ECCOMAS 2012.

[21] Küssner, H. G., "Zusammenfassender Bericht über den instationären Auftrieb von Flügeln," *Luftfahrtforschung*, Vol. 13, No. 2, 1936, pp. 410–424.

[22] Theodorsen, T., "General Theory of Aerodynamic Instability and the Mechanism of Flutter," *NACA Report*, , No. 496, 1935, pp. 413–433.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE REGULARISATION OF NON-REFLECTING BOUNDARY CONDITIONS NEAR ACOUSTIC RESONANCE

**Christian Frey, Hans-Peter Kersken**

German Aerospace Center (DLR)
Linder Hohe, 51147 Cologne, Germany
e-mail: christian.frey@dlr.de, hans-peter.kersken@dlr.de

**Keywords:** Nonreflecting Boundary Conditions, Acoustic Resonance, Turbomachinery Flutter

**Abstract.** *In this paper we discuss the phenomenon of acoustic resonance in the context of non-reflecting boundary conditions and its impact on the analysis of turbomachinery blade flutter. Acoustic resonance is observed on the one hand when flutter curves, i.e., the aerodynamic damping as a function of the interblade phase angle, displays singularities at certain angles. On the other hand, the implementation of appropriate inlet and outlet boundary conditions naturally leads to the normal mode analysis of the Euler equations on eiher cylindrical (2D) or annular ducts (3D). To formulate boundary conditions one typically carries out the normal mode analysis and thus decomposes the space of normal modes as a direct sum of incoming and outgoing modes. This decomposition, however, breaks down when the generalised eigenvalue problem no longer produces a complete set of eigenvectors and one obtains non-trivial Jordan blocks. Both phenomena are shown to be due to acoustic modes with vanishing normal group velocity.*

*The main goal of this paper is to outline how the numerical boundary conditions can be regularised near acoustic resonance. The impact of this regularisation on the prediction of flutter stability is then demonstrated. Moreover, we interpret the regularisation as an interpolation between the exact non-reflecting and the approximative characteristic, one-dimensional boundary conditions.*

## INTRODUCTON

The design of modern turbomachinery increasingly relies on unsteady flow simulations, especially to avoid engine failure due to blade flutter. Typically, the flutter analysis is based on the prediction of the unsteady flow response for a forced motion of the blades at various operating conditions [1]. Usually, this forced motion is defined by the structural displacement, $q^{\mathrm{s}}$, corresponding to a structural eigenmode $\psi$,

$$q^{\mathrm{s}}(t,x) = \mathrm{Re}\left[q^{\mathrm{s}}_{\mathrm{mod}}(t)\psi(x)\right],$$

where the modal displacement is the harmonic oscillation

$$q^{\mathrm{s}}_{\mathrm{mod}}(t) = e^{\mathrm{i}\omega t},$$

with the angular frequency $\omega$ being the eigenfrequency $\sqrt{k/m}$. From the unsteady pressure on the blade surface one can determine the aerodynamic damping,

$$\delta = -\frac{\mathrm{Re}\, W_{\mathrm{cyc}}}{\omega^2 m}, \quad W_{\mathrm{cyc}} = \int_0^{2\pi/\omega}\int_\Gamma \left(\dot{q}^{\mathrm{s}}_{\mathrm{mod}}(t)\psi(x)\right)^H \left(p(t,x)\vec{n}(t,x)\right) dS(x)dt, \tag{1}$$

and thus estimate if the coupled system is stable at rest. Whereas the real part of $W_{\mathrm{cyc}}$ represents the net work per period done on the structure by the unsteady flow, the imaginary part of $W_{\mathrm{cyc}}$ gives rise to the aerodynamic stiffness coefficient

$$\kappa = \frac{\mathrm{Im}\, W_{\mathrm{cyc}}}{\omega^2 m},$$

from which a frequency shift of the coupled fluid structure system can be determined. One

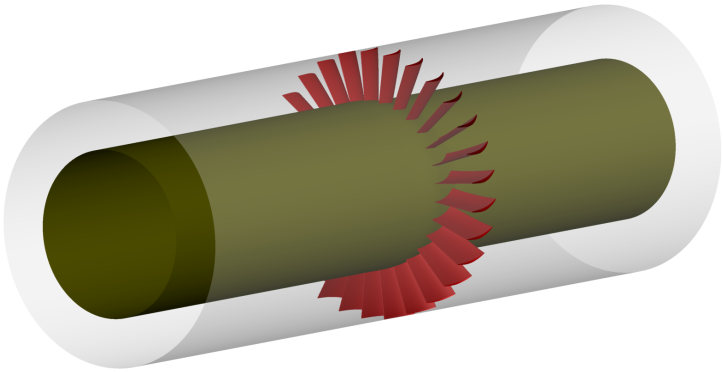


Figure 1: Idealised flow domain for the anaylsis of blade flutter

usually considers an idealised flow domain as in Fig. 1, i.e., the interactions with neighbouring blade rows are neglected.

Under the assumption that the blade row consists of $N$ structurally uncoupled blades, each eigenmode $\psi$ that one obtains from the structural analysis of one blade, gives rise to the full annulus modes

$$(\psi, e^{\mathrm{i}\sigma}\psi, e^{2\mathrm{i}\sigma}\psi, \ldots, e^{(N-1)\mathrm{i}\sigma}\psi),$$

where the $n$-th component represents the mode shape of the $n$-th blade, and $\sigma \in [-\pi, \pi)$ is such that $N\sigma = 2\pi$. $\sigma$ is called the *interblade phase angle*. The integer

$$N_\mathrm{d} = \frac{2\pi}{\sigma}$$

is called the *nodal diameter*. A full annulus mode has an interblade phase angle of $\sigma$ if the modal displacement of the blade $n$ lags that of blade $n+1$ by a time shift of $\sigma/\omega$. In this case the unsteady flow response $q(t, x)$ will exhibit the same space time symmetry, i.e.,

$$q(t, x, r, \vartheta + 2\pi/N) = q(t + \sigma/\omega, x, r, \vartheta),$$

Together with (1) one deduces that blade vibrations of different interblade phase angles are uncoupled. Therefore, one usually computes the aerodynamic damping for each interblade phase angle. The result is plotted in the so-called *damping curve*, although, strictly speaking, it consists of a discrete set of $N$ values.

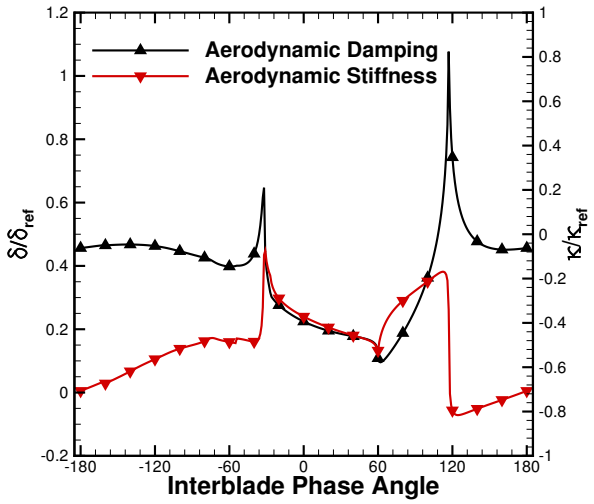



Figure 2: Aerodynamic damping and stiffness over interblade phase angle for a compressor blade.

Figure 2 shows the aerodynamic damping and stiffness curves for the academic compressor test case Standard Configuration 10, cf. [2]. The unsteady flow response is computed using the time-linearised approach [3, 4]. One observes that the damping appears to be "smooth" except for a few interblade phase angles.

It is easy to see that the flutter curve corresponds to the discrete Fourier transform of the influence coefficients, i.e., the modal forces on one blade which are due to the vibration of another one [5]. In many cases, the influence coefficents decay rapidly as the distance between two blades of the blade row increases and one obtains smooth flutter curves. For instance, if the influence coefficients of all but direct neighbours can be neglected the flutter curve can be approximated rather accurately by a function $\tilde{\delta}(\sigma) = c_0 + c_1 \cos(\sigma - \sigma_1)$, cf. [6]. However, this may not always be the case, as the spikes and the sudden drops in the curves in Fig. 2 show. When these singularities are observed near a minimum damping close to zero, the CFD based prediction of flutter stability becomes questionable, as the infinite slope of the flutter curve indicates a considerable uncertainty in the damping.

The occurence of these singularities is due to the phenomenon of *acoustic resonance* [6]. One speaks of acoustic resonance if a flow unsteadiness, in this case caused by a vibrating blade, is resonant with an acoustic wave whose energy does not propagate out of the configuration. As is well-known, the energy transport of a wave is closely related to its group velocity. For such a standing wave to be excited it is also necessary that the unsteadiness has the same circumferential wave number (or nodal diameter) as the acoustic mode. Consequently, one observes acoustic resonance only at certain interblade phase angles. In turbomachinery, acoustic resonance can occur at specific operating conditions. Often the source of the instability unknown. Alongside blade vibration possible sources of excitation include pressure disturbances due to the periodic blade passings, rotating instabilities and vortex shedding. For acoustic resonances observed during turbomachinery experiments, the reader is refered to [7, 8] and the literature cited therein.

The goal of this work is to study the issue of acoustic resonance in the context of non-reflecting boundary conditions. When acoustic resonance occurs, there no longer exists a complete set of duct eigenmodes so that the standard formulations of inlet and outlet boundary conditions (see e.g. [9, 10]) are no longer possible.

The paper is organised as follows. First we revisit the theory of non-reflecting boundary conditions for the linearised Euler equations. Then we show how to regularise the boundary conditions by shifting the frequency slightly to the complex lower half plane. This technique has been used by Moinier and Giles [11] to simplify the determination of the propagation direction of normal modes computed numerically. We demonstrate the impact of the regularisation on the prediction of flutter stability by means of a compressor blade. It is shown that the regularisation has the effect of a mollifier for the flutter curves. Most of the mathematical arguments carry over to symmetrisable hyperbolic problems. Therefore, we will use general arguments whenever possible.

We interpret the regularisation as an interpolation between the exact non-reflecting and the characteristic boundary conditions. Since two-dimensional non-reflecting boundary conditions are typically used for three-dimensional turbomachinery configurations, we discuss the question to what extent the results obtained with two-dimensional boundary conditions are representative in real applications.

## 1 NON-REFLECTING BOUNDARY CONDITIONS

We consider the Euler equations in $\mathbb{R}^n$ in conservative variables $q = (\rho, \rho U, \rho e_\mathrm{t})$

$$\frac{\partial q}{\partial t} + \operatorname{div} F(q) = 0 \tag{1.1}$$

with the inviscid flux

$$F(q) = \begin{pmatrix} \rho U \\ \rho U \otimes U + p\,\mathrm{Id} \\ \rho U h_\mathrm{t} \end{pmatrix}.$$

It is well-known that the Euler equations form a symmetrisable hyperbolic system of conservation laws [12]. In particular, the Hessian of the entropy density, viewed as a function

$$(\rho, \rho U, \rho e_\mathrm{t}) \to -\rho s,$$

is a symmetriser, i.e.,

$$T(q) = \left( \frac{\partial^2(-\rho s)}{\partial q_i \partial q_j}\Big|_q \right)_{i,j}$$

is symmetric positive definite and $T(q)\frac{\partial F^k}{\partial q}|_q$ is symmetric for all $q$ and $k$. Assume that the computational domain is the half space $(-\infty, 0) \times \mathbb{R}^{n-1}$ and that the solution is periodic in the $n-1$ tangential variables. Consider the linearised equations at some mean flow conditions, e.g., the area average of the conservative variables. Transformation of the linearised equations into the frequency and wave number domain thus yields a linear system for $\hat{q}_{\omega,\xi}$,

$$\left(\omega + \sum_k \xi_k \frac{\partial F^k}{\partial q}\right) \hat{q}_{\omega,\xi} = 0. \tag{1.2}$$

The matrix on the left-hand side is symmetric with respect to the inner product defined by $T$ and, up to a factor i, equals the principal symbol [13]. Its determinant is a homogeneous polynomial of degree $n+2$ in $(\omega, \xi) \in \mathbb{R}^{n+1}$, whose zeros form the so-called characteristic variety $\mathscr{C}$, cf. [14]. Suppose that at a certain point $(\omega, \xi) \in \mathscr{C}$ the characteristic variety is smooth and that the lines $\xi = \text{const}$ are, near $(\omega, \xi)$, transversal to $\mathscr{C}$. Then, the angular frequency $\omega$ such that Eqn. (1.2) has a non-trivial solution, can be written locally as a function of $\xi$, and the *group velocity* is then defined by

$$v_{\mathrm{g}}(\omega, \xi) = -\frac{\partial \omega}{\partial \xi}. \tag{1.3}$$

The Euler equations for an ideal gas can be rewritten in primitive variables, $q_{\mathrm{prim}} = (\rho, U, p)$,

$$\frac{\partial q_{\mathrm{prim}}}{\partial t} + \sum_j A^j \frac{\partial q_{\mathrm{prim}}}{\partial x^j} = 0$$

where

$$A^j = U^j \,\mathrm{Id} + \begin{pmatrix} 0 & \rho e_j^T & 0 \\ 0 & 0 & \rho^{-1} e_j \\ 0 & \gamma p e_j^T & 0 \end{pmatrix},$$

and $e_1, \ldots, e_n$ is the standard basis of $\mathbb{R}^n$. The characteristic variety of the linearised Euler equations is thus given by those $(\omega, \xi)$ satisfying

$$(\omega + \xi \cdot U)^n ((\omega + \xi \cdot U)^2 - a^2 \|\xi\|^2) = 0. \tag{1.4}$$

In primitive variables, the convective modes

$$r_1 = \begin{pmatrix} \rho \\ 0 \\ 0 \end{pmatrix}, \quad r_2 = \begin{pmatrix} 0 \\ a e_1 \\ 0 \end{pmatrix}, \ldots r_n = \begin{pmatrix} 0 \\ a e_{n-1} \\ 0 \end{pmatrix}, \tag{1.5}$$

together with the acoustic eigenvectors

$$r_{n+1} = \begin{pmatrix} \rho \\ a\frac{\xi}{\|\xi\|} \\ \gamma p \end{pmatrix}, \quad r_{n+2} = \begin{pmatrix} \rho \\ -a\frac{\xi}{\|\xi\|} \\ \gamma p \end{pmatrix}, \tag{1.6}$$

thus form a basis of eigenvectors for a given $\xi \in \mathbb{R}^n$. The convective modes correspond to the hyperplane $\omega + \xi \cdot U = 0$, and their group velocity is the flow velocity. The acoustic angular frequencies and wave numbers lie on the oblique cone $(\omega + \xi \cdot U)^2 - a^2 \|\xi\|^2 = 0$. Observe that the acoustic angular frequencies are

$$\omega_{n+1} = -\xi \cdot U - a\|\xi\|, \quad \omega_{n+2} = -\xi \cdot U + a\|\xi\|.$$

We note that the acoustic eigenvalues are simple except for $\xi = 0$.

For the non-reflecting boundary conditions, we consider Eqn. (1.2) as a generalised eigenvalue problem for $\xi_1$. We will denote the tangential parts of vectors by primed variables, so, for instance,

$$x = (x^1, x'), \quad \xi = (\xi_1, \xi'), \quad A = (A^1, A').$$

For simplicity we will assume that the flow is normally subsonic, $M_1 < 1$, and that the boundary $x_1 = 0$ is an outlet of the domain $\{x_1 < 0\}$, i.e., $M_1 > 0$. The case of an inlet can be treated analogously. In primitive variables, the generalised eigenvalue problem for $\xi_1$ thus reads

$$\left(\xi_1 A^1 - (-\omega - \xi' \cdot A')\right) r = 0, \tag{1.7}$$

with $r$ being a generalised right-eigenvector. Assume that a complete basis of eigenvectors

$$r_1(\omega, \xi'), \ldots, r_{n+2}(\omega, \xi'),$$

exists for $(\omega, \xi')$, and that the corresponding eigenvalues $\xi_1$ are, at least locally, smooth functions of $\omega$. Then the boundary conditions can be formulated in terms of the temporal and spatial Fourier coefficients of the flow along the boundary $x_1 = 0$ as follows. Decompose the Fourier coefficients $\hat{q}_{\omega,\xi'}$ into normal modes, i.e.,

$$\hat{q}_{\omega,\xi'} = \sum_{j}^{n+2} \hat{q}^j_{\omega,\xi',\mathrm{mod}} r_j(\omega, \xi').$$

The non-reflecting boundary conditions are satisfied if $\hat{q}^j_{\omega,\xi',\mathrm{mod}} = 0$ for all $j$ such that $r_j$ corresponds to an incoming mode. In case the generalised eigenvalue $\xi_1$ is a non-zero real and $\omega$ is locally a smooth function of $\xi_1$, the sign of the normal group velocity $-\frac{\partial \omega}{\partial \xi_1}$ can be taken as a criterion to classify the modes into incoming and outgoing ones [14].

Note that the generalised eigenvalue problem is indefinite, since the flux Jacobian $\frac{\partial F^1}{\partial q}$ is indefinite, if the flow is normally subsonic. Therefore, one cannot expect that a basis of eigenvectors exists for all $(\omega, \xi')$. In fact, the cone

$$(\omega + \xi \cdot U)^2 - a^2 \|\xi\|^2 = 0 \tag{1.8}$$

may or may not intersect the line defined by some fixed $\omega$ and $\xi'$. In case the corresponding quadratic equation for $\xi_1$ has negative discriminant, it will have two complex conjugate solutions. These so-called *cut-off* modes are viewed as incoming if they decrease exponentially at $-\infty$, i.e., if $\operatorname{Im} \xi_1 < 0$. Accordingly, a cut-off mode is called outgoing if $\operatorname{Im} \xi_1 > 0$.

The corresponding (real or complex) eigenvectors are used to define the spectral projections onto the incoming along the outgoing eigenmodes, $P_{\mathrm{inc}}$. The physical boundary condition can thus be written

$$P_{\mathrm{inc}} \hat{q}_{\omega,\xi'} = 0.$$

where

$$P_{\mathrm{inc}} = P_{\mathrm{inc}}(q, \omega, \xi') = R(q, \omega, \xi')^{-1} \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{Id} \end{pmatrix} R(q, \omega, \xi')$$

where the dimensions of the subblocks of the block matrix are given by the number of right and left runnning modes, respectively. The first columns of the right-eigenvector matrix $R(q, \omega, \xi')$ are the left-running eigenvectors of the generalised eigenvalue problem (1.7), the last columns are given by the left-running eigenvectors.

Special care has to be taken in order to define the acoustic eigenvectors. Here, we write the acoustic normal wave numbers as functions of $q$, $\omega$, and $\xi'$,

$$\xi_{1,\mathrm{a},\pm} = \begin{cases} \pm\frac{\mathrm{i}\|\xi'\|}{\sqrt{1-M_1^2}}, & \text{if } \widetilde{\omega}+\xi'\cdot M' = 0, \\ (M_1 \mp \sqrt{\Delta})\left(\frac{\widetilde{\omega}+\xi'\cdot M'}{1-M_1^2}\right), & \text{if } \widetilde{\omega}+\xi'\cdot M' \neq 0, \quad \Delta \geq 0, \\ (M_1 \pm \mathrm{i}\,\mathrm{sign}(\widetilde{\omega}+\xi'\cdot M')\sqrt{-\Delta})\left(\frac{\widetilde{\omega}+\xi'\cdot M'}{1-M_1^2}\right), & \text{if } \widetilde{\omega}+\xi'\cdot M' \neq 0, \quad \Delta < 0, \end{cases} \tag{1.9}$$

where $\widetilde{\omega} = \omega/a$, $\vec{M} = (M_1, M') = U/a$ and $\Delta$ is the discriminant of the corresponding quadratic equation for $\xi_1$, i.e.,

$$\Delta = 1 - \frac{1-M_1^2}{(\widetilde{\omega}+\xi' M')^2}\|\xi'\|^2.$$

The definition of $\xi_{1,\mathrm{a},\pm}$ ensures that $\operatorname{sign}\operatorname{Im}\xi_{1,\mathrm{a},\pm} = \pm 1$, if $\Delta < 0$. Moreover, it is straightforward to check that

$$\frac{\partial \xi_{1,\mathrm{a},\pm}}{\partial \omega} = \mp\frac{1}{(1-M_1^2)\sqrt{\Delta}}$$

for cut-on modes, i.e., $\Delta > 0$. Therefore, the $x^1$-group velocity

$$v_{\mathrm{g}}^1(\omega, \xi_{1,\mathrm{a},\pm}, \xi') = -\frac{\partial\omega}{\partial\xi_1}\Big|_{(\xi_{1,\mathrm{a},\pm},\xi')} = -\left(\frac{\partial\xi_{1,\mathrm{a},\pm}}{\partial\omega}\Big|_{(\omega,\xi')}\right)^{-1}$$

is positive for $\xi_{1,\mathrm{a},+}$ and negative for $\xi_{1,\mathrm{a},-}$. The acoustic eigenvectors can be defined by

$$r_{\mathrm{a},\pm}(q,\omega,\xi') = \begin{pmatrix} \rho \\ -\frac{a\xi_\pm}{\omega+\xi_\pm\cdot\vec{M}} \\ \gamma p \end{pmatrix} \tag{1.10}$$

where $\xi_\pm = (\xi_{1,\mathrm{a},\pm}, \xi')$.

**The Characteristic Variety**

Let us express the characteristic variety in terms of the (directional) Mach numbers and the normalised frequency,

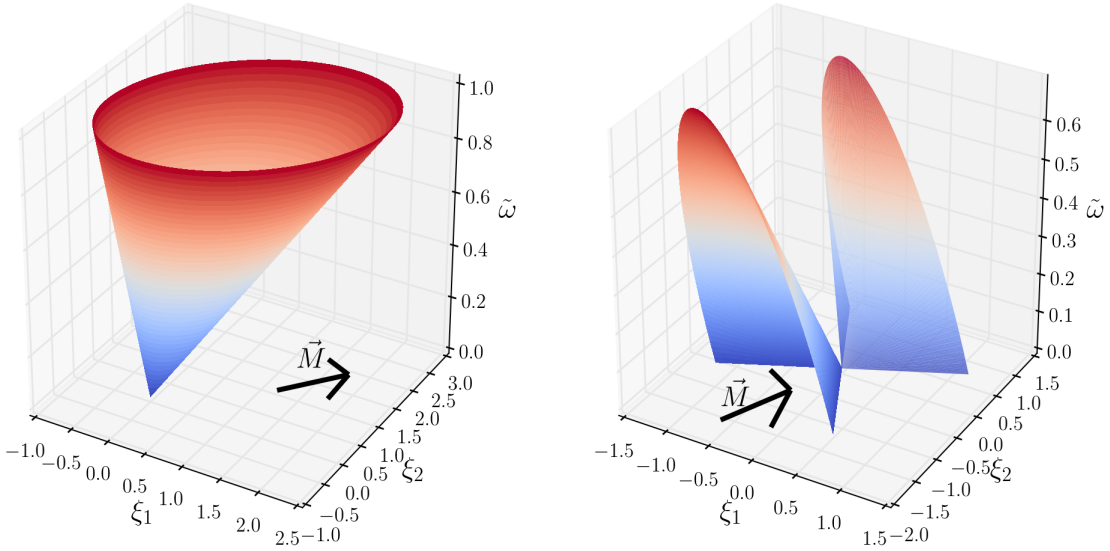
$$\vec{M} = (M_1, \ldots, M_n), \quad M_i = U_i/a, \quad M = \|\vec{M}\|, \quad \widetilde{\omega} = \omega/a.$$

As we have seen above the characteristic variety for the $n$-dimensional linearised Euler equations consists of:

- The plane $\widetilde{\omega} = -\sum_i M_i\xi_i$, corresponding to the convective modes.
- The cone

$$\left(\widetilde{\omega} + \vec{M}\cdot\xi\right)^2 = \|\xi\|^2. \tag{1.11}$$

  corresponding to acoustic modes.




(a) $M = 0.7$, $\alpha = 55°$.

(b) $M = 1.5$, $M_x = 0.5$.

Figure 3: Characteristic varieties for subsonic and supersonic cases.

To analyse the cone (1.11), consider an orthonormal basis $(\widetilde{e}_1, \ldots, \widetilde{e}_n)$ such that

$$\vec{M} = M\widetilde{e}_1.$$

Then, with respect to this basis, Eqn. (1.11) takes the form

$$(1 - M^2)\widetilde{\xi}_1^2 - 2M\widetilde{\omega}\widetilde{\xi}_1 + \sum_{i \neq 1} \widetilde{\xi}_i^2 - \widetilde{\omega}^2 = 0,$$

and therefore,

$$(1 - M^2)^2 \left(\widetilde{\xi}_1 - \frac{\widetilde{\omega} M}{1 - M^2}\right)^2 + (1 - M^2)\|\widetilde{\xi}'\|^2 = \widetilde{\omega}^2.$$

This shows that for $M < 1$ the acoustic part of the characteristic variety is an oblique cone as depicted in Fig. 3a. Both the inclination and the apex angle increase with the Mach number. When the flow is supersonic, $M > 1$, it consists of segments of a double cone, see Fig. 3b. In the subsonic case, for a given $\omega$, there is a bounded set of values for $\xi'$ such that the line defined by $(\omega, \xi')$ intersects the characteristic variety. In two-dimensions, for instance, the acoustic modes are cut-on if

$$\xi_2 \in \left[\tfrac{\widetilde{\omega}}{1-M^2}\left(M_2 - \sqrt{1 - M_1^2}\right), \tfrac{\widetilde{\omega}}{1-M^2}\left(M_2 + \sqrt{1 - M_1^2}\right)\right]. \tag{1.12}$$

In the supersonic case there will be a non-empty intersection outside a bounded set. In two dimensions the cut-on condition reads

$$\xi_2 \in \left(-\infty, \tfrac{\widetilde{\omega}}{1-M^2}\left(M_2 - \sqrt{1 - M_1^2}\right)\right] \cup \left[\tfrac{\widetilde{\omega}}{1-M^2}\left(M_2 + \sqrt{1 - M_1^2}\right), +\infty\right).$$

### Turbomachinery Boundary Conditions

In the case of a (rotational) turbomachinery configuration, one usually applies the above theory for $n = 2$ to cylinders around the engine axis. More precisely, the coordinate change

$$y = r\vartheta, \quad z = r$$

transforms the Euler flow in an annular duct with constant "hub" and "casing" radii to the three-dimensional flow in

$$\{(x, y, z) \in \mathbb{R}^3 \,|\, r_1 < z < r_2\}$$

as long as the $r$-components of the flow velocity can be neglected. Moreover, in the case of $N$ blades and a certain interblade phase angle $\sigma$, the resulting flow at radius $r$ will have the symmetry

$$q(t, x, y + 2\pi r/N, z) = q(t + \sigma/\omega, x, y, z).$$

The $y$-components of the wave numbers for which (1.8) must be considered, thus satisfy

$$e^{\mathrm{i}\xi_2(y+2\pi r/N)} = e^{\mathrm{i}\sigma} e^{\mathrm{i}\xi_2 y}.$$

Writing $m$ for the circumferential wave number, we thus have

$$\hat{q}_\omega(x, r, \vartheta) = \sum_{m=N_d \bmod N} \hat{q}_{\omega,m}(x, r) e^{\mathrm{i}m\vartheta}.$$

The cut-on condition for a subsonic flow at radius $r$ is

$$m \in \left[\tfrac{r\widetilde{\omega}}{1-M^2}\left(M_\vartheta - \sqrt{1 - M_x^2}\right), \tfrac{r\widetilde{\omega}}{1-M^2}\left(M_\vartheta + \sqrt{1 - M_x^2}\right)\right].$$

In the cell-centred, finite volume, time-linearised CFD solver discussed here, the boundary conditions are applied at bands of faces whose centre has constant radius. The modal amplitudes $q^j_{\omega,\xi',\mathrm{mod}}$ are computed using the inner cell values along the bands and extrapolated to the bands using the axial wave number $\xi$ for outgoing modes. The modal amplitudes for incoming modes are set to zero. The inverse Fourier transform yields flow states for each face which are then used to define flow states in the ghost cells by extrapolation.

## 2 ACOUSTIC RESONANCE AND REGULARISATION

The modal decomposition in the previous section is restricted to the case of non-zero group velocity $v^1_{\mathrm{g}}$ for all cut-on modes, i.e., we have assumed that the discrimimant $\Delta$ is non-zero. If

$$(1 - M_1^2)\|\xi'\|^2 = (\widetilde{\omega} + \xi' \cdot M')^2,$$

then the quadratic equation (1.8) has a real root of algebraic multiplicity 2. However, unless $\xi = 0$, the angular frequency $\omega$ is a simple root at the same point. Therefore the eigenspace, i.e., the solution space of (1.2) is one-dimensional.

We regularise the boundary conditions by adding a zeroth order dissipation term to the left-hand side of the model equations. Rather than the linearisation of Eqn. (1.1) we consider

$$\frac{\partial(\delta q)}{\partial t} + \mathrm{div}\left[\left.\frac{\partial F}{\partial q}\right|_q \delta q\right] + \varepsilon\delta q = 0, \tag{2.1}$$

where $\varepsilon > 0$ is a small regularisation parameter. The corresponding equation in dual variables thus reads

$$\left(\omega - \mathrm{i}\varepsilon + \sum_k \xi_k A^k\right) \hat{q}_{\omega,\xi} = 0. \tag{2.2}$$

If $\omega$ is replaced with $\omega - \mathrm{i}\varepsilon$, the formulas of Section 1 essentially remain valid. The crucial difference, however, is that there are no longer cut-on modes. More presicely, given real $\omega$ and $\xi'$, Eqn. (2.2) will have only trivial solutions for real $\xi_1$ since, otherwise, $\omega - \mathrm{i}\varepsilon$ would be a non-real eigenvalue of a symmetric operator. Since all eigenvalues $\xi_1$ have non-zero imaginary part, all eigenspaces correspond to either incoming or outgoing modes.

If $\xi$ is non-zero and $\xi_{1,\mathrm{a},\pm}(q,\omega,\xi')$ corresponds to a cut-on acoustic mode of the original system, then $\omega$ is locally a smooth holomorphic function of $\xi$. Moreover, if locally $\frac{\partial\omega}{\partial\xi_1}$ is non-zero, then the generalised eigenvalue $\xi_1$ is, locally, a holomorphic function of $\omega$. Using the Cauchy-Riemann equations, we infer

$$\begin{aligned}
\operatorname{sign} \operatorname{Im} \xi_{1,\mathrm{a},\pm}(q,\omega - \mathrm{i}\varepsilon,\xi') &= -\operatorname{sign} \frac{\partial \operatorname{Im} \xi_{1,\mathrm{a},\pm}}{\partial \operatorname{Im}\omega}\Big|_{(q,\omega,\xi')} \\
&= -\operatorname{sign} \frac{\partial \operatorname{Re} \xi_{1,\mathrm{a},\pm}}{\partial \operatorname{Re}\omega}\Big|_{(q,\omega,\xi')} \\
&= -\operatorname{sign} \frac{\partial \omega}{\partial \xi_{1,\mathrm{a},\pm}}\Big|_{(\xi_{1,\mathrm{a},\pm}(q,\omega,\xi'),\xi')} \\
&= \operatorname{sign} v_{\mathrm{g}}^1.
\end{aligned} \tag{2.3}$$

Therefore, defining the acoustic normal wavenumbers by the analytic continuation of the formulas given in Eqn. (1.9), we preserve the propagation direction in the sense that, e.g, right-running cut-on modes become right-running cut-off modes. Denoting by $\sqrt{z}$ the principal branch of the complex square root, it is now easy to verify that the analytic continuation of (1.9) is

$$\xi_{1,\mathrm{a},\pm} = (M_1 \mp \sqrt{\Delta}) \left(\frac{\widetilde{\omega} - \mathrm{i}\widetilde{\varepsilon} + \xi' \cdot M'}{1 - M_1^2}\right),$$

where

$$\widetilde{\varepsilon} = \frac{\varepsilon}{a}, \quad \Delta = 1 - \frac{(1 - M_1^2)\|\xi'\|^2}{(\widetilde{\omega} - \mathrm{i}\widetilde{\varepsilon} + \xi' \cdot M')^2}.$$

The definition of the right-eigenvectors carries over to the modified model problem. In particular, the acoustic eigenvectors are replaced with $r_{\mathrm{a},\pm}(q,\omega - i\varepsilon,\xi')$, cf. Eqn. (1.10). We emphasize, that for $\varepsilon > 0$, the above terms are holomorphic for all $\omega, \xi'$, including the acoustic resonance points.

## 3 ANALYSIS OF TURBOMACHINERY FLUTTER

To illustrate the modification of the boundary condition we will use the so-called tenth standard configuration (SC10) [2]. This configuration is a two-dimensional compressor test case consisting of a cambered NACA 0006 cascade at subsonic and transonic flow conditions. The eigenmode for which flutter stability is analysed corresponds to a rigid body rotation about the blade center. Flow conditions and reduced eigenfrequencies are summarised in Table 1. Observe that the reduced frequencies $k$ are computed using the full chord $c$ and the flow speed at the inlet $U_{\mathrm{inlet}}$, i.e.,

$$\omega_{\mathrm{red}} = \frac{\omega c}{\|U_{\mathrm{inlet}}\|}.$$

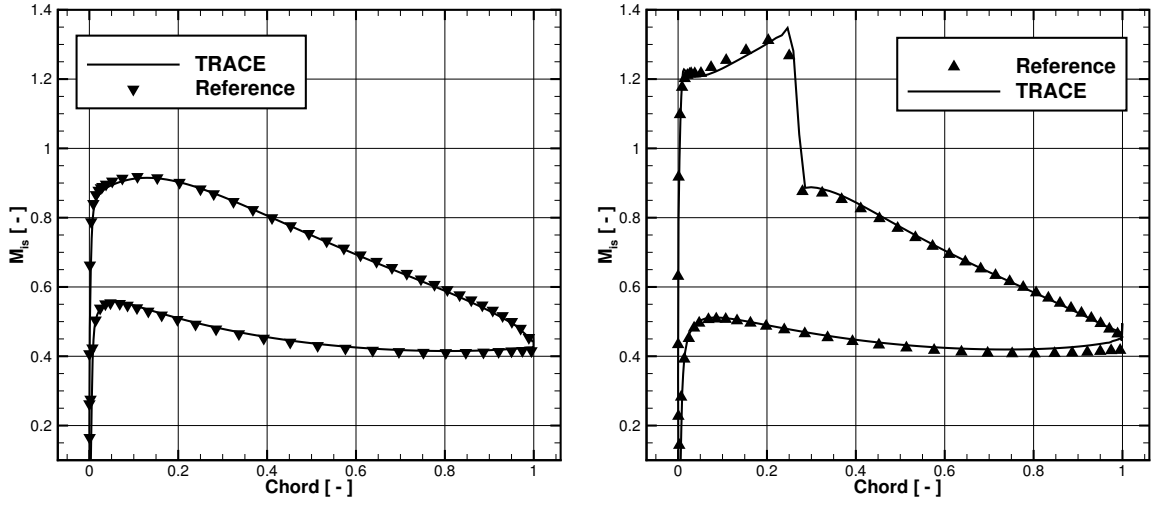
| Testcase | SC10 subsonic | SC10 transonic |
|---|---|---|
| Inflow Mach number | 0.7 | 0.8 |
| Inflow angle | $55°$ | $58°$ |
| Outflow Mach number | 0.446 | 0.443 |
| Outflow angle | $40.1°$ | $41.1°$ |
| Reduced frequency | 1 | 0.5 |
| Upstream acoustic resonances | $-26.9°$ | $-14.5°$ |
| | $117.1°$ | $100.8°$ |
| Downstream acoustic resonances | $-31.8°$ | $-17.8°$ |
| | $59.8°$ | $33.7°$ |

Table 1: Flow conditions, frequencies and acoustic resonances.

The steady isentropic Mach number distributions on the blade are shown in Fig. 4. Both the subsonic and the transonic background mean flow conditions, used for the linearised flutter simulations below, agree well with reference results obtained with a different flow solver, [16].

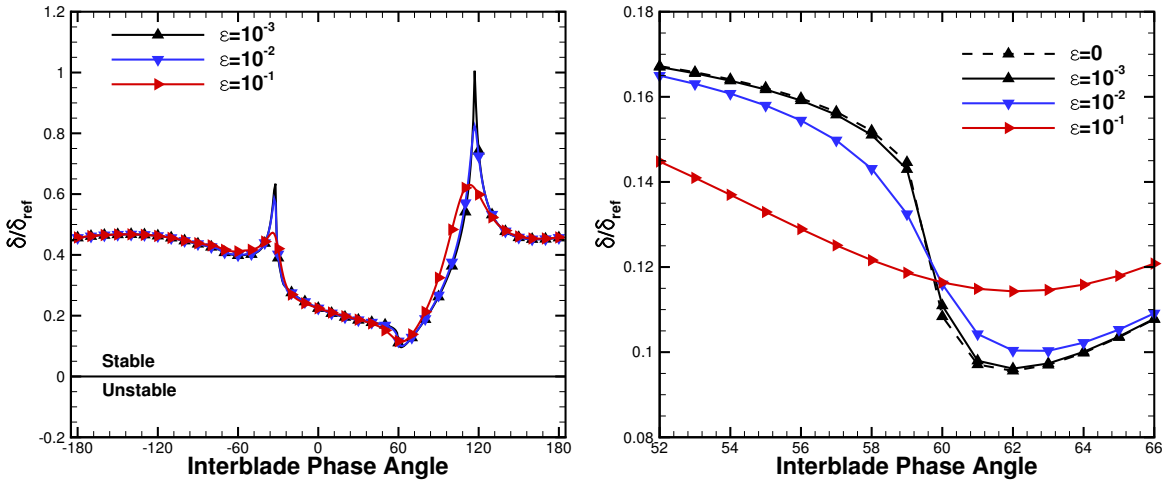
The flutter predictions using different values for the regularisation parameter $\varepsilon$ are shown in Figs. 5 and 6. The value of $\varepsilon$ is non-dimensionalised using the sound speed at the inlet and the chord length and thus equals, up to the inlet Mach number, the negative imaginary part of the modified reduced frequency. The results for $\varepsilon = 10^{-3}$ have been validated against reference results in previous publications by the authors [4, 17] and show good agreement with results obtained with a different solver [15, 16].

However, as $\varepsilon$ is increased, the damping curve becomes "smoother". This can be seen, for both the subsonic as the transonic configuration, near the interblade phase angles with minimal damping, which are close to the downstream acoustic resonance points $59.8°$ and $33.7°$. The zooms also show that the choice of $\varepsilon$ above $10^{-1}$ can have a significant impact on the prediction of the flutter stability and is therefore discouraged.




(a) Subsonic Case (b) Transonic Case

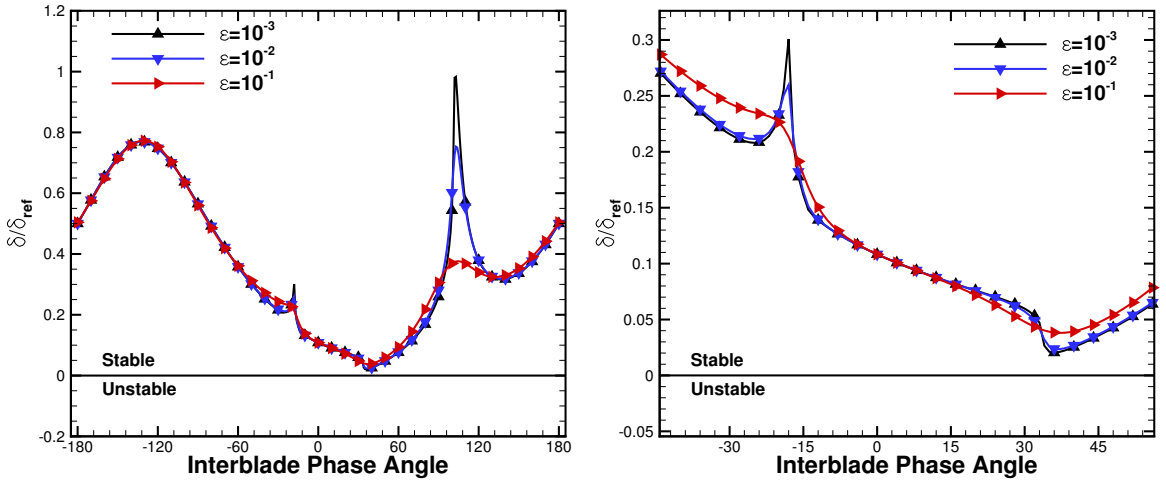
Figure 4: Comparison of steady solution with references given in [15]




(a) Normalised aerodynamic damping over interblade phase angle.

(b) Minimal damping region

Figure 5: Damping results for the subsonic case.




(a) Normalised aerodynamic damping over interblade phase angle.

(b) Minimal damping region

Figure 6: Damping results for the transonic case.

## 4 INTERPRETATION OF THE REGULARISATION

Let us study the behaviour of the regularised spectral projection $P_{\text{inc}}$ as $\varepsilon$ tends to infinity. Denoting by $\chi_\Lambda$ the characteristic function of a complex set $\Lambda$, we can write the regularised projection onto the incoming modes as

$$P_{\text{inc}} = \chi_{\{\operatorname{Im}\xi_1<0\}}\left(-(A^1)^{-1}\left(\omega - \mathrm{i}\varepsilon + \xi' A'\right)\right)$$

Observe that

$$\| -(A^1)^{-1}\left(\omega - \mathrm{i}\varepsilon + \xi' A'\right) \| \leq \varepsilon \|A^{-1}\| + \text{const}\,.$$

Therefore, for an appropriate $R > 0$ and sufficiently large $\varepsilon$, the contour $\gamma_\varepsilon$, depicted in Fig. 7, encircles all eigenvalues $\xi_1$ with negative imaginary part anticlockwise. Using holomorphic

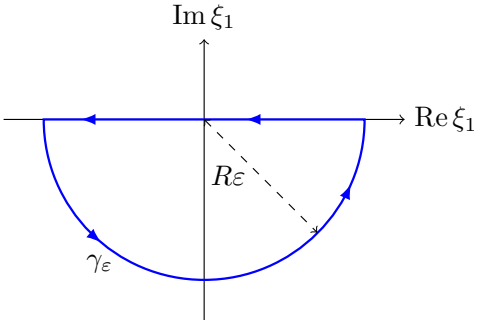



Figure 7: Closed contour encircling the acoustic normal wave numbers of the regularised system anticlockwise.

.

functional calculus, we have

$$\begin{aligned} P_{\text{inc}} &= \frac{1}{2\pi \mathrm{i}} \int_{\gamma_\varepsilon} \left(\xi_1 + (A^1)^{-1}(\omega - \mathrm{i}\varepsilon + \xi' A')\right)^{-1} d\xi_1 \\ &= \frac{1}{2\pi \mathrm{i}} \int_{\gamma_1} \left(z + (A^1)^{-1}(-\mathrm{i} + \varepsilon^{-1}(\omega + \xi' A'))\right)^{-1} dz, \end{aligned} \tag{4.1}$$

where we have made the substitution $\xi_1 = \varepsilon z$. Since

$$-(A^1)^{-1}(-\mathrm{i} + \varepsilon^{-1}(\omega + \xi' A'))$$

has no real eigenvalues and converges uniformly to $\mathrm{i}(A^1)^{-1}$ along $\gamma_1$ for $\varepsilon \to \infty$, it follows that $P_{\text{inc}}$ converges to the spectral projection

$$\chi_{\{\operatorname{Im} z<0\}}(\mathrm{i}(A^1)^{-1}) = \chi_{\{\operatorname{Re} z<0\}}(A^1).$$

Therefore, in the limit $\varepsilon \to +\infty$, the regularised boundary conditions converge to characteristic boundary conditions. This is confirmed by numerical experiments, see Fig. 8. As the parameter $\varepsilon$ is increased, the results for the aerodynamic damping approach the curve which is obtained by using one-dimensional, characteristics based boundary conditions. Note that for this numerical experiment, the modified normal wave number has been replaced by $0$ in the extrapolation of the outgoing mode amplitudes to the ghost cells.

## 5 DISCUSSION

So far, we have studied only a two-dimensional test case and two-dimensional boundary conditions which raises the question of the relevance for real three-dimensional turbomachinery configurations. Therefore, consider a computational domain representative of an annular duct in a real engine. Although two-dimensional (or even characteristic one-dimensional) non-reflecting boundary conditions are still predominantly used for turbomachinery simulations, three-dimensional non-reflecting boundary conditions have been developed by several authors [18, 19]. For a general mean flow distribution, the eigenmodes can no longer been computed explicitly. Rather, one performs a numerical spectral decomposition using e.g. appropriate LAPACK routines [20, 11]. In three dimensions, the equivalent of Eqn. (1.2), can be written in the form

$$(\mathrm{i}(\omega + \xi A^x + mC) + B\frac{\partial}{\partial r} + D)\hat{q}_{\omega,\xi,m}(r) = 0, \tag{5.1}$$

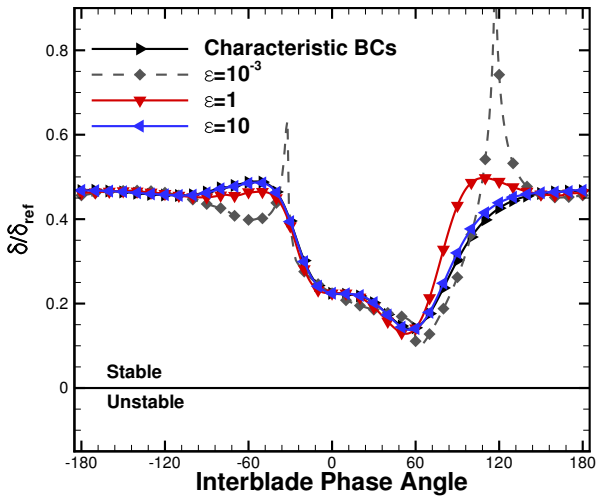



Figure 8: Normalised aerodynamic damping over interblade phase angle for subsonic case and large values for $\varepsilon$.

with suitable boundary conditions at the hub and tip radii, $r_{\min}$, $r_{\max}$, [20]. In case the mean flow is purely axial, the pressure harmonics satisfy

$$\left(-\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} + \xi_x^2 + \frac{m^2}{r^2} - (\widetilde{\omega} + \xi_x M_x)^2\right) \widehat{p}_{\omega,\xi,m}(r) = 0, \tag{5.2}$$

together with Neumann boundary conditions at $r_{\min}$ and $r_{\max}$. Eqn. (5.2) may thus be written in the form

$$\left(L(m) + \xi_x^2 - (\widetilde{\omega} + \xi_x M_x)^2\right) \widehat{p}_{\omega,\xi,m}(r) = 0$$

where $L(m)$ is an elliptic self-adjoint eigenvalue problem (with respect to the Riemannian metric $g(r) = r^2$) depending on $m$. From its discrete eigenvalues $(\lambda_n(m))_{n\geq 0}$, we can thus determine the relation between the axial wavenumber and $\widetilde{\omega}$ for all radial and circumferential mode orders, $(n, m)$:

$$\lambda_n(m) + \xi_x^2 + (\widetilde{\omega} + \xi_x \cdot M_x)^2 = 0. \tag{5.3}$$

Viewing $m$ as a continuous parameter, we can thus plot the generalisation of the two-dimensional characteristic varieties for each radial mode order $n$. Figure 9 shows the surface defined by (5.3) for the radial mode orders $n = 0, 1, 2$. Here, $M_x = 0.5$ and $r_{\min} = 1$, $r_{\max} = 2$. The resulting surfaces resemble the upper sheets of an oblique two-surface hyperboloid except for $n = 0$ in which case the surface is a slightly deformed oblique cone. In many cases the three-dimensional characteristic variety for $n = 0$ is very close to the two-dimensional one. For the above parameters the two-dimensional characteristic variety at $r = \frac{1}{2}(r_{\min} + r_{\max})$ is nearly identical to the three-dimensional one for $n = 0$, see Fig. 10.

If acoustic resonance occurs for certain values of $(\omega, \xi_x, m, n)$, then $\xi_x$ is a double zero of Eqn. (5.3), whereas $\omega$ is a simple one. Hence, similar arguments as in the two-dimensional case show that the geometric multiplicity of the generalised eigenvalue problem for $\xi_x$ is 1. This shows that the modal decomposition has to be regularised in the three-dimensional case as well. In their implementations of three-dimensional non-reflecting boundary conditions [19], the authors perform a numerical eigenvalue analyis of (5.1) where $\omega$ is replaced with $\omega - \mathrm{i}\varepsilon$ and $\varepsilon$ is typically set to $10^{-3} a_{\mathrm{ref}}/l_{\mathrm{ref}}$. Here $a_{\mathrm{ref}}$, $l_{\mathrm{ref}}$ denote reference sound speed and length,

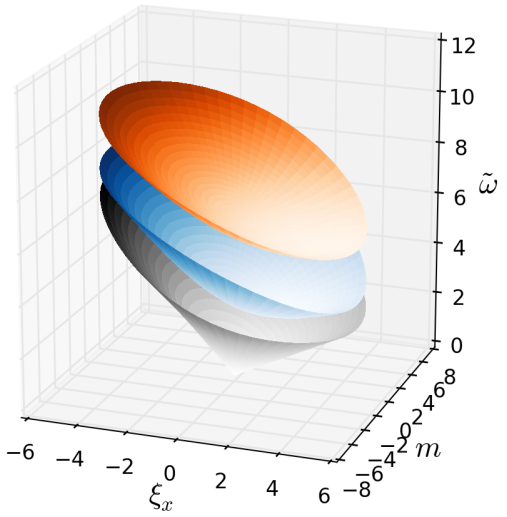



Figure 9: Three-dimensional characteristic variety for radial mode orders $0$ (grey), $1$ (blue), and $2$ (orange).

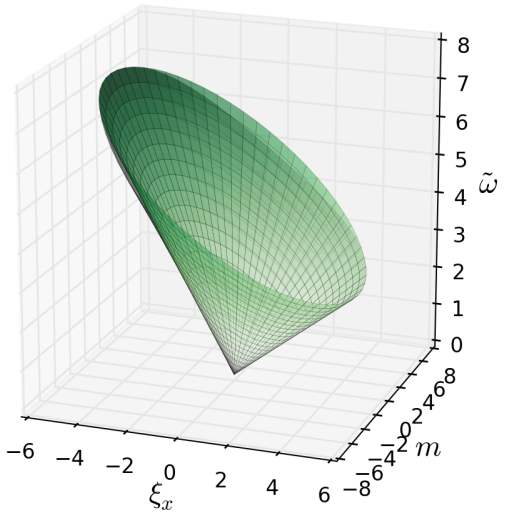


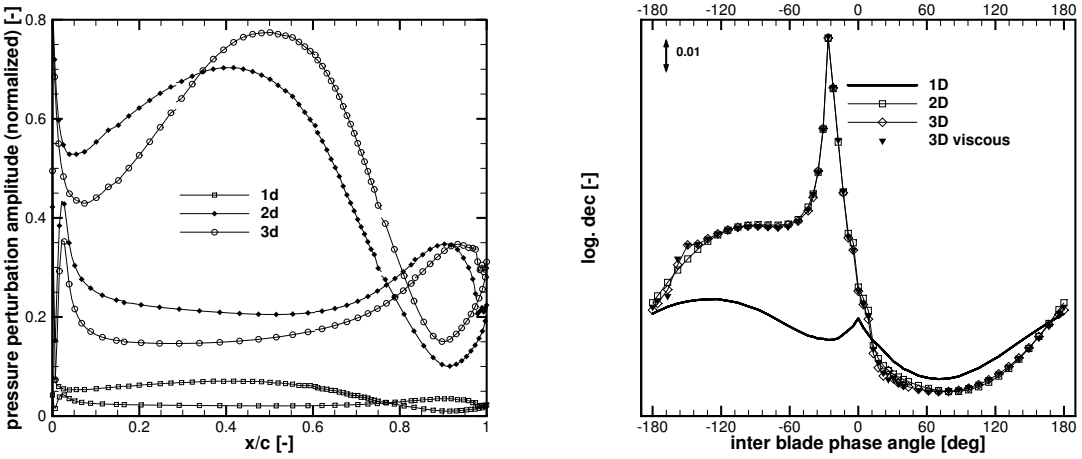

Figure 10: Three-dimensional Characteristic variety for radial mode orders $0$ (grey) compared to the two-dimensional one at $r = \frac{1}{2}(r_{\min} + r_{\max})$ (green).

respectively. The regularisation allows to circumvent the numerical computation of the group velocity, as pointed out in [11]. One can use the sign of $\operatorname{Im} \xi_x$ to determine the propagation direction of the modes.

Acoustic resonance occurs when one of the characteristic varieties has vanishing slope in the $\xi_x$-direction. For a given frequency, this can happen at several radial orders, although in practice, for only a few low values of $n$ there exist cut-on modes, i.e., the plane $\omega = \text{const}$ intersects only a few of the hyperboloids in Fig. 9. Therefore, one can expect that the use

of two-dimensional boundary conditions for a three-dimensional configuration might suppress acoustic resonance for some non-zero radial mode orders but will show qualitatively similar results as far as the acoustic resonance for $n = 0$ is concerned.

This is illustrated by the flutter analysis of a turbine blade carried out in [19]. The test case represents a low pressure turbine blade of a modern aeroengine. As can be seen in Figure 11, the use of three-dimensional instead of two-dimensional non-reflecting boundary conditions has a significant impact on the unsteady pressure distribution on the blade. However, the aerodynamic damping is very insensitive to the choice of boundary conditions, in particular near the resonance point located between $-30°$ and $-20°$.




(a) Pressure amplitude over chord for $\sigma = -88°$.

(b) Damping curve.

Figure 11: Flutter analysis of a turbine blade with 1D, 2D, and 3D non-reflecting boundary conditions. From [19].

## CONCLUSIONS

In this paper, a regularisation of the non-reflecting boundary conditions for unsteady turbomachinery flows has been investigated. The regularisation is necessary in order to ensure that the spectral projections onto outgoing along incoming modes remain bounded. The idea is to apply the usual theory of non-reflecting boundary conditions to the Euler equations modified by a small dissipation term which is controlled by some parameter $\varepsilon$. As the additional dissipation increases the resulting flutter curves become smoother. A choice of $\varepsilon = 10^{-3}$ with respect to the reference frequency has been shown to be a reasonable choice for the turbomachinery test cases presented in this paper. As $\varepsilon$ grows, the boundary conditions approach the one-dimensional characteristics based boundary conditions.

The issue of acoustic resonance occurs for both two-dimensional and three-dimensional non-reflection boundary conditions. It is suspected that using two-dimensional boundary conditions for a three-dimensional turbomachinery configuration might smooth out singularities in the damping curve which are related to acoustic resonance modes of non-zero radial order.

## ACKNOWLEDGEMENTS

This study was conducted as part of the joint research programme *AG-Turbo 2020*. The authors gratefully acknowledge financial supported from the German Federal Ministry for Economic Affairs and Energy under grant number 03ET2012G.

## NOMENCLATURE

| | |
|---|---|
| $a$ | speed of sound |
| $A^i$ | flux Jacobian in the $x^i$-direction |
| $A^1, A'$ | normal and tangential flux Jacobians |
| $e_\mathrm{t}$ | total (specific) energy |
| $f$ | frequency |
| $h_\mathrm{t}$ | total (specific) enthalpy |
| $k$ | modal stiffness |
| $m$ | modal mass, circumferential wave number |
| $\vec{M} = U/a$ | Mach vector |
| $M^1, M'$ | normal Mach number and tangential Mach vector |
| $N$ | number of blades |
| $N_\mathrm{d}$ | nodal diameter |
| $\vec{n}$ | unit normal vector (pointing out of the flow domain) |
| $p$ | pressure |
| $q$ | flow state |
| $q^\mathrm{s}$ | modal displacement |
| $s$ | specific entropy |
| $(x, r, \vartheta)$ | cylindrical coordinates |
| $x^1, x' = (x^2, \ldots, x^n)$ | normal and tangential coordinates |
| $U$ | flow velocity |
| $W_\mathrm{cyc}$ | aerodynamic work per cycle |
| $\delta$ | logarithmic decrement of aerodynamic damping |
| $\gamma$ | specific heat ratio |
| $\kappa$ | aerodynamic stiffness coefficient |
| | structural eigenmode (mode shape) |
| $\rho$ | density |
| $\sigma$ | interblade phase angle |
| $\chi_\Lambda$ | characteristic function of a set $\Lambda \subset \mathbb{C}$ |
| $\omega$ | angular frequency ($= 2\pi f$) |

## REFERENCES

[1] R. Kielb, "CFD for turbomachinery unsteady flows - an aeroelastic design perspective," in *Aerospace Sciences Meetings*, American Institute of Aeronautics and Astronautics, 2001.

[2] T. H. Fransson and J. M. Verdon, "Updated report on Standard Configurations for the Determination of unsteady Flow Through Vibrating Axial-flow Turbomachine-Cascades," Tech. Rep. TRITA/KRV/92.009, KTH, Stockholm, 1992.

[3] K. C. Hall and W. S. Clark, *Unsteady Aerodynamics, Aeroacoustics, and Aeroelasticity of Turbomachines and Propellers*, ch. Calculation of Unsteady Linearized Euler Flows in

Cascades Using Harmonically Deforming Grids, pp. 195–212. New York, NY: Springer New York, 1993.

[4] H.-P. Kersken, C. Frey, C. Voigt, and G. Ashcroft, "Time-Linearized and Time-Accurate 3D RANS Methods for Aeroelastic Analysis in Turbomachinery," *J. Turbomach.*, vol. 134, no. 5, pp. 051024–051024, 2012.

[5] G. Kahl, *Aeroelastic effects of mistuning and coupling in turbomachinery bladings*. PhD thesis, École polytechnique fédérale de Lausanne, 2002.

[6] J. J. Waite and R. E. Kielb, "Physical understanding and sensitivities of low pressure turbine flutter," *Journal of Engineering for Gas Turbines and Power*, vol. 137, pp. 012502–012502, Aug. 2014.

[7] T. R. Camp, "A study of acoustic resonance in a low-speed multistage compressor," *Journal of Turbomachinery*, vol. 121, pp. 36–43, Jan. 1999.

[8] B. Hellmich and J. R. Seume, "Causes of acoustic resonance in a high-speed axial compressor," *Journal of Turbomachinery*, vol. 130, pp. 031003–031003, May 2008.

[9] B. Engquist and A. Majda, "Absorbing boundary conditions for the numerical simulation of waves," *Math. Comp.*, vol. 31, pp. 629–651, May 1977.

[10] M. B. Giles, "Non-reflecting boundary conditions for the Euler equations," tech. rep., MIT Dept. of Aero. and Astr., 1988. CFDL Report 88-1.

[11] P. Moinier and M. Giles, "Eigenmode analysis for turbomachinery applications," *Journal of Propulsion and Power*, vol. 21, no. 6, pp. 973–978, 2005.

[12] K. Friedrichs and P. Lax, "Systems of conservation equations with a convex extension," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 68, no. 8, 1971.

[13] M. Taylor, *Partial Differential Equations I: Basic Theory*. Applied Functional Analysis: Applications to Mathematical Physics, Springer, 1996.

[14] R. L. Higdon, "Initial-boundary value problems for linear hyperbolic systems," *SIAM Rev.*, vol. 28, pp. 177–217, June 1986.

[15] P. J. Petrie-Repar, A. M. McGhee, and P. A. Jacobs, "Three-dimensional viscous flutter analysis of standard configuration 10," in *Proceedings of the ASME Turbo Expo 2007*, 2007.

[16] RPMTurbo Pty. Ltd., "RPMTurbo Standard Configuration 10." http://rpmturbo.com/testcases/sc10/index.html, 2005-2014.

[17] G. Ashcroft, C. Frey, and H.-P. Kersken, "On the development of a harmonic balance method for aeroelastic analysis," in *6th European Conference on Computational Fluid Dynamics (ECFD VI)*, 2014.

[18] P. Moinier, M. B. Giles, and J. Coupland, "Three-dimensional nonreflecting boundary conditions for swirling flow in turbomachinery," *Journal of Propulsion and Power*, vol. 23, no. 5, pp. 981–986, 2007.

[19] H.-P. Kersken, G. Ashcroft, C. Frey, N. Wolfrum, and D. Korte, "Nonreflecting boundary conditions for aeroelastic analysis in time and frequency domain 3D RANS solvers," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[20] V. Golubev and H. Atassi, "Sound propagation in an annular duct with mean potential swirling flow," *Journal of Sound and Vibration*, vol. 198, no. 5, pp. 601–616, 1996.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON GRID RESOLUTION REQUIREMENTS FOR LES OF WALL-BOUNDED FLOWS

**S. Rezaeiravesh**[1], **M. Liefvendahl**[1,2] **and C. Fureby**[2]

[1] Division of Scientific Computing, Uppsala University, Sweden.
e-mail: saleh.rezaeiravesh@it.uu.se

[2] Swedish Defence Research Agency (FOI), Sweden.
e-mail: mattias.liefvendahl@foi.se, christer.fureby@foi.se

**Keywords:** Large-eddy simulation, Near-wall modelling, Grid point estimates, Turbulent boundary layers

**Abstract.** *Grid requirements for LES of wall-bounded flows are considered. The setting is a zero pressure gradient turbulent boundary layer on a flat plate, but the results are intended to be of use generally for the simulation of flows with an important influence of turbulent boundary layers. The basis for the grid estimates are expressions for the thickness and the viscous length scale of a turbulent boundary layer. The literature is reviewed, and a new power law is proposed, the coefficients of which have been determined using recent high-Re experimental data. An estimation for the number of grid points required for NWM-LES is derived, which is more general than previously published such estimates. A complete simulation methodology, including a numerical tripping device for transition to turbulence in the boundary layer, is demonstrated for NWM-LES of a flat plate turbulent boundary layer. The predictive accuracy is assessed by comparison with DNS data.*

## 1 INTRODUCTION

The number of application areas in which large-eddy simulation (LES) of turbulent flows is employed is currently growing rapidly, a development which is facilitated by the level of computational resources available today. The main alternative to LES for turbulent flows is Reynolds Averaged Navier-Stokes (RANS) modelling, which generally requires less computational resources since only the mean flow and turbulent quantities are computed. For free shear flows, there are compelling arguments for the use of LES. The main reason is that the energy cascade makes it possible to resolve a large part of the turbulent kinetic energy with a grid which is relatively coarse as compared to the viscous (Kolmogorov) length scale. Thus turbulent momentum transport, largely determined by the resolved scales, can be expected to be modelled more accurately by LES than RANS. Furthermore, since the large scale flow structures are directly simulated, LES provides significantly more information about the flow, as compared to RANS.

The present paper considers wall-bounded flows which put very different requirements on the simulation, as compared to free shear flows. The main reason for this is that the structure of turbulent boundary layers (TBLs) is quite different, as compared to turbulence away from walls. For instance, the peak in the Reynolds stress component $\langle u'u' \rangle$ occurs at $y^+ \approx 12$ in a zero pressure gradient (ZPG) TBL, e.g. [13]. Here $u'$ are the fluctuations of the streamwise velocity, $\langle \cdot \rangle$ denotes averaging, and $y^+ = y/\delta_\nu$, where $\delta_\nu$ is the viscous length scale. This fact has two implications, which are closely related, for the application of LES: *i)* energetic flow structures are present in the TBL on a length scale which is not very much larger than the viscous length scale, *ii)* with increasing Reynolds (Re-)number, the length scale of these flow structures decreases. Both of these items are in contrast to the free shear case, and make the application of LES to wall-bounded turbulence more difficult.

Three different broad approaches to LES of wall-bounded turbulence have generally been considered. *i)* Hybrid RANS-LES methods, where the TBL is modelled using RANS and the flow away from wall is modelled using LES. This means that the turbulence in the TBL is handled by the turbulence model, and that there are no resolved fluctuations in the TBL in the simulation. *ii)* Near-wall modelled (NWM-)LES in which the computational grid is constructed to resolve fluctuations the size of fractions of the TBL-thickness $\delta$. The grid is, however, not adapted to the viscous length scale $\delta_\nu$, and the fluctuations in the innermost part of the TBL are handled by a near-wall model. This approach is in focus of the present paper. *iii)* Near-wall resolved (NWR-)LES in which the computational grid is constructed and adapted to resolve fluctuations the size of the viscous length scale.

For these three approaches it is clear that, from *i)* to *iii)*, more of the flow is directly resolved and less is handled by turbulence modelling. Associated with this is the rising computational cost caused by the increasing requirements on grid resolution. It is thus of utmost importance to precisely formulate the grid requirements and to have a good understanding of how they depend on the parameters of the flow problem, in particular the Re-number. It is the aim of the present paper to make a contribution to this, in particular for the case of NWM-LES.

Chapman derived grid requirements for RANS, NWM-LES and NWR-LES in a pioneering paper, [1]. However, as was pointed out by Choi and Moin [2] (see also section 3 of the present paper) the NWM-LES estimate is too optimistic, in that it leads to an underestimation of the required grid resolution. The reason for this is that in the derivation Chapman employed the mean (over the wall surface) thickness of the TBL. A more correct approach, as suggested by Spalart et al. [20], is to use the local TBL-thickness, and to obtain the grid estimate using

integration, as was done in [2], and also in section 3 of the present paper. A note on terminology: often these requirements are described as estimates on the number of grid points, or grid point estimates. The basis for the argument is however, both here and in [1, 2] is the grid length scale. Thus the derivations are equally applicable to grid points, grid elements (in the finite element framework), and grid cells (in the finite volume framework).

The paper is structured as follows. The basis for the grid estimates are expressions for the thickness $\delta$ and the viscous length scale $\delta_\nu$ of the TBL. In section 2, the literature is reviewed, and new coefficient values are proposed for a power law form of the increase of the length scales with the downstream coordinate. The basis for the new coefficient values is a data fit to recent high-Re experimental data. Then, in section 3, a grid estimate is derived for NWM-LES, using the same approach as Choi and Moin [2]. The derivation is however more general in that it holds for any power law for $\delta$, and the improved coefficients are then inserted in the final expressions. Furthermore, the form of the estimate is new, and it highlights the clustering of the grid in the initial length of the flat plate, for $\mathrm{Re}_x < 10^6$, approximately. In section 4, a complete simulation methodology is demonstrated for NWM-LES of a flat plate ZPG-TBL, with a length-based Re-number of $2.42 \cdot 10^6$, and a computational grid of $12 \cdot 10^6$ finite volume cells. This includes a numerical tripping device to induce resolved fluctuations in the boundary layer. The predictive accuracy of the simulation is assessed by comparison with data from DNS (direct numerical simulation). Throughout the paper, the setting is that of a ZPG-TBL on a flat plate. The grid estimates and the simulation methodology are however intended for application to a wide range of flow problems where TBLs have a significant role. This includes a large number of important applications in the marine, aeronautical and automotive engineering as well as other areas of research.

## 2 CORRELATIONS FOR TBL INTEGRAL QUANTITIES

For the purpose of grid estimation for NWM- and NWR-LES, it is required to express the boundary layer thickness $\delta$ and the viscous length scale $\delta_\nu = \delta/\mathrm{Re}_\tau = \nu/u_\tau$ in terms of the streamwise distance $x$ measured from the leading edge of the plate or, equivalently, in terms of $\mathrm{Re}_x = U_0 x/\nu$, in which $U_0$ denotes the free-stream velocity. These quantities are interconnected through the definition of the friction coefficient,

$$c_f = 2\left(\frac{u_\tau}{U_0}\right)^2 = 2\left(\frac{\mathrm{Re}_\tau}{\mathrm{Re}_x}\right)^2 \left(\frac{x}{\delta}\right)^2. \tag{1}$$

A vast number of correlations to formulate the dependency of $c_f$ on various Re-numbers have been proposed for the ZPG-TBL. They can be divided into several types according to their functional forms: *i)* power laws including the $1/5$th and $1/7$th ones [21] with the general form $c_f = c\mathrm{Re}^m$, *ii)* logarithmic laws $c_f = c_1(c_2 \log \mathrm{Re} + c_3)^m$ such as the suggestions of Schultz-Grunow [14], Schlichting [13], Frenholz-Finley [5], and White [21], and finally *iii)* other types such as the Prandtl-Kármán relation that can be found in [21, 13]. The coefficients and powers in each of these correlations were estimated using the experiential data available at the time, completely or partially combined with analytical methods. However, as more recent experimental data including those by Österlund [11] and Nagib et al. [9] came out, it was shown by Nagib et al. [10] that a major part of the existing correlations must be tuned up to achieve a good agreement with high Re-number experiments.

Among all the available correlations, we look for those having less deviation from available benchmark data, specifically at higher Re-numbers, while possessing a simple mathematical formulation making it possible to analytically evaluate integrals arising in the grid estimation

process, in section 3. The family of power law correlations is a suitable candidate satisfying the above demands. Dependency between each two quantities among $\delta/x$, $c_f$, and different Re-numbers can be established given two starting correlations along with combination of (1) and the integral form of the streamwise momentum equation for ZPG-TBL which reads

$$c_f = 2\frac{d\theta}{dx} = 2\frac{d\mathrm{Re}_\theta}{d\mathrm{Re}_x}. \tag{2}$$

In Table 1, a summary of power law type correlations is presented. The oldest one is the $1/5$th law proposed by Prandtl in 1927 (as quoted in [21]), adjusted by low-Re experiments. The $1/7$th law is the outcome of combining White's [21] correlation, $c_f = 0.020\mathrm{Re}_\delta^{-1/6}$, derived from Coles' law of the wake [3] along with an assumption for $\theta/\delta$ to be constant and equal to $7/72$, which corresponds to the $1/7$th velocity profile. In these relations, $\delta$ was supposed to be the theoretical thickness of the boundary layer. On the other hand, by combining the latter assumption with $c_f = 0.02358\mathrm{Re}_x^{-1/7}$ calibrated by Nagib et al. [10], the modified version of the $1/7$th law is derived.

In order to avoid making any presumption on the shape of velocity profile, it was chosen to tune up $c_f = c_1\mathrm{Re}_\delta^{-m_1}$ and $\delta/x = c_2\mathrm{Re}_x^{-m_2}$ based on Österlund's experimental observations [11] employing a least-squares technique and then derive other correlations using (1) and (2) without any further ad-hoc assumption. The boundary layer thickness was taken to be $\delta_{99}$ ($\delta$ at the point where TBL mean streamwise velocity becomes $0.99U_0$), which is a meaningful measurable quantity in an experiment or simulation. However, it was observed that employing the theoretical $\delta$ approximately obtained by plugging experimental values of $c_f$ in Coles' law [3], for instance, would lead to very similar results for the derived correlations between $c_f - \mathrm{Re}_x$ and $c_f - \mathrm{Re}_\theta$.

| Quantity | Suggestion | $1/5$th law, [21] | $1/7$th law, [21] | Modified $1/7$th law |
|---|---|---|---|---|
| $c_f$ | $0.0283\mathrm{Re}_x^{-0.1540}$ | • $0.058\mathrm{Re}_x^{-1/5}$ | $0.027\mathrm{Re}_x^{-1/7}$ | • $0.02358\mathrm{Re}_x^{-1/7}$, [10] |
| $\delta/x$ | • $0.1222\mathrm{Re}_x^{-0.1372}$ | • $0.37\mathrm{Re}_x^{-1/5}$ | • $0.16\mathrm{Re}_x^{-1/7}$ | • $0.16\mathrm{Re}_x^{-1/7}$ |
| $c_f$ | • $0.01947\mathrm{Re}_\delta^{-0.1785}$ | $0.0452\mathrm{Re}_\delta^{-1/4}$ | • $0.02\mathrm{Re}_\delta^{-1/6}$ | $0.0174\mathrm{Re}_\delta^{-1/6}$ |
| $\mathrm{Re}_\delta$ | $0.1222\mathrm{Re}_x^{0.8628}$ | $0.37\mathrm{Re}_x^{4/5}$ | $0.16\mathrm{Re}_x^{6/7}$ | $0.16\mathrm{Re}_x^{6/7}$ |
| $\mathrm{Re}_\tau$ | $0.0145\mathrm{Re}_x^{0.7858}$ | $0.063\mathrm{Re}_x^{7/10}$ | $0.0186\mathrm{Re}_x^{11/14}$ | $0.01737\mathrm{Re}_x^{11/14}$ |
| $\mathrm{Re}_\theta$ | $0.0167\mathrm{Re}_x^{0.8460} + c_\theta$ | $0.0363\mathrm{Re}_x^{4/5}$ | $0.01575\mathrm{Re}_x^{6/7}$ | $0.01376\mathrm{Re}_x^{6/7}$ |

Table 1: Summary of the power law correlations for ZPG flat plate turbulent boundary layer. Symbol • represents the starting relations in each set. In the first column, $c_\theta = 373.83$.

The variation of $c_f$ with $\mathrm{Re}_x$ and $\mathrm{Re}_\theta$ estimated by different correlations is compared to the experimental measurements as well as direct numerical simulation (DNS) data from different sources over $5 \cdot 10^5 \leq \mathrm{Re}_x \leq 10^{10}$ in Figure 1. It immediately follows from the the proposed correlations that

$$c_f = 0.0134\left(\mathrm{Re}_\theta - c_\theta\right)^{-2/11},$$

which with a constant-value $c_\theta$ exhibits a good performance in the whole range of Re-numbers compared to DNS data [12, 16, 17] and experiments [11, 9].

It is noteworthy that at low Re-numbers only $c_f = 0.024\mathrm{Re}_\theta^{-1/4}$ given by Smits et al. [18] agrees with DNS of Schlatter and Örlü [12], besides the correlation suggested here. However, it, along with the $1/5$th law, diverges from the benchmark values and other curves as Re increases.

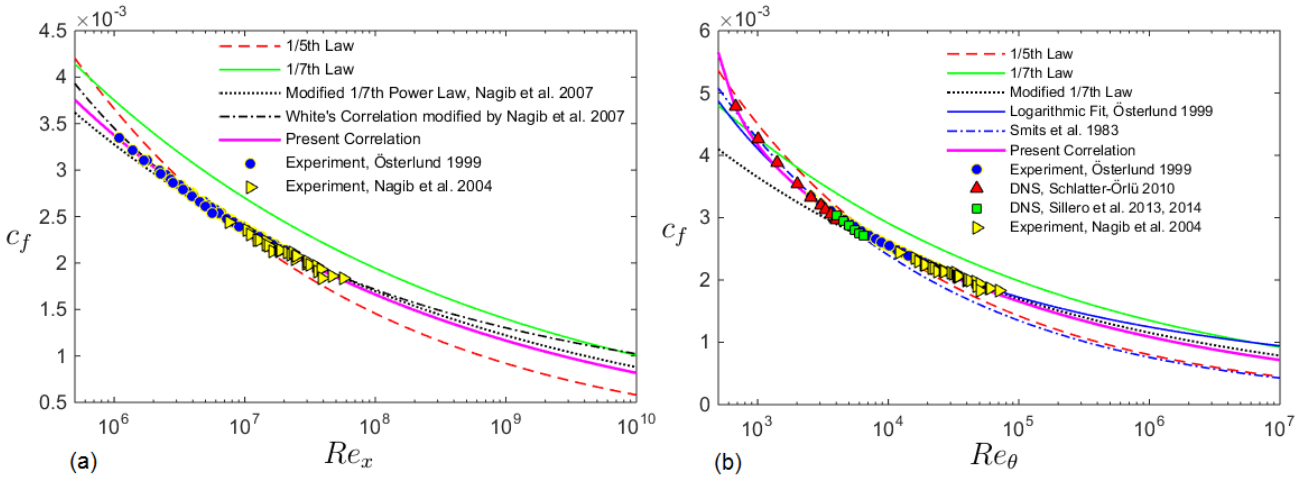



Figure 1: The friction coefficient, $c_f$, versus (a) $\mathrm{Re}_x$ and (b) $\mathrm{Re}_\theta$. White's correlation $c_f \approx 0.455\left[\ln(0.06\mathrm{Re}_x)\right]^{-2}$, [21], is modified by replacing $0.455$ with $0.4177$ as suggested by Nagib et al. [10]. Österlund's logarithmic fit: $c_f = 2\left[\frac{1}{\kappa} ln\mathrm{Re}_\theta + C\right]^{-2}$ with $\kappa = 0.384$ and $C = 4.08$ [11].

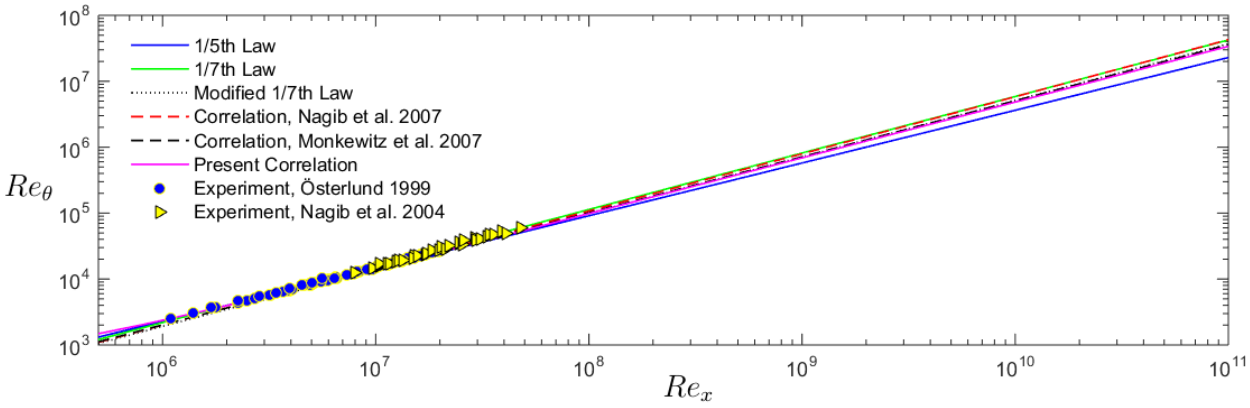



Figure 2: $\mathrm{Re}_\theta$ versus $\mathrm{Re}_x$ predicted by correlations and given by experimtns.

Relatively better performance of the suggested $c_f - \mathrm{Re}_\theta$ curve at low Re-numbers is mainly due to the additive constant appearing when taking the integral of (2). By using the lowest Re-number data among Österlund's measurements [11], $c_\theta$ is predicted to be $373.83$. This constant can also be seen as a single-point correction for $x$, taking into account the initial distance from the plate leading edge through which the boundary layer could have been laminar and then transited to turbulent. More detailed correction methods for the $\theta - x$ dependency can be found in [10, 8].

According to Figure 2, the deviation between various predictions increases with $\mathrm{Re}_\theta$ which, along with lack of experimental evidence, makes it impossible to draw a conclusion on the best performance. However, as suggested by Nagib et al [10], for numerical estimation no preference can be given to any correlation at high Re-numbers, although it is observed that the modified $1/7$th and the suggested correlations are respectively the closest curves to $\mathrm{Re}_\theta = 0.016\mathrm{Re}_x^{0.85}$ fitted in $10^6 \leq \mathrm{Re}_x \leq 10^9$ by Monkewitz et al [8] to their detailed asymptotic expansion for $\theta(x)$, taking into account corrections for the virtual origin.

On the other hand, at high Re-numbers (say $\mathrm{Re}_x > 10^9$), the correlation $\mathrm{Re}_\theta = 0.01277\mathrm{Re}_x^{0.8659}$ proposed by Nagib et al. [10] behaves very closely to the 1/7th power law. But it must be em-

phasized again that a concrete comment regarding accuracy of different correlations at very high Re-numbers cannot be made.

## 3 A GRID ESTIMATE FOR NWM-LES

A grid estimate is derived in this section for NWM-LES of turbulent boundary layers. The setting is a flat-plate ZPG-TBL, but the estimate is intended to be useful for NWM-LES of flows with TBLs at curved walls as well. The derivation is applicable to any general unstructured grid generation approach which allows for a change of grid resolution throughout the computational domain. Isotropic grid elements/cells are considered, i.e. there is no systematic directional stretching of grid cells.

As was also the case in [1, 2], the starting point is an expression for the growth of the thickness of the TBL, along the flat plate, of the following power law form.

$$\delta(x) \approx \alpha x \mathrm{Re}_x^{\beta} = \alpha \left(\frac{U_0}{\nu}\right)^{\beta} x^{1+\beta} \tag{3}$$

In the previous section, different suggestions for the coefficients $\alpha$ and $\beta$ were given and their agreement with experimental and DNS data was discussed. Note that another notation ($c_i$) was used there. The use of $\alpha$ and $\beta$ here avoids subscripts. The derivation below is carried out without fixing the coefficient values and can hence be used with all such suggestions of the power law type.

A Cartesian coordinate system is used such that the free-stream velocity is directed in the positive $x$-direction, and the plate is located at, $0 < x < l$, $y = 0$, $0 < z < b$. Thus the length of the plate is $l$, and its width is $b$. The volume is divided into three different regions based on the characteristics of the flow. The first region, denoted $V_1$, is the fully developed TBL, $x_0 < x < l$, where equation (3) is an accurate description of the boundary layer thickness. Here $x_0$ is a suitably chosen location which has an important effect on the resulting grid estimate, as discussed below. The second region, denoted $V_2$, is the initial part of the boundary layer, $0 < x < x_0$. Typically, this includes laminar flow and then transition to turbulence in the boundary layer. Here, the situation is considered when the focus is on the TBL (and not on transition). Thus, it is only required to "initialize" the resolved fluctuations in the NWM-LES. How this can be done with a numerical tripping device is briefly illustrated in the next section. Finally, the third region is what remains of the domain in which the flow is to be simulated. This region is discarded from the grid estimate, as it is well-known that the grid resolution requirements are driven by the TBL, [1]. In the context of unstructured grid generation, a very significant coarsening of the grid can then be used outside of the boundary layer.

In order to obtain the grid estimate, the local cell density is introduced, $\rho_N(x) = \Delta N/\Delta V$. Here, $\Delta N$ is the number of cells in the volume $\Delta V$. The total number of cells in a volume $V$ is then obtained by integration,

$$N = \int_V \rho_N \, \mathrm{d}V. \tag{4}$$

For an unstructured grid with isotropic cells, and the cell size adapted to the local boundary layer thickness, the cell density is given by $\rho_N = n_0/\delta^3(x)$. Here $n_0$ is the target value for the number of cells in a cube with a side length equal to $\delta$. In the literature, $10^3 < n_0 < 10^4$, is considered to be a suitable range for NWM-LES, [1, 20, 4]. For definiteness, $n_0 = 2\,500$, as suggested in the paper by Chapman, [1], is used below.

The required number of cells in region $V_1$ (the TBL) can now be estimated by,

$$N_1 = \int_0^b \int_0^{\delta(x)} \int_{x_0}^l \frac{n_0}{\delta(x)^3} \mathrm{d}x\mathrm{d}y\mathrm{d}z = bn_0 \int_{x_0}^l \frac{\mathrm{d}x}{\delta(x)^2} = \frac{bn_0}{\alpha^2} \left(\frac{\nu}{U_0}\right)^{2\beta} I_1, \tag{5}$$

where the integral $I_1$ is given by,

$$I_1 = \int_{x_0}^l \frac{\mathrm{d}x}{x^{2+2\beta}} = \frac{1}{1+2\beta}\left(\frac{1}{x_0^{1+2\beta}} - \frac{1}{l^{1+2\beta}}\right) < \frac{1}{1+2\beta}\frac{1}{x_0^{1+2\beta}} =: \bar{I}_1.$$

Here the last relation is used to define $\bar{I}_1$. The following somewhat surprising property is emphasized. In the limit $l \to \infty$, the integral $I_1$ is finite (and has the value $\bar{I}_1$), whereas it diverges as $x_0 \to 0$. Inserting the integral into equation (5) the grid estimate and its upper limit (denoted $\bar{N}_1$), are obtained in the following forms.

$$N_1 = \frac{n_0}{(1+2\beta)\alpha^2}\frac{b}{x_0}\left(\frac{1}{\mathrm{Re}_{x_0}^{2\beta}} - \frac{x_0}{l}\frac{1}{\mathrm{Re}_l^{2\beta}}\right) < \frac{n_0}{(1+2\beta)\alpha^2}\frac{b}{x_0}\mathrm{Re}_{x_0}^{-2\beta} =: \bar{N}_1 \tag{6}$$

It is clear from this estimate that there is a significant clustering of cells for $x \sim x_0$, whereas the grid cell coarsening for increasing $x$ is rapid.

Next, the number of cells in the initial part of the boundary layer, region $V_2$, is estimated by assuming a constant cell density, $\rho_N = n_0/\delta^3(x_0)$, in a layer of constant thickness $\delta(x_0)$, which leads to,

$$N_2 = \frac{bx_0n_0}{\delta^2(x_0)} = \frac{n_0}{\alpha^2}\frac{b}{x_0}\mathrm{Re}_{x_0}^{-2\beta}. \tag{7}$$

It is observed that the number of cells in this region is comparable to the number of cells in the TBL, i.e. $N_2/\bar{N}_1 = 1+2\beta$.

It is clear from the estimates (6) and (7) that $x_0$ is a crucial parameter. It was introduced above as a location after which the expression (3) is an accurate representation of the boundary layer thickness. This is however not an exactly defined location. For definiteness, $\mathrm{Re}_{x_0} = 5\cdot 10^5$ is chosen, which is an approximate location of transition, and it is also a suitable choice for the type of "tripped" LES which is illustrated in the next section. Thus, $\mathrm{Re}_{x_0}$ is considered fixed and, $b/x_0 = \mathrm{Re}_b/\mathrm{Re}_{x_0}$, is inserted into the estimate for $\bar{N}_1$, see equation (6), to obtain,

$$\bar{N}_1 = \frac{n_0}{(1+2\beta)\alpha^2}\mathrm{Re}_{x_0}^{-1-2\beta}\mathrm{Re}_b.$$

From this expression, it is seen that it is the Re-number based on the plate width which determines the increase in grid cells with increasing Re-number. This fact was not clearly pointed out in previous investigations, [1, 20, 2]. Another interesting fact is that $\bar{N}_1$ grows linearly with $\mathrm{Re}_b$, and that the coefficient $\beta$ only affects the proportionality constant.

The growth in the number of grid cells with increasing Re-number is illustrated in Figure 3. Because of the role of $x_0$ in the estimates, graphs are shown for two different aspect ratios ($b/l$) of the plate, 4 and 1/4, respectively. Also included is the estimate from the paper by Chapman [1],

$$N_{Chap} = 40\frac{b}{l}n_0\mathrm{Re}_l^{0.4}, \tag{8}$$

and the estimate derived by Choi and Moin, [2],

$$N_{CM} = 54.7\frac{b}{l}n_0\mathrm{Re}_l^{2/7}\left[\left(\frac{\mathrm{Re}_l}{\mathrm{Re}_{x_0}}\right)^{5/7} - 1\right]. \tag{9}$$

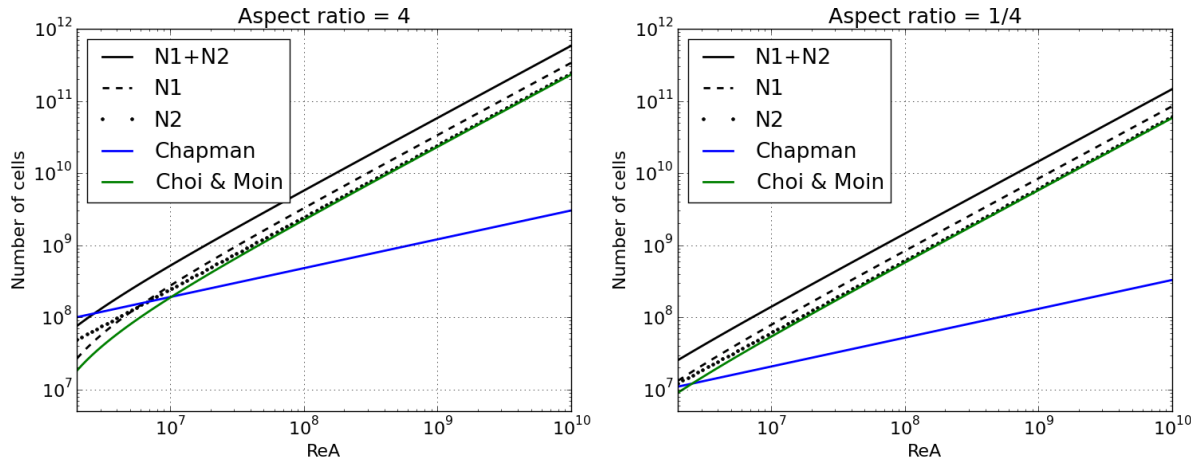



Figure 3: Grid estimates for a flat plate as a function of Re-number based on the plate area, $\mathrm{Re}_A = U_0\sqrt{bl}/\nu$. To the left the aspect ratio is $b/l = 4$, and to the right the aspect ratio is $b/l = 0.25$. The black lines (full, dashed, and dotted, respectively) show the estimates proposed in the present paper, where $N_1$ is given by equation (6), and $N_2$ is given by equation (7), and the parameters have the values $\alpha = 0.1222$, and $\beta = -0.1372$, see Table 1. The estimate by Chapman [1], blue line, is given by equation (8), and the estimate by Choi and Moin [2], green line, is given by equation (9).

As clearly illustrated in Figure 3, the estimate by Chapman is by far the most optimistic, in the sense that for high Re-numbers it leads to significantly lower estimates of the required number of grid cells. The main reason for this is that Chapman, [1], does not use integration, i.e. equation (4), to obtain the estimate. Instead an average boundary layer thickness over the whole plate is used. That it is more appropriate to use integration was pointed out in [20, 2], and the authors of the present paper agree with that assessment. Another remark is that the estimate by Choi and Moin [2], equation (9), only takes into account the part of the TBL downstream of $x_0$, whereas the estimate $N_1 + N_2$, also addresses the initial part of the boundary layer.

## 4 NWM-LES OF A FLAT PLATE TURBULENT BOUNDARY LAYER

This section contains a brief description of the computational set-up and results for one large-eddy simulation of a flat plate turbulent boundary layer. The first purpose is to illustrate what predictive accuracy can be expected with the mesh resolution levels discussed in the previous section. The second purpose is to demonstrate the practicality of the overall approach, including components such as the numerical tripping device used to induce resolved fluctuations in the boundary layer.

| Quantity | Notation | Value | Unit |
|---|---|---|---|
| Length | $l$ | 2.000 | m |
| Width | $b$ | 0.200 | m |
| Height | $h$ | 0.133 | m |
| Kin.visc. | $\nu$ | $1.65{\cdot}10^{-5}$ | $\mathrm{m}^2/\mathrm{s}$ |
| Velocity | $U_0$ | 20.4 | m/s |

Table 2: Parameters of the flat plate simulation. The plate has the aspect ratio 1/10, and the Reynolds numbers, $\mathrm{Re}_b$=$2.42{\cdot}10^5$, and $\mathrm{Re}_l$=$2.42{\cdot}10^6$.

The simulation case parameters are summarized in Table 4, and they lead to a length-based

Re-number of $\mathrm{Re}_l = 2.42 \cdot 10^6$. In fact, this corresponds to the TBL approaching the contoured ramp investigated by wind tunnel measurements in [19], and by LES in [7], which however is irrelevant for the present investigation. At the inflow boundary, a constant velocity $U_0$ is prescribed in the $x$-direction. A "strip" of grid cells is used for the numerical tripping device. In these cells the flow is perturbed by introducing a volume force in the momentum equation. The variation of the volume force is random in space and time. The TBL develops along the floor of the wind tunnel. On the lateral boundaries, periodic boundary conditions are used, and on the top boundary patch no-slip is used. The subgrid model employed is the One Equation Eddy Viscosity Model (OEEVM), [15, 7]. A wall model is used which modifies the viscosity in the cell-layer adjacent to the wall, as described in [6].

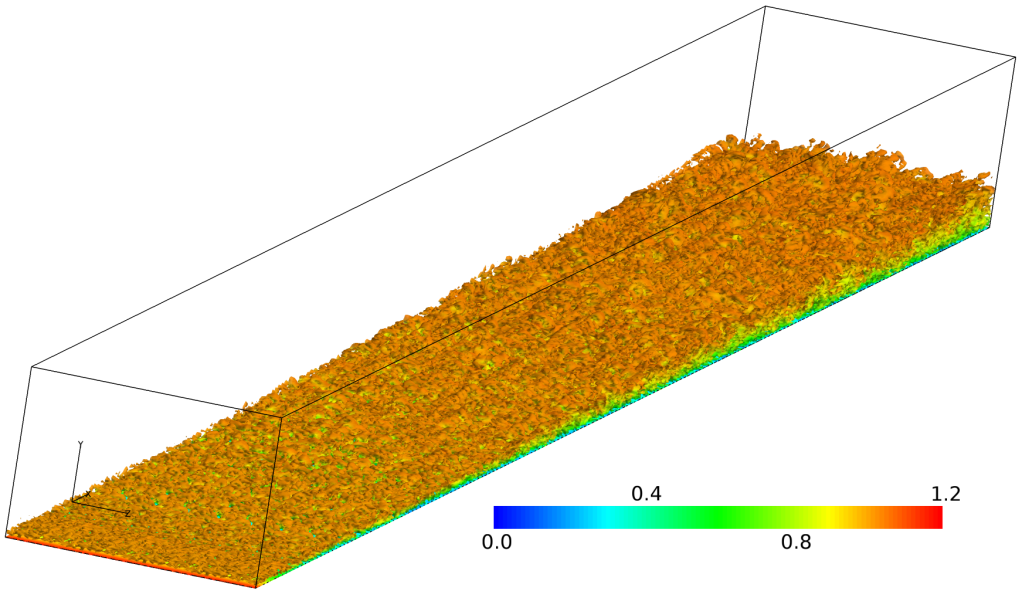



Figure 4: Illustration of the simulated TBL flow. An iso-surface of the second invariant of the velocity gradient is used to illustrate the turbulent structures in the boundary layer. The iso-surface is colored by the instantaneous normalized axial velocity, $v_x/U_0$. The black lines show the edges of the simulation domain.

A simple grid was used, which consists of hexahedral cells equal in size and shape throughout the domain. The number of cells was, $760 \times 160 \times 100$=12 160 000. This means that the grid adaption which is the basis of the grid estimates in the previous section, was not taken advantage of. However, the average grid resolution in the TBL in the simulation is similar to what is indicated in the grid estimates. Adopting the parameter values $\alpha = 0.1222$ and $\beta = -0.1372$, given in Table 1, $N_1$ and $N_2$ are estimated to be

$$\begin{aligned} N_1 &= 2.87 \cdot 10^6, \\ N_2 &= 3.03 \cdot 10^6, \end{aligned}$$

while employing Chapman's [1] and Choi-Moin's [2] approaches results in

$$\begin{aligned} N_{Chap} &= 3.61 \cdot 10^6, \\ N_{CM} &= 1.96 \cdot 10^6. \end{aligned}$$

Therefore, the simulation results give a rough indication of the level of accuracy which can be expected with the relevant grid resolution relative to the TBL thickness.

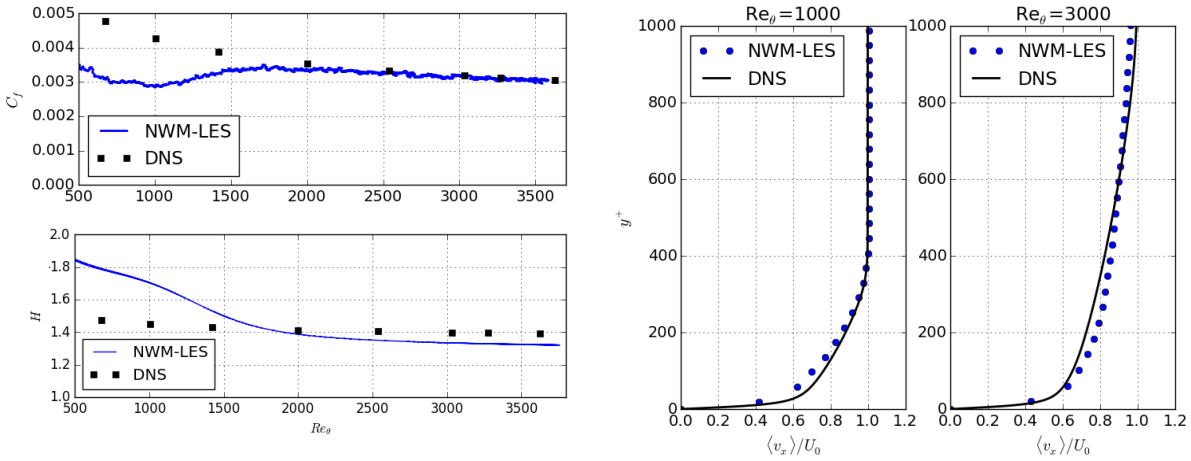



Figure 5: Left: Graphs for the friction coefficient $c_f$, and the shape factor $H$, as functions of $\mathrm{Re}_\theta$. Right: Profiles of the normalized mean velocity, $\langle v_x \rangle / U_0$, plotted versus $y^+$, at $\mathrm{Re}_\theta$ = 1000 and 3000, respectively. Comparison of NWM-LES with DNS data, [12].

The flow is shown in Figure 4, by an iso-surface of an invariant of the instantaneous velocity gradient, to illustrate vortical structures in the TBL. In Figure 5, results from the NWM-LES are compared with DNS-data, [12]. The comparison includes the friction coefficient, the shape factor $H = \delta^*/\theta$ (i.e. the quotient of the displacement thickness and the momentum thickness), as well as mean velocity profiles at two stations along the flat plate. The curves for the friction coefficient and the shape factor clearly indicate the adaption length of the NWM-LES boundary layer. Overall, considering the significantly lower computational cost of the NWM-LES, as compared to DNS, the agreement of the results is quite good. A detailed elucidation of the discrepancies between the NWM-LES and DNS results is outside the scope of the present paper.

## 5 CONCLUDING REMARKS

A grid estimate has been derived for NWM-LES, as summarized in the expressions (6) and (7), for the number of grid cells in the initial length and the fully developed TBL, respectively. The derivation follows the approach of Choi and Moin [2], but it is more general as it holds for any power law for $\delta$. Another novel feature is that an additional grid estimate is proposed for the initial length of the boundary layer, affected by transition. Furthermore, the grid estimate is formulated in a new way to highlight the clustering of the grid in the region starting from the leading edge of the plate up to $\mathrm{Re}_x \sim 10^6$.

New power law coefficients are also proposed, based on a data fit to recent high-Re experimental data. The coefficients are given in Table 1 together with a review of other proposed values from the literature.

In section 4, a complete simulation methodology is demonstrated for NWM-LES of a flat plate ZPG-TBL. The grid resolution level used in the simulation is comparable to that underlying the grid estimates. Note, however, that the type of grid coarsening and unstructured grid generation approach discussed in section 3 is not applied in the example simulation of section 4. The purpose of the simulation is to assess the predictive accuracy, with the relevant grid resolution relative to the boundary layer thickness. This is accomplished by comparing the friction coefficient, the shape factor and the mean velocity profile with DNS data.

## ACKNOWLEDGEMENT

The authors wish to thank Timofey Mukha, at Uppsala University, for fruitful scientific discussions concerning the simulation of TBLs, and also for providing us with very convenient post-processing scripts. Niklas Wikström, at FOI, is thanked for the implementation of the simple and robust trip-model which was used in the simulation described in section 4. The work was supported by Grant No 621-2012-3721 from The Swedish Research Council.

## REFERENCES

[1] D. R. Chapman. Computational aerodynamics development and outlook. *AIAA Journal*, 17:1293–1313, 1979.

[2] H. Choi and P. Moin. Grid-point requirements for large eddy simulation: Chapman's estimates revisited. *Physics of Fluids*, 24(1):011702, 2012.

[3] D. E. Coles. The law of the wake in the turbulent boundary layer. *J. Fluid Mech.*, 1:191–226, 1956.

[4] L. Davidson. Large Eddy Simulation: How to evaluate resolution. *International journal of Heat and Fluid Flow*, 30:1016–1025, 2009.

[5] H. H. Fernholz and P. J. Finley. The incompressible zero-pressure-gradient turbulent boundary layer: An assessment of the data. *Prog. Aerospace Sci.*, 32:245–311, 1996.

[6] C. Fureby, N. Alin, N. Menon, S. Svanstedt, and L. Persson. On Large Eddy Simulation of high Reynolds number wall bounded flows. *AIAA Journal*, page 457, 2004.

[7] C. Fureby, S. Zhu, and D. Jones. Large Eddy Simulation of the Flow over a Countoured Ramp. In *9th International Symposium on Turbulence and Shear Flow Phenomena*, Melbourne, Australia, 2015.

[8] A. Monkewitz, K. A. Chauhan, and H. M. Nagib. Self-consistent high-reynolds-number asymptotics for zero pressure- gradient turbulent boundary layers. *Phys. Fluids*, 19:115101, 2007.

[9] H. M. Nagib, C. Christophorou, and P. A. Monkewitz. High Reynolds number turbulent boundary layers subjected to various pressure-gradient conditions. In G. E. A. Meier and K. R. Sreenivasan, editors, *IUTAM Symp. on One Hundred Years of Boundary Layer Research*, pages 383–394. Springer, Berlin, 2004.

[10] H. N. Nagib, K. A. Chauhan, and P. A. Monkewitz. Approach to an asymptotic state for zero pressure gradient turbulent boundary layers. *Phil. Trans.R. Soc. A*, 365:755–770, 2007.

[11] J. M. Österlund. *Experimental studies of zero pressure-gradient turbulent boundary layer flow*. PhD thesis, Department of Mechanics, Royal Institute of Technology, Stockholm, 1999.

[12] P. Schlatter and R. Örlü. Assessment of direct numerical simulation data of turbulent boundary layers. *J. Fluid Mech.*, 659:116–126, 2010.

[13] H. Schlichting and K. Gersten. *Boundary Layer Theory*. Springer, Berlin, 8th edition, 2000.

[14] F. Schultz-Grunow. New frictional resistance law for smooth plates. *NASA Technical Memorandum*, 986, 1940.

[15] U. Schumann. Subgrid scale model for finite difference simulation of turbulent flows in plane channels and annuli. *J.Comp.Phys.*, 18(4):376–404, 1975.

[16] J. A. Sillero, Jiménez, and R. D. Moser. One-point statistics for turbulent wall-bounded flows at Reynolds numbers up to $\delta^+ \approx 2000$. *Phys. Fluids*, 25(105102), 2013.

[17] J. A. Sillero, Jiménez, and R. D. Moser. Two-point statistics for turbulent boundary layers and channels at Reynolds numbers up to $\delta^+ \approx 2000$. *Phys. Fluids*, 26(105109), 2014.

[18] A. J. Smits, N. Matheson, and P. N. Joubert. Low-Reynolds-number turbulent boundary layers in zero and favourable pressure gradients. *J. Ship Res.*, 27:147–157, 1983.

[19] S. Song, D. DeGraaff, and J. K. Eaton. Experimental study of a separating, reattaching and redeveloping flow over a smoothly contoured ramp. *Int. J. Heat and Fluid Flow*, 21(5):512–519, 2000.

[20] P. R. Spalart, W. H. Jou, M. Strelets, and S. R. Allmaras. Comments on the feasibility of LES for wings and on a hybrid RANS/LES approach. In *Advances in DNS/LES*, volume 1, pages 4–8, 1997.

[21] F. M. White. *Viscous Fluid Flow*. McGraw-Hill, New York, 3rd edition, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A HYBRID MESH HARMONIC BALANCE SOLVER FOR THE AEROELASTIC ANALYSIS OF TURBOMACHINERY

**Hans-Peter Kersken***, **Graham Ashcroft, Christian Frey**

German Aerospace Center (DLR)
Institute of Propulsion Technology
Linder Höhe, 51147 Cologne, Germany
email:{hans-peter.kersken, graham.ashcroft, christian.frey}@dlr.de

**Keywords:** CFD, Harmonic Balance Method, Unstructured Grid, Flutter, Aeroelasticity, Turbomachinery

**Abstract.** *An implementation of a Harmonic Balance method to simulate time-periodic, nonlinear flows in turbomachinery on hybrid grids comprising both structured and unstructured blocks is described. It has been developed with the aim to apply it to aeroelastic analysis problems which require support for deforming meshes. The necessary modifications of an existing implementation of a Harmonic Balance method on structured grids are presented. Especially block boundaries require a separate treatment. For the same spatial discretization order the solver for unstructured grids requires a moderate overhead in computational resources in terms of computation time compared to the structured solver when applied on grids with similar spatial resolution. The extended HB solver is applied to a standard turbomachinery aeroelastic test case for inviscid flows including both subsonic and transonic configurations. The computational results obtained are compared with results from a Harmonic Balance solver for structured grids and references from literature.*

# 1 INTRODUCTION

Harmonic Balance (HB) methods have been introduced into turbomachinery CFD to efficiently simulate periodic time-dependent flow phenomena by taking into account frequencies relevant for the phenomenon under investigation only. These phenomena have traditionally been simulated based on linearized governing equations starting from two-dimensional Euler equations to finally three-dimensional compressible Navier-Stokes equations. The linearization decouples the governing equations for the flow perturbations from those governing the mean flow and allows highly efficient solvers to be constructed for the perturbations. Whilst highly efficient this approach does not take into account nonlinear phenomena. Therefore such tools are not expected to give accurate results when applied to problems in which nonlinear effects are suspected to play an important role. HB methods, see for example [1, 2, 3, 4], overcome these restrictions and allow frequency-domain methods to be applied to nonlinear, time-periodic problems thereby taking into account the nonlinear coupling of perturbations and the mean flow.

The extension of the Harmonic Balance solver to unstructured and finally hybrid grids, comprising both structured and unstructured blocks, is driven by the requirement to include geometrically complex components of turbomachines like cooling channels or cavities, which are often modelled by unstructured grids. These geometrical details are a source of disturbances which may excite blade vibrations and therefore have to be taken into account when investigating aeroelastic phenomena. The newly developed solver for unstructured grids is based on previous work on aeroelastic analysis on unstructured grids with a time-linearized solver [5] and the development of a Harmonic Balance method on deforming grids [6] in the framework of DLR's block structured compressible URANS solver TRACE [7].

This paper is organized as follows: first an overview of the base flow solver, the HB method and the spatial discretization schemes is given. Then the modifications to the implementation are explained which are necessary when extending the existing HB implementation to unstructured grids. Finally results are compared to reference results when the solver for unstructured grids is applied to a duct and the Standard Configuration 10 aeroelastic test case.

We will restrict ourselves in this paper to inviscid flows as the initial validation step for the HB solver on hybrid grids.

# 2 NUMERICAL METHOD

## 2.1 Underlying flow solver

TRACE [7] is a parallel Navier-Stokes flow solver for structured and unstructured grids that has been developed at DLR's Institute of Propulsion Technology to model and investigate turbomachinery flows. The code solves the finite-volume discretization of the compressible Reynolds-averaged Navier-Stokes (RANS) equations in the relative frame of reference using a multi-block approach. For details of the discretization on structured grid refer to [8] and for a detailed description of the unstructured discretization see [9]. For the present work it is sufficient to note that following the discretization of the spatial operators in the Navier-Stokes equations the following system of ordinary differential equations (ODEs) is obtained

$$\frac{d\mathbf{q}}{dt} + \mathbf{R}\left(\mathbf{q}(t)\right) = 0 \tag{1}$$

where $\mathbf{q}$ is the vector of conservative variables, $\mathbf{R}$ is the discretized RANS residual vector and $t$ denotes the physical time.

### 2.2 Harmonic balance method

The Harmonic Balance method described in this work is part of the CFD code TRACE. For completeness we will summarize the HB algorithm as it is implemented in TRACE. For a detailed description see [10]. Time-periodic solutions of Eqn. (1) can often be described by a limited number of solution harmonics, i.e.,

$$\mathbf{q}(x,t) = \mathrm{Re}\left[\sum_{k=0}^{K} \hat{\mathbf{q}}_k(x) e^{ik\omega t}\right] \tag{2}$$

where $\hat{\mathbf{q}}_k$ are the complex valued solution harmonics and $\omega = 2\pi f$ is the fundamental angular frequency. In such cases it is computationally attractive to formulate the unsteady problem, Eqn. (1), in the frequency domain, i.e. to consider

$$\mathbf{i}k\omega\hat{\mathbf{q}}_k + \widehat{\mathbf{R}(\mathbf{q})}_k = 0 \tag{3}$$

for only a finite number of harmonics, $k = 0, \ldots, K$. For configurations in which non-linearity can be assumed to be negligible the coupling between the harmonics of $\mathbf{q}$, imposed by the nonlinear nature of $\mathbf{R}$, can be neglected and one obtains $K+1$ independent equations for the solution harmonics and the time-mean solution field. Here the equations for the harmonics are linear and only for the time-mean solution has a nonlinear equation to be solved. However, if one wishes to retain nonlinear effects an alternative approach is required in the modelling or computation of $\widehat{\mathbf{R}(\mathbf{q})}_k$. We compute $\widehat{\mathbf{R}(\mathbf{q})}_k$ as $\mathscr{F}(\mathbf{R}(\mathscr{F}^{-1}\hat{\mathbf{q}}))|_k$ and therefore solve

$$\mathbf{i}k\omega\hat{\mathbf{q}}_k + \mathscr{F}(\mathbf{R}(\mathscr{F}^{-1}\hat{\mathbf{q}}))|_k = 0, \tag{4}$$

where $\mathscr{F}$ denotes the Discrete Fourier Transform (DFT) and $\mathscr{F}^{-1}$ its inverse. Eqn. (4) is solved in the frequency domain to obtain the complex valued harmonics of the conservative variables $\hat{\mathbf{q}}_k$. To compute the harmonics of the RANS residual vector $\mathbf{R}$ the solution field vector is first reconstructed at $N$ sampling points within the period of oscillation from the Fourier coefficients of the conservative variables $\hat{\mathbf{q}}_k$ using the inverse DFT. With the reconstructed solution vectors the RANS residual vectors $\mathbf{R}$ are then computed at the $N$ sampling points to enable the DFT of $\mathbf{R}$ to be computed, i.e., the $\widehat{\mathbf{R}(\mathbf{q})}_k$. Note, since the RANS residual $\mathbf{R}$ is evaluated in the time-domain the standard flux and discretization schemes from the underlying nonlinear solver can be used.

As a hybrid time- and frequency-domain method the approach has the advantage, over methods formulated purely in the time- or frequency-domain, of being able to employ not only the possibly highly nonlinear time-domain flux functions (and their stabilizing numerical limiters) but also highly accurate nonreflecting boundary conditions formulated in the frequency domain [11]. Particularly in the context of aeroelasticity or aeroacoustics boundary conditions are of the utmost importance. The implementation in TRACE of these boundary conditions is described in [12].

For aeroelastic analysis the approach described above has to be modified to deal with time-dependent meshes as described in [6]. The governing equations are solved by an implicit pseudo-time approach. Discretizing the pseudo-time operator using the first-order Euler backward method and then linearizing the harmonic balance residual $\mathbf{R}^{\mathrm{HB}}$ in pseudo-time the following system of equations is obtained

$$\left[\left(\frac{1}{\Delta\tau} + \mathbf{i}k\omega\right) + \left.\frac{\partial\mathbf{R}}{\partial\mathbf{q}}\right|_{\hat{\mathbf{q}}_0^{(m)}}\right]\Delta\hat{\mathbf{q}}^{(m)} = -R_k^{\mathrm{HB}}\left(\hat{\mathbf{q}}^{(m)}\right), \tag{5}$$

where $\bar{V}$ is the unperturbed cell volume and

$$R_k^{\mathrm{HB}}\left(\hat{\mathbf{q}}^{(m)}\right) = \mathbf{i}k\omega\widehat{(Vq)}|_k + \mathscr{F}(V\mathbf{R}(\mathscr{F}^{-1}\hat{\mathbf{q}}))|_k, \quad \Delta\hat{\mathbf{q}}^{(m)} = \hat{\mathbf{q}}^{(m+1)} - \hat{\mathbf{q}}^{(m)}. \tag{6}$$

Here $m$ is the pseudo-time iteration counter and $V$ the time-dependent cell volume. The coupling terms have been neglected in the left-hand side and therefore it depends directly only on the time-mean solution $\hat{\mathbf{q}}_0^{(m)}$. As such the residual Jacobian $\frac{\partial \mathbf{R}}{\partial \mathbf{q}}$ is identical to that employed in the steady flow solver. The linear system of equations, Eqn. (5), is solved using either the incomplete lower upper (ILU) or successive overrelaxation (SSOR) methods.

### 2.3 Spatial discretization

So far no reference to a specific discretization method has been made because the algorithm described above can be applied unaltered to structured and unstructured meshes. To understand the differences between an HB implementation for structured and unstructured grids we have to look into the spatial discretization in more detail. For both types of grids the discretized residual vector $R_i$ at cell $i$ computed by a finite volume method is given by

$$\mathbf{R}_i = \frac{1}{V_i}\sum_{s\in B_i}\mathbf{F}_s(\mathbf{q}) - \mathbf{S}_i(\mathbf{q}). \tag{7}$$

Here $V_i$ is the volume of the cell, $B_i$ the set of its faces, $\mathbf{F}_s$ the numerical flux across face $s$ and $\mathbf{S}_i$ any source term. Apart from the source term, which depends on the state of the cell itself only, the residual is the sum of convective and viscous fluxes, $\mathbf{F}_{s,\mathrm{conv}}$ and $\mathbf{F}_{s,\mathrm{visc}}$, respectively:

$$\mathbf{F}_s(\mathbf{q}) = \mathbf{F}_{s,\mathrm{conv}}(\mathbf{q}_\mathrm{L}, \mathbf{q}_\mathrm{R}) + \mathbf{F}_{s,\mathrm{visc}}((\nabla\mathbf{q})_s) \tag{8}$$

where $\mathbf{F}_{s,\mathrm{conv}}$ is computed by the Roe upwind scheme [13] extended to second order accuracy using van Leers MUSCL technique [14] while $\mathbf{F}_{s,\mathrm{visc}}$ is discretized by a second order central scheme. The so-called left and right states, ($\mathbf{q}_\mathrm{L}$ and $\mathbf{q}_\mathrm{R}$) at face $s$ which is common to cells $i$ and $i'$ are computed with second order accurate extrapolation schemes, i.e., on structured grids,

$$\mathbf{q}_\mathrm{L} = \mathbf{q}_i + \frac{1}{2}(\widetilde{\Delta}\mathbf{q})_\mathrm{L}, \quad \mathbf{q}_\mathrm{R} = \mathbf{q}_{i'} - \frac{1}{2}(\widetilde{\Delta}\mathbf{q})_\mathrm{R}, \tag{9}$$

whereas, on unstructured grids,

$$\mathbf{q}_\mathrm{L} = \mathbf{q}_i + (\widetilde{\nabla}\mathbf{q})_\mathrm{L}\cdot(x_s - x_i), \quad \mathbf{q}_\mathrm{R} = \mathbf{q}_{i'} + (\widetilde{\nabla}\mathbf{q})_\mathrm{R}\cdot(x_s - x_{i'}) \tag{10}$$

is used. By $(\widetilde{\Delta}\mathbf{q})_{\mathrm{L/R}}$ and $(\widetilde{\nabla}\mathbf{q})_{\mathrm{L/R}}$ we denote the limited slopes and gradients at cell centers, respectively, see [15] and [9] for details. On unstructured grids the gradients are computed using the Green-Gauss approach.

## 3 IMPLEMENTATION FOR UNSTRUCTURED AND HYBRID GRIDS

The HB solver for structured grids has been implemented by reusing computational routines for the nonlinear residual $\mathbf{R}$ and its Jacobian $\frac{\partial \mathbf{R}}{\partial \mathbf{q}}$ including the computation of the numerical fluxes according to the previous section. As examination of Eqn. (5) and Eqn. (6) shows this implementation strategy carries over to the implementation of the HB solver for unstructured grids by using the respective routines from the nonlinear solver for unstructured grids.

Where different and additional implementations have to be considered is at block boundaries in a multi-block configuration. Fig. 1 depicts the situation for a two dimensional grid with the

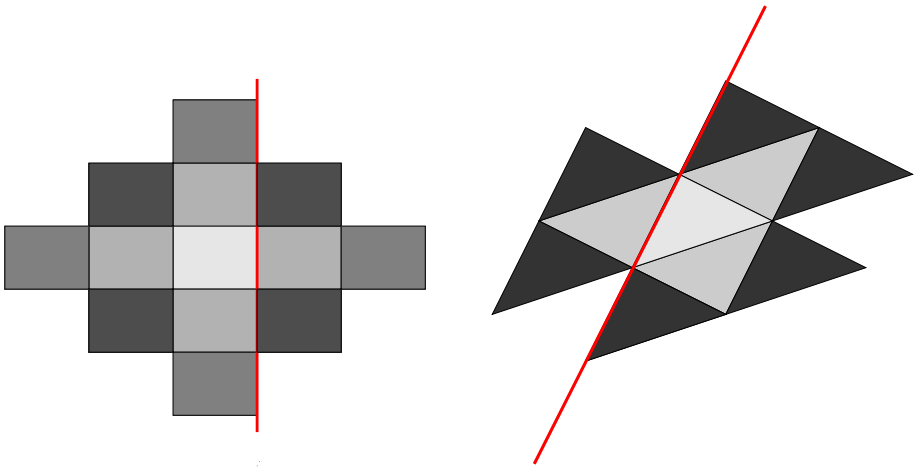


Figure 1: Stencil for the second order Navier-Stokes residual on structured and unstructured meshes [5].

focus on a single stencil where the red line marks the block boundaries. Neighbouring cells may reside on different blocks, however, their values are needed to maintain second order spatial accuracy across block boundaries, i.e., for the extrapolation given in Eqn. (9) and Eqn. (10). For completely structured grid this is solved by exchanging a two cell wide ghost cell layer surrounding each block which is filled with the respective values from neighbouring blocks. In Fig. 1 these are the cells to the right of the red line for the structured stencil. The second order extrapolation can be performed locally at block boundaries then. This procedure at block boundaries for structured grids does not carry over to unstructured grids because in TRACE we keep only one layer of ghost cells on unstructured grids which is constructed from cells having a common face with the local cell. The gradient in these cells however can not be computed locally now because information on the neighbourhood of the ghost cell, the black triangles to the left of the red line in the right picture in Fig. 1, is incomplete. Therefore in the nonlinear solver a two step process is used. First the states at the ghost cells are exchanged which allows the gradients at the cells adjacent to the boundary to be computed locally. Then the gradients are exchanged which in turn allows to interpolate gradients to boundary cell faces to compute the left and right states at the face required for second order spatial accuracy.

In the HB context we have to compute the nonlinear residual for every sampling point $t_p$ used to compute the harmonics of the residual $\widehat{\mathbf{R}(\mathbf{q})}_k = \mathscr{F}(\mathbf{R}(\mathscr{F}^{-1}\hat{\mathbf{q}}))|_k$ from the reconstructed states $\mathbf{q}(t_p) = \mathscr{F}^{-1}\hat{\mathbf{q}}$ and its gradients $\nabla\mathbf{q}(t_p)$ at the cell itself and its neighbours. If we follow the procedure for the nonlinear solver here this procedure has to be repeated for every sampling point, i.e., two communications steps are initiated each blocking the computation. It would be preferable to cluster or even avoid these communication steps. Taking the gradients at ghost cells not into account would reduce the spatial discretization order from second to first order. However, observe that the inverse DFT and the computation of the gradients are linear operations which commute, $\nabla\mathbf{q} = \nabla(\mathscr{F}^{-1}\hat{\mathbf{q}}) = \mathscr{F}^{-1}(\nabla\hat{\mathbf{q}})$. Therefore instead of computing the gradients of the states at ghost cells necessary to compute $\mathbf{R}$ at cells next to the boundary by the two step procedure they can be reconstructed from the gradients of the harmonics. This is achieved by first exchanging the harmonics $\hat{\mathbf{q}}_k$ and computing their gradients locally at cells which are ghost cells at neighbouring blocks. These gradients are then exchanged. Now all data are available to compute the nonlinear residual at every sampling point without further communication steps. For every sampling point first the states are reconstructed at the ghost cells and

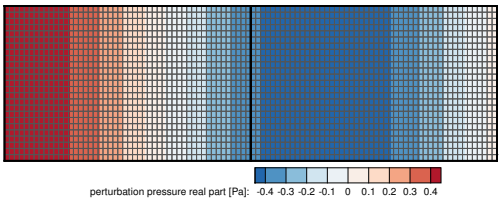



Figure 2: Grid and block topology of the 2D duct superimposed with the contours of the real part of the pressure perturbation.

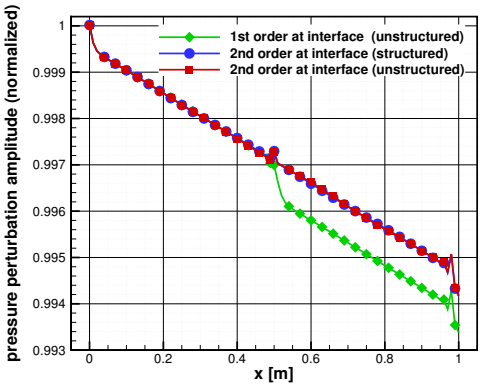



Figure 3: Normalized pressure amplitude along the center of the duct.

in the interior from their harmonics. Then the gradients at the ghost cells are reconstructed from their harmonics and finally the gradients at interior cells are computed from the reconstructed states. Now the limiter is applied and the states at the cell faces according to Eqn. (10) are computed. These steps can be performed on locally available data.

For hybrid grids comprising both structured and unstructured blocks non-matching interfaces may be generated, i.e., faces on one side of a block boundary between two blocks overlap with more than one face on the other side. In grids for turbomachinery simulation this happens especially where components with complex geometries like cooling channels or cavities are connected to the main flow geometry. Here a spatially first order coupling algorithm is implemented. The algorithm basically reconstructs the solution at the sampling points, interpolates the reconstructed states to the ghost cells at the neighbouring block and performs a DFT using all relevant frequencies at this block to update the harmonics. The algorithm is explained in detailed in [16].

## 4 VALIDATION

### 4.1 Spatial discretization order at block boundaries

To demonstrate the effectiveness of the mechanism described in section 3 a simple 2D duct configuration has been constructed comprising two blocks and 2500 cells. An acoustic wave is prescribed at the entry of the duct with an amplitude of 0.5Pa, a frequency of 400Hz and a circumferential wave number 0 in an axial mean flow with Mach number 0.6. The grid, together with the real part of the pressure perturbation, is shown in Fig. 2. The unstructured grid is constructed from the same cells as the structured grid. In Fig. 3 the pressure perturbation amplitude is plotted along the center of the duct. Without the new implementation of the gradient computation at block boundaries a jump at the block boundary can be seen due to the degradation of the spatial discretization order. With the gradients at the ghost cells included the solution using the HB solver for unstructured grids follows the solution computed with the HB solver for structured grids.

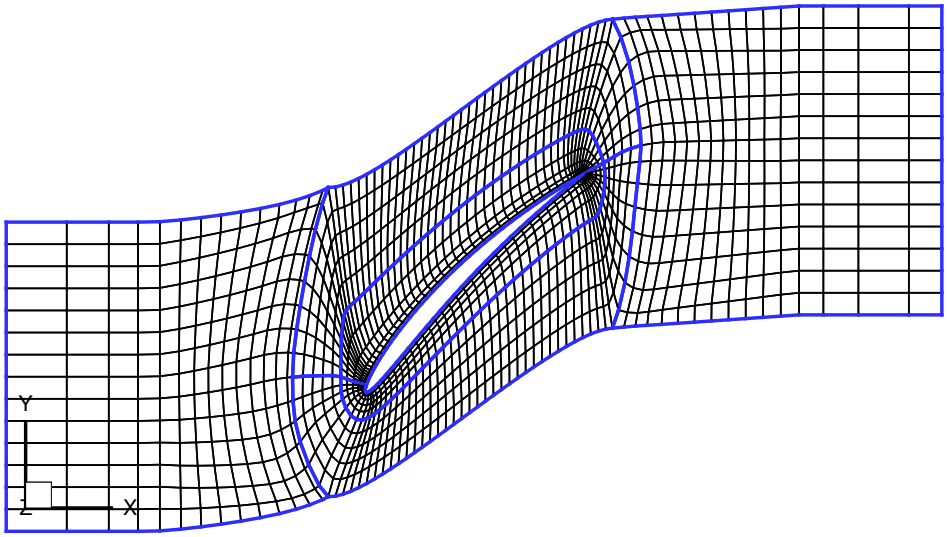



Figure 3: Computational domain, block topology and grid (every second mesh point only) for Standard Configuration 10.

## 4.2 Aeroelastic test case

To validate the HB solver for unstructured meshes for flutter problems the extended solver is applied to the well documented aeroelastic test case Standard Configuration 10 (STCF10) [17]. The test case is simulated by solving the nonlinear Euler equations. The two-dimensional computational domain comprises a total of 5600 cells distributed over 6 blocks, see Fig. 3. The entry and exit boundaries of the computational domain are located approximately one chord length upstream and downstream of the compressor blade, respectively. In [6] an excellent agreement of simulation result obtained with the HB solver in TRACE for block-structured grids with data from literature is reported. We will compare the results obtained with the solver for unstructured grids with these results and with results given in [18]. Fig. 4 shows the computed time-mean flow fields in terms of the isentropic Mach number $M_{is}$ along the surface of the blade. At both operating points the numerical results agree very well with the corresponding data for the solver on structured grids. The aeroelastic stability of the two-dimensional compressor cascade

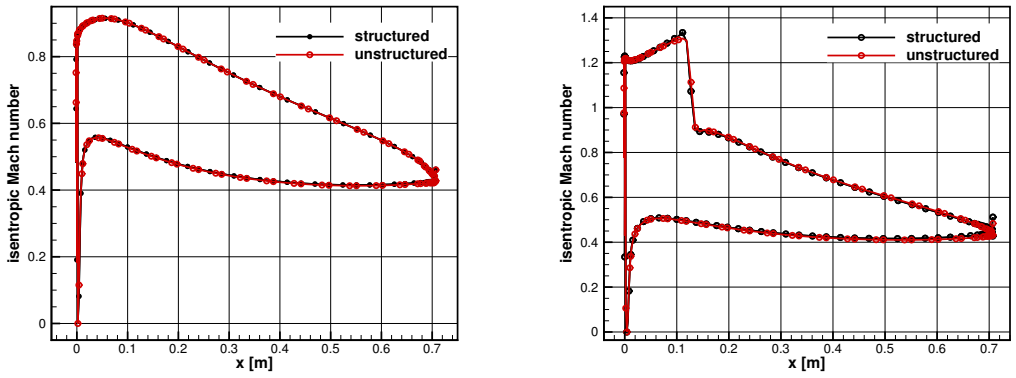



Figure 4: Isentropic Mach number distributions for the subsonic (left) and transonic cases (right) computed with the structured and the unstructured solvers.

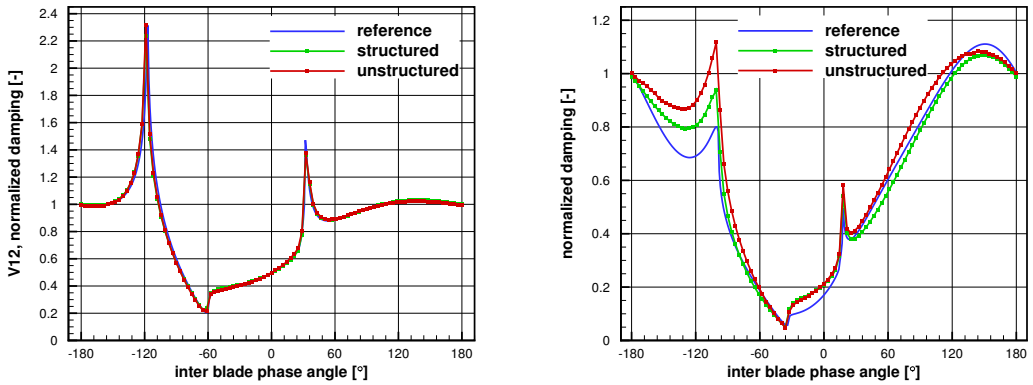



Figure 5: Normalized aerodynamic damping at subsonic (left) and transonic (right) flow conditions with $\omega^* = 1$ and $\omega^* = 0.5$, respectively.

is investigated for a torsional mode with an amplitude of $0.01°$ and axis of rotation located at midchord for a reduced frequency $\omega^*$ of 1.0 for the subsonic and 0.5 for the transonic case. $\omega^*$ is defined by $\omega^* = c\omega/|U|_{\text{Inlet}}$. Where $c$ is the chord length and $|U|_{\text{Inlet}}$ the velocity magnitude at the inlet. The simulation has been performed with a single higher harmonic, the prescribed frequency of the torsional mode. The computed values of aerodynamic damping, normalized against the value at an inter blade phase angle of 180°, are shown in Fig. 5 and compared against results obtained with the HB solver for structured grids and a reference from literature [18]. As can be seen in Fig. 5 the HB solutions show very good agreement with the structured solver results and the literature results for the subsonic case. The discrepancies observed for the transonic case can be attributed to the sensitivity of the position of the acoustic resonance at an inter blade phase angel of about -100°. Its position is very sensitive to the velocities at the exit boundary which in turn are sensitive to the position of the shock which depends to some extent on the specific solver parameters and the spatial resolution of the grid.

Structured and unstructured grids have the same size. Therefore we use this case to asses the overhead incurred by using the unstructured formulation instead of the structured. Fig. 6 shows the $L_1$ residual norm of the first harmonic for an inter blade phase angel of 60°. Having almost the same algorithmic convergence rate as seen in the right plot of the figure, from the left plot one deduces that the computation time to achieve the same residual norm about 15%

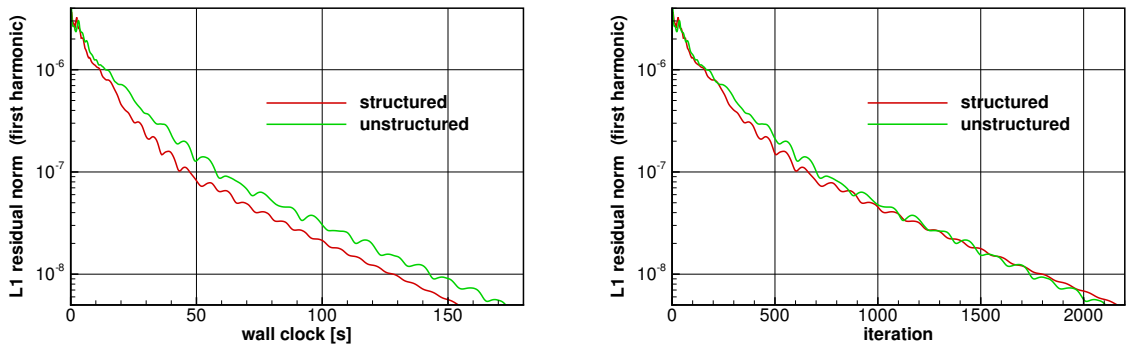



Figure 6: $L_1$ norm of the residual for the first harmonic.

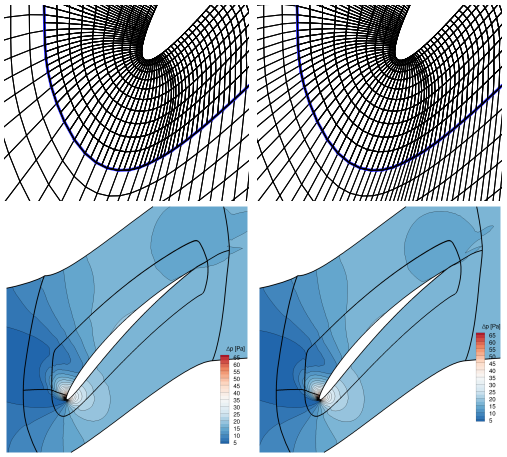


Figure 7: Mesh near the leading edge, with the interface indicated in blue, and the harmonic pressure amplitude on the hybrid grid (left) and on the refined structured grid (right).

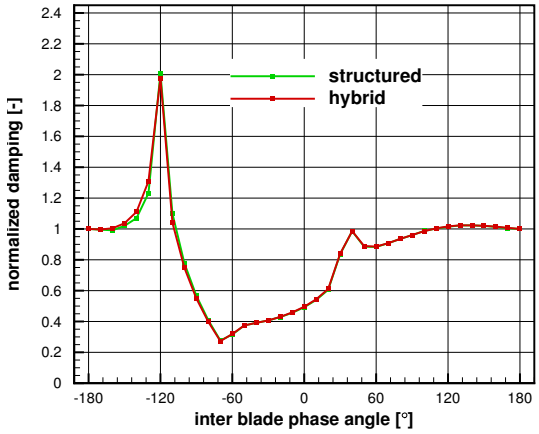



Figure 8: Damping curve normalized to value at 180° computed on the refined unstructured grid and the and the hybrid grid.

larger. This is acceptable with configurations in mind where a configuration due to its complex geometry can be simulated efficiently with an unstructured grid only.

To demonstrate the ability of the solver to handle hybrid grids the O-block surrounding the blade is refined by increasing the number of grids points by a factor of two such that at the block boundary the 1-1 connectivity is replaced by a non-matching connectivity. The O-block although having a structured topology is converted to an unstructured block, see the two plots at the top of Fig. 7. Reference results have been generated on a structured grid which is refined by a factor of two globally. As shown in Fig. 7 the pressure perturbation amplitude for an inter blade phase angle of 60° displays only small differences when computed on a grid with or without the non-matching interface. This holds for all other inter blade phase angles. The damping curve computed on the hybrid grid shows very good agreement with the results on the refined structured grid.

## 5 SUMMARY AND OUTLOOK

In this paper we described the steps to build a Harmonic Balance solver on hybrids grids for inviscid flows. It makes heavy use of functionality already available in the base solver. At block boundaries, both with 1-1 and non-matching connectivity, additional effort is necessary to maintain second order spatial accuracy at boundaries with 1-1 connectivities or handle the non-matching case. The new solver has been applied to the reference test case Standard Configuration 10. Very good agreement of the simulation results has been obtained with results from the HB solver for structured grids and reference results from the literature for the subsonic operation point. The differences observed for the transonic operation point can be explained by the sensitivity of the position of the acoustic resonance on the mean flow conditions. A simulation on a hybrid grid demonstrated the ability of solver to handle non-matching block boundaries.

The simulation results show that a HB method has been developed which includes the necessary ingredients for the aeroelastic analysis on hybrid grids. To apply the method to turbomachinery configurations with relevance to industry the solver will be extended to solve the URANS equations including a turbulence model by the HB method. For increased accuracy work is ongoing to maintain second order spatial discretization order at non-matching interfaces.

## ACKNOWLEDGEMENTS

This research was supported by the German Federal Ministry of Economic Affairs and Energy under grant numbers 03ET7040C as part of the AG Turbo project COOREFLEX.

## REFERENCES

[1] K. C. Hall, J. P. Thomas, and W. S. Clark, "Computation of unsteady nonlinear flows in cascades using a harmonic balance technique," *AIAA Journal*, vol. 40, pp. 879–886, MAY 2002.

[2] K. Ekici and K. C. Hall, "Nonlinear frequency-domain analysis of unsteady flows in turbomachinery with multiple excitation frequencies," *Collection of Technical Papers - AIAA Applied Aerodynamics Conference*, vol. 1, pp. 623–636, 2006.

[3] A. Gopinath, E. van der Weide, J. Alonso, A. Jameson, K. Ekici, and K. Hall, "Three-dimensional unsteady multi-stage turbomachinery simulations using the harmonic balance technique," in *45th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, January 8-11 2007.*, 2007.

[4] F. Sicot, G. Dufour, and N. Gourdain, "A time-domain harmonic balance method for rotor/stator interactions," *Journal of Turbomachinery*, vol. 134, no. 1, p. 011001, 2012.

[5] C. Frey and H.-P. Kersken, "A hybrid mesh linear harmonic solver for the aeroelastic analysis of turbomachinery," in *6th European Conference on Computational Fluid Dynamics (ECFD VI)*, 2014.

[6] G. Ashcroft, C. Frey, and H.-P. Kersken, "On the development of a harmonic balance method for aeroelastic analysis," in *6th European Conference on Computational Fluid Dynamics (ECFD VI)*, 2014.

[7] K. Becker, K. Heitkamp, and E. Kügeler, "Recent progress in a hybrid-grid CFD solver for turbomachinery flows," in *Proceedings Fifth European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, 2010.

[8] M. Franke, T. Röber, E. Kügeler, and G. Ashcroft, "Turbulence treatment in steady and unsteady turbomachinery flows," in *Proceedings Fifth European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, (Lisbon, Portugal), June 2010.

[9] K. Becker and G. Ashcroft, "A comparative study of gradient reconstruction methods for unstructured meshes with application to turbomachinery flows," in *52nd AIAA Aerospace Sciences Meeting*, (National Harbor, MD, USA), Jan. 2014.

[10] C. Frey, G. Ashcroft, H.-P. Kersken, and C. Voigt, "A harmonic balance technique for multistage turbomachinery applications," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[11] M. B. Giles, "Nonreflecting boundary conditions for Euler calculations," *AIAA Journal*, vol. 28, no. 12, pp. 2050–2058, 1990.

[12] H.-P. Kersken, G. Ashcroft, C. Frey, N. Wolfrum, and D. Korte, "Nonreflecting boundary conditions for aeroelastic analysis in time and frequency domain 3D RANS solvers," in *Proceedings of ASME Turbo Expo 2014*, 2014.

[13] P. L. Roe, "Approximate Riemann solvers, parameter vectors, and difference schemes," *Journal of Computational Physics*, vol. 43, no. 2, pp. 357–372, 1981.

[14] B. van Leer, "Towards the ultimate conservative difference scheme. V. a second-order sequel to Godunov's method," *Journal of Computational Physics*, vol. 32, no. 1, pp. 101–136, 1979.

[15] C. Hirsch, *Numerical Computation of Internal and External Flows – Computational Methods for Inviscid and Viscous Flows*, vol. 2. Wiley, 1 ed., 1990.

[16] C. Voigt and G. Ashcroft, "On the extension of a harmonic balance method for the simulation of compressor casing treatments," in *Proceedings of the ASME Turbo Expo 2016*, 2016.

[17] T. H. Fransson and J. M. Verdon, "Updated report on Standard Configurations for the Determination of unsteady Flow Through Vibrating Axial-flow Turbomachine-Cascades," Tech. Rep. TRITA/KRV/92.009, KTH, Stockholm, 1992.

[18] RPMTurbo Pty. Ltd., "RPMTurbo Standard Configuration 10." http://rpmturbo.com/testcases/sc10/index.html, 2005-2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL AND EXPERIMENTAL INVESTIGATIONS OF DYNAMIC STALL PROBLEM ON A 2D HELICOPTER ROTOR BLADE SECTION

**A. Marino[1], S. Russo[2], G. P. Reina[3], G. Esposito[1], F. Capizzano[1], C. de Nicola[2]**

[1] Italian Aerospace Research Center
Aircraft Department
Via Maiorise snc
81043, Capua (CE), Italy
e-mail: f.capizzano@cira.it, a.marino@cira.it, web page: http://www.cira.it

[2] Università degli Studi di Napoli Federico II - Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale
Piazzale V. Tecchio, 80
80125, Napoli, Italy
e-mail: aneres85@gmail.com, denicola@unina.it

[3] Seconda Università degli Studi di Napoli - Scuola Politecnica e delle Scienze di Base
Dipartimento di Ingegneria Industriale e dell'Informazione
Via Roma, 29
81031 Aversa (CE), Italy
e-mail: giovannipaolo.reina@unina2.it

**Keywords:** Dynamic Stall, Rotor Blade Section, Gurney Flap.

**Abstract.** *This work deals with the effects of an Active Gurney Flap device on the aerodynamic performance of oscillating rotor blade sections in different incidence ranges (pre-stall and dynamic stall conditions). Numerical and experimental simulations are carried out for different flow conditions, airfoil oscillation characteristics and Gurney flap activation laws, in order to assess the potential benefits of such a device on controlling blade loads and to get a deeper insight on the involved flow phenomena. This document is focused on the numerical simulation methodology and associated results obtained for some relevant cases selected from the wind-tunnel experiments test matrix.*

## 1 INTRODUCTION

Huge efforts have been dedicated to improve the performance of aircraft systems, including propulsion systems, propellers as well as aerodynamics systems. The main objectives are better efficiency, reduced fuel consumption and corresponding environmental gains. For helicopters, blade morphing technologies have become of interest to reduce noise and vibrations.

The study of the dynamic stall phenomenon on the rotor blade represents one of the major research topics in helicopter aerodynamics. The first analysis of the dynamic stall effects on helicopter rotor blades appeared in an experimental study in 1960 (see [1]). Empirical techniques were developed ([1][2]) that attempted to represent the primary effects of dynamic stall on airfoil lift and pitching-moment characteristics. Parametric experiments were run to expand the database used for improving these empirical methods ([3][4] [5]). The experiments were primarily directed toward obtaining aerodynamic force and moment coefficients as functions of frequency and angle, and the empirical corrections, which developed from these tests, were the prime source for improving helicopter aerodynamic prediction methods. The first investigation which attempted to isolate the fluid dynamic phenomena causing dynamic stall led to a correlation between experimental pressure distributions and the movement of a strong vortex along the surface of the airfoil ([6][7]). The presence of this vortex that forms on the upper surface of the airfoil can produce lift significantly greater than that obtained in steady flow ([8][9]) and dramatic changes in the lift and moment characteristics of the airfoil ([10]). The main consequence of these phenomena is a hysteresis loop developed in both lift and pitching moments with much larger magnitudes developed than in steady flow. At the end of the cycle, when the vortex leaves the airfoil, there is an abrupt drop in lift and moment.

This fact gives strong limitations for the operational envelope of helicopters during maneuvers, high-speed flight, and operations at high density-altitude. For all these reasons, several research activities are ongoing to control/delay the rotor blade dynamic stall. Improvements rely upon the optimization of the blade airfoil shape, or, as alternative, adopting active/passive systems. Due to the strong demand for faster helicopters/blade, in recent years, the dynamic stall phenomenon has become one of the more investigated topics in aerodynamic and aeroelasticity fields. Several research activities both in experimental and numerical fields are currently focused on the design and development of advanced rotor blades equipped with active and passive devices to mitigate the detrimental effects on the helicopter performance produced by the dynamic stall on the retreating side of the rotor ([11]).

Between a range of potential technologies that could be incorporated within active segments of a helicopter main rotor blade, the ‘Active Gurney Flap’ (AGF) seems to be the most interesting solution. The AGF is a small appendix perpendicular to the surface of the airfoil (typically on the lower surface) with the ability to alter its height from zero (fully retracted) to a maximum value (fully protruded). The AGF concept is attracting more and more interest because of its relatively low power demand and stronger resistance to blade centrifugal stresses when compared to other similar active devices. To assess the potential benefits deriving from the use of AGF on smart rotors in forward flight, as well as to mature the knowledge in the field of the complex unsteady phenomena involved, the development of adequate numerical and experimental tools is required, providing methodologies able to estimate the AGF influence on aerodynamic blade loads.

In the last 20 years, several wind tunnel facilities have been upgradated/modified to allow simulating the unsteady flow conditions around pitch oscillating airfoils. The most recent experimental research activities have been performed in the DNW-TWG transonic wind tunnel facility located at DLR-Göttingen, Germany (see [12][13]) or in the Icing Research Tunnel

(IRT) at the NASA Glenn Research Center ([14]). One of the major limit of the existing WT facilities is that they are able to cover only a limited "Mach-Reynolds-Pitching Reduced frequency" operative envelope with respect to the typical flight conditions of helicopters. With this work, we propose to reproduce, in CIRA Icing Wind Tunnel Test facility, Deep Dynamic Stall conditions acting on a full-scale retreating rotor blade section referred to a medium size helicopter (NH90 like) and analyze the effect of AGF devices on the performances of the airfoil. Moreover the experimental results will be compared with numerical data of computational fluid dynamic simulations.

This paper contains a description of the experimental apparatus able to simulate Deep Dynamic Stall situations and numerical results. It is structured as follow. Section 2 deals with a description of the Experimental Set-Up, including some details about the pitch oscillating airfoil and the characteristics of CIRA Icing Wind Tunnel. Section 3 is devoted to the Numerical Results founded in different flight conditions. Finally, Section 4 contains the Conclusions.

## 2 EXPERIMENTAL SET-UP

### 2.1 The Pitch Oscillating Airfoil

The pitch-oscillating system is composed by a wall-to-wall full span model (1.1m) having a NACA 0018 airfoil section and a chord of 0.28m. It is activated by one power system (electric motor) with a reduction gear system able to generate the required pitch-oscillation motion around the model quarter-chord axis. The model is integrated in the wind tunnel test section by means of two interface disks, for each side of the test section, one connected to the pitch-oscillating motor and the other to a mechanical support. A general view of the pitch-oscillating system is reported in Figure 1.

From a technical point of view, according with the given requirements, the pitch-oscillating system has been designed to generate a pitch-oscillating motion of the 2D model with the following laws:

$$\alpha = \alpha_M - \alpha_0 \sin(\omega t)$$

where $\alpha_M$, $\alpha_0$ and $\omega$ are, respectively, the mean angle, the oscillating angle and the angular velocity.

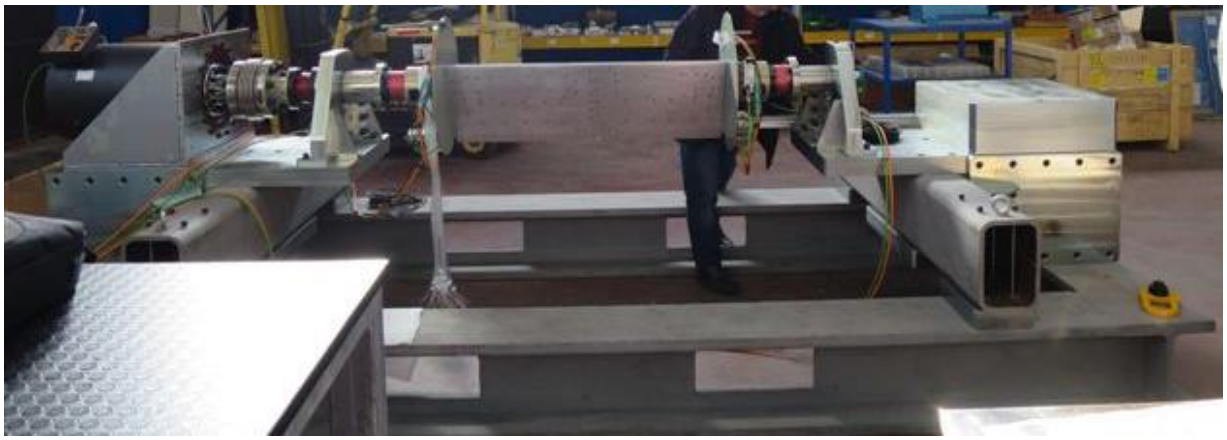


**Figure 1**: P*itch-oscillating system*

The system is integrated in the Secondary Test Section (STS) of the CIRA Icing Wind Tunnel (IWT). Officially inaugurated in September 2002, the IWT is one of the most advanced ground facilities for aerodynamic tests in icing conditions. The IWT is a closed loop-circuit, refrigerated, pressurized aerodynamic and icing wind tunnel equipped with four interchangeable test sections able to reproduce the atmospheric conditions up to a height of 7000 m and static temperature of -40°C. The test section can have both closed or slotted walls with 7% open area. Airflow refrigeration is obtained via a twin row heat exchanger; the minimum achievable static temperature in the STS is –40°C. In addition, an evacuation/pressurization air system allows static pressure being regulated between 39,000 Pa, corresponding to an altitude of about 7,000 m, and 145,000 Pa, for high Reynolds number aerodynamic tests.

The Pitching oscillating test rig is interfaced to the STS-Secondary Test Section frame (height 2.35m, width 1.15m, length 5.0m) allowing tests up to M=0.7. Figure 2 shows a general view of the IWT facility and of the STS. The details of the IWT facility and its operative envelope are provided in [15]**.**

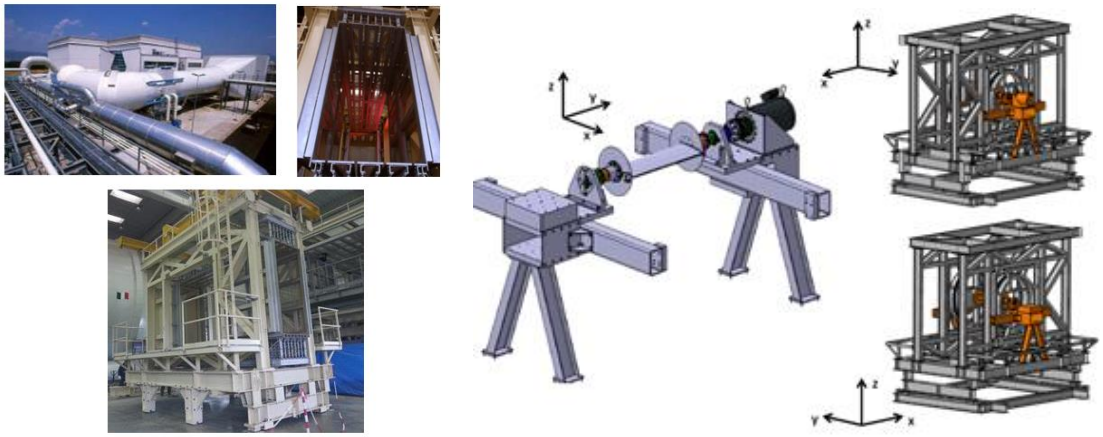


**Figura 2:** General View of the IWT (up/left and down) and view of the STS (right)

2D pitch oscillating tests will be performed on a NACA0018 airfoil having a chord length of 0.28m having a wing span length of 1.15m to be installed in wind tunnel in the "wall to wall" configuration. The 2D WT model has been designed to meet an adequate structural integrity under aerodynamic and inertial loads as well as model modularity to allow the setting of the different configurations. An overview of the 2D model is reported in Figure 3. The airfoil has been designed to guarantee sufficient access to the instrumentation installed inside the model, by minimizing the number of removable components and mitigating the risk of steps, gaps and/or imperfections on external skin, which can affect the natural flow development. At this regard, removable panels have been manufactured on the lower part of the surface, by leaving the upper surface of the model clean. The model has been designed to allow different configurations:

- Clean configuration (absence of GF);
- No. 2 model configurations having a fixed GF with 1.4 % and 2.8 % of protrusion length respectively, with GF located on the lower surface at 95% of the model chord.

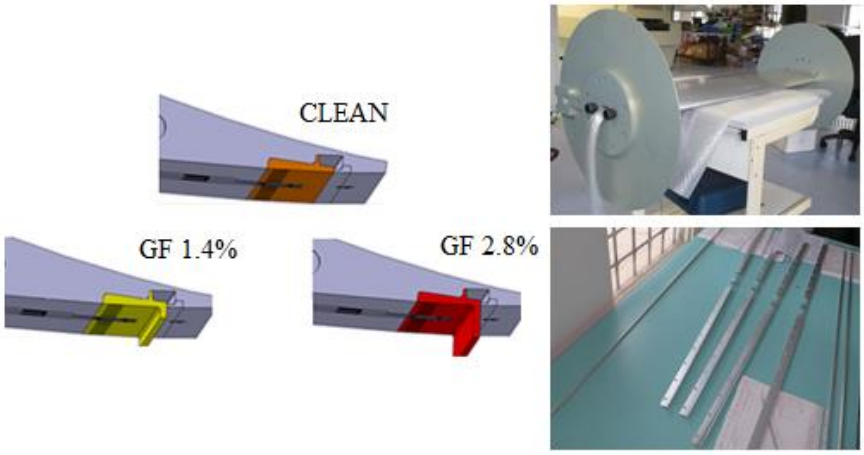



**Figure 3:** 2D Wind Tunnel Model Configurations

## 3 NUMERICAL SIMULATIONS

This section is devoted to the numerical study of the aerodynamic behavior of a NACA 0018 airfoil without and with a fixed Gurney Flap. First of all a description of the geometrical data and operating conditions is presented; then steady and unsteady results are detailed.

### 3.1 Geometrical data and selected operating conditions

The analyzed blade section is represented by a NACA 0018 airfoil (see [16] for coordinates set). For the present calculations, a blade section chord of 0.28 meters is considered and the airfoil angle of incidence is time-modulated according to the following harmonic oscillation law

$$\boldsymbol{\alpha = \alpha_M - \alpha_0 \sin(2\pi f t)}$$

where f is the frequency of pitch oscillations.
Computations include both steady and unsteady simulations according to Table 1, which represents the wind tunnel test matrix selected for cross-comparison of numerical predictions against experimental results.

| ***Blade Motion*** | **Mach Number** | **Reynolds Number** | $\boldsymbol{\alpha_M}$ | $\boldsymbol{\alpha_0}$ | **f** | **AGF** |
|---|---|---|---|---|---|---|
| **Fixed Airfoil** | 0.3 | $1 \times 10^6$ | $[-2° \rightarrow 20°]$ | 0 | N/A | OFF |
| | | | | | | FIXED (1.4-2.8%) |
| **Rotating Airfoil** | 0.3 | $1 \times 10^6$ | 8° | 6° | 3 Hz | OFF |
| | | | | | | FIXED (1.4-2.8%) |

**Table 1:** Summary of numerical simulations

The Gurney Flap device is positioned on the lower side of the airfoil, at 95% of the chord, and it reaches a maximum protrusion length equal to 1.4% and 2.8% of the chord (i.e. 4 or 8 mm). Unsteady simulations are also performed for the oscillating airfoil with Gurney Flap device fully extended during the whole pitch cycle (fixed protrusion length).

### 3.2 Domain discretization and boundary conditions

The fluid domain has a typical C-shape with outer boundaries located 30 chord length away from the body surface. The minimum spacing in the normal direction to the airfoil surface is set to an adequate value to solve the boundary layer down to the airfoil surface (unitary dimensionless wall distance with growth ratio around 1). Figure 4 shows a sketch of the leading edge grid on the airfoil.

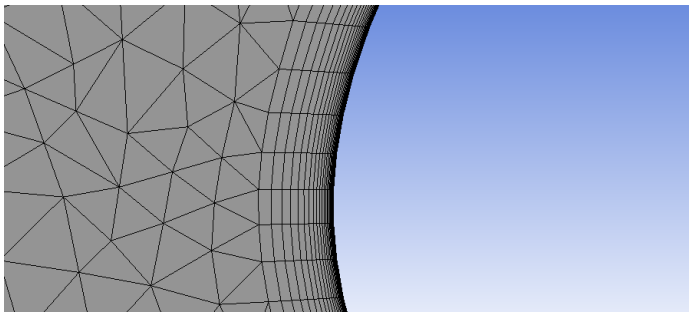


**Figure 4:** Mesh close–up view

All flow simulations listed in Table 1 use the same computational mesh and boundary conditions: cases with extended Gurney Flap (fixed length) are computed applying a solid wall boundary condition to both faces lying on the GF surface.

### 3.3 Flow solver and CFD approach

The Reynolds–Averaged Navier-Stokes equations are solved with a finite volume method (FV), so they are expressed as a system of conservation laws that relate rate of mass change, momentum, and energy in a control volume of area A to the spatial fluxes of these quantities through the volume. The flow field is computed using the commercial CFD code ANSYS Fluent R15.0. The presented computations solve URANS equations using a k-ω SST turbulence closure (see [18]) for all cases. According to the wind-tunnel model set-up, CFD simulations were run as fully turbulent, i.e. with no transition model and no regions with artificially modified turbulence production.

The novel element of this work is the application of the dynamic meshing methodology to obtain accurate time-dependent results because the computational mesh varies in time consistently with the changing positions of the different moving bodies. This dynamic approach allows the numerical solution to more closely follow the unsteady flow evolution and, thus, to better represent the time history of the forces acting on the bodies surfaces (see [19]). In order to avoid the deterioration of the mesh quality and/or the degeneration of existing FV cells, due to the geometry modification, two different methods are used, which are referred to as “smoothing” and “remeshing”. The former technique consists in moving the interior nodes of the mesh without changing their number and their connectivity. The latter technique allows for the local update of the mesh by either adding or deleting cells, where the boundary displacement would be

otherwise too large with respect to the local mesh size. After some preliminary tests demonstrating the feasibility of such an approach, the remeshing and the spring–based smoothing techniques are simultaneously used in the present work.

### 3.4 Results

A preliminary characterization of the airfoil aerodynamic performances at fixed incidence angles (fixed airfoil conditions) is performed to compare our numerical predictions with both available data from other experimental and numerical set-up and, so, to provide reference data for unsteady computations. Figure 5 depicts the trend of the lift coefficient against the angle of attack: our data (black curve) are compared with [20], [21] and [22]. The agreement between our results and the experimental data is good, especially in the linear part of the curve.

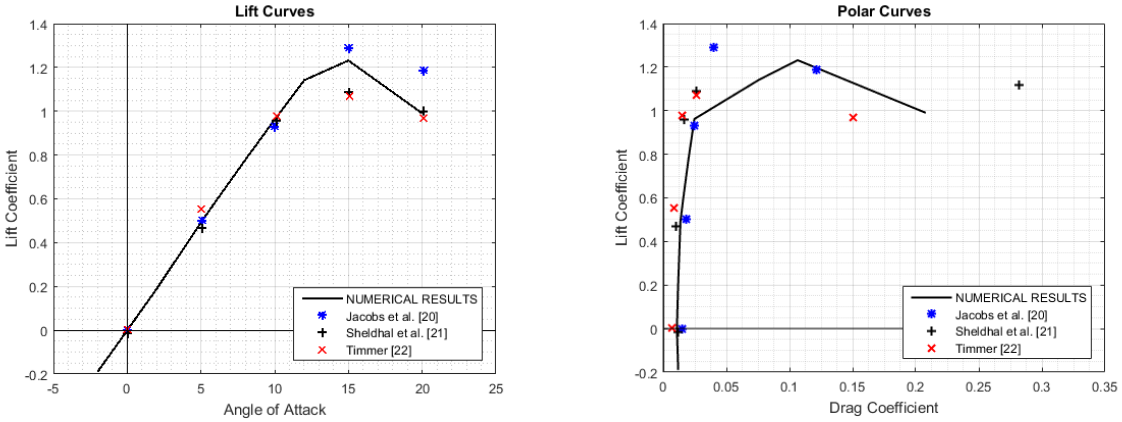



**Figura 5:** Lift and Polar Curves

Figures 7 and 8 refer to the cases in which the Gurney Flap assumes fixed positions. Watching the $C_l(\alpha)$ curve (Figure 6), it is noticeable, respectively, an increasing of 6% and 8% of the $C_l$ for the model with the 1.4% and 2.8% fixed Gurney flap in comparison with the clean configuration. Moreover, the stall angle diminishes from 15 deg to 12 deg. The $C_l(\alpha)$ slope doesn't show significant changes in comparison with the clean airfoil.

Comparing the $C_d(\alpha)$ curves (see Figure 7), a section drag increment is observed along the curve being more important in the range of high angles of attack. There is also a 10% increment

of the $C_{d0}$ in comparison with the clean airfoil. All of these conclusions have a good agreement with other papers focused on the same topic (see [23], [24]).

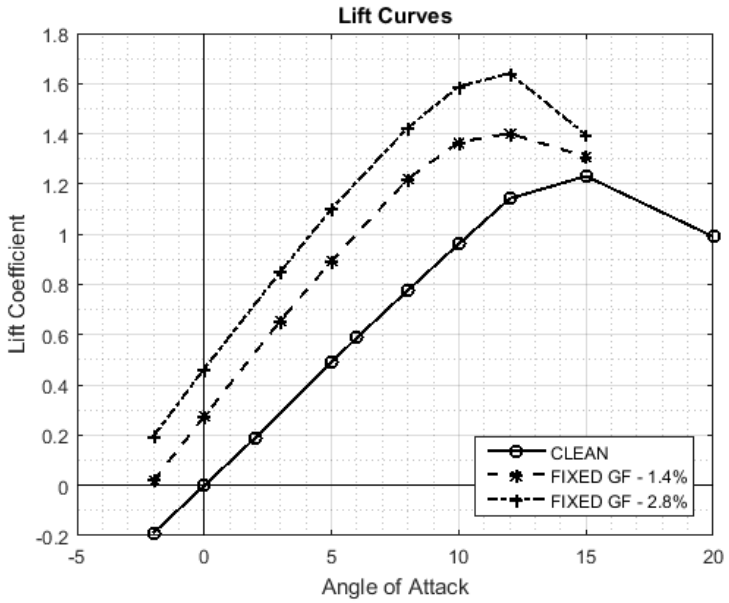



**Figure 6:** Lift curves

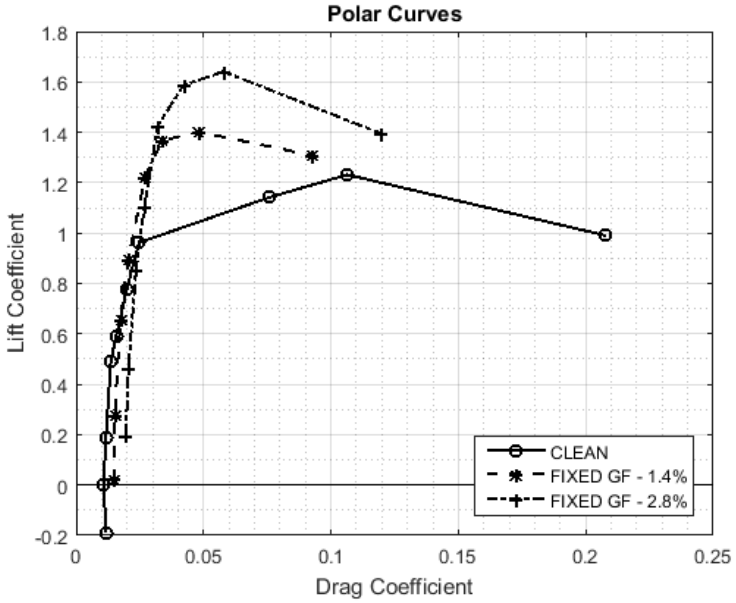



**Figure 7:** Polar curves

Figure 8 shows the trend of the pressure coefficient at $C_l = 0.78$. The main difference is the presence of a plateau of pressure in the rear part of the airfoil after the GF. Moreover the presence of small humps in the pressure coefficient for all configurations denotes the existence of laminar bubbles on the upper side of the airfoil. This fact is also confirmed by data presented in literature ([16]).

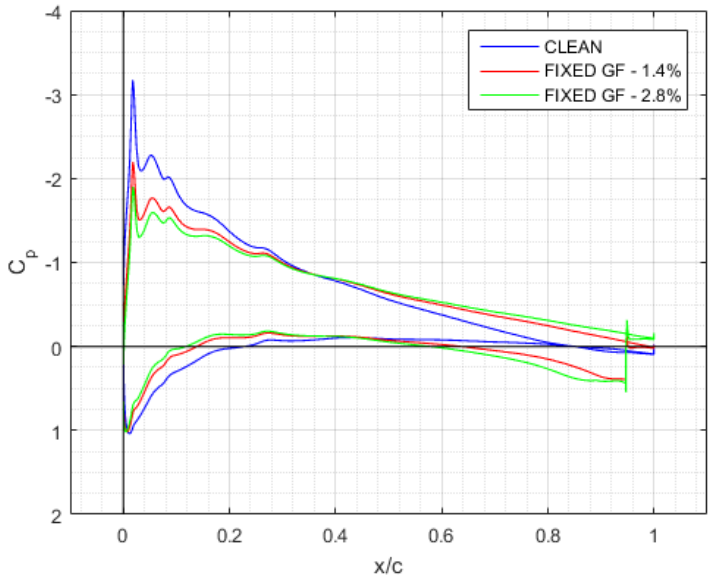



**Figura 8:** Pressure Coefficient ($C_l = 0.78$)

Figure 9 shows the comparison between the oscillating case and the corresponding steady case, for the configuration with 1.4% fixed Gurney Flap. As we expect, the maximum lift coefficient in the oscillating case is greater than that of the fixed case. In this case the Light Dynamic Stall is exploring in which the maximum angle of oscillation is not so far from the steady one.

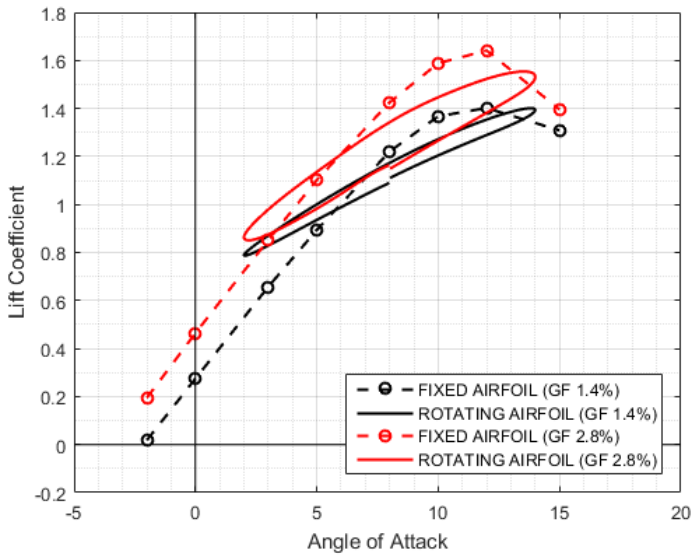



**Figure 9:** Lift Curves

## 4 CONCLUSIONS

This work addresses the characterization of the aerodynamics of the Gurney Flap from the experimental and numerical point of view. The first part describes the experimental apparatus to study Dynamic Stall conditions acting on a full-scale retreating rotor blade section referred to a medium size helicopter. In the second part, numerical results are performed in both static and dynamic configurations: all of them have demonstrated the capability of GF devices in increasing the maximum lift coefficient and, so, alleviating some negative aspects of the dynamic stall. Dynamic simulations are focused on the Light Dynamic Stall problem, in which the maximum angle of oscillation is not so far from the steady one.

### 4.1 Future works

A future extension of this work will contain the results of the experimental tests into two different conditions: clean and with fixed GF configurations. These results will be compared with the numerical simulations of this paper. In addition to the Light Dynamic Stall, next calculations will be focused on the characterization of the Deep Dynamic Stall in which the maximum angle achieved during the oscillation is much higher with respect to the Light Stall. Simulations will be, also, performed on an oscillating airfoil having an active Gurney Flap.

Future developments regarding the experimental apparatus will be the integration of an active Gurney Flap device in the pitching oscillating test rig.

## REFERENCES

[1] F. O. Carta, Experimental Investigation of the Unsteady Aerodynamic Characteristics of a NACA 0012 Airfoil, *UAC RL Report M-1283-1*, July 26, 1960.

[2] F. D. Harris, R. R. Pruyn, Blade Stall – Half Fact, Half Fiction. *J. Amer. Helicopter Soc*., Vol. 13, No. 2, 1968, pp. 27-48.

[3] J. Liiva, F.J. Davenport, L. Gray, I. C. Walton, Two-Dimensional Tests of Airfoils Oscillating Near Stall, *U.S. Army AVLABS, TR68-13*, 1968.

[4] L. Gray, J. Liiva, F. J. Davenport, Wind-Tunnel Tests of Thin Airfoils Oscillating Near Stall. *U.S. Army AVLABS, TR68-89A*, 1969.

[5] R. I. Windsor: Measurement of Aerodynamic Forces on an Oscillating Airfoil, *U. S. Army AVLABS TR69-98*, 1970.

[6] M. S. Garelick, Non-Steady Airloads on Dynamically Stalling Two-Dimensional Wings. *Master's Degree Thesis*, MIT, 1967.

[7] N.D. Ham, M.S. Garrelick, Dynamic Stall Considerations in Helicopter Rotors. *J. Amer. Helicopter Soc.*, Vol. 13, 1968, pp. 49-55.

[8] McAlister, K. W., and McCroskey, W. J., Analysis of the Development of Dynamic Stall Based on Oscillating Airfoil Experiments, *NASA TN-8382*, Jan. 1977.

[9] Carr, L. W, Progress in Analysis and Prediction of Dynamic Stall, *Journal of Aircraft*, Vol. 25, No. 1, 1988, pp. 6-17.

[10] Carr, L. W., M. S. Chandrasekharat, Design and Development of a Compressible Dynamic Stall Facility, *Journal of Aircraft* Vol. 29, No. 3, May-June 1992

[11] H. Mai, G. Dietz, W. Geissler, et al. Dynamic stall control by leading edge vortex generators, *J. Am. Helicopter. Soc.*, 2008; 56: 26–36.

[12] W. Geissler and al., Dynamic Stall Control Investigations on a Full Size Chord Blade Section, *European Rotorcraft Forum*, 2004

[13] Dynamic Stall and its Passive Control Investigations on the OA209 Airfoil Section, *31th European Rotorcraft Forum*, 2014

[14] T. Reinert and al., Oscillating Airfoil Icing Tests in the NASA Glenn Research Center Icing Research Tunnel, *2011-38-0016 SAE International*, 2011

[15] "Cira Icing Wind Tunnel User Manual", www.cira.it.

[16] I. H. Abbott, A. E. von Doenhoff, Theory of Wing Sections: Including a Summary of Airfoil Data. *Dover Books*, 1945.

[17] J.H. Argyris, M. Papadrakakis, L. Karapitta, Elastoplastic analysis of shells with the triangular element TRIC. M. Papadrakakis, A. Samartin, E. Oñate eds. *4th International Colloquium on Computation of Shell and Spatial Structures (IASS-IACM 2000)*, Chania, Crete, Greece, June 4-7, 2000.

[18] F. R. Menter, Two Equation Eddy Viscosity Turbulence Models for Engineering Applications. *AIAA Journal*, **32**, 1598–1605, 1994.

[19] S. Russo, G. P. Reina, C. de Nicola. RANS prediction of the flap unsteady aerodynamics using dynamic mesh. *Memories XXV Congresso AIDAA Torino*, 17-19 November (2015).

[20] E. Jacobs, N. A. Sherman, Airfoil section characteristics as affected by variations of the Reynolds number. *NACA Report 586*, 1937

[21] R. E. Shedahl, and P. C Klimas, Aerodynamic characteristics of seven airfoil sections through 180 degrees angle of attack for use in aerodynamic analysis of vertical axis wind turbines, *Technical Report No. SAND80-2114*, 1981

[22] W.A. Timmer, Two-dimensional low-Reynolds number wind tunnel results for airfoil NACA 0018, *Wind Engineering*, 32( 6), pp. 525–537, 2008

[23] M. E. Camocardi, J. M. Di Leo, J. S. Delnero, J. L. Colman Lerner, Experimental Study Of A Naca 4412 Airfoil With Movable Gurney Flap, *49th AIAA Aerospace Sciences Meeting*, 4 - 7 January 2011, Orlando, Florida, 2011

[24] D. Tang, E. H. Dowell, Aerodynamic Loading for an Airfoil with an Oscillating Gurney Flap, *JOURNAL OF AIRCRAFT*, Vol. 44, N. 4, 2007

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Editors)
Crete Island, Greece, 5–10 June 2016

# INVESTIGATION OF A MODERN & HYBRID TURBULENCE MODELLING APPROACH FOR TRANSIENT AUTOMOTIVE AERODYNAMICS SIMULATIONS

**René Devaradja[1], Petr Šimánek[1], Jacques Papper[1], and Pavla Polická[2]**

[1] ICON Technology & Process Consulting SAS
7 rue Auber, 31000 Toulouse, France
e-mail: {r.devaradja,p.simanek,j.papper}@iconcfd.com

[2] ŠKODA AUTO a.s.
tř. Václava Klementa 869, 293 60 Mladá Boleslav, Czech Republic
pavla.policka@skoda-auto.cz

**Keywords:** Turbulence Modelling, Automotive Aerodynamics, Transient Simulation, PANS, Zeta-f, iconCFD®.

**Abstract.** *In order to achieve accurate transient predictions of turbulent flow fields on industrial scale problems, it is necessary to use a turbulence model since Direct Numerical Simulation (DNS) would be prohibitively expensive. Unsteady Reynolds-Averaged Navier Stokes (URANS) is the most commonly used to fully model the turbulence as an additional eddy viscosity using the Boussinesq approximation. Large Eddy Simulation (LES) is another modelling approach, which can be more accurate than URANS as only the smallest scales of turbulence are modelled. However, LES mesh requirements are very demanding, and too constraining for simple use in industrial contexts. Therefore, a variety of hybrid models have been developed, attempting to use the provided mesh to its maximum potential by resolving as many turbulent scales as possible and only model those eddies for which the mesh is insufficiently fine. Detached Eddy Simulation (DES), is one example of hybrid model, where URANS is used in the attached boundary layer regions whilst LES is used in the largely separated flow regions where the larger scales are dominant. More recently, the Partially-Averaged Navier Stokes (PANS) method introduced by Girimaji et. Al [3] provides an interesting framework allowing the modification of existing URANS models to resolve the larger scales of turbulence whilst modelling the smaller ones.*

*A validation of the PANS zeta-f model using iconCFD® is presented on a variety of cases ranging from cylinder flows to fully detailed vehicle aerodynamics courtesy of ŠKODA AUTO. Comparison to experimental wind-tunnel testing of the Yeti, Fabia and Superb models are presented. Various modifications to the modelled-to-resolved scale ratio are proposed and tested. Successful performance of this model is demonstrated within the industrial context of vehicle aerodynamics, and superiority of this model is shown in the prediction of separation regions, off-body flow structures, and aerodynamic forces.*

1. INTRODUCTION ..... 3
2. PANS $\zeta - f$ MODEL DESCRIPTION. ..... 4
2.1. PANS GOVERNING EQUATIONS ..... 4
2.2. PANS $\zeta - f$ CLOSURE MODEL ..... 5
2.3. NEAR-WALL TREATMENT ..... 6
2.4. MODEL IMPLEMENTATION ..... 7
3. PANS VALIDATION ..... 9
3.1. GENERAL WORKFLOW ..... 9
3.2. ACADEMIC CASES ..... 9
3.2.1. FLOW PAST A SQUARE CYLINDER ..... 9
3.2.2. FLOW PAST A SURFACE-MOUNTED CUBE ..... 16
3.3. AUTOMOTIVE EXTERNAL AERODYNAMICS ..... 19
4. SUMMARY ..... 25
5. ACKNOWLEDGEMENTS ..... 26

## 1. Introduction

Although URANS is generally an attractive turbulence modelling approach for industrial applications owing to its reasonable computational cost, on the other hand its lack of accuracy in the prediction of unsteady bluff body separation flows – which is the main focus of the present paper - raises the need for alternative solution methods. The problematic limitation of URANS resides in its inherent inability to simulate a broad spectrum of turbulent scales since all eddies are modelled. The resulting drawback is a simplified representation of the turbulence physics suffering from a shortfall in large scale characteristics that are influential in the evolution of the turbulent coherent structures. The Large Eddy Simulation (LES) approach is a suitable candidate to overcome this deficiency in accuracy as it directly resolves for the large turbulent scales, dominant in largely separated flows, and also simulates their interaction with the smaller sub-grid scales. This method allows to capture most of the flow details – if not all of them like in Direct Navier Stokes (DNS) - and to represent the overall flow complexity with a satisfactory physical fidelity. However in the near-wall regions, the largest turbulence scales are still very small making the LES approach computationally too expensive. Therefore from an engineering perspective, the LES technique may not be an adequate choice if the turbulent structures of interest are essentially in the outer wall regions, which is typically the case for ground vehicles aerodynamics. Yet, the motivation to apply LES far away from the walls, and to combine it with an affordable but accurate-enough modelling approach valid in the wall vicinity, has led to the development of various and successful hybrid LES/RANS methods as described in [1]. One of the main addressed challenges for these hybrid approaches is the capacity to enable LES in the mesh regions which are fine enough for the local scales to be resolved and bridge it, in a seamless way, with RANS which is invoked in the other regions that do not satisfy the former criterion. In the same spirit of hybrid LES/RANS, a new emerging modelling approach, classified as "second generation URANS", is intended to resolve the energy-containing scales at reduced computational cost. It is called "Partially-Averaged Navier Stokes" (PANS) and was first introduced by Girimaji [2] as a variant of the standard $k-\varepsilon$ closure model. The key feature of this approach is based on the introduction of two new parameters, denoted by $f_k$ and $f_\varepsilon$, which control the relative amount (with respect to the total quantity) of the unresolved turbulent kinetic energy and dissipation rate respectively. Along with appropriate modifications of the closure coefficients, these modelled-to-total scale ratios allow to set the physical resolution of the turbulent fluctuations to any level ranging from pure URANS ($f_{k,\varepsilon}=1$) to pure DNS ($f_{k,\varepsilon}=0$). The mesh resolution to be employed will directly depend on the $f_{k,\varepsilon}$ value if the latter is prescribed as a constant parameter. To remove that dependency, Girimaji [3] suggested a spatially-varying formulation of $f_{k,\varepsilon}$, based upon the local grid dimensions, so that the PANS model can self-adjust its performance to its maximum resolution potential on any grid. In this paper, the PANS $\zeta-f$ model [4] will be retained for further investigations, within the iconCFD® framework, as it should theoretically show a better predictive capability for the near-wall flows in comparison to the original PANS $k-\varepsilon$

formulation. The implementation guidelines of the PANS model, and the different viable options for the $f_{k,\varepsilon}$ computation, are provided and discussed. The model is then evaluated for a set of different bluff body transient flows ranging from the academic square cylinder case to an external aerodynamics simulation on a ŠKODA car model. Finally, the validity of the new eddy-resolving method is assessed through qualitative analysis of turbulent structures and comparison of the numerical results with available experimental data.

## 2. PANS $\zeta - f$ model description.

### 2.1. PANS governing equations

The general PANS formulation is derived from the incompressible Navier-Stokes equations by applying an arbitrary filter, denoted by $< \cdots >$, to the instantaneous velocity $\boldsymbol{V}$ and pressure $p$ fields**.** The resulting filtered equations [2] read:

$$\partial_t U_i + U_j \partial_j U_i + \partial_j \tau_{ij} = \frac{1}{\rho} \partial_i p_f + \nu \partial_{jj}^2 U_i$$
$$\partial_{ii}^2 p_f = \partial_j U_i \partial_i U_j + \partial_{jj}^2 \tau_{ij}$$

Where $U_i =< V_i >$ and $p_f =< p >$ are the filtered velocity and pressure fields respectively. The sub-filter scale stress $\tau_{ij}$, assimilated to the Reynolds stress [5], can be calculated using the Boussinesq approximation:

$$\tau_{ij} = -\nu_u S_{ij}$$
$$S_{ij} = \frac{1}{2}\left(\partial_j U_i + \partial_i U_j\right)$$

The unknown variable $\nu_u$, which represents the eddy viscosity of the unresolved scales is then determined by any turbulence closure model. For models of the $k - \epsilon$ family, it is generally expressed as a function of unresolved turbulent kinetic energy $k_u$ and dissipation rate $\epsilon_u$:

$$\nu_u = f(k_u, \epsilon_u)$$

The relation between the unresolved (sub-filtered) and averaged (filtered) turbulence is established in the following way:

$$f_k = \frac{k_u}{k_{tot}} = \frac{k_u}{k_u + 0.5\overline{(\boldsymbol{U} - \bar{\boldsymbol{U}})^2}}$$

$$f_\epsilon = \frac{\epsilon_u}{\epsilon_{tot}} \approx 1$$

The parameters $f_k$ and $f_\epsilon$ represent the unresolved-to-total ratios of the turbulent kinetic energy and dissipation rate respectively. Inserting these into any existing URANS closure model allows its adaptation to the PANS framework and controls the level of physical resolution of the turbulence fluctuations ($\boldsymbol{U} - \overline{\boldsymbol{U}}$), with $\overline{\boldsymbol{U}}$ referring to the mean statistical velocity field. A unit value of $f_k$ means that the filtered fluctuations are zero. In other terms, the filtered velocity is equal to its time-averaged value and the PANS model completely switches to the standard URANS formulation. If $f_k = 0$, then $k_u = \nu_u = \tau_{ij} = 0$ and the PANS equations collapse to the DNS ones as the filtered and instantaneous velocities are the same. For intermediate $f_k$ values, the filtered velocity can be called "partially averaged" as the contribution of both resolved and unresolved turbulent fluctuations is non-zero. Finally, the parameter $f_\epsilon$ can be set to 1 [3] – at least for high Reynolds number flows - since the very small dissipation scales, which are isolated from the larger energy-containing ones, are not intended to be resolved.

### 2.2. PANS $\zeta - f$ closure model

Here the $\zeta - f$ model, suggested by Basara [4], is used to close the PANS equations. This model is derived from the URANS $\zeta - f$ model, introduced by Hanjalic [6], and which is known to deal better than its parent, the standard $k - \epsilon$ model, with anisotropic and curvature effects encountered in some wall-bounded turbulent flows. The equations of PANS $\zeta - f$ read:

$$\nu_u = C_\mu \zeta_\mu \frac{{k_u}^2}{\epsilon_u}$$

$$\partial_t k_u + U_j \partial_j k_u = P_u - \epsilon_u + \partial_j \left[ \left( \nu + \frac{\nu_u}{\sigma_{ku}} \right) \right] \partial_j k_u$$

$$\partial_t \epsilon_u + U_j \partial_j \epsilon = C_{\epsilon 1} P_u \frac{\epsilon_u}{k_u} - C_{\epsilon 2}^* \frac{\epsilon_u^2}{k_u} + \partial_j \left[ \left( \nu + \frac{\nu_u}{\sigma_{\epsilon u}} \right) \partial_j \epsilon_u \right]$$

$$\partial_t \zeta_u + U_j \partial_j \zeta_u = f_u - \frac{\zeta_u}{k_u} \epsilon_u (1 - f_k) + \partial_j \left[ \left( \nu + \frac{\nu_u}{\sigma_{\zeta u}} \right) \partial_j \zeta_u \right]$$

$$L_u^2 \nabla^2 \mathrm{f_u} - \mathrm{f_u} = \frac{1}{T_u} \left( C_{f1} + C_{f2} \frac{P_u}{\epsilon_u} \right) \left( \zeta_u - \frac{2}{3} \right)$$

The different terms from the previous equations are: $k_u$ - unresolved turbulent kinetic energy; $\epsilon_u$ - dissipation rate of $k_u$; $P_u = -\tau_{ik}\, \partial_k U_i$ - production rate of $k_u$; $\zeta_u$ - unresolved wall normal velocity scale; $f_u$ - elliptic relaxation function; $L_u = C_L \max \left[ \frac{k_u^{1.5}}{\epsilon}, C_\eta \left( \frac{\nu^3}{\epsilon} \right)^{0.25} \right]$ - turbulent length scale; $T_u = \max \left[ \frac{k_u}{\epsilon}, C_\tau \left( \frac{\nu}{\epsilon} \right)^{0.5} \right]$ - turbulent time scale. The closure coefficients are given in Table 1.

| $C_\mu$ | $C_{\epsilon 1}$ | $C^*_{\epsilon 2}$ | $C_{\epsilon 2}$ | $C_{f1}$ | $C_{f2}$ |
|---|---|---|---|---|---|
| 0.22 | $1.4\left(1+\frac{0.045}{\sqrt{\zeta_u}}\right)$ | $C_{\epsilon 1}+f_k(C_{\epsilon 2}-C_{\epsilon 1})$ | 1.9 | 0.4 | 0.65 |
| $\sigma_{ku}$ | $\sigma_{\epsilon u}$ | $\sigma_{\zeta u}$ | $\sigma_k$ | $\sigma_\epsilon$ | $\sigma_\zeta$ |
| $\sigma_k\frac{f_k^2}{f_\epsilon}$ | $\sigma_\epsilon\frac{f_k^2}{f_\epsilon}$ | $\sigma_\zeta\frac{f_k^2}{f_\epsilon}$ | 1.0 | 1.3 | 1.2 |
| $C_L$ | $C_\eta$ | $C_\tau$ | | | |
| 0.36 | 85 | 6.0 | | | |

**Table 1 Closure coefficients of PANS $\zeta-f$**

If $f_k=1$, then any PANS model should, in theory, be identical to its parent URANS version. Here a few differences can be noted between PANS $\zeta-f$ and URANS $\zeta-f$: (a) in the $\epsilon$ transport equation, the isotropic turbulence length scale $\frac{k}{\epsilon}$ is used instead of $T$; (b) the realizability constraints for $T$ and $L$ are removed; (c) the definition of $C_{\epsilon 1}$ differs from the original URANS one $C_{\epsilon 1}=1.4\left(1+\frac{0.012}{\zeta_u}\right)$.

Finally, the zero transport model [2] is used by Basara to calculate the unresolved turbulent Prandtl numbers ( $\sigma_{ku},\sigma_{\epsilon u},\sigma_{\zeta u}$ ). The maximum transport assumption [2] will be also considered in this paper. According to that model, the following assumptions can be made: $\sigma_{ku}=\sigma_k;\ \sigma_{\epsilon u}=\sigma_\epsilon;\ \sigma_{\zeta u}=\sigma_\zeta$.

### 2.3. Near-wall treatment

The $\zeta-f$ model is a low Reynolds model which requires, by its nature, a fine wall-normal mesh discretization in order to resolve the whole boundary layer. Thus the required $y^+$value for such a model is less than 1 to accurately predict the near-wall flow. Because these mesh-related constraints are not easy to satisfy in an industrial context, a hybrid wall treatment which combines the integration to the wall (ITW) with generalised wall functions (GWF), is incorporated into the PANS turbulence model. This method allows to cover all $y^+$ values, even those falling into the critical buffer region $(5<y^+<30)$. In this study, the compound wall treatment (CWT) approach, proposed by Popovac et al. [7], is adopted to define the wall boundary conditions of the turbulent variables. The general form of CWT, for any flow variable $\phi$, is given by:

$$\phi_P=\phi_v e^{-\Gamma}+\phi_t e^{-\frac{1}{\Gamma}}$$

The index "$P$" denotes the centre of the wall-adjacent cell. The indices "$v$" and "$t$" respectively stand for the viscous and turbulent contributions of $\phi$. Finally, $\Gamma$ represents the Kader's blending coefficient. The previous decomposition is applied to the fields $P_u,\epsilon_u,\ f_u$ and $\nu_u$ as

reported in Table 2. The different terms involved in the near-wall treatment are: $\nu_l$ - the laminar viscosity; $y$ - the wall distance; $y_P^* = \frac{C_{\mu w}^{0.25} k_P^{0.5} y_P}{\nu_l}$ ; $C_{\mu w} = 0.07$ - the non-dimensional wall distance; $u^* = C_{\mu w}^{0.25} k^{0.5}$ - the turbulent velocity scale; $\kappa = 0.41$ - the Von Karman constant; $u_\tau = \sqrt{(\nu_l + \nu_{u,P}) \frac{U_P}{y_P}}$ - the friction velocity; $E = 8.3$ - the log-law constant.

| $\boldsymbol{\phi}$ | $\boldsymbol{\phi_\nu}$ | $\boldsymbol{\phi_t}$ | $\boldsymbol{\Gamma}$ |
|---|---|---|---|
| $\epsilon_u$ | $\frac{2\nu_l k_u}{y_P^2}$ | $\frac{u_P^{*\,3}}{\kappa y_P}$ | $0.001 \frac{y_P^{*\,4}}{1 + y_P^*}$ |
| $P_u$ | $C_\mu \zeta_{u,P} \frac{k_{u,P}^2}{\epsilon_{u,P}} \left(\frac{U_P}{y_P}\right)^2$ | $\frac{u^{*2} u_\tau}{\kappa y_P}$ | $0.01 \frac{y_P^*}{1 + 5y_P^*}$ |
| $\nu_u$ | $\nu_l$ | $\frac{\kappa u_p^* y_P}{\ln(E y_P^*)}$ | |
| $f_u$ | $-\frac{2\nu_l \zeta_P}{y_P^2}$ | 0 | 0 |

**Table 2 CWT formulation for PANS $\zeta - f$**

2.4. Model implementation

The PANS $\zeta - f$ model along with the enhanced CWT method has been implemented in iconCFD® which is a finite volume CFD code. In this section, the different possible ways of modelling the PANS key parameter $f_k$ are presented and discussed.

The easiest option would be to treat $f_k$ as a user-defined constant coefficient. However this method is impractical for two main reasons. First, in a case where the turbulent flow structures locations are unknown in advance, the engineer will not know where to refine the mesh. Secondly, in some areas of the flow it may be easy to refine the mesh enough to satisfy the prescribed $f_k$ value while in other regions such as boundary layers the required refinement would become prohibitive. Alleviating these constraints, Girmaji [3] suggested another approach by which the $f_k$ variable self-adjusts its value based on the local grid dimensions. This new method, which is more suitable for an industrial process, allows the PANS model to be used at its maximum resolution potential on an arbitrary mesh. The spatially dynamic formulation of $f_k$ is given by:

$$f_{k,min}(x) = \frac{1}{\sqrt{C_\mu}} \left(\frac{\Delta}{\Lambda}\right)^{\frac{2}{3}}$$

The subscript "$min$" indicates that there is not any strict restriction to apply a larger value than the right-hand term of the above equality. The local grid length scale is denoted by $\Delta$. The variable $\Lambda = \frac{k_{tot}^{1.5}}{\epsilon_u}$ represents the Taylor scale of turbulence.

The only difficulty at this stage is to determine an accurate way to calculate $k_{tot}$ during the PANS simulation. In order to achieve this, it is necessary to know the time-averaged velocity field in advance, or in other terms to estimate that missing field by means of a pre-computed flow solution. Four different methods to estimate $k_{tot}$ are suggested hereafter. For all of them, it is assumed that a converged RANS $\zeta - f$ solution has been obtained beforehand and used for the initialisation of the transient simulation.

**Method 1 (GB1)** - At each time step $n$, the value of $k_{tot}$ is updated from the solution at the previous iteration $n-1$, following the definition of $f_k$:

$$k_{tot}^{n} = \frac{k_u^{n-1}}{f_k^{n-1}}$$

The initial value of $f_k$ is 1.

**Method 2 (GB2)** – The formula of $k_{tot}$ is simplified by removing the resolved contribution so that at each time step $n$:

$$k_{tot}^{n} = k_u^{n-1}$$

This simplification is acceptable as it just overestimates the value of $f_k^n$ but the counter-part of this assumption is that some turbulent scales that could have been physically resolved by the grid - with a more correct estimation of $f_k^n$ - will end up being modelled.

**Method 3 (GB3)** – At each time step $n$, $k_{tot}$ is estimated from the filtered velocity fluctuations:

$$k_{tot}^{n} = k_u^{n-1} + 0.5(\boldsymbol{U^n} - \overline{\boldsymbol{U}}^n)^2$$

The time-averaged velocity $\overline{\boldsymbol{U}}$ is first initialised with the steady RANS velocity field and then updated in the course of the simulation as $\overline{\boldsymbol{U}}^n$, the running average of $\boldsymbol{U}$.

**Method 4 (GB4)** – At each time step $n$, $k_{tot}$ is estimated using the real definition of the resolved turbulent kinetic energy:

$$k_{tot}^{n} = k_u^{n-1} + 0.5\left(\overline{u_i u_i}^{n-1}\right)$$

The time-averaged resolved fluctuations $\overline{u_i u_i}$ are initialised to 0 and then updated in the course of the simulation as $\overline{u_i u_i}^n$, the running average of $u_i u_i$.

After $f_k$ is computed, a few precautions need to be taken with respect to that variable. First, the near-wall cell field should be explicitly set to 1 in case it does not converge to that value. This insures, in the DES spirit, that the turbulent scales in the extreme wall vicinity will be all modelled by URANS. Secondly, the maximum allowed value for $f_k$ must not exceed 1 and its minimum value can be set to any user-defined value (as low as possible) in order to help stabilize the transient solution.

## 3. PANS Validation

### 3.1. General Workflow

The numerical simulations are performed with the CFD suite iconCFD®. The computational grids are created with the included automatic unstructured, hex-dominant mesh generator iconHexMesh. The flow is simulated using a two-step process in order to obtain a proper initial estimate of $f_k$ for the PANS calculation. First, a steady flow solution is computed using the RANS $\zeta - f$ model. For this preliminary computation, all the convection terms are approximated by a first-order upwind scheme to ensure solution stability. Once the steady solution is obtained, it is then used for the initialisation of the PANS simulation. For this second calculation phase, a second-order upwind convection scheme is applied to the momentum equation only. The maximum transport model is enabled for the calculation of the turbulent diffusion coefficients. The time step is automatically adjusted by the solver during the transient simulation so that the maximum value of the CFL number does not exceed 1.0.

### 3.2. Academic Cases

In this section, a primary verification study will be carried out to check the PANS performance in transient bluff-body flows that are commonly used as base benchmarks for LES simulations. In the following sections, the modelled flow is assumed to be incompressible and dependent on the Reynolds number only. The fluid density and kinematic viscosity are arbitrarily set to $1\ kg.m^{-3}$ and $1.5e^{-05}\ m^2.s^{-1}$ respectively. The reference velocity is then deduced from the characteristic length scale of the problem.

#### 3.2.1. Flow past a square cylinder

The different $f_k$ implementations should be initially evaluated in a simple turbulent flow with negligible wall interactions like the one past a square cylinder. The coherent vortical shedding developing in this configuration has been accurately analysed in many research works such as the experimental study conducted by Lyn et al. [8], [9] or the LES simulations performed by Rodi et al. [10]. Reproducing these well-known turbulence characteristics with the PANS model would be a first critical success factor.

## Case Description

The Reynolds number for the concerned flow, based on the square cylinder side $D$, is 21400. This gives a reference bulk velocity of 0.3 $m.s^{-1}$, if $D$ is set to $1m$. The geometric dimensions of the computational domain, are shown in Figure 1. Among the available reference data, those which are retained for comparative analysis are the Strouhal number ($Str$), the mean drag coefficient ($C_D$) and the mean stream-wise velocity profile on the centre plane of the cylinder. Further details with respect to the experimental protocol and data can be found on the ERCOFTAC website [12].

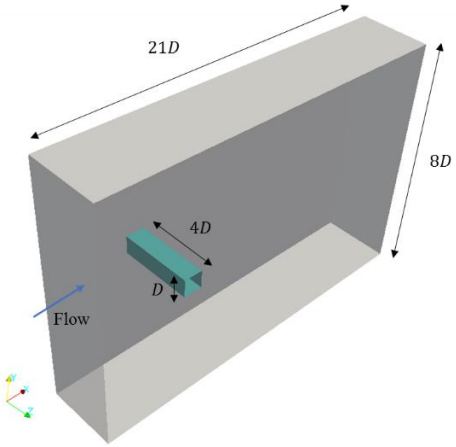



**Figure 1 Computational domain and geometry of the square cylinder**

## Computational Details

*Mesh* – The computational grid used for the numerical simulations contains 830000 cells. The mesh refinement is applied in the wake regions, as illustrated in Figure 2. The near-wall flow region of the cylinder is meshed with 5 layers and the $y^+$ surface-averaged value is around 3.

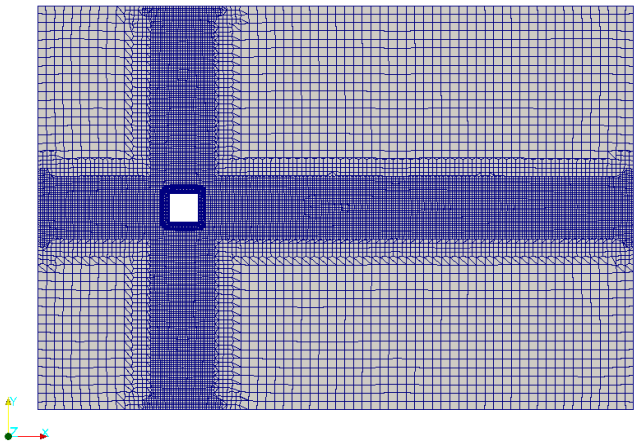


**Figure 2 Coarse mesh view on the centre plane of the square cylinder**

*Boundary Conditions* – The square cylinder surface is modelled as a non-slip wall while the upper and lower boundaries are applied a zero-friction condition. The side walls of the flow domain are treated as symmetry planes.

*Solver* - The transient simulation is carried out over 8 flow cycles - 1 flow cycle corresponds to the travelling time of the fluid from the inlet to the outlet of the computational domain. The flow fields are time-averaged over the last two flow cycles.

**Results**

*Constant $f_k$* - The second invariant of the velocity gradient, denoted by $Q$, is commonly used to identify the coherent structures of a transient turbulent flow. In fact, iso-surfaces of $Q > 0$ represent regions of high vorticity where the rotation rate is higher than the strain rate. The effect of the modelled-to-total scale ratio on these iso-surfaces is shown in Figure 3. Setting $f_k$ to a reasonable constant value of 0.4 over the whole computational domain results in the appearance of new vortical structures that do not exist in the URANS simulation ($f_k = 1$). The difference in the resolution of turbulence scales between PANS and URANS clearly affects the mean axial velocity profile as illustrated in Figure 4. Superiority of PANS over URANS is quantitatively demonstrated in that plot.

*Grid-based $f_k$*- The different calculation methods of the $f_k$ parameter are compared in Figure 5 and Figure 6. An overestimation of $f_k$ can be observed for GB1 and GB2. As a result, most of the turbulence characteristics obtained with these methods are modelled. On the other hand, GB3 and GB4 show more resolved scales in the wake region and the corresponding velocity profiles are closer to the experiment. The mean drag coefficient and the Strouhal number for the different cases are all reported in Table 3.

*Convection scheme* – The usage of a blended central differences-upwind convection scheme for the momentum equation sometimes helps stabilize a non-converging transient solution. Nevertheless for a PANS simulation, the contribution of the first-order scheme can have an inhibiting effect on the resolution of critical turbulence scales. Figure 7 and Figure 8 show that despite low $f_k$ values in the wake and apparent multi-scale vortices, the actual PANS solution is very similar to the URANS one.

*Unresolved turbulent Prandtl numbers* – The zero and maximum transport models are compared in Figure 9 and Figure 10. The difference in the spatial distribution of $f_k$ can be observed between the two cases. The zero transport model seems to be more diffusive as it produces lower $f_k$ values in the detached flow regions. This is probably due to a higher diffusion coefficient, resulting from a division by $f_k$ (ranging from 0 to 1), in the equation of the unresolved turbulent kinetic energy. Despite this difference, the overall flow structures and the mean axial velocity profile are similar for both transport models.

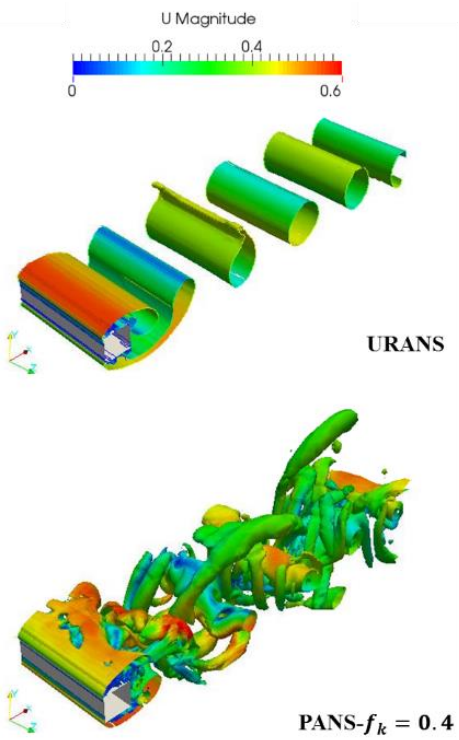



**Figure 3 Coarse mesh. Instantaneous iso-surfaces of the Q criterion ($Q = 0.01\ s^{-1}$)**

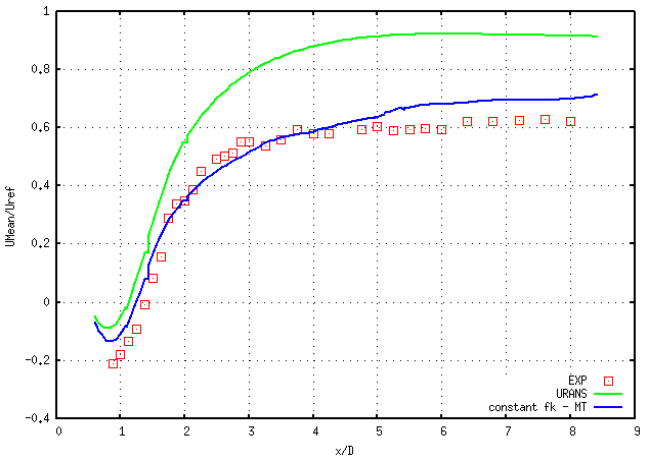



**Figure 4 Coarse mesh. Mean axial velocity profile - Experiment (red symbol), URANS (green), $f_k = 0.4$ (blue)**

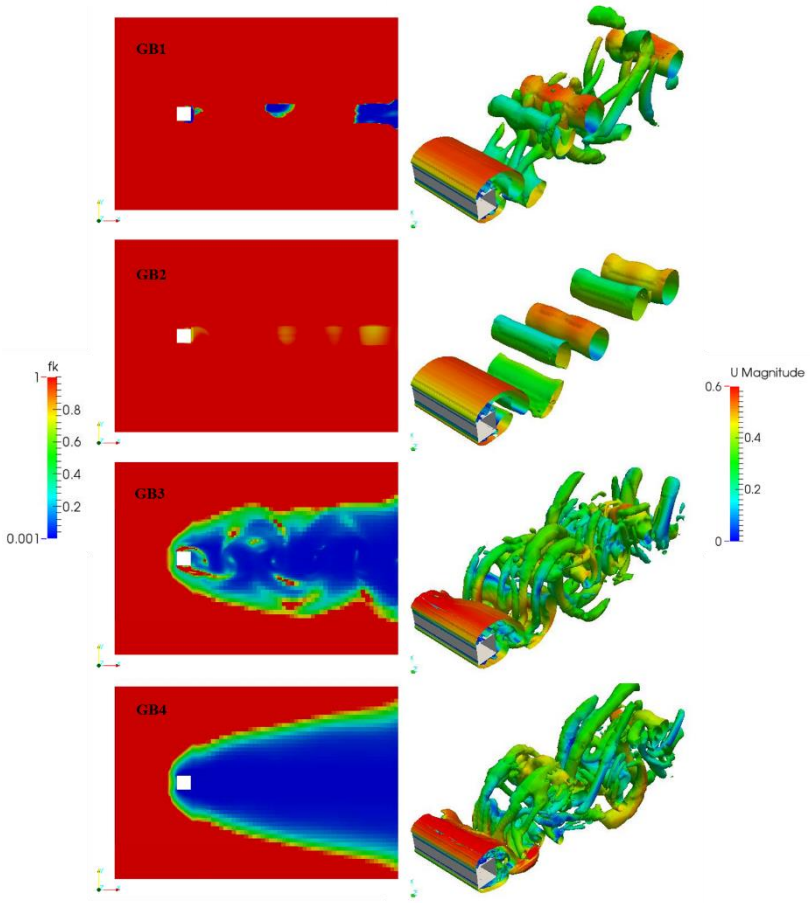



**Figure 5 Coarse mesh. Representation of $f_k$ in the symmetry plane of the cylinder (left), instantaneous iso-surfaces of $Q = 0.01\ s^{-1}$ (right) for GB1, GB2, GB3 and GB4 (from top to bottom)**

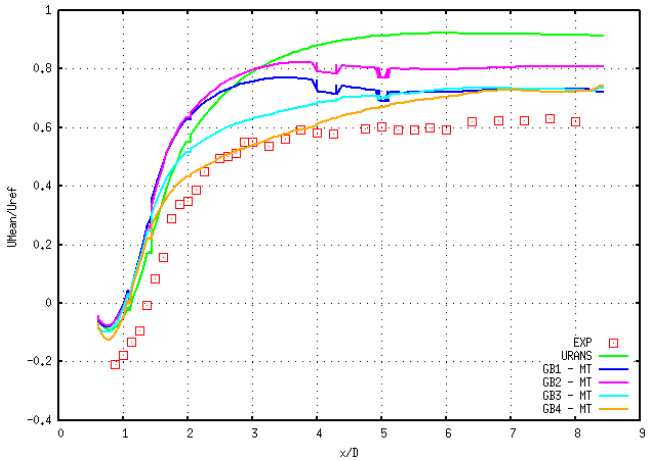



**Figure 6 Coarse mesh. Mean axial velocity profile - Experiment (red symbol), URANS (green), GB1 (dark blue), GB2 (pink), GB3 (light blue), GB4 (orange)**

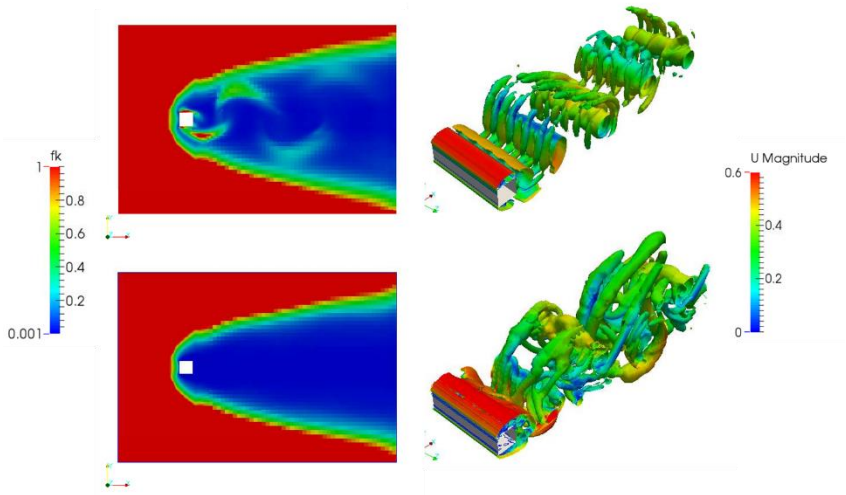



**Figure 7 Coarse Mesh. PANS GB4 –Blended central differences (0.4) – upwind scheme (top), Second-order upwind scheme (bottom)**

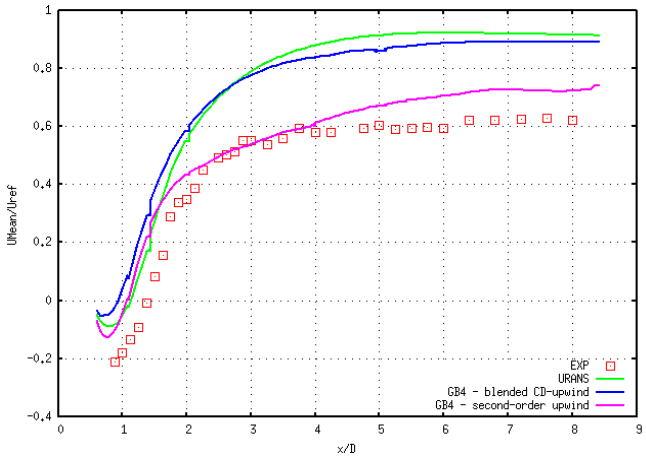



**Figure 8 Coarse Mesh. Mean axial velocity profile - Experiment (red symbol), URANS (green), GB4 second-order upwind (pink), GB4 Blended central differences (0.4) – upwind scheme (blue)**

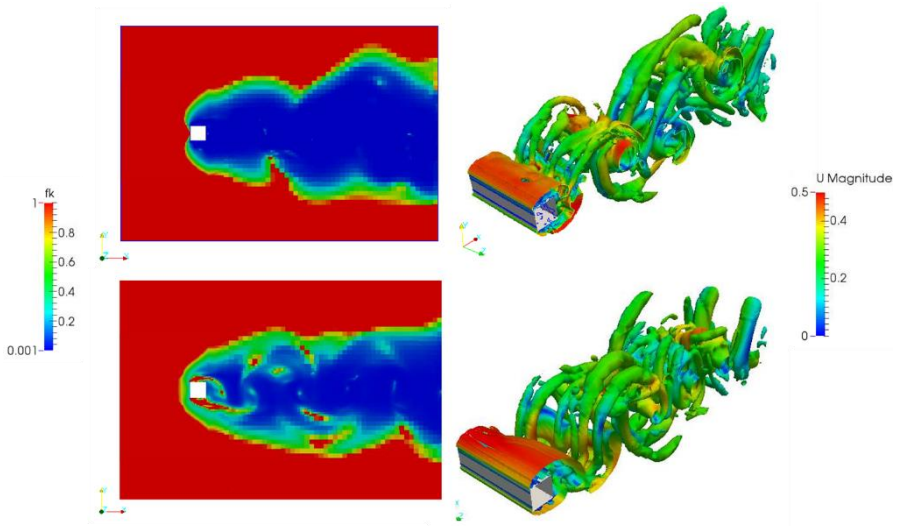



**Figure 9 Coarse Mesh. PANS GB3 – Zero Transport Model (top), Maximum Transport Model (bottom)**

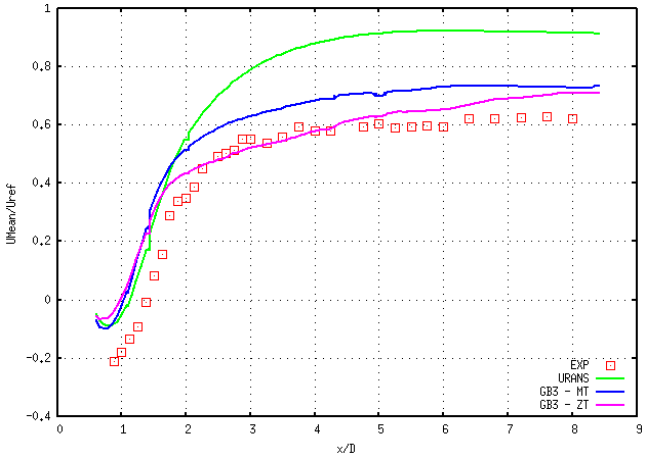



**Figure 10 Coarse Mesh. Mean axial velocity profile - Experiment (red square), URANS (green), GB3 Zero Transport Model (pink), GB3 Maximum Transport Model (blue)**

| | $C_D$ | $C_D$ − error (%) | $Str$ | $Str$ − error (%) |
|---|---|---|---|---|
| Exp. | 2.1 | - | 0.132 | - |
| URANS | 2.01 | -4.5 | 0.144 | 8.8 |
| Constant $f_k$ | 2.09 | -0.5 | 0.123 | -6.5 |
| GB1 | 2.08 | -1.1 | 0.141 | 7 |
| GB2 | 2.08 | -1.2 | 0.154 | 16.7 |
| GB3 | 2.09 | -0.4 | 0.127 | -3.3 |
| GB4 | 2.09 | -0.5 | 0.124 | -6 |

**Table 3 Coarse mesh. Mean drag and Strouhal number for the different cases**

### 3.2.2. Flow past a surface-mounted cube

It is important to assess the accuracy of the PANS model in simple turbulent wall-bounded flows before using it in the context of ground vehicles aerodynamics. The surface-mounted cube is a good test case for that as it involves unsteady separated flow regions and wake reattachment on the wall. A detailed description of the vortex structures as well as measurements of the flow statistics can be found in the experimental study of Martinuzzi and Tropea [11].

**Case Description**

The Reynolds number of the flow, based on the cube height $H$, is $80000$. This gives a reference bulk velocity of $48\,m.s^{-1}$, if $H$ is set to $0.025\,m$. The geometric dimensions of the computational domain, proportional to $H$, are represented in Figure 11. The velocity profile at the channel inlet is interpolated from the experimental data to simulate a fully developed turbulent upstream flow. The vortex structures and the velocity profiles in the symmetry plane of the cube will be compared to the experiment.

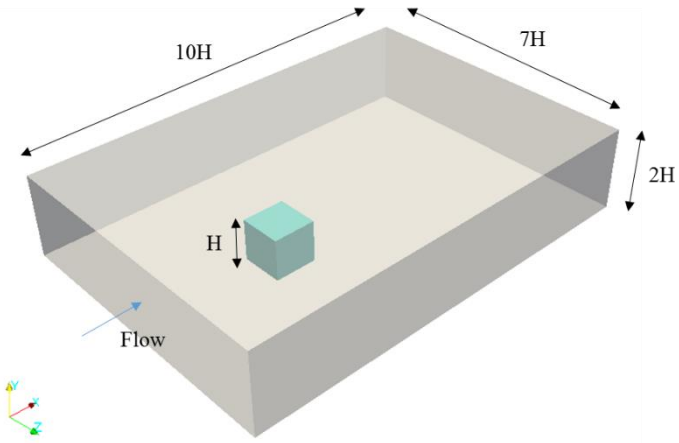



**Figure 11 Computational domain and geometry of the surface-mounted cube**

**Computational Details**

*Mesh* – In order to capture most of the flow details, the wake region of the cube is refined twice with respect to the coarsest mesh regions, as shown in Figure 12. The number of layers is set to 5 on all the non-slip walls and the $y^+$ surface-averaged value is equal to 10. The total cell count of the computational grid is 1800000.

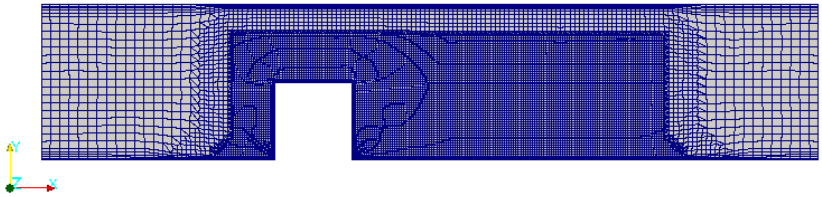


**Figure 12 Mesh view in the symmetry plane of the surface-mounted cube**

*Boundary Conditions* – The cube surface as well as the channel upper and lower boundaries are modelled as non-slip walls. The side boundaries of the flow domain are set to symmetry planes.

*Solver* – The total number of iterations for the transient simulation corresponds to 10 flow cycles - 1 flow cycle is defined as the travelling time of the fluid from the inlet to the outlet of the channel. The flow fields are time-averaged over the last two flow cycles.

**Results**

*Flow topology* – Iso-surfaces of the second invariant of the velocity gradient Q are compared between URANS and PANS-GB4 in Figure 14. In addition to the upstream trumpet vortex on the floor, which is present in both simulations, the PANS solution features richer multi-scale turbulent structures in the wake region of the cube.

*Velocity field* – Figure 15 shows the time-averaged streamlines in the symmetry plane of the flow. The experimental value of the downstream primary recirculation length is estimated at $1.612H$ [11]. In the PANS case, that length is better predicted as it is equal to $1.6H$ against $2H$ for URANS. Figure 17 shows the mean axial velocity profiles in the wall normal direction at five different locations of the flow represented in Figure 16. Overall, the PANS and URANS solutions are very similar except at $x/H = 4$ for which the PANS profile is closer the experiment.

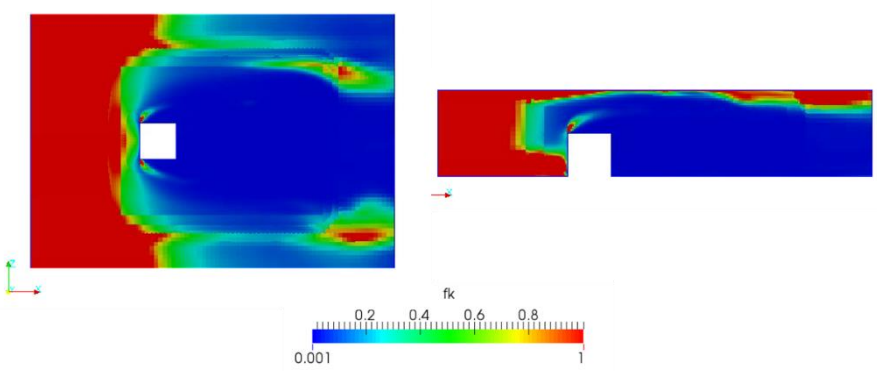



**Figure 13 $f_k$ field in the symmetry planes of the cube – $y = 0.5H$ (left), $z = 0H$ (right)**

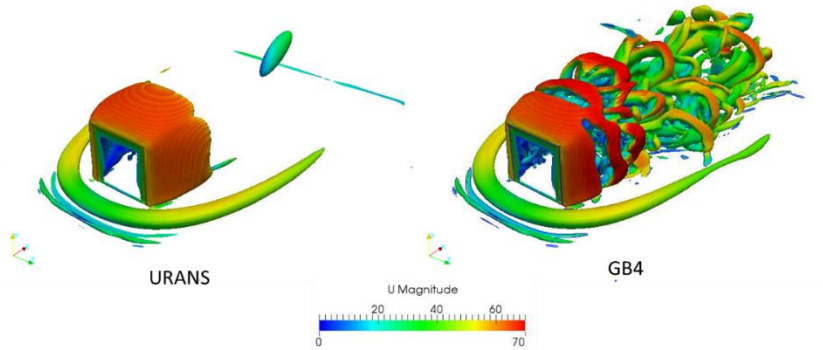



**Figure 14 Surface-mounted cube. Iso-surfaces of the Q criterion**

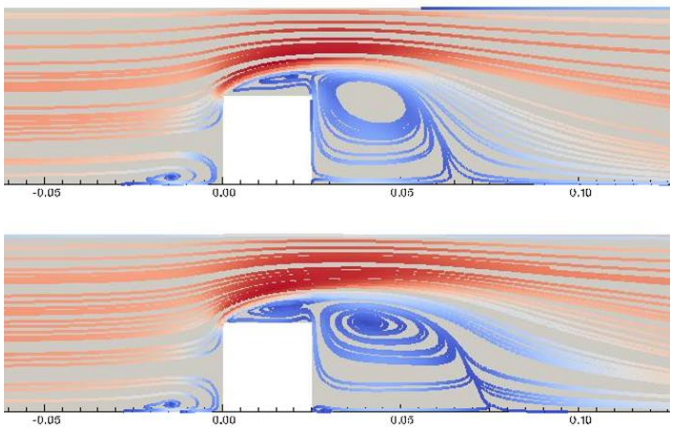



**Figure 15 Streamlines on the symmetry plane of the cube – PANS GB4 (top), URANS (bottom)**

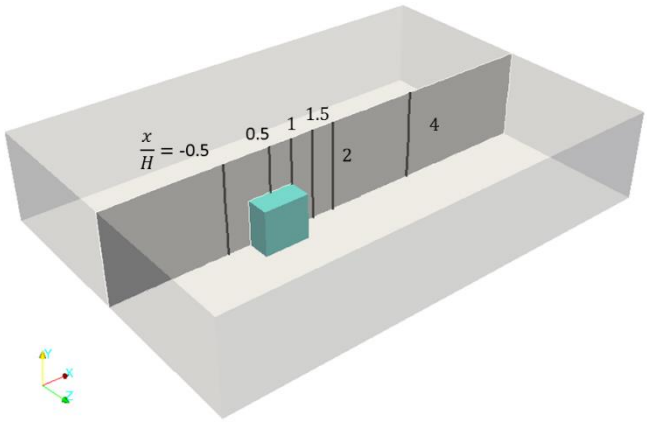



**Figure 16 Locations of experimental probes in the symmetry plane of the surface-mounted cube**

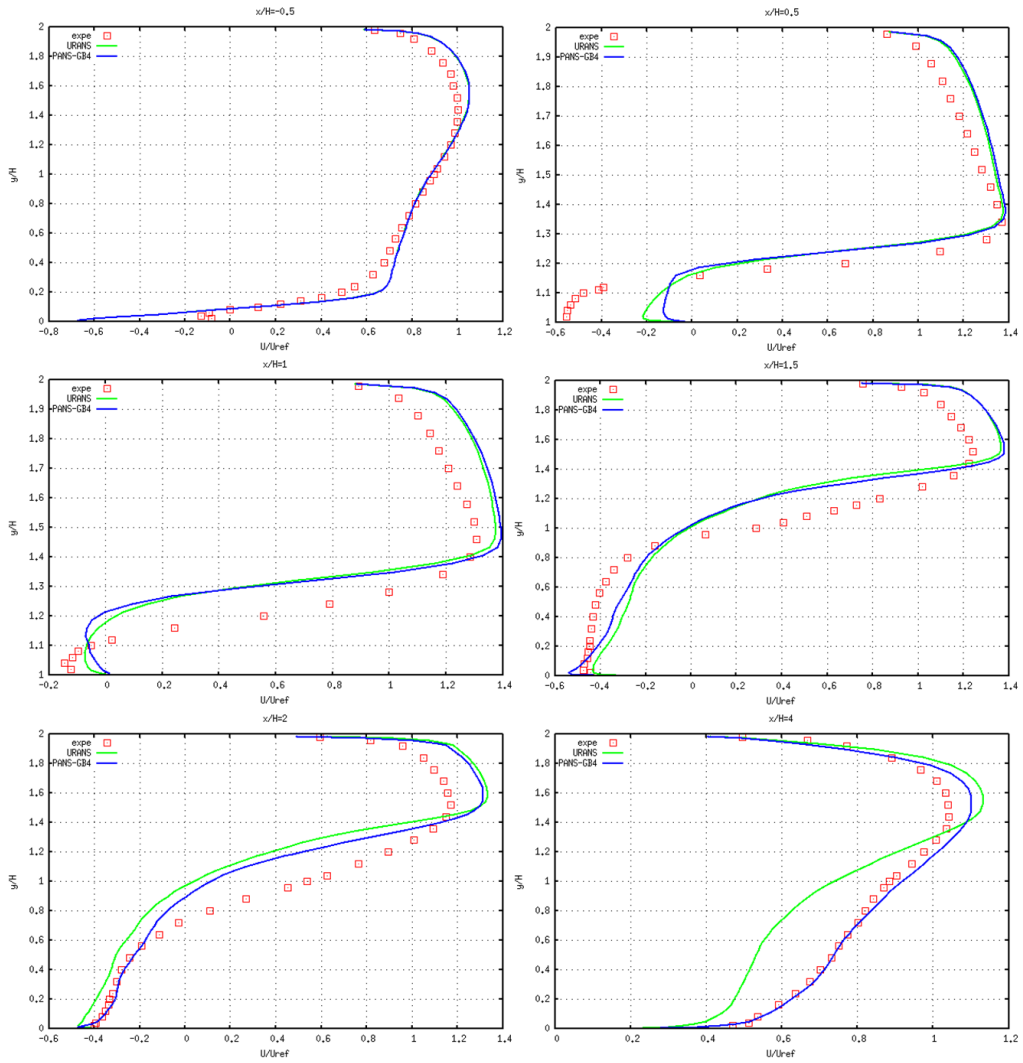



**Figure 17 Mean axial velocity in the symmetry plane of the surface-mounted cube – Experiment (red symbols), PANS GB4 (blue), URANS (green)**

### 3.3. Automotive External Aerodynamics

In this section, the PANS model is validated on a full-scale industrial case. The car model is of type hatchback. It is generally difficult to correctly predict the flow over this type of car by CFD methods. The flow in the wake of the car is very complicated and inherently unsteady. With the growing pressure on aerodynamic optimizations of road cars, the ability to quickly and accurately predict the aerodynamics is very important.

#### Case Description

The 1:1 detailed mock-up model is considered in the current validation study. The car model is displayed in Figure 18, the bulk dimensions of the car are approximately (3.8m × 1.7m × 1.6m). The computational domain is a rectangular numerical wind tunnel with dimensions (60m × 50m × 30m) displayed in Figure 18. At the inflow plane the flow velocity $U_\infty$ = 38.89 m/s was specified. At the top plane, side planes and outlet plane the zero value of the relative pressure and zero gradient boundary conditions for all turbulent variables were applied. The

floor plane is divided into parts where no-slip and symmetry boundary conditions were applied. The distance between car centre and start of the no-slip condition was chosen so to reach the thickness of the boundary layer corresponding to wind tunnel measurement. The computational domain is sufficiently large to reduce influence of the computational wind tunnel boundaries on the flow around the vehicle.

The computational mesh created with iconHexMesh consists of 68 million cells. The resulting mesh is polyhedral (hex-dominant) with multiple regions of refinement around the car and its wake as can be seen in Figure 19. To properly refine mesh near the car surface two consecutive distance-based refinements are done as displayed in Figure 19. The whole car surface is covered by 4 prismatic layers resulting in the dimensionless wall distance $y^+ \sim 30$. The surface mesh is refined in areas of high curvature.

The numerical schemes and the solution procedure were described in Section 3.1. The GB2 method for the calculation of $f_k$ was chosen for the industrial case. The time step was set to $\Delta t = 0.0001s$ in order to achieve a maximum CFL ($\sim U\Delta t/\Delta x$) under 5.

The same mesh, numerical schemes and solution procedure was used to predict the flow with URANS $\zeta - f$ model ( $f_k = 1$). In both PANS and URANS cases 1.5s of flow was calculated.

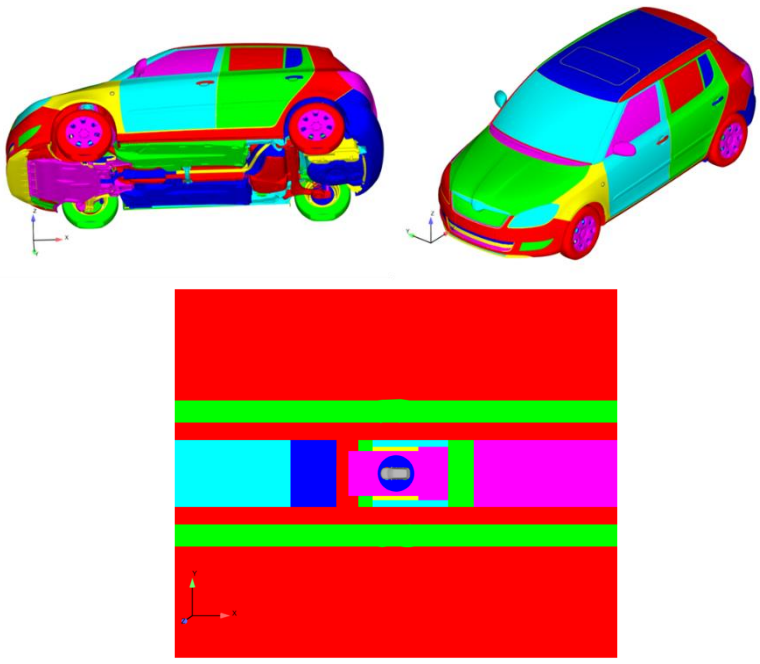


**Figure 18 Car geometry and top view of the computational domain**

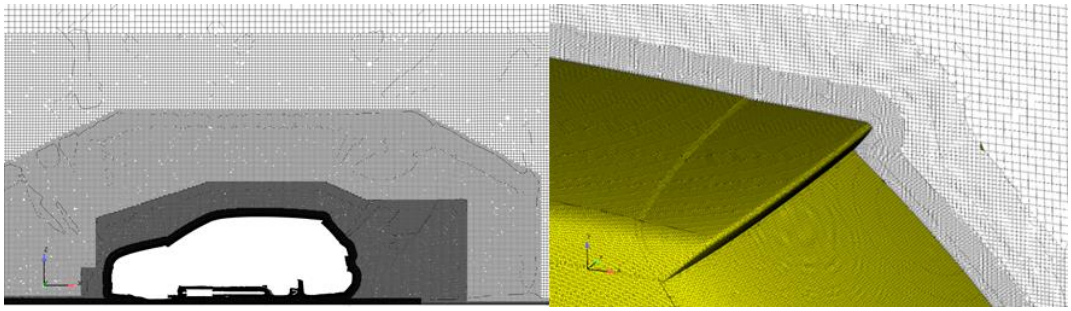


**Figure 19 Mesh refinement around the car**

## Results

To evaluate the performance of the PANS model in the industrial external flow we must first check if the mesh resolution is sufficient to trigger the seamless transition between URANS and DNS. This depends on the parameter $f_k$, if $f_k < 1$ then the PANS results depart from URANS results. It can be seen from Figure 20 that the spatial resolution is fine enough to differ substantially from the URANS limit. The discontinuities in the $f_k$ field are results of mesh refinement. It can be observed that the mesh resolution is not sufficient in some areas of the wake, especially in the far wake and in the areas with Helmholtz-Kelvin instability. To show capability of PANS modelling to predict strongly separated flow in the external car aerodynamic, we compare PANS results to URANS $\zeta - f$ model results. The overall solution time and HPC resources are the same between URANS and PANS.

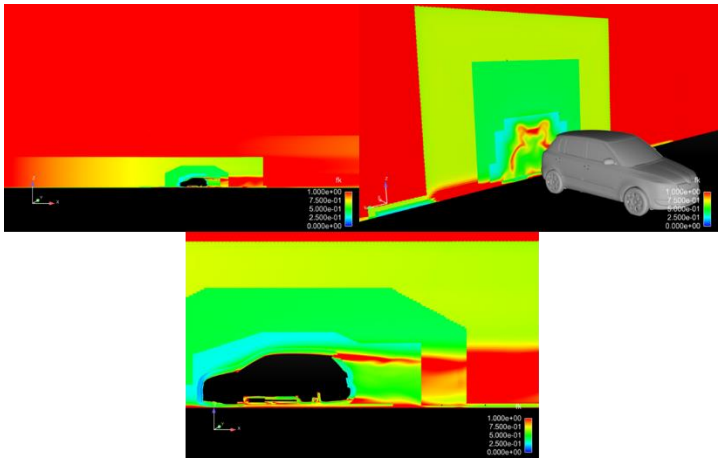


**Figure 20 Unresolved-to-total ratio of the turbulent kinetic energy $f_k$ instantaneous value**

The vortex structure within the wake region and around the car is visualized by Q-criterion ($Q = 1000s^{-2}$) representing basically the second invariant of the velocity gradient tensor, see Figure 21. The predicted vortex structure is richer with the PANS model in the wake area and also in wakes of the wheels and mirrors. The structure of A-pillar vortex is also finer than with URANS. Figure 22 displays the complicated structure of the wake behind the car.

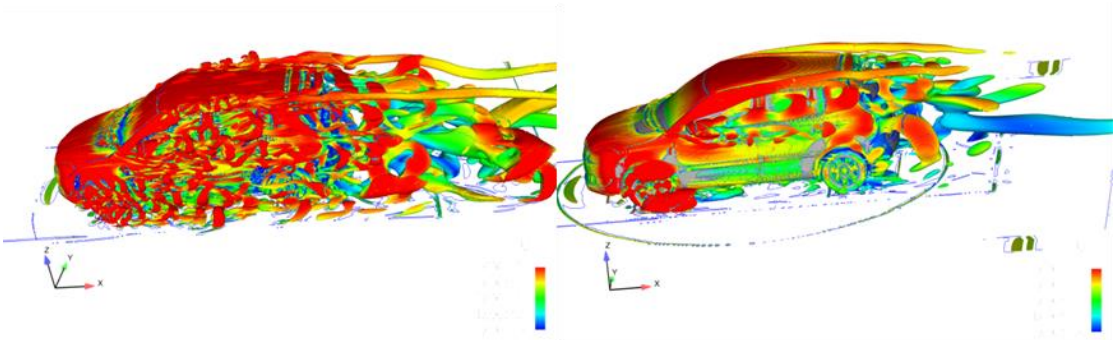


**Figure 21 Flow field structures predicted by PANS (left) and URANS (right), visualized by the Q-criterion ($Q = 1000s^{-2}$) and coloured by velocity magnitude**

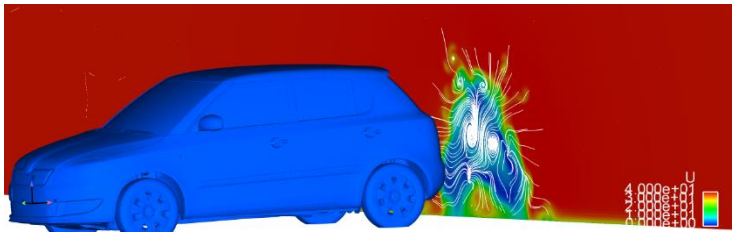


**Figure 22 Instantaneous streamline pattern behind the car coloured by magnitude of instantaneous velocity**

The CFD results are compared to experimental results. The flow around the car was experimentally examined in the VW wind tunnel. To have more information about the flow field the pressure strips measurement method was used, see Figure 25. Detailed view on the pressure strip is displayed in

Figure 23. This tool uses multiple connected pressure probes, details about the tool are described in [13]. The measured car was equipped with 24 pressure strips and each strip contains 24 pressure sensors; the distance between the probes is 1cm. The post-processing tool also described in [13] was used to compare pressure coefficient $Cp$ on different parts of the car between wind tunnel experiment and CFD, see Figure 25, Figure 26, Figure 28 and Figure 29. We compare value of $C_p$ and also pressure gradients on some parts of the car. The analysis of the pressure stripes shows very similar behaviour of the models in the front of the car; see Figure 25 and Figure 28. Also the comparison to experimental value of $C_p$is very good.

Differences between the models can be found on the back of the car, see Figure 27 and Figure 26. The PANS results are closer to experimental values of $C_p$on the rear bumper than the URANS results, see Figure 26. Larger discrepancies between the experimental values and the CFD result can be found on the fifth door; see Figure 27. This difference is reduced in PANS results. This discrepancy is possibly a result of a difference between real wind tunnel geometry and the numerical domain where some important effects (horizontal buoyancy, effect of nozzle and collector shape) are omitted. This issue will be addressed in the future by using more realistic wind tunnel geometry and also by using wind tunnel correction on the experimental results.

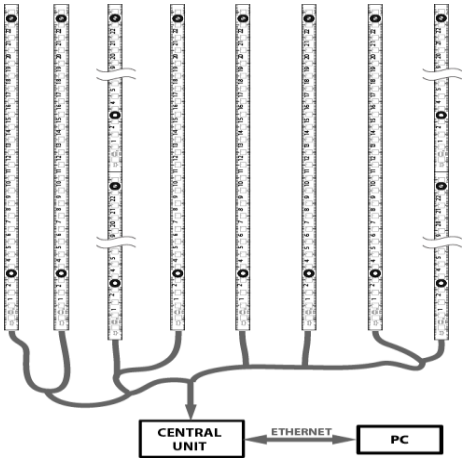



**Figure 23 Pressure strips**

Both methods resulted in similar averaged values of a drag coefficients $C_d$, both methods overpredicted the values measured in the wind tunnel, the delta between experiment and CFD was by 10% smaller with PANS than with URANS. Very similar improvement was observed also for the lift coefficient $C_l$.

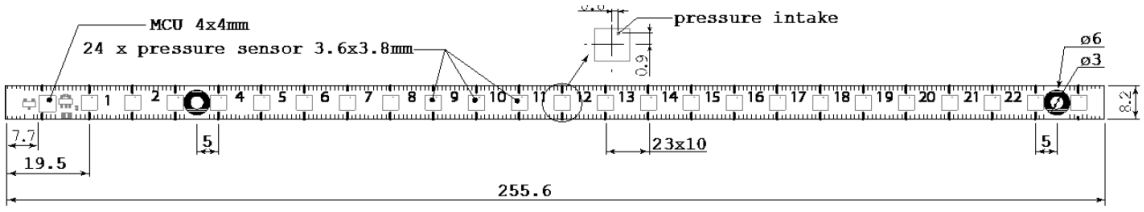



**Figure 24 Pressure strip**

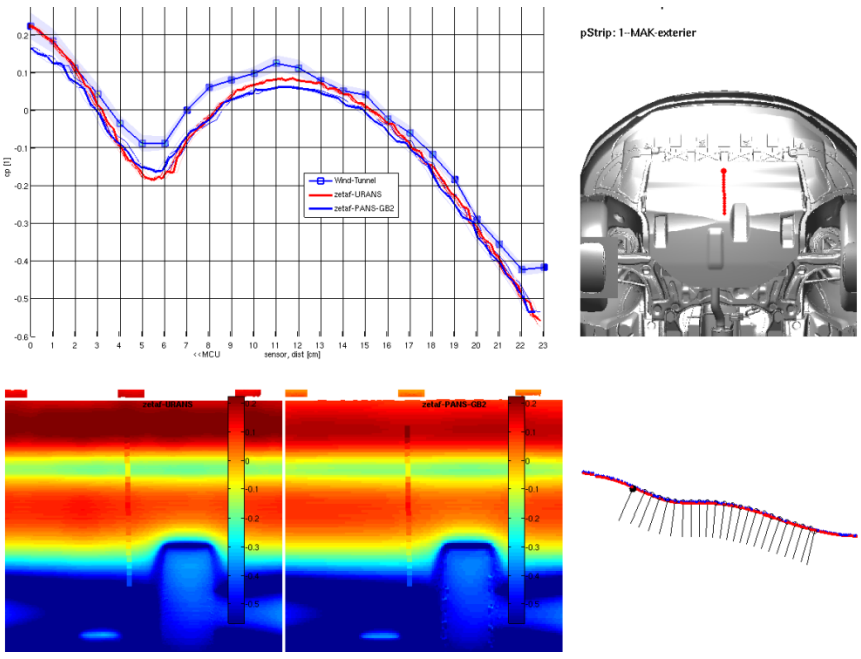



**Figure 25 Pressure strip on engine cover**

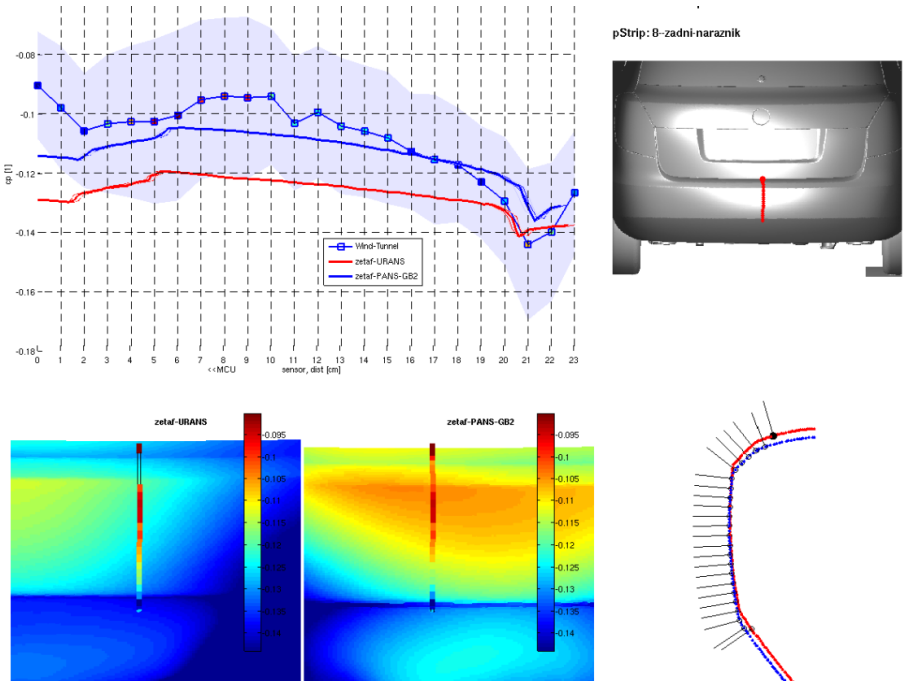



**Figure 26 Pressure strip on rear bumper**

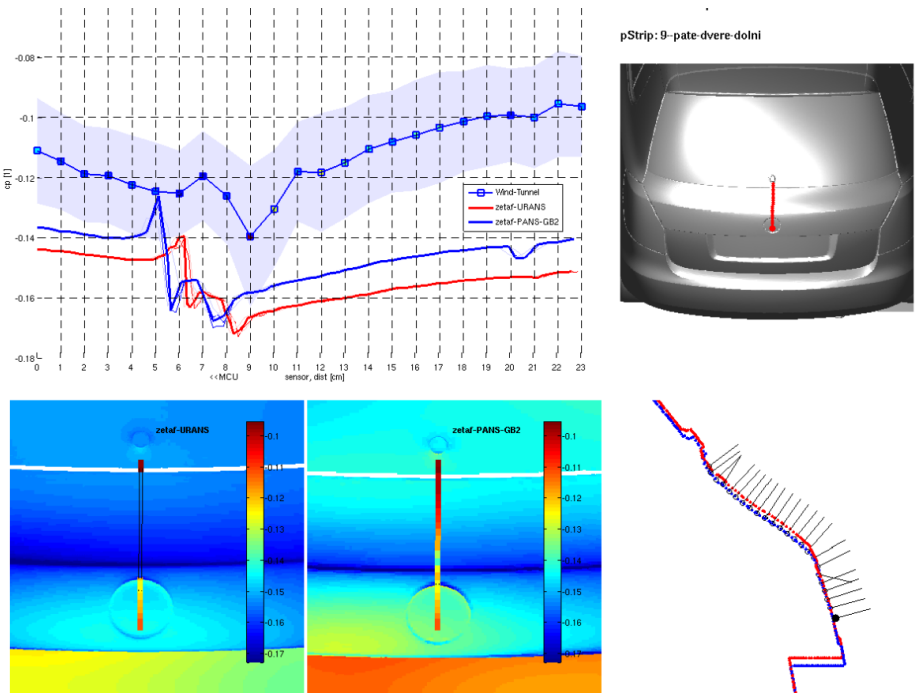



**Figure 27 Pressure strip on fifth door**

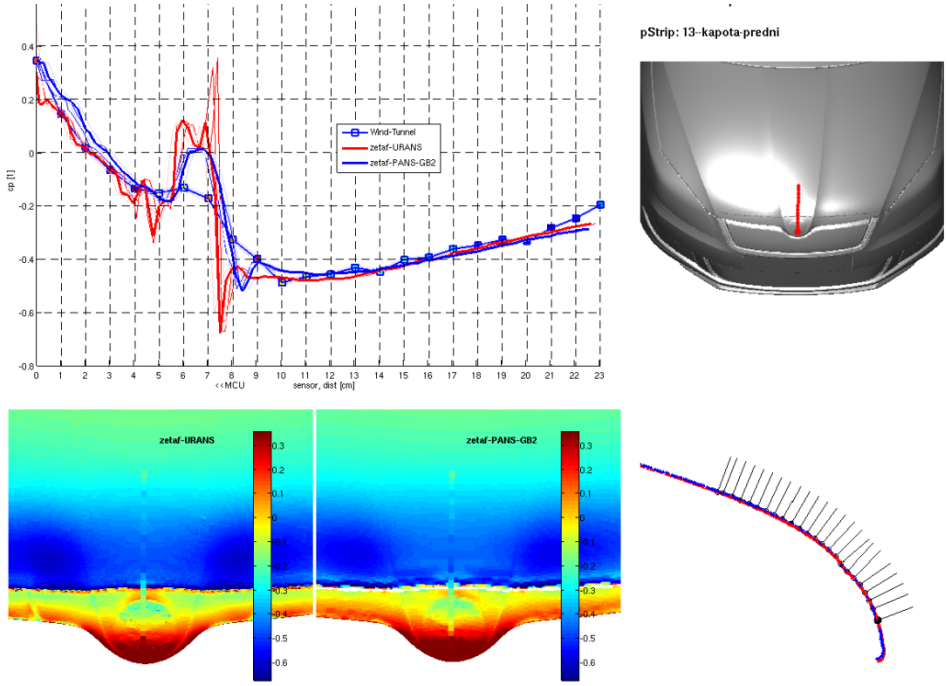



**Figure 28 Pressure strip in the front of the car**

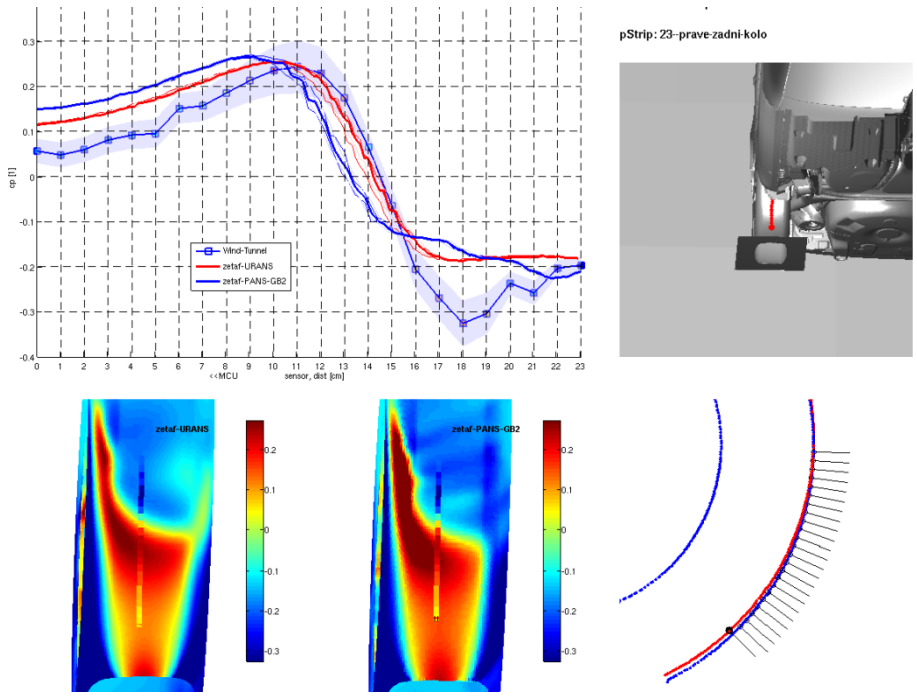



**Figure 29 Pressure strip on rear right wheel**

## 4. Summary

A PANS $\zeta - f$ turbulence modelling framework was implemented in the iconCFD® simulation toolkit. Four different methods for the calculation of $f_k$ were analysed on three cases: the square cylinder, the surface-mounted cube and a SKODA vehicle model. In all cases,

the PANS implementation has showed both in the richness of the turbulent content as well as in the improved accuracy of the results. Further investigation needs to be done in the stability issues of the GB4 method which provides the best accuracy. For external aerodynamics simulations, the authors currently recommend the use of the GB2 method which provides a good compromise between accuracy and stability.

## 5. Acknowledgements

The work presented here was carried out in collaboration with ŠKODA who kindly accepted to provide validation materials for testing the PANS $\zeta - f$ model. The authors of this paper are grateful for this greatly acknowledged contribution.

## REFERENCES


1. Jochen Frohlic, D. v. (2008). Hybrid LES/RANS methods for the simulation of turbulent flows. Progress in Aerospace Sciences 44, 349-377.
2. Girimaji S. S. (2006). Partially-Averaged Navier-Stokes Model for Turbulence: A Reynolds-Averaged Navier-Stokes to Direct Numerical Simulation Bridging Method. Journal of Applied Mechanics, 413-421.
3. Girimaji S. S. (2005). Partially-averaged Navier Stokes Model for Turbulence: Implementation and Validation. AIAA 2005-0502, 1-14.
4. Branislav Basara et al. (2011), Near-Wall Formulation of the Partially Averaged Navier-Stokes Turbulence Model. AIAA Vol. 49, No. 12, 2627-2635.
5. M. Germano (1992), Turbulence: the filtering approach. Journal of Fluid Mechanics Vol. 238, 325-336.
6. K. Hanjalic et al. (2004). A robust near-wall elliptic-relaxation eddy-viscosity turbulence model for CFD. International Journal of Heat and Fluid Flow 25, 1047-1051.
7. M. Popovac, K. Hanjalic (2007). Compound Wall Treatment for RANS Computation of Complex Turbulent Flows and Heat Transfer. Flow Turbulence Combust 78, 177-202.
8. Lyn D.A, Rodi W. (1994). The flapping shear layer formed by flow separation from the forward corner of a square cylinder. Journal of Fluid Mechanics, 267-353.
9. Lyn D.A et al. (1994). A laser-Doppler velocimetry study of ensemble-averaged characteristics of the turbulent near wake of a square cylinder. Rept. SFB 210 /E/100.
10. Rodi W. (1997). Comparisons of LES and RANS calculations of the flow around bluff bodies. J. Wind Eng. Ind. Aerodyn. 69-71, 55-75.
11. Martinuzzi R., Tropea C. (1993). The Flow Around Surface-Mounted, Prismatic Obstacles Placed in a Fully Developed Channel Flow. Journal of Fluids Engineering, Vol. 115, 85-92.
12. http://cfd.mace.manchester.ac.uk/ercoftac/
13. Filipský J., Polická P., Čížek J. (2015). Multipoint Pressure Strip Measurement for Advanced CFD Validation. 10th FKFS-Conference, Progress in vehicle Aerodynamics and Thermal Management.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FREE-SURFACE FLOW SIMULATIONS FOR MOORED AND FLOATING OFFSHORE PLATFORMS

**Arthur E.P. Veldman[1,4], Roel Luppes[1], Peter van der Plas[1], Henri van der Heiden[1], Bülent Düz[2,3], Henk Seubers[1], Joop Helder[3], Tim Bunnik[3] and René Huijsmans[2]**

[1] Institute for Mathematics and Computer Science, University of Groningen, Groningen, The e-mail: {a.e.p.veldman,r.luppes,p.van.der.plas,h.j.l.van.der.heiden,h.seubers}@rug.nl

[2] Department of Ship Hydrodynamis, Delft University of Technology, Delft, The Netherlands
e-mail: {r.h.m.huijsmans}@tudelft.nl

[3] Maritime Reseach Institute MARIN, Wageningen, The Netherlands
e-mail: {b.duz,j.helder,t.bunnik}@marin.nl

[4] Multiscale Modeling and Simulation, University of Twente, Enschede, The Netherlands

**Keywords:** Extreme waves, fluid–solid body interaction, CFD, local grid refinement, experimental validation.

**Abstract.** *During the development of the ComFLOW simulation method many challenges have to be tackled concerning the flow modelling and the numerical solution algorithm. Examples hereof are wave propagation, absorbing boundary conditions, fluid–solid body interaction, turbulence modeling and numerical efficiency. Some of these challenges will be discussed in the paper, in particular the design of absorbing boundary conditions and the numerical coupling for fluid–solid body interaction. As a demonstration of the progress made, a number of simulation results for engineering applications from the offshore industry will be presented: a wave-making oscillating buoy, a free-fall life boat dropping into wavy water, and wave impact against a semi-submersible offshore platform. For those applications, MARIN has carried out several validation experiments.*

## 1 INTRODUCTION

Over the centuries, understanding the motion and behavior of waves in nature has been a very popular subject among researchers from various fields of science. Even today, when we have highly capable numerical methods and computational power at our disposal, numerical modeling of water wave propagation remains a formidable challenge. During the development of the ComFLOW simulation method [1–5] many of the numerical challenges have been tackled:

- Waves should be allowed to freely enter or leave the domain, requiring absorbing or non-reflecting boundary conditions which are able to deal with the dispersive character of waves on deep water [6]. Such conditions will be discussed below in more detail.
- Accurate wave propagation requires much attention to the description (reconstruction and advection) of the free surface. Info on this aspect is to be found in the PhD thesis of Duz [7] and the forthcoming paper [8].
- The turbulence model not only has to deal with coarse grids, but should also recognize wall-bounded turbulence from free-surface turbulence. This is tackled with a new class of adaptive, minimum-dissipation turbulence LES models [9–11].
- The numerical coupling of the flow dynamics with the dynamics of a floating object has to deal with situations where the added mass is larger than the mass of the object. Hereto a more simultaneous type of numerical coupling is being developed. This paper contains more details about this issue.
- The efficiency of the simulations is enhanced by means of local grid refinement and parallelisation [12, 13].

The phenomena of interest are local but embedded in a vast spatial domain, like the interaction between free-surface waves and man-made structures. For efficient computational modeling, this spatial domain around the region of interest is truncated via artificial boundaries, leading to the fundamental question: What is the boundary condition to be imposed on these artificial boundaries in such a way that the solution in the compact domain coincides with the solution in the original domain? We thus enter the realm of absorbing boundary conditions (ABC), also known as non-reflecting, open, transparent or radiating boundary conditions. Following the pioneering research by Engquist and Majda [14], a wide variety of ABCs has been developed. These conditions not only have to be considered from a theoretical modeling point of view (well-posedness, modeling accuracy), but also from a numerical point of view (stability, numerical accuracy, efficiency). An impression of the work done thus far can be obained from several review papers [15–18]. For the work on ABCs done in the ComFLOW project we refer to the PhD theses of Wellens [19] and Düz [7]. Below we will present our latest contributions to this issue which focus on the treatment of dispersive waves, i.e. waves whose propagation speed is not known beforehand but has to be deduced from the (local) solution.

Another dimension to the above problem is added when the structures under study are moving, either free-floating or attached to a mooring system, where the interaction between the incoming waves and the dynamics of the structure comes into play. Physically, we can distinguish one-way or two-way interaction. In the former case the structure ‘simply’ reacts to the oncoming flow field. But in the latter case the interaction is such that the motion of the structure influences the flow field around the structure. The latter case also poses most challenges to the numerical coupling between flow and structure. Such a numerical coupling approach can be aggregated (monolithic) or segregated (partitioned). In the former case all discrete flow equations are combined into one single set of equations which is then solved simultaneously. In the

latter case two separate discrete systems (modules) can be recognized equipped with recipes to exchange information between the two separate modules. On the one hand this enhances the flexibility of the approach, but on the other hand it requires a (hierarchical) iterative exchange of information between the modules with its consequences for numerical stability and convergence. Below we describe our efforts to find a compromise between the robust monolithic approach and the more flexible but vulnerable partitioned approach. We will see that the ratio between the mass of the structure and its added fluid mass plays an essential role.

The organization of the paper is as follows. After an introduction of the flow equations, the new class of absorbing boundary conditions is presented. Thereafter we discuss the numerical coupling strategy between fluid and solid structure. The paper finishes with some practical applications: an oscillating buoy in a small domainn, a falling lifeboat and a semi-submersible in waves. Some of these applications are validated by experiments carried out at MARIN.



Figure 1: Drop test with free-fall lifeboat (from: www.verhoef.eu)

## 2 MATHEMATICAL FLOW MODEL

Incompressible, turbulent fluid flow can be modelled by means of the Navier–Stokes equations.

$$\mathcal{M}\mathbf{u} = 0, \qquad \frac{\partial \mathbf{u}}{\partial t} + \mathcal{C}(\mathbf{u})\mathbf{u} + \mathcal{G}p - \mathcal{V}\mathbf{u} = \mathbf{f}. \tag{1}$$

Here $\mathcal{M}$ is the divergence operator [1] which describes conservation of mass. Conservation of momentum is based on the convection operator $\mathcal{C}(\mathbf{u})\mathbf{v} \equiv \nabla(\mathbf{u} \otimes \mathbf{v})$, the pressure gradient operator $\mathcal{G} = \nabla$, the viscous diffusion operator $\mathcal{V}(\mathbf{u}) \equiv \nabla \cdot \nu \nabla \mathbf{u}$ and a forcing term $\mathbf{f}$. The kinematic viscosity is denoted by $\nu$.

Turbulence is modelled by means of large-eddy simulation (LES) using a low-dissipation QR-model as formulated by Verstappen [9] and extended in the PhD thesis of Rozema [10, 11]. For its use in maritime applications, see [20, 21].

The Navier–Stokes equations (1) are discretized on an Arakawa C-grid. The second-order finite-volume discretization of the continuity equation at the 'new' time level $\cdot^{n+1}$ is given by

$$M^0 \mathbf{u}_h^{n+1} = -M^\Gamma \mathbf{u}_h^{n+1}, \tag{2}$$

[1] Calligraphic symbols denote analytic operators, whereas their discrete counterparts are denoted by upper-case italic symbols.

where $M^0$ acts on the interior of the domain and $M^\Gamma$ acts on the boundaries of the domain. In the discretized momentum equation, convection $C(\mathbf{u}_h)$ and diffusion $D$ are discretized explicitly in time. The pressure gradient is discretized at the new time level. In this exposition, for simplicity reasons the first-order forward Euler time integration will be used. In the actual calculations, a second-order Adams–Bashforth method is being applied.

Letting the diagonal matrix $\Omega$ denote the matrix containing the volumes of the control volumes, gives the discretized momentum equation as

$$\Omega \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\delta t} = -C(\mathbf{u}^n)\,\mathbf{u}^n + V\mathbf{u}^n - Gp^{n+1} + \mathbf{f}. \tag{3}$$

The discrete convection operator is skew-symmetric, such that convection does not contribute to energy production or dissipation; see Verstappen and Veldman [22]. In particular, its discretization preserves the energy of the flow and does not produce artificial viscosity. To make the discretization fully energy-preserving, the discrete gradient operator and the divergence operator are each other's negative transpose, i.e. $G = -M^{0T}$, thus mimicking the analytic symmetry $\nabla = -(\nabla\cdot)^T$. In this way, also the work done by the pressure vanishes discretely.

The solution of the discrete Navier–Stokes equations is split into two steps. Firstly, an auxiliary variable $\tilde{\mathbf{u}}$ is defined through

$$\Omega \frac{\tilde{\mathbf{u}} - \mathbf{u}^n}{\delta t} = -C(\mathbf{u}^n)\,\mathbf{u}^n + D\mathbf{u}^n + \mathbf{f}. \tag{4}$$

With this abbreviation, the discrete momentum equation (3) can be reformulated as

$$\mathbf{u}^{n+1} = \tilde{\mathbf{u}} - \delta t\,\Omega^{-1} G p^{n+1}. \tag{5}$$

Secondly, by imposing discrete mass conservation (2) at the new time level, substitution of (5) results in a discrete Poisson equation for the pressure:

$$\delta t\, M^0 \Omega^{-1} G p^{n+1} = M^0 \tilde{\mathbf{u}} + M^\Gamma \mathbf{u}^{n+1}. \tag{6}$$

The liquid region and the free liquid surface are described by an improved VOF-method; see Hirt and Nichols [23] and Kleefsman et al. [1].

## 3 ABSORBING BOUNDARY CONDITIONS

Following the expose presentend in [7], a suitable ABC can be developed starting from the radiation condition studied by Sommerfeld [24]:

$$\frac{\partial}{\partial t}\phi + c\frac{\partial}{\partial x}\phi = 0. \tag{7}$$

This condition, formulated in the velocity potential $\phi$, suppresses left-running waves at a right-hand-side boundary. Therefore, it has been proposed as a non-reflecting boundary condition by Engquist and Majda [14]. It was later extended to higher-order versions using powers of the left-hand side operator; see e.g. the review by Givoli [18].

### 3.1 Three dimensions

In the three-dimensional case, Higdon [25, 26] showed it possible to allow the wave under an angle of incidence $\alpha$ with the outflow boundary. In the higher-order conditions even more

angles can be chosen: $\alpha_p,\ p = 1, \ldots, P$, where P denotes the order:

$$\prod_{p=1}^{P} \left( \cos\alpha_p \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right) \phi = 0. \tag{8}$$

To show the benefit of using (8) as opposed to (7), the amount of spurious reflection is investigated as a function of the angle of incidence $\theta$. Hereto, at the artificial boundary the solution can be expressed as the sum of the outgoing and reflected waves

$$\phi(x, y, t) = e^{i(k_x x + k_y y - \omega t)} + R\, e^{i(-k_x x + k_y y - \omega t)}, \tag{9}$$

where the first term represents the wave with amplitude equal to unity impinging on the boundary, and the second term represents the spuriously reflected wave with amplitude $R$. The latter amplitude can be evaluated by substituting (9) into the general Higdon boundary condition (8):

$$|R_H| = \prod_{p=1}^{P} \left| \frac{\cos\alpha_p - \cos\theta}{\cos\alpha_p + \cos\theta} \right|. \tag{10}$$

Here, $\theta$ is the 'real' angle of incidence measured in the clockwise or counter-clockwise direction from the positive $x$-direction ($|\theta| < \pi/2$). Higdon's boundary condition (8) is non-reflecting as soon as $\theta$ equals one of the $\pm\alpha_p$'s, whereas (7) is non-reflecting for $\theta = 0$ only. A comparison of the theoretical reflection coefficients for the various ABC variants is shown in Fig. 2.

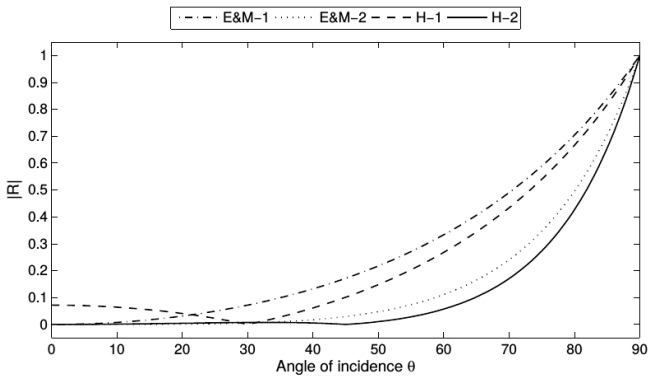



Figure 2: Reflection coefficient of four absorbing boundary conditions versus the angle of incidence $\theta$. E&M-1 and E&M-2 stand for the $1^{st}$- and $2^{nd}$-order Engquist-Majda ABC. The $1^{st}$- and $2^{nd}$-order Higdon ABC are denoted H-1 (with $\alpha_1 = 30°$) and H-2 (with $\alpha_1 = 0°,\ \alpha_2 = 45°$), respectively [7].

For use in a Navier–Stokes context, the potential has to be related to the velocity (via the gradient of the potential) and pressure (via the dynamic free-surface condition). Doing so, it becomes a boundary condition featuring velocity and pressure. It can be embedded in the pressure Poisson equation as a boundary condition in a similar way as the substitution of (5) in (2); see also [7].

### 3.2 Dispersive ABC

Because of dispersion each individual wave component has its own phase speed $c$. Therefore, a boundary condition like (8) cannot annihilate all these wave components. We expect some amount of reflection for such a boundary condition, but it will hopefully be restricted within certain limits.

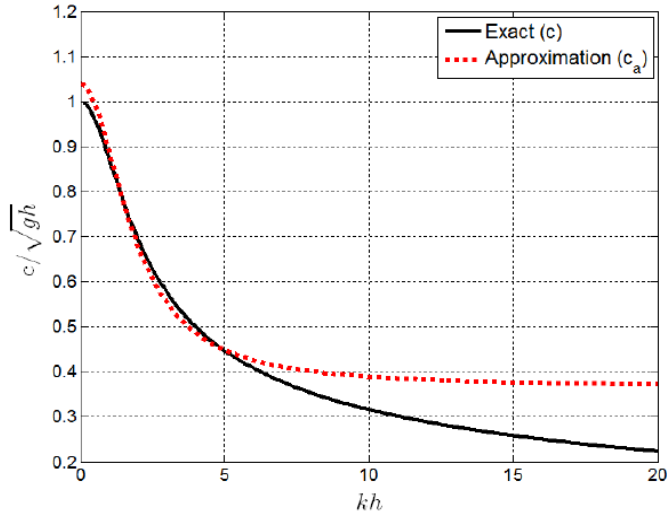



Figure 3: Padé approximation of the dispersion relation. For the coefficients in (11), $a_0 = 1.04$, $a_1 = 0.106$ and $b_1 = 0.289$ are used [7].

Firstly, we introduce a rational Padé expression which approximates the dispersion relation as

$$\text{exact: } c^2 = gh\,\frac{\tanh(kh)}{kh}; \qquad \text{approximation: } c_a = \sqrt{gh}\,\frac{a_0 + a_1(kh)^2}{1 + b_1(kh)^2}. \tag{11}$$

A proper choice of coefficients $a_0$ , $a_1$ and $b_1$ leads to a close approximation for the targeted range of $kh$ values. In Fig. 3 the difference between the two curves gives an indication for the amount of reflection caused by the rational approximation. Note that the higher wave numbers will be damped by the physics of the problem, hence some reflection will disappear quickly.

A further improvement is introduced into the design of the boundary condition, by exploiting the exponential behavior of the wave potential in the $z$-direction. The wave number $k$ is computed locally from the potential itself, via

$$k^2\phi = \frac{\partial^2}{\partial z^2}\phi. \tag{12}$$

This relation can be substituted into (11) and combined with (8) for $P = 1$, to reach the final form of the $1^{st}$-order absorbing boundary condition ABC-1:

$$\cos\alpha\left(1 + b_1h^2\frac{\partial^2}{\partial z^2}\right)\frac{\partial\phi}{\partial t} + \sqrt{gh}\left(a_0 + a_1h^2\frac{\partial^2}{\partial z^2}\right)\frac{\partial\phi}{\partial x} = 0. \tag{13}$$

A stability analysis of this boundary condition can be found in the PhD thesis of Wellens [19].

### 3.3 Dispersive directional ABC

A further modification of the dispersive ABC accounts for both dispersive and directional effects of the waves. As the $2^{nd}$-order Higdon ABC has superior performance over the $1^{st}$-order one in terms of directional effects, we will incorporate the improvements that we made in the previous section by adding dispersive effects. As this ABC consists of the product of two operators, and considering the relations (11) and (12), we realize that only one of the operators can include the approximation for the dispersion relation. Otherwise, the product of two approximations would yield a fourth-order derivative in the $z$-direction which will cause difficulties

when discretized at the boundaries. Therefore, we substitute the relations (11) and (12) in only one of the two operators. The resulting expression for the $2^{nd}$-order ABC-2 becomes

$$\left(\cos\alpha_1\frac{\partial}{\partial t}+c\frac{\partial}{\partial x}\right)\left[\cos\alpha_2\left(1+b_1h^2\frac{\partial^2}{\partial z^2}\right)\frac{\partial\phi}{\partial t}+\sqrt{gh}\left(a_0+a_1h^2\frac{\partial^2}{\partial z^2}\right)\frac{\partial\phi}{\partial x}\right]=0. \quad (14)$$

An in-depth analysis of the implementation and effectiveness of these boundary conditions in practical situations can be found in the PhD thesis of Duz [7, 27].

### 3.4 Example

As an illustration of the new type of absorbing boundary conditions we show some simulations of an oscillating sphere, with prescribed motion, producing a concentric circular wave pattern. A sphere (radius 4m) is initially located 4m above a free water surface and moves vertically along the z-direction with a sinusoidal motion $z_s(t) = 2 + 2\cos(2.4t)$. After the impact of the sphere on the water surface, a series of circular concentric waves is produced radiating outward from the center. Both ABC-1 and ABC-2 are used at the outflow boundaries and their performance in absorbing these waves is compared.

A reference solution is obtained by repeating the simulation in a large domain: see Fig. 4 for the setup of the two domains. The small domain $\Omega_S$ has $Lx_{\Omega_S} = Ly_{\Omega_S} = 50$m, its depth is $Lz_{\Omega_S} = 10$m. The larger domain $\Omega_L$ has the same depth but different length and width, $Lx_{\Omega_L} = Ly_{\Omega_L} = 400$m. The size of $\Omega_L$ is chosen in such a way that radiating circular waves do not reach the outflow boundaries of $\Omega_L$ during the simulation.

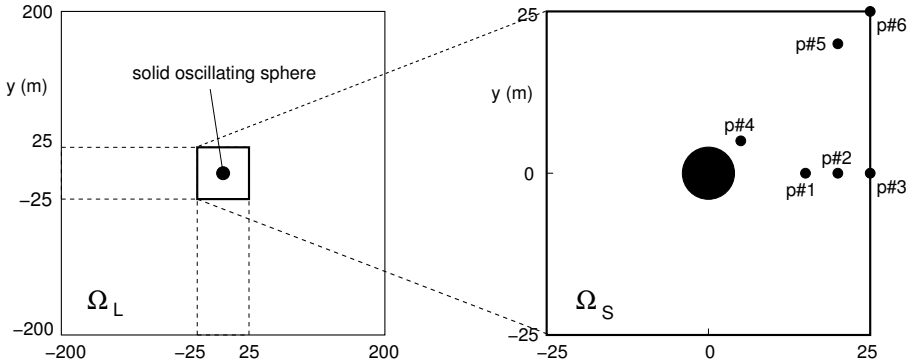



Figure 4: Setup for the test with the oscillating solid sphere showing the domain sizes (left) and the positions of the wave probes (right). The solution in $\Omega_L$ is considered as the reference solution for the solution in $\Omega_S$.

Simulations are performed for $30$s at two uniform grid resolutions of $0.25$m and $0.50$m. Several probes are placed in the domain to compare the free surface elevation records; see Fig. 4. Due to the prescribed motion of the sphere the generated circular wave is regular, and both the ABC-1 and ABC-2 are tuned to absorb this regular wave. But because the wave is circular, it reaches the outflow boundaries at different angles at different positions. To account for this directional effect, the angle coefficients in both boundary conditions are set to $45^\circ$ (combined with $0^\circ$ for the second order condition).

Figure 5 shows the free surface elevation history from the first three probes on both grid resolutions. The results in the other three probes are comparable [7]. As the circular concentric wave travels out from the center, the amplitude of the wave decreases because the energy of the wave is spread over a larger area. The probes show a significant difference between the

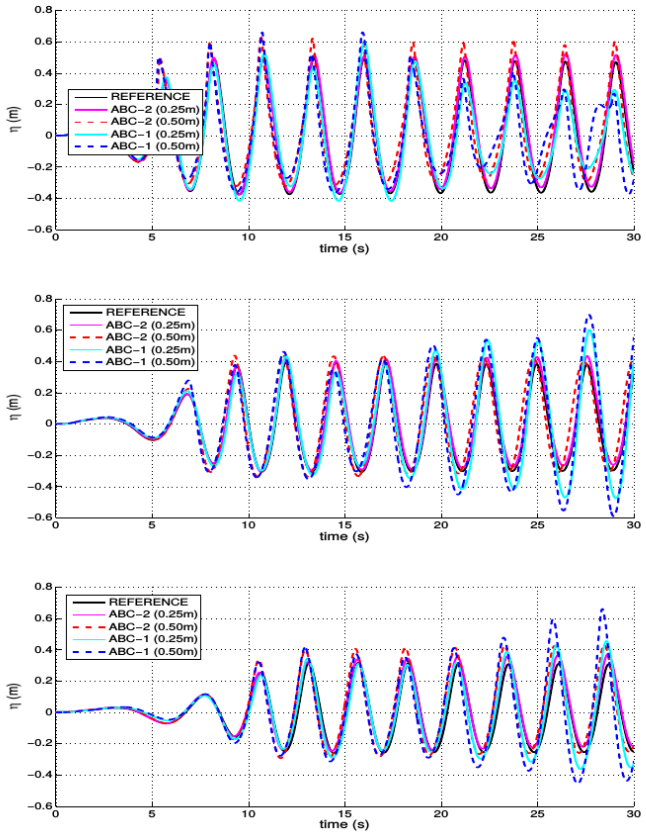
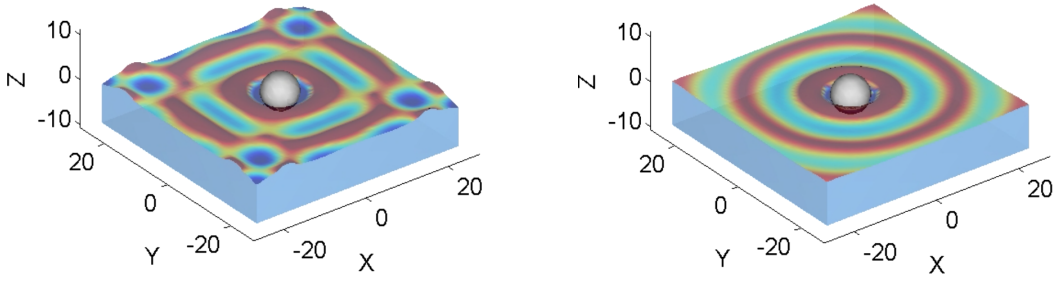



Figure 5: Free surface elevations as a function of time at various locations. From top to bottom results are shown at p#1, p#2 and p#3 [7].

performance of the ABC-1 and ABC-2, and clearly demonstrate the superiority of the ABC-2 over ABC-1.

Fig. 6 shows snapshots of the simulations at $t \approx 25$s when ABC-1 and ABC-2 are used on the fine grid. With ABC-1 different amounts of reflection at different locations result in a graphically-interesting but highly disturbed free surface. In contrast, with ABC-2 the free surface is considerably less disturbed.




(a) ABC-1 (b) ABC-2

Figure 6: Snapshots of the simulations of an oscillating sphere at $t \approx 25$s (fine grid with resolution of $0.25$m). The colors correspond with the vertical velocity at the free surface.

## 4 FLUID–SOLID BODY COUPLING

A further step is to allow the moving object to interact with the fluid dynamics, e.g. it is floating on the water surface. This physical two-way coupling has to be mirrorred in the numerical coupling algorithm between the flow solver and the solid-body solver. The coupling takes place along the common interface $\Gamma$ between the fluid and the solid body; quantities involved are the dynamics (position and acceleration) of the solid body, and the dynamics of the fluid (in particular its pressure forces); see Fig. 7.

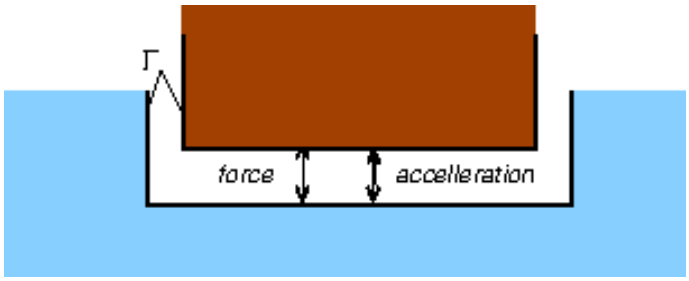



Figure 7: Schematic partitioning between fluid and solid body, with information exchange along their common interface $\Gamma$.

1. The coupling conditions at the interface between fluid and solid body basically express 'continuity' of the physics on both sides of the common interface $\Gamma$. The *kinematic* condition expresses that the boundary of the liquid region (partially) coincides with the surface of the solid body. In particular, material particles on both sides have the same velocity and acceleration (when the no-slip condition does not hold, only the normal component is continuous). Hence, it makes sense to talk about the velocity at the interface $\mathbf{u}_\Gamma$. Additionally, the *dynamic* coupling condition is based on Newton's $3^{rd}$ law "*action* $= -$*reaction*", which expresses equilibrium of forces. In particular, we denote the fluid pressure by $p_\Gamma$. In the sequel, we will formulate the coupled problem in terms of the two interface variables: $\mathbf{u}_\Gamma$ and $p_\Gamma$.

2. The solid-body dynamics is governed by an equation describing the accelaration $\dot{\mathbf{u}}$ of the solid body when reacting to the forces and moments $\mathbf{f}$ excerted by the fluid. The latter are found by integration of the liquid pressure over the common interface $\Gamma$. In abstract terms we denote this relation by

$$\text{solid body dynamics:} \qquad \dot{\mathbf{u}}_\Gamma = D p_\Gamma. \tag{15}$$

   The dynamics operator $D$ involves the integration of the fluid pressure over the interface to obtain the force and moments acting on the body. Through the dynamics equations of the solid body these lead to an acceleration of the body. Thus the eigenvalues of $D$ are inversely related to the body mass and its moments of inertia. Note that we do not show eventual external forces, as they play no essential role in the coupling algorithm.

3. The fluid dynamics governs the reaction of the fluid to the motion of the solid body. The latter creates a boundary condition along the interface $\Gamma$ which has to be added to the Navier–Stokes equations (1). As a result, the pressure along the interface and acting on the solid body can be computed. In abstract notation we write

$$\text{fluid dynamics:} \qquad p_\Gamma = -M_a \dot{\mathbf{u}}_\Gamma, \tag{16}$$

where $M_a$ is, by definition, the so-called *added mass* operator. Again, terms in $\mathbf{u}$ have been omitted because of their minor role in the analysis of the coupling process.

The above formulation (15)+(16) in principle shows two equations for the two unknowns along the interface: pressure and acceleration. Their coupling can be done in an *aggregated / monolithic* or *segregated / partitioned* way. An aggregated (or strong) coupling recombines both equations (or modules) into one single global system which is solved simultaneously. In contrast, a segregated (or weak) coupling keeps the two equations apart and solves them in an iterative way. The former approach may not always be possible, e.g. due to their 'black-box' character, or due to the large complexity of the modules; the latter approach may diverge.

### 4.1 Weak coupling

In practice often a segregated approach is followed. The usual way is to let the fluid determine a pressure field $p_\Gamma^{\rm old}$ which moves the solid body:

$$\dot{\mathbf{u}}_\Gamma^{\rm new} = D p_\Gamma^{\rm old}. \tag{17}$$

This motion of the solid body is then transferred to the fluid to react with a new pressure field:

$$p_\Gamma^{\rm new} = -M_a \dot{\mathbf{u}}_\Gamma^{\rm new}. \tag{18}$$

In this way, we effectively have created an iterative process

$$p_\Gamma^{\rm new} = -M_a D p_\Gamma^{\rm old}. \tag{19}$$

The iterations in such a weak coupling method will converge if and only if the spectral radius of the iteration matrix $\rho(M_a D) < 1$. Recalling that $D$ inversely scales with the mass and inertia of the solid body, this condition becomes a requirement for the ratio between added mass and body mass. Roughly speaking, the solid body should be heavy enough. If it is not, underrelaxation can help achieving convergence, but this will require (many) additional (sub)iterations.

In practice, such iterations can be implemented as follows in the solution process for the Navier–Stokes equation (3) during the time step from $n \to n+1$. An additional subiteration process (with iteration count $k$) is included:

$$\text{solid-body dynamics:} \qquad (\dot{\mathbf{u}}_\Gamma^{n+1})^k = D(p_\Gamma^{n+1})^k; \tag{20}$$

$$\text{Navier–Stokes:} \qquad (p_\Gamma^{n+1})^{k+1} = M_a(\dot{\mathbf{u}}_\Gamma^{n+1})^k. \tag{21}$$

If necessary, these subiterations can be made convergent by applying (severe) underrelaxation without disturbing the time accuracy of the time integration method, but often at a considerable computational price.

Below we will demonstrate with an example of a falling life boat how many subiterations can be required. Because of this inefficiency, we would like to apply an aggregated approach (which does not require subiterations), or at least to be as close as possible to such an approach. Such an approach is the quasi-simultaneous method, originally developed for interacting aerodynamic boundary layers along airplane wings [28]; a historic overview is provided in [29, 30].

### 4.2 Quasi-simultaneous coupling

For iterative efficiency reasons, we would like to follow the monolithic, simultaneous approach as much as possible. With the full dynamics operator $D$ being too complex, we look for a good approximation $\tilde{D}$ which is simple enough to be used as a boundary condition inside the Navier–Stokes solver (1). Such an approximation has been termed an *interaction law* [28–30]

$$\dot{\mathbf{u}}_\Gamma = \tilde{D} p_\Gamma. \tag{22}$$

The interaction law (22) is then incorporated in the time integration process in defect formulation:

$$\text{interaction law:} \qquad \dot{\mathbf{u}}_\Gamma^{n+1} - \tilde{D} p_\Gamma^{n+1} = (D - \tilde{D}) p_\Gamma^n \tag{23}$$

$$\text{Navier–Stokes:} \qquad p_\Gamma^{n+1} - M_a \dot{\mathbf{u}}_\Gamma^{n+1} = 0. \tag{24}$$

Another interpretation in terms of time integration is that the part $\tilde{D}$ of the dynamics equation (15) is treated implicitly and the remaining part $D - \tilde{D}$ explicitly.

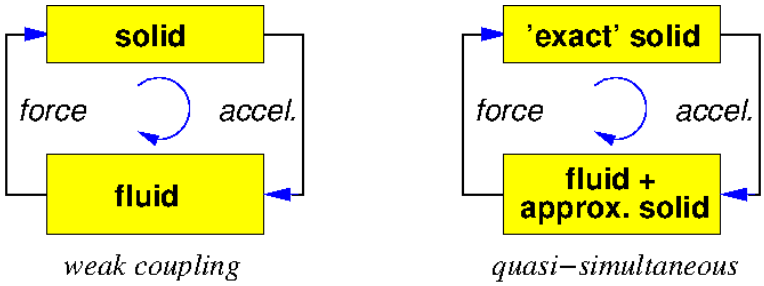



Figure 8: Weak versus quasi-simultaneous coupling.

The interaction law is a boundary condition for the Navier–Stokes equations along the interface $\Gamma$. More precisely, it will be implemented as a boundary condition for the pressure Poisson equation (6). The latter is derived by first rewriting the boundary condition (23) for $\mathbf{u}$ as

$$\mathbf{u}_\Gamma^{n+1} - \delta t \, \tilde{D} p_\Gamma^{n+1} = \mathbf{u}_\Gamma^n + \delta t \, (D - \tilde{D}) p_\Gamma^n.$$

Then the discrete momentum equation is written in the form (2), wherein the above relation is substituted through the term $M^\Gamma \mathbf{u}$. Finally a relation is obtained featuring $p_\Gamma$ as a boundary condition for the Poisson equation; see also [31].

The above *quasi-simultaneous* integration can be analysed by eliminating $\dot{\mathbf{u}}$ from (23)+(24), which leads to

$$(I + M_a \tilde{D}) p_\Gamma^{n+1} = -M_a (D - \tilde{D}) p_\Gamma^n. \tag{25}$$

For $\tilde{D} \equiv 0$, we recover the iteration process (19) which breaks down when $M_a$ is 'too large'. But with the term $M_a \tilde{D}$ on the left-hand side, it will be clear that this process will converge when $\tilde{D}$ is sufficiently close to $D$, in spite of a possibly 'large' $M_a$.

### 4.3 Example

To show the performance of the quasi-simultaneous approach, we now present an example of a the simulation of a life boat falling into a large breaking wave. A snapshot of the simulations

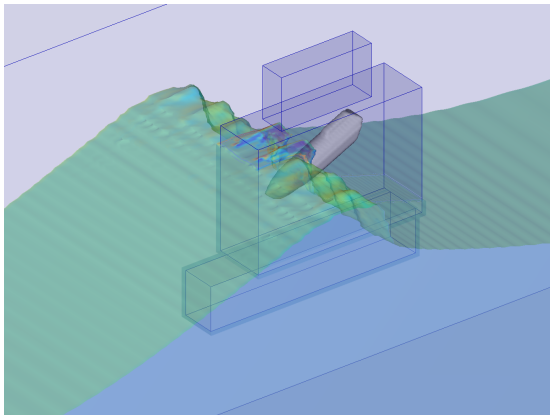


Figure 9: A snapshot of the life boat falling into a large wave. Also the locally-refined grid regions are shown.

is shown in Fig. 9. The dynamics of the life boat is modelled by means of a 6-DOF mechanical model.

The fluid flow is modelled with the Navier-Stokes equations and solved on a grid consisting of about 0.7 million active (i.e. within the fluid) grid points, with local grid refinement [12] around the life boat (Fig. 9). For physical accuracy this grid is rather coarse, but the focus in these simulations is on the numerical behaviour of the coupling process. Thus both the weak coupling procedure (19) as well as the quasi-simultaneous procedure (24) have been applied. In the latter case, the interaction law is based on the under-water part of the lifeboat (as the Poisson equation is only solved under water).

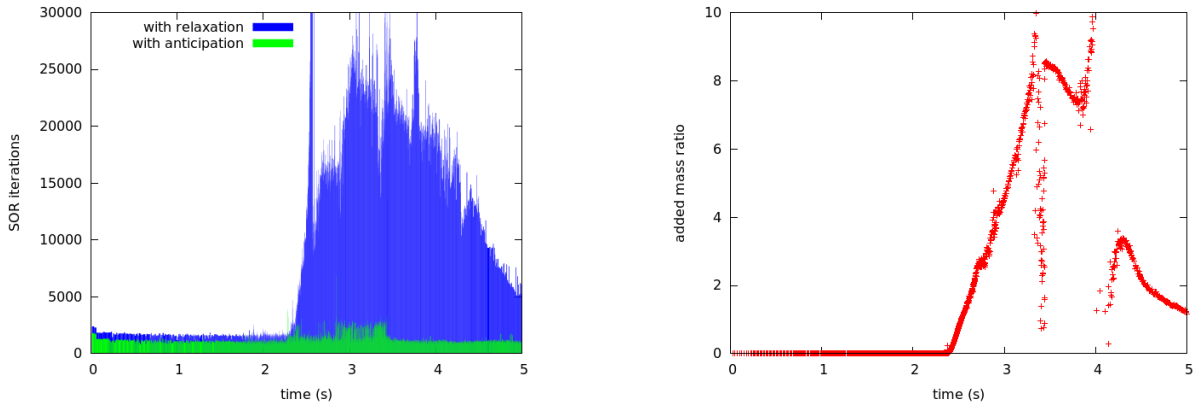



Figure 10: *Left:* The number of SOR iterations per time step for the underrelaxed weak coupling method (blue) and the anticipating quasi-simultaneous method (green)). *Right:* The estimated added mass for the falling lifeboat as a function of time. The moment of impact is clearly visible.

The most important result concerns the amount of work that is needed per time step to achieve the coupling between solid-body dynamics and fluid flow. The weak method often requires dozens of subiterations, in each of which a Poisson equation has to be solved. This number is dependent on the amount of fluid that is moved aside by the moving body, represented by the added-mass operator $M_a$. Fortunately, the later subiterations have a good initial guess so they are not as expensive as the earlier ones. Thus the amount of work is better represented by the total number of SOR-iterations [32] that is needed for all Poisson solves within one time step. This amount of work is shown in Fig. 10(left). The relation with the added mass becomes visible when plotting the time history of the estimated added mass in Fig. 10(right).

Comparison with Fig. 10(left) shows clearly that the number of iterations grows rapidly when the 'added-mass ratio' $\rho(DM_a)$ growths beyond 1. In contrast, the quasi-simultaneous method requires 1 or 2 subiterations (with additional Poisson solver), resulting in much less work per time step (Fig. 10).

## 5 WAVE IMPACT

As a third we present simulations and experiments for run-up against a semi-submersible offshore platform. The setup of the experiment, carried out at MARIN on a scale of $1:50$, is sketched in Fig. 11. A simplified semi-submersible is located in the center of the flow domain. On full scale it measures $114.5\,\text{m} \times 17.5\,\text{m} \times 28.0\,\text{m}$ with a draft of $16.0\,\text{m}$. The waves are generated by a flap-type wave generator. The basin width is $4\,\text{m}$, i.e. $200\,\text{m}$ on full scale, with solid side walls. The incoming wave has a full scale wave height of $14.7\,\text{m}$ and a wave period of $11.0\,\text{s}$. The wave elevation and pressure are measured at several positions, indicated in Fig. 11.




Figure 11: Wave run-up against a semi-submersible: snapshot of experiment (left) and position of sensors (right).

To reduce computational costs, the boundaries of the flow domain are brought closer to the structure than in the experiment, with an inflow boundary located at $240\,\text{m}$ from the center of the semi-submersible. To account for this decreased distance between the wave maker and the semi-submersible in the simulations, the incoming waves have been analyzed by wave calibration tests (without semi-submersible in the flow) [33]. The incoming computational wave thus is modeled as a $5^{th}$-order Stokes wave. Note that the shape of the experimental wave differs from a theoretical Stokes wave: the wave crests can be made to correspond nicely, but a difference in the troughs is left (see Fig. 13). This has to be taken into account when comparing the simulations with the model tests.

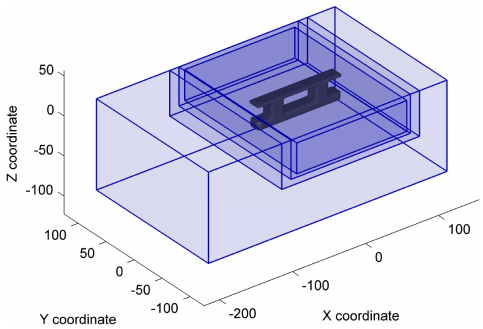



Figure 12: The grid with local refinements around the free surface and around the semi-submersible. Each box contains a refined grid level (no grid lines are shown).

The ComFLOW simulations used a one-phase flow model (as shown above) with a modern minimum-dissipation turbulence LES model [20]. Central spatial discretization with Adams-Bashforth time stepping is used throughout. The numerical simulations are shown for two different, stretched grids: $180 \times 40 \times 60 = 430$k and $360 \times 80 \times 120 = 3.5$M grid cells. The grid sizes near the semi-submersible are about $80$ cm and $40$ cm, respectively, in all three directions. An absorbing boundary condition, as discussed above, is applied at the outflow boundary and the simulations are carried out for several wave periods (5-10 periods, depending upon the size of the computational grid). Also results for a locally-refined grid [12, 13] are shown, with cells of around $10$ m away from the structure and four refinement levels, down to $50$ cm near the object, making a total of $0.5M$ cells (Fig. 12).

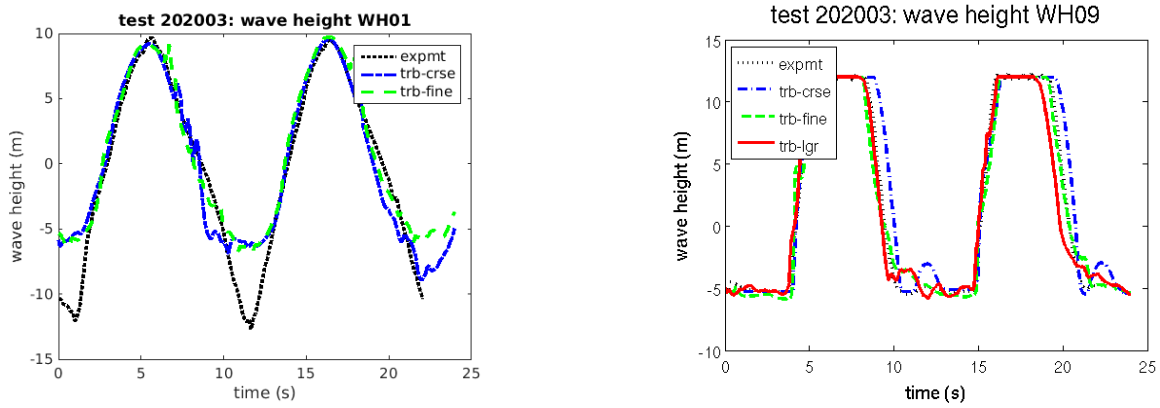



Figure 13: The wave height development in front of the semi-submersible (WH01) and at the first column (WH09) for several grids compared with experimental data.

In the presentation of results we focus on the region in front of the first column, i.e. wave-height probes WH01 and WH09, and pressure probes P11 and P15; see Fig. 11. Firstly, in Fig. 13 the wave run-up against the column is shown for the first two wave periods in the simulation. The results from the locally-refined simulation are pretty similar to those of the finest grid, although the amount of grid points is about $7\times$ less.

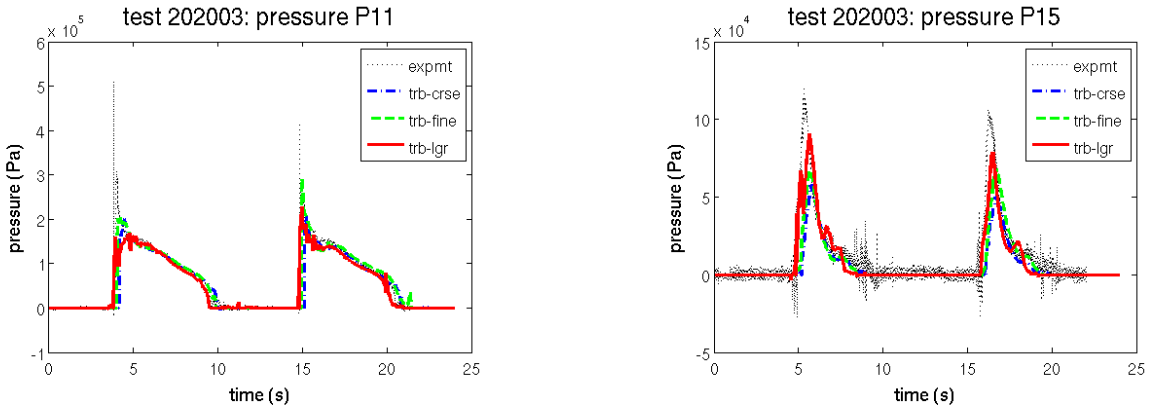



Figure 14: The pressure at the first column for several grids compared with experimental data. Left: P11 near the foot of the column; right: P15 near the top.

The pressure development near bottom (P11) and top (P15) of the first column of the semi-submersible is shown in Fig. 14. Near the top of the first column, at transducer P15, the peak

pressure values in the simulations are relatively low, which can be attributed to the limited number of grid cells along the column. The locally-refined grid, with its higher resolution near the semi-submersible, seems to catch the pressure peaks somewhat better. More detailed simulations can be found in the PhD thesis of Wemmenhove [5, 34] and the forthcoming PhD thesis of Van der Plas [12, 13].

## 6 CONCLUSIONS

Some recent developments in the ComFLOW simulation method have been presented. It is designed to simulate and study extreme waves and their impact on falling, floating and moored structures. Firstly, an absorbing boundary condition has been presented that adapts itself to the (local) wave number and phase speed of dispersive waves. Waves generated by an oscillating buoy show that the computational boundaries can be put close to the wave-generating object. Secondly, a life boat falling into an extreme wave reveals that the interaction between solid-body dynamics and fluid dynamics poses serious numerical challenges for their numerical coupling. It is demonstrated that a quasi-simultaneous coupling strategy can efficiently handle these challenges. Thirdly, simulations of large turbulent waves against a fixed semi-submersible, validated through experiments, demonstrate the efficiency of local grid refinement.

## 7 ACKNOWLEDGMENTS

The research is supported by the Dutch Technology Foundation STW, applied science division of NWO and the technology programma of the Ministry of Economic Affairs in The Netherlands (contracts 10475 and 13267).

## REFERENCES

[1] K.M.T. Kleefsman, G. Fekken, A.E.P. Veldman, B. Iwanowski and B. Buchner, A Volume-of-Fluid based simulation method for wave impact problems'. *Journal of Computational Physics*, **206**, 363–393, 2005.

[2] A.E.P. Veldman, J. Gerrits, R. Luppes, J.A. Helder and J.P.B. Vreeburg, The numerical simulation of liquid sloshing on board spacecraft. *Journal of Computational Physics*, **224**, 82–99, 2007.

[3] A.E.P. Veldman, R. Luppes,T. Bunnik, R.H.M. Huijsmans, B. Düz, B. Iwanowski, R. Wemmenhove, M.J.A. Borsboom, P.R. Wellens, H.J.L. van der Heiden and P. van der Plas (2011). Extreme wave impact on offshore platforms and coastal constructions'. *Proceedings 30th Conference on Ocean, Offshore and Arctic Engineering OMAE2011*, Rotterdam (The Netherlands), 19–24 June, 2011, paper OMAE2011-49488.

[4] A.E.P. Veldman, R. Luppes, H.J.L. van der Heiden, P. van der Plas, B. Düz and R.H.M. Huijsmans (2014). Turbulence modeling, local grid refinement and absorbing boundary conditions for free-surface flow simulations in offshore applications. *Proceedings 33rd International Conference on Ocean, Offshore and Arctic Engineering*, San Francisco (USA), 8–13 June, 2014, paper OMAE2014-24427.

[5] R. Wemmenhove, R. Luppes, A.E.P. Veldman and T. Bunnik, Numerical simulation of hydrodynamic wave loading by a compressible two-phase flow method. *Computers and Fluids*, **114**, 218–231, 2015 (doi: 10.1016/j.compfluid.2015.03.007).

[6] B. Düz, R.H.M. Huijsmans, A.E.P. Veldman, M. Borsboom and P. Wellens, An absorbing boundary condition for regular and irregular wave simulations. In: L. Eça, E. Oñate, J. Garçia-Espinosa, T. Kvamsdal and P. Bergan (Eds.), *Computational Methods in Applied Sciences*, **29**, Springer, 2014.

[7] B. Düz, *Wave generation, propagation and absorption in CFD simulations of free surface flows.* PhD thesis, Technical University Delft, 2015.

[8] B. Düz, M.J.A. Borsboom, A.E.P. Veldman, P.R. Wellens, R.H.M. Huijsmans, The effect of different volume-of-fluid (VOF) methods on energy dissipation in simulations of propagating water waves. In preparation.

[9] R.W.C.P. Verstappen, When does eddy viscosity damp subfilter scales sufficiently? *Journal of Scientific Computing*, **49**, 94–110, 2011.

[10] W. Rozema, *Low-dissipation methods and models for the simulation of turbulent subsonic flow.* PhD thesis, University of Groningen, 2015.

[11] W. Rozema, H.J. Bae, P. Moin and R. Verstappen, Minimum-dissipation models for large-eddy simulation. *Physics of Fluids* **27**, 085107, 2015.

[12] P. van der Plas, A.E.P. Veldman, H.J.L. van der Heiden and R. Luppes, Adaptive grid refinement for free-surface flow simulations in offshore applications. *Proceedings 34th International Conference on Ocean, Offshore and Arctic Engineering OMAE2015*, St. John's (Canada), 31 May–5 June, 2015, paper OMAE2015-42029.

[13] P. van der Plas, *Local grid refinement for free-surface flow simulations.* PhD thesis, University of Groningen, 2016.

[14] B. Engquist and A. Majda, Absorbing boundary conditions for the numerical simulation of waves". *Mathematics of Computation*, **31**, 629–651, 1977.

[15] D. Givoli, Non-reflecting boundary conditions. *Journal of Computational Physics*, **94**, 1–29, 1991.

[16] S. Tsynkov, Numerical solution of problems on unbounded domains. a review. *Applied Numerical Mathematics*, **27**, 465–532, 1998.

[17] T. Hagstrom, Radiation boundary conditions for the numerical simulation of waves. *Acta Numerica*, **8**, 47–106, 1999.

[18] D. Givoli, High-order local non-reflecting boundary conditions: a review. *Wave Motion*, **39**, 319–326, 2004.

[19] P. Wellens, *Wave simulation in truncated domains for offshore applications*. PhD thesis, Delft University of Technology, 2012.

[20] H.J.L. van der Heiden, A.E.P. Veldman, R. Luppes, P. van der Plas, J. Helder and T. Bunnik, Turbulence modeling for free-surface flow simulations in offshore applications. *Proceedings 34th International Conference on Ocean, Offshore and Arctic Engineering OMAE2015*, St. John's (Canada), 31 May–5 June, 2015, paper OMAE2015-41078.

[21] A.E.P. Veldman, R. Luppes, P. van der Plas, H.J.L. van der Heiden, J. Helder and T. Bunnik (2015). Turbulence modeling for locally-refined free-surface flow simulations in offshore applications. *Proceedings Int Symp Offshore and Polar Eng ISOPE2015*, Kona (Hawaii, USA), 23-27 June, 2015, paper 2015-TPC-0282.

[22] R.W.C.P. Verstappen and A.E.P. Veldman, Symmetry-preserving discretization of turbulent flow. *Journal of Computational Physics*, **187**, 343–368, 2003.

[23] C.R. Hirt and B.D. Nichols, Volume of fluid (VOF) method for the dynamics of free boundaries. *Journal of Computational Physics*, **39**, 201–225, 1981.

[24] A. Sommerfeld, *Partial differential equations in physics*. Academic Press, 1949.

[25] R. Higdon, Absorbing boundary conditions for difference approximations to the multi-dimensional wave equation. *Mathematics of Computation*. **47**, 437–459, 1986.

[26] R. Higdon, Numerical absorbing boundary conditions for the wave equation. *Mathematics of Computation*, **49**, 65–90, 1987.

[27] B. Düz, M.J.A. Borsboom, P.R. Wellens, A.E.P. Veldman, R.H.M. Huijsmans, An absorbing boundary condition for free-surface water waves. In preparation.

[28] A.E.P. Veldman, New quasi-simultaneous method to calculate interacting boundary layers. *AIAA Journal*, **19**, 79–85, 1981.

[29] A.E.P. Veldman, Matched asymptotic expansions and the numerical treatment of viscous-inviscid interaction. *Journal of Engineering Mathematics*, **39**, 189–206, 2001.

[30] A.E.P. Veldman, A simple interaction law for viscous-inviscid interaction. *Journal of Engineering Mathematics*, **65**, 367–383, 2009 (doi: 10.1007/s10665-009-9320-0).

[31] A.E.P. Veldman, Missing boundary conditions? discretize first, substitute next, combine later. *SIAM Journal of Scientific and Statistical Computing,*. **11**, 82–91, 1990.

[32] E.F.F. Botta and M. Ellenbroek, A modified sor method for the poisson equation in unsteady free-surface flow calculations. *Journal of Computational Physics*, **60**, 119–134. 1985.

[33] B. Iwanowski, M. Lefranc and R. Wemmenhove, CFD simulation of wave run-up on a semi-submersible and comparison with experiment.*Proceedings 28th Conference on Ocean, Offshore and Arctic Engineering OMAE2009*, Honolulu (USA), 31 May–5 June, 2009, paper OMAE2009-79052.

[34] R. Wemmenhove, *Numerical simulation of two-phase flow in offshore environments*. PhD thesis, University of Groningen, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MODELLING OF UNSTEADY SECONDARY VORTICES GENERATED BEHIND THE RADIAL GAP OF THE AXIAL TURBINE BLADE WHEEL

**Straka P.**[1]

[1]Aerospace Research and Test Establishment, Plc.
Beranových 130, 19905 Prague, Czech Republic
e-mail: straka@vzlu.cz

**Keywords:** EARSM, hybrid RANS/LES, axial turbine stage, secondary flows.

**Abstract.** *The paper deals with application of the linear eddy viscosity model, the nonlinear explicit algebraic Reynolds stress model and the hybrid RANS/LES model for computation of flow in an axial turbine stage with radial gaps under the stator and above the rotor blade. Results show that the nonlinear explicit algebraic Reynolds stress model allows to describe deformation of the secondary vortices and their breakdown into smaller structures while the linear eddy viscosity model doesn't have this capability. Only hybrid RANS/LES model describes decay of the secondary vortices which interact with the rotor blade. This capability is necessary for good prediction of distribution of the turbine stage efficiency behind the rotor blade.*

## 1 INTRODUCTION

The axial steam turbine of low power (around hundreds kilowatts to units megawatts) are typically designed in configuration with the drum-type rotor. It means that the rotor blades are carried directly by the rotor of large diameter as it is depicted in figure 1. In case of low power turbines, in order to reduce a production costs a blades with free ends that are not equipped with the shroud are still in use. For more reduction of the production cost the prismatic blades are sometimes used as it is in the present work. In this configuration it is necessary to maintain the radial gap under the hub-end of the stator blades and above the tip-end of the rotor blades.

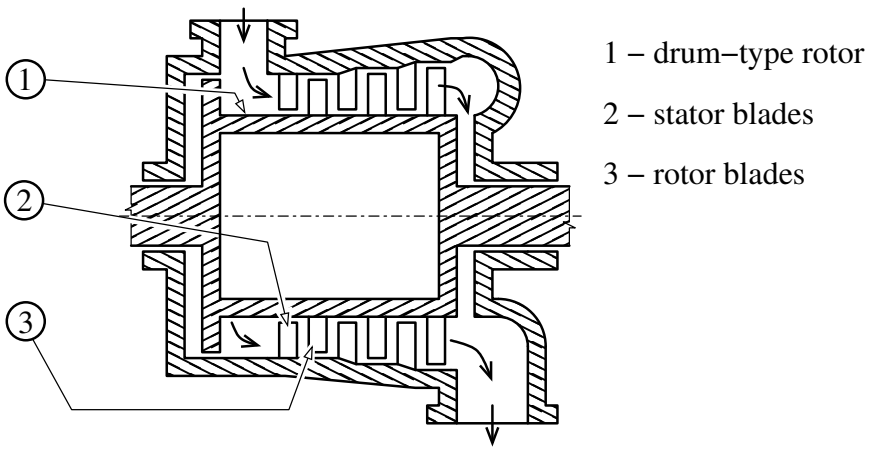



Figure 1: Scheme of the low power axial steam turbine with the drum-type rotor.

Flow through the radial gap generates an intensive secondary vortices which interacts with following blade wheel. This interaction leads to strong swirl of flow field in the peripheral parts of the blade span and it has major impact on the efficiency of the turbine stage. In [1] there was shown that drop in the efficiency due to interaction of the rotor blades with the secondary flows is up to ten percent. Prediction of the interaction of blade with the secondary flows strongly depends on used turbulence model. In industrial applications, such as this, it is usual to solve system of RANS (Reynolds Averaged NavierStokes) equations, which is closed by two-equation turbulence model (such as $k-\omega$). In [2] there was studied mechanism of the interaction between the rotor blades and the secondary vortices generated behind the radial gap under the stator blade using the $k-\omega$ model of Kok (2000) [3]. This model is based on linear relation for the Reynolds stress tensor. Although this model has provided interesting results, comparison with the experimental data revealed some differences in span-wise distribution of the efficiency behind the rotor blade. Therefore in [4] there was compared prediction of the secondary vortices development using the linear turbulence model of Kok (2000) [3] and the nonlinear EARSM (Explicit Algebraic Reynolds Stress Model) model of Rumsey & Gatski (2001) [5]. It was found that the nonlinear EARSM model predicts deformation and unsteady behaviour of the secondary vortices, while the linear model doesn't have this capability. However, for better prediction of the interaction between the rotor blade and the secondary vortices generated behind the radial gap under the stator blade it is necessary to model not only deformation of large vortices but also their decay. Therefore in this work a hybrid RANS/LES model is used for modeling of the rotor blade/secondary vortices interaction beside the LEVM (Linear Eddie Viscosity Model) model of Kok (2000) and the nonlinear EARSM model of Rumsey and Gatski (2001).

## 2 GEOMETRIC SETUP

Figure 2 shows the computational domain with marked inlet and outlet boundaries and sliding mesh interface between stator and rotor parts of the computational domain. It is seen that the computational domain contains only one stator and one rotor blade. Such a reduction of the actual number of blades is achieved by method of scaling. The scale factor of the stator blade is given by $2N_s/(N_s + N_r)$, where $N_s$ is real number of the stator blades and $N_r$ is real number of the rotor blades. The scale factor of the rotor blade is defined analogously. In our case the scale factor is $0.9\overline{09}$ for the stator blade and $1.\overline{09}$ for the rotor blade. Note that scaling has no effect on static aerodynamics loading of the blades, because the pitch to chord ratio of blades is preserved. Scaling of the blade profile causes little change of the chord based Reynolds number, but it is minor problem in case of values of scale factors near 1. The method of scaling has, however, major impact on prediction of dynamic aerodynamic loading of the blades. Therefore this method is unusable in cases where such that prediction is a main scope. In our case we want to compare prediction of the rotor blade interaction with secondary vortices generated behind the radial gap under the hub-end of the stator blade using turbulence models with different closure formulas. So the scaling method is applicable here. After scaling, the chord of the stator blade is 29.0 mm and the chord of the rotor blade is 28.1 mm. The hub diameter is 330 mm and diameter of the outer casing is 430 mm. Between the hub-end of the stator blade and the hub-wal there is the radial gap of 1.4 mm. Also between the tip-end of the rotor blade and the outer casing there is the radial gap of 1.4 mm. The axial distance between the stator trailing edge and the rotor leading edge is 8 mm. The inlet boundary is placed 22 mm in front the stator leading edge and the outlet boundary is placed 22 mm behind the rotor trailing edge. The sliding mesh interface between stator and rotor parts of the computational domain is placed in the middle of the axial gap between the stator and the rotor blades.

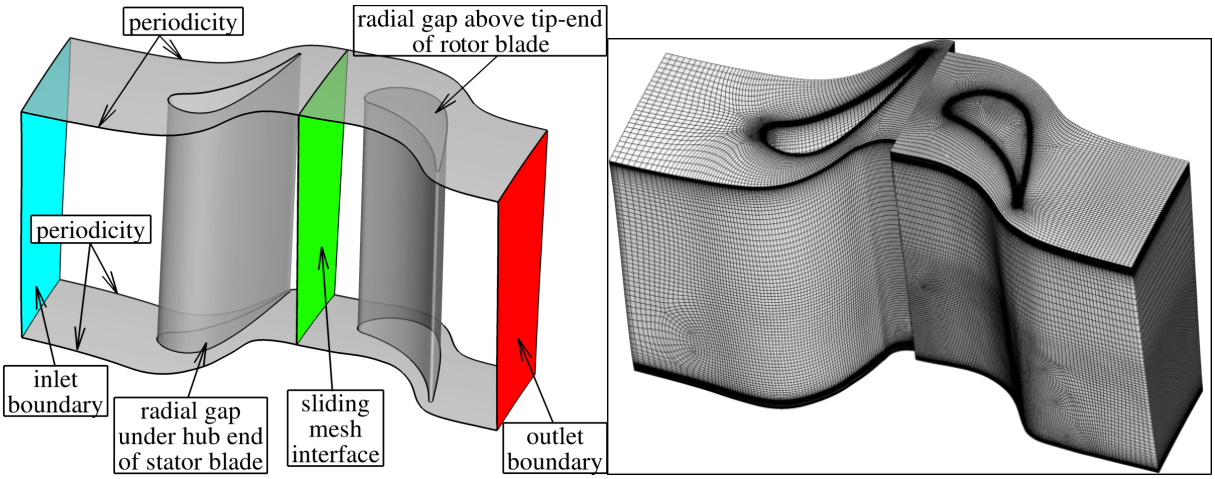



Figure 2: Scheme of the computational domain (left); computational mesh (right).

## 3 FLOW CONDITIONS

Presented results were calculated for the total/static expansion ratio 1.4, the velocity ratio $u/c_{is} = 0.6$ ($u$ is the circumferential velocity at middle diameter and $c_{is}$ is the isentropic velocity at the hub diameter behind the rotor trailing edge). It corresponds with the inlet total pressure 120 kPa, the outlet pressure 85.7 kPa, speed of revolutions 7430 RPM, the total inlet temperature

of 303 K. The isentropic outlet Mach number is $M_{is} = 0.7$, while the outlet isentropic Reynolds number based on blade chord is $Re_{is} \approx 3 \times 10^5$.

## 4 NUMERICAL APARATUS

Flow through the axial turbine stage is modeled as unsteady, 3D, compressible, viscous, fully turbulent flow of the perfect gas. Simulation is performed using the in-house numerical code based on solution of system of the RANS equation and two-equation $k - \omega$ turbulence model. Used numerical code [6, 7] is focused on turbomachinery application. The interface between stator and rotor part of the computational domain is performed via the "sliding mesh" method, where communication between stationary stator domain and rotating rotor domain is based on interpolation.

System of the governing equations reads:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}\left(\rho u_j\right) = 0\,, \tag{1}$$

$$\frac{\partial}{\partial t}\left(\rho u_i\right) + \frac{\partial}{\partial x_j}\left(\rho u_j u_i + p\delta_{ij}\right) - \frac{\partial \tau_{ij}}{\partial x_j} = 0\,, \tag{2}$$

$$\frac{\partial e}{\partial t} + \frac{\partial}{\partial x_j}\left(\rho u_j(e+p) - u_k\tau_{kj} + q_j - d_j^t\right) = 0\,, \tag{3}$$

$$\frac{\partial}{\partial t}\left(\rho k\right) + \frac{\partial}{\partial x_j}\left(\rho u_j k - d_j^t\right) = P_k - D_k\,, \tag{4}$$

$$\frac{\partial}{\partial t}\left(\rho \omega\right) + \frac{\partial}{\partial x_j}\left(\rho u_j \omega - o_j^t\right) = P_\omega - \beta\rho\omega^2 + C_D, \tag{5}$$

where

$$p = (\kappa - 1)\left[e - \frac{\rho}{2}\left(u_1^2 + u_2^2 + u_3^2\right) - \rho k\right]\,,\; q_j = -\left(\frac{\mu}{\mathrm{Pr}} + \frac{\mu_t}{\mathrm{Pr_t}}\right)\frac{\kappa r}{\kappa - 1}\frac{\partial T}{\partial x_j}\,, \tag{6}$$

$$\tau_{ij} = \tau_{ij}^{lam} + \tau_{ij}^t\,,\; \tau_{ij}^{lam} = 2\mu\left(S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right)\,, \tag{7}$$

$$d_j^t = \left(\mu + \sigma_k \mu_t^{evm}\right)\frac{\partial k}{\partial x_j}\,,\; o_j^t = \left(\mu + \sigma_\omega \mu_t^{evm}\right)\frac{\partial \omega}{\partial x_j}\,, \tag{8}$$

$$P_k = \min\left(2\mu_t S^2,\, 2\mu_t \Omega^2,\, 20 D_k\right)\,,\; P_\omega = \gamma\frac{\omega}{k}P_k\,,\; C_D = \sigma_d\frac{\rho}{\omega}\max\left(\frac{\partial k}{\partial x_j}\frac{\partial \omega}{\partial x_j},\, 0\right)\,, \tag{9}$$

$$\mu = 1.72 \times 10^{-5}\left(\frac{T}{273.15}\right)^{0.754}\,, \tag{10}$$

$$\mu_t^{evm} = \rho\frac{k}{\omega}\,, \tag{11}$$

$$S = \sqrt{2S_{ij}S_{ij}}\,,\; \Omega = \sqrt{2\Omega_{ij}\Omega_{ij}}\,,\; S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\,,\; \Omega_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}\right)\,. \tag{12}$$

The turbulent viscosity $\mu_t$, the Reynolds stress tensor $\tau_{ij}^t$ and the destruction term $D_k$ will be described later in caption 5.

System of equations (1) to (5) we can formally rewrite in vector form as follows:

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}_j}{\partial x_j} = \mathbf{Q}\,, \tag{13}$$

where $\mathbf{W}$ is vector of the conservative variables, $\mathbf{F}$ is vector of the inviscid and viscous fluxes and $\mathbf{Q}$ is vector of the source terms. System (13) is discretized using the cell-centered finite-volume method on multi-block structured mesh of hexahedral cells. Temporal discretization is

performed using the second-orded backward Euler formula in implicit form, which is realized through a dual iterative process:

$$\left(\frac{3}{2}\mathbf{I} + \Delta t \, \frac{\partial \mathbf{R}_i^{n,\nu}}{\partial \mathbf{W}_i^{n,\nu}}\right) \Delta \mathbf{W}_i^{n,\nu+1/2} + \Delta t \sum_{j=1}^{6} \frac{\partial \mathbf{R}_i^{n,\nu}}{\partial \mathbf{W}_j^{n,\nu}} \Delta \mathbf{W}_j^{n,\nu+1/2} =$$

$$= -\frac{3\mathbf{W}_i^{n,\nu} - 4\mathbf{W}_i^{n} + \mathbf{W}_i^{n-1}}{2} - \Delta t \, \mathbf{R}_i^{n,\nu} \,, \tag{14}$$

where $\mathbf{W}^{n,0} := \mathbf{W}^n$, $\Delta \mathbf{W}^{n,\nu+1/2} = \Delta \mathbf{W}^{n,\nu+1} - \Delta \mathbf{W}^{n,\nu}$ and $\mathbf{W}^{n+1} := \mathbf{W}^{n,\nu^*}$, where $\mathbf{W}^{n,\nu^*}$ is a steady solution in the dual iterative process. In scheme (14) there $n$ means the time level, $\Delta t$ is the time step, $\nu$ is index of the dual iteration, $i$ is index of current cell in the computational mesh, $j$ is index of the neighbouring cells and $\mathbf{R}$ is the numerical approximation of the inviscid and the viscous fluxes:

$$\mathbf{R}_i = \frac{1}{V_i} \sum_{j=1}^{6} [\Phi^{inv}(\mathbf{W}_j^L,\ \mathbf{W}_j^R,\ \mathbf{n}_j) - \Phi^{vis}(\mathbf{W}_j^C,\ (\nabla \mathbf{W})_{D_j^{dual}},\ \mathbf{n}_j)] \, s_j \,. \tag{15}$$

The inviscid numerical fluxes $\Phi^{inv}$ are calculated using the exact solution of the 1D Rieman problem in normal direction to the cell edges. The viscous numerical fluxes $\Phi^{vis}$ are calculated using the central scheme and the gradient of the state vector $\nabla \mathbf{W}$ is calculated using the Green-Gauss theorem on a dual cells. Higher order of accuracy in space is obtained using linear reconstruction with the Van Leer's slope limiter.

The computational domain was discretized by structured multiblock mesh of 4.5 millions hexahedral cells which is displayed in figure 2.

## 5 CLOSURE FORMULA

As it was mentioned above, this work is focused on comparison of prediction of the blade/secondary flow interaction calculated using LEVM model of Kok (2000) [3], EARSM model of Rumsey and Gatski (2001) [5] and the hybrid RANS/LES method. The hybrid RANS/LES method according to Davidson and Peng (2003) [8] and Kok et al. (2004) [9] is used in this work. This approach uses one transport equation for the kinetic energy which plays role of the turbulent kinetic energy in RANS mode and the subgrid scale kinetic energy in LES mode. This kind of the hybrid RANS/LES methods is also called XLES (from eXtra Large Eddy Simulation).

### 5.1 LEVM model

This model is based on the Boussinesq approximation. The Reynolds stress tensor $\tau_{ij}^t$ is modeled as a linear function of the strain rate tensor $S_{ij}$:

$$\tau_{ij}^t = 2\mu_t \left(S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right) - \frac{2}{3}\rho k \delta_{ij} \,, \tag{16}$$

where the turbulent viscosity $\mu_t = \mu_t^{evm}$ is given by eq. (11). The destruction term $D_k$ in eq. (4) and (9) is given as

$$D_k = \beta^* \rho k \omega \,. \tag{17}$$

The model constants: $\beta^* = 0.09$, $\beta = 0.075$, $\gamma = 0.5532$, $\sigma_k = 0.5$, $\sigma_\omega = 0.5$ and $\sigma_d = 0.5$ corresponds with the TNT $k - \omega$ turbulence model proposed by Kok (2000) [3].

### 5.2 EARSM model

In this model according to Rumsey and Gatski (2001) [5] the Reynolds stress tensor $\tau_{ij}^t$ is modeled as a quadratic function of the strain rate tensor $S_{ij}$:

$$\begin{aligned}\tau_{ij}^t = 2\mu_t\left(S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right) - \frac{2}{3}\rho k\delta_{ij} + 2\mu_t\Big[a_2a_4\left(S_{ik}\Omega_{kj} - \Omega_{ik}S_{kj}\right) \\ - 2a_3a_4\left(S_{ik}S_{kj} - \tfrac{1}{3}S_{kl}S_{lk}\delta_{ij}\right)\Big]\,.\end{aligned} \tag{18}$$

It is clear that eq. (18) extends relation (16) about anisotropy part (term in square brackets). The turbulent viscosity $\mu_t$ is now given as

$$\mu_t = C_\mu\rho k\tau\,, \tag{19}$$

where the turbulent time scale $\tau$ is defined as

$$\tau = \max\left(\frac{1}{\beta^*\omega},\, C_\tau\sqrt{\frac{\mu}{\beta^*\rho k\omega}}\right)\,, \tag{20}$$

where $C_\tau = 6.0$ is a model constant. The nominal level of the variable coefficient $C_\mu$ in a zero-pressure-gradient log-layer is approximately 0.09. In general case value of $C_\mu$ is obtained by solving following cubic equation:

$$\left(-C_\mu\right)^3 + q_1\left(-C_\mu\right)^2 + q_2\left(-C_\mu\right) + q_3 = 0\,, \tag{21}$$

where

$$q_1 = -\frac{\gamma_1^*}{II_S\tau^2\gamma_0^*}\,,\; q_2 = \frac{1}{\left(2II_S\tau^2\gamma_0^*\right)^2}\left(\gamma_1^{*2} - 2II_S\tau^2\gamma_0^*a_1 - \tfrac{2}{3}II_S\tau^2a3_2 - 2II_\Omega\tau^2a_2^2\right), \tag{22}$$

$$q_3 = \frac{\gamma_1^*a_1}{\left(2II_S\tau^2\gamma_0^*\right)^2}\,,\; II_S = S_{kl}S_{lk}\,,\;\; II_\Omega = \Omega_{kl}\Omega_{lk}\,. \tag{23}$$

The correct root to choose from this equation is the root with the lowest real part. See [5] for more details. Other parameters are gived by:

$$a_1 = \tfrac{1}{2}\left(\tfrac{4}{3} - C_2\right)\,,\; a_2 = \tfrac{1}{2}\left(2 - C_4\right)\,,\; a_3 = \tfrac{1}{2}\left(2 - C_3\right)\,,\; a_4 = \tau/\left(\gamma_1^* + 2\gamma_0^*C_\mu S^2\tau^2\right), \tag{24}$$

$$\gamma_0^* = C_1^1/2\,,\;\; \gamma_1^* = \tfrac{1}{2}C_1^0 + \left(\tfrac{C_{\varepsilon2} - C_{\varepsilon1}}{C_{\varepsilon1} - 1}\right)\,, \tag{25}$$

where $C_{\varepsilon1} = 1.44$, $C_{\varepsilon2} = 1.83$, $C_1^0 = 3.4$, $C_1^1 = 1.8$, $C_2 = 0.36$, $C_3 = 1.25$ and $C_4 = 0.4$.

The destruction term $D_k$ in eq. (4) and (9) is given by eq. (17).

The model constants are now: $\beta^* = 0.09$, $\beta = 0.0747$, $\gamma = 0.53$, $\sigma_k = 1$, $\sigma_\omega = 0.5339$ and $\sigma_d = 0.5$.

### 5.3 XLES model

The hybrid RANS/LES method according to Davidson and Peng (2003) [8] and Kok et al. (2004) [9] switches between the system of RANS equations closed by two-equation turbulence model of Kok (2000) [3] and LES when computational mesh is fine enough to simulation of large turbulent eddies. The sub-grid scale model of LES is based on transport equation for the sub-grid scale turbulent energy given by eq. (4), where the destruction term $D_k$ is given as

$$D_k = \max\left(\beta^*\rho k\omega,\; \frac{\rho\sqrt{k^3}}{C_{DES}\Delta}\right)\,, \tag{26}$$

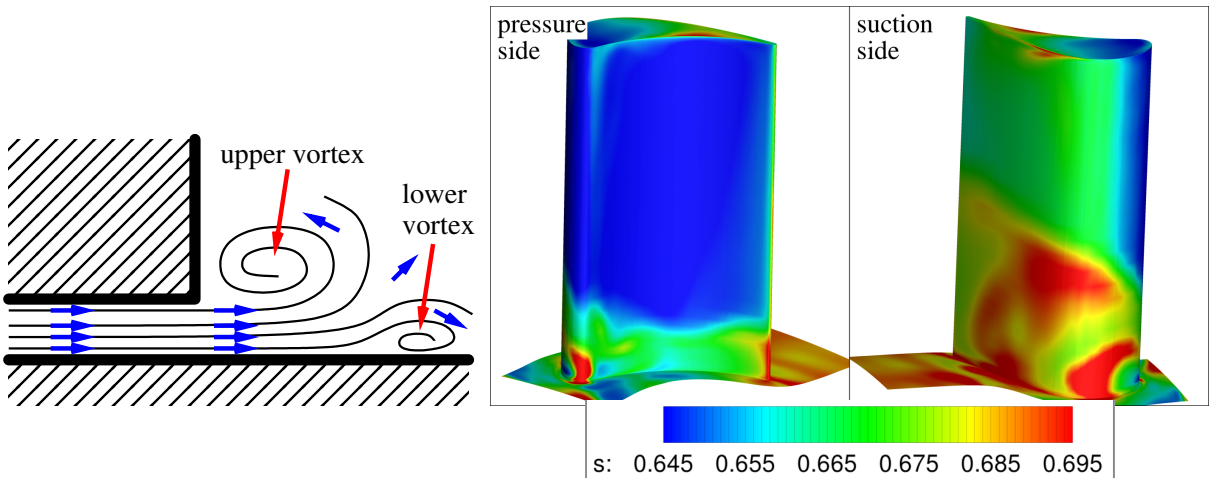



Figure 3: Scheme of flow behind the radial gap under the stator blade (left); distribution of the dimensionless entropy index $s = p/\rho^\kappa$ on the rotor blade surface (right).

where the local grid size $\Delta$ is defined by $\Delta = \max(\Delta_\xi, \Delta_\eta, \Delta_\zeta)$ where $\Delta_\xi$, $\Delta_\eta$ and $\Delta_\zeta$ are the distances between the cell faces in local $\xi$, $\eta$ and $\zeta$ grid line directions. Constant $C_{DES} = 0.6086$ is chosen according to Kubacki et al. (2013) [10]. Note that in XLES approach $k$ is either the turbulent energy in RANS mode or the sub-grid scale energy in LES mode.

The modeled stress tensor (Reynolds stress tensor in RANS mode, sub-grid scale stress tensor in LES mode) is given by eq. (16) where

$$\mu_t = \min\left(\rho k/\omega,\ \rho\beta^*\sqrt{k}C_{DES}\Delta_{LES}\right) \tag{27}$$

is either the turbulent viscosity in RANS mode or the sub-grid scale viscosity in LES mode. In eq. (27) there is $\Delta_{LES} = (\Delta_\xi\Delta_\eta\Delta_\zeta)^{1/3}$. It is evident that the solution is independent on equation (5) in LES mode although equation (5) is solved.

## 6 ADAPTATION OF NUMERICAL METHOD FOR XLES MODELLING

As it was mentioned above, the discretization scheme is based on the finite volume method in cell-centered form, where the inviscid numerical fluxes $\Phi^{inv}$ are calculated using the exact solution of the 1D Rieman problem in normal direction to the cell edges. Although numerical methods based on exact or approximate solution of the Riemann problem have good features for RANS modelling, they are too dissipative for using in LES approach. Therefore a central differencing is used in this work for suppression of the native numerical dissipation.

Let $\Phi^{inv}(\mathbf{W}^L_{i+1/2},\ \mathbf{W}^R_{i+1/2},\ \mathbf{n}_{i+1/2})$ is the inviscid numerical flux through cell face between i-th and (i+1)-th cells which is based on solution of the Riemann problem where $\mathbf{W}^L_{i+1/2}$ and $\mathbf{W}^R_{i+1/2}$ denotes state vectors extrapolated to the cell face from the left and right in means of linear reconstruction with the slope limiter. Let $\mathbf{W}^C_{i+1/2}$ is the state vector at the cell face computed as an average value of the state vectors in centers of i-th and (i+1)-th cells. The central differencing is obtained by replacing of $\mathbf{W}^L_{i+1/2}$ and $\mathbf{W}^R_{i+1/2}$ by $\mathbf{W}^{CL}_{i+1/2}$ and $\mathbf{W}^{CR}_{i+1/2}$ where $\mathbf{W}^{CL}_{i+1/2} = \psi\mathbf{W}^C_{i+1/2} + (1-\psi)\mathbf{W}^L_{i+1/2}$ and $\mathbf{W}^{CR}_{i+1/2} = \psi\mathbf{W}^C_{i+1/2} + (1-\psi)\mathbf{W}^R_{i+1/2}$. Parameter $\psi$ is given as

$$\psi = \begin{cases} 0 & \text{for } L_t < L_{DES} \quad \text{(RANS mode)}, \\ [0.5 - 0.5\cos(2\pi(1 - L_{DES}/L_t))]^2 & \text{for } L_t \geq L_{DES} \quad \text{(LES mode)}, \end{cases} \tag{28}$$

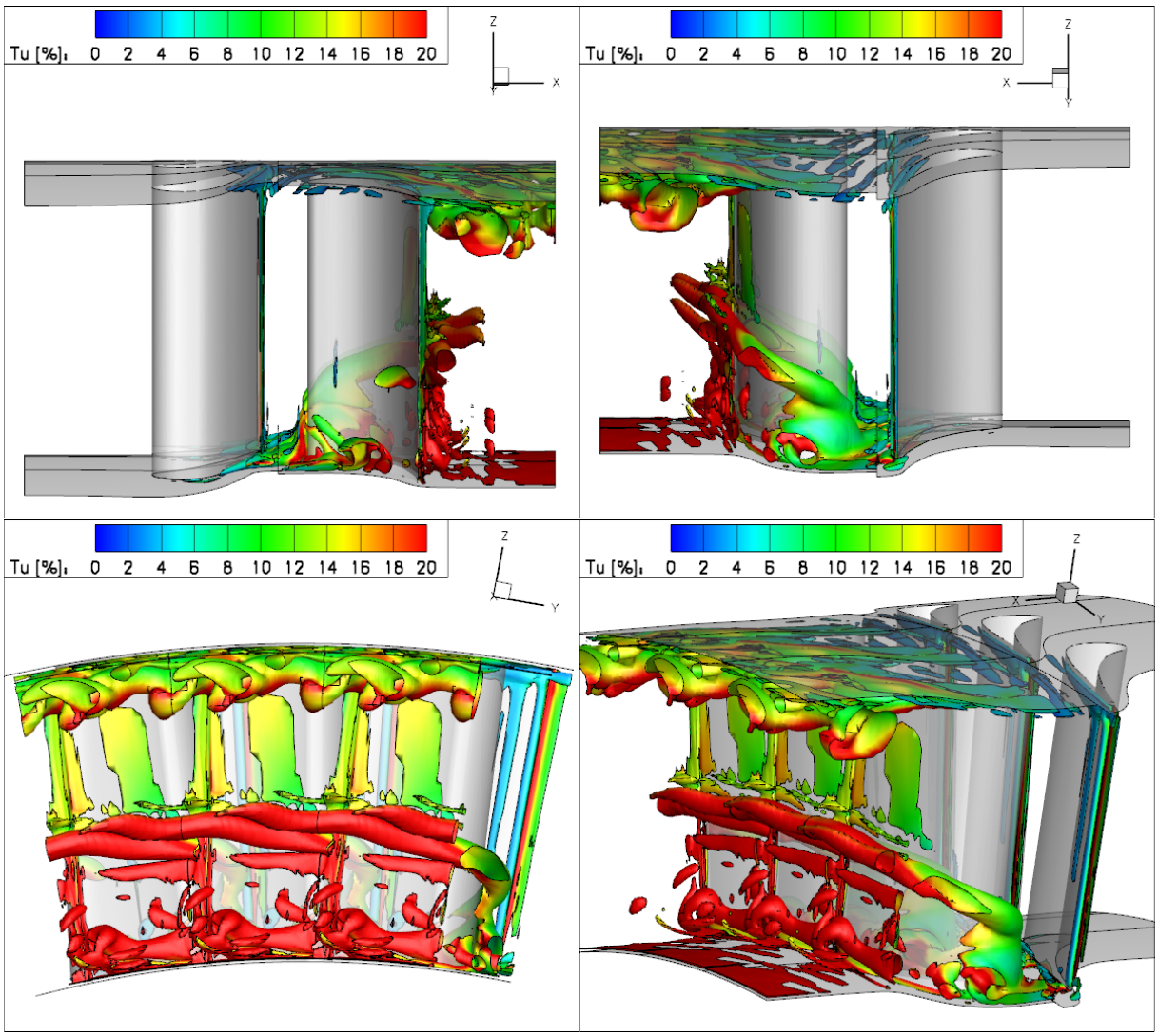



Figure 4: LEVM – instantaneous isosurfaces of $(|\Omega|^2 - |S|^2)$ colored by local turbulence intensity.

where $L_t = \sqrt{k}/\beta^*\omega$ is the turbulent length scale and $L_{DES} = C_{DES}\Delta_{DES}$ is the sub-grid length scale.

## 7 COMPUTATIONAL RESULTS

In figures 4 to 6 there are shown an instantaneous isosurfaces of $(|\Omega|^2 - |S|^2)$ colored by local turbulence intensity for LEVM, EARSM anf XLES models. The local turbulence intensity is related to the modeled part of the turbulent or sub-grid scale kinetic energy. In figures 4 to 6 there we can recognize two strong sources of secondary vortices. These are the radial gaps under the hub-end of the stator blade and above the tip-end of the rotor blade. We focus now on the radial gap under the stator blade. The effect of flow through the radial gap under the stator blade is such that stream flowing from the radial gap generates (in interaction with main stream in blade channel) massive vortex above the stream and beside the vertical surface of the suction side as illustrates a scheme in left part of figure 3. This upper vortex is in given regimes and geometrical configurations intensive enough to deform the stream flowing from the radial gap which leads to separation of the stream from the wall and to generation of the

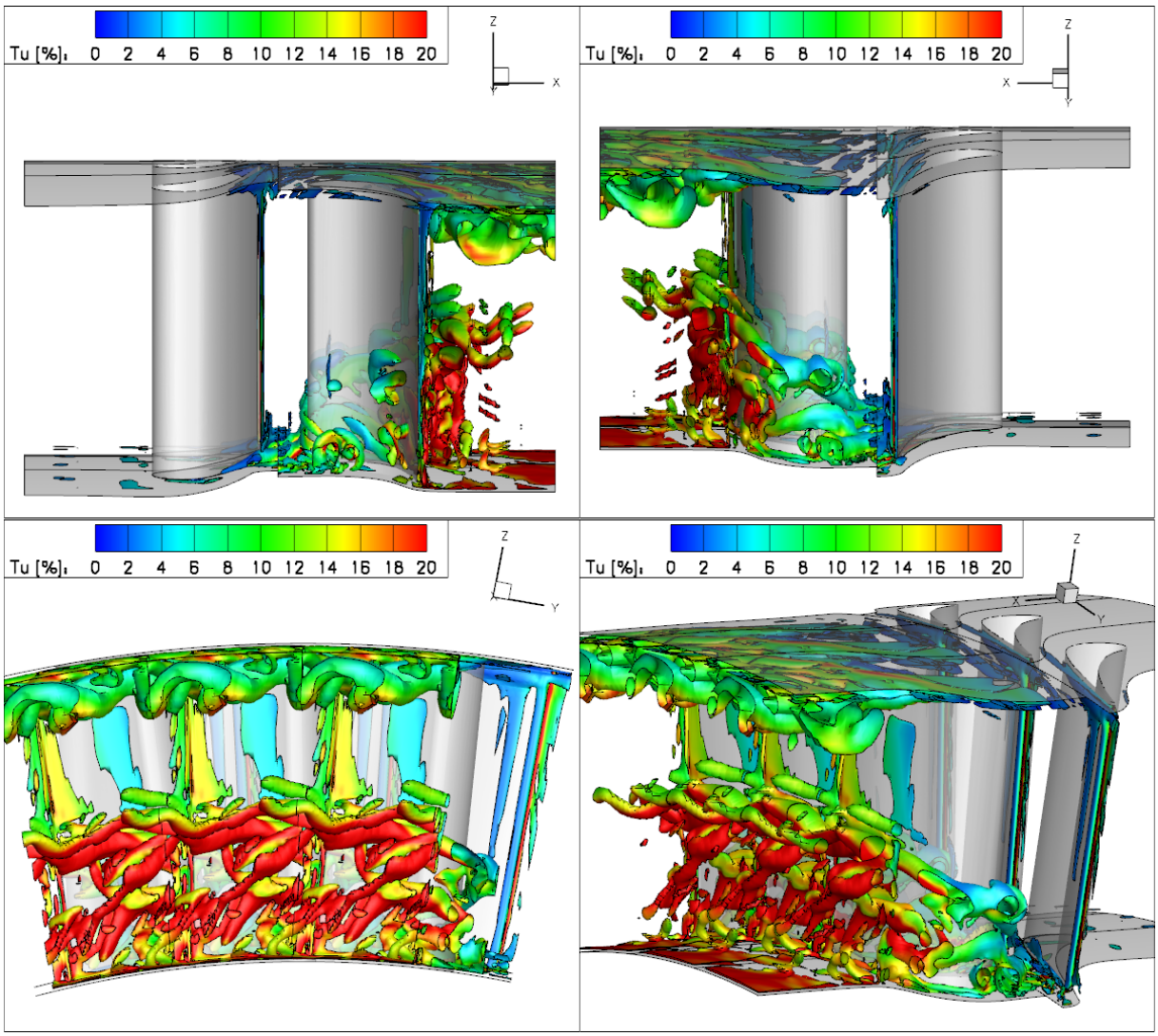



Figure 5: EARSM – instantaneous isosurfaces of $(|\Omega|^2 - |S|^2)$ colored by local turbulence intensity.

other vortex under the stream. Both vortices then interact with the rotor blade so that they deforms in upward direction and they are carried on the suction side almost up to half of the span. This causes increase of the dissipation rate on the lower half of span on the suction side, as illustrates distribution of the dimensionless entropy index $s = p/\rho^\kappa$ (computed from the normalized pressure and density) on the rotor blade surface that is shown in right part of figure 3. The LEVM model predicts the interaction of the rotor blade with the secondary vortices so that the secondary vortices remains more or less compact and undeformed. We can identify them in figure 4 behind the rotor blades at the middle of span. The EARSM model predicts deformation of the secondary vortices which interact with the rotor blades and their breakdown into smaller structures as is shown in figure5. In case of the XLES model we can observe a decay of the secondary vortices during the interaction with the rotor blade as is shown in figure 6. We can also see that in case of the XLES model the local turbulence intensity, that relates to the modeled kinetic energy, is smaller compared to the LEVM and the EARSM models. This is because a larger part of the kinetic energy is directly resolved as a nonstationary fluctuations. We can also recognize in figures 4 to 6 vortex structures behind the rotor blade at upper twenty

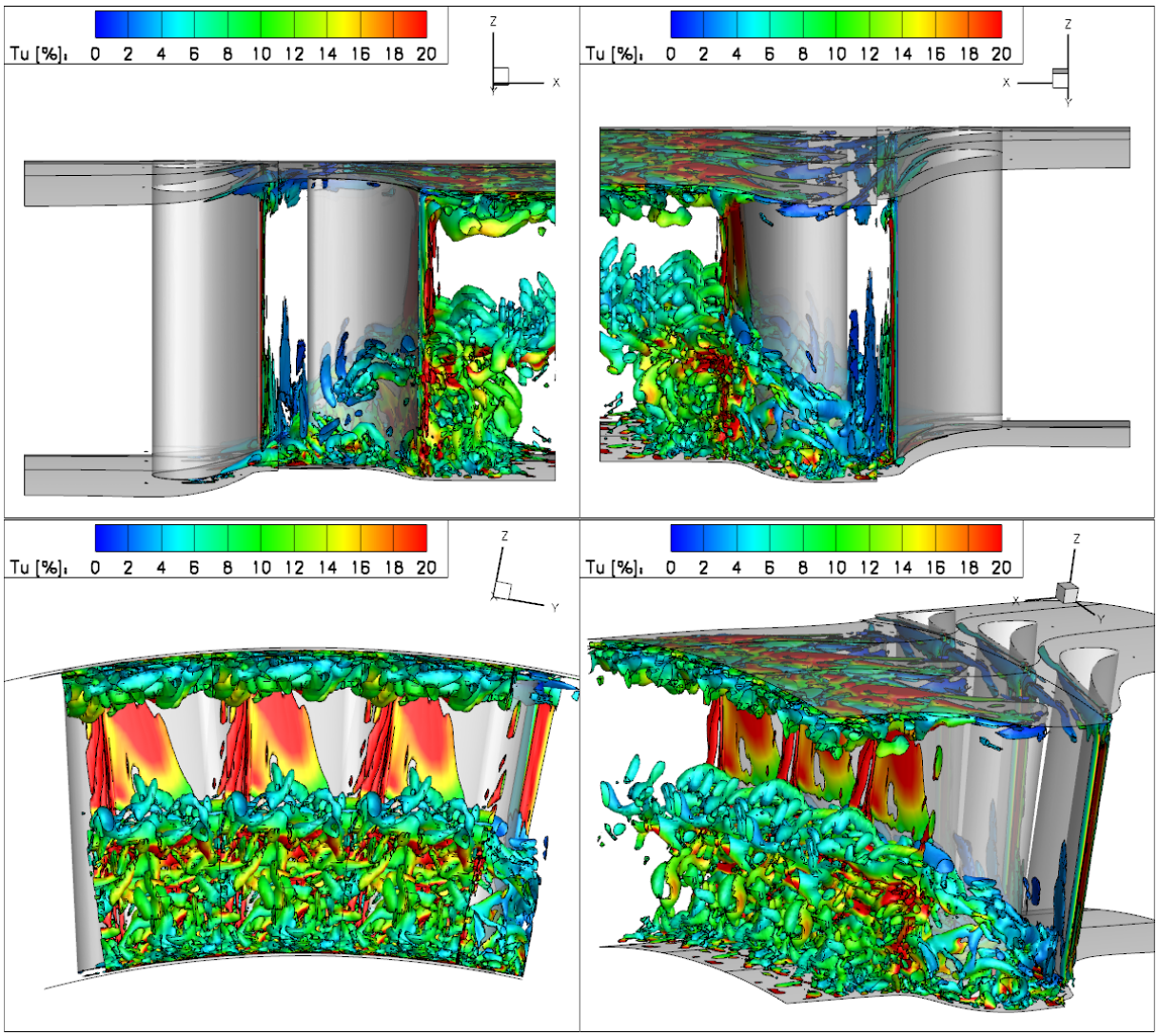



Figure 6: XLES – instantaneous isosurfaces of $(|\Omega|^2 - |S|^2)$ colored by local turbulence intensity.

percent of span, that arise behind the radial gap above the rotor blade.

In figure 7 there is shown an instantaneous and time averaged (in rotating frame) distribution of the total-total efficiency $\eta_{TT}$ in an angular section 11 mm behind the trailing edge of the rotor blade. The total-total efficiency is defined as $\eta_{TT} = (T_{T0} - T_T)/(T_{T0} - T_{Tis})$ where $T_{T0}$ is the inlet total temperature, $T_T$ is local total temperature and $T_{Tis}$ is the total isentropic temperature. On time averaged distribution there is a local maximum of the efficiency approximately at one quarter of the span in case of the LEVM and the EARSM models. This locale maximum is supressed in case of the XLES model due to decay of the secondary vortices into smaller and smaller structures which fill almost uniformly lower half of the span. This behaviour much better corresponds with the experimental data (from [1, 2]) than results of the LEVM and the EARSM models as shown in figure 8, where is shown a span-wise distribution of the circumferential averaged total-total efficiency. From figure 8 it is evident that all used models overpredict the efficiency about four percent between sixty and eighty percent of the span. It is an area which is not affected by secondary vortices. The overprediction of the efficiency may be due to insufficient resolving of the wakes.

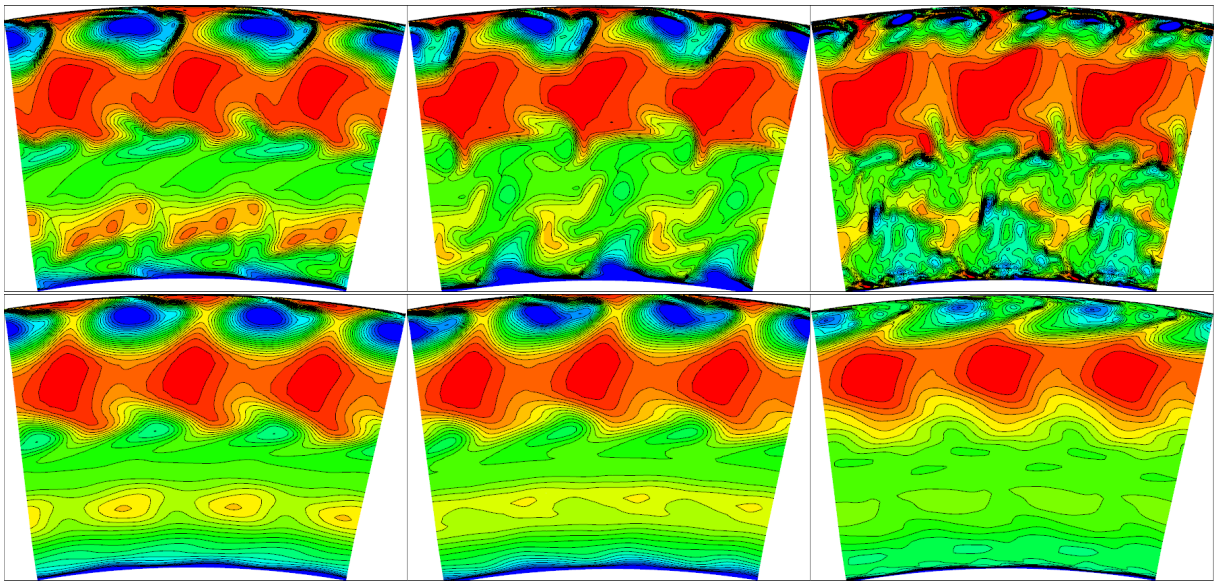


Figure 7: Distribution of the total-total efficiency in angular section 11 mm behind the trailing edge of the rotor blades. LEVM (left column); EARSM (middle column); XLES (right column); instantaneous distribution (top row); time averaged distribution – in rotating frame (bottom row).

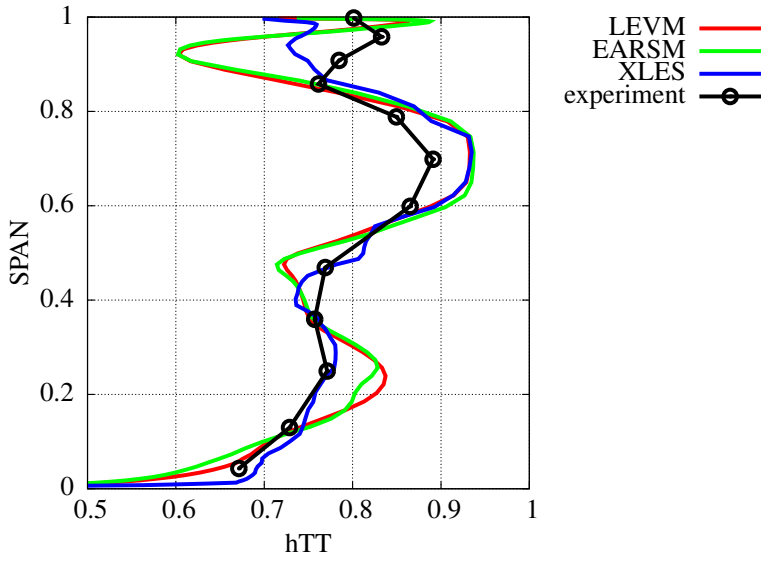



Figure 8: Span-wise distribution of the circumferentially averaged total-total efficiency in section 11 mm behind the trailing edge of the rotor blade – comparison with the experimental data.

## 8 CONCLUSIONS

Three kinds of turbulence models were applied for computation of flow in axial turbine stage with prismatic blades which are not equipped with the shroud. The LEVM model which is based on linear relation for the Reynolds stress tensor, the nonlinear EARSM model which uses a quadratic relation for the Reynolds stress tensor and the XLES model which combines RANS approach near the wall and LES approach for simulation of large vortices. Results shown that the LEVM model does not predict deformation and breakdown of the secondary vortices generated behind the radial gap under the stator blade which interact with the rotor blade. Deformation of this vortices and their breakdown into smaller structures predicts the nonlinear EARSM model. Better agreement with the experimental data is achieved using the XLES model which describes decay of the secondary vortices.

## ACKNOWLEDGEMENT

This work was supported by the Long-term Framework Advancement Plan provided by the Ministry of Industry and Trade of the Czech Republic.

## REFERENCES

[1] P. Straka, M. Němec, T. Jelínek, Investigation of flow in axial turbine stage without shroud-seal. *EPJ Web of Conferences,* , **92**, 02088, 2015.

[2] P. Straka, M. Němec, Influence of the radial gap under the stator blade on flow around the hub-end of the rotor blade. *Applied Mechanics and Materials*, **821**, 120–128, 2016.

[3] J. C. Kok, Resolving the dependence of freestream values for the k turbulence model. *AIAA Journal*, **38**, 7, 1292–1295, 2000.

[4] P. Straka, Numerical simulation of high-swirl flow in axial turbine stage. P. Dančová, M. Veselý eds. *Proc. Conf. Experimental Fluid Mechanics 2015*, 904–910, Prague, Czech Republic, 2015.

[5] C. L. Rumsey, T. B. Gatski, Recent turbulence model advances applied to multielement airfoil computations. *Journal of Aircraft*, **38**, 5, 904–910, 2001.

[6] P. Straka, Simulation of a 3D unsteady flow in an axial turbine stage. *EPJ web of Conferences*, **25**, 01090, 2012.

[7] P. Straka, Calculation of 3D unsteady flow in axial stage of low power experimental turbine. *Technical Report VZLU R-5904*, Prague, 2013, (in Czech).

[8] L. Davidson, S. H. Peng, Hybrid LES-RANS modelling: a one-equation SGS model combined with a $k-\omega$ model for predicting recirculating ows. *Int. Jour. Numer. Meth. Fluids*, **43**, 1003-1018, 2003.

[9] J. C. Kok, H. S. Dol, B. Oskam, H. van der Ven, Extra-large eddy simulation of massively separated flows. *AIAA paper*, 2004-264, 2004.

[10] S. Kubacki, J. Rokicki, E. Dick, J. Degroote, J. Vierendeels, Hybrid RANS/LES of plane jets impinging on a flat plate at small nozzle-plate distances. *Arch. Mech.*, **65**, 2, 143–166, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# PATHWAYS TO IMPROVED AERODYNAMIC DESIGN

**James H. Page[1], Paul Hield[2], Callum Mantell[1] and Paul G. Tucker[1]**

[1]University of Cambridge
Department of Engineering
Trumpington Street
Cambridge, CB2 1PZ
United Kingdom
e-mail: {jhp41,crm61,pgt23}@cam.ac.uk

[2]Fan Systems Engineering
Rolls-Royce plc
Filton, Bristol, BS34 7QE
United Kingdom
e-mail: Paul.Hield@Rolls-Royce.com

**Keywords:** Graphics Processing Unit, Computational Fluid Dynamics, Turbomachinery, Distortion, Design.

**Abstract.** *Large-scale unsteady graphical processor unit (GPU) based calculations are used to study the flow in realistic transonic fan geometries and shown to produce useful insights into the complex unsteady flow physics in systems of practical engineering relevance. Inlet flow distortion produces complex flow patterns which can significantly reduce stall margin. In order to design efficient, well targeted fan stabilizing technologies, it is essential to gain understanding into which specific features of distorted flow patterns are important in causing instability.*

*In this paper, the use of large-scale, high-fidelity 3D unsteady Reynolds-Averaged Navier Stokes (URANS) calculations with sliding planes is shown to allow the isolation of a wide range of inlet distortion features and a high throughput of calculations is achieved. This facilitates a comprehensive investigation into stall margin loss and the observation of complex instability processes, which is in turn used to support the development of specialized design solutions. The costs of simulations and turnaround times are given. The potential for mixed fidelity simulations, in particular mixing one-dimensional low order models and large eddy type simulations to generate inlet conditions, is discussed.*

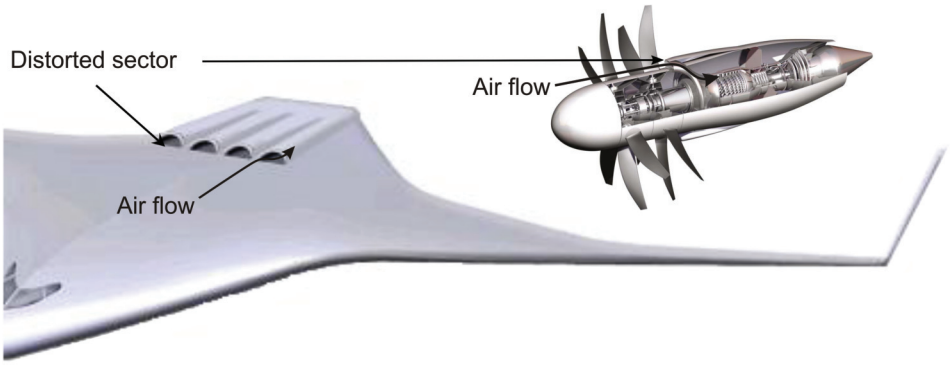



Figure 1: Aerospace airframe and propulsion technology under development.

## 1 INTRODUCTION

The world aircraft fleet is predicted to double by 2030.[1] At the same time, mankind faces the immense challenge of reducing $CO_2$ and other emissions, along with other pressing tasks such as reducing noise.[1] The Advisory Council for Aeronautics Research in Europe (ACARE) has set targets to reduce $CO_2$ and noise emissions $50\%$ by 2020.[1] These challenges have proved to be a game changer in encouraging the development of promising alternative technologies and new concepts in aircraft design. Examples, among others, are the open rotor gas turbine engine and the boundary layer ingestion (BLI) Silent Aircraft, shown in figure 1. The emergence of such novel architecture raises a considerable deficit in aerodynamic understanding and extensive design spaces to explore. While experimental rigs lack flexibility and scope before such a challenge, high fidelity computational fluid dynamics (CFD) has the agility to aggressively explore and narrow down the vast design spaces.

CFD is a rapidly developing tool used to support aerodynamic component design in the aerospace industry. It is, however, not an exact science and its successful use requires prudent consideration of numerous sources of error, listed by Denton [2] as numerical errors due to finite difference approximations, modelling errors for example turbulence modelling, unknown boundary conditions, unknown geometry such as tip clearances or leading edge shapes, and assumptions of steady flow. Improvements in the capability of high performance computing and advances in CFD codes have made practical the deployment of unsteady 3D CFD on an industrial scale.

3D unsteady CFD comprises, in order of increasing fidelity, unsteady Reynolds-Averaged Navier-Stokes (URANS) calculation, detached eddy simulation (DES), large eddy simulation (LES) and Direct Numerical Simulation (DNS). While computational expense confines the last three in the list to academic study and very small scale industrial use, URANS has now become practical for large-scale studies in industrial aerodynamic design processes.

An important topic in jet engine design for futuristic aircraft architecture is inlet distortion due to curved intake ducts, which has the potential to reduce fan stall margin and negatively affect fan stability. There is a press need to gain in-depth understanding of distorted intake flow in order to design efficient, well targeted fan stabilizing technology.

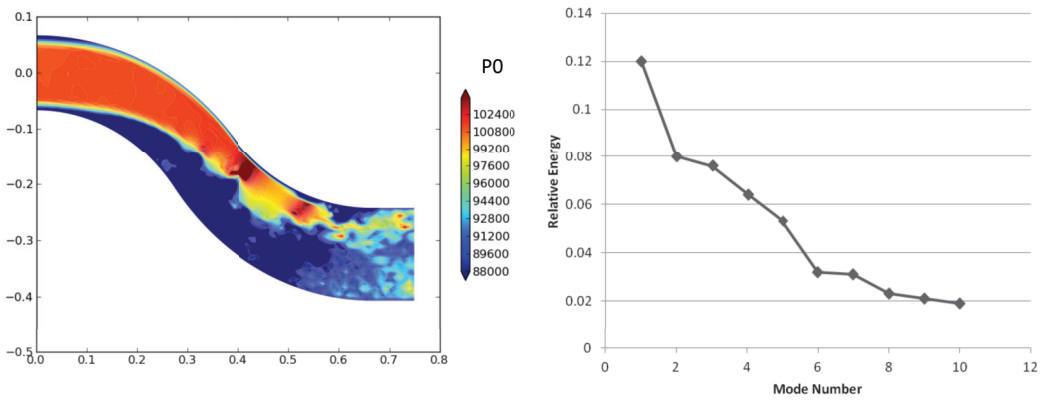



Figure 2: Azimuthal cut showing a typical curved duct LES flow.

### 1.1 Inlet Distortion

Research into the effects of an inlet total pressure circumferential square wave on stall margin was carried out with and without VIGVs by Shaw *et al.* [3]. They found that the distortion significantly reduced stall margin and that some of the reduction could be recovered by using VIGVs to reduce the circumferential variation in inlet absolute swirl. Other literature applicable to fan inlet distortion also looked at square waves. Mistry *et al.* [4] carried out experiments on a contra-rotating rotor rig and analysed effects on aerodynamics and performance. Gunn *et al.* [5] conducted experiments on a low speed fan stage and studied the effect on loss mechanisms. Gunn *et al.* [6] later studied the effect of BLI inlet profile on fan stage aerodynamics and loss using CFD and experiments. More recently, Perovic *et al.* [7] investigated stall inception in the same BLI environment, gaining insight in to the growth and decay of stall cells. So far, research has focussed on more or less idealized distortions while leaving out much of the detail in real flows.

What is now well known is that fan flow and intake flow are strongly coupled. Future design of curved intakes and fans will become increasingly integrated to ensure good performance in non-uniform unsteady intake flow. A large eddy simulation (LES) of flow in an S shaped curved duct based on that used by Garnier reveals a highly complex, unsteady distortion pattern emerging downstream of the curves.[8] The distortion is characterised by a static pressure and velocity distortion, 3D secondary flow vortices, total pressure losses in the boundary layer and main flow —both due to separation— and unsteady vortex shedding. These are visible in the azimuthal snapshot of total pressure ($P0$) in figure 2.[1]

A key question with respect to fan rotor stabilization is which particular distortion features are most damaging and what are the stalling mechanisms that they promote? To answer, it may be helpful to deconstruct inlet distortion and test each feature separately. A popular method for analyzing complex flows is proper orthogonal decomposition (POD) which deconstructs the cross-sectional unsteady velocity field into modes and quantify their relative energies.[8] Where one or a few modes dominate, this can be a powerful method of identifying dominant axially aligned vortical flow structures. However, the POD relative energy plot in figure 2 shows a broad spectrum with the most energetic 10 modes accounting for only half of the total energy. Therefore as an alternative, visual inspection of the flow suggests the following decomposition:

[1]Results kindly provided by R. Watson.

1. inlet vortices of various sizes and orientations,

2. the static pressure and associated velocity distortion due to curved streamlines,

3. a total pressure deficit due to separation in the curved intake in addition to the above static pressure and velocity distortion,

4. a time-averaged distortion profile from the Garnier duct LES, which includes more realistic total pressure 2D spatial and temporal profile due to separation in addition to the curved streamline static pressure and velocity distortion,

5. the unsteady Garnier duct LES profile which adds a complex unsteady dimension to the time-averaged version.

Further research on complex distorted flows and methods of analysis is underway.

Building on item 5, a way to couple the URANS fan flow to the LES duct flow is to input the duct outflow directly into the fan inlet. Results for this will be shown in section 3.4. However, a far better way would be to replicate the particular fan rotor in the duct LES downstream conditions by means of an inexpensive body force model such as that proposed by Cao *et al.* [9]. This would allow the fan to impose its own influential static pressure field on the duct flow which would generate more realistic inlet conditions for the fan URANS calculation.

A full account of the analyses of each part is beyond the scope of this paper. Therefore, calculations for items 2, 3 and 4 will be discussed in section 3.

### 1.2 The Use of URANS in Aerodynamic Design and Fan Stall

An extensive validation of the URANS solver with consistent settings and the Spalart-Allmaras turbulence model was carried out under the same project umbrella on a Rolls-Royce single stage fan rig by Shaw *et al.* [3]. The validation fan geometry was similar in pressure ratio and aspect ratio to the geometry used in this paper. The CFD was found to show good agreement with experimental results in radial and circumferential total temperature and total pressure profiles, and changes in stall margin. This is regarded to provide sufficient validation for the further project work in this paper.

In an internally reported work, further studies were carried out using two different CFD solvers each with the Spalart-Allmaras and mixing length turbulence model. Clean stall was predicted to within $9\%$ of the test nozzle area in $95\%$ of about $40$ different fan geometries. Most importantly, however, the results for change in stall margin under the influence of inlet distortion were within $3\%$ of the experimental results.[3]

The key point is that, while predicting single values on a large scale is still out of reach, URANS is sufficiently powerful method to investigate behaviour patterns of the level of complexity of stall. This will be shown to be useful in supporting design choices and providing significant understanding of the design space prior to experimental investigation.

## 2 METHODOLOGY

### 2.1 Fan Geometry

A series of CFD experiments have been carried out on the rotor of an advanced two-stage transonic fan geometry, which was inverse designed by the methodology and code described in Page *et al.* [10]. The geometry, shown in figure 3, was designed to specified radial profiles for stage pressure ratio and exit flow angle, and a specified mass flow at a nominal $100\%$ speed.

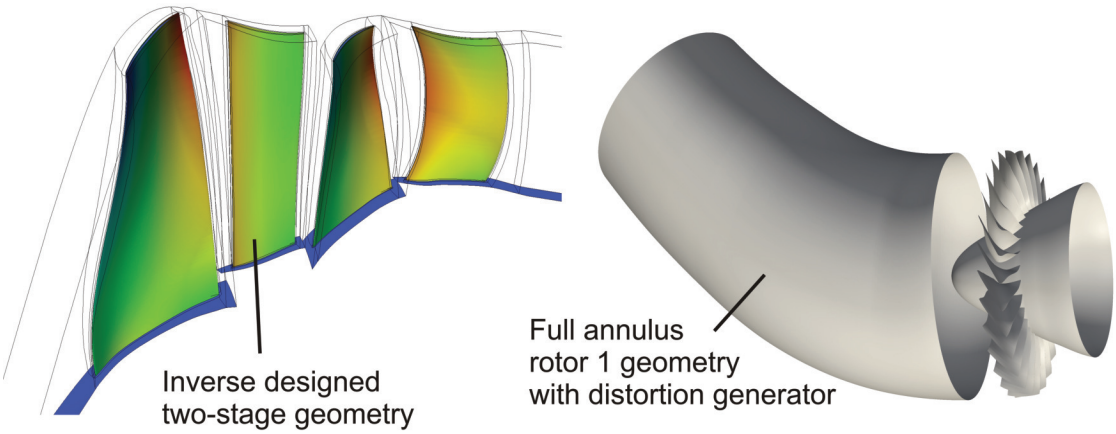



Figure 3: The two-stage fan geometry designed by a dual point inverse design algorithm and the 3D URANS rotor geometry with distortion generator.

Using the dual point algorithm developed by the authors, the fan was designed to *simultaneously* achieve these performance targets at $100\%$ speed and desired rotor shock positioning, blade loading distribution and, therefore, peak efficiency at $90\%$ speed. The design to mass flow and radial stage pressure ratio and exit flow angle profiles for this geometry were chosen to be representative of typical Rolls-Royce fans. The design is described in greater detail in [10].

For the CFD calculations in this paper, the first rotor blade was converted to a multipassage geometry. Figure 3 also visualizes the full annulus rotor geometry.

### 2.2 Inlet Distortion Generator

The method used to generate distortion items 2 and 3 is a curved intake duct attached upstream of the full annulus rotor as shown in figure 3. The duct is area ruled with the cross-section transitioning from elliptical to circular and features a bullet formed by a second order Bezier surface of revolution. The walls can be defined as either inviscid slip walls or viscous no-slip. Part way along the duct an impermeable fence can be positioned in the flow if so desired.

The purpose of the inviscid slip walls was to facilitate the generation of the static pressure and velocity distortion associated with curved streamlines without the addition of total pressure distortion, especially a thicker boundary layer on the inside of the curve. With the no-slip walls a circumferentially non-uniform boundary layer was permitted.

### 2.3 Stall Point Calculation

The CFD calculations in this paper were carried out unsteady using Turbostream, a URANS CFD solver.[11] The code uses the finite volume method to solve for the flow on structured O-H multiblock meshes with hexahedral cells. The temporal discretization used in unsteady mode is based on the second-order Adams-Bashforth scheme. For this paper, all the calculations were carried out unsteady with a CFL number of 0.3 and using the Spalart-Allmaras turbulence model. Wall functions were used at solid boundaries. Sliding planes were used to interface rotating rotor mesh blocks with static mesh blocks for static blade rows or intake ducts.

In order to calculate the stall point, a choked nozzle was used to control the exit flow function and thus the position on the characteristic. Starting from a nozzle area away from the stall point, unsteady CFD was calculated over six revolutions, representing about three through-flows, with

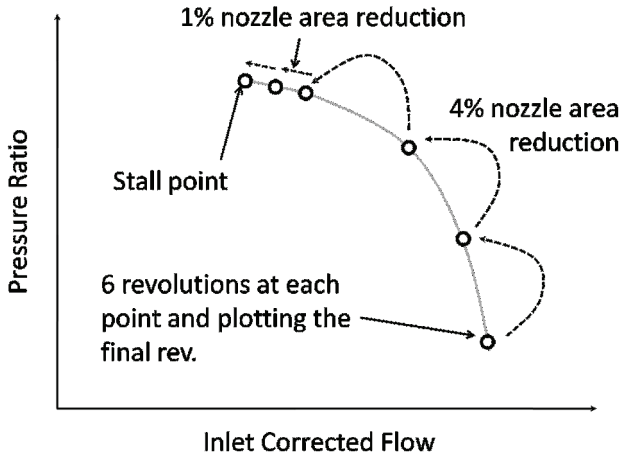



Figure 4: METHOD USED TO PLOT A CHARACTERISTIC CURVE AND CALCULATE THE STALL POINT.

flow results being saved once every blade passing during the final rev. The nozzle area was then reduced by $4\%$ and the six rev procedure repeated. The stall point was calculated by successive $4\%$ nozzle area reductions in this manner, followed by a targeted $1\%$ nozzle area reduction high resolution pass. The procedure is illustrated in figure 4

The pressure ratio versus corrected inlet mass flow for each blade passing provided a cloud of points for each nozzle area showing the point on the characteristic curve along with the level of unsteadiness.

### 2.4 High Performance Computing Architecture

Turbostream is an updated version of TBLOCK, recoded and optimized for Graphics Processing Units (GPUs), delivering a speedup of typically up to 20 times.[11] The hardware used to carry out the calculations was a cluster comprising 48 Nvidia K10 GPUs. Each K10 is made up of two GK104 GPUs, each of which has 1536 cores. The investigation required the calculation of six nozzle areas on a one-third annulus geometry and 24 nozzle areas on a full annulus geometry. In total there were the equivalent of 156 full annulus revolutions.

## 3 THE EFFECT OF INLET DISTORTION FEATURES ON ROTOR STALL MARGIN

### 3.1 Clean Inlet Stalling Behaviour

The first CFD experiment calculated the stall point of the ‘clean inlet’ case. As the clean inlet was axisymmetric, the unsteady calculation was run on one-third of the annulus. The stall inception was observed to be due to tip leakage flow disturbing the flow at the leading edge of the blade[12] near the tip, leading to separation and the formation of a rotating stall vortex[13] which develops into a rotating stall cell. Figure 5 uses the Q-criterion vortical structure visualization to show separation close to the tip and the rotating stall vortex which is formed moments afterwards. Figure 6 shows the velocity field superimposed on axial velocity contours, highlighting the strong negative axial component of velocity (blue) interfering with the inlet flow in the tip region near the leading edges.

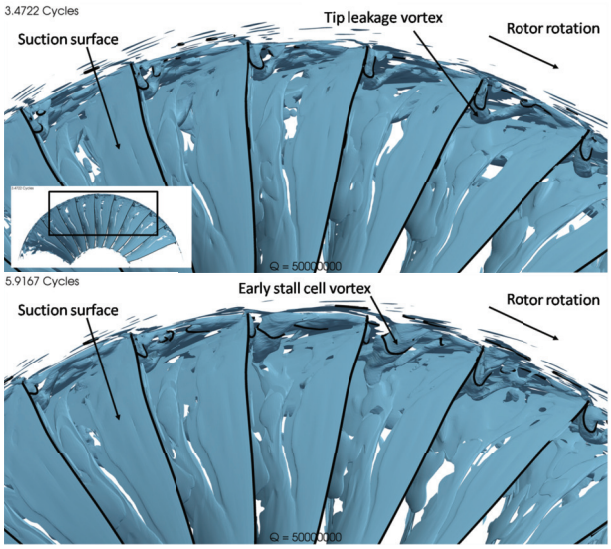



Figure 5: Vortical structure visualization using Q-criterion at stall inception.

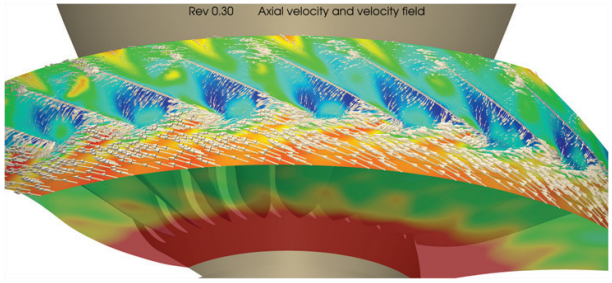



Figure 6: Tip region velocity field and axial velocity contours.

### 3.2 Curved Streamline Static Pressure and Velocity Distortion

The second CFD experiment compares the stall point of the inviscid slip-wall distortion generator case with the clean inlet case. The azimuthal plot in figure 7 shows the distortion generated by the curved duct. The variation between the higher static pressure and lower axial velocity on the outside of the bend and the lower static pressure and higher axial velocity on the inside of the bend is clearly visible. It is also noted that there is the absence of an inlet boundary layer or separation and the total pressure profile remains largely uniform. The curved duct isolates the static pressure-velocity distortion feature. The axial velocity seen by the rotor is shown again in the axial cut in figure 7. The upward trajectory of the incoming flow leads to a swirl distribution with regions of co-swirl and contra-swirl entering the rotor. This is evident in the tangential velocity plot in figure 7 which shows a negative swirl velocity, co-rotating with the rotor, on the left and a positive swirl velocity, contra-rotating against the rotor, on the right. The absence of a boundary layer is again conspicuous.

Figure 7 shows a remarkably well behaved flow at $95\%$ span via relative Mach at the last stable point on the characteristic. This case required six revolutions to be calculated at each of six nozzle areas. The wall clock time per revolution averaged 3 hours 11 minutes on 24 K10 GPUs, giving 19 hours 6 minutes per nozzle area and 114 hours 36 minutes (less than 5 days on half of the cluster) to calculate the stall point. This case showed no significant change in stall

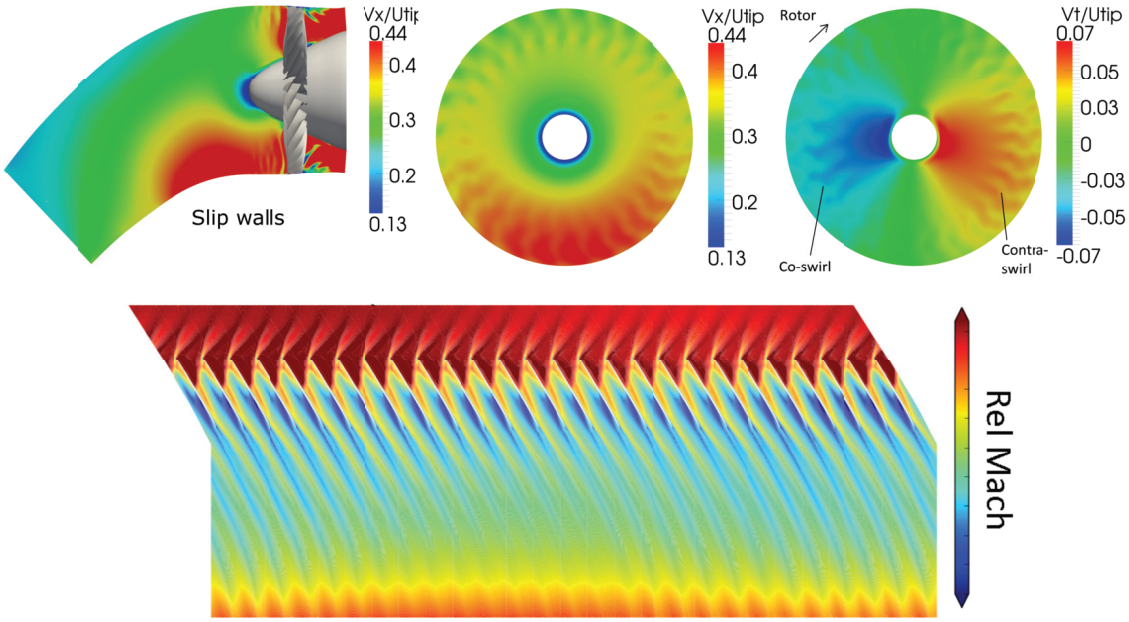



Figure 7: Overview of the inlet and fan flow with curved streamline distortion.

margin from the clean inlet, ruling out the importance of distortion due to curved streamlines alone at this magnitude.

### 3.3 Curved Streamline Static Pressure and Velocity Distortion with Boundary Layer

The third CFD study assessed the effect of adding a circumferentially varying boundary layer to the curved streamline distortion pattern. No-slip walls were added to the curved distortion generator to modify the inlet profile to include the thin boundary layer which is visible on the inside of the curve in figure 8. The circumferential profile of the new distortion and the swirl is shown in figure 8 to be very similar to the previous case but with the boundary layer visible. A key feature of the inlet profile is that is now contains total pressure loss near the tip, which is most pronounced downstream of the inside of the curve.

The radial plot of relative Mach at the final stable nozzle area shows how much more separated the flow is compared with the previous case. The turnaround time for this case was only fractionally longer than for the previous case. The stalling point for this distortion was at a nozzle area $3\%$ larger than with clean, giving a key result in identifying low momentum tip flow as a important distortion feature.

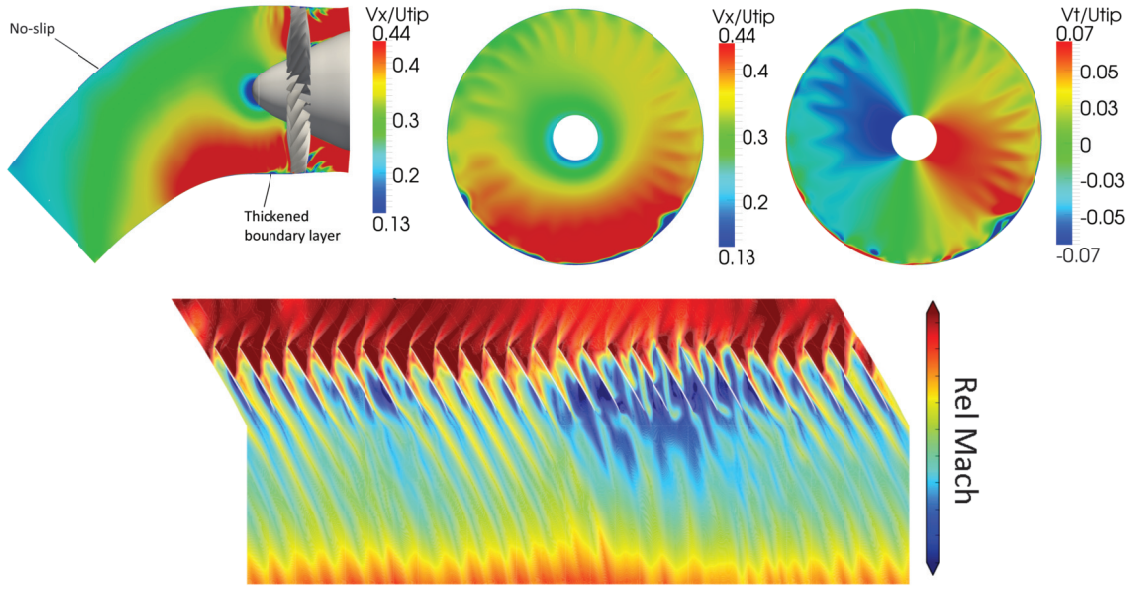



Figure 8: Overview of the inlet and fan flow with curved streamline distortion and boundary layer.

### 3.4 Garnier outlet flow as fan inlet condition

The final increment in the increasing complexity of inlet conditions shown in this paper is where the Garnier LES outflow is input directly into the rotor inlet. Six revolutions of unsteady Garnier duct flow data were input to straight duct section upstream of the fan rotor. Figure 9 shows snapshots of the inlet and fan flow, clearly demonstrating its chaotic nature.

The turnaround wall clock time for calculations increased to an average of 5 hours 35 minutes per revolution on 24 K10 GPUs, giving 33 hours 30 minutes per nozzle area or 134 hours to find the stall point with four nozzle areas. The increase in time per revolution was due to the highly unsteady flow field requiring a greater average number of inner time-steps to converge. In this case each point in the characteristic displayed increased unsteadiness but there was no significant reduction in stall margin compared with the curved duct with boundary layer. The Garnier duct flow is essentially a combination of distortion features including low momentum tip region flow. The lack of a further drop in stall margin in the presence of this inlet flow provided an early suggestion that the low momentum flow in the tip region encountered in the previous case was could still by far the most important flow feature.

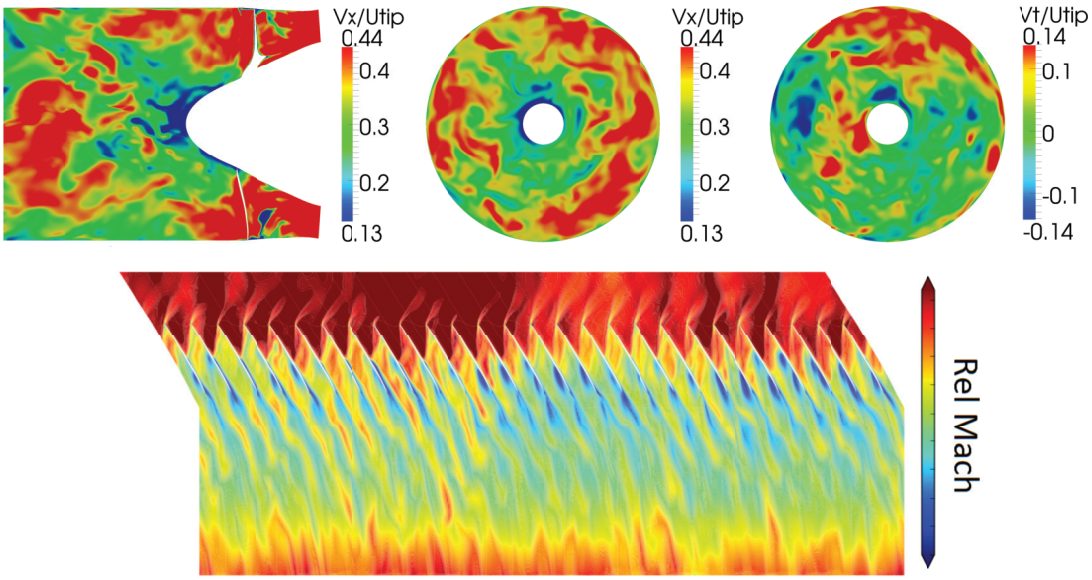



Figure 9: Overview of the inlet and fan flow with a straight inlet the Garnier results as inlet conditions.

## 4 Conclusions and Discussion

This paper discusses developments in the use of CFD in gas turbine engine design and demonstrates the practicality and value of large-scale 3D URANS in the design process via a non-exhaustive outline of an investigation into fan stall with inlet distortion. Results from LES of the Garnier duct were used to support the design of idealized distorted inlet conditions and to provide a complex unsteady inlet profile. This highlighted low momentum tip region flow as a key distortion pattern.

Resulting from this work, a recirculation passage geometry similar to that described by Guinet *et al.* [14, 15] has been designed and simulated. Design features of the passage are currently under URANS investigation by the authors and are providing early evidence of a significant increase in stall margin. Figure 10 shows a the plan view of the recirculation passages on a 3D multipassage geometry. The passage was preliminarily designed using 1D diffuser theory [16] to diffuse the flow and reduce losses in the passage, before reaccelerating to maintain attachment around a low radius curve and achieve high velocity flow injection. The diffuser sections are marked in the side view of the passage in figure 11. The diffuser sections 3D URANS in a one-third annulus geometry was then used to calculate stall points for variations in passage width, injection yaw angle and circumferential positional distribution.

One improvement in the current design process should be the calculation of more realistic inlet conditions for the fan. A duct flow LES using a body force representation of the fan rotor, as mentioned in section 1.1, would create far more realistic outlet conditions than a static pressure condition by coupling duct and rotor flows. Figure 12 compares schematic diagrams of the current and future setups.

A second improvement would be to increase the fidelity of the CFD local to stalling mechanisms, and in the case with the recirculation passage, local to the region affected by flow injection and where geometry optimization is to be carried out. For the rotor case in this paper, a zonal LES calculation, where LES is used in the fluid domain at high span and URANS ev-

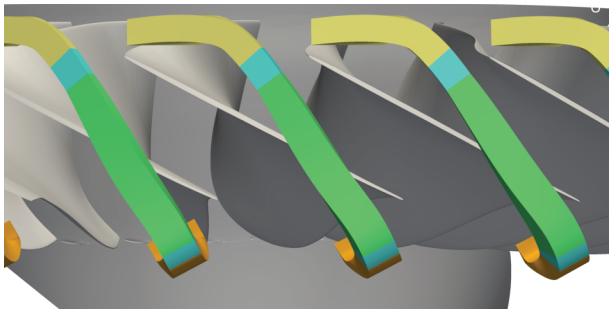


Figure 10: Plan view of the recirculation passage in the casing of a one-third annulus rotor geometry.

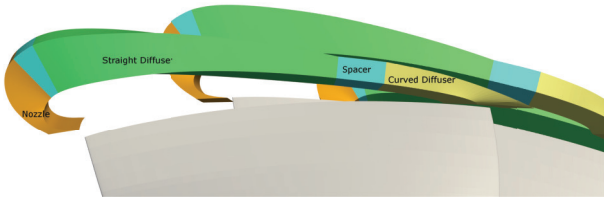



Figure 11: Side view of the recirculation passage showing the diffuser sections.

erywhere else, would capture the complex tip leakage stall inception mechanism at even higher fidelity. This would lead to a more detailed understanding of the stall mechanisms for different distortions. High span zonal LES would also give a higher fidelity calculation of the turbulent mixing of the injected jet from a recirculation passage. This would move practical unsteady CFD closer to the predictive capability of full LES and may facilitate detailed design optimization of the passage itself.

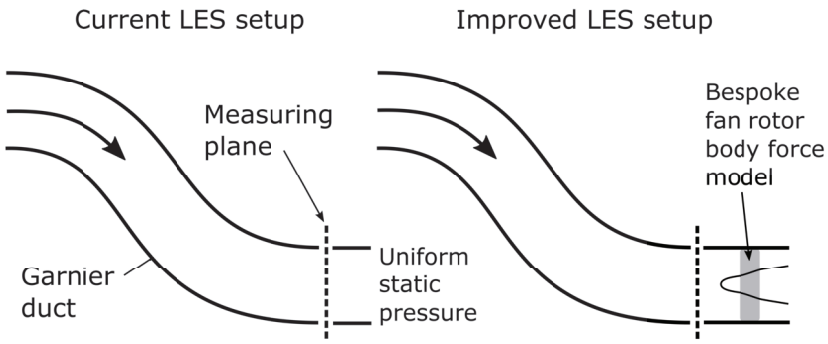



Figure 12: Schematic diagrams of the current and future setups of duct LES.

## ACKNOWLEDGEMENT

The authors would like to thank Rolls-Royce plc for permission to publish and to use in-house software. Rob Watson is gratefully acknowledged for permission to use LES results, and also Tobias Brandvik, Graham Pullen and John Denton for permission to use Turbostream.

## REFERENCES

[1] P. G. Tucker. Unsteady Computational Fluid Dynamics in Aeronautics. *Fluid Mechanics and its Applications*, 2014.

[2] J. D. Denton. Some Limitations of Turbomachinery CFD. In *Proc. ASME Turbo Expo 2010*, June 2010. ASME Paper No. GT2010-22540.

[3] M. J. Shaw, P. Hield, and P. G. Tucker. The Effect of Inlet Guide Vanes on Inlet Flow Distortion Transfer and Transonic Fan Stability. *ASME J. Turbomach.*, **136**, Feb 2014.

[4] C. S. Mistry and A. M. Pradeep. Investigations on the Effect of Inflow Distortion on the Performance of a High Aspect Ratio, Low Speed Contra Rotating Fan Stage. In *Proc. ASME Turbo Expo 2013*, June 2013. ASME Paper No. GT2013-94311.

[5] E. J. Gunn, S. E. Tooze, C. A. Hall, and Y. Colin. An Experimental Study of Loss Sources in a Fan Operating With Continuous Inlet Stagnation Pressure Distortion. *ASME J. Turbomach.*, **135**, Sep 2013.

[6] E. J. Gunn and C. A. Hall. Aerodynamics of Boundary Layer Ingesting Fans. In *Proc. ASME Turbo Expo 2014*, June 2014. ASME Paper No. GT2014-26142.

[7] D. Perovic, C. A. Hall, and E. J. Gunn. Stall Inception In A Boundary Layer Ingesting Fan. In *Proc. ASME Turbo Expo 2015*, June 2015. ASME Paper No. GT2015-43025.

[8] R. Watson, P. G. Tucker, and K. Menzies. Unsteady Simulations of the Wellborn Diffusing S-Duct. In *54th AIAA Aerospace Sciences Meeting, AIAA Science and Technology Forum and Exposition*, Jan 2016.

[9] T. Cao, Hield P., and P. G. Tucker. Hierarchical Immersed Boundary Method with Smeared Geometry. In *54th AIAA Aerospace Sciences Meeting, AIAA Science and Technology Forum and Exposition*, Jan 2016.

[10] J. H. Page, P. Hield, and P. G. Tucker. Inverse Design of 3D Multistage Transonic Fans at Dual Operating Points. *ASME J. Turbomach.*, **136**, Apr 2014.

[11] T. Brandvik and G. Pullen. An Accelerated 3D Navier-Stokes Solver for Flows in Turbomachines. *ASME J. Turbomach.*, **133**, 2011.

[12] J. H. Page, P. Hield, and P. G. Tucker. Effect of Vortex Ingestion on Transonic Fan Stability. In *54th AIAA Aerospace Sciences Meeting, AIAA Science and Technology Forum and Exposition*, Jan 2016.

[13] G. Pullan, A.M. Young, I.J. Day, E.M. Greitzer, and Z.S. Spakovsky. Origins and Structure of Spike-Type Rotating Stall. *Journal of Turbomachinery*, **137**, 2015.

[14] C. Guinet, J. A. Streit, and V. Gümmer. Tip Gap Variation on a Transonic Rotor in the Presence of Tip Blowing. In *Proc. ASME Turbo Expo 2014*, June 2014. ASME Paper No. GT2014-25042.

[15] C. Guinet, A. Inzenhofer, and V. Gümmer. Influencing Factors of Tip Blowing Interacting with Rotor Tip Flow. In *Proc. ASME Turbo Expo 2015*, June 2015. ASME Paper No. GT2015-42039.

[16] D.S. Miller. *Internal Flow Systems*. Miller Innovations, 2 edition, April 2009.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SIMULATIONS OF BLAST WAVE PROPAGATION IN OPEN AND CLOSED SPACE

**Natalya N. Fedorova**[1,2], **Svetlana A. Valger**[1], **Yulia V. Zakharova**[1,2]

[1]Novosibirsk State University of Architecture and Civil Engineering (Sibstrin),
113, Leningradskaya st., Novosibirsk, Russia,
e-mail: swetla-ya@mail.ru

[2] Khristianovich Institute of Theoretical and Applied Mechanics SD RAS
4/1 Institutskaya, Novosibirsk, Russia
{nfed,julsemen}@itam.nsc.ru

**Keywords: blast, explosion, shock waves, numerical modelling, cityscape**

***Abstract.*** *The paper presents results of 3D numerical simulations of shock wave spreading in cityscape area modeled by prismatic bodies placed on a flat plate and inside a closed room. ANSYS Autodyne software is used for the computations. Four different test cases are investigated numerically. The simulation results have been compared to experimental data. The ability of two numerical schemes is studied to correctly predict the pressure history in several gauges placed on walls of obstacles.*
.

## 1 INTRODUCTION

Nowadays, a lot of industrial accidents accompanied by explosions are happening throughout the world. Also, increase in the number of terrorist acts committed by means of explosions is observed. For improving safety of buildings and structures it is necessary to raise their resistance to explosive effects, as well as to be able to predict degree of potential damage upon explosive loads of various intensities [1 - 3]. One of the principal goals in designing the structure resistant to explosive effects is to determine the dynamic response of structures to the impact of the blast wave. To this end, the transient pressure loads on the walls of the civil engineering structures are to be determined.

The simulation of explosion is highly complicated, involving an explosion causing the shock wave propagation in air and then interaction with a structure. In recent years, many efforts have been devoted to the development of reliable methods and algorithms for a more realistic analysis of structures subjected to blast loading. Blast predictions are most commonly conducted in one of two ways: simplified analytical and engineering methods [4 - 6] or sophisticated CFD models [7 -8].

The engineering-level techniques permit one to estimate an explosive shock impact only for isolated buildings. The complexity of the building, the presence of nearby structures and the surrounding environment can't be taken into account. To overcome these limitations, various CFD instrumentations are used for blast impact on the civil engineering objects including in-house and commercial software. Advanced computer aid engineering (CAE) software techniques combined with the latest methods of discrete three-dimensional city modeling permits one to simulate and analyze the effects of explosions in urban areas with a precision which previously was not possible [9, 10].

In the previous papers by the authors [11, 12], the adaptation of LS-DYNA and AUTODYN software was performed for the 3D problem of the blast wave impact on the single prism installed on the flat plate. Comparison of the simulation results to the experimental data [13] has shown that LS-DYNA and AUTODYN allow one to perform numerical modelling of explosive impact on the environment with an acceptable accuracy. The simplified engineering formulas gave a good coincidence with the experimental data only for the gauges situated on the windward surface of the model.

In the paper, ANSYS AUTODYN software was tested on the problems of blast wave propagation in the open and closed spaces. The simulation results are compared to the experimental and numerical data [13 - 17]. The ability of two numerical schemes [18-20] is studied to predict the pressure history in gauges placed on walls of obstacles.

## 2 PROBLEM SETUP

Four different test cases are investigated numerically: three of them are for an open space with prismatic bodies placed on a flat plate, and the fourth is for the closed space with/without opening. In all the computations performed, the computational domains were the volume of air with the initial parameters corresponding to normal atmospheric conditions, namely, density $\rho = 1.225$ kg/m$^3$, temperature T = 298.15 K, static pressure P = 101325 Pa, the heat capacity $C_p = 1004$ J/kg·K. The geometries, problem setup and computational grids for the test cases are described below.

### 2.1 Test Case #1

The example has been taken from [13]. The prismatic body with 0.163x0.061 m cross-section and of 0.163 m height is placed on the flat plate. The charge of 23.7 g Demex explosive that corresponds to 27.26 g TNT is situated above the ground on height of 0.1 m and 1.5

m distance from the body. The geometry is shown in Figure 1. Two pressure gauges T1 and T2 are installed on the windward and leeward faces of the prism. Because of problem symmetry, only half of the physical domain was modeled to reduce the computation costs.

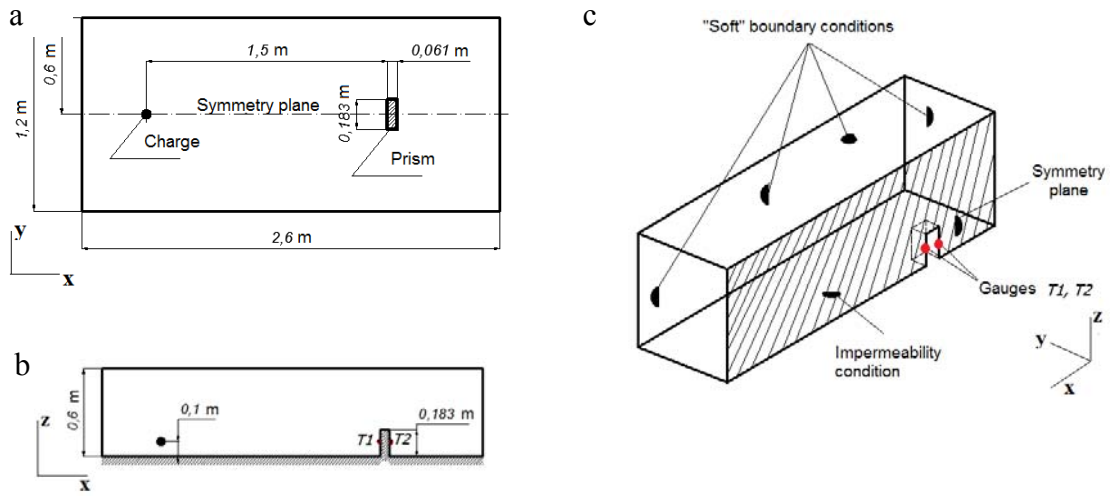



Figure 1: Plane view (a), central *y* cross-section (b) and 3D computational domain (c) for the test case #2

## 2.2 Test Case #2

This case has no any experimental data. The example was proposed and studied numerically in [14]. The geometry of Case #2 is shown in Figure 2. Two prisms of different heights are installed on the plate substrate. At initial moment, the charge of 189 g TNT explosive is detonated at the point placed on the ground by 0.2 m distance from the front face of the smaller prism. The virtual gauges are put on the windward and leeward faces of the prisms. This case has also been computed for a half of the domain taking into account symmetry of the problem.

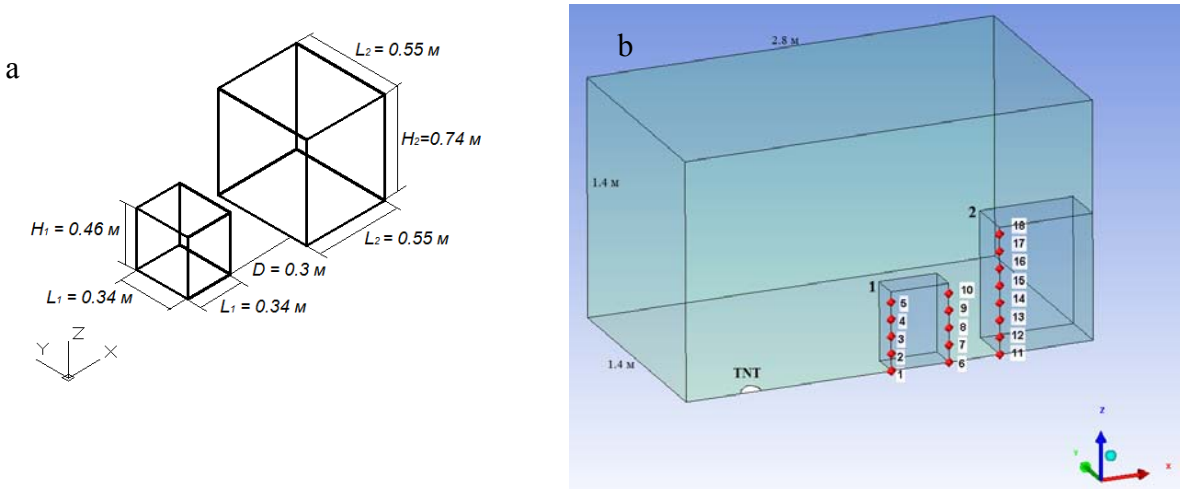



Figure 2: Prism geometry (a), and computational domain for the test Case #2

## 2.3 Test Case #3

The test Case #3 experimentally studied in [15, 16]. Seven building shown in Figure 3 as A - F placed on the plate ground surface were imitating a cityscape. The heights prisms were as follows: $H_A = H_G = 0.45$ m; $H_B = H_C = H_E = 0.4$ m; $H_D = H_F = 0.3$ m. The TNT charge of 16.0 g capacity is detonated at the point located by 0.04 m above the ground between buildings C and F. The distance between adjacent objects is less than or comparable to a linear scale of the prisms. In contrast to the test problems #1 and #2, for this configuration simulation cannot be carried out in a symmetrical approach and requires full 3D consideration. The static pressure behavior was recorded in several gauges placed at the prism walls and compared to the experimental data [15].

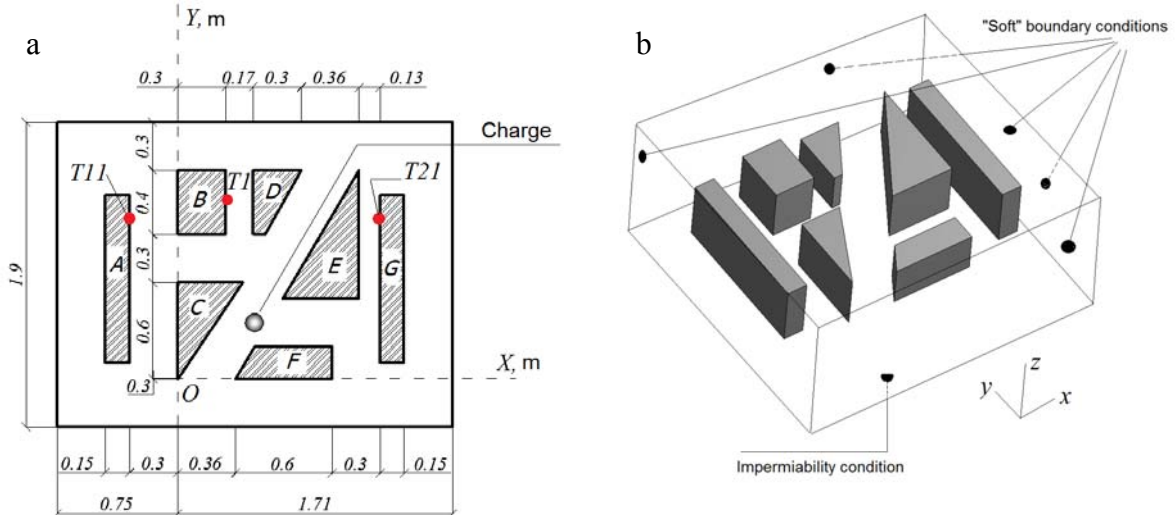



Fig. 3. Plane (a) and isometric (b) views of test Case #3 geometry

## 2.4 Test Case #4

In the test Case #4, (Figure 4), the internal explosions of TNT charge are studied for the closed cubical region of $3 \times 3 \times 3$ m without and with ventilation opening. The numerical simulation was carried out under the conditions of the experiments performed at the National Building Research Institute, Technion, Haifa, Israel [17]. The contact pressure was recorded during the computations at gauges shown in the Figure 4.

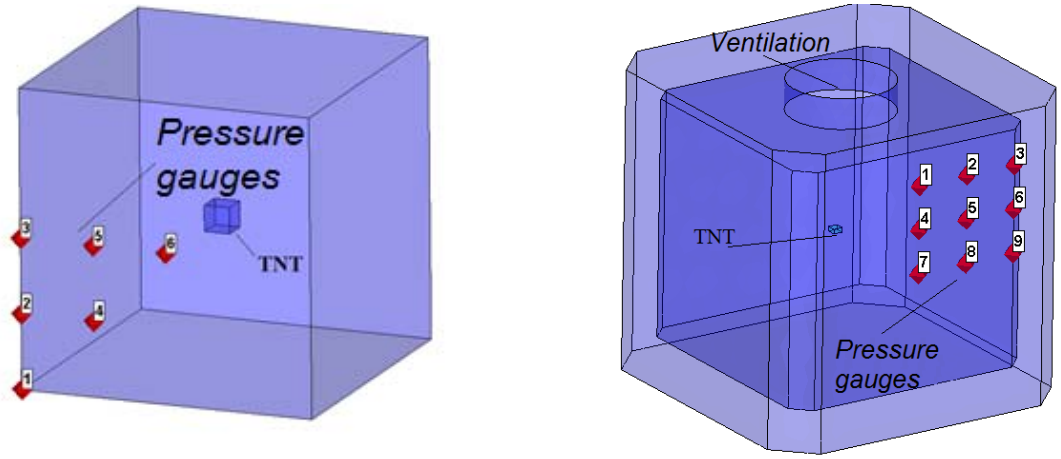



Figure 4: Geometries for test Case #4 without (a) and with ventilation (b)

## 3 MATHEMATICAL MODEL AND METHODS OF COMPUTATIONS

The air region was generated around the prismatic bodies in Cases #1 - 3 including the explosive charge. For the Case #4, the computational domain was confined by the room walls. At the initial time $t = 0$ the charge detonates, and detonation and shock wave spreads in open or closed space. Blast wave formation and propagation was computed with AUTODYN ANSYS 15.0 software on the basis of a hydrodynamic multi-material approach. The air and TNT properties used in the computations were retrieved from the standard AUTODYN library. The 3D Euler equations were used for computations complemented with the ideal gas equation of state for air and the JWL equation of state [21] for TNT:

$$\rho = A\left(1-\frac{\lambda\eta}{R_1}\right)e^{-\frac{R_1}{\eta}} + B\left(1-\frac{\lambda\eta}{R_2}\right)e^{-\frac{R_2}{\eta}} + \lambda\rho e, \qquad (1)$$

The values of the empirical constants A, B, R1, R2, □ are shown in Table1, e is the specific internal energy and □ = □/□0 is the relative specific density.

| Constant | *A,* kPa | *B,* kPa | $R_1$ | $R_2$ | $\lambda$ | *E*, kJ/m$^3$ | kg/m$^3$ |
|---|---|---|---|---|---|---|---|
| Value | $3{,}7377\cdot10^8$ | $3{,}7471\cdot10^6$ | 4,15 | 0,9 | 0,35 | $6{,}0\cdot10^6$ | 1630 |

Table 1: Values of empirical constants in (1).

At the external boundaries chosen far enough from region of interest, so called “soft” boundary conditions were set. All walls were supposed to be non-deformable, and “symmetry” conditions were used at the walls that guaranteed the absence of the flow through these boundaries. The second-order Godunov [18, 19] and FCT [20] finite-difference schemes were applied for space approximation. For the temporal approximation, the explicit scheme of the second order was used in compliance with the stability conditions. As the initial process of detonation in the open space is well described in the assumption of axial symmetry, to the moment when the primary shock wave reaches the plate or some prism surface, it is carried out in a 2D approximation. Then, the data obtained in the 2D calculations are interpolated at a 3D computational domain and simulations were continuing taking into account the shock wave interaction with the substrate and prism surfaces.

## 4 RESULTS OF COMPUTATIONS

### 4.1 Test Case #1

The grid convergence study was first performed using the second order sequel to Godunov’s type finite difference scheme. Four various grids with hexagonal cells of equal size were used. Table 2 shows the details of the grids.

| Grid # | Cell size in *x*-direction, cm | Cell size in *y*-direction, cm | Cell size in *z*-direction, cm | *Nx* | *Ny* | *Nz* | Total cell number |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 260 | 60 | 60 | 9.36E+05 |
| 2 | 0.5 | 1 | 1 | 520 | 60 | 60 | 1.872E+06 |
| 3 | 0.25 | 1 | 1 | 1040 | 60 | 60 | 3.744E+06 |
| 4 | 0.5 | 0.5 | 0.5 | 520 | 120 | 120 | 7.488E+06 |

Table 2: Computational grids for test case #1.

The computed pressure history in the T1 gouge placed on the windward prism face presented in Figure 5 shows that all grids give similar results. All further computations were performed on the grid #4.

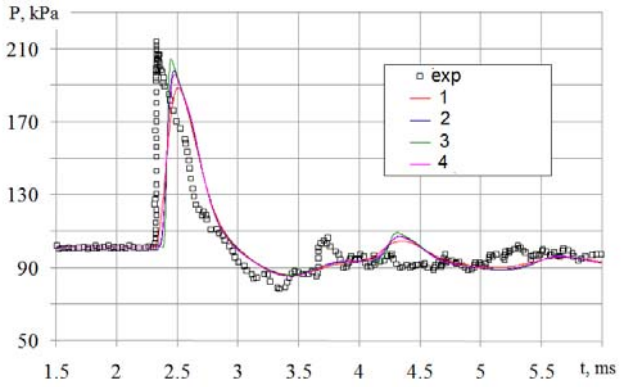



Figure 5: Grid convergence study for test case #1

Figure 6 shows the computed static pressure field in a symmetry plane in various time moments. Pressure histories written in gauges T1 and T2 are shown in Figure 7. As can be seen from Figure 6, a that corresponding $t$ = 0.018 ms, shock (1) is reflected regularly from the surface. In Figure 6, b one can see the irregular mode of shock reflection from the substrate. The wave front consisted of the original shock (1) and the Mach stem (2) propagates toward the prism. At the same time, reflected shocks (3) and (4) are moving toward the epicenter. After their interaction, a new secondary shock (5) is formed. In Figure 6, f pressure contours are shown at the moment when the primary shock front reaches the front wall of the prism. After reflection, the shock (6) appears, and negative phase starts with pressure level lower that 1 atm. Figures 6, g, h show the diffraction of shock wave (2) along the prism walls and spreading of reflected shock (6) toward the epicenter. Further, shocks (6) and (5) interact that lead to formation of secondary wave of lower intensity.

In Figure 7, besides the experimental data, three computation data are presented. The two of lines are obtained on the basis of 3D Euler equation computations using Godunov-type scheme of the second approximation order [18, 19] and FCT scheme [20]. The data obtained with empirical CONWEP function [4]. It can be seen that engineering-based approach permits one to predict the pressure maximum and shock arrival time to the frontal face of the prism which is in direct view of the explosion epicenter. But the CONWEP function is not able to describe the shielding effect as well as the secondary front formation. So an implementation of this approach is only possible to determine impact of the explosion on a single building.

As Figure 7 shows, there are several pressure peaks at frontal and rare faces of the prism. The first maximum $P_{max} \approx 213$ kPa in Figure 7, a, is due to the impact of the shock (2) on the frontal face of the prism. The secondary shocks (5) and (6) cause the local maxima of 115 kPa on the plot of pressure history taken in T1 gauge. It should be noted that Godunov-type scheme underpredicts the pressure peaks and gives some delay in shock arrival moments on the windward and leeward faces. The first pressure peak on the leeward face is caused by action of primary shock wrapped the prism surface, while the second peak is due to the action of the same wave reflected from the ground surface. The second maximum is underpredicted by both schemes. Nevertheless, the plot shows that FCT scheme works better comparing the Godunov-type scheme and gives adequate prediction of the main pressure peaks and shock arrival times.

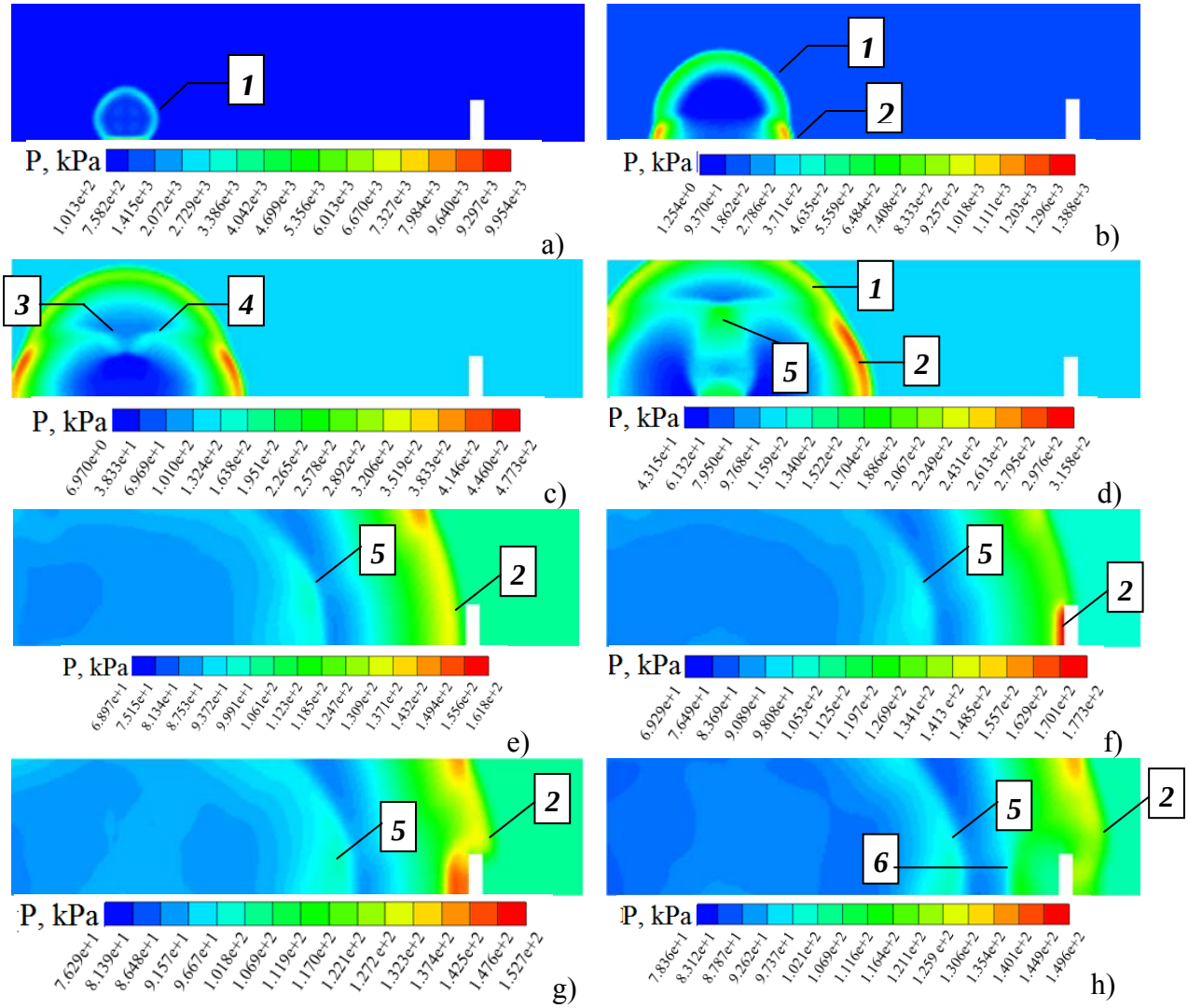



Figure 6: Instantaneous pressure fields at various time moments: 0.018 (a), 0.13 (b); 0.36 (c); 0.62 (d); 2.32 (e); 2.45 (f); 2.72 (g); 2.95 (h) ms.

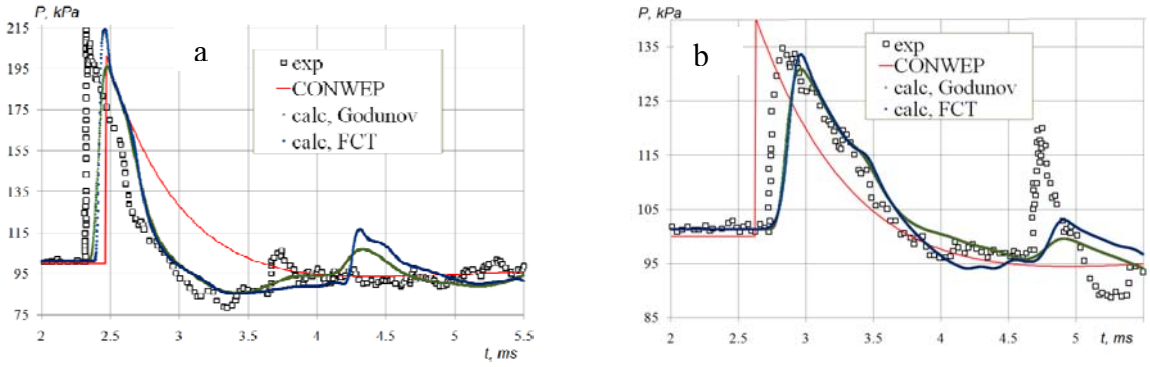



Figure 7: Computed (lines) and experimental (symbols) pressure history in T1 (a) and T2 (b) gauges

## 4.2 Test Case #2

The computation domain was of 1.4 × 2.8 × 1.4 m size and consisted of 3.8 million cells with the cell size of 5 mm. Visualization of pressure contours available during the post-processing stage allows a better understanding of the complex process of blast pressure interaction with a group of buildings. Figure 8 shows the computed static pressure at the ground level ($z = 0$) for different time moments after detonation. The pressure histories in the gauges placed on the front walls of the buildings are shown in Figure 9.

From the figures, it could be seen that the shock (1) is spherically propagating till it reaches the front surface of the first building at time $t$=0.076 ms. Pressure history written in gauges on front face of the first building shows that maximal pressure of 8.5e+3 kPa is observed in gauge #1. The shock (1) interacts with shock (2) formed as a result of reflection. The common wave front propagates along the wall and wraps the whole building 1. During the diffraction, in a vicinity of the front face of the first building pressure falls below the atmospheric level. High pressure levels shown in figures by red circles are observed in triple points of shock interactions and reflections from the walls and symmetry line. This pressure rise causes the second peaks observed in gauges 1-5 at t ≈ 0.8 ms. After shock coming to the frontal face of the second building, all the gauges placed on the surface show significant pressure increase.

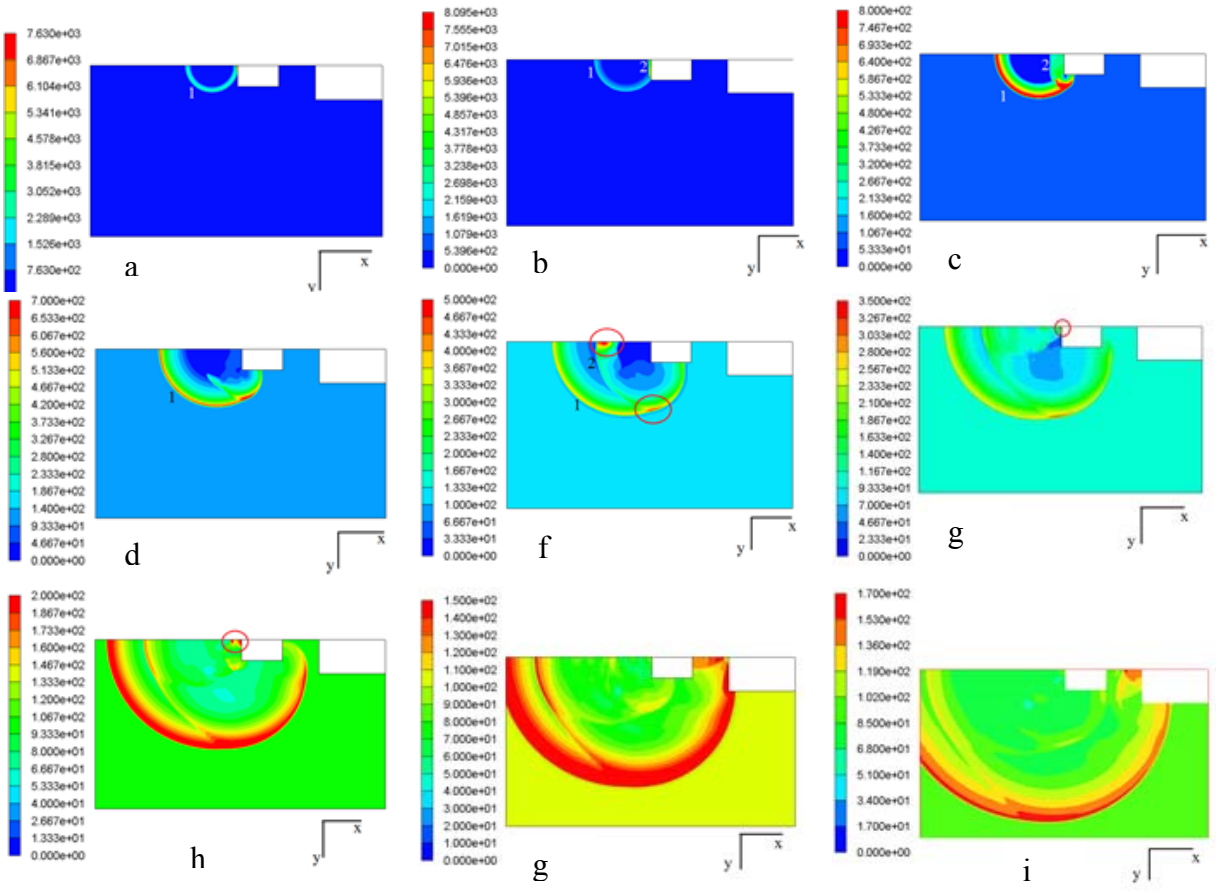



Figure 8: Instantaneous pressure fields for test case 2 at various time moments: 0.07 ms (a), 0.1 ms (b); 0.2 ms (c); 0.3 ms (d); 0.5 ms (e); 0.8 ms (f); 1.04 ms (g); 1.4 ms (h), 1.8 ms (i).

Nevertheless, it should be noted that amplitudes of the peaks here is lower by the order of magnitude comparing those on the frontal face of the first building due to the shield effect. Because of a narrow space between the first and the second buildings, some additional shocks of lower intensity arise due to multiple reflections from the canyon walls. The traces of these secondary shocks can be seen in the pressure history plots shown in Figure 9, b.

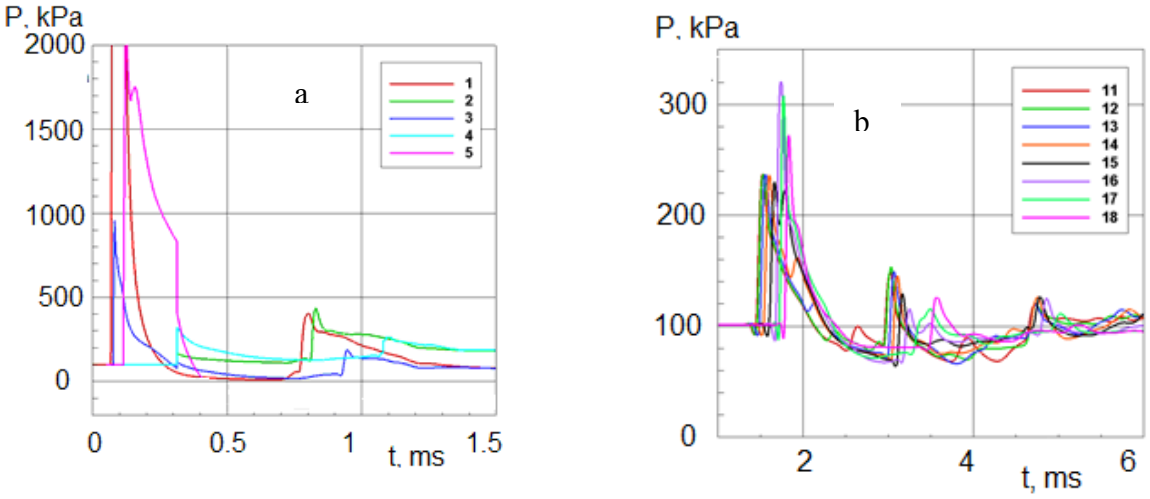



Figure 9: Computed (lines) pressure history in gauges places on the front face of Building 1 (a) and front face of Building 2 (b).

### 4.3 Test Case #3

The computations of this test case were performed in the computational domain with about 6 million hexa-cells each of 1×1×1 cm size.

Let's consider the shock-wave structure in the vicinity of the building complex. In Figure 10 the instantaneous static pressure contours in a horizontal plane-section z = 0.105 m. Spherical shock (1), formed as a result of detonation propagating in an open space till it reaches the corner of the building F (Figure 10, a). After reflection, the shock wave intensity is about 6 MPa. In some time moment the shock (1) falls on the wall of the building C and corner of the building E and reflects by irregular manner forming the triple points denoted by 3, 4 and 5 in Figure 10, b . The shock (2) arises after reflection of the shock (1) from the building C and propagates toward the epicenter. The interaction of shocks (6) and (2) results in formation of the shock (7). At the same time, as Figures 10, c, d show, the shock (2) goes around a sharp corner of the building C and interacts with the shock (1).

In Figures 10, e- one can see how the joint front of shocks 1 and 2 comes through the channel between building B and D. The action of this front leads to a fast pressure rise in the gauge T1 that can be observed in Figure 11,a. The second pressure peak is formed due to action of secondary shocks (11) formed in the epicenter in result of numerous interactions of reflected shocks.

The first peak is pressure history plot in T21 gauge (Figure 11,b) is a result of shock (8) passing between buildings E and G. Afterwards, a negative phase is observed. Corresponding rarefaction wave formed in result of shock diffractions can be seen in Figure 10,h.

Further wave structure analysis is complicated by a large number of shocks and rarefaction waves formed in result of numerous interferences, reflections and diffraction of the waves.

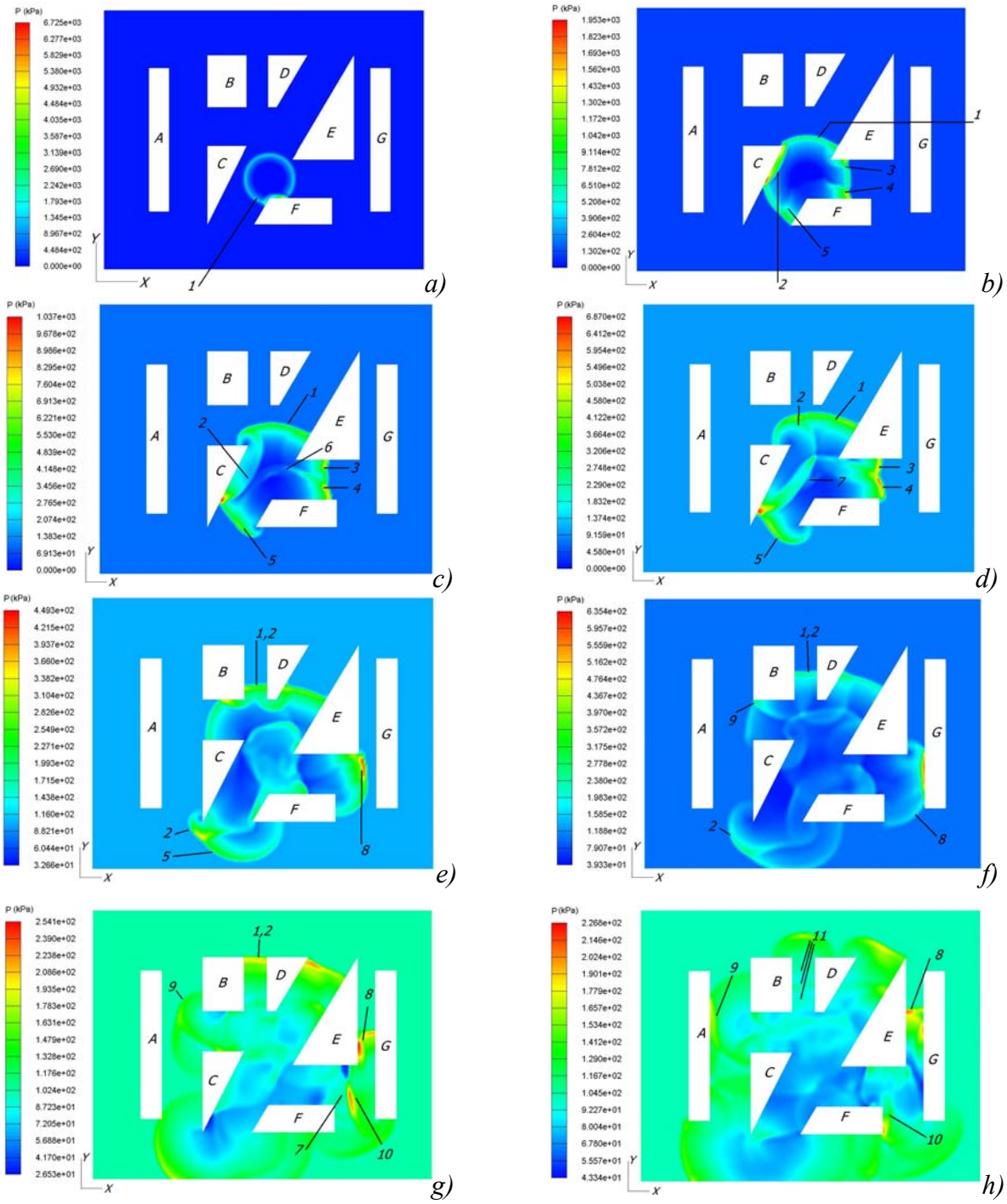



Figure 10: Instantaneous pressure fields for test case #3 at various time moments: 0.1 (a), 0.2 (b); 0.3 (c); 0.4 (d); 0.7 (e); 0.9 (f); 1.3 (g); 1.7 (h) ms.

The comparison of computed results and experimental data is shown in Figure 11. The symbols stand for experimental data [15] and the lines show the pressure history computed with two different schemes. The picture shows that both schemes predict the pressure loads on the building walls including positive and negative phases. The FCT scheme, as in previous case, is better in predicting the peak amplitudes and secondary waves.

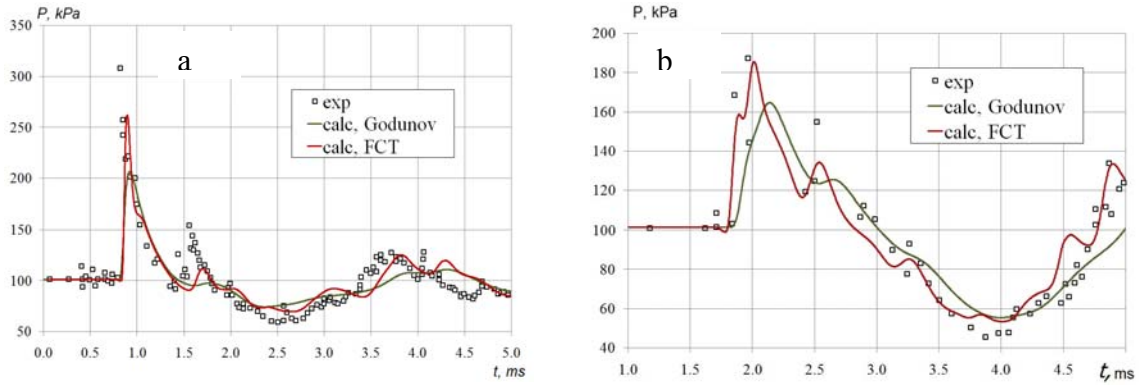



Figure 11: Computed (lines) and experimental (symbols) pressure history in T1 (a) and T21 (b) gauges.

### 4.4 Test Case #4

First, 3D modeling of the explosion was carried out of for the 16 kg TNT charge located in the center of the room of 300×300×300 cm size in fully confined space. During the calculation, the static pressure history was recorded at control points located on one of the walls which were assumed to be non-deformable. Computational grid for the task contained 343,000 cells. The characteristic size of the computational cell was 4 cm.

Figure 12 presents the pressure contours on the wall (x=0) showing the shock wave propagation and reflection. Figure 13, a presents computed pressure histories at various gauges on the wall $x = 0$. It can be seen that because of a big charge mass, very high pressure levels were obtained in this case.

The blast wave expands from the epicenter and first comes to the gauge #6 situated in the wall center. As the shock spreads along the wall, the pressure rises in gauges #4 and 5 while decreases in gauge #6. In time moment $t = 1.5$ ms shock front reaches wall edges and reflects from perpendicular walls that lead to significant increase of the pressure peak. The highest pressure level of about 20 MPa is observed at gauge #1 situated at a very corner of the room. After reflection from the wall, the shock front moves toward the room center, and after symmetrical reflection, the whole process is repeated, but the second pressure spikes are lower and average pressure in the room increases.

As the next stage, computations were performed under the conditions of [17] for the TNT charge, located inside the room of 290×290×270 cm with the vent opening of 120 cm diameter in the ceiling. All the internal corners of the room had a 20 cm chamfer. In the center of the room TNT explosive was located. The mass of the explosive chards was varying (Table 4). It the experiment, the static pressure was measured in nine sensors installed on one of room walls. Computational grid for the task contained 2 million cells with the size of 2 cm.

| Charge mass, kg | Charge dimensions, mm | | |
|---|---|---|---|
| | $\mathbf{L_1}$ | $\mathbf{L_2}$ | $\mathbf{L_3}$ |
| 0.5 | 104 | 67 | 52 |
| 1 | 134 | 104 | 52 |
| 1.5 | 156 | 134 | 52 |

Table 4: Charge parameter for Test case #4.

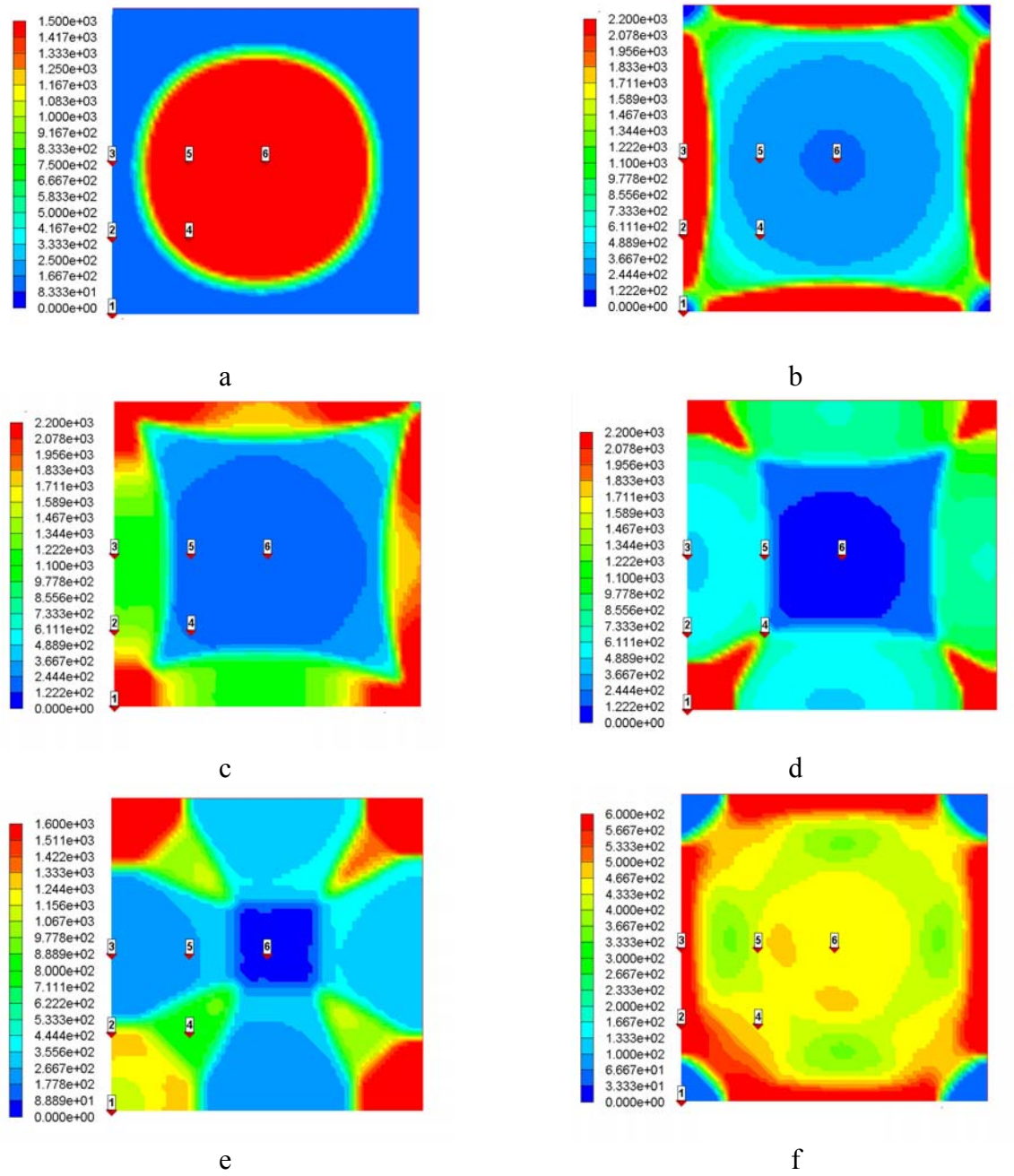



Figure 12: Instantaneous pressure fields for test case #4 (without ventilation) at various time moments: 0.9 (a), 1.5 (b); 1.7 (c); 1.97 (d); 2.3 (e); 4.8 (f) ms.

The computed pressure history for the charge mass of 0.5 kg is shown in Figure 13, b, and pressure plots at various time moments are presented in Figure 14. Because of significantly smaller charge mass, the amplitude of pressure spikes is much lower than in the previous computations. The highest pressure is observed in the gauge #7 located close to the wall center. Static pressure contours at the wall $z = 0$ for test Case #4 with vent, $m_{TNT}$ =0.5 kg at various time moments are presented in Figure 14. At the beginning, the whole shock wave structure is similar to that observer for a confined room. But due to the presence of vent opening, big part of the shock wave energy goes out of the room, an after the second reflection, pressure equalized, but the third pressure spike is observed in the gauge #7. The computed data were compared to experimental data [17] and a satisfactory agreement was obtained.

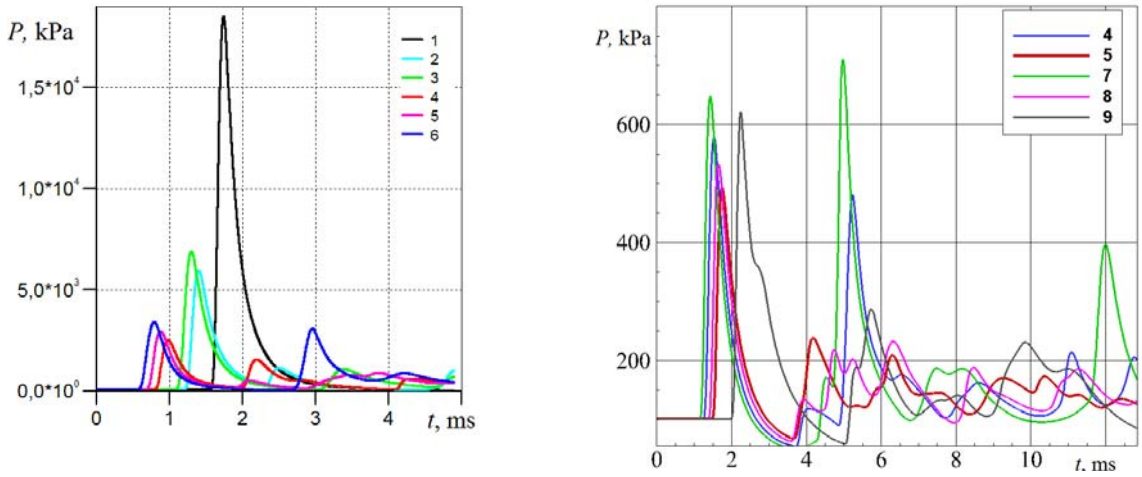



Figure 13: Computed pressure history at various gauges for test Case #4, $m_{TNT}$=16 kg without vent (a) and $m_{TNT}$ =0.5 kg with vent (b).

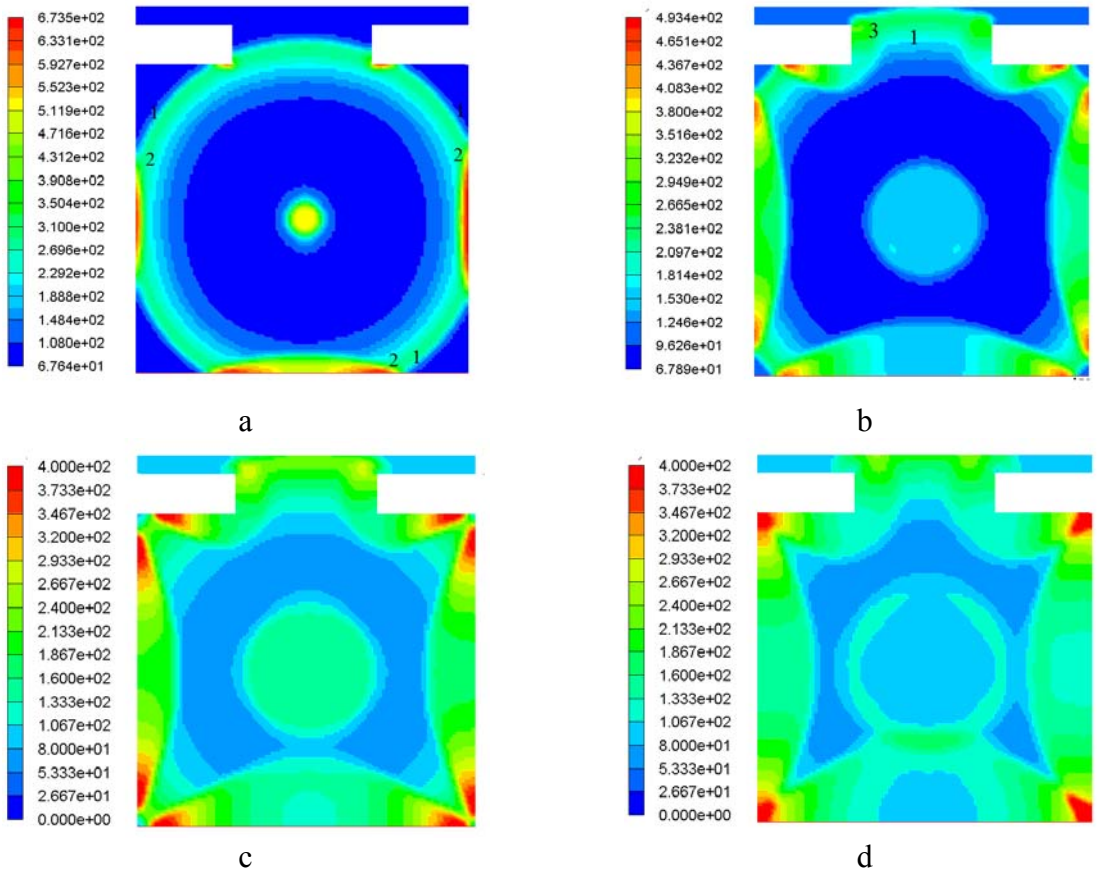



Figure 14: Instantaneous pressure fields for test case #4 (with ventilation) at various time moments: 1.4 (a), 1.9 (b); 2.0 (c); 2.2 (d) ms.

## 5 CONCLUSIONS

- Four different test cases of explosive charge detonation and blast wave propagation in open and fully or partly confided space were investigated numerically using ANSYS Autodyn software;

- The list of test cases includes: (1) single prismatic body installed on the flat substrate; (2) two prismatic body of different size installed on a flat plate; (3) generic cityscape of several prismatic bodies installed on a flat plate; (4) fully or partly confined room.
- On a basis of simulation results, a complex wave structure was analyzed, and all the peculiarities of flows and pressure history records on building / room walls were described and explained.
- Juxtaposition of simulation results with experimental data available were performed showing a satisfactory agreement.
- The comparison of two numerical scheme abilities has shown the better FCT properties to predict fine flow details
- The use of ANSYS AUTODYN tool can provide an effective approach to determining blast loads in an urban environment.
- The results indicate the necessity to take all the surrounding buildings into account when computing the blast loads on buildings in an urban environment.

## ACKNOWLEDGMENTS

The work was supported by the Ministry of Education of the Russian Federation (Project No. 211, Task No. № 2014/140 for executing scientific activities within the basic part of government order) and by Russian Foundation for Basic Research, grant № 15-07-06581.

## REFERENCES

[1] Donald O. Dusenberry, ed., *Handbook of blast resistant design of buildings.* John Wiley & Sons, 2010.

[2] W. E. Baker, P. A. Cox, P. S. Westine, J. J. Kulesz, R, A. Strehlow, *Explosion Hazards and Evaluation*. Elsevier Scientific Publishing Company, 1983.

[3] T. Ngo, P. Mendis, A. Gupta, and J. Ramsay, Blast loading and blast effects on structures - an overview. *EJSE international*, Special Issue: Loading on structures, 76 – 91, 2007.

[4] Hyde D., *Users Guide for Microcomputer Programs CONWEP and FUNPRO – Applications of TM 5-855-1*. U.S. Army Engineer Waterways Experimental Station, Vicksburg, 1988.

[5] Kingery C.N. and Bulmash G., *Airblast Parameters from TNT Spherical Air Burst and Hemispherical Surface Burst, Report ARBL-TR-02555*, U.S. Army BRL, Aberdeen Proving Ground, MD, 1984.

[6] *Design and Analysis of Hardened Structures to Conventional Weapons Effects*. Washington, DC: The Departments of the Army, Air Force, Navy and The Defense Special Weapons Agency Defense Special Weapons Agency (DSWA), 1998.

[7] Randers-Pehrson G. *Airblast Loading Model for DYNA2D and DYNA3D. ARL-TR-1310*. Army Research Laboratory, Aberdeen Proving Ground. MD, 1997.

[8] Rose T. A. A Computational Tool for Airblast Calculations, *Air3d version 9 users' guide, Engineering Systems Department* Cranfield University, 2006.

[9] Smith, P. D. and Rose, T. A., Blast wave propagation in city streets—an overview. *Prog. Struct. Engng Mater*., **8,** 16 – 28. 2006.

[10] http://www.virtualcitysystems.de/en/114-english/references/research/352-detorba-2

[11] Valger S.A., Danilov M.N., Fedorov A.V., Fedorova N.N. The comparison of results of simulation of shock wave action on structures using ANSYS AUTODYN and LS-DYNA software // *News of Higher Educational Institutions. Construction*. No. 11 (671). 77 – 92, 2014.

[12] Fedorov A.V., Fedorova N.N., Fomin P.A., Valger S.A. *Propagation of explosive shock waves in block-up spaces*. Novosibirsk: Novosibirsk State University of Architecture and Civil Engineering (Sibstrin), 2015.

[13] Rose T.A. *An approach to the evaluation of blast loads on finite and semi-infinite structures.* PhD thesis, Engineering Systems Department, Cranfield University, Royal Military College of Science, February 2001.

[14] Remennikov A. M., Rose T. A., Modelling blast loads on buildings in complex city geometries. *Journal Computers and Structures,* **83**, (27), 2197 - 2205, 2005.

[15] Matthew A. Brittle, *Blast propagation in a geometrically complex environment. MSc dissertation*, Craneld University, Defence College of Management and Technology, Defence Academy of the UK, Shrivenham, Swindon, SN6 8LA, UK, Jul 2004.

[16] Smith P.D., Rose T.A., and Brittle M.A. Analysis of a generic cityscape using an adaptive mesh CFD code. *Proceedings of the 12th International Symposium on Interaction of the Effects of Munitions with Structures,* New Orleans, USA, 13th–16th September. 2005.

[17] V.R. Feldgun, Y.S. Karinski, I. Edri, D.Tsemakh and D.Z. Yankelevsky, On Blast Pressure Analysis Due to a Partially Confined Explosion: I – Experimental Studies. *International Journal of Protective Structures*, **2**(1), 1 – 20, 2011.

[18] S. K. Godunov, A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics, *Mat. Sb*., **47**, 271–290, 1959.

[19] Van Leer B., Towards the ultimate conservative difference scheme V: a second – order sequel to Godunov‘s method. *Journ. of Comput. Phys.* **32** (1), 101 – 136, 1983.

[20] Zalesak S. T., Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids. *Journal of Comp. Phys.* **31**, 335–362, 1979.

[21] Baum F.A., Orlenko L.P., Stanukevich K.P., Chelyshev V.P., Shaehter B.I. *Physics of explosion.* Moscow: Nauka, 1975.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON A DYNAMIC SCHEDULING ALGORITHM FOR MASSIVELY PARALLEL COMPUTATIONS OF ATOMIC ISOTOPES

**Elizaveta Dorofeeva**[1], **Jatin Arora**[2], **Stefan Typel**[3], **Gevorg Poghosyan**[1], **Peter Sanders**[4], **Achim Streit**[1]

[1] Steinbuch Centre for Computing, Karlsruhe Institute of Technology
Hermann-von-Helmholtz-Platz 1, 76344, Eggenstein-Leopoldshafen, Germany
e-mail: elizaveta.dorofeeva@kit.edu

[2] PEC University of Technology
Sector -12, Chandigarh, 160012, India

[3] GSI Helmholtz Center for Heavy Ion Research
Planckstraße 1, 64291, Darmstadt, Germany

[4] Institute of Theoretical Informatics, Karlsruhe Institute of Technology
Kaiserstraße 12, 76131, Karlsruhe, Germany

**Keywords:** High-performance computing, Scalability, Scheduling, Nuclear clusters.

**Abstract.** *The goal of the work is to enhance the execution scalability of the code called AtomicClusters for evaluations of in-medium properties of nuclear clusters to extreme scales. In order to fully exploit the computing power of massively parallel supercomputing systems, the code was supplemented with parallel output and dynamic scheduling system based on task-stealing technique. The scheduling system was implemented for state-of-the-art distributed-memory high-performance computers (HPC) using the advanced features of Message Passing Interface (MPI) as an independent adjustable module. The parallel output was integrated into the code using MPI IO. A number of strong scaling tests was performed for the resulting parallel software. An almost linear scalability was reached on up to 4000 cores. The code scales up to at least 38400 processes, but with lower speedup. The obtained results are discussed in the fifth section of the paper.*

## 1 INTRODUCTION

The dynamical evolution of core-collapse supernovae, the static properties of neutron stars, and the conditions for nucleosynthesis could be determined by Nuclear Equation of States (EoS) of stellar matter [1], the chemical composition of stellar matter. At densities below nuclear saturation density the nuclear matter is a mixture of nucleons, nuclei and electrons that could be considered as huge atoms also known as nuclei clusters. Their abundances depend strongly on the thermodynamic conditions such as density, temperature, and electron fraction. Since nuclei, which are composite objects, are immersed in a dense medium, their properties change significantly with the parameters mentioned due to the nuclear and electro-magnetic interaction with the surrounding particles and the action of the Pauli Exclusion Principle [1]. The latter is the main cause for the dissolution of nuclei with increasing density (Mott effect) [1]. Hitherto, several thousand clusters have to be considered for various values of the thermodynamic parameters as well as atomic mass and charge number to identify stable ones.

The whole range of parameters for the evaluation of in-medium properties of all possible nuclear clusters is almost 500 million variations. Thus, the massive parallel calculations are crucial to complete full analyses. To reach good load balance of such computations at extreme scales, a middleware between cluster scheduling system and the simulation software, based on scalable scheduling algorithm has to be developed. Since the scalability of the commonly used single-queue master-managed technique is limited due to the Amdahl's law [2] (the work distribution can be done only sequentially as there is only one master), and the calculation time for each set of parameters of computation for nuclear clusters cannot be accurately predicted (see chapter 3.1), the usage of static scheduling algorithms proved to be inefficient and a principally different scheduling approach known as task stealing was used.

The task stealing suspects each process having own task stack, accessible concurrently by all the processes in system. Shared-memory systems meet the requirements of such an approach, but do not give enough computational resources for the simulations (the Blacklight machine of the Pittsburgh Supercomputing Center (PSC), the world's largest shared-memory computing system [3], features 512 eight-core Intel Xeon 7500 processors). Distributed-memory systems offer enough performance, but do not support direct access to the operative memory of another process. Therefore, a hybrid concept of the shared-memory-like synchronization based on MPI-3 passive-target remote memory access [4] was implemented.

The scalability tests of the designed parallel system were performed on the computational resource ForHLR Phase I (SCC/KIT, Karlsruhe) on up to 4000 cores and on the Cray XC40 Hazel Hen system (HLRS Stuttgart, #8 in Top500) on up to 38400 cores, using renowned MPI libraries.

The paper is organized as follows. Section 2 shortly describes physical models used in the code. Section 3 presents overview of applicable scheduling techniques. Section 4 gives more technical details of the implementation of the chosen approach. Section 5 is devoted to the quality assessment of the re-designed software. Conclusions and future work prospects are given in the section 6.

## 2 CODE "ATOMICCLUSTERS" FOR THE EVALUATION OF IN-MEDIUM PROPERTIES OF NUCLEAR CLUSTERS

A parallel *AtomicClusters* [5] code for the evaluation of in-medium properties of nuclear clusters was developed in a joint R&D project of the Steinbuch Centre for Computing of KIT with the Nuclear Theory group of GSI and is used to model the behavior of composite particles at zero and different non zero temperatures. The code allows searching for stable isotopes

and defining EoS consisting of such isotopes at extremely dense matter. This kind of EoS is an essential ingredient in astrophysical model simulations.

The main objective of the *AtomicClusters* code is to calculate the change of properties of nuclei, mainly energy and radii, embedded in a medium of nucleons and electrons, required to have charge neutrality of the total system as in stellar matter, for given temperature and (baryon) density. This is achieved by comparing the results of two calculations: one of homogeneous nuclear matter and one where the formation of a cluster is calculated inside a spherical Wigner-Seitz cell (inhomogeneous matter). For this we use a relativistic energy density functional (EDF) that is based on a phenomenological relativistic mean-field (RMF) model [6] for nuclei with density-dependent nucleon-meson couplings. The exchange of mesons models the interaction between the nucleons. From the EDF the relevant field equations, i.e. coupled differential equations, for the particles are derived. They have to be solved self-consistently. This is done using the relativistic Thomas-Fermi approximation. The results for energy shifts etc. are needed for the complete chart of nuclei, including unstable exotic nuclei, for different densities and temperatures. They will be used then in a second step as input for a model for the equation of state [1] of stellar matter (generalized relativistic density functional) that incorporates the full table of nuclei.

So far, only simple approximations for the energy shifts are used, but they should be replaced in the future by more realistic results based on the underlying EDF.

The parallelization technique used in the initial version of the *AtomicClusters* code was based on static domain decomposition, and proved to be inefficient (see section 5).

## 3 SCHEDULING SYSTEM

The necessity of optimal and effective use of limited and expensive computational time on any HPC system is crucial. Efficient resource utilization can seldom be reached without long term re-engineering of the scientific simulation codes. The code rebuilt may result in changes of the scientific concepts and architecture, practically leading to fundamentally new software and possibly different scientific results. In order to advance the *AtomicClusters* code for state-of-the-art HPC systems and avoid affecting the underlying physical models, the external scheduling system was developed. As distinguished from code parallelization techniques, the scheduling system operates code-defined computational units instead of mathematical operations, thereby not influencing the simulation results.

For the *AtomicClusters* code the scheduler-managed unit (hereinafter task), is one set of possible isotope parameters, namely density, temperature, atomic number and electron fraction. Since the calculation time for each task is unknown, the simple statement of the scheduling problem in case of heterogeneous computational system can be represented as follows: the project consists of N independent tasks of variable size; the tasks must be distributed among M: M<<N equal resources so that the latest are used in a most efficient way: the makespan of the project is minimized and the resources utilization is maximized; the distribution method must scale up to M>>10000. Standard static algorithms as those solving the bin packing problem [7] cannot be used to schedule tasks of undefined size. To solve the problem, the following scheduling approaches were implemented and evaluated.

### 3.1 Static scheduling techniques

The first approach is a static distribution of equal N/M number of tasks among processes, or static domain decomposition. The advantages of the approach are the simplicity and the lack of bottlenecks. Since the input range of parameters is represented in an incremental form, it is small enough to fit into the memory of a single process irrespective of the problem size,

and the distribution of parameters among processes is done in parallel, namely the range of parameters to be processed is calculated on each process according to its rank. The disadvantages of the approach are the non-optimized completion time and the high load imbalance on the computational systems with static resource management. This technique gives better result on systems with dynamic resource management, where the finished processes can be freed immediately without waiting for the whole computation to complete [8]. Herewith, fulfilling the resource occupation condition we are still not satisfying the makespan minimization criteria.

The second approach is a statistics-based grouping technique. The sizes of new tasks are estimated using the 3D spline-polynomial interpolation of the collected statistical data. The groups of tasks with similar computational intensity (cumulative size) are formed and scheduled as single units of equal size. Figure 1 demonstrates the estimated task size for the range of atomic numbers from 50 to 300 and temperatures from 25 to 140 MeV.

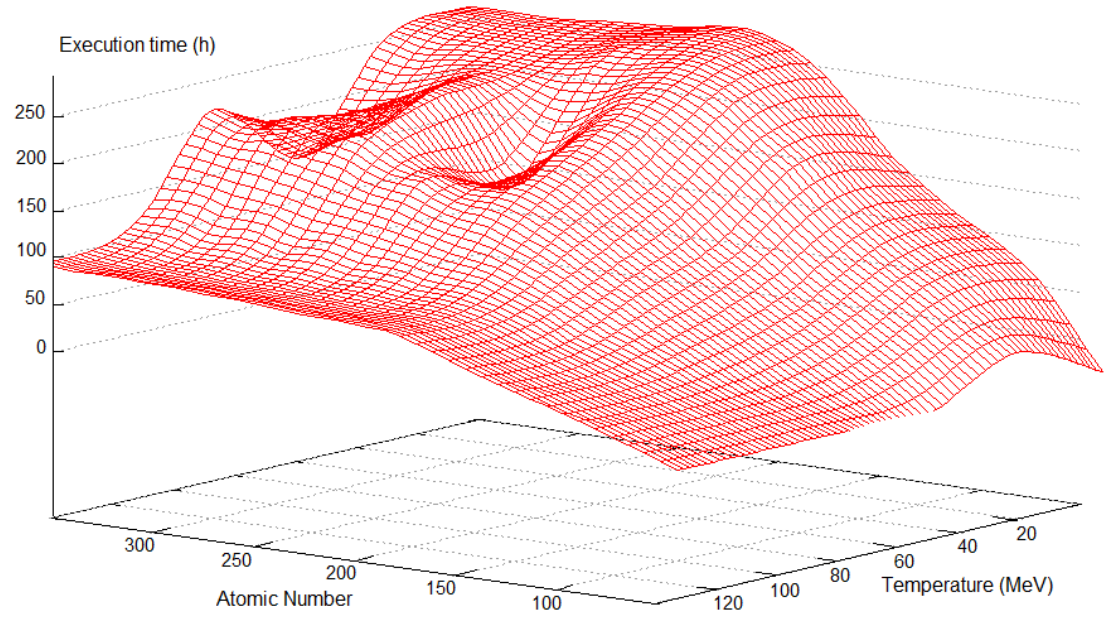



Figure 1: Estimation of task sizes for given temperature and atomic number.

On particular range of parameters, where the size function is smooth, the approach shows feasible results. However, the scalability of the approach is limited due to statistics I/O bottleneck, and the prediction is not possible on the whole range of parameters. This technique remains useful for the simulations where the task size dependency on the input parameters can be approximated with monotonous function, such as the software to search for periodic gravitational wave signals [9], where we successfully applied the grouping approach. In case of nuclear clusters study, where the computation of each set of parameters must be done only once, this approach must be advanced.

### 3.2 Dynamic task-stealing scheduling technique

Taking into account specifics of the *AtomicClusters* code such as non-predictable task sizes and independency of the tasks, the approach based on a task-stealing technique was chosen for the scheduling system.

The idea of the task-stealing technique is derived from work-stealing. In [10] work-stealing was opposed to work-sharing – the technique where a single scheduler is responsible for the work distribution. Both work-stealing and work-sharing do not meet the requirements we

have set. The work-sharing technique is master-managed and does not scale due to Amdahl's law. The work-stealing supposes that the tasks can be subdivided, that contradicts the idea of operating code-defined computational units.

The following scheduling algorithm was chosen to be implemented. The tasks are initially distributed among processes as in the first approach, but as soon as the process completes the tasks from own queue, it starts "stealing" tasks from other queues. The task-stealing algorithm is initially parallel since each process is following the same steps. Provided the communication time is negligible compared to computation time, the load imbalance is not higher than the size of the biggest task, since each process is occupied while there are tasks still left in the system. Contrary to the simple distribution, this approach conduces to the completion time minimization.

To balance the load, two techniques were applied: the random polling [11, 12], where the stealing requests are send to a randomly chosen target, and the ordered polling with match memorization, which means repetitive stealing from the same target until it is not empty. The first is quite efficient despite its simplicity and match unpredictability, but requires supplemental termination detection mechanism. The second is finite, since no new tasks are arising during the computational process, and the cycle can be subdivided into nested loops, first searching within the immediately accessible area (e.g. shared memory region), thus profiting from the HPC architecture. However, ordered polling can cause work maldistribution, and therefore higher communicational load, since the number of non-empty targets will decrease faster.

## 4 IMPLEMENTATION OF THE TASK-STEALING ALGORITHM

In the era of HPC one of the biggest problem limiting parallel applications scalability is the synchronization problem. The first goal of the scheduler implementation was to minimize interprocessor communications, therethrough reaching higher scalability. The task stealing algorithm is naturally suitable for the shared-memory systems, since it suspects each process having its own task stack, accessible concurrently by all the processes in system. To apply the task-stealing technique for distributed memory systems and optimize the interconnection between processes, shared-memory-like synchronization based on MPI passive-target remote memory access (RMA) [4] was implemented.

The RMA is a new MPI feature, requiring an external progress engine to remain truly passive-target [13], and not fully supported by all MPI libraries. Normally the MPI progress engine on each process activates only when the process itself is doing an active MPI communication. Therefore, once initiated the passive-target access epoch stalls and can be processed only when the target calls an MPI routine. The external progress engine is responsible for the background progression, irrespective of the processes behavior, and allows avoiding deadlocks during passive-target access.

For the scheduling system to remain efficient irrespective of the HPC environment, the additional optionally activated synchronization mechanism was integrated. The mechanism periodically simulates MPI activity during the computational process, therethrough poking the MPI progress engine and activating stalled calls from other processes. Considering that no additional communication is done within the synchronization call, the increase of the single task computation time is negligible. The disadvantage of such a mechanism is a conflict of load balancing and synchronization. On the one hand, passive-target access epochs happening between the synchronization calls still stall, causing load imbalance, but on the other hand the meaningful decrease of synchronization periods may badly affect the computation time of each task. Therefore, the usage of the MPI libraries supporting asynchronous progress is preferable.

## 5 EVALUATION OF THE DYNAMIC SCHEDULING SYSTEM

The quality assessment of the re-designed parallel software was performed in a number of scalability tests. Figure 2 demonstrates the results of strong scaling of the code on the ForHLR Phase I system. The ordered polling with match memorization was chosen as a balancing technique. The tests were run with all three MPI libraries installed on the system: MVAPICH [13], Intel MPI and Open MPI. The default settings of the libraries were used. The speedup was measured relating to the sequential runtime with the same library and tuning parameters.

Among the libraries available on the system only MVAPICH fully supports the asynchronous progress. However, minimizing interconnection time to a negligible value by use of the external progress engine caused the decrease of performance. The cumulative runtime of all tasks on all processes increased. The Intel MPI library shows better results for the simulations with additional synchronization, which can be explained with optimal library configuration for the system.

The magenta line is given for comparison and refers to the scalability results of the original (based on static domain decomposition) parallel version of the *AtomicClusters* code without integrated scheduling system.

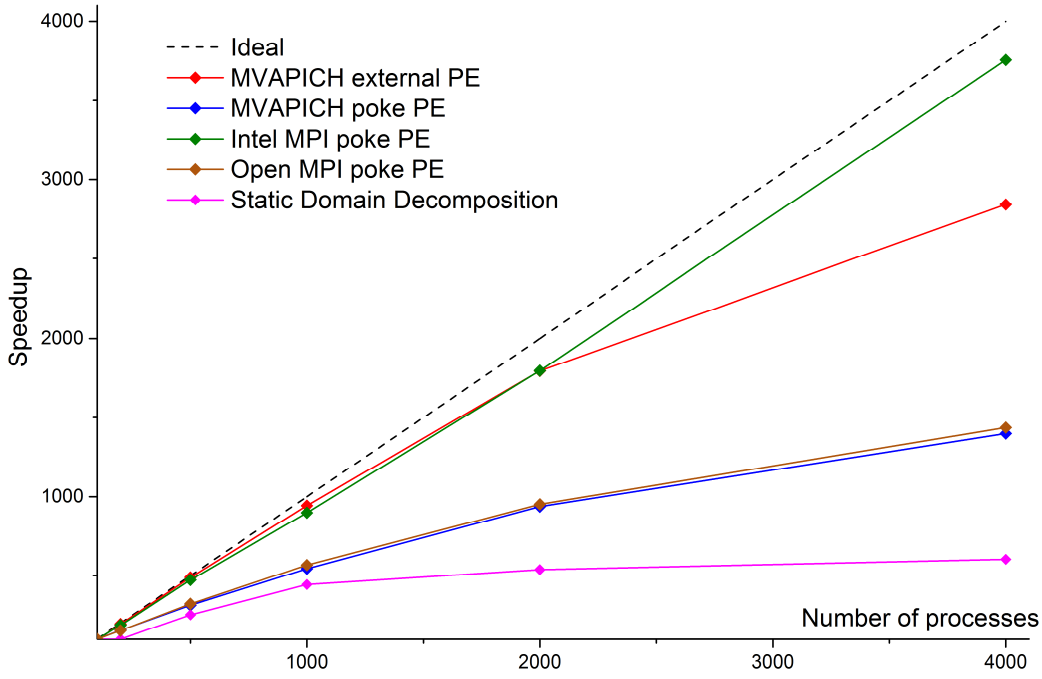



Figure 2: Scalability tests on ForHLR system with different MPI libraries.

The diversity of the results proves that to reach higher efficiency of the parallel application at any system, the initial tuning of the system software in order to optimize its settings for the used hardware and application must be done.

Figure 3 demonstrates the results of strong scaling of the code on the Cray XC40 Hazel Hen system with the default settings of the Cray environment.

Both load balancing mechanisms were tested on higher scales: the ordered stealing with repetitive target choice on match and the random polling. Additional synchronization was used to provide non-stalling passive target access.

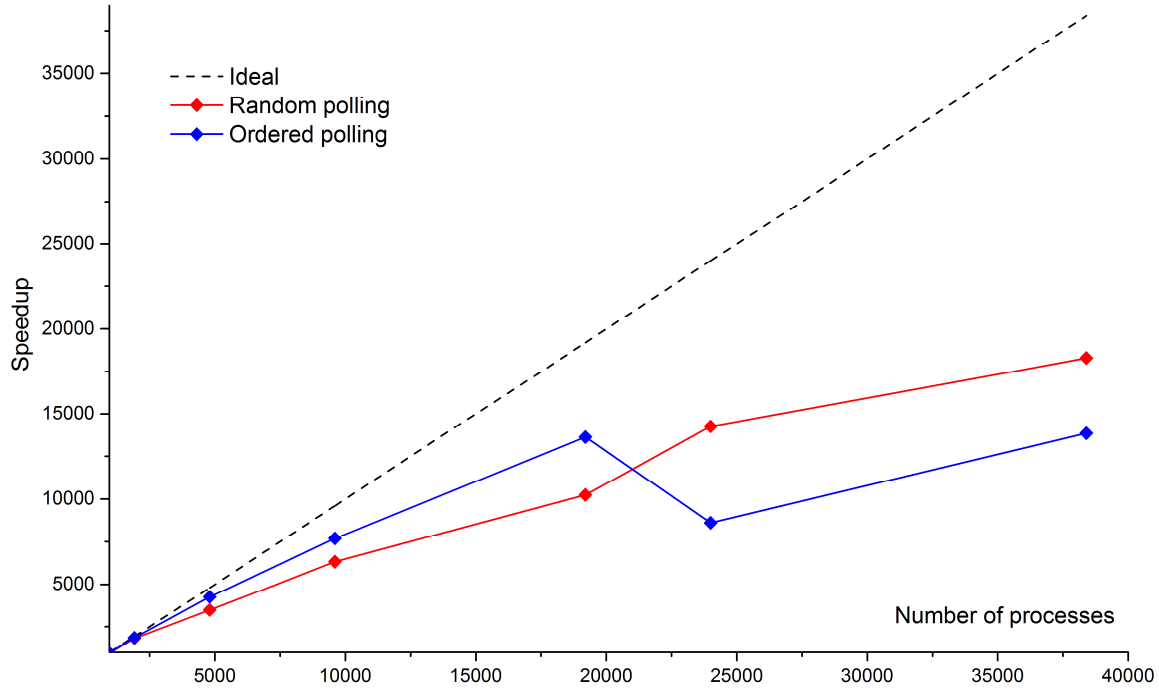



Figure 3: Scalability tests on Cray XC40 Hazel Hen system with different load balancing mechanisms.

As was predicted, the random target choice showed better results on scale-up. We explain it with the tasks maldistribution, caused by faster decrease of the number of non-empty targets, which is the consequence of repetitive target choice on match. This imbalance leads to higher number of access epochs to non-empty targets left and increases the number of concurrent access epochs. For random polling, the possibility of simultaneous epoch to the same target is quite low.

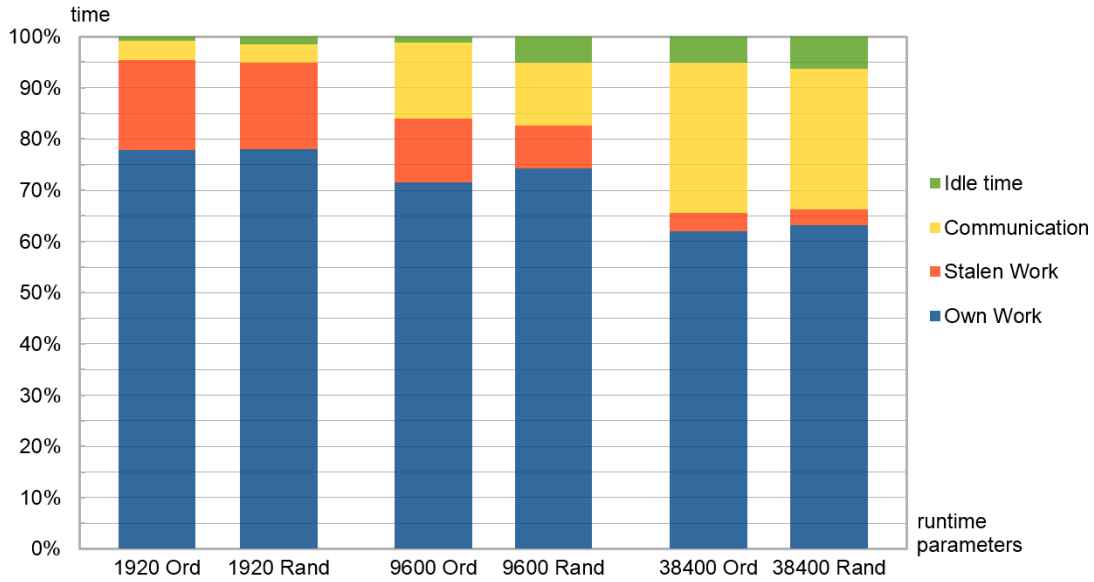



Figure 4: Relative work/idle/communication time distribution.

The overall result on Hazel Hen appeared to be worse than on ForHLR. Detailed evaluation of the results allowed discovering a bottleneck in the parallel output system, which can be

one of the reasons of speedup decrease. Here the optimization of the parallel output for *AtomicClusters* code is necessary and is under development. Another reason can be a not optimal (too small) size of the problem being solved. The performance of a single computational unit (core) of Hazel Hen system is almost two times greater than one of ForHLR according to the official documentation [14, 15], and is approximately three times greater according to the measurement of the single-process runtime of the same task on each machine. Thus the problem size must be enlarged to profit from Gustafson's law principles [16].

Higher performance also causes decrease of the computational time of a single task, which means that communication intensity increases. Figure 4 illustrates this increase, demonstrating the average distribution of processor time among work on own (received by initial distribution) tasks, work on stolen tasks, communication and staying idle, for different runtime parameters (number of processors and stealing order).

To minimize communications, the tasks grouping will be used so that groups rather than single tasks could be managed by scheduler as one unit. The optimal size of such a group as well as the preferable size of the whole simulated problem is a subject of further investigation.

## 6 CONCLUSIONS

The existing *AtomicClusters* code was supplemented with a dynamic scheduling system based on task stealing technique. The scheduling system was implemented using the advanced methods of MPI for the distributed memory system. The tests performed on the ForHLR system up to 4000 cores showed good scalability of the designed software. Results, obtained for higher amount of cores on the Hazel Hen system, allowed us appointing the following prospects of further system optimization:

- the revision of the parallel output system and its re-integration from code into scheduler;
- the estimation of the optimal size of the units operated by the scheduling system depending on the system performance;
- the investigation of the tuning parameters of the environments used and an evaluation of the optimal configuration.

Summarizing, one can claim that task stealing based scheduling concept appeared to be a good alternative to code parallelization and re-designing parallel software. Being used with optimal parameters it can provide almost linear speedup.

## REFERENCES

[1] S. Typel, H.H. Wolter, G. Röpke, and D. Blaschke, Effects of the liquid-gas phase transition and cluster formation on the symmetry energy, *The European Physical Journal A*, **50:17**, 2014.

[2] G.M. Amdahl, Validity of the Single Processor Approach to Achieving Large-Scale Computing Capabilities, *AFIPS Conference Proceedings*, **30**, 483–485, 1967.

[3] *Blacklight, the World's Largest Coherent Shared-Memory Computing System*, https://www.psc.edu/publicinfo/news/2010/101110_Blacklight.php, 2010, retrieved on 28-02-2016.

[4] W.D. Gropp, R. Thakur, Revealing the performance of MPI RMA implementations, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, Springer Berlin Heidelberg, 272-280, 2007.

[5] G. Poghosyan, et al., High-Level Support Activities of Simulation Laboratory E&A Particles, *High Performance Computing in Science and Engineering*, **14**, 121-133, 2015.

[6] S. Typel, H.H. Wolter, Relativistic mean field calculations with density dependent meson nucleon coupling, *Nuclear Physics A*, **656**, 331-364, 1999.

[7] R. Lewis, A General-Purpose Hill-Climbing Method for Order Independent Minimum Grouping Problems: A Case Study in Graph Colouring and Bin Packing, *Computers and Operations Research*, **36 (7)**, 2295–2310, 2009

[8] W. Gropp, E. Lusk, and R. Thakur, *Using MPI-2 Advanced Features of the Message-Passing Interface*, The MIT Press, 1999.

[9] G. Poghosyan, S. Matta, A. Streit, M. Beiger, A. Krolak, Architecture, implementation and parallelization of the software to search for periodic gravitational wave signals. *Computer Physics Communications*, **188**, 167-176, 2014.

[10] R.D. Blumofe, C.E. Leiserson, Scheduling multithreaded computations by work stealing, *Journal of the ACM*, **46:5**, 720-748, 1999.

[11] P. Sanders, Randomized Receiver Initiated Load Balancing Algorithms for Tree Shaped Computations, *The Computer Journal*, **45:5**, 561-573, 2002.

[12] R. Finkel, U. Manber. DIB - A distributed implementation of back-tracking. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, **9(2)**, 235-256, 1987.

[13] G. Santhanaraman, S. Narravula, D.K. Panda, Designing passive synchronization for MPI-2 one-sided communication to maximize overlap. *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, 1-11, IEEE, 2008.

[14] *ForHLR Phase I - Hardware and Architecture*, https://www.bwhpc-c5.de/wiki/index.php/ForHLR_Phase_I_-_Hardware_and_Architecture, retrieved on 28-02-2016.

[15] Hazel Hen - Cray XC40, Xeon E5-2680v3 12C 2.5GHz, Aries interconnect, http://top500.org/system/178446, retrieved on 28-02-2016.

[16] J.L. Gustafson, Revaluating Amdahl's Law, *Communications of the ACM*, **31(5)**, 532-533, 1988.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# APPLICATION THE SPARSE MATRIX METHOD TO CALCULATE THE METAL ELASTIC STRESS-STRAIN STATE, USING THE FINITE ELEMENT METHOD

**Inna I. Barankova**[1], **Uliana V. Mikhailova**[1], **Victor B. Demidovich**[2]

[1] Nosov Magnitogorsk State Technical University (NMSTU)
38 Lenina str, Magnitogorsk, 455000, Russian Federation
e-mail: inna_barankova@mail.ru, ylianapost@gmail.com

[2] Sankt-Petersburg Electrotechnical University (LETI)
5 Prof.Popov str, St.Petersburg, 197376, Russian Federation
e-mail: vbdemidovich@mail.ru

**Keywords:** sparse matrix, elastic stress-strain state, compliance, solid mechanics.

**Abstract.** *The metal stress-strain state simulation needs high precision computations. The use of any simplifying assumptions, such as the plane section hypothesis, the deformable material relative compressibility, uniform temperature distribution over the cross and longitudinal sections of the strip, etc. reduces the computation accuracy of strain and stress at the nodes. Standard programs often use such assumptions. It is acceptable for solving the routine problems. For the manufacture of critical elements such assumptions can lead to undesirable serious consequences. Modeling of the metal forming processes by using the finite element method gives a more accurate solution of the problem. However, the solutions require to solve a linear equations system of high dimensionality. The coefficient matrix is sparse, namely it consists of mainly zero elements, randomly scattered. This leads to considerable computation time for solution methods. In general, the number of arithmetic operations is proportional to the cube of the system's order. In this paper, we propose to increase the computational efficiency analysis through the use of the high-sparse matrix properties. In this case computer time will be proportional to the system order to the 1.2 power. The proposed solution SLAE comprises sparse matrix methods, which include techniques for optimal ordering of matrix rows and columns by using the criterion of the rows in order to increase the first non-zero elements number. This paper shows the application of the computation method for the process of the Belleville spring upset. Belleville springs can be classified as critical elements. One of its possible applications is the rail joint. Unlike conventional parts the elastic elements are characterized by the presence of large deformation under loads. The main factor that determines the service properties of the elastic element is the amount of strain energy that it can be accumulated. Accumulation and dispersing of the Belleville springs strain energy is directly depends to the relation of the meridian cross-section size and the spring working area. The establishment of this dependence is necessary for its rational design.*

## 1. INTRODUCTION

The metal stress-strain state computation is widely used in the metal forming processes simulation, as well as for the investigation of finished product fatigue strain. The study of stresses and strains arising during the various process stages, will improve the manufacturing process of metal products. Simulation identifies areas with critical stress. Then through the modernization of stamping tools and products specifications arising stresses will be reduced in these areas. This in turn will lead to increase the lifetime of the stamping equipment and manufactured metal products. To obtain the required results, numerical methods should be used.

Consideration of the historical development of numerical methods for the metal stress-strain state computation, shows the gradual improvement of these methods. The purpose of this is to improve the any task computations accuracy in any conditions. To achieve this, first of all it is necessary to provide a solution to the problem without the plane section hypothesis, the deformable material relative compressibility, uniform temperature distribution over the cross and longitudinal sections of the strip, etc. In plastic metal working modeling by using the finite element method the majority of simplifying assumptions is no need to use. In most FEM codes usage of the computer methods the stress-strain state is based on the variational computation, created by O.C. Zienkiewiczand R.C. Taylor[1], R. Hill [2]-[4], D.H. Norrieand G. deVries [5], D.Banabic [6], F. Barlat [7]-[12], J.L. Bassani [13], et al. However, the use of FEM codes for irregular shaped parts requires considerable computer resources. One possible solution to this problem is the 2D model prediscovery and then on the problem areas more accurate 3D simulation [14].

While there is great diversity in the formulations of the FEM, in general the method can be characterized by the following properties:

1. The task physical area is divided into sub-areas, or finite elements.
2. The dependent variable is approximated by the specialized functions on each finite element and, consequently, in the entire region. The parameters of these approximations in the subsequent are unknown the problem parameters.
3. Substitution approximations to defining equation gives the equations set system with unknown parameters. Solving these equations, we can determine the values of these parameters and, consequently, obtain an approximate the problem solution.

As a result of solving the SLAE for the elastic stress-strain state computation is obtained stress and strain in each finite element. However, the solutions require to solve a linear equations system of high dimensionality. The coefficient matrix is sparse, namely it consists of mainly zero elements, randomly scattered. This leads to considerable computation time for solution methods. In general, the number of arithmetic operations is proportional to the cube of the system's order. In this paper, we propose to increase the computational efficiency analysis through the use of the high-sparse matrix properties. This computer time will be proportional to the system order to the 1.2 power. The proposed solution SLAE comprises sparse matrix methods, which include techniques for optimal ordering of matrix rows and columns by using the criterion of the rows in order to increase the first non-zero elements number.

This paper shows the application of the numerical method for the process of the Belleville spring upset. One of its possible applications in the rail joint. Belleville springs are also commonly used in engineering for damping dynamic and shock loads encountered in operation of machines (anti-vibration damper and other buffers). So the Belleville springs can be classified as critical elements. However, according to numerous studies and practical testing of a significant portion of the Belleville springs during exploitation under repeated

loading does not stand above the claimant the amount of cycles ($2 \cdot 10^6$). Damage occurs due to development of fatigue cracks in the edge hole due to wear of their contact surfaces. Defects in the Belleville spring edges surface, appearing in the typical stamping and accelerates the fatigue cracks nucleation. In standard Belleville springs during operation there are significant the stresses value, and this their service life is reduced. Contact spring surface wears out quickly due to its small area. To increase the spring service life necessary to reduce the stresses during exploitation strain, to increase the contact surface area and eliminate the occurrence of surface defects during stamping. Reducing stress possible by optimizing the design parameters of the spring based on the stress-strain state analysis results of its compression during use [15].

## 2. APPLICATION OF THE NUMERICAL METHODS TO ANALYSIS THE STRESS-STRAIN STATE

### 2.1. Application of the finite element method to compute the stress-strain state problems

Consider the application of finite element method for solving the computer problem the Belleville spring elastic stress-strain state. The computer problem the spring elastic stress-strain state resolved on the basis of the minimum potential energy principle [5], [16]. The potential energy functional for solid mechanics bidimensional problems looks as follows:

$$\chi = \int_V \frac{1}{2} \varepsilon^T \sigma \, dV - \int_V U^T P \, dV - \int_{A_T} U^T T \, dA_T, \tag{1}$$

where $\varepsilon = \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \gamma_{xy} \end{bmatrix}$ – the strain vector, $\sigma = \begin{bmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{bmatrix}$ – the stress vector, $U^T$ – the transpose $U = \begin{bmatrix} u \\ v \end{bmatrix}$ of the movement at point matrix, $\mathrm{P} = \begin{bmatrix} \mathrm{P}_x \\ \mathrm{P}_y \end{bmatrix}$ – the mass force per unit volume matrix, $T = \begin{bmatrix} T_x \\ T_y \end{bmatrix}$ – surface stress at point matrix, $V$ – solid volume, $A_T$ –solid surface.

Values $\varepsilon_x$, $\varepsilon_y$ and $\sigma_x$, $\sigma_y$ are strains and stresses axial $x$ and $y$, respectively. In constructing a mathematical model of the elastic stress-strain state in a spring-loaded by the linearity of the elastic problem and small strain velocities do not take into account the temperature changes effects and the mass forces emergence [16]. These factors can be taken into account additionally if necessary.

Consequently:

$$\begin{aligned} F &= 0 \\ \delta T &= 0, \end{aligned} \tag{2}$$

where $F$ –mass force potential, $\delta T$ – solid temperature increment.

Due to the fact that the Belleville spring is in equilibrium, and consider strain state, the true system of solid material points displacement $u$ by the loads action $T$, as well as taking into account the condition (2), the potential energy functional (1) transform as follows:

$$\chi = \frac{1}{2} \int_V \varepsilon \sigma dV - \int_{A_T} u T dA_T, \tag{3}$$

where $V$ – solid volume, $T = \begin{bmatrix} T_x \\ T_y \end{bmatrix}$ – surface stress vector, $A_T$ – outer contact surface.

Since the Belleville spring has a rotation solid shape (Figure 1), then the decision we consider axisymmetrical stress-strain state.

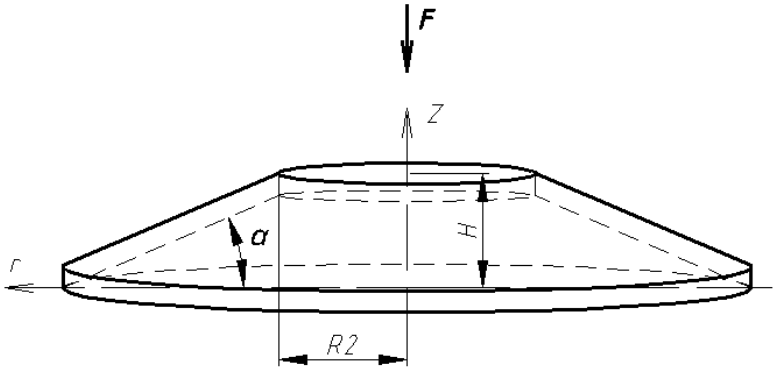



Figure 1: Parameters Belleville spring and the cylindrical system position.

After the conversion, taking into account characteristics of the Belleville spring parameters, the potential energy functional (3) for the spring will be:

$$\chi = \frac{1}{2}\int_0^{2\pi} d\varphi \int_{R_2}^{R_1}\int_{z_2}^{z_1} \varepsilon^T \cdot D \cdot \varepsilon \cdot r \cdot dr dz - k \cdot p_z \cdot \delta \int_0^{2\pi} R_2 \cdot u d\varphi \quad , \tag{4}$$

where $z_1 = H_1 - \frac{H_1}{R_1} r;$, $z_2 = H_2 - \frac{H_1}{R_1} r$, $H_1 = R_1 \cdot tg\alpha$, $H_2 = H_1 + \frac{S}{\cos\alpha}$, $H$ – the Belleville spring height, $k$ – friction coefficient, $R_1$ и $R_2$ – the Belleville spring radius and spring holes, respectively (Figure 1).

Then displacement to each of the finite element is approximated by linear functions. Then using the differentiation matrix rules, we compute the partial derivatives of the elemental energies along the nodal displacements vectors. The Belleville spring cross section break in the 48 triangular elements. Then, after securing the condition of potential energy functional minimum (4) and implement the united elemental matrix equations (association for nodes) in the matrix equation system with the defined movement. A result we get linear equations system for the nodal displacements values:

$$\sum_{e=1}^{48} K^e U^e - \frac{F \cdot S \cdot (R_1 - R_2)}{288 \cdot \pi \cdot R_2 \cdot \cos\alpha} = 0,$$

where $K^e = 120{,}859 \cdot S(R_1 - R_2) \cdot \begin{pmatrix} \kappa_{11} & \kappa_{12} & \kappa_{13} & \kappa_{14} & \kappa_{15} & \kappa_{16} \\ \kappa_{21} & \kappa_{22} & \kappa_{23} & \kappa_{24} & \kappa_{25} & \kappa_{26} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} & \kappa_{34} & \kappa_{35} & \kappa_{36} \\ \kappa_{41} & \kappa_{42} & \kappa_{43} & \kappa_{44} & \kappa_{45} & \kappa_{46} \\ \kappa_{51} & \kappa_{52} & \kappa_{53} & \kappa_{54} & \kappa_{55} & \kappa_{56} \\ \kappa_{61} & \kappa_{62} & \kappa_{63} & \kappa_{64} & \kappa_{65} & \kappa_{66} \end{pmatrix}$,

where $\kappa_{i1} = \begin{pmatrix} (z_2 - z_3)^2 + 0{,}35 \cdot (r_3 - r_2)^2 \\ 0{,}3 \cdot (r_3 - r_2)(z_2 - z_3) + 0{,}35 \cdot (z_2 - z_3)(r_3 - r_2) \\ (z_3 - z_1)(z_2 - z_3) + 0{,}35 \cdot (r_1 - r_3)(r_3 - r_2) \\ 0{,}3 \cdot (r_1 - r_3)(z_2 - z_3) + 0{,}35 \cdot (z_3 - z_1)(r_3 - r_2) \\ (z_1 - z_2)(z_2 - z_3) + 0{,}35 \cdot (r_2 - r_1)(r_3 - r_2) \\ 0{,}3 \cdot (r_2 - r_1)(z_2 - z_3) + 0{,}35 \cdot (z_1 - z_2)(r_3 - r_2) \end{pmatrix}$; $\kappa_{i2} = \begin{pmatrix} 0{,}3 \cdot (r_3 - r_2)(z_2 - z_3) + 0{,}35 \cdot (z_2 - z_3)(r_3 - r_2) \\ 0{,}35 \cdot (z_2 - z_3)^2 + (r_3 - r_2)^2 \\ 0{,}3 \cdot (z_3 - z_1)(r_3 - r_2) + 0{,}35 \cdot (r_1 - r_3)(z_2 - z_3) \\ (r_1 - r_3)(r_3 - r_2) + 0{,}35 \cdot (z_3 - z_1)(z_2 - z_3) \\ 0{,}3 \cdot (z_1 - z_2)(r_3 - r_2) + 0{,}35 \cdot (r_2 - r_1)(z_2 - z_3) \\ (r_2 - r_1)(r_3 - r_2) + 0{,}35 \cdot (z_1 - z_2)(z_2 - z_3) \end{pmatrix}$;

$\kappa_{i3} = \begin{pmatrix} (z_3 - z_1)(z_2 - z_3) + 0{,}35 \cdot (r_1 - r_3)(r_3 - r_2) \\ 0{,}3 \cdot (z_3 - z_1)(r_3 - r_2) + 0{,}35 \cdot (r_1 - r_3)(z_2 - z_3) \\ (z_3 - z_1)^2 + 0{,}35 \cdot (r_1 - r_3)^2 \\ 0{,}3 \cdot (r_1 - r_3)(z_3 - z_1) + 0{,}35 \cdot (z_3 - z_1)(r_1 - r_3) \\ (z_1 - z_2)(z_3 - z_1) + 0{,}35 \cdot (r_2 - r_1)(r_1 - r_3) \\ 0{,}3 \cdot (r_2 - r_1)(z_3 - z_1) + 0{,}35 \cdot (z_1 - z_2)(r_1 - r_3) \end{pmatrix}$; $\kappa_{i4} = \begin{pmatrix} 0{,}3 \cdot (r_1 - r_3)(z_2 - z_3) + 0{,}35 \cdot (z_3 - z_1)(r_3 - r_2) \\ (r_1 - r_3)(r_3 - r_2) + 0{,}35 \cdot (z_3 - z_1)(z_2 - z_3) \\ 0{,}3 \cdot (r_1 - r_3)(z_3 - z_1) + 0{,}35 \cdot (z_3 - z_1)(r_1 - r_3) \\ 0{,}35 \cdot (z_3 - z_1)^2 + (r_1 - r_3)^2 \\ 0{,}3 \cdot (z_1 - z_2)(r_1 - r_3) + 0{,}35 \cdot (r_2 - r_1)(z_3 - z_1) \\ (r_1 - r_3)(r_2 - r_1) + 0{,}35 \cdot (z_1 - z_2)(z_3 - z_1) \end{pmatrix}$

;

$\kappa_{i5} = \begin{pmatrix} (z_1 - z_2)(z_2 - z_3) + 0{,}35 \cdot (r_2 - r_1)(r_3 - r_2) \\ 0{,}3 \cdot (z_1 - z_2)(r_3 - r_2) + 0{,}35 \cdot (r_2 - r_1)(z_2 - z_3) \\ (z_1 - z_2)(z_3 - z_1) + 0{,}35 \cdot (r_2 - r_1)(r_1 - r_3) \\ 0{,}3 \cdot (z_1 - z_2)(r_1 - r_3) + 0{,}35 \cdot (r_2 - r_1)(z_3 - z_1) \\ (z_1 - z_2)^2 + 0{,}35 \cdot (r_2 - r_1)^2 \\ 0{,}3 \cdot (z_1 - z_2)(r_2 - r_1) + 0{,}35 \cdot (z_1 - z_2)(r_2 - r_1) \end{pmatrix}$; $\kappa_{i6} = \begin{pmatrix} 0{,}3 \cdot (r_2 - r_1)(z_2 - z_3) + 0{,}35 \cdot (z_1 - z_2)(r_3 - r_2) \\ (r_2 - r_1)(r_3 - r_2) + 0{,}35 \cdot (z_1 - z_2)(z_2 - z_3) \\ 0{,}3 \cdot (r_2 - r_1)(z_3 - z_1) + 0{,}35 \cdot (z_1 - z_2)(r_1 - r_3) \\ (r_2 - r_1)(r_1 - r_3) + 0{,}35 \cdot (z_1 - z_2)(z_3 - z_1) \\ 0{,}3 \cdot (z_1 - z_2)(r_2 - r_1) + 0{,}35 \cdot (z_1 - z_2)(r_2 - r_1) \\ 0{,}35 \cdot (z_1 - z_2)^2 + (r_2 - r_1)^2 \end{pmatrix}$.

Coefficient matrix of the linear equations system resulting contains a large number of zero elements (a sparse matrix). Therefore, the use of accurate computation methods (for example, the Gauss method) provides more computing error and has low computing speed. Since they depend on the positioning of columns and rows with null elements. This leads to high computations processing power. To improve the efficiency computations used special methods to transform a sparse coefficient matrix, include a method of arranging the rows and columns, reduce the sparse matrix filling, namely a modified Gauss method that provides a numerical method stability [17]. As a result of its solutions obtained nodal displacements, and then determine the stress and strain in each element [18].

## 2.2 Application of the sparse matrix method to compute the SLAE

Using iterative methods for solving linear systems with sparse matrices is not very effective, as to ensure their convergence necessary to satisfy the convergence conditions, in addition, if the problem solving linear systems is incorrect, then the iterative methods do not give a clear answer. Exact methods devoid of these shortcomings, however, their use gives greater computational errors and a higher dimension takes more memory [19]. The errors are caused by the transition from the mathematical model to the numerical method. They are connected with the fact that every numerical method reproduces the original mathematical model approximately. To improve the efficiency computations, there are special methods to transform a sparse coefficient matrix, including the optimal ways to organize the sparse matrix rows and columns.

The Gauss method - the successive unknowns elimination method from a equations system [20]. With the unknown $k_x$ exclusion, the equations system:

$$Ax = b, \tag{5}$$

where all coefficients $a_{ij}$ when $i > k$ and $j > k$ computed by the formula:

$$a_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \quad (6)$$

In general, the arithmetic operations numbers to Gaussian solve (5) be proportional to the system order to the 3 power. This leads to considerable cost of computer time. It is possible to improve significantly the computational efficiency of the analysis, if application the property of the high matrix sparseness *A* to model (5). Sparse matrix processing efficiency is high because it does not require, firstly, conversion of the formula (6), if at least one of the elements $a_{ik}$ or $a_{kj}$ is zero, secondly, the memory cost to store zero elements. Although sparse matrix algorithms are more complex, but as a result the computer time cost will be closer to linear. When using the sparse matrix methods need to take into account the dependence of the computational efficiency of the coefficient matrix submission *A*, more precisely, on the order in which recorded its rows and columns. Sparse matrix methods should include ways to the optimal streamline of rows and columns matrix. It is possible to use several ordering optimality criteria. The simplest of these is the criterion of the rows in order to increase the number of primary non-zero, more complex criteria take into account not only the primary non-zero, but also the emerging secondary non-zero.

To implement the method of sparse matrices it was developed software in the Delphi language. Input data the program are placed in a separate file containing non-zero values of the matrix coefficients in a packed form. The principle of reading the input data is as follows: at the beginning indicates the matrix order, each subsequent line is in the form kind of $a_1$, $a_2$, $a_3$ ..., where the first and the third number - a columns number, where the data is non-zero elements, and the second, the fourth number - data is non-zero elements. This type of the data presentation allows the large dimensions matrices can be stored and processed in the computer's memory. In this case gain in storage costs is significant. Using this software allows you to get a solution to a linear equations system with sparse coefficients matrices, to estimate the error and make a conclusion about the decision correctness [21].

### 2.3 Application of the method to analysis the stress-strain state in the spring-loaded

On the basis of the discussed above computer method the mathematical elastic stress-strain state model of the Belleville spring loaded was worked out. For the numerical implementation of the model it was developed software designed to compute strains and stresses in the spring-loaded. The program was developed in programming environment C# [18]. Figure 2 shows the stresses in the Belleville spring-loaded with design parameters: $S$ = 1 mm, $R1$ = 17.5 mm, $R2$ = 4.98 mm, $\alpha = 5.47^{\circ}$, $H$ = 1.2 mm, when a load is applied $F$ = 110 N.

Problem is solved by using a CAE system «ANSYS» to check the model verification (Figure 3).

On the basis of the stress-strain state analysis the computation method of disc springs optimal design parameters and form was worked out.

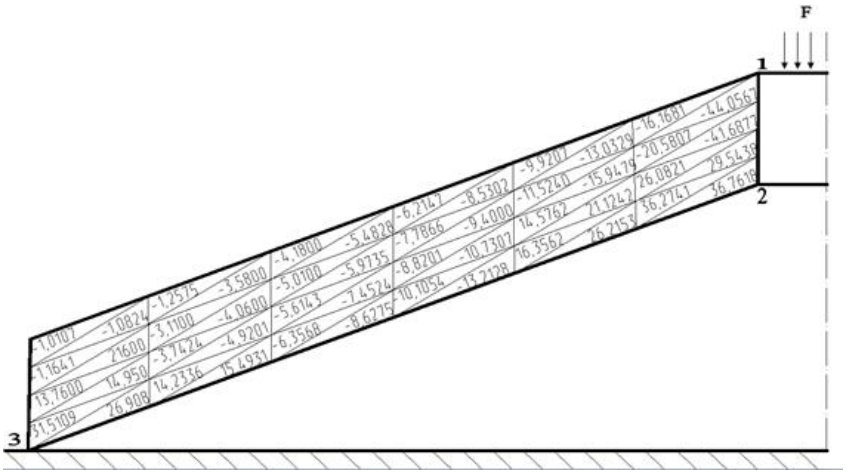



Figure 2: Stresses $\sigma_r$ in the Belleville spring-loaded.

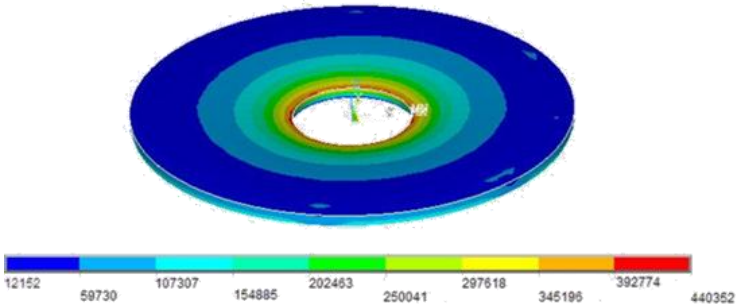



Figure 3: Stresses distribution field in a Belleville spring-loaded.

## 3. CONCLUSIONS

- A mathematical model of elastic springs compression during operation is developed.
- Sparse matrix method is used to improve computation performance.
- For the numerical implementation of the model it was developed software designed to compute strains and stresses in the spring-loaded.
- The stress-strain state of the spring-loaded is analyzed.
- On the basis of the stress-strain state analysis the computation method of optimal springs design parameters and form is worked out.

## REFERENCES

[1] O.C. Zienkiewicz, R.C. Taylor, *The finite element method, Vol. I, 4th Edition*. McGrawHill, 1989.

[2] R. Hill, *A theory of the yielding and plastic flow of anisotropic metals*. Proc. R. Soc. Lond. A193, 281–297, 1948.

[3] R. Hill, *Theoretical plasticity of textured aggregates*. Math.Proc. Cambr. Philos. Soc. 85, 179–191, 1979.

[4] R. Hill, *Constitutive modeling of orthotropic plasticity in sheet metals*. J. Mech. Phys. Solids 38, 405–417, 1990.

[5] D.H. Norrie, G. de Vries, *An Introduction to Finite Element Analysis*. Academic Press, London, 1978.

[6] D. Banabic, H. Aretz, D.S. Comsa, et al., *An improved analytical description of orthotropy in metallic sheets*. Int. J. Plast. 21, 493–512, 2005.

[7] F. Barlat, J. Lian, *Plastic behavior and stretchability of sheet metals. Part I: a yield function for orthotropic sheets under plane stress conditions.* Int. J. Plast. 5, 51–66, 1989.

[8] F. Barlat, D.J. Lege, J.C. Brem, *A six-component yield function for anisotropic materials.* Int. J. Plast. 7, 693–712, 1991.

[9] F. Barlat, Y. Maeda, K. Chung, et al., *Yield function development for aluminum alloy sheet.* J. Mech. Phys. Solids 45, 1727–1763, 1997.

[10] F. Barlat, J.C. Brem, J.W. Yoon, et al., *Plane stress yield function for aluminum alloy sheets. Part I: theory*. Int. J. Plast. 19, 1297–1319, 2003.

[11] F. Barlat, H. Aretz, J.W. Yoon, et al., *Linear transformation based anisotropic yield functions.* Int. J. Plast. 21, 1009–1039, 2005.

[12] F. Barlat, J.W. Yoon, O. Cazacu, *On linear transformation of stress tensors for the description of plastic anisotropy.* Int. J. Plast. 23, 876–896, 2007.

[13] J.L. Bassani, *Yield characterisation of metals with transversally isotropic plastic properties.* Int. J. Mech. Sci. 19, 651–660, 1977.

[14] J. Wang, J Sun, Plane strain transversely anisotropic analysis in sheet metal forming simulation using 6-component Barlat yield function. *International Journal of Mechanics and Materials in Design, Vol. 8, Issue 4*, 327-333, 2012.

[15] U.V. Mikhailova, *Improvement technology manufacturing special Belleville spring with elastic medium. Dissertation for the degree of technical sciences candidate*. Magnitogorsk, Nosov Magnitogorsk state technical university, 2011.

[16] O.S. Zhelezkov, S.V. Kochukov, U.V. Mikhailova, Modeling compress the Belleville springs by using finite element method. *Processes and equipment of metallurgical production.* Magnitogorsk, NMSTU, 239-243, 2006.

[17] U.V. Mikhailova, Application of finite element method for the metal stress-strain state computation. *Mathematics. Mathematics applications in the economic, technical and pedagogical research.* Magnitogorsk, NMSTU, 258–260, 2006.

[18] U.V. Mikhailova, *The elastic stress-strain state Belleville spring modeling using the finite element method.* The certificate of registration of the branch development № 8128. M., 2007. № 50200700_808.

[19] I.P. Norenkov, *Fundamentals of computer-aided design*. Moscow, Bauman MSTU, 2002.

[20] R. Tyuarson, *Sparse matrix*. Moscow, 1977.

[21] G.M. Korinchenko, U.V. Mikhailova, K.A. Kislitsin, Sparse matrix application for the design solution analysis design elements in mechanical engineering. The application of mathematics to economic and technical studies№ 1 (1), Magnitogorsk, NMSTU, 142-145, 2007.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FLUID ANALYSIS USING FICTITIOUS DOMAIN FINITE ELEMENT METHOD

**Y. Terakado[1], and T. Kurahashi[2]**

[1] Graduate school of Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata, Japan
e-mail: s123054@stn.nagaokaut.ac.jp

[2] Department of Mechanical Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata, Japan
kurahashi@mech.nagaokaut.ac.jp

**Keywords:** Fictitious domain method, Finite element method, Lagrange multiplier method, Incompressible viscose flow, Background domain, Foreground domain.

**Abstract.** *When the fluid analysis is carried out in the case that there is a moving object in the flow field, the mesh regeneration method is applied due to the Adaptive mesh refinement method. In addition, the Shear-Slip mesh update method is employed in case of rotating body problems. On the other hand, a method using the flow domain overlapped with the moving object is also proposed. In the finite difference method, a foreground grid is represented by moving object, and a main-grid is used for flow field calculation. The Overset-grid is applied to the fluid analysis to reflect the foreground grid to main-grid by the least-squares method. Furthermore, the Immersed boundary method based on the finite volume method is also employed. In the finite element analysis, the fictitious domain method is often adopted for the moving body problems. In this method, the computational domain is divided into foreground domain and back-ground domain. In addition, the formulation is carried out based on the Lagrange multiplier method, and is applied to consider the velocity condition in the foreground domain. The finite element fluid analysis is carried out to reflect the physical quantity of the foreground domain to back-ground domain. Physical quantities at any points of the foreground domain can be obtained from the physical quantities at each node by interpolation method. In this study, the flow field analysis using the fictitious domain finite element method is carried out.*

## 1 INTRODUCTION

When fluid analysis is carried out, the finite element method (FEM) is often employed, and is widely used as one of the most common analysis techniques. However, if the object moves in the flow field, it is necessary to re-generate the element mesh for the calculation each time, there is a problem that the computational cost increases. On the other hand, the fictitious domain method [1, 2] is employed as an analysis technique using two element areas within the object and the flow field in the background. Therefore, it is not necessary to re-generate the background mesh even if the object moves. This approach formulated based on the method of Lagrange multiplier [3, 4], representing the background and foreground meshes. Analysis of the flow field to perform a physical quantity of the foreground mesh is reflected to the background mesh by using the interpolation method. This study is to verify the numerical analysis of incompressible viscous fluid with a fictitious domain method.

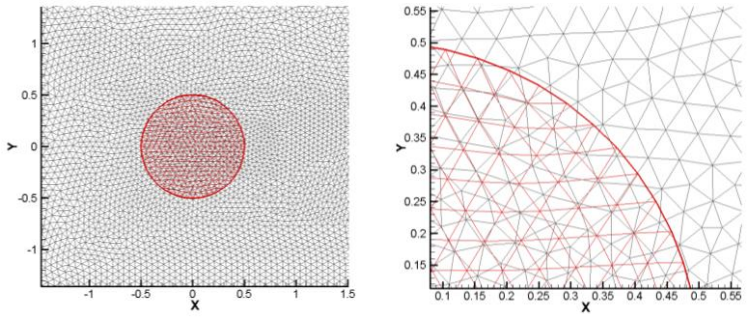



Figure 1: Example of finite element mesh in the fictitious domain method.

## 2 DISCRETIZATION OF THE GOVERNING EQUATIONS

### 2.1 Governing equation

The Navier-Stokes and the continuity equations are used to represent flow behavior and are written as Equations (1) and (2):

$$\dot{u}_i + u_j u_{i,j} - \frac{1}{\rho} P_{,i} - \nu\left(u_{i,j} + u_{j,i}\right)_{,j} = \frac{1}{\rho} f_i \tag{1}$$

$$u_{i,i} = 0 \tag{2}$$

where $V_i$, $P$, $Re$, and $f_i$ denote the flow velocity, pressure, Reynolds number, and body force per unit volume, respectively.

### 2.2 Discretization of the governing equation

As a method to solve the equation, the splitting method is applied. In this approach, it is possible to solve the equation by separating the pressure and velocity.

First, to get the solution of the pressure field, we derive the pressure Poisson equation. Performing discretization in the time direction using the Euler method to Equation (1), to give Equation (3).

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + u_j^n u_{i,j}^n + \frac{1}{\rho} P_{,i}^{n+1} - \nu\left(u_{i,j}^n + u_{j,i}^n\right)_{,j} = \frac{1}{\rho} f_i^n \tag{3}$$

Since the flow velocity is interpolated using triangular primary element, the term of the third order differentiation and differential term of constant value are vanished. The equation which was rewritten to Poisson equation of pressure is shown in Equation (4).

$$P_{,ii}^{n+1} = \rho\left( \frac{1}{\Delta t} u_{i,i}^{n} - \left( u_j^n u_{i,j}^n \right)_{,i} \right) \tag{4}$$

When it is assumed that the weighting function $P^*$ relates to a pressure $P$, the weighted residual equation is as shown in Equation (5).

$$\begin{aligned} &\int_{\Omega e} P_{,i}^{*} P_{,i}^{n+1} d\Omega = -\rho\left( \frac{1}{\Delta t} \int_{\Omega e} P^{*} u_{i,i}^{n} d\Omega + \bar{u}_j^{\,n} \int_{\Omega e} P_{,i}^{*} u_{i,j}^{n} d\Omega \right) + \rho \bar{u}_j^{\,n} \int_{\Gamma e} P^{*} u_{i,j}^{n} n_i d\Gamma \\ &+ \int_{\Gamma e} P^{*} P_i^{n+1} n_i d\Gamma \end{aligned} \tag{5}$$

Next, if the Crank-Nicolson method is applied to the momentum equation, Equation (6) is obtained.

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + u_j^n u_{i,j}^{n+\frac{1}{2}} + \frac{1}{\rho} P_{,i}^{n+1} - \nu\left( u_{i,j}^{n+\frac{1}{2}} + u_{j,i}^{n+\frac{1}{2}} \right)_{,j} = \frac{1}{\rho} f_i^{\,n} \tag{6}$$

Multiplying the weighting function $u_i^*$ on the flow velocity for both sides of Equation (6), integrating over the domain $\Omega_e$, and introducing the term of the Lagrange multiplier $\lambda_i$ [3, 4], Equation (7) is consequently obtained.

$$\begin{aligned} &\int_{\Omega e} u_i^{*} \left( \left( u_i^{n+1} - u_i^n \right) + \Delta t u_j^n u_{i,j}^{n+\frac{1}{2}} + \frac{\Delta t}{\rho} P_{,i}^{n+1} - \Delta t \nu \left( u_{i,j}^{n+\frac{1}{2}} + u_{j,i}^{n+\frac{1}{2}} \right)_{,j} \right) d\Omega + \frac{1}{\rho} \int_{\omega w} u_i^{*} \lambda_i^{n+1} d\omega \\ &= \int_{\Omega e} u_i^{*} \frac{\Delta t}{\rho} f_i^{\,n} d\Omega \end{aligned} \tag{7}$$

In addition, by introducing an interpolation function by the triangular and the bubble function elements, Equations (8) and (9) are finally obtained.

$$\left( \mathbf{C}_{xx} + \mathbf{C}_{yy} \right) \mathbf{p}^{n+1} = -\left( \frac{1}{\Delta t} \left( \mathbf{S}_x \mathbf{v}_x^n + \mathbf{S}_y \mathbf{v}_y^n \right) + \left( \mathbf{D}_{xx} + \mathbf{D}_{xy} \right) \mathbf{v}_x^n + \left( \mathbf{D}_{yx} + \mathbf{D}_{yy} \right) \mathbf{v}_y^n \right) + \mathbf{b}^n \tag{8}$$

$$
\begin{bmatrix} \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \\ \mathbf{r}_x \\ \mathbf{r}_y \end{Bmatrix}^{(n+1)} + \begin{bmatrix} \frac{\Delta t}{2}\mathbf{H}_{xx} & \frac{\Delta t}{2}\mathbf{H}_{xy} & \mathbf{B}_x^T & \mathbf{0} \\ \frac{\Delta t}{2}\mathbf{H}_{yx} & \frac{\Delta t}{2}\mathbf{H}_{yy} & \mathbf{0} & \mathbf{B}_y^T \\ \mathbf{B}_x & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_y & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \\ \mathbf{r}_x \\ \mathbf{r}_y \end{Bmatrix}^{(n+1)}
$$

$$
= \begin{bmatrix} \mathbf{M} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \\ \mathbf{r}_x \\ \mathbf{r}_y \end{Bmatrix}^{(n)} - \begin{bmatrix} \frac{\Delta t}{2}\mathbf{H}_{xx} & \frac{\Delta t}{2}\mathbf{H}_{xy} & \mathbf{0} & \mathbf{0} \\ \frac{\Delta t}{2}\mathbf{H}_{yx} & \frac{\Delta t}{2}\mathbf{H}_{yy} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{v}_x \\ \mathbf{v}_y \\ \mathbf{r}_x \\ \mathbf{r}_y \end{Bmatrix}^{(n)} \tag{9}
$$

$$
+ \begin{Bmatrix} \Delta t\mathbf{A}_x\mathbf{p} \\ \Delta t\mathbf{A}_y\mathbf{p} \\ \mathbf{0} \\ \mathbf{0} \end{Bmatrix}^{(n+1)} + \begin{Bmatrix} \Delta t\mathbf{t}_x \\ \Delta t\mathbf{t}_y \\ \mathbf{f}_x \\ \mathbf{f}_y \end{Bmatrix}^{(n)}
$$

The matrices and vectors in the finite element equation for the pressure Poisson equation and the momentum equation are shown in Equations (10) and (11), respectively.

$$
\begin{aligned}
&\mathbf{C}_{xx} = \int_{\Omega e} \mathbf{N}_{,x}\mathbf{N}_{,x}^T d\Omega, \quad \mathbf{C}_{yy} = \int_{\Omega e} \mathbf{N}_{,y}\mathbf{N}_{,y}^T d\Omega \\
&\mathbf{S}_x = \int_{\Omega e} \mathbf{N}\mathbf{N}_{B,x}^T d\Omega, \quad \mathbf{S}_y = \int_{\Omega e} \mathbf{N}\mathbf{N}_{B,y}^T d\Omega \\
&\mathbf{D}_{xx} = \int_{\Omega e} \mathbf{N}_{,x}\left(\mathbf{N}_B^T\mathbf{v}_x\right)\mathbf{N}_{B,x}^T d\Omega, \quad \mathbf{D}_{xy} = \int_{\Omega e} \mathbf{N}_{,x}\left(\mathbf{N}_B^T\mathbf{v}_y\right)\mathbf{N}_{B,y}^T d\Omega \\
&\mathbf{D}_{yx} = \int_{\Omega e} \mathbf{N}_{,y}\left(\mathbf{N}_B^T\mathbf{v}_x\right)\mathbf{N}_{B,x}^T d\Omega, \quad \mathbf{D}_{yy} = \int_{\Omega e} \mathbf{N}_{,y}\left(\mathbf{N}_B^T\mathbf{v}_y\right)\mathbf{N}_{B,y}^T d\Omega \\
&\mathbf{b} = \int_{\Gamma e} \left(\left(V_x^n V_{x,x}^n + V_y^n V_{x,y}^n + P_{,x}^{n+1}\right)n_x + \left(V_y^n V_{x,y}^n + V_y^n V_{y,y}^n + P_{,y}^{n+1}\right)n_y\right)\mathbf{N}d\Gamma \\
&\mathbf{p} = \begin{Bmatrix} P_1 & P_2 & P_3 \end{Bmatrix}^T \\
&\mathbf{v}_x = \begin{Bmatrix} V_{x1} & V_{x2} & V_{x3} & \tilde{V}_{x4} \end{Bmatrix}^T, \quad \mathbf{v}_y = \begin{Bmatrix} V_{y1} & V_{y2} & V_{y3} & \tilde{V}_{y4} \end{Bmatrix}^T
\end{aligned} \tag{10}
$$

$$
\begin{aligned}
&\mathbf{M} = \int_{\Omega e} \mathbf{N}_B \mathbf{N}_B^T d\Omega \\
&\mathbf{H}_{xx} = \int_{\Omega e} \mathbf{N}_B \left(\mathbf{N}_B^T \mathbf{v}_x\right) \mathbf{N}_{B,x}^T d\Omega + \int_{\Omega e} \mathbf{N}_B \left(\mathbf{N}_B^T \mathbf{v}_y\right) \mathbf{N}_{B,y}^T d\Omega \\
&+ \tfrac{1}{Re}\left(2\int_{\Omega e} \mathbf{N}_{B,x} \mathbf{N}_{B,x}^T d\Omega + \int_{\Omega e} \mathbf{N}_{B,y} \mathbf{N}_{B,y}^T d\Omega\right) \\
&\mathbf{H}_{xy} = \tfrac{1}{Re} \int_{\Omega e} \mathbf{N}_{B,y} \mathbf{N}_{B,x}^T d\Omega \\
&\mathbf{H}_{yx} = \tfrac{1}{Re} \int_{\Omega e} \mathbf{N}_{B,x} \mathbf{N}_{B,y}^T d\Omega \\
&\mathbf{H}_{yy} = \int_{\Omega e} \mathbf{N}_B \left(\mathbf{N}_B^T \mathbf{v}_x\right) \mathbf{N}_{B,x}^T d\Omega + \int_{\Omega e} \mathbf{N}_B \left(\mathbf{N}_B^T \mathbf{v}_y\right) \mathbf{N}_{B,y}^T d\Omega \\
&+ \tfrac{1}{Re}\left(\int_{\Omega e} \mathbf{N}_{B,x} \mathbf{N}_{B,x}^T d\Omega + 2\int_{\Omega e} \mathbf{N}_{B,y} \mathbf{N}_{B,y}^T d\Omega\right) \\
&\mathbf{A}_x = \int_{\Omega e} \mathbf{N}_{B,x} \mathbf{N}^T d\Omega, \quad \mathbf{A}_y = \int_{\Omega e} \mathbf{N}_{B,y} \mathbf{N}^T d\Omega \\
&\mathbf{B}_x = \begin{Bmatrix} N_1(\mathbf{a}) & N_2(\mathbf{a}) & N_3(\mathbf{a}) & N_4(\mathbf{a}) \end{Bmatrix} \\
&\mathbf{B}_y = \begin{Bmatrix} N_1(\mathbf{a}) & N_2(\mathbf{a}) & N_3(\mathbf{a}) & N_4(\mathbf{a}) \end{Bmatrix} \\
&\mathbf{p} = \begin{Bmatrix} P_1 & P_2 & P_3 \end{Bmatrix}^T \\
&\mathbf{v}_x = \begin{Bmatrix} V_{x1} & V_{x2} & V_{x3} & \tilde{V}_{x4} \end{Bmatrix}^T, \quad \mathbf{v}_y = \begin{Bmatrix} V_{y1} & V_{y2} & V_{y3} & \tilde{V}_{y4} \end{Bmatrix}^T \\
&\mathbf{r}_x = \lambda_x(\mathbf{a}), \quad \mathbf{r}_y = \lambda_y(\mathbf{a}) \\
&\mathbf{t}_x = \int_{\Gamma e} \mathbf{T}_x \mathbf{N} d\Gamma, \quad \mathbf{t}_y = \int_{\Gamma e} \mathbf{T}_y \mathbf{N} d\Gamma \\
&\mathbf{f}_x = V_{x\alpha}(\mathbf{a}), \quad \mathbf{f}_y = V_{y\alpha}(\mathbf{a})
\end{aligned}
\tag{11}
$$

## 3 DETECTION OF THE FOREGROUND NODES NODE

We consider the conditions under which the nodes of the foreground is present in the elements of the background area. From the definition of the cross product, the node determines whether inside or outside. It is known that determinants $D_1$, $D_2$ and $D_3$ expressed by Equation (12) must be positive value, if there is node as shown in Figure 2, $a = (x_a, y_a)$, of foreground in an element consists of three nodes, $(x_1, y_1)$, $(x_2, y_2)$, and $(x_3, y_3)$.

$$
\begin{cases}
D_1 = \begin{vmatrix} x_1 - x_a & y_1 - y_a \\ x_2 - x_a & y_2 - y_a \end{vmatrix} \\
D_2 = \begin{vmatrix} x_2 - x_a & y_2 - y_a \\ x_3 - x_a & y_3 - y_a \end{vmatrix} \\
D_3 = \begin{vmatrix} x_3 - x_a & y_3 - y_a \\ x_1 - x_a & y_1 - y_a \end{vmatrix}
\end{cases}
\tag{12}
$$

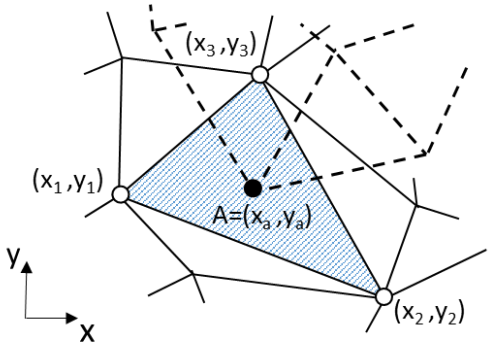



Figure 2: Example of overlapped domain.

## 4 NUMERICAL EXPERIMENTS

### 4.1 Comparison of the conventional FEM and the fictitious domain FEM

In the flow field around the cylinder, we compered the analysis using the conventional FEM and the analysis using the fictitious domain FEM. Computational model is shown in Figure 3. In addition, the computational condition for conventional FEM and fictitious domain FEM is shown in Table 1.

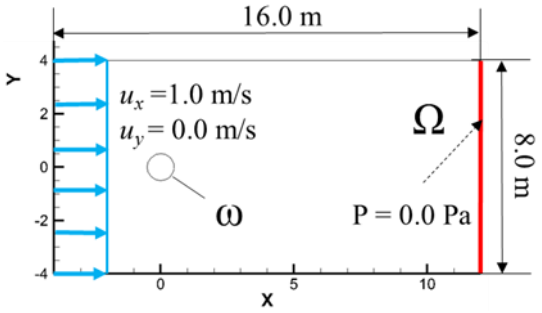



Figure 3: Computational model.

| | Conventional FEM | Fictitious domain FEM |
|---|---|---|
| Density | 1.00 kg/m$^3$ | 1.00 kg/m$^3$ |
| Kinematic viscosity | 0.01 m$^2$/s | 0.01 m$^2$/s |
| Time increment Δt | 0.001 | 0.001 |
| Nodes in background domain | 1636 | 1907 |
| Elements in background domain | 3116 | 3716 |
| Nodes in background domain | - | 331 |
| Elements in foreground domain | - | 600 |

Table 1: Computational conditions.

The analysis results are shown in Figures 4-6. Figures 4 and 5 are shown the pressure distribution and the flow velocity vector at $T$ = 10 [s] in the conventional FEM and the fictitious domain FEM. There are no elements inside the moving object in case of the conventional method. On the other hand, in case of the fictitious domain FEM, the physical quantity of the foreground domain is reflected in the background. Therefore, it is seem that the flow velocity vector obtained by the fictitious domain FEM is distributed around the cylinder as same as that obtained by the conventional method. Figure 6 shows the comparison of the time history of the drag force between the conventional FEM and the fictitious domain FEM. From this result, it is found that the result by the fictitious domain FEM is good agreement with that by the conventional FEM.

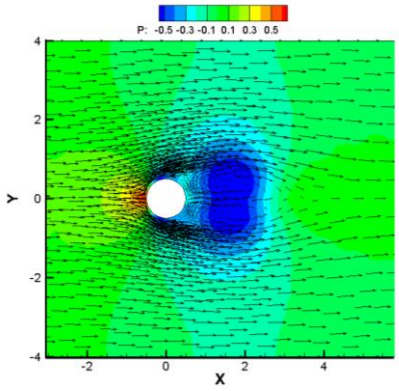



Figure 4: Pressure and velocity distributions in case of conventional FEM.

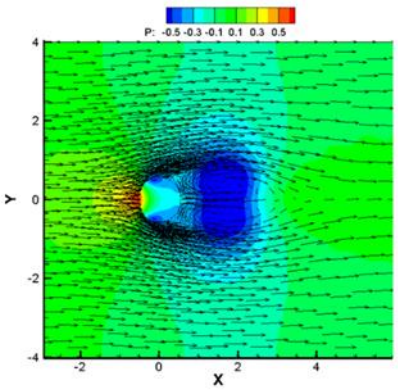



Figure 5: Pressure and velocity distributions in case of fictitious domain FEM.

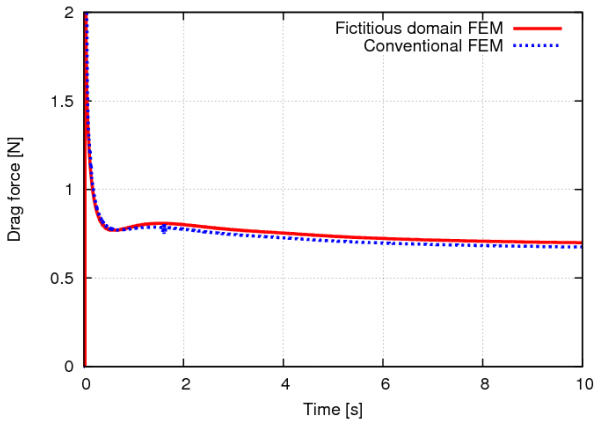



Figure 6: Comparison of time history of drag force between conventional FEM and the fictitious domain FEM.

## 4.2 Comparison of the fluid force for the case of changing the shape of the foreground domain

It is necessary to move the nodes freely in the computational domain, when the appropriate shape is obtained such that the drag force minimizes. In this study, the shape expressed by the nodes of the foreground mesh is changed by the solution of the Laplace equation shown in Equation (13) under the boundary condition Equation (14), the comparison of the drag force is carried out in each shape case.

$$\begin{cases} \dfrac{\partial^2 \Delta x}{\partial x^2} + \dfrac{\partial^2 \Delta x}{\partial y^2} = 0 \\ \dfrac{\partial^2 \Delta y}{\partial x^2} + \dfrac{\partial^2 \Delta y}{\partial y^2} = 0 \end{cases} \tag{13}$$

$$\begin{cases} x_i^{new} = \hat{x}_i \\ y_i^{new} = \hat{y}_i \end{cases} \quad \text{on } \partial\omega \tag{14}$$

Figure 7 shows the comparison of the foreground meshes in case of the cylinder and the ellipse. Figure 8 is the comparison of time history of drag force between cylinder and ellipses. The drag force decreases gradually with decreasing ellipticity (minor axis radius / major axis radius). It is found that when the drag force in case of cylinder and ellipses at $T$ = 10 [s], the drag force of ellipse 1 and ellipse 2 decreases 18% and 30%, respectively, in comparison with that in case of circular cylinder.

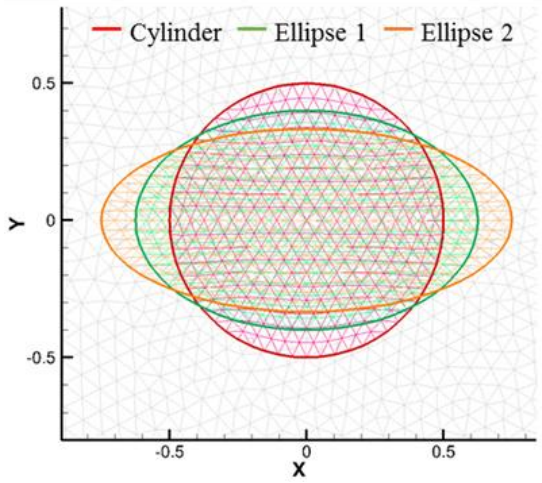



Figure 7: Foreground meshes in each examination case.

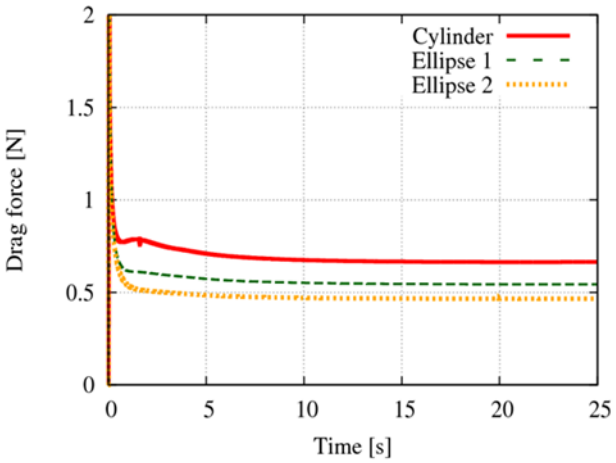



Figure 8: Comparison of time history of drag force between cylinder and ellipses.

## 5 CONCLUSIONS

The flow analysis around the body based on the fictitious domain FEM is shown in this study. Navier–Stokes equations and equation of continuity are employed for the governing equation. The FEM using the mixed interpolation by the triangular and the bubble function elements is used to discretize the governing equation for space direction, and the Crank-Nicolson method are used to discretize the governing equation for time direction. The conclusions are shown as follows.

- Similar calculation results were obtained in the conventional FEM and the fictitious domain FEM from the result of drag force distributions. It means that the present method is suitable in this study.
- It was found that even if the shape of foreground mesh was changed, the fluid force can be accurately obtained by using the fictitious domain FEM.

## REFERENCES

[1] R. Glowinski, T.W. Pan, Numerical simulation and oprimal shape for viscous flow by a fictitious domain method, *International Journal for Numerical Methods in Fluids*, vol.20:695-711, 1995.

[2] R. Glowinski, T.W. Pan, T.I. Hesla , D.D. Joseph, and J. Periaux, A Fictitious Domain Method with Distributed Lagrange Multipliers for the Numerical Simulation of Particulate Flow, *Contemporary Mathematics*, vol.218, 121-137, 1998.

[3] M. Nagai, and M. Kawahara, A Fictitious Domain Method with Distributed Lagrange Multiplier for Particulate Flow, *International Journal for Computational Methods in Engineering Science and Mechanics*, vol.8, 115-122, 2007.

[4] C. Borgers, Domain embedding method for the Stokes equations，*Numerische Mathematik*, vol.57 (1), 435-451, 1990.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE SIMULATION OF AERODYNAMIC NOISE WITH DIFFERENT TURBULENCE MODELS

**X. Huang**[1,2], **M. Schäfer**[1,2]

[1]Graduate School Computational Engineering
Technische Universität Darmstadt
Dolivostrasse 15, 64293 Darmstadt, Germany
{huang, schaefer}@gsc.tu-darmstadt.de

[2] Institute of Numerical Methods in Mechanical Engineering
Technische Universität Darmstadt
Dolivostrasse 15, 64293 Darmstadt, Germany
{huang, schaefer}@fnb.tu-darmstadt.de

**Keywords:** Aeroacoustics, Turbulence Model, LEE.

**Abstract.** *This paper investigates the performance of different turbulence models for simulating the aeroacoustic field generated by turbulent flow. The acoustic field is calculated by solving the linearized Euler equations (LEE), while the acoustic source is obtained by solving the incompressible Navier-Stokes equations. Three different turbulence models are adopted for the flow simulation: the large eddy simulation (LES), the delayed detached-eddy simulation (DDES) and unsteady Reynolds Averaged Navier-Stokes equations (URANS). The acoustic sources are calculated on a flow grid and then transferred and interpolated onto an acoustic grid, enabling the acoustic computation of the far field with a reasonable computational cost. We present the acoustic sources calculated using different turbulence models. Furthermore, we compare the sound pressure level (SPL) spectra of different turbulence models with the experimental data.*

# 1 INTRODUCTION

In recent decades, the simulation of aerodynamic noise has become increasingly significant for engineering applications. The prediction of aeroacoustic fields can be realized using the direct noise computation (DNC) [1], which solves the compressible Navier-Stokes equations directly for flow and acoustics. However, due to the significant difference at both length and time scales of flow and acoustic fields, the DNC method is considered inefficient especially for low Mach number flows. A hybrid scheme is commonly adopted, in which the simulations of flow and acoustics are separated and coupled through the acoustic sources.
Regarding the simulation of acoustics, several integral formulations, such as Lighthill's analogy [8, 9] and Curle's analogy [2] have been successfully applied. An alternative is the linearized Euler equations (LEE) method, which is obtained by using the Expansion about Incompressible Flow (EIF) technique [12, 13].
In order to acquire the acoustic sources, the incompressible Navier-Stokes equations are utilized. For turbulent flows, direct numerical simulation (DNS) in most cases cannot be applied due to its unaffordable computational cost. A turbulence model is usually adopted to characterize the unresolved turbulence scales, leading to a significant reduction in computational cost. Large eddy simulation (LES) resolves the large eddies and implicitly accounts for the small eddies by using a sub-grid scale model. Nevertheless, the computational cost of LES for engineering problems is mostly beyond the current computational power. Reynolds Averaged Navier-Stokes (RANS) equations only resolve the time averaged flow, leading to a drastic reduction of computational cost. However, the lack of accuracy makes it inappropriate for aeroacoustic problems. By adding an additional unsteady term to the RANS equations, one obtains the unsteady RANS (URANS) model. The idea of hybrid methods is to combine the advantages of LES model and RANS models. Detached-eddy simulation (DES) is one of the most popular hybrid turbulence models. The DES model switches between LES and RANS modes according to the numerical resolution [15, 16].
In the present work, the results of three turbulence models are compared with the experimental data to determine their suitability in the prediction of aerodynamic noise.

# 2 GOVERNING EQUATIONS

## 2.1 Linearized Euler equations for aeroacoustic simulation

The acoustic field is described through three quantities: the acoustic pressure $p^{\mathrm{ac}}$, the acoustic density fluctuation $\rho^{\mathrm{ac}}$ and the particle velocity $u_i^{\mathrm{ac}}$. These are the perturbation of pressure, density and velocity caused by sound waves, respectively. Since the humans' sensation of sound is proportional to the logarithm of the acoustic pressure, the acoustic pressure is rarely used in practice. The sound pressure level (SPL) $L_p$ with unit decibel (dB) is commonly adopted, which is defined by

$$L_p = 10 \cdot \lg \frac{\left(p_{\mathit{eff}}^{\mathrm{ac}}\right)^2}{p_{\mathit{ref}}^2} = 20 \cdot \lg \frac{p_{\mathit{eff}}^{\mathrm{ac}}}{p_{\mathit{ref}}}, \tag{1}$$

where $p_{\mathit{ref}}$ is a reference acoustic pressure and $p_{\mathit{eff}}^{\mathrm{ac}}$ is the effective acoustic pressure given by

$$p_{\mathit{eff}}^{\mathrm{ac}} = \sqrt{\frac{1}{T}\int_{t=0}^{T} (p^{\mathrm{ac}}(t))^2 \,\mathrm{d}t}. \tag{2}$$

In order to derive the governing equations for the acoustic quantities, a decomposition technique called Expansion about Incompressible Flow method (EIF) is utilized. This technique is

proposed by Hardin, Pope [6] and improved by Shen and Sørensen [12, 13]. It assumes that the compressible flow field at low Mach number can be decomposed into an incompressible flow field and an acoustic field

$$u_i = u_i^{\mathrm{inc}} + u_i^{\mathrm{ac}}, \tag{3}$$

$$p_i = p_i^{\mathrm{inc}} + p_i^{\mathrm{ac}}, \tag{4}$$

$$\rho_i = \rho_i^{\mathrm{inc}} + \rho_i^{\mathrm{ac}}, \tag{5}$$

where $u_i$, $p_i$ and $\rho_i$ are the velocity, pressure and density of compressible flow and superscripts inc and ac represent the components of incompressible flow and acoustic field respectively. Based on the EIF decomposition, one can obtain the linearized Euler equations (LEE) as follows

$$\frac{\partial \rho^{\mathrm{ac}}}{\partial t} + \rho^{\mathrm{inc}} \frac{\partial u_i^{\mathrm{ac}}}{\partial x_i} + u_i^{\mathrm{inc}} \frac{\partial \rho^{\mathrm{ac}}}{\partial x_i} = 0, \tag{6}$$

$$\rho^{\mathrm{inc}} \frac{\partial u_i^{\mathrm{ac}}}{\partial t} + \rho^{\mathrm{inc}} u_j^{\mathrm{inc}} \frac{\partial u_i^{\mathrm{ac}}}{\partial x_j} + \frac{\partial p^{\mathrm{ac}}}{\partial x_i} = 0, \tag{7}$$

$$\frac{\partial p^{\mathrm{ac}}}{\partial t} + c^2 \rho^{\mathrm{inc}} \frac{\partial u_i^{\mathrm{ac}}}{\partial x_i} + c^2 u_i^{\mathrm{inc}} \frac{\partial \rho^{\mathrm{ac}}}{\partial x_i} = -\frac{\partial p^{\mathrm{inc}}}{\partial t}. \tag{8}$$

For more details regarding the derivation of the above equations, the reader is referred to [7].

### 2.2 Turbulent flow

For turbulent flows, the governing equations read

$$\frac{\partial \overline{u_i^{\mathrm{inc}}}}{\partial x_i} = 0, \tag{9}$$

$$\frac{\partial \overline{u_i^{\mathrm{inc}}}}{\partial t} + \overline{u_j^{\mathrm{inc}}} \frac{\partial \overline{u_i^{\mathrm{inc}}}}{\partial x_j} = \frac{\partial}{\partial x_j} \left( \nu \frac{\partial \overline{u_i^{\mathrm{inc}}}}{\partial x_j} - \tau_{ij} \right) - \frac{1}{\rho^{\mathrm{inc}}} \frac{\partial \overline{p^{\mathrm{inc}}}}{\partial x_i}. \tag{10}$$

For URANS models, $\overline{u_i^{\mathrm{inc}}}$ is the mean velocity, while for LES models $\overline{u_i^{\mathrm{inc}}}$ is the filtered velocity. This similarity of the equations of URANS and LES facilitates the development of hybrid RANS/LES models [4].

**$k-\omega$ SST model** For URANS models, $\tau_{ij}$ in (10) is the Reynold's stress, which is usually calculated using the Boussinesq approximation

$$\tau_{ij} = \nu_{\mathrm{t}} \left( \frac{\partial \overline{u_i^{\mathrm{inc}}}}{\partial x_j} + \frac{\partial \overline{u_j^{\mathrm{inc}}}}{\partial x_i} \right) - \frac{2}{3} k \delta_{ij}, \tag{11}$$

where $\nu_{\mathrm{t}}$ is the turbulent viscosity that has to be modeled. The $k-\omega$ SST model is a URANS model applying the $k-\omega$ model in the near wall region and the $k-\varepsilon$ model in the free stream region. Therefore, it combines the advantages of both basic URANS models [10]. The $k-\omega$ SST model is formulated as

$$\frac{\partial k}{\partial t} + \overline{u_j^{\mathrm{inc}}} \frac{\partial k}{\partial x_j} = P_k - \beta^* k\omega + \frac{\partial}{\partial x_j} \left[ (\nu + \sigma_k \nu_t) \frac{\partial k}{\partial x_j} \right], \tag{12}$$

$$\frac{\partial \omega}{\partial t} + \overline{u_j^{\mathrm{inc}}} \frac{\partial \omega}{\partial x_j} = \frac{\gamma}{\nu_t} P_k - \beta \omega^2 + \frac{\partial}{\partial x_j} \left[ (\nu + \sigma_\omega \nu_t) \frac{\partial \omega}{\partial x_j} \right] + 2 (1 - F_1) \frac{\sigma_{\omega 2}}{\omega} \frac{\partial k}{\partial x_j} \frac{\partial \omega}{\partial x_j}, \tag{13}$$

where $k$ is the turbulent kinetic energy, $\omega$ is a specific dissipation, $P_k$ is the production term given as

$$P_k = \tau_{ij} \frac{\partial \overline{u_i^{\text{inc}}}}{\partial x_j}. \tag{14}$$

The turbulent viscosity is calculated using

$$\nu_t = \frac{a_1 k}{\max(a_1 \omega, \Omega F_2)}, \tag{15}$$

where $\Omega$ is the vorticity magnitude

$$\Omega = \sqrt{2\Omega_{ij}\Omega_{ij}}, \quad \Omega_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u_i^{\text{inc}}}}{\partial x_j} - \frac{\partial \overline{u_j^{\text{inc}}}}{\partial x_i}\right). \tag{16}$$

**$\zeta - f$ based DDES model** The $\zeta - f$ model was described in [5] and the equations are given as

$$\frac{\partial k}{\partial t} + \overline{u_j^{\text{inc}}} \frac{\partial k}{\partial x_j} = P_k + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_k}\right)\frac{\partial k}{\partial x_j}\right] - \varepsilon, \tag{17}$$

$$\frac{\partial \varepsilon}{\partial t} + \overline{u_j^{\text{inc}}} \frac{\partial \varepsilon}{\partial x_j} = \frac{C_{\varepsilon 1} P_k - C_{\varepsilon 2}\varepsilon}{\tau} + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_k}\right)\frac{\partial \varepsilon}{\partial x_j}\right], \tag{18}$$

$$\frac{\partial \zeta}{\partial t} + \overline{u_j^{\text{inc}}} \frac{\partial \zeta}{\partial x_j} = f - \frac{\zeta}{k} P_k + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_k}\right)\frac{\partial \zeta}{\partial x_j}\right], \tag{19}$$

$$L^2 \nabla f - f = \frac{1}{\tau}\left(C_1 + C_2' \frac{P_k}{\varepsilon}\right)\left(\zeta - \frac{2}{3}\right), \tag{20}$$

where $\varepsilon$ is the dissipation, $f$ is the elliptic relaxation term, $L$ is the length scale and $\tau$ is the time scale of turbulence. A detailed description of the equations and the value of the coefficients can be found in [5].

The delayed detached-eddy simulation (DDES) model modifies URANS models to achieve an LES content, enabling the DDES model to switch between URANS and LES according to the numerical resolution [16]. For the $\zeta - f$ model, the $\varepsilon$ in the $k$ equation is modified to

$$\varepsilon = \frac{k^{3/2}}{l_{\text{turb}}}, \tag{21}$$

where $l_{\text{turb}}$ is the length scale of DDES

$$l_{\text{turb}} = l_{\text{RANS}} - f_d \max(0, d - C_{\text{DES}}\Delta\phi). \tag{22}$$

The turbulent viscosity is then calculated using

$$\nu_t = C_\mu \zeta k \tau. \tag{23}$$

**Large eddy simulation** For LES, $\tau_{ij}$ in (10) is called the sub-grid scale stress, which is commonly modeled using the Smagorinsky model [14]

$$\tau_{ij}^{\mathrm{Smag}} - \frac{1}{3}\tau_{kk}\delta_{ij} = -2\nu_t\overline{S_{ij}}, \quad \nu_t = (C_s\Delta)^2|\overline{S}|, \tag{24}$$

where $|\overline{S}|$ is the magnitude of the strain rate tensor, which is represented by

$$\overline{S_{ij}} = \frac{1}{2}\left(\frac{\partial\overline{u_i^{\mathrm{inc}}}}{\partial x_j} + \frac{\partial\overline{u_j^{\mathrm{inc}}}}{\partial x_i}\right). \tag{25}$$

## 3 NUMERICAL REALIZATION

This research is conducted within an in-house code FASTEST, which adopts a fully conservative finite-volume method to solve the incompressible NSE on a block-structured grid. The work flow of the aeroacoustic simulation is shown in Fig.1.

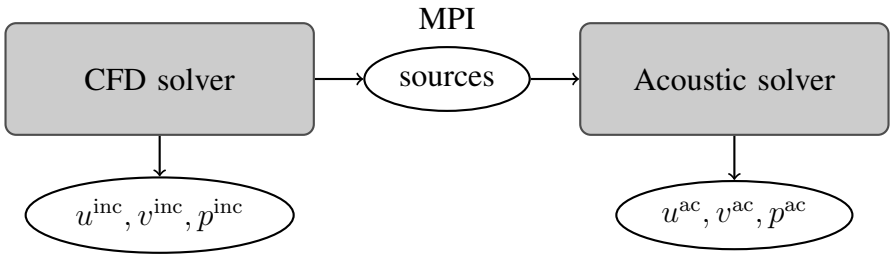



Figure 1: Numerical realization of flow solver and acoustic solver.

First, we use the different turbulence models to calculate the acoustic sources on a flow grid, which has a fine resolution. Second, we transmit and interpolate the acoustic sources onto an acoustic grid, which has a relatively coarse resolution. For the coupling of the two grids, trilinear interpolation is used owing to the simplicity of implementation. Finally, we compute the acoustic quantities by solving the LEE.
The linearized Euler equations are solved using a high-resolution scheme, which combines the first order Godunov method and the second order Lax-Wendroff method [7]. Since the flow and the acoustic variables have different time scales, the flow and the acoustic field are not computed using the same time step. Instead, each time step of the flow is further divided into several time steps for the acoustic computation.

## 4 TEST CASE

The main objective of this paper is to study the aeroacoustic performance of different turbulence models. For this purpose, the aeroacoustic experimental data of Etkin et al. [3] are used to compare different turbulence models. This test case investigates the flow past a cylinder and the acoustic quantities generated by the turbulent flow. The inlet flow velocity is 68.6 m/s, implying a Mach number of Ma=0.2 such that the incompressibility assumption is reasonable. The diameter of the cylinder is D=0.0125m. The Reynolds number is approximately 60000, indicating that the flow is fully turbulent.
As can be seen in Fig.2, the domain has a length of 22D in $x$-direction and 10D in $y$-direction. In order to capture the three dimensional turbulent features, the $z$-direction is modeled with a length of 4D.

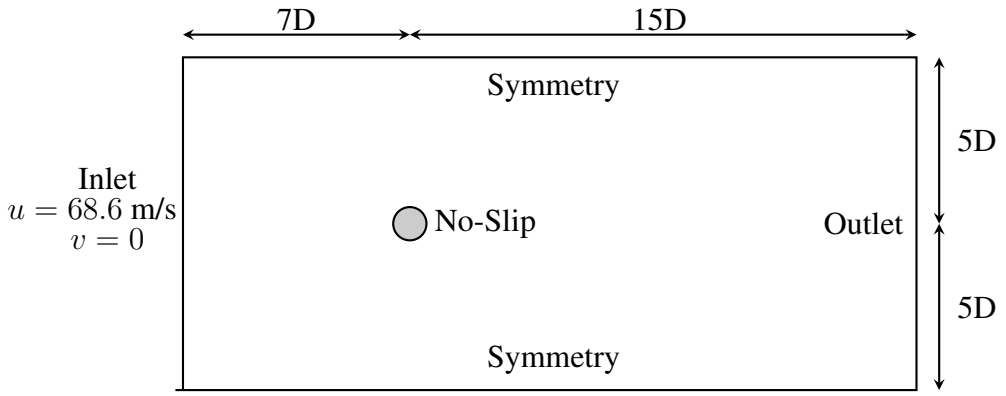
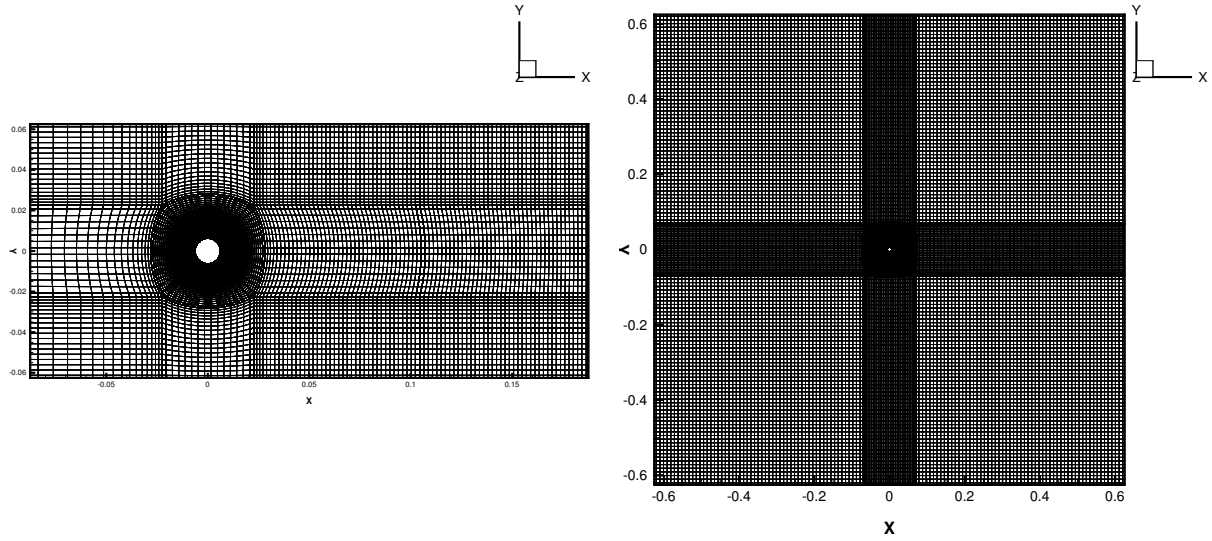



Figure 2: Sketch and boundary conditions of flow domain.

The problem domain is discretized using a block-structured grid. Figure 3a shows the spatial discretization of the flow domain. The flow is resolved up to the boundary layer, resulting in the dimensionless wall distance $y^+ < 1$. Figure 3b illustrates the discretization of the acoustic domain. Considering that the experimental data are collected at a distance of 48D from the cylinder's center, the acoustic domain has a length of 100D in each direction.




(a) Flow grid (b) Acoustic grid

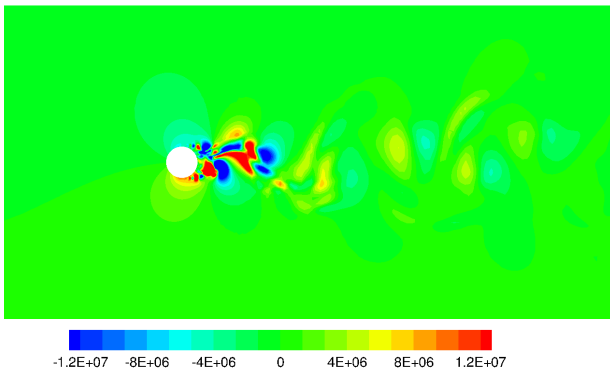
Figure 3: Discretization of flow domain and acoustic domain

The number of CVs for the flow and acoustic domains are listed in Table 1. In the LES model, 80% of the turbulence energy should be resolved. Hence, it needs a relatively fine grid to achieve adequate simulation results. A flow grid with about 2.5 million cells is adopted in this work. For the URANS model, a relatively coarse grid is utilized. The grid for the DDES model is created according to the grid resolution requirements from [17]. Regarding the time discretization, the second order fully implicit scheme is adopted, which has no stability problem when the CFL number is greater than 1 [11]. In order to fairly compare the SPL spectrum of different models, the sampling frequency is kept the same, which is realized by setting the time step for the flow to $3 \times 10^{-6}$s for all cases.

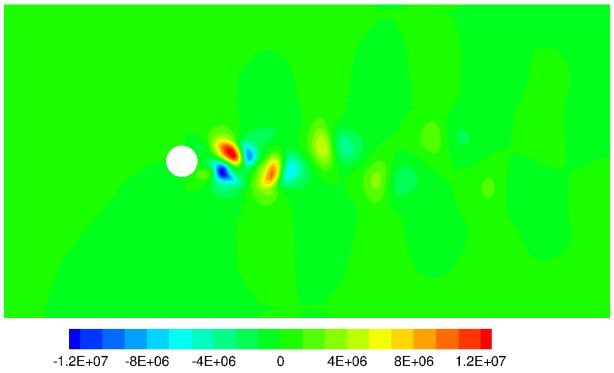
Table 1: Turbulence models and number of CVs.

| Simulation | Turbulence model | No. of CVs of flow domain | No. of CVs of acoustic domain |
|---|---|---|---|
| No.1 | LES | 2500368 | 158496 |
| No.2 | $\zeta - f$ based DDES | 1233648 | 158496 |
| No.3 | $k - \omega$ SST | 184480 | 158496 |

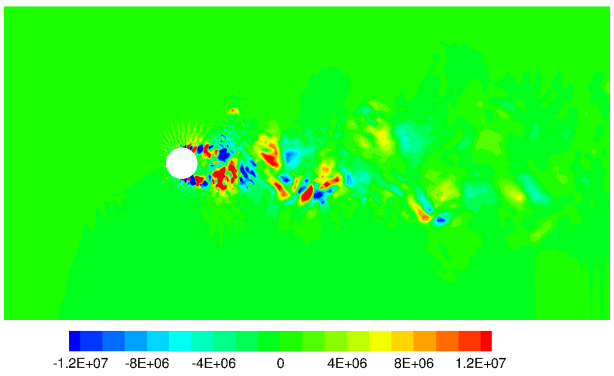
## 5 RESULTS AND DISCUSSION




(a) Acoustic source of LES model in [Pa/s]




(b) Acoustic source of URANS model in [Pa/s]




(c) Acoustic source of DDES model in [Pa/s]

Figure 4: Acoustic sources obtained by different turbulence models

In order to acquire a fully developed von Kármán vortex street and a stable propagation process of acoustics, 30000 time steps are observed. For the calculation of this time period, the LES model requires approximately 4 days, the DDES model needs 2.5 days, while the URANS model needs only 10 hours.
Figure 4 shows the acoustic sources predicted by three different turbulence models. According to the simulation of LES, as shown in Fig.4a, the acoustic sources appear predominantly in the wake region and the shear layer. The URANS model is not suitable to predict the acoustic source in the shear layer, as indicated in Fig.4b. In the wake region, only large periodic sources can be predicted to some extent. Figure 4c illustrates that the DDES turbulence model can predict the acoustic sources in the shear layer accurately. However, the acoustic sources in the wake region can only be partially predicted. This may be caused by the "grey area" problem. The DDES model switches between LES and RANS modes depending on the numerical resolution. However, there often exist regions where it is neither LES nor RANS mode, resulting in a large simulation error in such areas. This problem can emerge after the flow separation, where the mode should be changed from RANS to LES. Due to the insufficient turbulence information from the RANS mode from the upstream, the LES mode can not develop immediately after the transition.
As in the experiment of Etkin et al. [3], the acoustic data at an observer point are collected, which is placed at a distance of 48D from the cylinder's center perpendicular to the flow direction. The $1/3$ octave spectrum of the simulations and the experimental data are given in Fig.5.

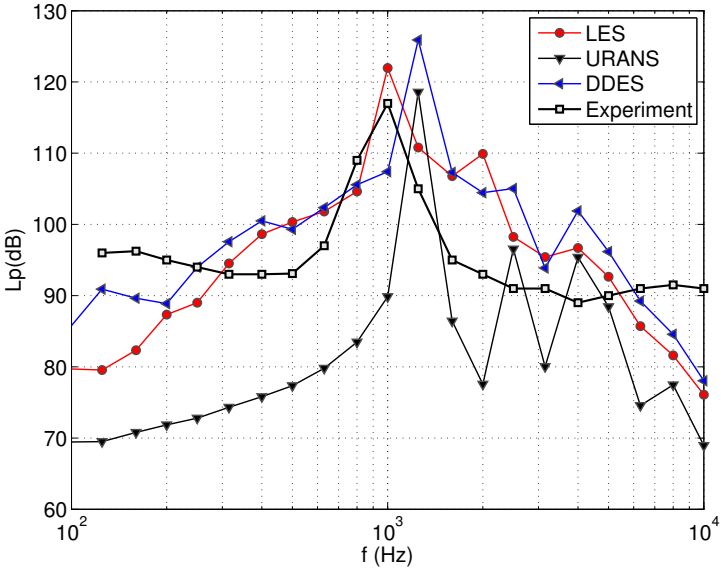



Figure 5: Comparison of SPL of different turbulence models with experimental data

The LES model offers the closest comparison with the experimental data. The underprediction in the high frequency region is caused by the filtering process in LES, which leads to a loss of the high frequency fluctuation component in the flow. In addition, as shown in Table 2, the fundamental frequency obtained by LES is 1061Hz, which has a relative error of only 6.1%. As illustrated in Fig.5, the SPL of the $k-\omega$ SST model has an obvious deviation from the experimental data. Nevertheless, this URANS model can provide the fundamental frequency with a relative error of 22%.
For the $\zeta - f$ based DDES model, a similar SPL spectrum to the LES model is observed.

However, it can only predict the fundamental frequency with a similar accuracy as the URANS model, as indicated in Table 2.

Table 2: Fundamental frequency.

| | Frequency (Hz) | Error |
|---|---|---|
| Experiment | 1000 | - |
| LES | 1061 | 6.1% |
| URANS ($k-\omega$ SST) | 1222 | 22% |
| $\zeta - f$ based DDES | 1200 | 20% |

## 6 CONCLUSION

Three different turbulence models have been investigated in the context of aeroacoustic simulation. The SPL spectrum of the LES model shows the best agreement with the experimental data. The underprediction in the high frequency region is also reasonable. The URANS model can only provide the fundamental frequency with a modest error. Its SPL spectrum deviates obviously from the experimental data. In comparison to URANS, the DDES model has been shown to be able to predict the SPL spectrum with a satisfactory accuracy. However, only a minimal improvement in predicting the fundamental frequency is observed.

## 7 ACKNOWLEDGMENT

This work is supported by the 'Excellence Initiative' of the German Federal and State Governments within the Graduate School of Computational Engineering at Technische Universität Darmstadt.

## REFERENCES

[1] C. Bailly, C. Bogey, O. Marsden, Progress in direct noise computation. *International Journal of Aeroacoustics*. 9(1):123-143, 2010.,

[2] N. Curle, The influence of solid boundaries upon aerodynamic sound, *Proceedings of the Royal Society (London)*. 231(A):505-514, 1955.

[3] B. Etkin, G.K. Korbacher, R.T. Keefe, Acoustic Radiation from a Stationary Cylinder in a Fluid Stream (Aeolian Tones). *Journal of the Acoustical Society of America*. 2930-36, 1957.

[4] J. Fröhlich, D. Terzi, Hybrid LES/RANS methods for the simulation of turbulent flows, *Progress in Aerospace Sciences*, 44(2008):349-377, 2008.

[5] K. Hanjalić, M. Popovac, M. Hadžiabdić, A robust near-wall elliptic-relaxation eddy-viscosity turbulence model for CFD. *International Journal of Heat and Fluid Flow*, 25(6):1047-1051, 2004.

[6] J.C. Hardin, D.S. Pope, An acoustic/viscous splitting technique for computational aeroacoustics. *Theoretical and Computational Fluid Dynamics*. 6(5):323-340, 1994.

[7] M. Kornhaas, M. Schäfer, D.C. Sternel. Efficient numerical simulation of aeroacoustics for low Mach number flows interacting with structures. *Computational Mechanics*, 55:1143-1154, 2015.

[8] M.J. Lighthill, On Sound Generated Aerodynamically. I. General Theory. *Proceedings of The Royal Society of London A*. 211:564-587, 1952.

[9] M.J. Lighthill, On Sound Generated Aerodynamically. II. Turbulence as a source of sound. *Proceedings of The Royal Society of London A*.Volume 222, pp. 1-34, 1954.

[10] F.R. Menter, Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications *AIAA Journal*, 32(8):1598-1605, 1994.

[11] M. Schäfer, *Computational Engineering - Introduction to Numerical Methods*. Springer, 2006.

[12] W.Z. Shen, J.N. Sørensen, Aeroacoustic Modelling of Low-Speed Flows *Theoretical and Computational Fluid Dynamics*. Volume 13, Issue 4, pp. 271-289, 1999.

[13] W.Z. Shen, J.N. Sørensen, Comment on the aeroacoustic formulation of Hardin and Pope. *AIAA Journal*. 37(1):141-143, 1999.

[14] J. Smagorinsky, General circulation experiments with the primitive equations I: The basic experiment *Monthly weather review*, 91(3):99-163, 1963.

[15] P.R. Spalart. Detached-eddy simulation. *Annual Review of Fluid Mechanics*, 41(1):181-202, 2009.

[16] P.R. Spalart, S. Deck, M.L. Shur, K.D. Squires, M.Kh Strelets, A. Travin. A new version of detached-eddy simulation, resistant to ambiguous grid densities. *Theoretical and Computational Fluid Dynamics*, 20(3):181-195, 2006.

[17] P.G. Tucker, Unsteady Computational Fluid Dynamics in Aeronautics, *Fluid Mechanics and Its Applications*, Volume 104, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL INVESTIGATION OF THERMOACOUSTIC ENGINE USING IMPLICIT LARGE EDDY SIMULATION

**M. Mahdaoui[1,2], R. Bennacer[2] and S. Kouidri[1, 3], N. Martaj [4]**

[1] LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France
Mahdaoui@limsi.fr

[2] LMT /ENS-Cachan/CNRS/Université Paris Saclay,
61 avenue du Président Wilson, 94235 Cachan, France
Rachid.bennacer@ens-cachan.fr

[3] Sorbonne University, UPMC Univ, Paris 6
UFR 919, 75005 Paris, France
Smaine.Kouidri@limsi.fr

[4] EPF Campus de Troyes
2, rue F. Sastre 10430 Rosières-près-Troyes, France
nadia.martaj@epf.fr

**Keywords:** Thermoacoustic engine, Acoustic streaming, Implicit Large Eddy Simulation, CFD.

**Abstract.** *Thermoacoustic engines are device that converts heat energy into work in the form of acoustic energy. The energy conversion using thermoacoustic techniques has the advantages of constructional simplicity, absence of moving parts, reliable operation, long service life and no environmental pollution. However, efficiency of thermoacoustic systems still needs to be improved (which is typically 40% of the Carnot coefficient of performance).*
*To improve the efficiency of thermoacoustic engine, a thorough understanding of the flows in machines, onset temperature, acoustic velocity and streaming velocity is necessary. In this study, numerical simulation of a thermoacoustic heat engine is performed. The possibility of using Implicit Large Eddy Simulation (ILES) to simulate the flows in the thermoacoustic engine has been examined by using COMSOL Multiphysics. Unlike the Large Eddy Simulation (LES), ILES is neither explicit filtering nor explicit subgrid model: small-scale fluctuations are damped by the numerical diffusion, which acts both as an implicit filter and an implicit built-in subgrid model for a given grid. As the grid is refined, the simulation converges to a DNS.*
*The results show that the ILES has predicted flows in the thermoacoustic engine and especially time-averaged flow called acoustic streaming, which can lead to undesired heat convection and loss of efficiency.*

## 1 INTRODUCTION

The thermoacoustic machines are a new type machines based on the interaction between the thermodynamic and acoustic phenomena. In recent years, these machines occupy an important place in both the industrial development and in the comfort of habitat. This diversification of the mode of use of thermoacoustic machines is due to a number of benefits, recorded on these machines: no moving parts, high reliability and low environmental friendliness compared with traditional heat engine. [1-3]. However, efficiency of thermoacoustic systems still needs to be improved, especially for the standing wave thermoacoustic engine (which is typically 40% of the Carnot coefficient of performance (Swift [3])) and their performance needs to be better understood. To improve the efficiency of thermoacoustic engine and refrigerators apart from improved resonator, stack and loudspeaker design, it is necessary to better understand the flow and thermal fields. Many strategies are proposed to improve the efficiency of thermoacoustic machines.

In thermoacoustic devices, the energy is carried by an acoustic wave. This acoustic wave can induce a time-averaged flow, called acoustic streaming, which can lead to undesired heat convection and loss of efficiency. Different classes of acoustic streaming can be distinguished based on the underlying physical mechanisms. Rayleigh streaming, as first analytically described by Lord Rayleigh in 1884 [4], is the phenomenon of a net mean flow in a standing wave resonator driven by the viscous stresses close to the solid boundary [5]. In the traveling-wave thermoacoustic engine usually has a loop configuration, which may cause an acoustic streaming named as Gedeon streaming [6].

Acoustic streaming can have a large impact on the performance of thermoacoustic prime movers and refrigerators as the generated mean flow can lead to undesirable mean convective heat transport [7, 8]. There are many experimental and theoretical studies to better understand and to suppress acoustic streaming [8-14]. They investigated mechanisms of streaming generation, and they introduced elements to suppress the streaming like a jet pump and phase adjuster.

Several numerical calculations recently addressed the modeling of the flow field inside a thermoacoustic couple [15-19]. Cao et al. [15] presented numerical simulations of the flow behavior in the thermoacoustic couple, and calculated the energy flux density in a parallel-plate stack.

To understand the evolution of the physical parameters (pressure, velocity, streaming, etc.) in thermoacoustic engine, a 2D model was used in view of the geometric configuration of the machine. This model is based on solving the Navier Stokes equations, the heat equation and the ideal gas equation. The Implicit Large Eddy Simulation (ILES) was used to simulate the flows in thermoacoustic machine [20-22]. This method uses the full Navier-Stokes equations instead of Reynolds Averaged Navier-Stokes (RANS) and does not use an explicit turbulence model. Unlike the Large Eddy Simulation (LES), ILES is neither explicit filtering nor explicit subgrid model: small-scale fluctuations are damped by the numerical diffusion, which acts both as an implicit filter and an implicit built-in subgrid model for a given grid. As the grid is refined, the simulation converges to a DNS. ILES is an effective method for transitional flows, especially when coupled with heat transfer. It captures accurately the energy exchanges between various modes, namely convective, acoustic and thermal, in a wide range of velocity and turbulence intensity.

## 2 PHYSICAL AND NUMERICAL MODEL

The thermoacoustic engine illustrated in Fig. 1 is made of a resonator, opened to the environment on its one side, a stack and two heat exchangers whose function is to create an appropriate thermal gradient. A standing acoustic wave is generated in the device. Its velocity node is at the closed end of the tube and the velocity antinode – at the opened cross section of the engine. The stack, of length 10mm, formed by a stack of rectangular plates of 0.5 mm wide and spaced 0.5mm between them. Our model problem is similar to the one considered in [17].

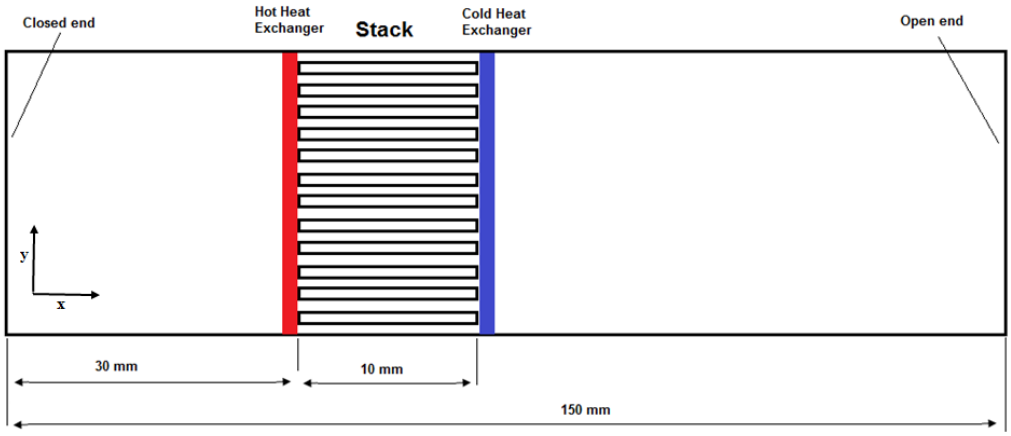



**Figure 1:** The thermoacoustic engine and its basic components.

To model the thermoacoustic generator, we used the mass conservation, momentum, energy equations and the equation of state.

The evolution of pressure, velocity and density is governed by the Navier-Stokes equations applied to the gas circulating in the resonator and in the stack.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0 \tag{1}$$

$$\rho \left( \frac{\partial u}{\partial t} + (u \cdot \nabla)\, u \right) = \nabla \cdot [-p\ \ I + \mu\, (\nabla u + (\nabla\ u)^T) - \frac{2}{3}\, \mu\, (\nabla \cdot u)\, I] + F \tag{2}$$

The energy equation used is as follows:

$$\rho C_P \left( \frac{\partial T}{\partial t} + u \cdot \nabla T \right) = \nabla \cdot (k \nabla T) + Q_{vh} + W_p \tag{3}$$

Where $W_p$ is the pressure work term and $Q_{vh}$ is the contribution of viscous heating given by:

$$W_p = \alpha_p T \left( \frac{\partial p_A}{\partial t} \right) \tag{4}$$

$$\alpha_p = -\frac{1}{\rho} \left( \frac{\partial \rho}{\partial T} \right)\Bigg|_p \tag{5}$$

$$Q_{vh} = \tau : \nabla u \tag{6}$$

$\alpha_p$ is thermal expansion coefficient.

To solve the system of partial differential equations described above, it is necessary to provide additional equation, linking the physical parameters of the fluid.

We consider the following state of the ideal gas equation for the working gas (air):

$$\rho = \frac{P_A}{R_S T} \tag{7}$$

Firstly, a steady simulation was performed with 10 Pa pressure as inlet condition at the closed end of the resonator and atmospheric pressure at the output of the resonator.
In the present study, all walls are adiabatic, except horizontal surfaces in the stack. The boundary condition of the temperature along such horizontal stack surfaces is determined by the following expression:

$$T(x) = 500 - 200 \cos(\pi \frac{x_{stack}}{L_{st}}) \tag{8}$$

This steady simulation was used to introduce pressure disturbance and an initial flow for transient simulation. In this case, adiabatic wall replaced pressure inlet condition on the compliance side.
A mesh based triangular elements was used with a total cell number of about 160000 cells. The mesh was refined near wall in the boundary viscous layer of the stack. The resolution of this boundary layer is necessary since the formation of acoustic streaming is principally induced by the interaction between the acoustic wave and the viscous boundary.

## 3 RESULTS ANS DISCUSSION

The model described above is used to study a thermoacoustic engine using the implicit large eddy simulation implemented in COMSOL Multiphysics software. As implicit modeling we denote the situation when the unmodified conservation law is discretized and the numerical truncation error acts as an SGS model. Since this SGS model is implicitly contained within the discretization scheme, an explicit computation of model terms becomes unnecessary.
The process of the acoustic wave excitation is presented in Fig 2. The basic time step of $10^{-5}$ second was adopted for the unsteady calculations. It can be noticed that in the initial stage of the pressure oscillation stems from an initial disturbance until the balance between the acoustic power dissipation and acoustic power generation is reached. The peak-to-peak value of thermoacoustic oscillation was 8 kPa. The FFT analysis of the pressure oscillation is shown in the Fig. 3. The main frequency component is equal to about 614 Hz and is in a good agreement with results of [17]. Two additional harmonics, at 1222 Hz and 1830 Hz are visible.
Fig. 4 shows the acoustic velocity distribution, u, versus time at fixed position of x= 20 mm, the velocity amplitude at this position is about 2.5 m/s. The acoustic velocity u along the length of the thermoacoustic engine for individual phase of half period T/2 of oscillation is presented in Fig. 5. The characteristics were determined along the line y= 4.23mm. We find that the acoustic velocity is maximum at x = 30mm and equal to 11m / s. This maximum value is due to the presence of the stack, which generates a constricting section. The velocity amplitude was determined at the open end of the device where the velocity antinode located.
To study the distribution of pressure in the thermoacoustic engine Figs. 6 and 7 present the acoustic pressure p distribution in function of time at x= 20 mm and The acoustic pressure u along the length of the thermoacoustic engine for individual phase of half period Tp/2 of oscillation along the line y =4.23mm. In the close end of the device, the pressure amplitude is equal 4 kPa in this place the pressure antinode appears. It can also be noticed in the outlet section of the thermoacoustic engine a pressure node is located.

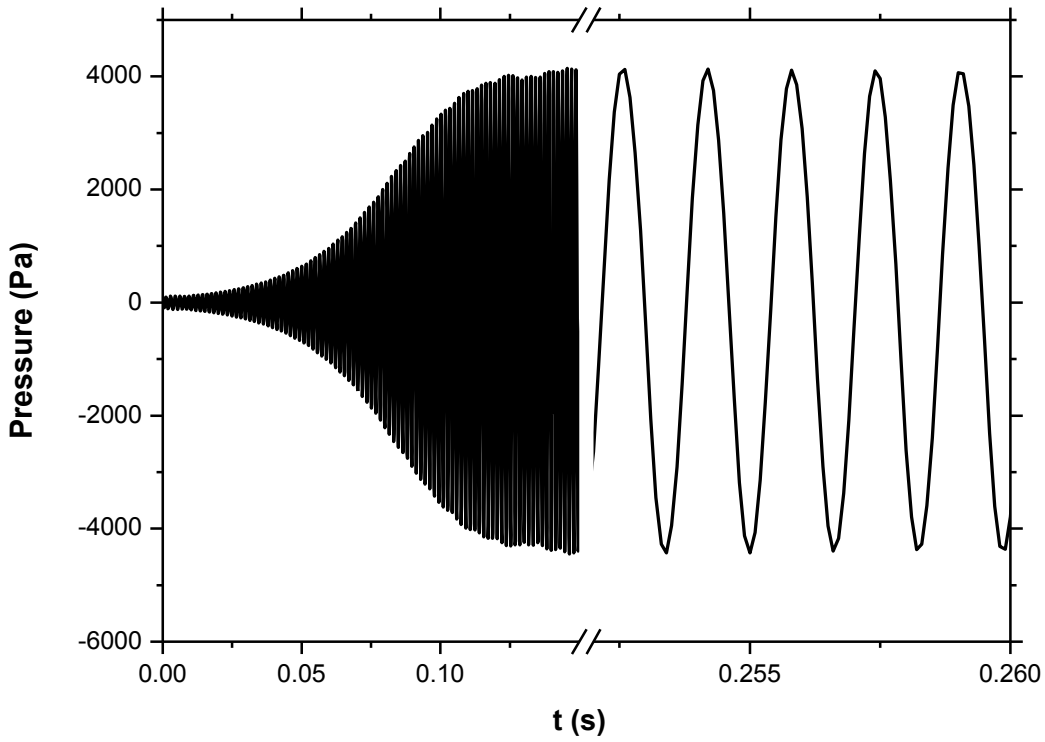



**Figure 2:** The evolution of the pressure oscillation at a x= 20mm.

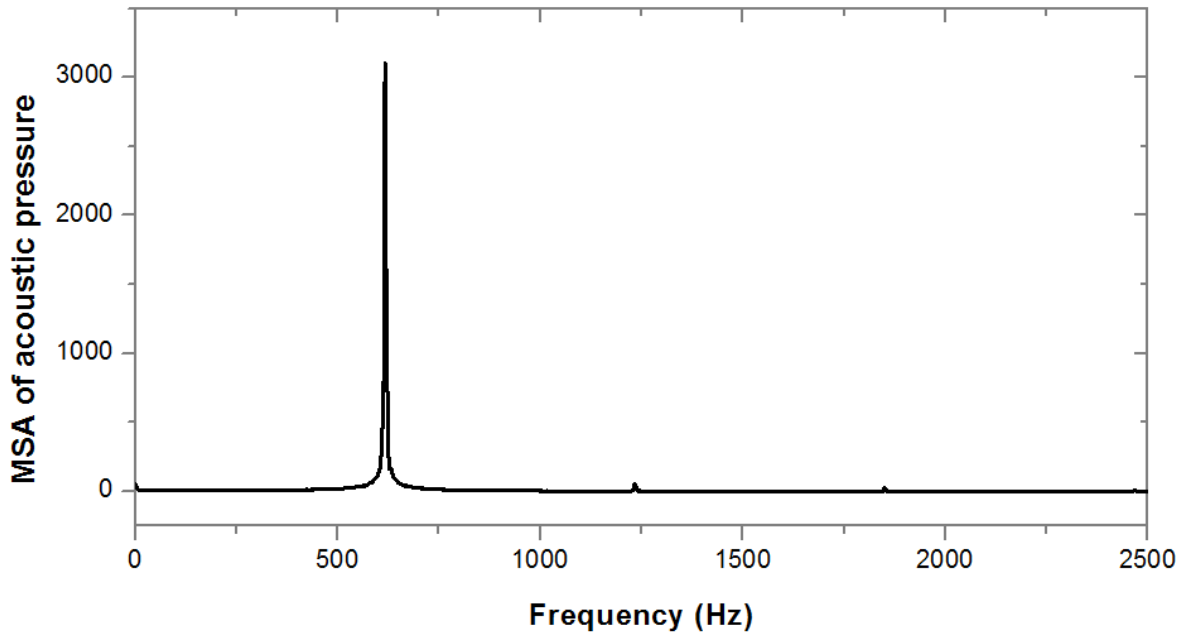



**Figure 3:** FFT analysis of the pressure oscillation.

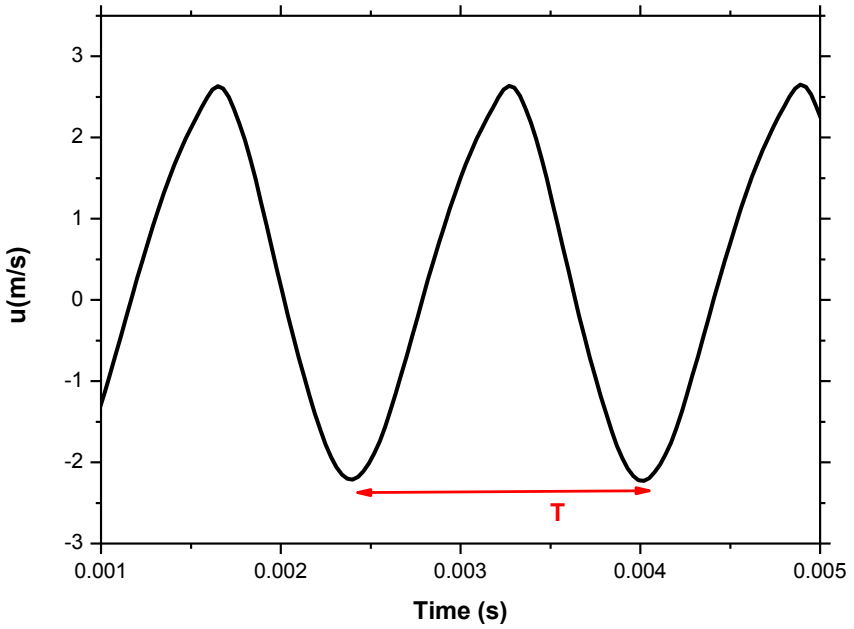



**Figure 4:** Time evolution of the acoustic velocity at a point in the vicinity of the stack.

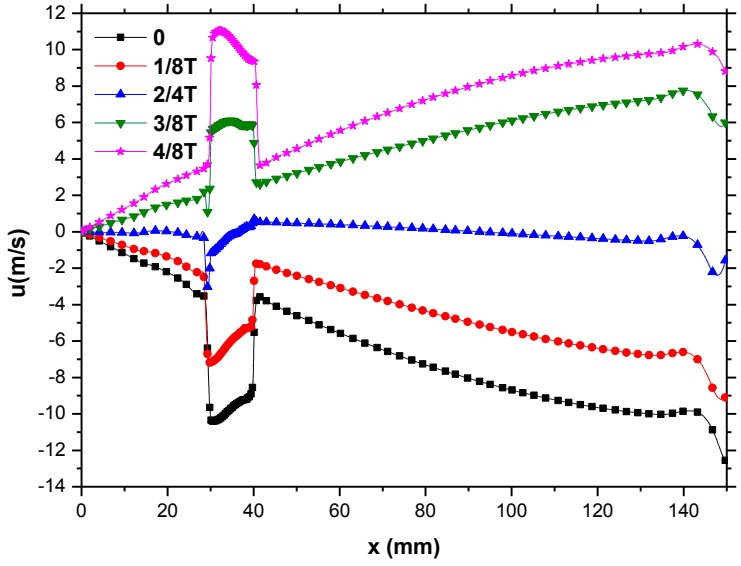



**Figure 5:** Distribution of instantaneous velocity depending on the length of thermoacoutic engine.

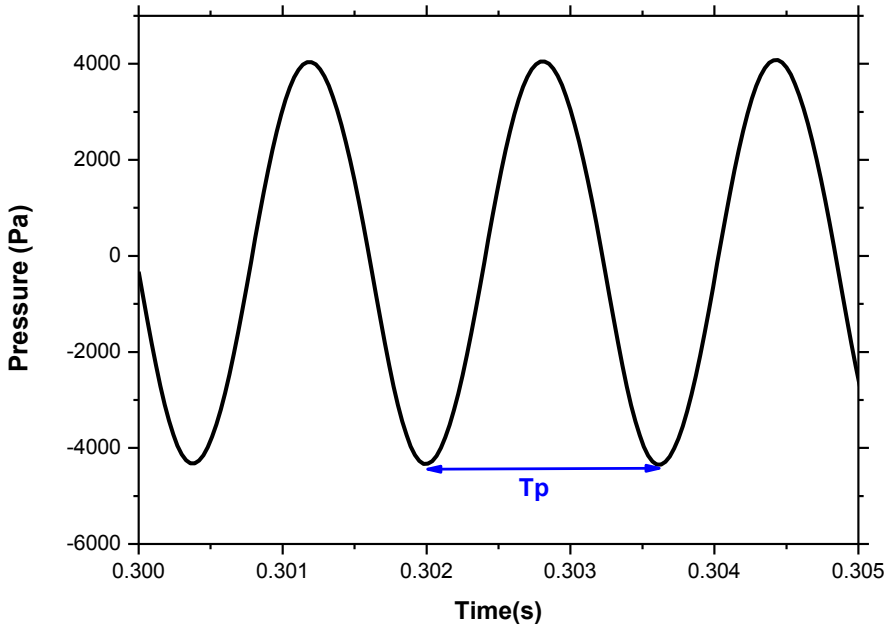



**Figure 6:** The evolution of the pressure oscillation.

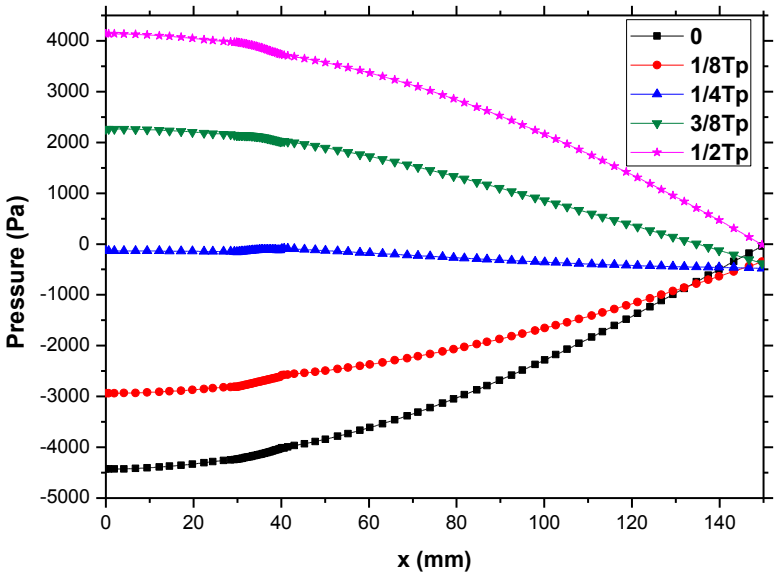



**Figure 7:** Distribution of instantaneous pressure depending on the length of thermoacoutic engine.

Secondary flows are mean flows that reduces the energy efficiency of thermoacoustic machine by creating thermal bridges particularly between stack ends. Fig. 8 depicts the cycle averaged flow field in the thermoacoustic engine. These steady flow structures (acoustic streaming) are computed from on the average mass transport velocities defined as;

$$u_2 = \frac{\langle \rho u \rangle}{\langle \rho \rangle} \qquad v_2 = \frac{\langle \rho v \rangle}{\langle \rho \rangle} \tag{10}$$

Here $u_2$ and $v_2$ are components of the acoustic streaming velocity. The instantaneous density, u and v values are time averaged during the 60$^{th}$ cycle and the streaming velocities obtained at the time which the mass transport velocities reach quasi steady values and do not change with increasing number of periods.

From Fig. 8 it can be seen that there are several secondary flow into the machine. These streaming appear in the resonator of thermoacoustic engine and near close end. Also small vortex are seen next to the stack and between the plates of the stack. These streaming reduce the useful gradient and therefore the engine efficiency. To make a quantitative study of streaming velocity, the profile of $u_2$ was plotted for three different axial positions as illustrated in Fig. 9. From this figure, it can be seen the symmetry of the streaming velocity curves over the axis that passes through the engine center. the acoustic velocity varies between 0.3 m/s and -0.3 m/s. In both areas, next to the stack, we plotted the velocity profile along y (see Fig. 10). After this analysis, the ability of the ILES approach to capture the small structures in the oscillating flow in the thermoacoustic engine is confirmed.

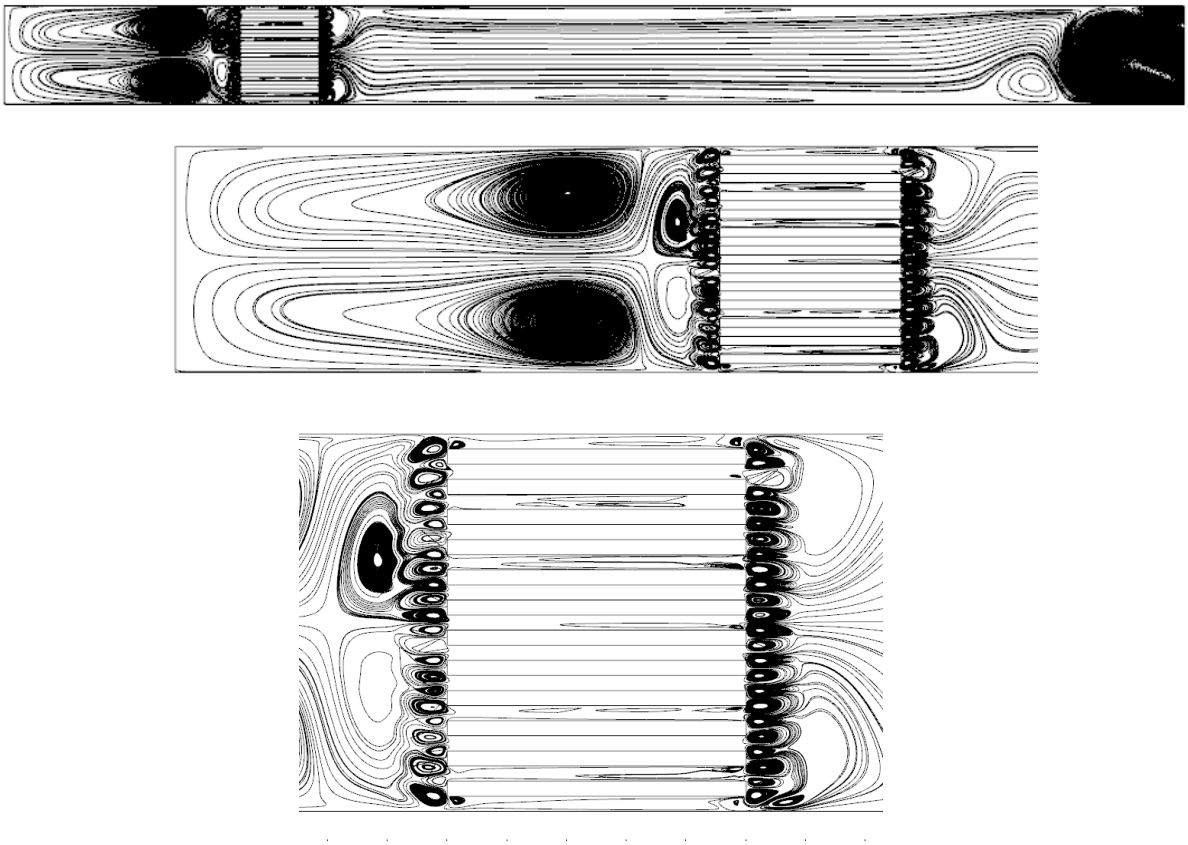


**Figure 8:** Stream line for streaming velocity

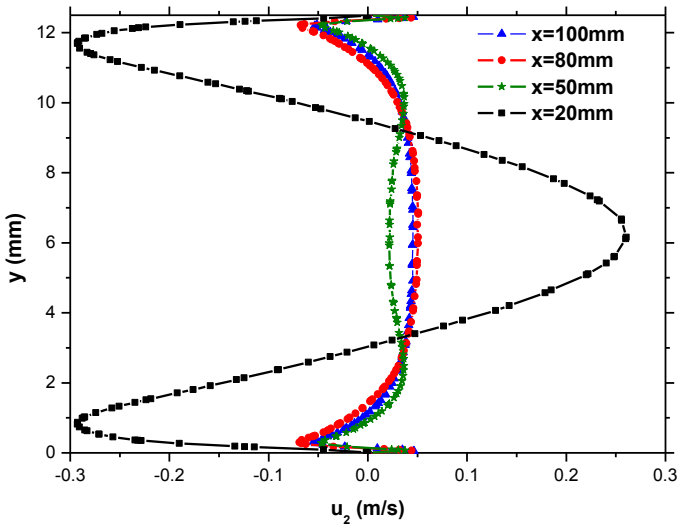



**Figure 9:** Streaming velocity profile $u_2$ at three axial positions.

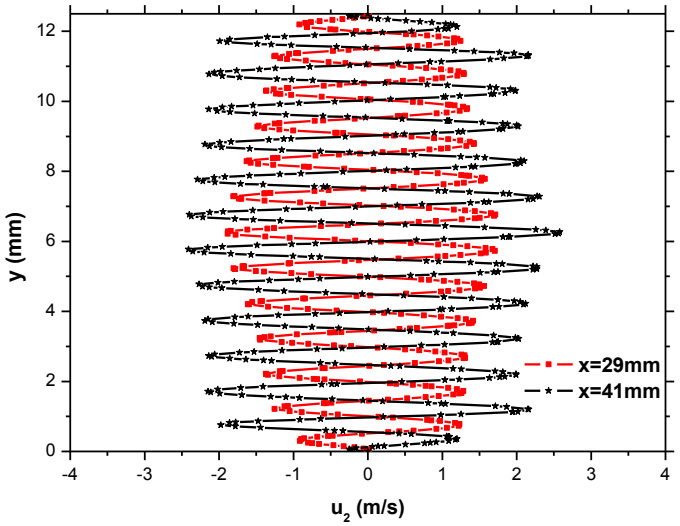



**Figure 10:** Streaming velocity profil at two positions near the stack**.**

## 4 CONCLUSIONS

This work presents the study of streaming velocity in a thermoacoustic engine using Implicit Large Eddy Simulation. The model used is based on solving the Navier Stokes equations, the heat equation and the ideal gas equation. First order velocity and second order velocity (streaming velocity) are obtained and analyzed. The results show that the ILES has predicted flows in the thermoacoustic engine and particularly acoustic streaming, which can lead to undesired heat convection and loss of efficiency. These streaming appear into thermoacoustic engine, especially between the plates of the stack.

## REFERENCES


[1] G.W. Swift, Thermoacoustic engines, *Journal of the Acoustical Society of America 84* 1145–1180, 1988.

[2] J.C. Wheatly, T. Hofler, G.W. Swift, A. Miglori, Understanding some simple phenomena in thermoacoustics with applications to acoustical heat engines, *American Journal of Physics 53*, 147–162, 1985.

[3] G.W. Swift, Thermoacoustic engines and refrigerators, *Physics Today*, 22–28, 1995.

[4] L Rayleigh, On the circulation of air observed in Kundt's tubes, and on some allied acoustical problems, *Philosophical Transactions of the Royal Society of London* 175, 1–21, 1884.

[5] S. Boluriaan and P. J. Morris, Acoustic streaming: from Rayleigh to today, *International Journal of Aeroacoustics* 2, 255–292, 2003.

[6] D. Gedeon, *DC gas flows in Stirling and pulse tube cryocoolers*. In: Ross Jr RG, editor. Cryocoolers 9. New York: Plenum Press, p. 385–92, 1997.

[7] S. Backhaus, GW. Swift, A thermoacoustic Stirling heat engine. *Nature,* 399(6734), 335–8, 1999.

[8] S. Backhaus, GW. Swift. A thermoacoustic-Stirling heat engine: detailed study. *J Acoust Soc Am,* 107(6), 3148–66, 2000.

[9] N. Rott, Thermoacoustics, *Adv. Appl. Mech*. 20, 135–175, 1980.

[10] MF. Hamilton, YA Ilinskii, EA Zabolotskaya. Thermal effects on acoustic streaming in standing waves. *J Acoust Soc Am*,114:3092–101, 2003.

[11] IE. Idelchik *Handbook of hydraulic resistance*. *3rd ed*. New York: Begell House; 1996.

[12] R. Paridaens, S. Kouidri, & F. J. Jerbi, Investigation on the generation mechanisms of acoustic streaming in a thermoacoustic prime mover. *Cryogenics*, 58, 78-84, 2013.

[13] Y. W. Liu, & P. Yang, Influence of inner diameter and position of phase adjuster on the performance of the thermo-acoustic Stirling engine. *Applied Thermal Engineering*, 73(1), 1141-1150, 2014.

[14] M. Chen, & Y. L. Ju, Effect of different working gases on the performance of a small thermoacoustic Stirling engine. *International Journal of Refrigeration*, 51, 41-51, 2015.

[15] N. Cao, JR. Olson, GW. Swift, S. Chen, Energy flux density in a thermoacoustic couple. *J Acoust Soc Am,* 99(6):3456–64, 1996.

[16] N. Lycklama, J. L., M. E. H. Tijani, & S. Spoelstra. Simulation of a traveling-wave thermoacoustic engine using computational fluid dynamics. *The Journal of the Acoustical Society of America*, 118(4), 2265-2270, 2005.

[17] F. Zink, J. Vipperman, & L. Schaefer, CFD simulation of a thermoacoustic engine with coiled resonator. *International Communications in Heat and Mass Transfer*, 37(3), 226-229, 2010.

[18] S. H. Tasnim & R. A. Fraser, Computation of the flow and thermal fields in a thermoacoustic refrigerator. *International Communications in Heat and Mass Transfer, 37*(7), 748-755, 2010.

[19] G. Yu, W. Dai, & E. Luo, CFD simulation of a 300Hz thermoacoustic standing wave engine. *Cryogenics*, *50*(9), 615-622, 2010.

[20] F.F. Grinstein, L.G. Margolin, W.J. Rider, *Implicit Large Eddy Simulation*, Cambridge University Press, 2007.

[21] COMSOL Multiphysics 5.1, *COMSOL Reference Guide*, 2015.

[22] M. Hashiguchi, Possibility of Implicit LES for Two-Dimensional Incompressible Lid-Driven Cavity Flow Based on COMSOL Multiphysics. *Accessed online* 11, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SEMI-DISCRETE COINCIDENCE IN THE MID-FREQUENCY SOUND TRANSMISSION THROUGH RIB-STIFFENED PANELS

**Edwin Reynders, Cédric Van hoorickx and Arne Dijckmans**

KU Leuven Dept. of Civil Engineering
Kasteelpark Arenberg 40, B-3001 Leuven, Belgium
edwin.reynders@kuleuven.be

**Keywords:**

**Abstract.** *Plates with attached beams or ribs, rather than monolithic plates, are commonly employed in civil and mechanical engineering for reasons of weight and static stiffness. Unfortunately, attaching stiffeners to a plain plate generally increases its radiation efficiency and reduces its airborne sound insulation. This contribution aims at gaining insight into the sound insulation of rib-stiffened plates in the mid-frequency range, where the modal behavior of the plate is still important, but the neighboring sound fields can already be considered as diffuse. A detailed finite element model of a PMMA plate with steel stiffeners attached is constructed and coupled to a reverberant sound field model of the adjoining room(s) within the recently developed hybrid finite element - statistical energy analysis framework. This framework is computationally very efficient, it enables to compute the coupling loss factors between the sound fields in a rigorous, straight forward way, and the uncertainties due to random wave scattering - which is the physical origin of the diffuse field - can be quantified. The finite element model of the plate is first calibrated by minimizing the difference between its lowest natural frequencies and mode shapes and the corresponding measured values. The hybrid model is then employed for predicting the sound reduction index across the building acoustics frequency range. These predictions are validated against airborne sound insulation measurements. Within a wide frequency range, pronounced dips are observed at specific natural frequencies of the plate, at which the wavelength of the corresponding mode shape is close to the free acoustic wavelength. Since there is an obvious connection with the coincidence phenomenon for infinite plates, the observed phenomenon is termed the semi-discrete coincidence phenomenon. It is shown that in the mid-frequency range, an important increase in airborne sound insulation can be achieved by suppressing only a few particular resonances of the considered rib-stiffened plate.*

## 1 INTRODUCTION

Rib-stiffened plates are commonly employed in civil and mechanical engineering as they achieve a similar strength and stiffness as plain plates with a substantially lower weight. Unfortunately, attaching stiffeners to a plain plate generally increases its radiation efficiency and reduces its airborne sound insulation [3].

Early investigations into these phenomena led to two approximate physical explanations based on infinite plate theory. Firstly, when the wavelength of deformation is much larger than the distance between the stiffeners, the main effect of the stiffeners is to introduce orthotropy into the plate. The effect of the stiffeners is then that the frequency range in which coincidence has a detrimental effect on radiation efficiency and transmission loss is broadly smeared out towards lower frequencies when compared to the plane plate [6]. Secondly, when the wavelength of deformation is much smaller than the distance between the stiffeners, the stiffeners have the approximate acoustic effect of dividing the plate into smaller panels, the radiation efficiency of which is much larger than the radiation efficiency of the large unstiffened plate [12]. These investigations were followed by several more refined analyses; see, e.g., [1, 2, 5, 11, 13].

The present contribution aims at gaining additional insight into the sound transmission through finite rib-stiffened plates by investigating the behavior of the room-wall-room system in the mid-frequency range. This is the frequency range in which the sound fields in the rooms can already be considered as diffuse (in contrast to the low-frequency range), while the modal behavior of the wall is still important (in contrast to the high-frequency range). A detailed finite element model of the rib-stiffened plate is constructed and coupled to a reverberant model of the adjoining rooms within the recently developed hybrid finite element - statistical energy (FE-SEA) analysis framework [16, 10, 15]. This framework is substantially more flexible and general than conventional statistical energy analysis as it allows, for example, (i) incorporating diffuse and non-diffuse vibro-acoustic system components within a single model in a computationally efficient way, (ii) assessing the uncertainty inherent in the assumption of diffuse sound fields in the rooms, and (iii) computing coupling loss factors in a rigorous and straightforward way, accounting automatically for both resonant and non-resonant transmission [9].

As a specific case, a polymethyl methacrylate (PMMA) plate with steel stiffeners attached is considered. By analysis of the model predictions, it is found that the finite dimensions of the plate result in a semi-discrete coincidence phenomenon, i.e., the wavelength of a mode shape of the rib-stiffened plate is close to the free acoustic wavelength at the corresponding natural frequency. This is seen to result in an oscillating narrow-band transmission loss over a wide frequency range, which is in contrast to the discrete coincidence phenomenon for finite homogenous isotropic plates [4, 7]. Only a few of the plate modes in this wide frequency range exhibit semi-discrete coincidence, which suggests that the sound insulation in that frequency range can be substantially increased by suppressing only a few structural modes.

The reminder of this article is structured as follows. In section 2, the dynamic structural model of the stiffened plate is discussed in detail. In section 3, the finite element model is incorporated in a hybrid FE-SEA model of the transmission suite and the transmission loss of the rib-stiffened plate is computed and validated against measurements. A physical interpretation in terms of semi-discrete coincidence is given in section 4, while the potential of suppression a few particular modes is investigated in section 5. Conclusions are drawn in section 6.

## 2 FINITE ELEMENT MODEL OF THE STIFFENED PLATE

The example structure that is considered in this paper consists of a base plate made of PMMA, to which 11 steel L30 stiffeners are attached. The base plate has a width of $L_\mathrm{x} = 1.25\,\mathrm{m}$, a height of $L_\mathrm{y} = 1.5\,\mathrm{m}$ and a thickness of $t_\mathrm{p} = 15\,\mathrm{mm}$. The steel L-shaped stiffeners have an outer leg length of $L_\mathrm{s} = 30\,\mathrm{mm}$ and a thickness of $t_\mathrm{s} = 3\,\mathrm{mm}$. The center-to-center spacing between the stiffeners is $d_\mathrm{x} = 100\,\mathrm{mm}$, the distance between a vertical edge of the plate and a vertical edge of the closest stiffener is $a_\mathrm{x} = 110\,\mathrm{mm}$, and the distance between a horizontal edge of the plate and the closest end section of a stiffener is $a_\mathrm{y} = 52.5\,\mathrm{mm}$. The stiffeners are both glued to the base plate over their entire length, and additionally screwed to the base plate at four points.

A finite element (FE) model of this structure was made using the ANSYS software package. Both the base plate and the steel stiffeners were modeled with four-node linear thin shell elements (of the SHELL181 type). The mesh is shown in Fig. 1. When connecting the plate and the horizontal stiffener legs, the offset of the stiffeners with respect to the midplane of the plate was taken into account. The out-of-plane displacements were restrained at the plate edges. No other boundary restrictions were applied, and the three in-plane rigid body modes were removed from the model. This set of boundary conditions agrees with the dynamic plate boundary conditions that are applied in the sound transmission suite of the KU Leuven Laboratory of Acoustics. For the steel profiles, a Young's modulus of $E_\mathrm{s} = 210\,\mathrm{GPa}$, Poisson's ratio of $\nu_\mathrm{s} = 0.3$ and density of $\rho_\mathrm{s} = 7850\,\mathrm{kg/m^3}$ were taken. For the PMMA base plate, a Young's modulus of $E_\mathrm{p} = 4.564\,\mathrm{GPa}$, a Poisson's ratio of $\nu_\mathrm{p} = 0.35$ and a density of $\rho_\mathrm{p} = 1170\,\mathrm{kg/m^3}$ were taken.

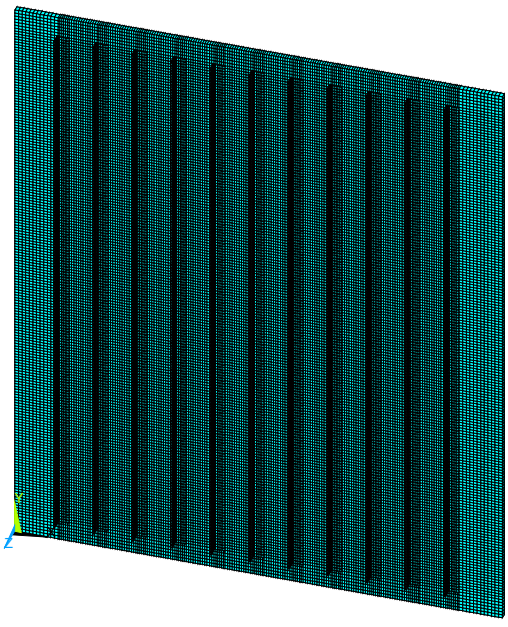


Figure 1: Finite element mesh. The volumetric representation is for visualization purposes only; both the plate and the stiffeners are modelled with two-dimensional shell elements.

## 3 SOUND INSULATION PREDICTIONS AND EXPERIMENTAL VALIDATION

For the sound insulation predictions, a hybrid finite element - statistical energy analysis (FE-SEA) approach [10, 15, 16] is adopted. This approach allows coupling diffuse field (or SEA

| $f$ [Hz] | $\eta[-]$ | $f$ [Hz] | $\eta[-]$ | $f$ [Hz] | $\eta[-]$ | $f$ [Hz] | $\eta[-]$ |
|---|---|---|---|---|---|---|---|
| 100 | 0.049 | 250 | 0.046 | 630 | 0.060 | 1600 | 0.044 |
| 125 | 0.049 | 315 | 0.056 | 800 | 0.055 | 2000 | 0.038 |
| 160 | 0.056 | 400 | 0.062 | 1000 | 0.059 | 2500 | 0.035 |
| 200 | 0.065 | 500 | 0.040 | 1250 | 0.054 | 3150 | 0.027 |

Table 1: Damping loss factor measured in 1/3 octave bands.

subsystem) models of modally dense components to deterministic models of modally sparse components in a rigorous way. A diffuse field model is a probabilistic wave field model, and therefore the predicted sound transmission loss is also a probabilistic quantity.

The model of the considered transmission suite is divided into three components: emitting room, partition wall, and receiving room. The rooms each have a volume of about $87\,\mathrm{m}^3$ and a reverberation time of about $1.5\,\mathrm{s}$. They are modeled to carry a diffuse wave field, as SEA subsystems. The values $c = 343\,\mathrm{m/s}$ and $\rho_\mathrm{a} = 1.2\,\mathrm{kg/m}^3$ are adopted for the sound speed and air density, respectively. For the partition wall, which consists of the baffled rib-stiffened plate, the finite element model of section 2 is employed. Measured values are used for the damping loss factor of the wall as reported in Table 3. Full details on sound transmission modeling with the hybrid method can be found in [15]. For completeness, it is mentioned that the acoustic direct field dynamic stiffness matrices of the rooms (SEA subsystems) were evaluated by means of a wavelet approach [8], using a square interpolation grid covering the flat plate surface and a distance between the grid points of $2.08\,\mathrm{cm}$.

Fig. 2 displays the model predictions and compares these with narrow-band (1/48 octave) sound reduction index measurements, which have been performed in the transmission suite of the KU Leuven Laboratory of Acoustics. Not only the mean of the predicted sound transmission loss values are plotted, but also the related $\pm 2\sigma$ confidence intervals, which correspond to $95\,\%$ confidence intervals when a Gaussian probability distribution can be assumed. One should keep in mind that this uncertainty is due to the assumption of a harmonic diffuse field model for the rooms. This assumption can be justified when the acoustic pressure field in the rooms is sensitive to small wave scattering elements, and for the transmission rooms at KU Leuven this is the case from about $140\,\mathrm{Hz}$ onwards [15]. At lower frequencies, the uncertainty due to the presence of small wave scattering elements will be overestimated with a diffuse field model. Furthermore, individual room resonances, which are absent in the adopted hybrid model, will be important at such frequencies. The present analysis however is concerned with mid-frequency sound transmission.

The critical frequency of the infinite panel without stiffeners would be at $2051\,\mathrm{Hz}$. The critical frequency of the ribbed plate in the weak bending direction (i.e., for bending around an axis parallel to the ribs) can be estimated at $2735\,\mathrm{Hz}$ by neglecting the stiffness of the steel ribs but accounting for their mass. The critical frequency in the strong bending direction can be estimated at $461\,\mathrm{Hz}$ by taking both the stiffness and the mass of the ribs into account. Between both frequencies, a broad coincidence zone exists for infinite plates. For the finite plate that is studied here, a strong oscillation of transmission loss with frequency is observed in this zone. Such oscillation is not predicted by, and cannot be explained from, infinite plate theory. It is therefore investigated with a more refined model in the next sections.

A very good agreement between the measurements and the hybrid model predictions is observed in the frequency range in which the model is valid (above $140\,\mathrm{Hz}$). The oscillation

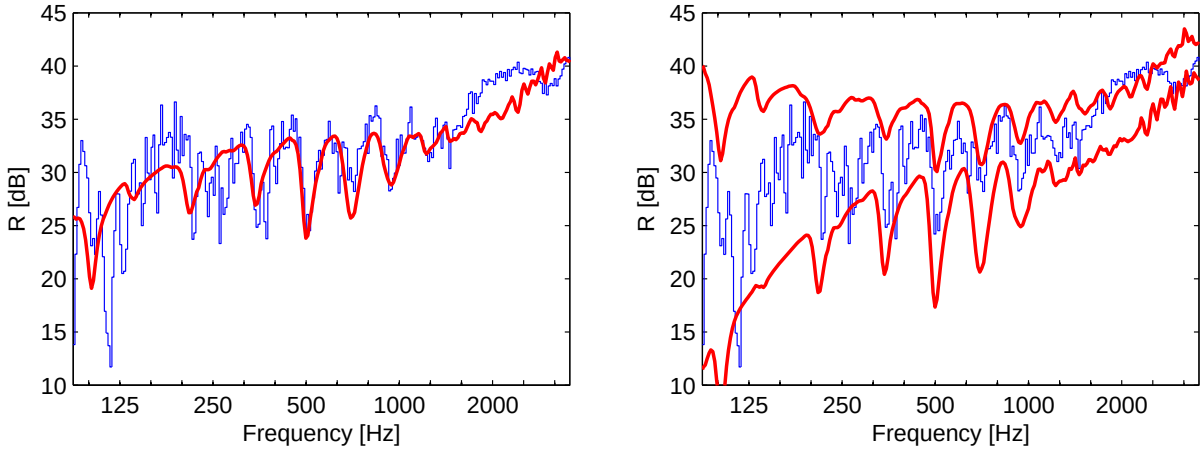



Figure 2: Measured sound reduction index of a rib-stiffened plate in 1/48-octave bands (blue) vs. the harmonic mean values (red, left figure) and 95 % confidence intervals (red, right figure) as predicted with the hybrid FE-SEA method.

of the transmission loss is very well predicted, and the measurements fall almost completely within the 95 % confidence interval which represents the uncertainty caused by the diffuse field assumption for the sound fields in the rooms.

## 4 SEMI-DISCRETE COINCIDENCE

The pronounced dips that are observed in both the measured and the predicted sound reduction index above 140 Hz correspond with some of the eigenfrequencies of the finite-sized rib-stiffened plate. They do not correspond with eigenfrequencies of the bare rooms, as these are not included in the hybrid FE-SEA model. In order to get more insight into why certain eigenfrequencies of the plate result in a pronounced reduction of the airborne sound insulation while others do not, both the in-vacuo modes of the rib-stiffened plate (as determined from the finite element model) and the acoustic dispersion curve are plotted in the frequency-wavenumber domain (Fig. 3). The structural wavenumber of mode $j$ was computed as the amplitude of $[k_{\mathrm{x}j}, k_{\mathrm{y}j}]$, where $k_{\mathrm{x}j}$ denotes the horizontal and $k_{\mathrm{y}j}$ the vertical wavenumber. $k_{\mathrm{x}j}$ and $k_{\mathrm{y}j}$ are determined through counting the number of approximate half sine wavelengths in the horizontal and vertical directions, respectively, of mode $j$.

This counting process was automated as follows. First, the correlation between a computed mode shape and all possible mode shapes that consist of between 0 and 15 half sine wavelengths in both horizontal directions was determined by computing the MAC value [14]. The combination that yielded the highest MAC value was then retained for determining the wavenumber corresponding to the mode shape. In Fig. 3, a different symbol is used for wavenumbers that are computed based on a high MAC value and those based on a low MAC value.

Subsequently, the frequency-wavenumber relation (also called dispersion curve) for sound waves was plotted in the same figure. It can be seen that this dispersion curve comes close to a mode of the rib-stiffened plate at 212 Hz, 345 Hz, 502 Hz, 693 Hz, and 964 Hz, which are precisely the frequencies at which the sound transmission loss of the plate dips. The dips can therefore be attributed to a semi-discrete coincidence phenomenon which involves only specific plate modes, namely, those modes for which the mode shapes exhibit one half wavelength in the horizontal direction. From around 1000 Hz, many plate modes match closely the acoustic wavelength at their resonance frequency, and pronounced dips in the sound reduction index are

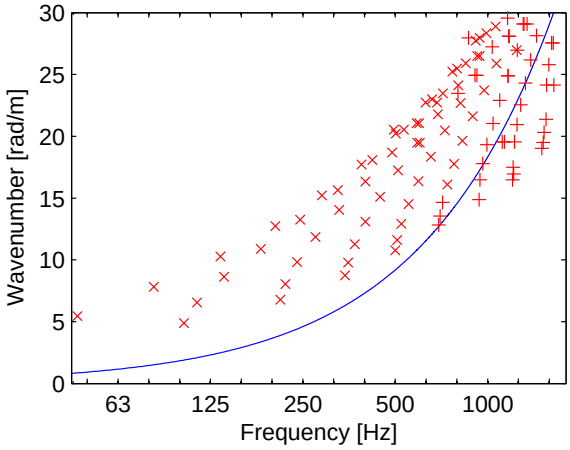



Figure 3: Modes of the rib-stiffened plate as predicted with the updated finite element model (crosses) in the frequency-wavenumber domain, together with the frequency-wavenumber relation for sound waves in an unbounded domain (solid lines). x: MAC value between the mode shape and the closest half-sines mode shape is higher than 0.7. +: MAC value is lower than 0.7.

no longer observed.

## 5 INFLUENCE OF MODE SUPPRESSION

In the light of this physical interpretation, it is instructive to investigate by how much the sound insulation of the plate improves when these five modes are suppressed. Therefore, the sound transmission loss of the plate is re-computed but without taking these five modes into account. The sound insulation improvement that is computed in this way is an upper bound, because in reality mode suppression measures - such as local stiffening, attachment of tuned mass dampers, and active vibration control - are never perfect.

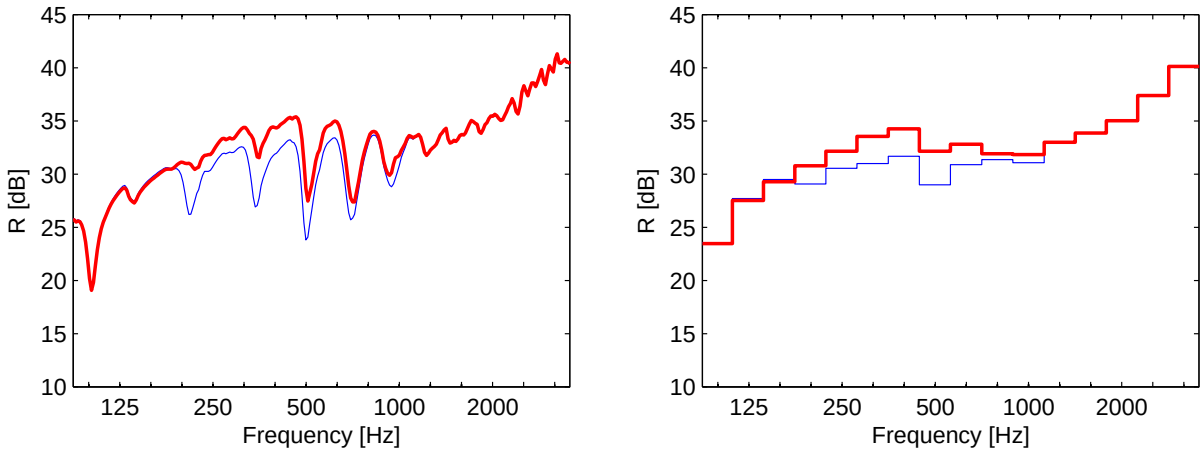



Figure 4: Predicted sound transmission loss of the rib-stiffened plate with (red) and without (blue) mode suppression. Left: harmonic values, right: 1/3 octave band averages.

Fig. 4 compares the computed sound transmission loss with and without mode suppression. Both harmonic values and one-third octave band averages are shown. Suppressing the modes with natural frequencies of $212\,\mathrm{Hz}$, $345\,\mathrm{Hz}$ has the largest influence: the dips in the transmission loss are largely reduced, resulting in an increase of the transmission loss with about $5\,\mathrm{dB}$ around

those frequencies. Suppressing the modes at $502\,\mathrm{Hz}$, $693\,\mathrm{Hz}$ and $964\,\mathrm{Hz}$ has a smaller effect. This can be understood from Fig. 3: around these frequencies, the bending wavelengths of several other modes also match the acoustic wavelength well, and as a result, suppressing only one of these modes is not effective anymore.

## 6 CONCLUSIONS

In this paper, valuable insight into the airborne sound insulation of finite-sized rib-stiffened plates was obtained by constructing a detailed finite element model of the structure and coupling this to diffuse field models for the adjoining rooms within the computationally efficient hybrid FE-SEA framework. For the specific rib-stiffened plate that was considered here, a pronounced oscillation in the narrow-band sound transmission loss was observed over a broad frequency range. The oscillation dips correspond to only a few natural frequencies of the plate, at which the wavelength of the corresponding mode shape is close to the acoustic wavelength, resulting in a semi-discrete coincidence phenomenon. This observation is important with respect to vibration control because within this broad frequency range, an increase in the airborne sound insulation can theoretically be achieved by suppressing only a few particular plate resonances.

## REFERENCES

[1] J.S. Anderson and M. Bratos-Anderson. Radiation efficiency of rectangular orthotropic plates. *Acta Acustica united with Acustica*, 91(1):61–76, 2005.

[2] J. Brunskog. Near-periodicity in acoustically excited stiffened plates and its influence on vibration, radiation and sound insulation. *Acta Acustica united with Acustica*, 90(2):301–312, 2004.

[3] F. Fahy and P. Gardonio. *Sound and structural vibration*. Academic Press, Oxford, UK, 2nd edition, 2007.

[4] L. Gagliardini, J. Roland, and J.L. Guyader. The use of a functional basis to calculate acoustic transmission between rooms. *Journal of Sound and Vibration*, 145(3):457–478, 1991.

[5] C. Guigou-Carter and M. Villot. Modelling of sound transmission through lightweight elements with stiffeners. *Building Acoustics*, 10(3):193–209, 2003.

[6] M. Heckl. Untersuchungen an orthotropen Platten. *Acustica*, 10:109–115, 1960.

[7] R. Josse and C. Lamure. Transmission du son par une paroi simple. *Acustica*, 14(5):266–280, 1964.

[8] R.S. Langley. Numerical evaluation of the acoustic radiation from planar structures with general baffle conditions using wavelets. *Journal of the Acoustical Society of America*, 121(2):766–777, 2007.

[9] R.S. Langley and J.A. Cordioli. Hybrid deterministic-statistic analysis of vibro-acoustic systems with domain couplings on statistical components. *Journal of Sound and Vibration*, 321(3–5):893–912, 2009.

[10] R.S. Langley and V. Cotoni. Response variance prediction for uncertain vibro-acoustic systems using a hybrid deterministic-statistical method. *Journal of the Acoustical Society of America*, 122(6):3445–3463, 2007.

[11] J.-H. Lee and J. Kim. Analysis of sound transmission through periodically stiffened panels by space-harmonic expansion method. *Journal of Sound and Vibration*, 251(2):349–366, 2002.

[12] G. Maidanik. Response of ribbed panels to reverberant acoustic fields. *Journal of the Acoustical Society of America*, 34(6):809–826, 1962.

[13] D.J. Mead. Plates with regular stiffening in acoustic media: vibration and radiation. *Journal of the Acoustical Society of America*, 88(1):391–401, 1990.

[14] E. Reynders. System identification methods for (operational) modal analysis: review and comparison. *Archives of Computational Methods in Engineering*, 19(1):51–124, 2012.

[15] E. Reynders, R.S. Langley, A. Dijckmans, and G. Vermeir. A hybrid finite element - statistical energy analysis approach to robust sound transmission modelling. *Journal of Sound and Vibration*, 333(19):4621–4636, 2014.

[16] P.J. Shorter and R.S. Langley. Vibro-acoustic analysis of complex systems. *Journal of Sound and Vibration*, 288(3):669–699, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SOLVING UNILATERAL CONTACT PROBLEMS IN MULTIBODY DYNAMICS USING A PRIMAL-DUAL INTERIOR POINT METHOD

**Dario Mangoni**[1], **Alessandro Tasora**[1]

[1]Università degli Studi di Parma
Parco Area delle Scienze 181/A 43124 Parma, Italy
e-mail: dario.mangoni@studenti.unipr.it, alessandro.tasora@unipr.it

**Keywords:** interior-point, multibody, contacts

**Abstract.** *Rigid unilateral contacts in multibody problems lead to non-smooth dynamics: this requires the development of new types of time integrators that draw on the solution of a variational inequality at each time step. In most settings, this variational inequality is a complementarity problem. In the context of non-smooth multibody dynamics, projected fixed-point, spectral and Nesterov methods have been successfully used to solve those complementarity problems, especially for granular problems or real-time simulations [2] [3], but large scale and ill-conditioned problems still advocate for solvers with better convergence. The promising class of interior-point method has been applied only recently to multibody dynamics [4]. Their implementation is more complicate, but this comes at the benefit of high precision and good convergence, moreover they can be applied to problems with both deformable and rigid bodies [5]. We present an algorithm based on a dual-primal interior point method. The solution is advanced along the central path using the Mehotra predictor-corrector scheme, in a small number of sub steps that leverage on a parallel direct solver for computing the updates from highly sparse saddle-point problems. Although the original method addresses unilateral constraints only, it is possible to restate it in order to fit also equality constraints, as might happen when introducing joints between colliding parts. The algorithm is implemented in C++ language in our open-source simulator. Results show that the method regularly converges to small residuals even in cases (for instance, odd mass ratios) that degraded the convergence of other methods.*

## 1 INTRODUCTION

Interior-Point algorithms are born to provide a solution for minimization problems, mainly linear and quadratic, that can arise from many different scenarios. In this paper we would show some improvements related to multibody simulations involving unilateral constraints (contacts) without friction. In this case the solver is called to solve a Quadratic Programming problem (QP from now on) that can be written as:

$$\min \quad q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T G\mathbf{x} + \mathbf{x}^T\mathbf{c} \tag{1a}$$

$$\text{s.t.} \quad A\mathbf{x} \geq b \tag{1b}$$

where $G \in \mathbb{R}^{n\times n}$ is a square, symmetric, positive definite matrix and $A \in \mathbb{R}^{m\times n}$ is the constraints matrix. Although we will show only the case when only inequality constraints hold, it is possible to restate the equality constraints to fit this formulation.

As demonstrated in [1], there are some necessary and sufficient conditions for which a vector $\mathbf{x}^*$ is solution (i.e. feasible minimizer) of the QP problem. These conditions, known as *Karush-Kuhn-Tucker conditions*, are here reported:

$$G\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} = 0 \tag{2a}$$

$$A\mathbf{x} - vectb \geq 0 \tag{2b}$$

$$(A\mathbf{x} - \mathbf{b}) \circ \boldsymbol{\lambda} = 0 \tag{2c}$$

$$\boldsymbol{\lambda} \geq 0 \tag{2d}$$

where the $\circ$ operator stands for the Hadamard (or component-wise) product. The condition 2c can also be seen as an orthogonality condition $\mathbf{y} \perp \boldsymbol{\lambda}$ where the operator $\perp$ is applied component-wise.
Introducing *slack* variables in $A\mathbf{x} - \mathbf{b} \geq 0$ let us move the inequality constraints over a variable (namely $\mathbf{y}$).

$$G\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} = 0 \tag{3a}$$

$$A\mathbf{x} - \mathbf{y} - \mathbf{b} = 0 \tag{3b}$$

$$\mathbf{y} \circ \boldsymbol{\lambda} = 0 \tag{3c}$$

$$\boldsymbol{\lambda} \geq 0 \tag{3d}$$

$$\mathbf{y} \geq 0 \tag{3e}$$

If a set of variables $(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ respects the conditions of eq.2 (or, that is the same, eq.3) then $\mathbf{x}$ is a minimizer for eq.1a that respects eq.1b i.e. is solution for the QP problem. This conditions in fact are not only necessary but, if $G$ is positive semidefinite as in this case, are also sufficient.

## 2 INTERIOR-POINT BASICS

Since the algorithm we are going to show is based on a *path-following* method we introduce the *perturbed* KKT system:

$$F(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}; \sigma) = \begin{bmatrix} G\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} \\ A\mathbf{x} - \mathbf{y} - \mathbf{b} \\ \mathbf{y} \circ \boldsymbol{\lambda} - \sigma\mu\mathbf{e} \end{bmatrix}_{(\mathbf{y},\boldsymbol{\lambda})\geq 0} \tag{4}$$

where $\mathbf{e} = [1, \dots, 1] \in \mathbb{R}^n$ and $\sigma$, the *centering* parameter and $\mu$ the *duality measure* that will define in a moment.

It is easy to observe that the problem stated in 3 can be summarized as $F(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}; 0) = 0$ so we can make the easy assessment that we can operate on such multivariate function using a Newton step (applied to a non-linear equation) in order to find the solution $(\mathbf{x}^*, \mathbf{y}^*, \boldsymbol{\lambda}^*)$ to bring its value down to zero. This operation, also called *affine scaling step*, gives the foundation to the Interior-Point algorithm.

As we will notice in section 2.1, pointing directly to the actual solution is *not* always the smartest option available. In fact, although the first step could provide a big reduction of $q(\mathbf{x})$, it also brings the approximated solution very close to the boundaries of the feasible region so that only very short steps can be taken in the following iteration. Because of this, the Newton step is made towards a different point that lays on the *central path*.

## 2.1 Basic Newton step

### 2.1.1 Newton step direction

Since we want $F(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}; 0) = 0$ the most trivial idea is to make a fixed-point iteration like:

$$F_{k+1} = F_k + J_{F_k} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\boldsymbol{\lambda} \end{bmatrix} = 0 \quad \Rightarrow \quad J_{F_k} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\boldsymbol{\lambda} \end{bmatrix} = -F_k \tag{5}$$

where $F_k = F(\mathbf{x}_k, \mathbf{y}_k, \boldsymbol{\lambda}_k; 0)$. Expanding the Jacobian $J_{F_k}$, that is valid also for the more generic $F_k = F(\mathbf{x}_k, \mathbf{y}_k, \boldsymbol{\lambda}_k; \sigma)$:

$$\begin{bmatrix} G & 0 & A^T \\ A & -I & 0 \\ 0 & \Lambda & Y \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\boldsymbol{\lambda} \end{bmatrix} = - \begin{bmatrix} G\mathbf{x}_k - A^T\boldsymbol{\lambda} + \mathbf{c} \\ A\mathbf{x}_k - \mathbf{y}_k - \mathbf{b} \\ \mathbf{y}_k \circ \boldsymbol{\lambda}_k - \sigma\mu\mathbf{e} \end{bmatrix} \tag{6}$$

since we assume that $\sigma$ and $\mu$ are, in any case, constant parameters. If $F$ were linear, the solution of 6 would have been also the solution of the original KKT system. Unfortunately this is not true with a QP problem, where

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{y}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \mathbf{y}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} + \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\boldsymbol{\lambda} \end{bmatrix} \tag{7}$$

is only an approximation. Now that we have the *directions* of the step, the step length(s) must be calculated too.

### 2.1.2 Newton step length

In order to determine the step length, we use the non-negativity constraints $\mathbf{y} \geq 0$ and $\boldsymbol{\lambda} \geq 0$ for which:

$$\begin{aligned} \mathbf{y}_{k+1} &= \mathbf{y}_k + \alpha_p^{prim} \Delta\mathbf{y}_k \geq 0 \\ \boldsymbol{\lambda}_{k+1} &= \boldsymbol{\lambda}_k + \alpha^{dual} \Delta\boldsymbol{\lambda}_k \geq 0 \end{aligned} \tag{8}$$

that gives, with simple math:

$$\begin{aligned} \alpha^{prim} &= \min\left(1, \eta \min_{i\,:\,\Delta\mathbf{y}<0} -\frac{\mathbf{y}_i}{\Delta\mathbf{y}_i}\right) \\ \alpha^{dual} &= \min\left(1, \eta \min_{i\,:\,\Delta\boldsymbol{\lambda}<0} -\frac{\boldsymbol{\lambda}_i}{\Delta\boldsymbol{\lambda}_i}\right) \end{aligned} \tag{9}$$

We notice another control parameter $\eta$. The standard version considers this value equal to 1, but we will use *also* another update method that improve the robustness. The variation is trivial: it simply takes the minimum step length between the two $\alpha$ parameters and adds that control parameter $\eta \in (0, 1)$ that approaches 1 as the iterations come closer to the solution. Also in this case the control parameter is used to soften the constraint.

### 2.2 The central path

Any iterative solver creates a sequence in where each element is progressively closer to the solution. However, the paths they follow are not unique and give different behaviours in term of convergence speed and robustness. We will focus on the most used one, the *central* path, that is characterized by softening the orthogonality condition while keeping the residuals $\mathbf{r}_p$ and $\mathbf{r}_d$ equal to zero.

Moreover this path is *strictly feasible* so that it follows the definition of strict feasibility given here:

$$\mathcal{F}_0 = \{(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \,|\, \mathbf{r}_p = 0, \mathbf{r}_d = 0, (\mathbf{y}, \boldsymbol{\lambda}) > 0\}$$

The equation of the KKT system that is not fully respected is the orthogonality constraint 3c that becomes:

$$\mathbf{y} \circ \boldsymbol{\lambda} = \tau \mathbf{e} \tag{10}$$

giving the *central path* definition:

$$\mathcal{C}(\tau) = \{(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda}) \in \mathcal{F}_0 \,|\, \mathbf{y} \circ \boldsymbol{\lambda} = \tau \mathbf{e}\} \tag{11}$$

Let imagine the $\mathbb{R}^m$ space in which $\mathbf{y} \circ \boldsymbol{\lambda}$ lays. The inequalities $\boldsymbol{\lambda} \geq 0$ and $\mathbf{y} \geq 0$ still hold so we are restricting our analysis only over the non-negative orthant. The path that equation 11 draws is simply its bisector; because of this is called *central*. As we notice, given $\tau$, the iteration keeps the approximated solution as far as possible from the boundaries. However, this step is not generally the best in terms of residual/error reduction and it will never bring the iteration to the actual solution of the problem unless $\tau$ is zero. Actually, this is just want we want to do: as soon as the algorithm advance, $\tau$ is progressively reduced to zero, so that the iteration points toward the actual solution of the KKT system.

Another derived problem is how to choose $\tau$. The concept is that it should tend to zero as the approximated solution is closer to the actual one. Since the only part of the error that is not reduced yet is the orthogonality condition, we could take this parameter to drive $\tau$. An useful criteria to understand how far the constraint hold for a given point is the *complementarity measure* (that has the same meaning as the *duality measure* in LP). Given a set $(\mathbf{x}, \mathbf{y}, \boldsymbol{\lambda})$ that satisfies both $\boldsymbol{\lambda} \geq 0$ and $\mathbf{y} \geq 0$ we can compute:

$$\mu = \frac{\mathbf{y}^T \boldsymbol{\lambda}}{m} \tag{12}$$

At the optimal solution the complementarity measure should equal to zero so it can be easily adopted to choose $\tau$

$$\tau = \sigma \mu \tag{13}$$

where $\sigma$ is called *centering* parameter and it is normally chosen as:

$$\sigma = \left( \frac{\mu_p}{\mu_c} \right)^3 \tag{14}$$

as we will describe later.

### 2.3 Predictor-Corrector Scheme

As we have already defined the concept of *central path* in section 2.2 and the basic Newton iteration in section 2.1 we can assemble the practical algorithm that has been implemented. A more in-depth review of starting point strategies is showed in section 2.4.

#### 2.3.1 Prediction

Basically the prediction phase is based on a bare Newton step on the original eq.3 in the form previously showed in 6 without any perturbation, i.e. $\sigma = 0$. We will name the unknown vector as $[\Delta\mathbf{x}_p \quad \Delta\mathbf{y}_p \quad \Delta\boldsymbol{\lambda}_p]^T$, omitting the subscript $k$ for the sake of clarity. The lengths of this step are obtained from 9 with $\eta = 0$. Considering the new approximation given by eq.8, we obtain the new value of the complementarity measure $\mu_p = \mathbf{y}_p^T\boldsymbol{\lambda}_p/m$. The centering parameter can now be evaluated by mean of the equation 14, remembering that the previous value $\mu$ comes directly from the previous iteration.

#### 2.3.2 Correction

From the prediction step we observe that the updated value of the orthogonality condition is, component-wise:

$$\begin{aligned}(\mathbf{y}_i + \Delta\mathbf{y}_i)(\boldsymbol{\lambda}_i + \Delta\boldsymbol{\lambda}_i) &= \mathbf{y}_i\boldsymbol{\lambda}_i + \mathbf{y}_i\Delta\boldsymbol{\lambda}_i + \boldsymbol{\lambda}_i\Delta\mathbf{y}_i + \Delta\boldsymbol{\lambda}_i\Delta\mathbf{y}_i \\ &= \Delta\boldsymbol{\lambda}_i\Delta\mathbf{y}_i\end{aligned} \tag{15}$$

instead of zero. The correction step takes into consideration this residual and aims to delete it, putting it on the right-hand side of the KKT system that now becomes:

$$\begin{bmatrix} G & 0 & A^T \\ A & -I & 0 \\ 0 & \Lambda & Y \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\mathbf{y} \\ \Delta\boldsymbol{\lambda} \end{bmatrix}_c = - \begin{bmatrix} G\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} \\ A\mathbf{x} - \mathbf{y} - \mathbf{b} \\ \mathbf{y} \circ \boldsymbol{\lambda} + \Delta\mathbf{y}_p \circ \Delta\boldsymbol{\lambda}_p - \sigma\mu\mathbf{e} \end{bmatrix} \tag{16}$$

where $\mathbf{y}$, $\boldsymbol{\lambda}$ and $\mathbf{x}$ comes from the previous iteration (so they are *not* the updated approximations from the prediction step). It is also true that the entire matrix on the left is kept untouched as well as the upper part of the right-hand side of the equation. So only the product $\Delta\mathbf{y}_p \circ \Delta\boldsymbol{\lambda}_p$ must be effectively evaluated.

As before these considerations are useful only to obtain the directions along which the step will be made. The length is evaluated using eq.9 with $\eta \in (0, 1)$ and taking as $\alpha = \min(\alpha^{prim}, \alpha^{dual})$. We used this simple equation to give adaptivity:

$$\eta = \exp(-0.1\mu) + 0.9 \tag{17}$$

The approximated solution is given by eq.8.

At this stage, we would like to update the upper part of the right-hand side of eq.6 — used both for the prediction and correction step — without expensive operations. We will notice that, calling:

$$\begin{aligned}\mathbf{r}_d &= G\mathbf{x} - A^T\boldsymbol{\lambda} + \mathbf{c} \\ \mathbf{r}_p &= A\mathbf{x} - \mathbf{y} - \mathbf{b}\end{aligned} \tag{18}$$

we have:

$$\begin{aligned}\mathbf{r}_{d_{k+1}} &= (1 - \alpha^{dual})\,\mathbf{r}_{d_k} + (\alpha^{prim} - \alpha^{dual})\,G\,\Delta\mathbf{x}_k \\ \mathbf{r}_{p_{k+1}} &= (1 - \alpha^{prim})\,\mathbf{r}_{p_k}\end{aligned} \tag{19}$$

Different formulations of eq.6 and of the more general eq.16 exists. One of the most useful is given by substituting $\Delta\mathbf{y} = A\Delta\mathbf{x} + \mathbf{r}_p$ in the last row of these systems, leading to:

$$\begin{bmatrix} G & A^T \\ A & \Lambda^{-1}Y \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\lambda} \end{bmatrix}_c = - \begin{bmatrix} \mathbf{r}_d \\ \mathbf{r}_p + \mathbf{y} - \Lambda^{-1}(\Delta\mathbf{y}_p \circ \Delta\boldsymbol{\lambda}_p - \sigma\mu\mathbf{e}) \end{bmatrix} \tag{20}$$

This *augmented* system will be extensively used in our calculations.

### 2.4 Starting point

One drawback of IP methods is their sensitivity to starting points. In fact, depending on where the algorithm starts, it could converge to the solution in few iterations or it could diverge as well, depending on the initial given value. Because of this particular behaviour the choice of a reliable starting point technique is one of the most critical. Different approaches are shown in the next sections and have been implemented in order to test their performances (see section 4).

#### 2.4.1 Nocedal's starting point

In [1, §16.6] the priority is given to moving away from the boundary of the non-negative orthant while trying to reduce primal-dual residuals and complementarity product. This is done by a classic affine/prediction step (that will compute $\Delta\mathbf{y}$ and $\Delta\boldsymbol{\lambda}$) that will drive the algorithm towards the solution and indirectly will also take the distance from the boundaries.

$$\begin{aligned} \mathbf{x}_0 &= [1 \dots 1]^T \\ \mathbf{y}_0 &= \max(1, |\mathbf{y} + \Delta\mathbf{y}|) \\ \boldsymbol{\lambda}_0 &= \max(1, |\boldsymbol{\lambda} + \Delta\boldsymbol{\lambda}|) \end{aligned} \tag{21}$$

However, also with trivial examples, this technique is not conclusive and some failures have been experienced in our tests. Initial values of $\mathbf{x}_0$, $\mathbf{y}_0$ and $\boldsymbol{\lambda}_0$ cannot be chosen randomly. In fact, changing their value of an order of magnitude could give divergence problems.

#### 2.4.2 D'Apuzzo's starting point

Another possibility is shown in [6]. This method is by far one of the less computational expensive and it grants robustness in a wide range of scenarios. This technique is based on the idea that the the neighbourhood in where the approximated solutions are kept is influenced by the initial residuals/complementarity gap ratio. Because of this, the starting point has to show not only small residuals and a small complementarity product, but they should be, at at the same time, comparable. In our case the strategy leads to this conclusion:

$$\begin{aligned} \mathbf{x}_0 &= [1 \dots 1]^T \\ \mathbf{y}_0 &= \max(1, A\mathbf{x} - \mathbf{b}) \\ \boldsymbol{\lambda}_0 &= 1/\mathbf{y}_0 \end{aligned} \tag{22}$$

where the $\max()$ operator is applied component-wise.

## 3 MULTI-BODY FORMULATION

In this section we will introduce a multi-body formulation for the non-smooth dynamics of a system subject to an arbitrary number of contacts between multiple parts. The notation in section 3.1 is intentionally kept similar to the references while in section 3.2 we will adapt symbols to the current IP notation.

### 3.1 Motion equation and constraints

The non-smooth nature of mechanical systems subject to contacts stems from the fact that contact constraints are unilateral, as the contact force cannot be negative. More in general, in case of frictional contacts, the contact force must satisfy some set-valued law, that cannot be translated in simple equality constraints: for instance for the Coloumb-Amontons friction, the contact force $\mathbf{F}_c$, expressed in the local contact reference must satisfy $\mathbf{F}_c \in \mathsf{F}_c \subset \mathbb{R}^3$ with $\mathsf{F}_c$ being a second-order Lorentz cone.

In sake of clarity, in this work we will deal with the simpler friction-less contact case. Therefore, the contact constraint can be expressed with the following Signorini complementarity condition at the position level:

$$F_{c,n} \geq 0 \quad \perp \quad \Phi_{c,n} \geq 0 \tag{23}$$

In the previous constraint, $F_{c,n}$ is the surface-normal component of the force, and $\Phi_{c,n}$ is the normal displacement; the $\perp$ complementarity symbol means:

$$\mathbf{a} \geq \mathbf{0} \quad \perp \quad \mathbf{b} \geq \mathbf{0} \quad \Leftrightarrow \quad a_i \geq 0, \quad b_i \geq 0, \quad a_i b_i = 0 \tag{24}$$

As we will see in the following, this introduces a major difficulty in the formulation, which requires the solution of complementarity problems at discrete time steps, instead of solving trivial linear problems as those required by classical smooth systems with bilateral constraints. In fact, when bilateral constraints are added to a system of ordinary equations (ODE), it turns into a Differential Algebraic Equation (DAE) problem. In our case, the addition of complementarity constraints lead to a Differential Variational Inequality (DVI).

Assume we have multiple contacts and multiple parts. Each $i-th$ part is a rigid body with position $\{\mathbf{x}_i, \rho_i\}$, using three dimensional translation vectors $\mathbf{x}_i \in \mathbb{R}^3$ and four-dimensional unit quaternions $\rho_i \in \mathbb{H}^1$ for rotation in $\mathsf{SO}(\mathbb{R}, 3)$. Velocity is represented by $\{\dot{\mathbf{x}}_i, \omega_i\}$, with $\omega \in \mathbb{R}^3$ angular velocity in local reference of the body. Each body has a mass $m_i$ and a tensor of inertia $I \in \mathbb{R}^{3x3}$.

We can arrange all the states in single vectors $\mathbf{q} = \{\mathbf{x}_1, \mathbf{q}_1, \mathbf{x}_2, \mathbf{q}_2, ..., \mathbf{x}_n, \mathbf{q}_n, \}$ and $\mathbf{v} = \{\dot{\mathbf{x}}_1, \omega_1, \dot{\mathbf{x}}_2, \omega_2, ..., \dot{\mathbf{x}}_n, \omega_n, \}$. Using simple quaternion algebra, there is a linear map relation between $d\mathbf{q}/dt$ and $\mathbf{v}$ as $d\mathbf{q}/dt = \Gamma(\mathbf{q}, \mathbf{v})$. Similarly, all masses can be assembled in a single matrix $M$, that is very sparse since it includes just $m$ and $I$ values along its diagonal.

Forces and torques can be applied to bodies, we collect them into a $\mathbf{f}_t = \mathbf{f}_e + \mathbf{f}_g$ term, which includes the $\mathbf{f}_g$ gyroscopic term and generic external loads $\mathbf{f}_e$.

For each $j$-th contact, we have a contact distance $\Phi_j$, where $j \in \mathcal{G}_\mathcal{A}$ given a set $\mathcal{G}_\mathcal{A}$ of active unilateral contacts. We make the assumption that for each system state, the distance $\Phi_j$ can be computed as a (smooth, differentiable) [1] function of state $\Phi_j = \Phi_j(\mathbf{q}, t)$, and we assume that for every integration time step an algorithm can update the $\mathcal{G}_u$ set by adding potential contact points that come within a distance threshold $\eta$. [2].

[1] This is not always the case when sharp and non-concave shapes come into contact, where some regularization or heuristic workarounds might be needed especially interpenetrating $\Phi_j < 0$, but the rest of the discussion still holds.

[2] This process is handled by the *collision engine*, and can be very CPU-intensive when shapes are numerous and complex. Also, defining a proper threshold $\eta$ is not trivial, but it is not discussed here.

We relate contact-orthogonal displacements to state changes via jacobians:

$$\begin{aligned} d\Phi_j(\mathbf{q},t)/dt &= [\partial\Phi_j/\partial\mathbf{q}]d\mathbf{q}/dt & (25)\\ &= \nabla_{\mathbf{q}}{\Phi_j}^T\Gamma(\mathbf{q},\mathbf{v}) & (26)\\ &= \nabla{\Phi_j}^T\mathbf{v} & (27)\end{aligned}$$

We remark that the $\nabla{\Phi_j}^T$ jacobian is a single-row matrix with high sparsity: since the $j$-th contact depends on two bodies only, the $\nabla{\Phi_j}^T$ jacobian is a row full of zeroes, with only two 1x6 non-zero blocks.

The contact-orthogonal force $\mathbf{F}_{c,j}$ can be expressed by introducing a $\hat{\gamma}_j$ scalar multiplier per each contact, thus in absolute coordinates it holds:

$$\mathbf{F}_{c,j} = \nabla\Phi_j\hat{\gamma}_j \quad (28)$$

The full system that expresses the mechanical model subject to unilateral constraints is the following DVI:

$$\begin{aligned} M\frac{d\mathbf{v}}{dt} &= \sum_{j\in\mathcal{G}_\mathcal{A}} \nabla\Phi_j\hat{\gamma}_j + \mathbf{f}_t(t,\mathbf{q},\mathbf{v}) & (29)\\ \Phi_j(\mathbf{q},t)\geq 0 \quad &\perp \quad \hat{\gamma}_j \geq 0, \quad j\in\mathcal{G}_\mathcal{A} & (30)\\ \dot{\mathbf{q}} &= \Gamma(\mathbf{q},\mathbf{v}) & (31)\end{aligned}$$

The time integration of the model above can be performed using a stabilized discrete time-stepping method, as presented in [2]. This requires the introduction of a time step $h$ to advance the state from $\mathbf{q}^{(l)}, \mathbf{v}^{(l)}$ to $\mathbf{q}^{(l+1)}, \mathbf{v}^{(l+1)}$. The DVI problem is discretized at the velocity-level, in terms of measures to include impulsive effects if any, so we define $\gamma_j = h\hat{\gamma}_j$ and we write the following system where the unknowns are the changes in velocities and the contact impulses $\gamma_j$:

$$\begin{aligned} M(\mathbf{v}^{(l+1)} - \mathbf{v}^{(l)}) &= \sum_{j\in\mathcal{G}_\mathcal{A}} \nabla\Phi_j\gamma_j + h\mathbf{f}_t(t,\mathbf{q},\mathbf{v}) & (32)\\ \frac{1}{h}\Phi_j(\mathbf{q},t) + \nabla{\Phi_j}^T\mathbf{v} \geq 0 \quad &\perp \quad \gamma_j \geq 0, \quad j\in\mathcal{G}_\mathcal{A} & (33)\\ & & (34)\end{aligned}$$

followed by an update $\mathbf{q}^{(l+1)} = \boldsymbol{\lambda}(\mathbf{q}^{(l)}, \mathbf{v}^{(l+1)}, h)$.

We note that the system above can be expressed as a (mixed) Linear Complementarity System (M-LCP), using a compact matrix notation:

$$\begin{aligned} \begin{bmatrix} M & D \\ D^T & 0 \end{bmatrix}\begin{bmatrix} \mathbf{v}^{(l+1)} \\ \boldsymbol{\gamma} \end{bmatrix} - \begin{bmatrix} h\mathbf{f}_t + M\mathbf{v}^{(l)} \\ -\frac{1}{h}\boldsymbol{\Phi}(\mathbf{q},t) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{C} \end{bmatrix} & (35)\\ \mathbf{C}\geq\mathbf{0}\perp\boldsymbol{\gamma}\geq\mathbf{0} & (36)\end{aligned}$$

where $D$ is a sparse matrix that contains all the jacobians $\nabla\Phi_j$ aligned column-by-column, as $D = \{\nabla\Phi_1, \nabla\Phi_2, ..., \nabla\Phi_{n_c}\}$. Also $\boldsymbol{\Phi} = \{\Phi_1, \Phi_2, ..., \Phi_{n_c}\}^T$ and $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, ..., \gamma_{n_c}\}^T$.

This M-LCP is also the Karush-Kuhn-Tucker optimality condition of a quadratic program with convex (linear, unilateral) constraints, a class of problems that have been solved successfully with Interior-Point algorithms.

We note that, if the constraints were bilateral, as happens in revolute joints, mating constraints etc., the MLCP would turn into a simple linear problem of the type:

$$\begin{bmatrix} M & D \\ D^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v}^{(l+1)} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} h\mathbf{f}_t + M\mathbf{v}^{(l)} \\ -\frac{1}{h}\boldsymbol{\Phi}(\mathbf{q}, t) \end{bmatrix} \tag{37}$$

as often met in the solution of ODEs and DAEs in smooth multibody dynamics. In such special case, the discussed integration scheme would turn into a linearized implicit Euler integration scheme. The difficulty, in our more general case, is the fact that the solution of a MLCP is by far more complex than the solution of a linear system.

### 3.2 Fitting the MB system in Interior-Point algorithm

We would now provide another perspective from which the system 35 can be observed. Its solution can be thought as the minimizer of the QP problem (as stated for example in 3) considering:

$$\begin{aligned} \mathbf{x} &\Leftrightarrow \mathbf{v}^{(l+1)} \\ \mathbf{c} &\Leftrightarrow -\left(h\mathbf{f}_t + M\mathbf{v}^{(l)}\right) \\ \mathbf{b} &\Leftrightarrow -\frac{1}{h}\boldsymbol{\Phi}(\mathbf{q}, t) \\ \mathbf{y} &\Leftrightarrow \mathbf{C} \\ \boldsymbol{\lambda} &\Leftrightarrow \boldsymbol{\gamma} \\ G &\Leftrightarrow M \\ A &\Leftrightarrow D^T \end{aligned}$$

Because of this strict similarity, the multibody problem formulated in 35 can be solved through the IP algorithm described in the previous sections.

## 4 RESULTS

The algorithm presented in section 2 has been implemented in C++ language in order to be tested in a multibody dynamic simulation software, Chrono::Engine. The hardware on which these tests are performed consists in an Intel i7-4710HQ processor, 2x4GB DDR3-1600K RAM, 1TB 2.5" SATA-300 5400RPM. The software is compiled for x64 architecture with MSVC compiler with full optimization. The Intel MKL Pardiso direct linear solver is used for the solution of the linear system showed in 20. The Interior-Point algorithm exits whenever the residual norms and the complementarity product are below a threshold that in the cases below are set as:

$$\frac{\|\mathbf{r}_p\|}{m} < 1 \times 10^{-8} \tag{38a}$$

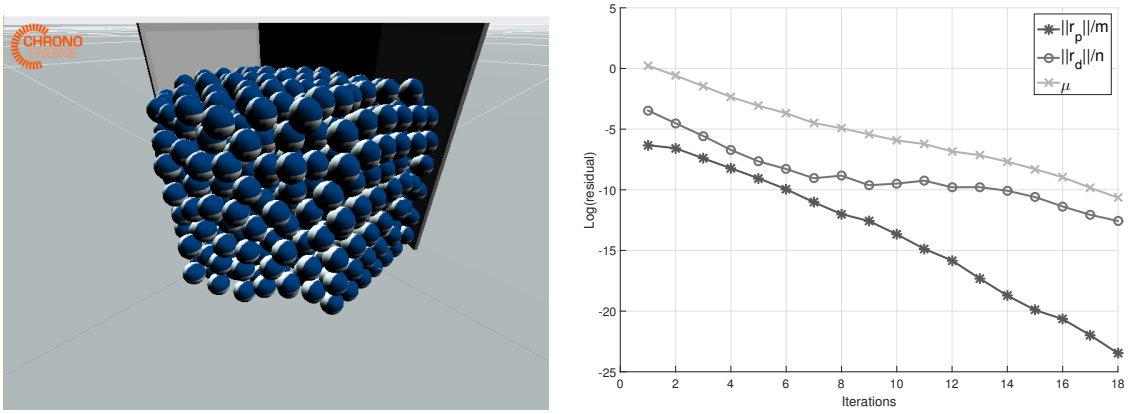
$$\frac{\|\mathbf{r}_d\|}{n} < 1 \times 10^{-8} \tag{38b}$$

$$\mu < 1 \times 10^{-7} \tag{38c}$$

### 4.1 Convergence and speed results

The first test involves unilateral contacts of spheres that drop inside a box. As we can see in table 1 increasing the number of bodies leads to a higher number of degrees of freedom (bigger mass/stiffness matrices) and higher number of contacts/constraints (bigger Jacobian matrices).

However, the test demonstrate that the number of iterations needed to get the approximated solution is almost invariant. Obviously the increased size negatively affects the time and computational expense of the linear solver.




(a) *Ball in the box test.* (b) *Convergence result.*

Figure 1: Balls falling in a box.

| Matrix size | Mean iterations |
|---|---|
| 13800 | 18.9 |
| 4500 | 19.4 |
| 3200 | 18.6 |
| 1200 | 16.7 |

Table 1: Iterations needed at each timestep for different matrix sizes (Ball in the box example).

### 4.2 Warm start

In many different case-scenario it would be quite useful if the solver could take advantage of warm start. In fact it could happen that the problem matrices are quite similar between consecutive timesteps so it would be obvious to re-use the solution of the previous step as starting point of the following. This choice often improves the convergence speed of many iterative solvers and it would be desirable if also IP methods could manifest this behaviour.

In order to test if this feature is natively supported by the proposed algorithm we set up a couple of trivial examples. The first one is a brick supported by four contact points. Although this problem can be easily solved statically we tested it in a dynamic simulation. This means that the collision detection algorithm will slightly change the constraint equations at each timestep while keeping the mass matrix unchanged.
In the second case the brick is substituted with a ball so that the contacts are no more redundant and the solution is unique.

Classic starting point algorithms usually reset the initial vectors to some default value (usually 1) that are further manipulated in order to reduce the ratio between residuals and complementarity product like in [6] or in other used to make a step away from the non-negative orthant boundary like in [1, §16.6]. Adapting these methods to use a trivial warm start is quite straightforward although it cannot be considered a state-of-the-art solution. The tests performed aim

to point out if, at least in a controlled environment, the IP algorithm could take advantage of a solution recycle.

It has to be underlined that, in order to reuse the full solutions from previous timesteps, the matrix must have kept the same dimension and moreover the constraint equations should not have undergone big modifications. However, this is not always true especially when the problem involves unilateral constraint so, in such cases, a *partial* warm start has been applied. In fact, although the Jacobian matrix $A$ has suffered not negligible changes, the mass/stiffness matrix $G$ is usually kept unchanged and because of this, the $\mathbf{x}$ vector keeps the same dimension and it can be warm started. In this case $\mathbf{y}$ and $\boldsymbol{\lambda}$ have to be reinitialized.

| Test (IP calls) | None | Partial WarmStart | Warm Start |
|---|---|---|---|
| Balls in Box (50) | 935 | 927 | 747 |
| Single Brick (100) | 499 | 370 | 81 |
| Single Ball (100) | 720 | 515 | 90 |

Table 2: IP iterations for solving unilateral contact problems.

### 4.3 Redundant constraints and softened contacts

Formally the IP method needs a full-rank constraint matrix $A$ i.e. the Jacobian. Unfortunately it is a common situation to have redundant constrained bodies for which the Jacobian is rank-deficient and no unique solution exists. Under those circumstances the matrix that is composed to make a Newton step is more prone to become singular as soon as the components of $\mathbf{y}/\boldsymbol{\lambda}$ tend to zero. We investigated this case and we present here some interesting results.

Before trying classical workarounds let us consider how the IP is approaching the contact problem. As we can see in 20 the $\mathbf{y}/\boldsymbol{\lambda}$ vector occupies the diagonal of the south-east corner of the augmented matrix, so that the Lagrangian multipliers i.e. the forces that acts on the bodies will affect the constraints equations. Trying to give a physical meaning to this phenomena we can easily come to the conclusion that these elements acts as contact springs i.e. they give a reaction that tends to push away the bodies that are going to collide. In some way we can say that the hard constraints that were permitted by the DVI approach are now softened, introducing a spring behaviour to the contacts. However, this is also an interesting feature considering the numerical issue given by redundant constraints: the non-zero, although small, elements that are inserted yield to a better conditioned matrix, adding the required linear independence to the possibly rank-deficient $A$. This approach permitted to get a reliable solution also in such cases where the Jacobian matrix was highy compromised. Equipping the IP algorithm with a direct solver is, in such cases, the best choice since they are less affected by the ill-conditioning of the matrix.

Moreover, it has been tested another useful feature in which a compliance matrix is added in place of the diagonal $\Lambda^{-1}Y$ so that, when the algorithm converges, the matrix still has a non-zero diagonal matrix in its south-east corner:

$$\begin{bmatrix} G & A^T \\ A & \Lambda^{-1}Y + E \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\lambda} \end{bmatrix}_c = - \begin{bmatrix} \mathbf{r}_d \\ \mathbf{r}_p + \mathbf{y} - \boldsymbol{\Lambda}^{-1}(\Delta\mathbf{y}_p \circ \Delta\boldsymbol{\lambda}_p - \sigma\mu\mathbf{e}) \end{bmatrix} \tag{39}$$

where $E$ is a diagonal matrix with elements $1/k$ with $k$ stiffness of the contact spring.

The more robust workaround of removing redundant lines with SVD or QR revealing methods is, in our tests, not needed since the IP methods always converged with the required accuracy to the solution.

## REFERENCES

[1] J. Nocedal, S. Wright, *Numerical Optimization, 2nd Edition*, Springer New York, 2006.

[2] A. Tasora, M. Anitescu, A Matrix-Free Cone Complementarity Approach for Solving Large-Scale, Nonsmooth, Rigid Body Dynamics *Computer Methods in Applied Mechanics and Engineering*, **200.5**, 439–453, 2011.

[3] T. Heyn, et al. Using Krylov Subspace and Spectral Methods for Solving Complementarity Problems in ManyBody Contact Dynamics Simulation *International Journal for Numerical Methods in Engineering*, **95.7**, 541–561, 2013.

[4] J. Klinert, S. Bernd, M. Obermayr An Inexact Interior Point Method for the Large-Scale Simulation of Granular Material *Computer Methods in Applied Mechanics and Engineering*, **278**, 567–598, 2014.

[5] R. Kučera et al., An Interior-Point Algorithm for the Minimization Arising from 3D Contact Problems with Friction *Optimization Methods and Software*, **28.6**, 1195–1217, 2013.

[6] M. D'Apuzzo, V. De Simone, D. Di Serafino, Starting-Point Strategies for an Infeasible Potential Reduction Method, *Optimization Letters*, **4.1**, 131–146, Springer, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# 3D DRBEM MODELING FOR ROTATING INITIALLY STRESSED ANISOTROPIC FUNCTIONALLY GRADED PIEZOELECTRIC PLATES

**Mohamed Abdelsabour Fahmy*, †**

* Jummum University College
Umm Al-Qura University
Alazizya, behind Alsalam Souq, 21955 Makkah, Saudi Arabia.
e-mail: maselim@uqu.edu.sa, Web page: https://uqu.edu.sa/staff/ar/4340548

† Faculty of Computers and Informatics
Suez Canal University
New Campus, 4.5 Km, Ring Road, El Salam District, 41522 Ismailia, Egypt.
e-mail: Mohamed_fahmy@ci.suez.edu.eg
Web page: http://mohamed_fahmy_ci.staff.scuegypt.edu.eg

**Keywords:** Rotation, Initial Stress, Anisotropic, Functionally Graded Piezoelectric Plates, Dual Reciprocity Boundary Element Method.

***Abstract.*** Functionally graded piezoelectric plates (FGPPs) considered in the current study have been received much attention in recent years due to their various applications including sensors and actuators, aerospace, nuclear energy, chemical plant, electronics, biomaterials, piezoelectric motors, reduction of vibrations and noise, infertility treatment, ultrasonic micromotors, micropumps and microvalves, photovoltaics.

The time-stepping dual reciprocity boundary element method modeling was proposed to study the 3D dynamic response of an anisotropic rotating initially stressed functionally graded piezoelectric plate (FGPP). The FGPP is assumed to be graded through the thickness. The main aim of this paper is to evaluate the effects of initial stress and rotation on the displacement components in an anisotropic FGPP. Then we studied the effect of inhomogeneity on the displacement components in presence of initial stress and rotation. In the end, the accuracy of the proposed method was examined and confirmed by comparing the obtained results with those known previously.

## 1. Introduction

Functionally graded piezoelectric plates (FGPPs) considered in the current study have been received much attention in recent years due to their various applications including sensors and actuators, aerospace, nuclear energy, chemical plant, electronics, biomaterials, piezoelectric motors, reduction of vibrations and noise, infertility treatment, ultrasonic micromotors, micropumps and microvalves, photovoltaics. Since it is very difficult to find the analytical solution to the considered problem, therefore, an important number of engineering and mathematical papers devoted to the numerical solution have been studied to describe the global behavior of such problems.

Piezoelectric ceramics are being widely used in electromechanical devices such as sensors, filters, ultrasonic generators and actuators because they offer excellent coupling properties between the mechanical and electrical fields of these devices, and the fracture of these piezoelectric materials has therefore been receiving a great deal of attention (Suo et al. [1], Pan [2], Jin and Zhong, [3], Zhang et al. [4], Lin et al. [5], Fang et al. [6], Abd-Alla and Al-Sheikh [7, 8], Kuna [9], Zaman et al. [10], Zhong et al. [11], Akbarzadeh et al. [12], Davì and Milazzo [13], Abd-Alla et al. [14], Abd-Alla and Askar [15], Alibeigloo and Liew [16] and Rafiee et al. [17]). It is well known that extension of the current fundamental fracture concepts and criteria from pure elasticity to piezoelectricity is not straightforward because of the coupling between the mechanical and electric fields.

The advantages in the boundary element method (BEM) arises from the fact that the BEM can be regarded as boundary–based method that uses the boundary integral equation formulations where only the boundary of the domain of the partial differential equation (PDE) is required to be meshed. But in the domain-based methods such the finite element method (FEM), finite difference method (FDM) and element free method (EFM) that use ordinary differential equation (ODE) or PDE formulations, where the whole domain of the PDE requires discretisation. Thus the dimension of the problem is effectively reduced by one, that is, surfaces for three–dimensional (3D) problems or curves for two-dimensional (2D) problems. And the equation governing the infinite domain is reduced to an equation over the finite boundary. Also, the BEM can be applied along with the other domain-based methods to verify the solutions to the problems that do not have available analytical solutions. Presence of domain integrals in the formulation of the BEM dramatically decreases the efficiency of this

technique. One of the most frequently used techniques for converting the domain integral into a boundary one is the so-called dual reciprocity boundary element method (DRBEM). This method was initially developed by Nardini and Brebbia [18] in the context of two-dimensional (2D) elastodynamics and has been extended to deal with a variety of problems wherein the domain integral may account for linear-nonlinear static-dynamic effects. A more extensive historical review and applications of dual reciprocity boundary element method may be found in Brebbia et al. [19], Wrobel and Brebbia [20], Partridge and Brebbia, [21], Partridge and Wrobel [22], Frijns et al. [23], Gaul [24] and Fahmy [25-34].

In this paper the governing equations of an anisotropic FGPP under the influence of gravity are solved by means of a time-stepping dual reciprocity boundary element method (DRBEM) to describe the displacement behavior of the homogeneous and functionally graded plates in an anisotropic FGPP under the influence of gravity. The accuracy of the proposed method was examined and confirmed by comparing the obtained results with those known previously.

## 2 FORMULATION OF THE PROBLEM

The governing equations for the stress wave propagation in anisotropic functionally graded piezoelectric plate may be written in the following form

$$\sigma_{fg,f} = \Gamma_{fg} + \rho(x+1)^m \ddot{u}_g + \rho(x+1)^m \omega^2 x_g \tag{1}$$

$$D_{f,f} = 0 \tag{2}$$

In which

$$\sigma_{fg} = (x+1)^m \left[C_{fghi}\varepsilon_{hi} - e_{ifg}E_i\right] \tag{3}$$

$$D_f = (x+1)^m \left[e_{fhi}\varepsilon_{hi} + \epsilon_{fi}E_i\right] \tag{4}$$

$$\Gamma_{fg} = P_0(x+1)^m \left(u_{g,f} - u_{f,g}\right) \tag{5}$$

where

$$\varepsilon_{hi} = \frac{1}{2}\left(u_{h,i} + u_{i,h}\right) \quad , \quad E_i = -\Phi_{,i} \tag{6}$$

Substituting (3)-(5) into (1) and (2) leads to

$$\left[C_{fghi}u_{h,if} + e_{ifg}\Phi_{,if}\right] + \frac{m}{x+1}\left[C_{fghi}u_{h,i} + e_{ifg}\Phi_{,i}\right] = \rho\ddot{u}_g$$
$$+P_0\left(u_{g,f} - u_{f,g}\right) + \rho\omega^2 x_g \tag{7}$$

$$\left[e_{fhi}u_{h,if} - \epsilon_{fi}\Phi_{,if}\right] + \frac{m}{x+1}\left[e_{fhi}u_{h,i} - \epsilon_{fi}\Phi_{,i}\right] = 0 \tag{8}$$

Where $E_i$ is the electric field vector, $\Phi$ is the electric potential, $D$ is the electric displacement, $C_{fghi}$ is the elasticity tensor $\left(C_{fghi} = C_{gfhi} = C_{hifg}\right)$, $e_{ifg}$ is the piezoelectric tensor $\left(e_{ifg} = e_{igf}\right)$, $\epsilon_{fi}$ is the permittivity tensor $\left(\epsilon_{fi} = \epsilon_{if}\right)$, $\rho$ is the density and $\aleph = \frac{m}{x+1}$

A superposed dot denotes differentiation with respect to the time and a comma followed by a subscript denotes partial differentiation with respect to the corresponding coordinates.

The governing equations (7) and (8) can be written in the following form

$$C_{fghi}u_{h,if} = \rho\ddot{u}_g + P_0\left(u_{g,f} - u_{f,g}\right) + \rho\omega^2 x_g - C_{fghi}\aleph u_{h,i} - e_{ifg}\left[\Phi_{,if} + \aleph\Phi_{,i}\right] \tag{9}$$

$$e_{fhi}u_{h,if} = \epsilon_{fi}\Phi_{,if} - e_{fhi}\aleph u_{h,i} + \epsilon_{fi}\left[\Phi_{,if} + \aleph\Phi_{,i}\right] \tag{10}$$

## 3 NUMERICAL IMPLEMENTATION

The governing equations (9) and (10) can now be written in operator form as follows

$$L_{fg}u_h = b_{fg} \tag{11}$$

$$L_{fh}\Phi = b_{fh} \tag{12}$$

Where

$$L_{fg} = C_{fghi}\frac{\partial}{x_i}\frac{\partial}{x_f} \tag{13}$$

$$b_{fg} = \rho\ddot{u}_g + P_0\left(u_{g,f} - u_{f,g}\right) + \rho\omega^2 x_g - C_{fghi}\aleph u_{h,i} - e_{ifg}\left[\Phi_{,if} + \aleph\Phi_{,i}\right] \tag{14}$$

$$L_{fh} = e_{fhi}\frac{\partial}{x_i}\frac{\partial}{x_f} \tag{15}$$

$$b_{fh} = \epsilon_{fi}\Phi_{,if} - e_{fhi}\aleph u_{h,i} + \epsilon_{fi}\left[\Phi_{,if} + \aleph\Phi_{,i}\right] \tag{16}$$

It is convenient to use the contracted notation to introduce generalized piezoelectric vectors and tensors, which contain corresponding elastic and electric variables as follows:

$$U_H = \begin{cases} u_h & h = H = 1,2,3 \\ \Phi & H = 4 \end{cases} \tag{17}$$

$$Z_G = \begin{cases} t_g & g = G = 1,2,3 \\ q & G = 4 \end{cases} \tag{18}$$

$$C_{fGHi} = \begin{cases} C_{fghi}, & g = G = 1,2,3; h = H = 1,2,3 \\ e_{ifg}, & g = G = 1,2,3; H = 4 \\ e_{fhi}, & G = 4; \quad h = H = 1,2,3 \\ -\epsilon_{fi}, & G = 4; \quad H = 4 \end{cases} \tag{19}$$

Using the following Kronecker delta representation

$$\delta_{GH} = \begin{cases} \delta_{gh} & g = G = 1,2,3, \quad k = K = 1,2,3 \\ 0 & otherwise \end{cases} \tag{20}$$

The governing equations can be combined as follows:

$$L_{\mathrm{GH}} U_{\mathrm{H}} = \rho \delta_{GH} \ddot{U}_{\mathrm{H}} - \mathrm{B_G} \tag{21}$$

where

$$L_{\mathrm{GH}} = C_{f\mathrm{GH}i} \frac{\partial}{x_i} \frac{\partial}{x_f}$$

$$\mathrm{B_G} = \begin{cases} -\rho \ddot{u}_g - \mathrm{P}_0 \left( u_{g,f} - u_{f,g} \right) - \rho \omega^2 x_g + C_{fghi} \bar{\bar{u}} + \mathrm{e}_{ifg} \bar{\bar{\Phi}}, & g = G = 1,2,3 \\ \epsilon_{fi} \Phi_{,\mathrm{i}f} - \mathrm{e}_{fhi} \bar{\bar{u}} + \epsilon_{fi} \bar{\bar{\Phi}}, & G = 4 \end{cases}$$

$$\bar{\bar{u}} = \aleph u_{h,\mathrm{i}} \quad , \qquad \bar{\bar{\Phi}} = \Phi_{,if} + \aleph \Phi_{,i}$$

Now, we choose the fundamental solution $U^*_{MH}$ as weighting function as follows

$$L_{GH} U^*_{MH}(x, \xi) = -\delta_{GM} \delta(x, \xi) \tag{22}$$

The weighted residual formula is integrated by parts twice to obtain the following piezoelectric reciprocity relation

$$\int_R \left( L_{\mathrm{GH}} \mathrm{U}_H U^*_{MG} - L_{\mathrm{GH}} U^*_{MH} \mathrm{U_G} \right) dR = \int_\Gamma \left( U^*_{MG} \mathrm{Z_G} - \mathrm{Z}^*_{MG} \mathrm{U_G} \right) d\Gamma \tag{23}$$

where

$$\mathrm{Z}^*_{MG} = C_{f\mathrm{GH}i} U^*_{MH,i} n_f$$

Making use of the sifting property, we obtain from equation (23) the piezoelectric representation formula

$$U_{\mathrm{M}}(\xi) = \int_\Gamma \left( U^*_{MG} \mathrm{Z_G} - \mathrm{Z}^*_{MG} \mathrm{U_G} \right) d\Gamma - \int_R U^*_{MG} (\rho \delta_{GH} \ddot{U}_{\mathrm{H}} - \mathrm{B_G}) dR \tag{24}$$

The DRBEM is employed in equation (24) to transform the domain integrals into boundary integrals, hence we may deduce the following piezoelectric dual reciprocity representation formula

$$U_H(\xi) = \int_\Gamma \left( U^*_{HG} \mathrm{Z_G} - \mathrm{Z}^*_{HG} \mathrm{U_G} \right) d\Gamma$$

$$+ \sum_{q=1}^{N} \left( U^q_{HN}(\xi) + \int_\Gamma \left( T^*_{HG} U^q_{GN} - \mathrm{U}^*_{HG} T^q_{GN} \right) d\Gamma \right) \alpha^{\mathrm{q}}_{\mathrm{N}} \tag{25}$$

Now the source term in equation (24) is approximated by a series of tensor functions $f^{\mathrm{q}}_{GN}$ and unknown coefficients $\alpha^{\mathrm{q}}_{\mathrm{N}}$ as follows

$$\rho\delta_{GH}\ddot{U}_{\mathrm{H}} - \mathrm{B_G} \approx \sum_{q=1}^{N} f_{GN}^{\mathrm{q}}\alpha_{\mathrm{N}}^{\mathrm{q}} \tag{26}$$

Using the thin plate splines (TPS) as in Fahmy [27], we can write the particular solution of the displacement as follows

$$\mathrm{U_{GN}^{q}} = \begin{cases} -\dfrac{4}{\lambda^4}\left[K_0(\lambda r) + \log(r) - \dfrac{r^2\log r}{\lambda^2} - \dfrac{4}{\lambda^4}\right], & r > 0 \\ \dfrac{4}{\lambda^4}\left[\gamma + \log\left(\dfrac{\lambda}{2}\right)\right] - \dfrac{4}{\lambda^4}\,, & r = 0 \end{cases} \tag{27}$$

where $K_0$ is the Bessel function of the third kind of order zero and $\gamma = 0.5772156649015328$, which is known as Euler's constant, $r = \|x - \xi\|$ is the Euclidean distance between the field point $x$ and the load point $\xi$.

Hence, the traction particular solution $T_{GN}^{q}$ and the source function $f_{GN}^{\mathrm{q}}$ can be obtained by evaluating

$$T_{GN}^{q} = C_{f\mathrm{GH}i}U_{HN,i}^{q}n_f\,,\;\; L_{\mathrm{GH}}U_{HN}^{q} = f_{GN}^{\mathrm{q}} \tag{28}$$

According to Fahmy [35-38] and Fahmy et al. [39-42], the dual reciprocity boundary integral equation (25) can be written in the following system of equations

$$\zeta\breve{U}(t) - \eta\breve{\mathrm{T}}(\mathrm{t}) = \left(\zeta\breve{\mathrm{U}}(t) - \eta\breve{\mathrm{t}}(\mathrm{t})\right)\alpha(t) \tag{29}$$

where $\zeta$, $\eta$ are BEM system matrices, $\breve{U}$, $\breve{\mathrm{T}}$ contain the nodal values of the generalized displacements and fluxes, and $\breve{\mathrm{U}}$, $\breve{\mathrm{t}}$ contain the particular solutions

$$\begin{bmatrix} \rho\breve{\ddot{u}} - \rho\breve{b} \\ 0 \end{bmatrix} = \begin{bmatrix} F_{11} & F_{14} \\ F_{41} & F_{44} \end{bmatrix}\begin{bmatrix} \alpha_1 \\ \alpha_4 \end{bmatrix} \tag{30}$$

Now, the piezoelectric dual reciprocity representation formula can be written as

$$C_{HG}(\xi)\mathrm{U_G}(\xi) + \mathrm{CPV}\int_{\Gamma} \mathrm{Z}_{HG}^{*}\mathrm{U_G}\,d\Gamma = \int_{\Gamma} U_{HG}^{*}\mathrm{Z_G}\,d\Gamma \tag{31}$$

Where CPV means the Cauchy Principal Value.

The coefficient vector $\alpha_{\mathrm{s}}(\mathrm{t})$ can be calculated by setting up a system of N equations from (26) using the point collocation procedure, which yield

$$\rho\breve{\ddot{u}} - \rho\breve{b} = \mathcal{F}_{11}\alpha_{\mathrm{s}}(\mathrm{t}) \tag{32}$$

where

$$\mathcal{F}_{11} = F_{11} - F_{14}F_{44}^{-1}F_{41} \tag{33}$$

$$\alpha_s(t) = \mathcal{F}^{-1}\left(\rho\ddot{\breve{U}}(t) - \breve{B}(t)\right) \tag{34}$$

$$\mathcal{F}^{-1} = \begin{cases} \mathcal{F}_{11}^{-1} & s = 1 \\ -\mathcal{F}_{44}^{-1}F_{41}\mathcal{F}_{11}^{-1} & s = 4 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

Substitution of (34) into (29) yields the system

$$M\ddot{\breve{U}}(t) + \zeta\breve{U}(t) = \eta\breve{T}(t) + \breve{\mathcal{B}}(t) \tag{36}$$

where the volume matrix V, piezoelectric mass matrix $M$ and source vector $\breve{\mathcal{B}}(t)$ are as follows:

$$V = \left(\eta\check{\tau}(t) - \zeta\breve{\mathfrak{V}}(t)\right)\mathcal{F}^{-1}, \quad M = \rho V, \quad \breve{\mathcal{B}}(t) = V\breve{B}(t). \tag{37}$$

In order to solve system (36), the nodal vectors are subdivided into known and unknown parts denoted by the superscripts k and u.

$$\{U^k, T^u\} \in \Gamma_u, \qquad \{U^u, T^k\} \in \Gamma_T \tag{38}$$

The following matrix equation is obtained from Eq. (36).

$$\begin{bmatrix} M^{11} & M^{12} \\ M^{21} & M^{22} \end{bmatrix}\begin{bmatrix} \ddot{U}^k \\ \ddot{U}^u \end{bmatrix} + \begin{bmatrix} K^{11} & K^{12} \\ K^{21} & K^{22} \end{bmatrix}\begin{bmatrix} U^k(t) \\ U^u(t) \end{bmatrix} = \begin{bmatrix} \eta^{11} & \eta^{12} \\ \eta^{21} & \eta^{22} \end{bmatrix}\begin{bmatrix} T^k(t) \\ T^u(t) \end{bmatrix} + \begin{bmatrix} \mathcal{B}^1(t) \\ \mathcal{B}^2(t) \end{bmatrix} \tag{39}$$

The unknown fluxes $T^u(t)$ are obtained from the first row of matrix equation (24) and are expressed as follows.

$$T^u(t) = (\eta^{12})^{-1}\big[M^{11}\ddot{U}^k(t) + M^{12}\ddot{U}^u(t) + K^{11}U^k(t) + K^{11}U^k(t) + K^{12}U^u(t) - \eta^{11}T^k(t) - \mathcal{B}^1(t)\big] \tag{40}$$

Making use of Eq. (40), we can write the second row of matrix equation (39) as follows

$$M^u\ddot{U}^u(t) + K^uU^u(t) = Q^k(t) \tag{41}$$

where

$$Q^k(t) = \bar{\mathcal{B}}^k(t) + \eta^kT^k(t) - M^k\ddot{U}^k(t) - K^kU^k(t)$$

$$M^u = M^{22} - \eta^{22}(\eta^{12})^{-1}M^{12}$$

$$M^k = M^{21} - \eta^{22}(\eta^{12})^{-1}M^{11}$$

$$K^u = K^{22} - \eta^{22}(\eta^{12})^{-1}K^{12}$$

$$K^k = K^{21} - \eta^{22}(\eta^{12})^{-1}K^{11}$$

$$\eta^k = \eta^{21} - \eta^{22}(\eta^{12})^{-1}\eta^{11}$$

$$\bar{\mathcal{B}}^k(t) = \mathcal{B}^2(t) - \eta^{22}(\eta^{12})^{-1}\mathcal{B}^1(t)$$

We now split the system (41) into elastic and electric parts as follows:

$$\begin{bmatrix} M_{uu}^{u} & 0 \\ M_{\varphi u}^{u} & 0 \end{bmatrix}\begin{bmatrix} \ddot{u}^{u}(t) \\ \ddot{\varphi}^{u}(t) \end{bmatrix}+\begin{bmatrix} \mathrm{K}_{uu}^{u} & \mathrm{K}_{u\varphi}^{u} \\ \mathrm{K}_{\varphi u}^{u} & \mathrm{K}_{\varphi\varphi}^{u} \end{bmatrix}\begin{bmatrix} \mathrm{u}^{u}(\mathrm{t}) \\ \varphi^{u}(\mathrm{t}) \end{bmatrix}=\begin{bmatrix} \mathrm{Q}_{u}^{\mathrm{k}}(\mathrm{t}) \\ \mathrm{Q}_{\varphi}^{\mathrm{k}}(\mathrm{t}) \end{bmatrix} \tag{42}$$

The unknown electric potential $\varphi^{\mathrm{u}}$ can be obtained from the second row of Eq. (42) as follows

$$\varphi^{\mathrm{u}}(t)=\left(K_{\varphi\varphi}^{u}\right)^{-1}\left[\mathrm{Q}_{\varphi}^{\mathrm{k}}(\mathrm{t})-M_{\varphi u}^{u}\ddot{u}^{u}(t)-\mathrm{K}_{\varphi u}^{u}\mathrm{u}^{u}(\mathrm{t})\right] \tag{43}$$

With the use of Eq. (43) into the first row of Eq. (42) we obtain

$$\bar{M}^{u}\ddot{u}^{u}(t)+\bar{K}^{u}u^{u}(t)=\bar{Q}^{k}(t) \tag{44}$$

**where**

$$\bar{Q}^{k}(t)=Q_{u}^{k}(t)-\mathrm{K}_{u\varphi}^{u}\left(\mathrm{K}_{\varphi\varphi}^{u}\right)^{-1}\mathrm{Q}_{\varphi}^{\mathrm{k}}(\mathrm{t})$$

$$\bar{M}^{u}=M_{uu}^{u}-\mathrm{K}_{u\varphi}^{u}\left(\mathrm{K}_{\varphi\varphi}^{u}\right)^{-1}M_{\varphi u}^{u}$$

$$\bar{K}^{u}=K_{uu}^{u}-\mathrm{K}_{u\varphi}^{u}\left(\mathrm{K}_{\varphi\varphi}^{u}\right)^{-1}K_{\varphi u}^{u}$$

Now, writing equation (44) for the n+1th time step

$$\bar{M}^{u}\ddot{u}_{\mathrm{n+1}}^{u}+\bar{K}^{u}u_{\mathrm{n+1}}^{u}=\bar{Q}_{\mathrm{n+1}}^{\mathrm{k}} \tag{45}$$

where

$$\bar{Q}_{\mathrm{n+1}}^{\mathrm{k}}=\bar{\mathcal{B}}_{\mathrm{n+1}}^{\mathrm{k}}+\eta^{\mathrm{k}}\mathrm{T}_{\mathrm{n+1}}^{\mathrm{k}}-M^{k}\ddot{u}_{\mathrm{n+1}}^{\mathrm{k}}-K^{k}u_{\mathrm{n+1}}^{\mathrm{k}} \tag{46}$$

The Newmark time integration algorithm was used to reduce the system of ordinary differential equations (45) to an algebraic system. The displacements $u_{n+1}$ and velocities $\dot{u}_{\mathrm{n+1}}$ used in this algorithm are approximated at time step n+1 as follows:

$$\dot{u}_{\mathrm{n+1}}\approx\dot{u}_{\mathrm{n}}+[(1-\delta)\ddot{u}_{\mathrm{n}}+\delta\ddot{u}_{\mathrm{n+1}}]\Delta\mathrm{t} \tag{47}$$

$$u_{n+1}\approx u_{n}+\dot{u}_{\mathrm{n}}\Delta\mathrm{t}+\left[\left(\frac{1}{2}-\alpha\right)\ddot{u}_{\mathrm{n}}+\alpha\ddot{u}_{\mathrm{n+1}}\right]\Delta\mathrm{t}^{2} \tag{48}$$

The acceleration at time step n+1 may be expressed from Equation (48) as:

$$\ddot{u}_{\mathrm{n+1}}\approx\frac{1}{\alpha\Delta\mathrm{t}^{2}}(u_{n+1}-u_{n})-\frac{1}{\alpha\Delta\mathrm{t}}\dot{u}_{\mathrm{n}}-\left(\frac{1}{2\alpha}-1\right)\ddot{u}_{\mathrm{n}} \tag{49}$$

Upon substitution of (49) into (45) we obtain the following algebraic system

$$\mathbb{R}u_{\mathrm{n+1}}^{\mathrm{u}}=\mathcal{M}_{\mathrm{n+1}} \tag{50}$$

where the stiffness matrix $\mathbb{R}$ and effective load vector $\mathcal{M}_{\mathrm{n+1}}$ are given by

$$\mathbb{R}=\frac{1}{\alpha\Delta\mathrm{t}^{2}}\bar{M}^{u}+\bar{K}^{u} \tag{51}$$

$$\mathcal{M}_{\mathrm{n+1}}=\bar{Q}_{\mathrm{n+1}}^{\mathrm{k}}+\bar{M}^{u}\left[\frac{1}{\alpha\Delta\mathrm{t}^{2}}u_{n}^{u}+\frac{1}{\alpha\Delta\mathrm{t}}\dot{u}_{n}^{u}+\left(\frac{1}{2\alpha}-1\right)\ddot{u}_{n}^{u}\right] \tag{52}$$

Once we have solved (50) for the unknown displacements at time step n+1, we can compute the accelerations and velocities from equations (49) and (47) respectively. Finally, the electric potential $\varphi^{\mathrm{u}}(t)$ can be obtained from (43) and the unknown generalized tractions $\mathrm{T}^{u}(\mathrm{t})$ can be determined using equation (40).

## 4. Numerical results and discussion

With the view of illustrating the numerical results calculated by method presented in this paper, the material chosen for the plate is the piezoelectric ceramic Lead Zirconate Titanate (PZT), and the physical data for which is given as follows:

The elasticity tensor $C$ , piezoelectric tensor $e$ and relative permittivity $\epsilon^{\mathrm{rel}}$

$$C_{fghi} = \begin{pmatrix} 107.6 & 63.10 & 63.90 & 0.000 & 0.000 & 0.000 \\ 63.10 & 107.6 & 63.90 & 0.000 & 0.000 & 0.000 \\ 63.90 & 63.90 & 100.4 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 19.60 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 19.60 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 22.20 \end{pmatrix} \tag{53}$$

$$e = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 12.0 & 0.00 \\ 0.00 & 0.00 & 0.00 & 12.0 & 0.00 & 0.00 \\ -9.6 & -9.6 & 15.1 & 0.00 & 0.00 & 0.00 \end{pmatrix} \tag{54}$$

$$\epsilon^{\mathrm{rel}} = \begin{pmatrix} 1936 & 0.00 & 0.00 \\ 0.00 & 1936 & 0.00 \\ 0.00 & 0.00 & 2109 \end{pmatrix} \tag{55}$$

The present work should be applicable to any anisotropic FGPP deformation problem. The application is for purpose of illustration; we do not intend to validate the results in a quantitative way because we have no experimental data at hand; this may be justified because our objective is to introduce a viable numerical technique for studying a model rather than to study any physical behaviors of it. Such a technique was discussed in Fahmy [43] who solved the special case from this study in 2D in the absence of rotation and initial stress. To achieve better efficiency than the technique described in Fahmy [43], we use thin plate splines into a code, which is proposed in the current study. We extend the study of Fahmy [43], to solve 3D in the presence of rotation and initial stress. Thus, it is perhaps not surprising that the numerical values obtained here are in excellent agreement with those obtained by Fahmy [43].

The results of calculations are presented in Figs. 2–4. Comparison of the results is presented graphically for the following different cases: the solid line represents the solution in absence of initial stress ($\mathrm{P} = 0$) and rotation ($\omega = 0$), the dashed line represents the solution in the

presence of initial stress $(P = 0.5)$ and in the absence of rotation $(\omega = 0)$, the dotted line represents the solution in the absence of initial s tress $(P = 0)$ and in the presence of rotation $(\omega = 0.5)$ and the dashed-dotted line represents the solution in presence of initial stress $(P = 0.5)$ and rotation $(\omega = 0.5)$. It can be seen from these figures that the effect of initial stress and rotation is very pronounced.

To evaluate the influence of the inhomogeneity on the displacements in an anisotropic FGPP, the inhomogeneity parameter is taken to be $m = 0$ for the homogeneous plate, we assume that $m = 1$ for functionally graded plate. The computed results are presented graphically in Figs. 5-7, the figures show the difference between homogeneous and functionally graded plates.
In the special case under consideration, the results are plotted in Figs. 8–10 to show the validity of the DRBEM. These results obtained with the DRBEM have been compared graphically with those obtained using the Meshless Local Petrov–Galerkin (MLPG) method of Stanak et al. [45] and also the results obtained using the Peano-Baker Series (PBS) method of Liu et al. [46] are shown graphically in the same figures to confirm the validity of the proposed method. It can be seen from these figures that the DRBEM results are in excellent agreement with the results obtained by MLPG and PBS methods, thus confirming the accuracy of the DRBEM.

**ACKNOWLEDGEMENT**

The author would like to express his deepest gratitude to the Institute of Consulting Research and Studies (ICRS) at Umm Al-Qura University, Makkah, Saudi Arabia, for supporting and encouraging him during the development of this paper.

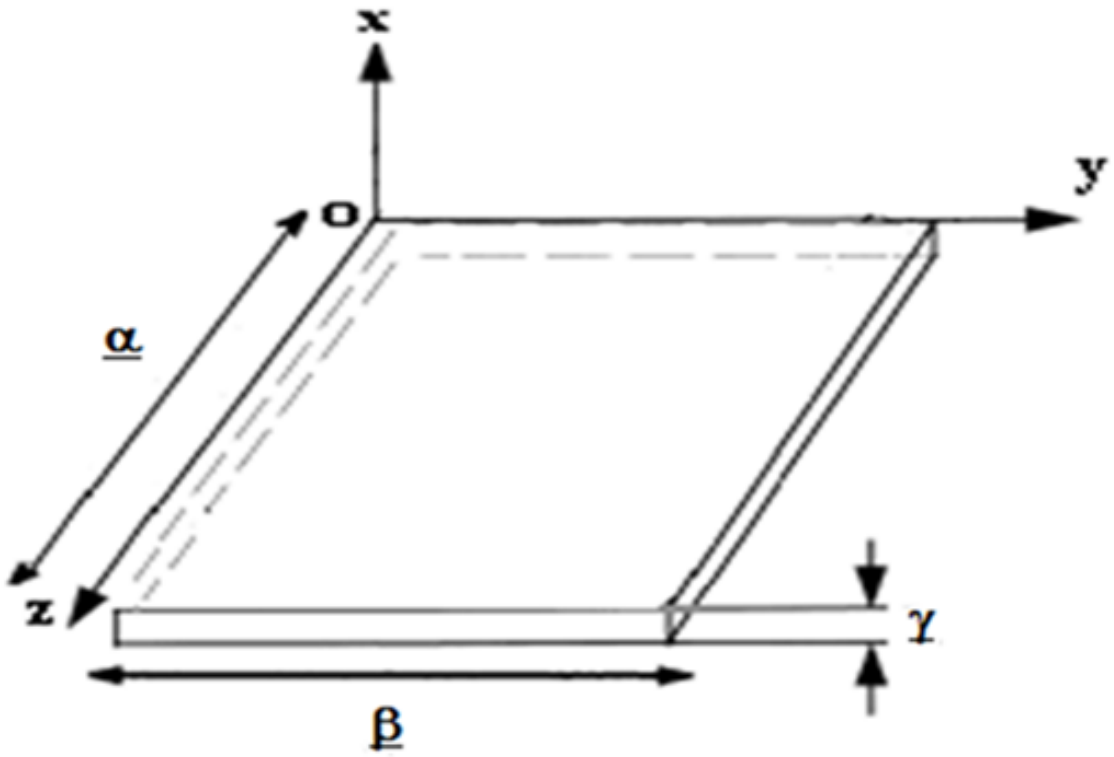



**Fig. 1. The coordinate system of the FGPP.**

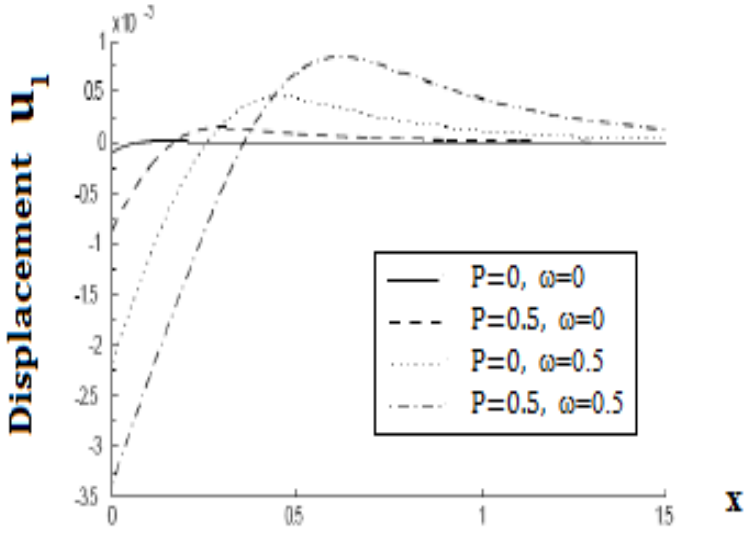



**Fig. 2. Variation of the displacement $u_1$ with x coordinate.**

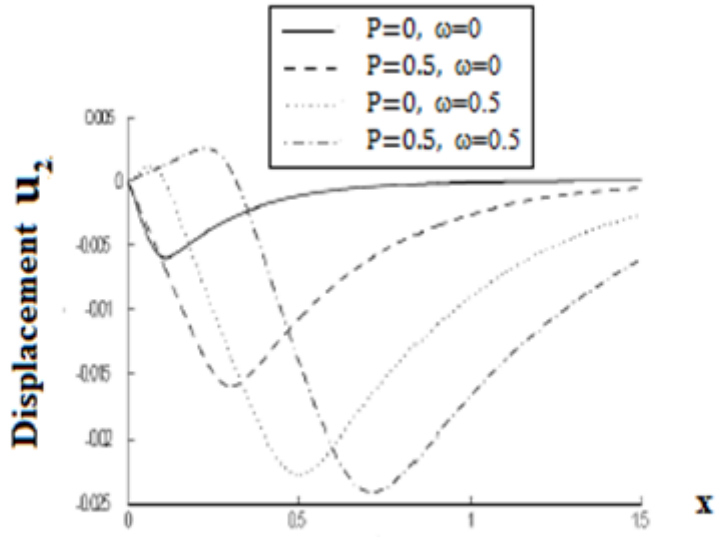



Fig. 3. Variation of the displacement $u_2$ with x coordinate.

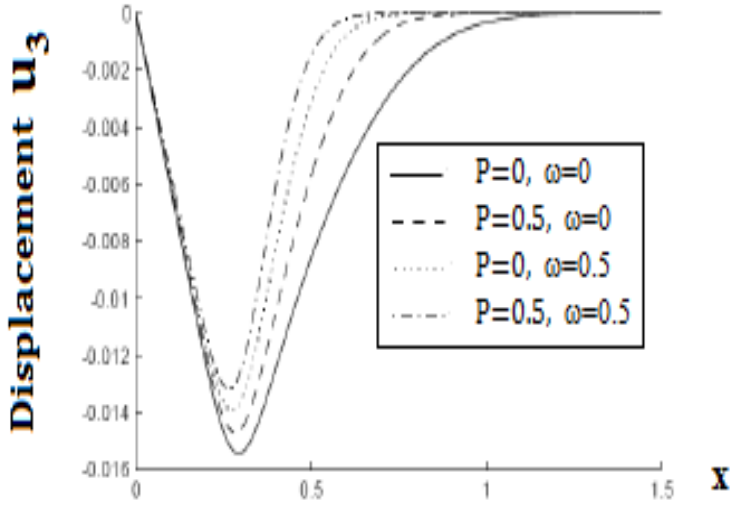



Fig. 4. Variation of the displacement $u_3$ with x coordinate.

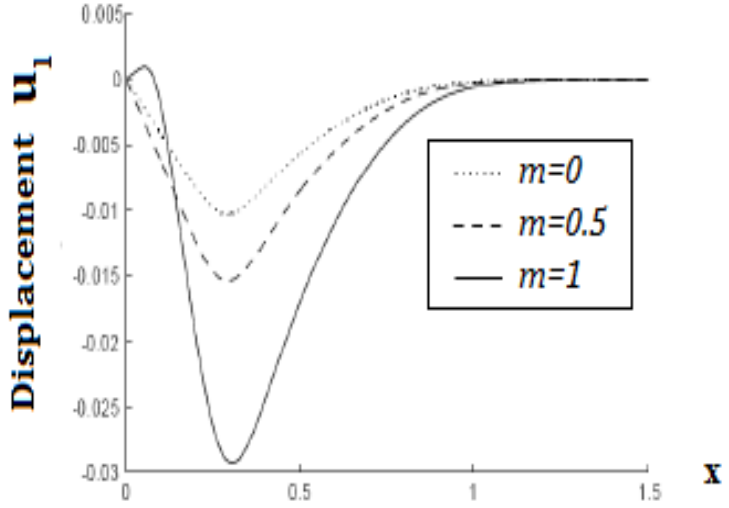



**Fig. 5. Variation of the displacement $u_1$ with x coordinate.**

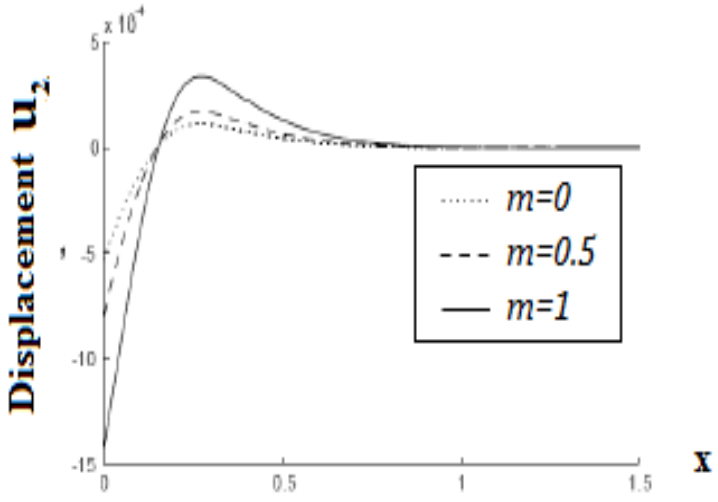



**Fig. 6. Variation of the displacement $u_2$ with x coordinate.**

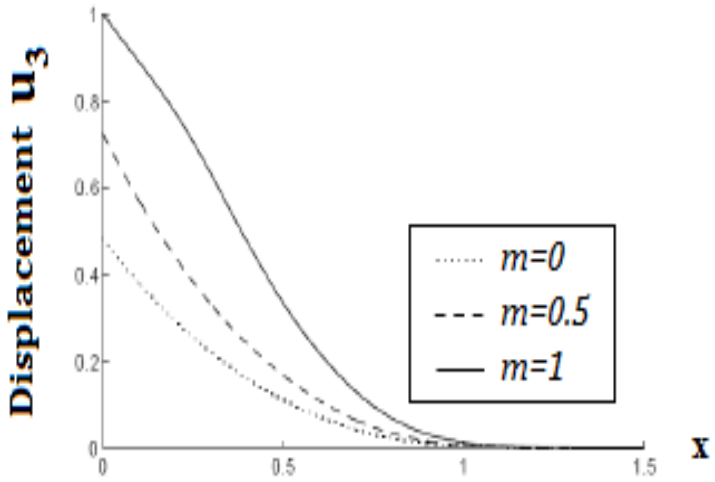



Fig. 7. Variation of the displacement $u_3$ with x coordinate.

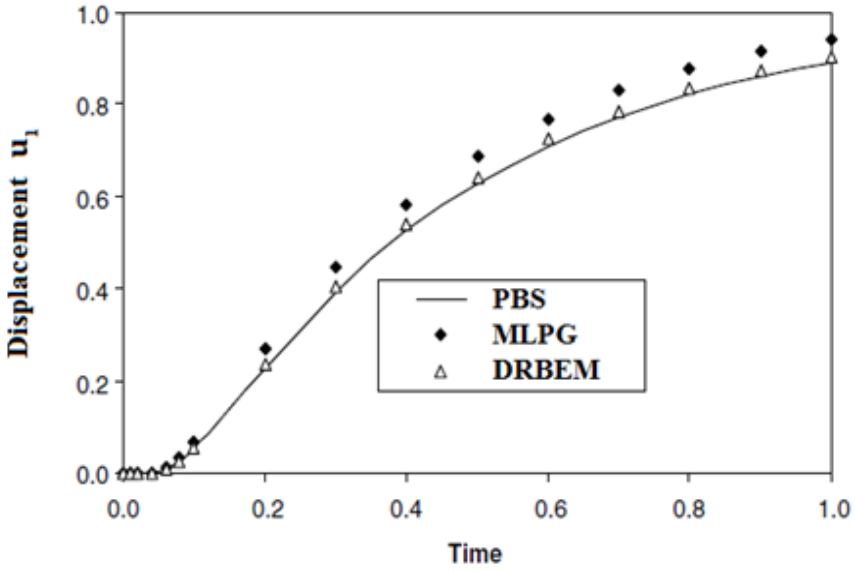



Fig. 8. Variation of the Displacement $u_1$ with time for three methods.

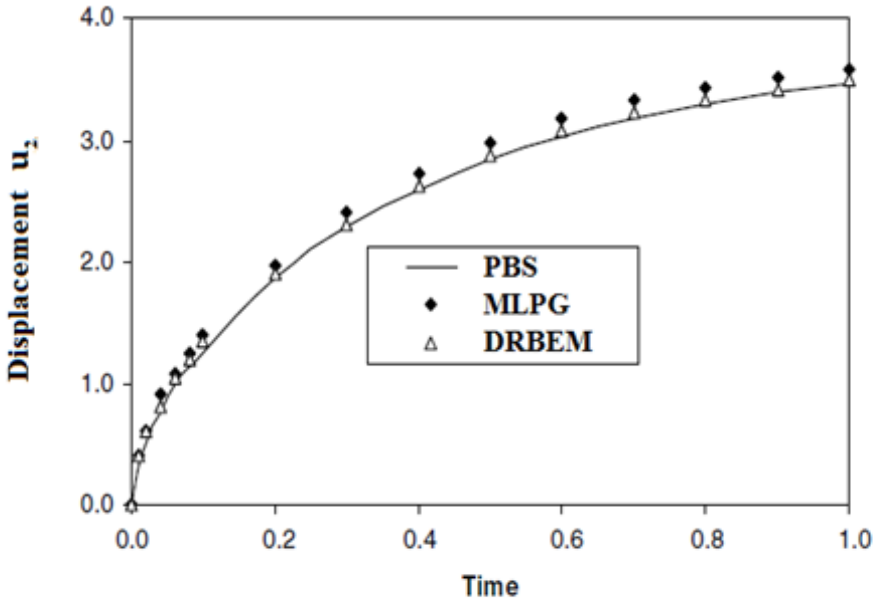



**Fig. 9. Variation of the displacement $u_2$ with time for three methods.**

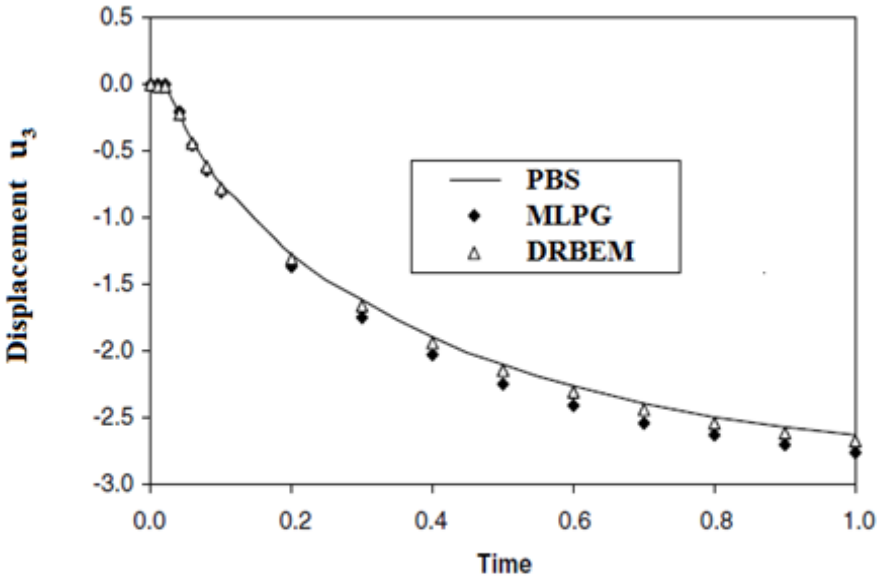



**Fig. 10. Variation of the displacement $u_3$ with time for three methods.**

## REFERENCES


[1] Z. Suo, C.M. Kuo, D.M. Barnett, J.R. Willis, Fracture mechanics for piezoelectric ceramics. Journal of the Mechanics and Physics of Solids, 40, 739–65, 1992.

[2] E.A. Pan, BEM analysis of fracture mechanics in 2D anisotropic piezoelectric solids. Engineering Analysis with Boundary Elements, 23, 67–76, 1999.

[3] B. Jin, Z. Zhong, A moving mode-III crack in functionally graded piezoelectric material: permeable problem," Mechanics Research Communications, 29, 217–24, 2002.

[4] T.Y. Zhang, M.H. Zhao, P. Tong, Fracture of piezoelectric ceramics. Advances in Applied Mechanics, 38, 147–289, 2002.

[5] S. Lin, F. Narita, Y. Shindo [2003] "Electroelastic analysis of a penny-shaped crack in a piezoelectric ceramic under mode I loading. Mechanics Research Communications 30, 371–86, 2003.

[6] D.N. Fang, Y.P. Wan, A.K. Soh, Magnetoelastic fracture of soft ferromagnetic Materials. Theoretical and Applied Fracture Mechanics, 42, 317–34, 2004.

[7] A.N. Abd-Alla, F.A. Alsheikh [2009a] "Reflection and refraction of plane quasi-longitudinal waves at an interface of two piezoelectric media under initial stresses," Archive of Applied Mechanics 79, 843–857, 2009.

[8] A.N. Abd-Alla, F.A. Al-sheikh [2009b] "The effect of the initial stresses on the reflection and transmission of plane quasi-vertical transverse waves in piezoelectric materials," World Academy of Science, Engineering and Technology 50, 660-668, 2009.

[9] M. Kuna [2010] "Fracture mechanics of piezoelectric materials – where are we right now?" Eng Fract Mech 77(2), 309–26, 2010.

[10] M. Zaman, Z. Yan, L. Jiang [2010] " Thermal effect on the bending behavior of curved functionally graded piezoelectric actuators," International Journal of Applied Mechanics 02(04), 787-805, 2010.

[11] X.C. Zhong, K.S. Zhang, Electroelastic analysis of an electrically dielectric Grif!th crack in a piezoelectric layer. International Journal of Engineering Science, 48, 612–23, 2010.

[12] A.H. Akbarzadeh, M.H. Babaei and Z.T. Chen, Thermopiezoelectric analysis of a functionally graded piezoelectric medium. International Journal of Applied Mechanics, 3, 47-68, 2011.

[13] G. Davì, A. Milazzo, A regular variational boundary model for free vibrations of magneto-electro-elastic structures. Engineering Analysis with Boundary Elements, 35, 303-312, 2011.

[14] A.N. Abd-Alla, F.A. Al-Sheikh, A.Y. Al-Hossain, The reflection phenomena of quasi-vertical transverse waves in piezoelectric medium under initial stresses. Meccanica, 47, 731-744, 2012.

[15] A.N. Abd-Alla, N.A. Askar, Calculation of bulk acoustic wave propagation velocities in trigonal piezoelectric smart materials. Applied Mathematics and Information Sciences 8, 1-8, 2014.

[16] A. Alibeigloo, K.M. Liew [2015] "Elasticity Solution of Free Vibration and Bending Behavior of Functionally Graded Carbon Nanotube-Reinforced Composite Beam with

Thin Piezoelectric Layers Using Differential Quadrature Method," International Journal of Applied Mechanics 7(1), 1550002, 2015.

[17] M. Rafiee, X.Q. He, S. Mareishi and K.M. Liew, Nonlinear Response of Piezoelectric Nanocomposite Plates: Large Deflection, Post-Buckling and Large Amplitude Vibration. International Journal of Applied Mechanics, 7, 1550074, 2015.

[18] D. Nardini, C.A. Brebbia [1982] "A new approach to free vibration analysis using boundary elements, in: C. A. Brebbia (Eds.), Boundary elements in engineering, Springer, Berlin, 312-326, 1982.

[19] C.A. Brebbia, J.C.F. Telles and L. Wrobel, Boundary element techniques in Engineering. (Springer-Verlag, New York), 1984.

[20] L.C. Wrobel, C.A. Brebbia, The dual reciprocity boundary element formulation for nonlinear diffusion problems. Computer Methods in Applied Mechanics and Engineering, 65, 147-164, 1987.

[21] P.W. Partridge, C.A. Brebbia [1990] "Computer implementation of the BEM dual reciprocity method for the solution of general field equations. Communications in Applied Numerical Methods, 6, 83-92, 1990.

[22] P.W. Partridge, L.C. Wrobel, The dual reciprocity boundary element method for spontaneous ignition. International Journal for Numerical Methods in Engineering, 30, 953–963, 1990.

[23] J.H. Frijns, S.L. de Snoo and R. Schoonhoven, Improving the accuracy of the boundary element method by the use of second-order interpolation functions. IEEE Transactions on Biomedical Engineering, 47, 1336-1346, 2000.

[24] L. Gaul, M. Kögl and M. Wagner [2003] "Boundary element methods for engineers and scientists," (Springer-Verlag, Berlin), 2003.

[25] M.A. Fahmy, Application of DRBEM to non-steady state heat conduction in non-homogeneous anisotropic media under various boundary elements. Far East Journal of Mathematical Sciences, 43, 83-93, 2010.

[26] M.A. Fahmy, A time-stepping DRBEM for magneto-thermo-viscoelastic interactions in a rotating nonhomogeneous anisotropic solid. Int. J. Appl. Mech., 3, 1-24, 2011.

[27] M.A. Fahmy, Influence of inhomogeneity and initial stress on the transient magneto-thermo-visco-elastic stress waves in an anisotropic solid. World Journal of Mechanics, 1, 256-265, 2011.

[28] M.A. Fahmy, A time-stepping DRBEM for the transient magneto-thermo-visco-elastic stresses in a rotating non-homogeneous anisotropic solid. Engineering Analysis with Boundary Elements, 36, 335-345, 2012.

[29] M.A. Fahmy, Numerical modeling of transient magneto-thermo-viscoelastic waves in a rotating nonhomogeneous anisotropic solid under initial stress. International Journal of Modeling, Simulation, and Scientific Computing, 3, 125002, 2012.

[30] M.A. Fahmy, Transient magneto-thermo-elastic stresses in an anisotropic viscoelastic solid with and without moving heat source. Numer. Heat Transfer, Part A., 61, 547-564, 2012.

[31] M.A. Fahmy, Transient magneto-thermoviscoelastic plane waves in a non-homogeneous anisotropic thick strip subjected to a moving heat source, Applied Mathematical Modelling, 36, 4565-4578, 2012.

[32] M.A. Fahmy, Transient magneto-thermo-viscoelastic stresses in a rotating nonhomogeneous anisotropic solid with and without a moving heat source. Journal of Engineering Physics and Thermophysics, 85, 874-880, 2012.

[33] M.A. Fahmy, The DRBEM solution of the generalized magneto-thermo-viscoelastic problems in 3D anisotropic functionally graded solids. I. Idelsohn, M. Papadrakakis, B. Schrefler Eds. 5th International conference on coupled problems in science and engineering (Coupled Problems 2013), Ibiza, Spain, pp. 862-872, June 17-19, 2013.

[34] M.A. Fahmy, Generalized Magneto-Thermo-Viscoelastic Problems of Rotating Functionally Graded Anisotropic Plates by the Dual Reciprocity Boundary Element Method. Journal of Thermal Stresses, 36, 284-303, 2013.

[35] M.A. Fahmy, A Three-Dimensional Generalized Magneto-Thermo Viscoelastic Problem of a Rotating Functionally Graded Anisotropic Solids with and without Energy Dissipation. Numerical Heat Transfer, Part A: Applications, 63, 713–733, 2013.

[36] M.A. Fahmy, Implicit–Explicit Time Integration DRBEM for Generalized Magneto-Thermoelasticity Problems of Rotating Anisotropic Viscoelastic Functionally Graded Solids. Engineering Analysis with Boundary Elements, 37, 107-115, 2013.

[37] M.A. Fahmy, A Comuputerized DRBEM model for generalized magneto-thermo-viscoelastic stress waves in functionally graded anisotropic thin film/substrate structures. Latin American Journal of solids and structures, 11, 386 – 409, 2014.

[38] M.A. Fahmy M. A. [2014b] "A 2D Time Domain DRBEM Computer Model for Magneto-Thermoelastic Coupled Wave Propagation Problems. International Journal of Engineering and Technology Innovation, 4, 138-151, 2014.

[39] M.A. Fahmy, A.M. Salem, M.S. Metwally and M.M. Rashid, Computer Implementation of the DRBEM for Studying the Generalized Thermoelastic Responses of Functionally Graded Anisotropic Rotating Plates with One Relaxation Time. International Journal of Applied Science and Technology, 3, 130-140, 2013.

[40] M.A. Fahmy, A.M. Salem, M.S. Metwally and M.M. Rashid, Computer Implementation of the Drbem for Studying the Classical Uncoupled Theory of Thermoelasticity of Functionally Graded Anisotropic Rotating Plates. International Journal of Engineering Research and Applications, 3, 1146-1154, 2013.

[41] M.A. Fahmy, A.M. Salem, M.S. Metwally and M.M. Rashid, Computer Implementation of the DRBEM for Studying the Classical Coupled Thermoelastic Responses of Functionally Graded Anisotropic Plates. Physical Science International Journal, 4, 674-685, 2014.

[42] M.A. Fahmy, A.M. Salem, M.S. Metwally and M.M. Rashid, Computer Implementation of the DRBEM for Studying the Generalized Thermo Elastic Responses of Functionally Graded Anisotropic Rotating Plates with Two Relaxation Times. British Journal of Mathematics & Computer Science, 4, 1010-1026, 2014.

[43] M.A. Fahmy, Boundary Element Solution of 2D Coupled Problem in Anisotropic Piezoelectric FGM Plates. Proceedings of the, B. Schrefler, E. Oñate and M. Papadrakakis eds. 6th International Conference on Computational Methods for Coupled Problems in

Science and Engineering (Coupled Problems 2015), Venice. Italy, pp. 382-391, May 18-20, 2015.

[44] M.A. Fahmy, A comuputerized DRBEM model for 3D dynamic coupled problems in anisotropic piezoelectric FGM plates, Presented in 2nd International Conference on New Horizons in Basic and Applied Science (ICNHBAS 2015), Hurghada, Egypt, August 1-6, 2015.

[45] P. Stanak, J. Sladek, V. Sladek, A. Tadeu, Three-Dimensional Meshless Modelling of Functionally Graded Piezoelectric Sensor. in Mechatronics 2013, ed. T. Březina and R. Jabloński (Springer International Publishing Switzerland), pp. 425-432.

[46] W. Liu, S. Ma, H. Wu, Three-dimensional analysis of functionally graded piezoelectric plate with arbitrarily distributed material properties. Journal of Wuhan University of Technology-Mater. Sci. Ed., 29, 712-720, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DIRECT EDDY CURRENT METHOD FOR VOLUMETRIC FLAWS OF CYLINDRICAL SHAPE

**Valentina Koliskina[1], Andrei Kolyshkin[1], Rauno Gordon[2] and Olev Märtens[2]**

[1]Riga Technical University
2 Daugavgrivas street Riga LV 1007 Latvia
e-mail: v.koliskina@gmail.com, andrejs.koliskins@rtu.lv

[2] Tallinn University of Technology
Ehitajate tee 5 19086 Tallinn Estonia
e-mail: rauno.gordon@ttu.ee,olev@elin.ttu.ee

**Keywords:** Eddy Currents, Truncated Eigenfunction Expansion

**Abstract.** *A quasi-analytical method for the solution of direct eddy current testing problems for the case of cylindrical volumetric flaws is presented in the paper. The method is based on a simple physical assumption that the electromagnetic field induced by a coil carrying alternating current is exactly equal to zero at a sufficiently large radial distance from the coil. The axis of the coil concides with the axis of a cylindrical flaw. The method of truncated eigenfunction expansions is used to compute the change in impedance of the coil. Complex eigenvalues are computed numerically using the method which does not require initial approximation for the eigenvalue. Computations are presented for different values of the parameters of the problem. Calculated change in impedance is compared with numerical results obtained by means of Comsol Multiphysics software. Good agreement between quasi-analytical method and numerical solution is found.*

## 1 INTRODUCTION

Eddy currents are widely used in nondestrutive testing of electrically conducting materials. Identification of flaws in conducting media is one of the goals of nondestructive testing. This is a complicated inverse problem which can be solved if a solution of a direct problem is available. In cases where a conducting medium is unbounded in one or two spatial dimensions solutions of a direct problem can be found by the method of integral transforms [1], [2]. In practice, however, finite size of a conducting medium has to be taken into account especially in cases where the size of an excitation coil is comparable to the size of the medium. Method of truncated eigenfunction expansions (known as the TREE method in the literature) is suggested in [3], [4] in order to construct solutions of direct problems for media of finite size. The method is based on a simple physical assumption that the electromagnetic field induced by a coil carrying alternating current is exactly equal to zero at a sufficiently large radial distance from the coil. This assumption allows one to extend the class of problems which can be solved by a quasi-analytical method i.e., by the method of truncated eigenfunction expansions. Examples of the application of the TREE method to the solution of direct eddy current testing problems with cylindrical symmetry can be found in [3], [5], [6], [7]. The case of an asymmetric infinite cylindrical flaw is considered in [8]. In the present paper we present the solution of eddy current testing problem for the case of a cylindrical flaw in a two-layer plate. Such a model can be used to assess the effect of corrosion in metal coatings.

## 2 MATHEMATICAL FORMULATION OF THE PROBLEM

Consider an air core coil located above an electrically conducting nonmagnetic two-layer plate of finite size. A flaw in the form of a circular cylinder is located in the medium. It is assumed that the axis of the flaw coincides with the axis of the coil. An example of the flaw at the surface of a two-layer plate is shown in Fig. 1.

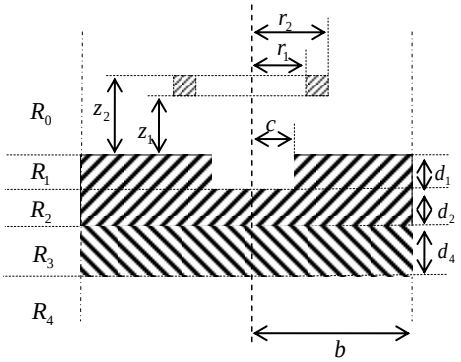



Figure 1: A coil carrying alternating current above a conducting two-layer plate with a flaw.

It is assumed that the electromagnetic field is equal to zero at a sufficiently large distance $b$ from the axis of the coil. Using the cylindrical symmetry of the problem we assume that the vector potential has only one non-zero component in the azimuthal direction. As usual in such type of problems, we consider a single-turn coil of radius $r_0$ located at the distance $h$ above the top surface of the plate. The solution for the case of a coil with finite dimensions can be found by the superposition principle. The system of the Maxwell's equations in each of the regions $R_0$–$R_4$ shown in Fig. 1 has the form:

$$\frac{\partial^2 A_0}{\partial r^2} + \frac{1}{r}\frac{\partial A_0}{\partial r} - \frac{A_0}{r^2} + \frac{\partial^2 A_0}{\partial z^2} = -\mu_0 I \delta(r - r_0)\delta(z - h), \qquad (1)$$

$$\frac{\partial^2 A_1}{\partial r^2}+\frac{1}{r}\frac{\partial A_1}{\partial r}-\frac{A_1}{r^2}+\frac{\partial^2 A_1}{\partial z^2}-j\omega\sigma\mu_0 A_1 = 0, \tag{2}$$

$$\frac{\partial^2 A_2}{\partial r^2}+\frac{1}{r}\frac{\partial A_2}{\partial r}-\frac{A_2}{r^2}+\frac{\partial^2 A_2}{\partial z^2}-j\omega\sigma_1\mu_0 A_2 = 0, \tag{3}$$

$$\frac{\partial^2 A_3}{\partial r^2}+\frac{1}{r}\frac{\partial A_3}{\partial r}-\frac{A_3}{r^2}+\frac{\partial^2 A_3}{\partial z^2}-j\omega\sigma_2\mu_0 A_3 = 0, \tag{4}$$

$$\frac{\partial^2 A_4}{\partial r^2}+\frac{1}{r}\frac{\partial A_4}{\partial r}-\frac{A_4}{r^2}+\frac{\partial^2 A_4}{\partial z^2} = 0, \tag{5}$$

where $\sigma_1$ and $\sigma_2$ are the electrical conductivities of regions $R_2$ and $R_3$, respectively, $\sigma=\sigma_1$ if $c<r<b$ and $\sigma=0$ if $0\le r<c$.

The boundary conditions at $r=b$ are

$$A_i|_{r=b}=0, \quad i=0,1,...,4. \tag{6}$$

The boundary conditions at $r=c$ have the form

$$A_1^{air}|_{r=c}=A_1^{con}|_{r=c}, \quad \frac{\partial A_1^{air}}{\partial r}|_{r=c}=\frac{\partial A_1^{con}}{\partial r}|_{r=c}, \tag{7}$$

where we use the notations $A_1^{air}$ and $A_1^{con}$ in regions $0\le r<c$ and $c<r<b$, respectively.

The conditions at the surfaces $z=0$, $z=-d_1$, $z=-d_3$ and $z=-d_5$, where $d_3=d_1+d_2$, $d_5=d_3+d_4$ are

$$A_0|_{z=0}=A_1^{air}|_{z=0}, \quad \frac{\partial A_0}{\partial z}|_{z=0}=\frac{\partial A_1^{air}}{\partial z}|_{z=0} \quad 0\le r<c, \tag{8}$$

$$A_0|_{z=0}=A_1^{con}|_{z=0}, \quad \frac{\partial A_0}{\partial z}|_{z=0}=\frac{\partial A_1^{con}}{\partial z}|_{z=0} \quad c<r<b. \tag{9}$$

$$A_1^{air}|_{z=-d_1}=A_2|_{z=-d_1}, \quad \frac{\partial A_1^{air}}{\partial z}|_{z=-d_1}=\frac{\partial A_2}{\partial z}|_{z=-d_1} \quad 0\le r<c. \tag{10}$$

$$A_1^{con}|_{z=-d_1}=A_2|_{z=-d_1}, \quad \frac{\partial A_1^{con}}{\partial z}|_{z=-d_1}=\frac{\partial A_2}{\partial z}|_{z=-d_1} \quad c<r<b. \tag{11}$$

$$A_2|_{z=-d_3}=A_3|_{z=-d_3}, \quad \frac{\partial A_2}{\partial z}|_{z=-d_3}=\frac{\partial A_3}{\partial z}|_{z=-d_3}, \tag{12}$$

$$A_3|_{z=-d_5}=A_4|_{z=-d_5}, \quad \frac{\partial A_3}{\partial z}|_{z=-d_5}=\frac{\partial A_4}{\partial z}|_{z=-d_5}. \tag{13}$$

In addition, the solution is bounded as $z\to\pm\infty$:

$$A_0|_{z\to+\infty}=0, \quad A_4|_{z\to-\infty}=0. \tag{14}$$

## 3 PROBLEM SOLUTION

Problem (1)-(14) is solved by the method of separation of variables. It is convenient to represent the solution in region $R_0$ in the form

$$A_{00}(r,z) = \sum_{i=1}^{\infty}(D_{1i}\,e^{-\lambda_i z}J_1(\lambda_i r), \tag{15}$$

$$A_{01}(r,z) = \sum_{i=1}^{\infty}(D_{2i}\,e^{-\lambda_i z}+D_{3i}\,e^{\lambda_i z})J_1(\lambda_i r), \tag{16}$$

where $\lambda_i = \alpha_i/b$ and $\alpha_i$ are the roots of the equation

$$J_1(\alpha_i) = 0. \tag{17}$$

Here $A_{00}$ and $A_{01}$ represent the solutions in subregions $R_{00} = \{0 < z < h\}$ and $R_{01} = \{z > h\}$, respectively. The vector potential is continuous at $z = h$:

$$A_{00}|_{z=h} = A_{01}|_{z=h}. \tag{18}$$

Another condition at $z = h$ can be obtained from equation (1). Integrating (1) with respect to $z$ from $h - \varepsilon$ to $h + \varepsilon$ and considering the limit as $\varepsilon \to +0$ we obtain

$$\frac{\partial A_{01}}{\partial z}|_{z=h} - \frac{\partial A_{00}}{\partial z}|_{z=h} = -\mu_0 I \delta(r - r_0). \tag{19}$$

Using (15), (16), (18) and (19) we obtain

$$A_{00}(r,z) = \sum_{i=1}^{\infty} (D_{2i}\, e^{-\lambda_i z} J_1(\lambda_i r) + \frac{\mu_0 I r_0}{b^2} \sum_{i=1}^{\infty} \frac{J_1(\lambda_i r_0) J_1(\lambda_i r)\, e^{\lambda_i (h-z)}}{\lambda_i J_0^2(\lambda_i b)}, \tag{20}$$

$$A_{01}(r,z) = \sum_{i=1}^{\infty} (D_{2i}\, e^{-\lambda_i z} J_1(\lambda_i r) + \frac{\mu_0 I r_0}{b^2} \sum_{i=1}^{\infty} \frac{J_1(\lambda_i r_0) J_1(\lambda_i r)\, e^{\lambda_i (z-h)}}{\lambda_i J_0^2(\lambda_i b)}. \tag{21}$$

The solution in region $R_1$ can be written in the form

$$A_1^{air}(r,z) = \sum_{i=1}^{\infty} (D_{4i} J_1(p_i r)\, e^{p_i z} + D_{5i} J_1(p_i r)\, e^{-p_i z}, \tag{22}$$

$$A_1^{con}(r,z) = \sum_{i=1}^{\infty} [(D_{6i} J_1(q_i r) + D_{7i} Y_1(q_i r))\, e^{p_i z} + (D_{8i} J_1(q_i r) + D_{9i} Y_1(q_i r))\, e^{-p_i z}], \tag{23}$$

where $p_i = \sqrt{q_i^2 + j\omega\sigma_1\mu_0}$. Using (22), (23) and the first condition in (7) we obtain

$$A_1^{air}(r,z) = \sum_{i=1}^{\infty} J_1(p_i r) T_1(q_i c) [\hat{D}_{6i}\, e^{p_i z} + \hat{D}_{8i}\, e^{-p_i z}], \tag{24}$$

$$A_1^{con}(r,z) = \sum_{i=1}^{\infty} J_1(p_i c) T_1(q_i r) [\hat{D}_{6i}\, e^{p_i z} + \hat{D}_{8i}\, e^{-p_i z}], \tag{25}$$

where

$$T_1(q_i r) = J_1(q_i r) Y_1(q_i b) - J_1(q_i b) Y_1(q_i r)$$

and $\hat{D}_{6i}$, $\hat{D}_{8i}$ are arbitrary constants. Using the second condition in (7) the following equation is obtained

$$q_i T_1'(q_i r) J_1(p_i c) = p_i J_1'(p_i c) T_1(q_i c). \tag{26}$$

Equation (26) is used to calculate complex eigenvalues $p_i$ and the corresponding values $q_i$.

The solution in region $R_2$ is

$$A_2(r,z) = \sum_{i=1}^{\infty} (D_{10i}\, e^{p_{1i} z} + D_{11i}\, e^{-p_{1i} z}) J_1(\lambda_i r), \tag{27}$$

where $p_{1i} = \sqrt{\lambda_i^2 + j\omega\sigma_1\mu_0}$.

Similarly, the solution in region $R_3$ can be written in the form

$$A_3(r,z) = \sum_{i=1}^{\infty} (D_{12i}\, e^{p_{2i}z} + D_{13i}\, e^{-p_{2i}z}) J_1(\lambda_i r), \tag{28}$$

where $p_{2i} = \sqrt{\lambda_i^2 + j\omega\sigma_2\mu_0}$.

Finally, the bounded solution in region $R_4$ is

$$A_4(r,z) = \sum_{i=1}^{\infty} D_{14i}\, e^{\lambda_i z} J_1(\lambda_i r). \tag{29}$$

Unknown constants in (20)–(25), (27)–(29) can be found from the boundary conditions (8)–(13). It can be shown that the corresponding system of equations can be reduced to system of two linear algebraic equations with respect to the unknowns $\hat{D}_{6i}$ and $\hat{D}_{8i}$. It is neccesary to truncate the infinite series to a finite number of terms, $m$. Computational details for the solution of similar problems can be found elsewhere (see, for example, [3] and [5]).

Assuming that the coefficients $\hat{D}_{6i}$ and $\hat{D}_{8i}$ are calculated (this can be done using any linear solver) we obtain the induced vector potential of the single-turn coil in the form

$$A_0^{ind}(r,z,r_0,h) = \sum_{k=1}^{m} D_{2k}\, e^{-\lambda_k z} J_1(\lambda_k r). \tag{30}$$

Induced vector potential for the coil of finite dimensions can be computed as follows

$$A_{coil}^{ind}(r,z) = \int_{r_1}^{r_2}\int_{z_1}^{z_2} A_0^{ind}(r,z,r_0,h)\, dr_o dh. \tag{31}$$

The induced change in impedance of the coil is given by the formula (see [3]):

$$Z_{ind} = \frac{2\pi j\omega N}{I(r_2-r_1)(z_2-z_1)} \int_{r_1}^{r_2}\int_{z_1}^{z_2} r A_{coil}^{ind}(r,z)\, dr dz, \tag{32}$$

where $N$ is the number of turns in the coil.

Using (30)–(32) we obtain

$$\begin{aligned} Z_{ind} \;=\; & \frac{2\pi j\omega N^2\mu_0}{(r_2-r_1)^2(z_2-z_1)^2} \sum_{i=1}^{n} \frac{e^{-\lambda_i z_1} - e^{-\lambda_i z_2}}{\lambda_i^3} \int_{\lambda_i r_1}^{\lambda_i r_2} y J_1(y)\, dy \\ & \times \sum_{k=1}^{n} Y_{ik} \frac{e^{-\lambda_k z_1} - e^{-\lambda_k z_2}}{\lambda_k^3} \int_{\lambda_k r_1}^{\lambda_k r_2} \xi J_1(\xi)\, d\xi. \end{aligned} \tag{33}$$

The coefficients $Y_{ik}$ in (33) are bulky and are not shown here for brevity.

## 4 NUMERICAL RESULTS

Calculations of the change in impedance of the coil are performed with Mathematica. Integrals in (33) are evalauted in closed form using Bessel and Struve functions (see [9]). Complex eigenvlaues (the roots of equation (26)) are computed using the method described in [10] and [11]. Numerical results are presented for the following values of the parameters of the problem: $z_1 = 0.3$ mm, $z_2 = 2.6$ mm, $\sigma_1 = 3$ Ms/m, $\sigma_2 = 7$ Ms/m, $d_1 = 0.4$ mm, $d_2 = 0.5$ mm,

$d_3 = 0.8$ mm, $c = 2.2$ mm. The real amd imaginary parts of the change in impedance for the seven frequencies of the excitation current (from 1 kHz to 7 kHz with a step size of 1 kHz) are shown in Figs. 2 and 3. The inner and outer radii of the coil in Fig. 2 are $r_1 = 1.5$ mm and $r_2 = 2.5$ mm, respectively. The same problem is also solved numerically using Comsol Multiphysics software. The details of the finite element modeling of similar problems with Comsol Multiphysics can be found in [12].

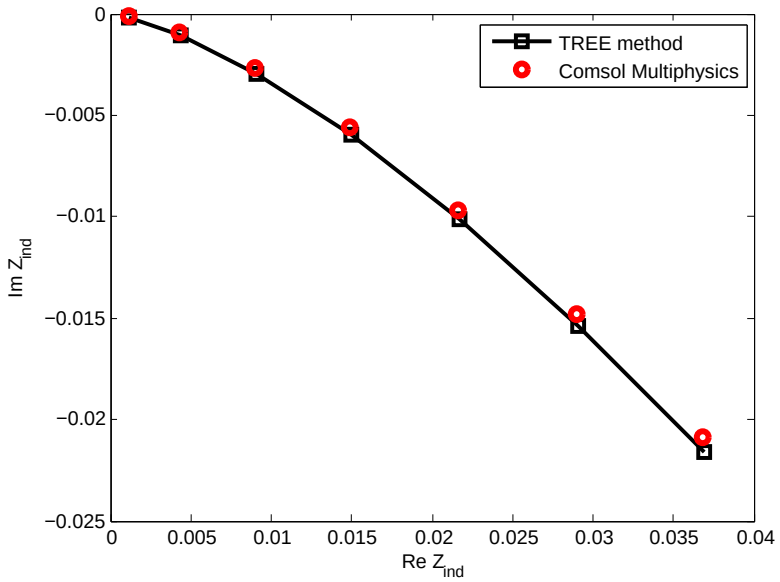



Figure 2: The change in impedance of the coil with $r_1 = 1.5$ mm, $r_2 = 2.5$ mm.

The change in impedance of the coil is also shown in Fig. 3 for the case $r_1 = 2.5$ mm, $r_2 = 4.5$ mm (the other parameters are the same as in Fig. 2).

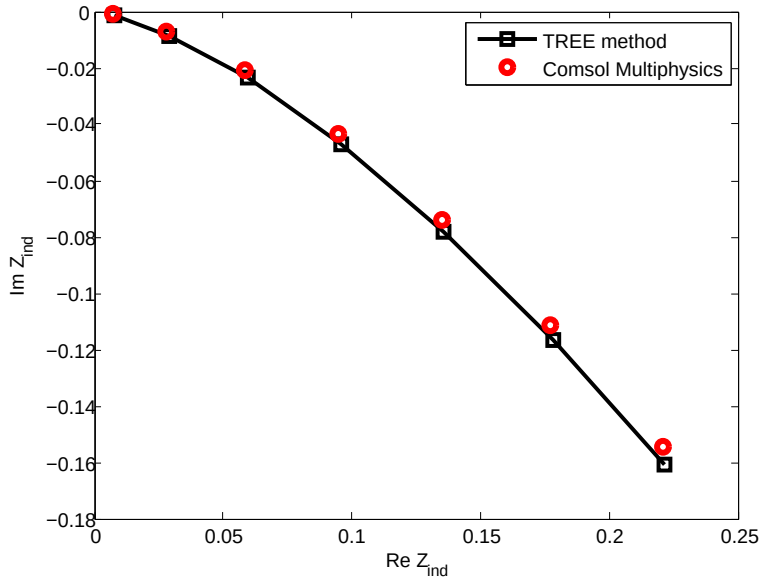



Figure 3: The change in impedance of the coil with $r_1 = 2.5$ mm, $r_2 = 4.5$ mm.

As can be seen from the graphs good agreement is found between the results obtained by the TREE method and finite element modeling with Comsol Multiphysics.

## 5 CONCLUSION

The model described in the paper represents the continuation of the authors' work related to the solution of direct eddy current testing problems for the case of cylindrical flaws. Good agreement is found between quasi-analytical solution and finite element modeling for all cases considered. The results obtained in the paper can be generalized for the case of magnetic materials (assuming that the magnetic permeability of a medium is constant).

## REFERENCES

[1] J.A. Tegopoulos, E.E. Kriezis, *Eddy currents in linear conducting media*. Elsevier, 1985.

[2] M.Ya. Antimirov, A.A. Kolyshkin, R. Vaillancourt, *Mathematical models for eddy current testing*. CRM, 1997.

[3] T.P. Theodoulidis, E.E. Kriezis, *Eddy current canonical problems (with applications to nondestructive evaluation*. Tech Science, 2006.

[4] T.P. Theodoulidis, E.E. Kriezis, Series expansions in eddy current nondestructive evaluation models. *Journal of Materials Processing Technology*, **161**, 343–347, 2005.

[5] V. Koliskina, *Analytical and quasi-analytical solutions of direct problems in eddy current testing*. PhD Thesis, Riga Technical Univesrity, 2013.

[6] V. Koliskina, A. Kolyshkin, Mathematical model for eddy current testing of metal plates with two cylindrical flaws Z. Leonowicz ed. *15th International Conference on Environment and Electrical Engineering (EEEIC 2015)*, Rome, Italy, June 10-13, 2015.

[7] V. Koliskina, A. Kolyshkin, Mathematical model for eddy current testing of surface flaws in a two-layer metal plate. J.F. Silva Gomes, Shakar A. Meguid Eds. *Proceedings of the 6th International Conference on Mechanics and Materials in Design*, P. Delgada, Portugal, July 26-30, 2015.

[8] A. Scarlatos, T. Theodoulidis, Solution to the eddy current induction problem in a conducting half-space with a vertical cylindrical borehole. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **468(2142)**, 1758–1777, 2012.

[9] M. Abramowitz, I.A. Stegun, *Handbook of mathematical functions with formulas, graphs and mathematical tables*. National Bureau of Standards, 1964.

[10] J.N. Lyness, Numerical algorithms based on the theory of complex variable. *Proceedings ACM*, 125–133, 1967.

[11] L.M. Delves, J.N. Lyness, A numerical method for locating the zeros of an analytic function. *Mathematics of Computation*, **21**, 543–560, 1967.

[12] R. Gordon, O. Märtens, A. Kolyshkin, Comparison of eddy current theory and finite element method for metal evaluation. *Lecture Notes in Impedance Spectroscopy*, **3**, 41–45, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF BUOYANT PLUMES USING A FIXED POINT ITERATIVE METHOD

**B. Bermúdez[1], A. Rangel-Huerta[1], W. Fermín Guerrero S.[2] and J. D. Alanís [3]**

[1] F. C. C. BUAP
14 Sur y San Claudio
Puebla, México
e-mail: bbj@cs.buap.mx, arangel@cs.buap.mx
[2] F. C. F. M. BUAP
18 Sur y San Claudio
Puebla, México
e-mail: wferming@gmail.com
[3] Universidad Tecnológica de Puebla
Carrera de Tecnologías de la Información y Comunicación, UTP
Antiguo Camino a la Resurrección 1002-A Col. Zona Industrial Oriente
Puebla, México
e-mail:david.alanis@utpuebla.edu.mx

**Keywords:** Buoyant plumes, Boussinesq approximation, Rayleigh number, high convection dominated flows.

**Abstract.** *A numerical simulation of unsteady plumes for driven thermal high convection dominated flows is presented in this work. The problem is of practical interest for example in the cooling of electronic devices and heat transfer from pipes in heat exchange systems. This kind of thermal flows may be modeled using the unsteady Boussinesq approximation in the Stream function-vorticity formulation.*
*Results are obtained with a simple numerical method previously reported for isothermal/natural and mixed convection flows. The numerical method is based on a fixed point iterative process to solve the non-linear elliptic system that results after an appropriate discretization in time.*
*The iterative process leads us to the solution of uncoupled, well-conditioned and symmetric linear elliptic problems for which very efficient solvers are known to exist regardless of the space discretization.*

## 1 INTRODUCTION

We present a numerical simulation of unsteady plumes for driven thermal high convection dominated flows. The iterative process used in this work, leads us to the solution of uncoupled, well-conditioned and symmetric linear elliptic problems.

The numerical method has been previously reported in [1] for mixed convection and in [2] for natural convection in tilted cavities. This scheme has shown to be robust enough to handle high Rayleigh numbers and different aspect ratios (ratio of the height to the width) of the cavity.

The evolution of the flows we are dealing with in this paper depends mainly on the variation of the parameters. The buoyancy forcing of the flow can be characterized by the Rayleigh number $Ra$, which in the case of this work, is taken in the range $10^6 \leq Ra \leq 2 * 10^8$. For this range of the Rayleigh number, numerical results are not easy to obtain, since in this case we have a highly convection-dominated thermally-driven problem as can be seen in [3] and [4].

Numerical simulation has been applied to this kind of flows and results with a 2D Direct Numerical Simulation (DNS) have been reported in [5] and [6]. For the validation of our results we are comparing them with those reported in the cited bibliography.

## 2 MATHEMATICAL MODEL

Let $\Omega \subset R^N$ $(N = 2, 3)$ be the region of a non-steady, viscous, incompressible flow, and $\Gamma$ its boundary. Under the hypothesis of the Boussinesq approximation, this kind of fluids may be modeled by the following adimensional system:

$$\begin{aligned} \mathbf{u}_t - \nabla^2\mathbf{u} + \nabla p + (\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{Ra}{Pr}\theta\mathbf{e} \quad (a) \\ \nabla\cdot\mathbf{u} = 0 \quad (b) \\ \theta_t - \frac{1}{Pr}\nabla^2\theta + \mathbf{u}\cdot\nabla\theta = 0 \quad (c) \end{aligned} \tag{1}$$

in $\Omega$, $t > 0$; where $\mathbf{u}$, $p$ and $\theta$ are the velocity, pressure and temperature of the fluid, respectively, $\mathbf{e}$ is the unitary vector in the gravitational direction. The parameters $Ra$ y $Pr$ are the Rayleigh and the Prandtl number respectively, which are given by: $Ra = \frac{\beta l^3 \kappa g \rho_0^2}{\mu^3 c_p}(T_l - T_0)$, $Pr = \kappa/\mu c_p$, where $T_0$ and $T_1$ are reference temperatures, $T_0 < T_1$, which may be the temperatures of the side walls when the region is a rectangular cavity. The adimensional temperature $\theta$ is given by $\theta = \frac{T - T_0}{T_1 - T_0}$.

This system must be provided with appropiate initial and boundary conditions:$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x})$ and $\theta(\mathbf{x}, 0) = \theta_0(\mathbf{x})$ in $\Omega$; $\mathbf{u} = \mathbf{f}$ and $B\theta = 0$ in $\Gamma$, $t \geq 0$, where $B$ is a boundary operator for the temperature and may involve Dirichlet, Neumann and mixed boundary conditions.

Restricting ourselves to a bidimensional region $\Omega$, taking the curl in both sides of equation $(1a)$ and taking into account that:

$$u_1 = \frac{\partial\psi}{\partial y}, \quad u_2 = -\frac{\partial\psi}{\partial x}, \tag{2}$$

which follows from $(1b)$, with $\psi$ the stream function and $(u_1, u_2) = \mathbf{u}$; the component in the $\mathbf{k} = (0, 0, 1)$ direction, we get the following scalar system:

$$\begin{aligned} \nabla^2\psi &= -\omega \quad (a) \\ \omega_t - \nabla^2\omega + \mathbf{u}\cdot\nabla\omega &= \frac{Ra}{Pr}\frac{\partial\theta}{\partial x} \quad (b) \\ \theta_t - \gamma\nabla^2\theta + \mathbf{u}\cdot\nabla\theta &= 0;, \quad (c) \end{aligned} \tag{3}$$

where $\gamma = 1/Pr$ and $\omega$ is the vorticity, which, from $\omega\mathbf{k} = \nabla\times\mathbf{u} = -\nabla^2\psi\mathbf{k}$, gives $(3a)$ and $\omega = \frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y}$ also.

Then, system (3) becomes the Boussinesq approximation in stream function-vorticity variables. The incompressibility condition $(1b)$, by (2), is automatically satisfied and the pressure $p$ has been eliminated.

A translation of the boundary conditions in terms of the velocity primitive variable $\mathbf{u}$ to the $\psi - \omega$ variables has to be performed. Following [7], $\psi$ is a constant function on solid and fixed walls; at the moving wall $y = b$, a constant function for $\psi$ is also obtained, then $\psi = 0$ is chosen in $\Gamma$. By Taylor expansion of $(3a)$ on the boundary, with $h_x$ and $h_y$ the space steps, one obtains

$$\begin{aligned} \omega(0,y,t) &= -\frac{1}{2h_x^2}[8\psi(h_x,y,t) - \psi(2h_x,y,t)] + O(h_x^2) \\ \omega(a,y,t) &= -\frac{1}{2h_x^2}[8\psi(a-h_x,y,t) - \psi(a-2h_x,y,t)] + O(h_x^2) \\ \omega(x,0,t) &= -\frac{1}{2h_y^2}[8\psi(x,h_y,t) - \psi(x,2h_y,t)] + O(h_y^2) \\ \omega(x,b,t) &= -\frac{1}{2h_y^2}[8\psi(x,b-h_y,t) - \psi(x,b-2h_y,t)] - \frac{3}{h_y} + O(h_y^2). \end{aligned} \tag{4}$$

.

where $h_x$ and $h_y$ denote the size of the spatial discretization in the $x$ and $y$ directions. It should be observed that the boundary values for $\omega$ are given by values in $\Omega$ and $t > 0$, still unknown, of the stream function $\psi$. This problem will be solved using a fixed point iterative process.

## 3 NUMERICAL METHOD

The time derivatives with respect to $\omega$ and $\theta$ in (3) are approximated using the following second order approximation:

$$f_t(\mathbf{x}, (n+1)\Delta t) \approx \frac{3f^{n+1} - 4f^n + f^{n-1}}{2\Delta t}, \tag{5}$$

where $n \geq 1,\ \mathbf{x} \in \Omega$, $\Delta t > 0$ is the time step, and $f^r \approx f(\mathbf{x}, r\Delta t)$; at each time level $t = (n+1)\Delta t$ the following semidiscrete system in $\Omega$ is obtained , with the corresponding boundary conditions on $\Gamma$:

$$\begin{aligned} \nabla^2\psi^{n+1} &= -\omega^{n+1}, & \psi^{n+1}|_\Gamma &= 0, \\ \alpha\omega^{n+1} - \nabla^2\omega^{n+1} + \mathbf{u}^{n+1}\cdot\nabla\omega^{n+1} &= \frac{R_a}{Pr}\frac{\partial\theta^{n+1}}{\partial x} + f_\omega, & \omega^{n+1}|_\Gamma &= \omega_{bc}^{n+1}, \\ \alpha\theta^{n+1} - \gamma\nabla^2\theta^{n+1} + \mathbf{u}^{\mathbf{n+1}}\cdot\nabla\theta^{n+1} &= f_\theta, & B\theta^{n+1}|_\Gamma &= 0. \end{aligned} \tag{6}$$

where $\alpha = \frac{3}{2\Delta t}$, $f_\omega = \frac{4\omega^n - \omega^{n-1}}{2\Delta t}$, and $f_\theta = \frac{4\theta^n - \theta^{n-1}}{2\Delta t}$; $\omega_{bc}$ denotes the boundary condition of $\omega$ , $B$ denotes the boundary operator for $\theta$ mentioned above, and $u_1$ and $u_2$ the components of $\mathbf{u}$ in terms of $\psi$ are given by (2).

Denoting $(\psi^{n+1}, \omega^{n+1}, \theta^{n+1})$ by $(\psi, \omega, \theta)$ the following non-linear elliptic system is obtained:

$$\begin{aligned} \nabla^2\psi &= -\omega, & \psi|_\Gamma &= 0 & (a), \\ \alpha\omega - \nabla^2\omega + \mathbf{u}\cdot\nabla\omega &= \frac{Ra}{Pr}\frac{\partial\theta}{\partial x} + f_\omega, & \omega|_\Gamma &= \omega_{bc} & (b), \\ \alpha\theta - \gamma\nabla^2\theta + \mathbf{u}\cdot\nabla\theta &= f_\theta, & B\theta|_\Gamma &= 0 & (c). \end{aligned} \tag{7}$$

To obtain $(\omega^1, \theta^1, \psi^1)$ in (6), a first order approximation for the time derivatives may be applied through a subsequence with asmaller time step; a stationary system is then also obtained.

Denoting by

$$\begin{aligned} R_\omega(\omega,\psi) &\equiv \alpha\omega - \nabla^2\omega + \mathbf{u}\cdot\nabla\omega - \frac{Ra}{Pr}\frac{\partial\theta}{\partial x} - f_\omega, \\ R_\theta(\theta,\psi) &\equiv \alpha\theta - \gamma\nabla^2\theta + \mathbf{u}\cdot\nabla\theta - f_\theta. \end{aligned} \tag{8}$$

Then, system (7) is equivalent, in $\Omega$, to:

$$\begin{aligned} \nabla^2\psi &= -\omega, & \psi|_\Gamma &= 0, \\ R_\omega(\theta,\psi) &= 0, & \omega|_\Gamma &= \omega_{bc} \\ R_\theta(\omega,\psi) &= 0, & B\theta|_\Gamma &= 0 \end{aligned} \tag{9}$$

To solve this system at each time level $(n+1)\Delta t$, the following fixed point iterative process is applied in $\Omega$:[8]

With $\{\theta^0, \omega^0\} = \{\theta^n, \omega^n\}$ given, solve "until convergence" on $\theta$ and $\omega$

$$\begin{aligned} \nabla^2\psi^{m+1} &= -\omega^m, \quad \psi^{m+1} = 0 \text{ on } \Gamma, \\ \theta^{m+1} &= \theta^m - \rho_\theta(\alpha I - \gamma\nabla^2)^{-1} R_\theta(\theta^m, \psi^{m+1}), \\ B\theta^{m+1} &= 0 \text{ on } \Gamma, \quad \rho_\theta > 0, \\ \omega^{m+1} &= \omega^m - \rho_\omega(\alpha I - \nabla^2)^{-1} R_\omega(\omega^m, \psi^{m+1}), \\ \omega^{m+1} &= \omega_{bc}^{m+1} \text{ on } \Gamma, \quad \rho_\omega > 0, \end{aligned} \tag{10}$$

and then take $(\omega^{n+1}, \psi^{n+1}, \theta^{n+1}) = (\omega^{m+1}, \psi^{m+1}, \theta^{m+1})$.

With "until convergence" we mean until two consecutive values of $\theta$ and $\omega$, that is, $\theta^{m+1}$ and $\theta^m$, and $\omega^{m+1}$ and $\omega^m$, differ by less than a certain tolerance, tol, given.

Then, at each iteration, *three* linear elliptic problems associated with the operators: $\nabla^2$, $\alpha I - \gamma\nabla^2$, and $\alpha I - \nabla^2$ have to be solved in $\Omega$. For the discretization of such problems, either finite elements or finite differences may be used, as long as rectangular domains, as in the case of this work, are considered.

## 4 NUMERICAL RESULTS

Our numerical experiments take place in a rectangular cavity $\Omega = [0,1] \times [0,1]$. The walls of the cavity are solid and fixed. For the temperature, the source in the bottom is taken as a smooth Gaussian hill: $\frac{\partial T}{\partial n} = e^{-F(x-x_0)^2}$, with $x_0$ the center. The side walls are adiabatic, and on the top wall, zero temperature is prescribed. No-slip boundary conditions for the velocity are imposed on all the walls of the cavity. A translation of the boundary condition in terms of the velocity primitive variable $u$ to the $\psi - \omega$ variables has been performed using (4). The cavity is supposed to be filled with air.

In the following Figures we show the Stream Function (left) and the isotherms (right) for different values of the Rayleigh number.

In Figure 1 we show results for $Ra = 10^6$, at $t = .05$ with $h_x = h_y = 1/256$ and $\Delta t = .00001$. In the graph of the streamline function two vortices in counter flow, but of the same intensity, can be observed. The isotherms, in this case, are symmetric with respect to the source of heat, occupying only part of the cavity.

Increasing the Rayleigh number to $10^7$, in Figure 2, we show results at $t = .05$ with $h_x = h_y = 1/256$ and $\Delta t = .00001$. In this case, we can observe just one vortex in the graph of the streamline function with its maximum located at the center of the cavity. Nevertheless, the isotherms begin to show some asymmetry, bending towards the right wall, and

presenting a fold at the left top part without filling still the cavity. This indicates that the flow of heat is being transported by the movement of the fluid flow.

Increasing, additionally, the Rayleigh number to $Ra = 10^8$, we show, in Figure 3 results at $t = 0.05$ with $h_x = h_y = 1/256$ and $\Delta t = 0.00001$. We can observe here the restoration of two vortices of counterflow, although one dominates the circulation and the other one is smaller. The isotherms present well-defined asymmetric contours with a plume to the right near the heat source and irregular contours in the top right of the cavity.

In Figure 4 we show results for $Ra = 2 * 10^8$ at $t = 0.01$ with $h_x = h_y = 1/256$ and $\Delta t = 0.00001$. The Figure shows changes to double the Rayleigh number and simulation time given. The numerical method shows to be robust enough to obtain these results. This changes show a growth on the small vortex although the circulation is still being dominated by a dominant vortex. The isotherms present more structure what can be interpreted as a greater influence of its direct buoyancy.

In Figures 5 and 6 an increase in time is considered for the same Rayleigh number. Results are shown at $t = 0.05$ with $h_x = h_y = 1/256$ and $\Delta t = 0.00001$. No significant changes in the flow pattern of the two vortices or in the asymmetric structure of the isotherms are observed.

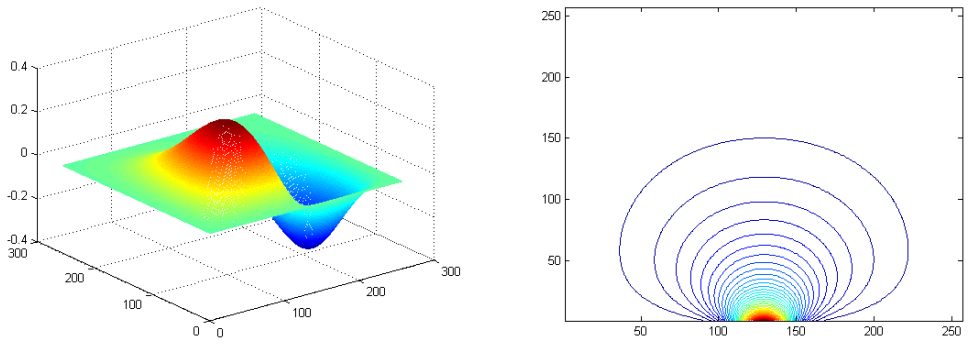



Figure 1: Stream function and isotherms for $Ra = 10^6$, $t = 0.05$, $F = 200$

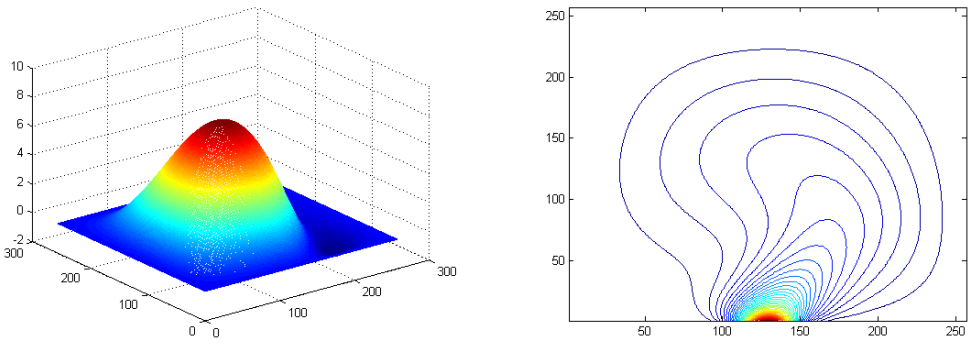



Figure 2: Stream function and isotherms for $Ra = 10^7$, $t = 0.05$, $F = 200$

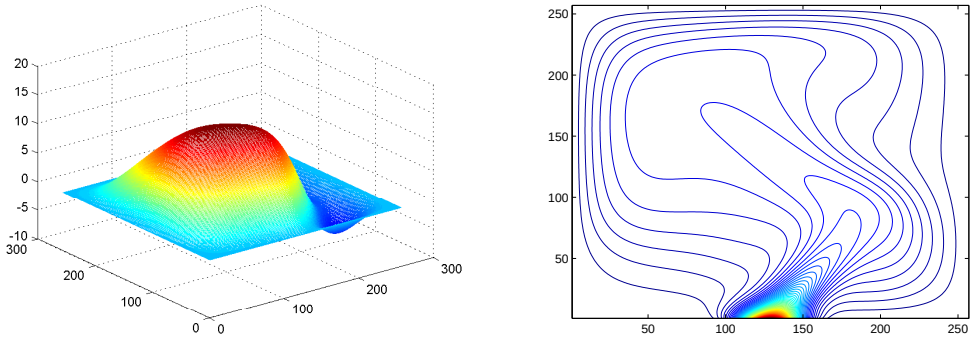



Figure 3: Stream function and isotherms for $Ra = 10^8$, $t = 0.05$, $F = 200$

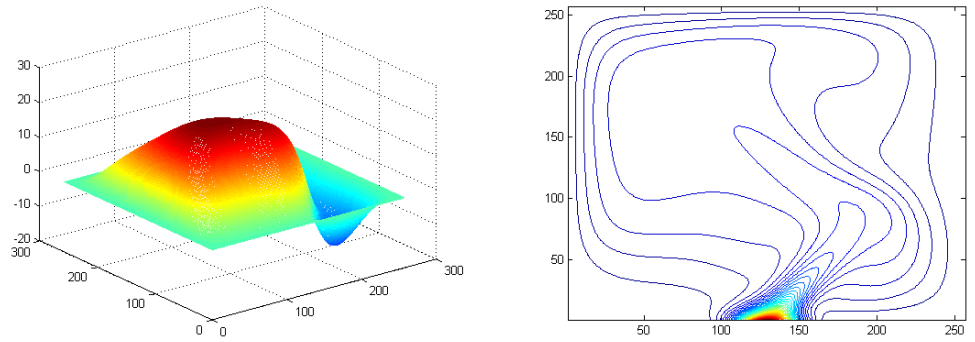



Figure 4: Stream function and isotherms for $Ra = 2 * 10^8$, $t = 0.01$, $F = 200$

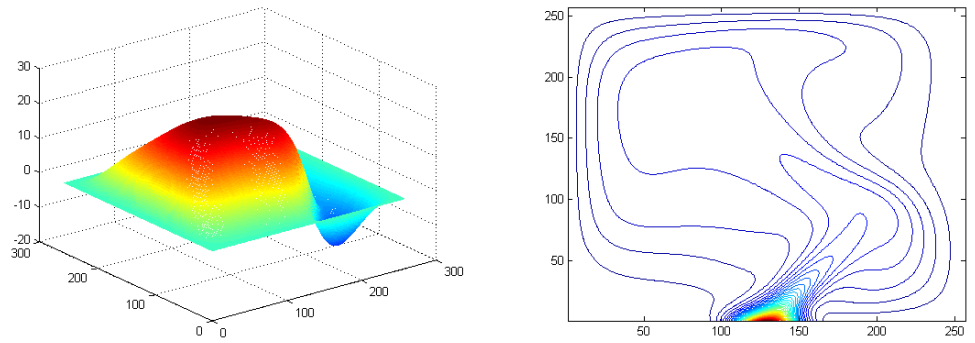



Figure 5: Stream function and isotherms for $Ra = 2 * 10^8$, $t = 0.05$, $F = 200$

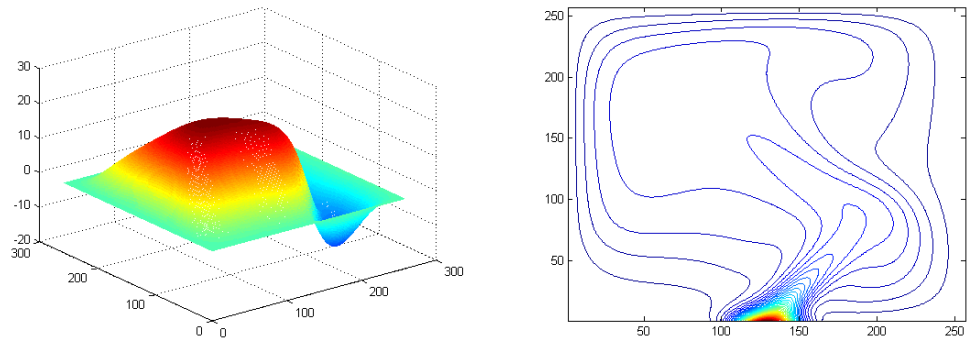



Figure 6: Stream function and isotherms for $Ra = 2 * 10^8$, $t = 0.075$, $F = 200$

## 5 CONCLUSIONS

The numerical method implemented here lets us distinguish diffusive and convective heat turbulent flow patterns generated in accordance with the increase of Rayleigh number within the cavity. Indeed, to relatively small values of this number the heat diffusion determines the flow profiles (Figure 1) up to $Ra = 10^7$ with only a vortex flow (Figure 2). However, for higher values of this number ($Ra = 10^8$) the isotherm profiles are shifted towards the walls asymmetrically with a dominant convective flow (Figures 3-5) as result of the heat source and the fluid recirculation generated by the turbulent flows within the cavity.

The fixed point iterative process used here has shown to be robust enough to handle very high values of the parameters appearing in the Boussinesq system, such as the Rayleigh number. We are looking forward to improve the method with respect to the time needed for the simulation, since till now, the time needed to solve the numerical systems of equations appearing after the discretization process is very large, especially when working with high Rayleigh numbers and having to arrive to a large value of t (time), since at each iteration in time, a very large system of linear equations has to be solved.

## REFERENCES

[1] Nicolás, A. and Bermúdez, B., 2D thermal/isothermal incompressible viscous flows, *Int. J. Numer. Meth. Fluids*,48, pp. 349-366, 2005.

[2] Báez E., Nicolás A., 2D natural convection flows in tilted cavities: porous media and homogeneous fluids, *Int. J. of heat and Mass Transfer*, vol. 49, 4773-4785, 2006.

[3] Minev, P.D., van de Vosse F.,N., Timmermans L. J. P., Rindt, C. C. M., van Steenhoven, A. A., Numerical simulation of buoyant plumes using a spectral element technique, Proceedings of Heat Transfer 94, Advanced Computational Methods in Heat Transfer,(ed. Wrobel, Brebbia and Nowak), Computational Mechanics publications, Southampton,Boston, 147-154, 1994.

[4] Bastiaans, R.G.M., Rindt, C.C.M., van Steenhoven, A.A. and Nieuwstadt, F.T.M., Direct and large eddy simulation of the transition of two and three-dimensional plane plumes in a confined enclosure, Int. Journal of Heat and Mass Transfer 43 2375-2393, 2000.

[5] Desrayaud G., G. Lauriat, Unsteady confined buoyant plumes, J. Fluid Mech 158, 245-268, 1985.

[6] Minev, P.D., van de Vosse F.,N., Timmermans L. J. P., van Steenhoven, A. A., A second order splitting algorithm for thermally-driven flow problems, Int. J. Num. Meth Heat Fluid Flow 6(2), 51-60, 1995.

[7] R. Peyret, T.D. Taylor, Computational Methods for Fluid Flow, Springer-Verlag, NY, 1983.

[8] Nicolas, A., A finite element approach to the Kuramoto-Sivashinski equation, Advances in Numerical Methods, SIAM, 1991.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# INVESTIGATION OF THE FLOW OVER AN OSCILLATING CYLINDER WITH THE VERY LARGE EDDY SIMULATION MODEL

**A. Kondratyuk[1,2], M. Schäfer[2], and A. Ali[2]**

[1]Graduate School of Computational Engineering
Technische Universität Darmstadt,
Dolivostr. 15, 64293 Darmstadt, e-mail: kondratyuk@gsc.tu-darmstadt.de

[2] Institute of Numerical Methods in Mechanical Engineering
Technische Universität Darmstadt,
Dolivostr. 15, 64293 Darmstadt
e-mail: schaefer@fnb.tu-darmstadt.de, ali@fnb.tu-darmstadt.de

**Keywords:** moving grid, hybrid turbulence modeling, oscillating cylinder, very large eddy simulation

**Abstract.** *The focus of this study is the numerical simulation of the forced oscillation of a circular cylinder with a Very Large Eddy Simulation model (VLES). Five different oscillating frequencies are considered to cover the so-called jump, where abrupt changes in drag and lift coefficients are observed. The results of the simulations are compared to available experimental data and numerical results from other studies.*

# 1 INTRODUCTION

The prediction of complex unsteady flows with a Direct Numerical Simulation (DNS) is too expensive in sense of computational costs, regardless of the advancements in computational technology over the past decades. DNS can only be applied to flows with very simple geometries and small Reynolds numbers, because a complete resolution of the turbulent structures is necessary. In contrary, with a Reynolds-Average-Navier-Stokes (RANS) model it is possible to simulate complex industrial problems, because all of the turbulent structures are modeled. However this approach shows problems with massively separating flows. Another possibility to simulate the turbulent flows is a Large Eddy Simulation (LES), where the big energy containing turbulent scales are resolved, while the small ones are modeled. In this case the computational costs increase very quickly with increasing Reynolds number. Because of these reasons in the last decade so-called hybrid turbulence models became increasingly popular. The underlying idea is to combine the advantages of different modeling approaches, more specifically, to deliver satisfactory results while demanding reduced computational costs. These new models also make it possible to solve complex industrial problems.

The Detached Eddy Simulation (DES) model, which was first proposed by Spalart [10] is the most popular hybrid turbulence model which has also been successfully used for many complex turbulent flow tasks. In this approach a RANS mode is applied near solid boundaries, while in separated flow regions and regions far from the wall the method switches to LES. The crucial issue in the application of DES is the "gray area", in which an undefined modeling zone exists. In this area the solution is neither pure RANS nor pure LES [7].

One of the new hybrid models, the so-called Very Large Eddy Simulation (VLES) model, was provided by Speziale [11]. This hybrid turbulence approach switches seamlessly between fully-modeled RANS and fully-resolved DNS modes depending on the numerical grid resolution. However, the original VLES model damps the Reynolds stress too much and requires a fine mesh resolution. Therefore, Han et al. [7] provided a modified VLES approach. This new hybrid model shows high efficiency and robustness in many applications already on relatively coarse grids [2], [7], [8].

Moving grid systems play an essential role in many engineering fields, for instance, for fluid-structure interaction. In such cases the problems become much more demanding in terms of computational cost [1]. Therefore, a reduction of computing times is especially important. Nowadays numerical investigations of such flows are rare, especially for hybrid turbulent models.

The focus of this study lies on the investigation of $k-\varepsilon$ and $\zeta-f$ VLES models in the context of moving grids. First, the VLES models are validated by computing the separating flow over periodic hills at a Reynolds number of $Re = 10,600$. Finally, the models are applied to investigate the flow over a forced oscillating circular cylinder at a Reynolds number of $Re = 10,000$. This simulation considers five different oscillating frequencies to cover the so-called jump, where abrupt changes in drag and lift coefficient are observed. The results of the simulation are compared to the available experimental data from [5] and numerical results from other studies.

# 2 VERY LARGE EDDY SIMULATION MODEL

## 2.1 CONCEPT

In the last twenty years hybrid turbulence models become more and more popular. They combine the advantages of different basic modeling techniques, such as RANS, LES or DNS.

In this study the Very Large Eddy Simulation (VLES) approach from Chang [2] is investigated. This hybrid turbulence model provides a transition from the fully-modeled RANS to the fully-resolved DNS modes depending on the numerical grid resolution. For the switching between these models in the VLES approach the Reynolds stress of a generalized RANS model is rescaled through a so-called resolution control function $F_r$:

$$\tau_{ij}^{sub} = F_r \tau_{ij}^{RANS}. \tag{1}$$

$F_r$ depends on the two length scales: the turbulent length scale $L_c$ related to the spectral cut-off and the integral length scale $L_i$ ($\propto k^{3/2}/\varepsilon$):

$$F_r = min\left[1, \left(\frac{L_c}{L_i}\right)^{\frac{3}{4}}\right]. \tag{2}$$

$F_r$ takes on a value between one and zero. When $F_r$ approaches zero, then (theoretically) all of the scales are resolved and the VLES model behaves like a DNS. In the near-wall region $F_r \rightarrow 1$, as in the case of a coarser mesh and the model works as a RANS model, similar to the DES concept. In [2] a detailed description of the VLES approach and the resolution control function $F_r$ can be found.

## 2.2 COMPATIBILITY WITH RANS MODELS

The VLES model is well compatible with different RANS turbulence models. In this study a VLES approach based on two different RANS models is used: the $k-\varepsilon$ VLES model based on the Chien $k-\varepsilon$ model [3] and the elliptic-relaxation eddy-viscosity $\zeta - f$ VLES model [9]. The $\zeta - f$ RANS model developed by Hanjalic et al. [9] uses a transport equation for the velocity scale ratio $\zeta = \overline{v^2}/k$ and an equation for the so-called elliptic relaxation function $f$, in addition to the equations for the turbulent kinetic energy $k$ and its dissipations rate $\varepsilon$:

$$\frac{\partial k}{\partial t} + u_j \frac{\partial k}{\partial x_j} = P_k - \varepsilon + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_k}\frac{\partial k}{\partial x_j}\right)\right], \tag{3}$$

$$\frac{\partial \varepsilon}{\partial t} + u_j \frac{\partial \varepsilon}{\partial x_j} = \frac{C_{\varepsilon_1} P - C_{\varepsilon_2}\varepsilon}{T} + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_\varepsilon}\frac{\partial \varepsilon}{\partial x_j}\right)\right], \tag{4}$$

$$\frac{\partial \zeta}{\partial t} + u_j \frac{\partial \zeta}{\partial x_j} = f - \frac{P}{k}\zeta + \frac{\partial}{\partial x_j}\left[\left(\nu + \frac{\nu_t}{\sigma_\zeta}\frac{\partial \zeta}{\partial x_j}\right)\right], \tag{5}$$

$$L^2 \Delta f - f = \frac{1}{T}\left(C_1 + C_2 \frac{P}{\varepsilon}\left(\zeta - \frac{2}{3}\right)\right). \tag{6}$$

The corresponding turbulent viscosity is defined as

$$\nu_t = C_\mu^\zeta \zeta k T, \tag{7}$$

where $T$ is the turbulent time scale and $C_\mu^\zeta$ is a model constant. The coefficients and a detailed description of this model can be found in [8].

This eddy-viscosity-based model yields better results in comparison to other RANS models for wall-bounded flows [8]. Since the predictive accuracy of VLES depends on the specific RANS turbulence model [6], the application of the $\zeta - f$ model as a background RANS model for VLES appears to be promising.

Compared to basic RANS models, in the VLES approach only the formulation of the turbulent viscosity is modified. For example, for the $\zeta - f$ model the turbulent viscosity takes the form

$$\nu_t = F_r C_\mu^\zeta \zeta k T. \tag{8}$$

### 2.3 NEW FILTER WIDTH

The definition of the resolution control function $F_r$ contains the turbulent cut-off $L_c = C_x\Delta$, which depends on the filter width $\Delta$. The filter width determines the value of the resolution control function and therefore implies a kind of interface between RANS and DNS. Thus, the choice of the expression for $\Delta$ may play an important role, especially in the case of strongly anisotropic grids.
In the original VLES model the standard LES filter-width $\Delta_{vol} = (\Delta_x\Delta_y\Delta_z)^{1/3}$ is used. In the present work the behavior on the VLES model is investigated with a modified filter width, the so-called IDDES filter width, which was introduced by Travin [13] for the Detached Eddy Simulation model:

$$\Delta_{IDDES} = min(max[C_w d_w, C_w h_{max}, h_{wn}], h_{max}), \tag{9}$$

where $d_w$ is the distance to the wall, $h_{wn}$ is a grid distance in the wall-normal direction, $h_{max}$ is a maximum local grid spacing and $C_w = 0.15$ is a constant.

## 3 FLOW OVER THE PERIODIC HILL

For validation of the VLES model on a static grid the flow over a two-dimensional periodic hills at a Reynolds number of $Re_b = 10,600$ based on the bulk velocity is considered. In this flow the separation is not clearly determined by the geometry, therefore, it is especially challenging to predict the correct separation point.
The size of the computational domain is defined as $L_x = 9h, L_y = 3.036h$ and $L_z = 4.5h$, where $h$ is the hill height. On the top and bottom edges of the domain the wall boundary condition is applied, while a periodic boundary condition is used in the streamwise and spanwise directions. The computational domain with the 80x100x30 mesh is shown in the figure 1. The results are compared with reference LES data [12] obtained by a highly-resolved simulation with $4.6 \cdot 10^6$ grid nodes.

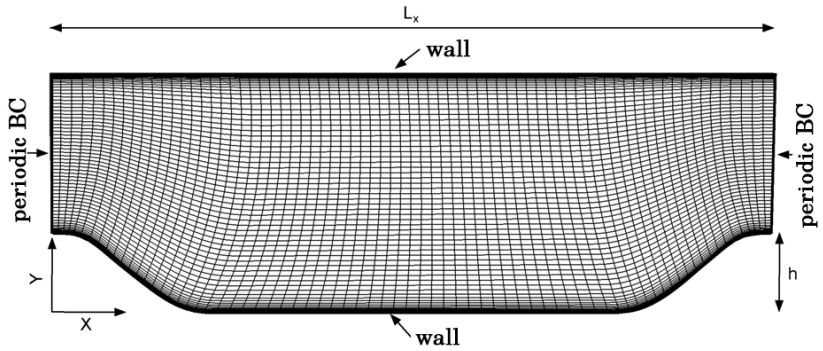



Figure 1: The flow over the two dimensional periodic hills. Computational domain and mesh

The flow separates on the top of the hill and a recirculation zone generates in the downstream of the hill side.The detailed separation and reattachment positions obtained with the different models are shown in the table 1. The VLES approach predicts the position of the separation well even on the coarse grid compared to the LES.

The velocity profiles for the VLES $k-\varepsilon$ model in comparison to the reference data are shown in figure 2. It can be seen that the VLES simulation shows very good agreement with the LES reference results.

| Model | $x_s$ | $x_r$ |
|---|---|---|
| LES [12] | 0.22 | 4.72 |
| VLES $k-\varepsilon$ ($\Delta_{vol}$) | 0.18 | 4.48 |
| VLES $k-\varepsilon$ ($\Delta_{IDDES}$) | 0.20 | 4.65 |
| VLES $\zeta-f$ ($\Delta_{vol}$) | 0.20 | 4.62 |

Table 1: Separation and reattachment points for different models

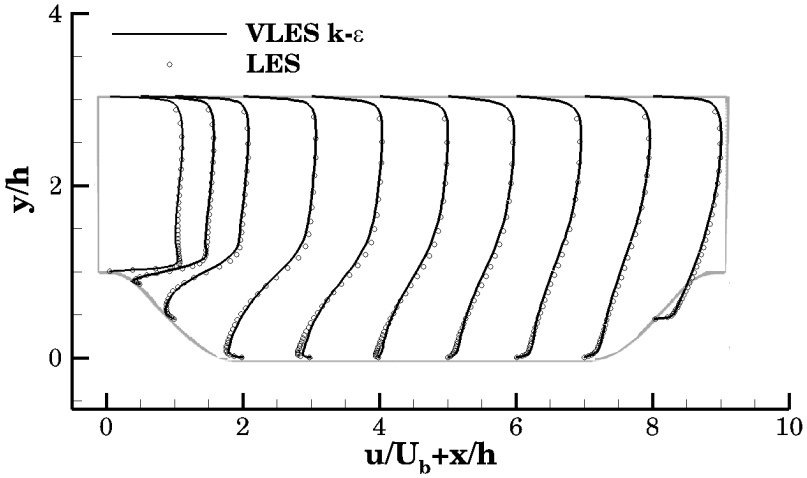



Figure 2: Velocity profiles for the LES and the VLES $k-\varepsilon$ models

With the IDDES filter width the $k-\varepsilon$ VLES approach yields better results than with the standard $\Delta_{vol}$ filter width (see table 1).

## 4 FLOW OVER THE OSCILLATING CYLINDER

Next the flow past an oscillating circular cylinder is investigated. The Reynolds number is $Re = 10,000$ based on the inflow velocity $U$ and the cylinder diameter $D$.

The following equation describes the cylinders displacement in the cross flow direction:

$$y = Y_0 \sin\left(2\pi f_0 t\right), \tag{10}$$

where $Y_0$ is the cylinder displacement amplitude and $f_0$ is the oscillation frequency of the cylinder.

To cover the abrupt changes in drag and lift coefficients in this simulation five different oscillating frequencies $f_0 D/U = 0.14; 0.17; 0.19; 0.21$ and $0.25$ are considered. The results are compared to the experimental data from Gopalkrishnan [5]

The lift and drag coefficients $C_L$ and $C_D$ are presented in figure 3 as a function of the dimensionless frequency for the $k-\varepsilon$ VLES model. This model captures the sharp change in the lift coefficient around the vortex shedding frequency. It under-predicts the value of $C_L$ for oscillation frequencies $f_r > 0.17$, while the value at the low frequency of $f_r = 0.14$ shows good agreement with the experiment data from Gopalkrishnan [5]. For comparison also the results for the lift coefficient for the $k-\varepsilon$ RANS model are given for two frequencies i.e. $f_r = 0.17$ and $f_r = 0.21$. As figure 3 (right) shows, this model yields results that are very similar to the VLES model results for both frequencies.

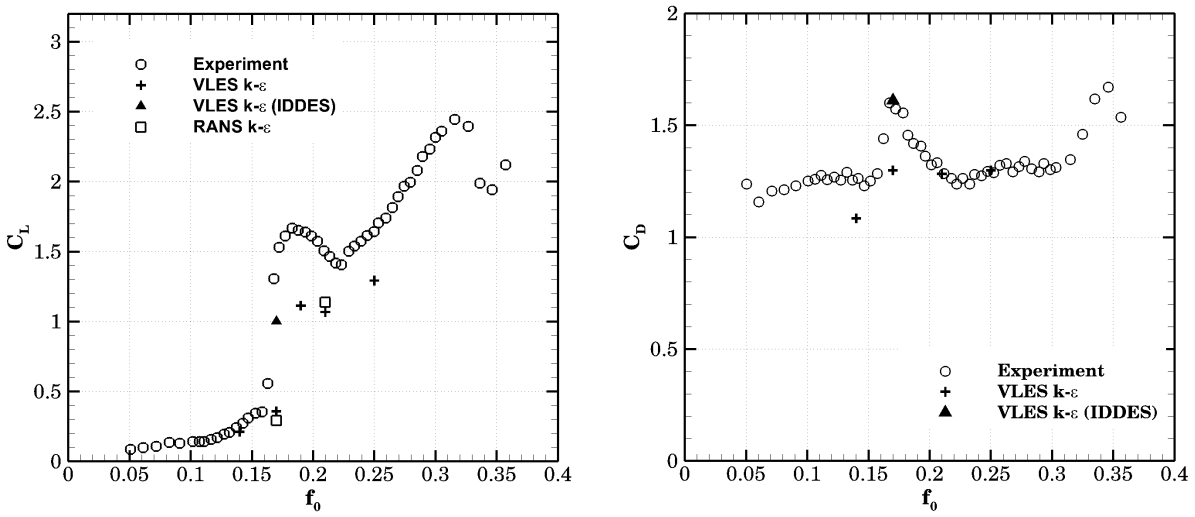



Figure 3: Lift and drag coefficients as a function of non-dimensional frequency for the $k-\varepsilon$ VLES model

To see the influence of the filter width on moving grid simulations, results for the $k-\varepsilon$ VLES with the IDDES filter are given for the oscillation frequency $f_r = 0.17$ in in figure 3. In this case the results show a better agreement with experimental data, in comparison to the standard $k-\varepsilon$ VLES.

The agreement between the simulation results for the VLES model and the experimental data for the drag coefficient is reasonably good (see figure 3, right). The standard VLES under-predicts the drag coefficient at low frequencies, but for higher frequencies the prediction is fairly accurate. While the prediction of the drag coefficient with the IDDES filter width in VLES, shows a very good agreement with the experimental data for the investigated frequency of $f_r = 0.17$.

The results from the $\zeta - f$ VLES model have similar characteristics (figure 4) as for the $k-\varepsilon$ VLES model. The lift coefficient is under-predicted for higher frequencies, while the drag coefficient is in very good agreement with the experimental data, without the application of the modified filter width $\Delta_{IDDES}$.

## 5 CONCLUSION

In this study the $\zeta - f$ and $k-\varepsilon$ VLES models were investigated for the moving grids. Firstly, both models are validated with a static grid test-case. The flow over a periodic hill at $R_e = 10,600$. It was shown that the results of the VLES model can be improved by the introduction of the new IDDES filter width instead of the standard volume filter width. Subsequently both models were applied on a moving grid test-case, where the flow past a harmonically oscillating circular cylinder has been investigated at a Reynolds number of $Re = 10,000$. It was shown that the $\zeta - f$ and $k-\varepsilon$ VLES model can predict the abrupt changes in drag and lift coefficients. The values of the lift coefficient were under-predicted for both VLES models, while the values for the drag coefficient were in very good agreement with the reference data. The results on the moving grid, as on the static grid, can be improved by applying the IDDES filter width in the expression for the resolution control function $F_r$.

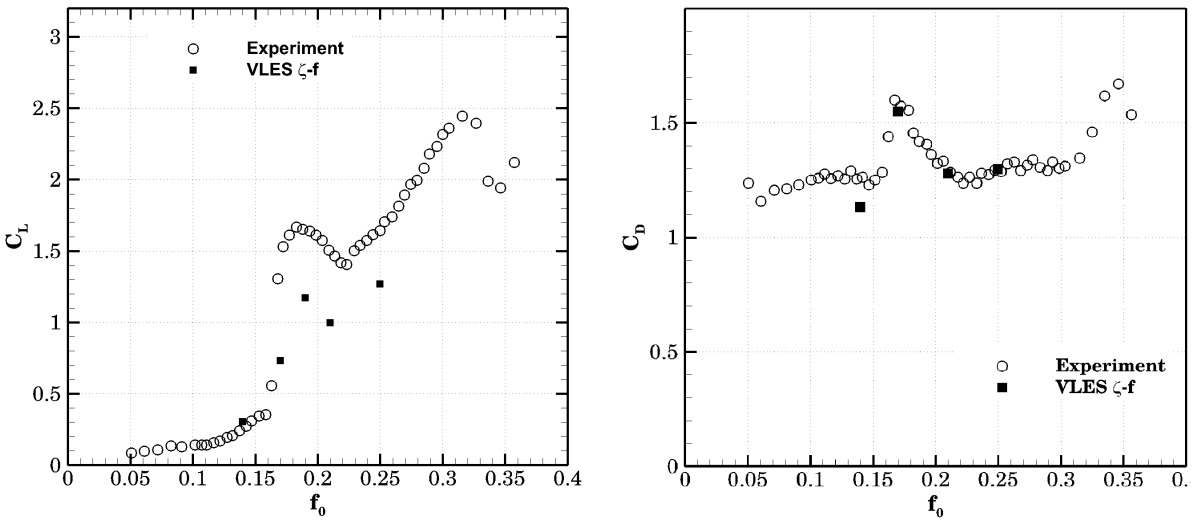



Figure 4: Lift and drag coefficients as a function of non-dimensional frequency for the $\zeta - f$ VLES model

## 6 ACKNOWLEDGMENT

This work is supported by the Excellence Initiative of the German Federal and State Governments within the Graduate School of Computational Engineering at Technische Universität Darmstadt.

## REFERENCES

[1] Bunge, U. Numerische Simulation turbulenter Strömungen im Kontext der Wechselwirkung zwischen Fluid und Struktur. *Univ.-Verlag der TU, Univ.-Bibliothek*, 2005

[2] Chang, C. Y., Jakirlic, S., Dietrich, K., Basara, B. and Tropea, C. Swirling flow in a tube with variably-shaped outlet orifices: An LES and VLES study. *International Journal of Heat and Fluid Flow*, 49 pp. 28-42. 2014

[3] Chien, K.-Y. Predictions of Channel and Boundary-Layer Flows with a Low-Reynolds-Number Turbulence Model, *AIAA Journal*, Vol. 20, No. 1, pp. 33-38, 1982.

[4] Dong, S., and Karniadakis, G. E. DNS of flow past a stationary and oscillating cylinder at. *Journal of Fluids and Structures*, 20.4 pp. 519-531, 2005.

[5] Gopalkrishnan, R. Vortex-induced forces on oscillating bluff cylinders. *WOODS HOLE OCEANOGRAPHIC INSTITUTION MA*, No. WHOI-92-38, 1993.

[6] Han, X., and Krajnović, S. A new very large eddy simulation model for simulation of turbulent flow. *Progress in Hybrid RANS-LES Modeling*. Springer Berlin Heidelberg, pages 131-140, 2012.

[7] Han, X. and Krajnović, S. Validation of a novel very large eddy simulation method for simulation of turbulent separated flow. *International Journal for Numerical Methods in Fluids*, 73(5), pages 436-461, 2013

[8] Han, X. and Krajnović, S. An efficient very large eddy simulation model for simulation of turbulent flow. *International Journal for Numerical Methods in Fluids*, 71(11), pages 1341-1360, 2013

[9] Hanjalić, K., Popovac, M. and Hadžiabdić, M. A robust near-wall elliptic-relaxation eddy-viscosity turbulence model for CFD. *International Journal of Heat and Fluid Flow*, 25(6), pages 1047-1051, 2004

[10] Spalart, P. R., Jou, W. H., Strelets, M., and Allmaras, S. R. Comments on the feasibility of LES for wings, and on a hybrid RANS/LES approach. *Advances in DNS/LES, 1*, pages 4-8, 1997

[11] Speziale, C. Turbulence modeling for time-dependent RANS and VLES: a review.*AIAA Journal*, 36(2), pages 173-184, 1998

[12] Temmerman, L. and Leschziner, M. A. Large eddy simulation of separated flow in a streamwise periodic channel constriction. *TSFP DIGITAL LIBRARY ONLINE. Begel House Inc.*, 2001.

[13] Travin, A. K., Shur, M. L., Spalart, P. R. and Strelets, M. K. Improvement of delayed detached-eddy simulation for LES with wall modelling. *ECCOMAS CFD 2006: Proceedings of the European Conference on Computational Fluid Dynamics, Egmond aan Zee, The Netherlands, September 5-8*, 2006. Delft University of Technology; European Community on Computational Methods in Applied Sciences (ECCOMAS), 2006.

VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL STUDY OF TWO OPTIMIZED COUPLING INTERFACE TREATMENTS FOR STEADY CONJUGATE HEAT TRANSFER PROBLEMS

**M-P Errera[1] and R. El Khoury[2]**

[1] ONERA - The French Aerospace Lab - DMFN -
29 Avenue de Division Leclerc 92322 Châtillon Cedex, France
e-mail: marc.errera@onera.fr

[2] Centre for energy efficiency of systems (CES), MINES ParisTech, PSL Research University,
5 rue Léon Blum, 91120 Palaiseau, France
e-mail: roch.khoury@gmail.com

**Keywords:** Conjugate heat transfer, stability, optimal, fluid-solid, Dirichlet, Neumann, Robin.

**Abstract.** *This paper presents two transmission interface treatments, Dirichlet-Robin and Neumann-Robin procedures, that may be employed for conjugate heat transfer problems. These conditions are analyzed on the basis of a 1D simplified model problem. In the first part of the paper, the Dirichlet-Robin procedure is presented. This interface treatment is the most widely employed in the literature. The same analysis is then performed with a Neumann-Robin procedure. On the basis of the model problem, the general expression of the amplification factor, the stability bounds and the optimal coefficients are provided. It is shown that the two interface treatments are opposite and complementary. Moreover, the so-called optimal coefficient provides the best results in terms of stability and convergence in the Dirichlet-Robin procedure. A criterion is expressed to choose the most appropriate transmission procedure and its importance is underlined by a test case.*

# 1 INTRODUCTION

Conjugate heat transfer (CHT) analysis refers to the ability to address the thermal interaction between a body and a fluid flowing over or through it. The objective of CHT modeling is the accurate prediction of temperature and heat flux distribution in space and time in a body and on its boundaries. Conjugate heat transfer problems occur whenever the fluid convection and the solid material conduction are taken into account simultaneously. This mutual interaction is becoming increasingly important because in many numerical simulations it is no longer acceptable to consider heat transport in the fluid only, with some ideal conditions at the fluid-solid interface.

Even if coupled procedures are the most correct and reasonable way to use when accurate heat transfer predictions are needed, in most cases, arbitrary relaxation parameters are used to stabilize the coupling procedure. This may have a significant negative impact on the numerical properties (low convergence rate, oscillations, and instabilities). Our goal in this paper is to present two interface conditions derived from a stability analysis and to present their main numerical characteristics.

Recently, we have shown [1] using a 1D thermal model problem that in a coupled system, a numerical transition can be identified. This fundamental result has been derived from a normal mode stability analysis based on the theory of Godunov-Ryabenkii [2][3]. This transition can be regarded as an optimal choice in terms of stability and convergence. Ideally, a Robin condition on each side of the interface should be considered. But in this case, a general Robin-Robin interface condition leads to a too large family of schemes. Thus, in this paper we will confine ourselves to two commonly used conditions :

- Dirichlet-Robin procedure : the temperature coming from the solid is applied on the fluid side and a "relaxed heat flux" is in turn used as a boundary condition for the solid.

- Neumann-Robin procedure : the heat flux coming from the solid is applied on the fluid side and a Robin condition is in turn used as a boundary condition for the solid.

These conditions have been chosen in order to deal with most situations likely to arise. They will be analysed on the basis of a 1D coupled aerothermal model problem. In the first part of the paper, Dirichlet-Robin interface treatments are presented. This transmission interface treatment is the most widely employed in the literature. The same analysis is then performed in the second part with a Neumann-Robin procedure. These two treatments will then be compared.

# 2 MODEL EQUATIONS

## 2.1 Model problem

It has been often been stressed that the nature of instabilities derived from a 1D model can give insight into the potential instabilities in 2D/3D computations. As a result, the behavior of interface conditions in CHT is often studied using a normal mode analysis. This is because one may reasonably assume that the modes that may be unstable are those whose variation is in the direction normal to the coupled interface. A 1D model is composed of two partitions with a shared interface. Partitioned techniques [4][5] are very popular, because they allow the direct use of specifically designed solvers for different fields and may offer significant benefits in terms of efficiency over the monolithic techniques. In this study, the fluid-solid system will be decomposed into two partitions: the (+) subdomain **(**$\Omega_+$, index $j \geq 0$ ; $x \geq 0$ ) and the (–) subdomain **(**$\Omega_-$, index $j \leq 0$ ; $x \leq 0$ ) as illustrated in Figure 1.

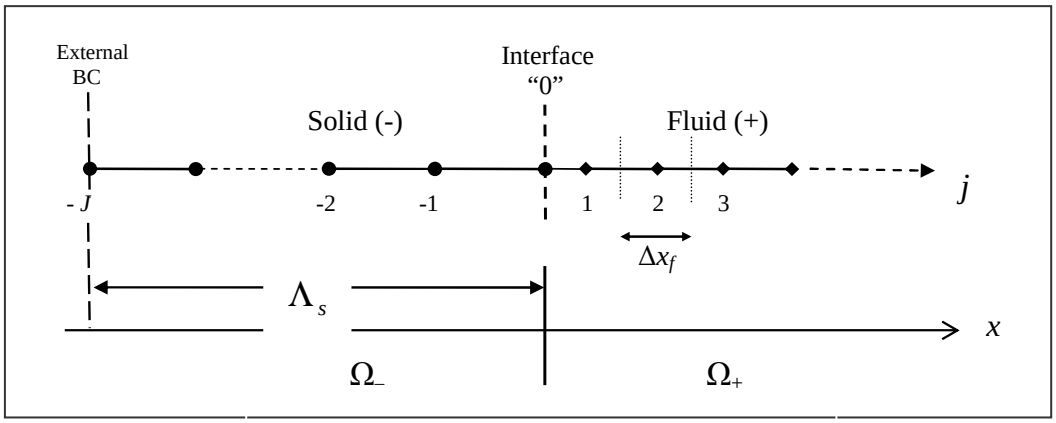



Fig.1 : Schematic of the fluid & solid domains

## 2.2 Interface conditions on the solid side

Interface conditions are needed on either side of the shared interface ($x$ = 0), where coupling conditions are applied. Our goal is to ensure a stable CHT process and to avoid destabilizing effects. It is well known that Robin conditions have many attractive features since such a condition introduces an interface stiffness forcing the boundary to behave in the same way as the boundary of the other domain resulting in much better stability properties. A Robin condition on the solid side is written simply as

$$\left[\hat{Q}_s + \alpha_f \hat{T}_s \right] = \left[Q_f + \alpha_f T_f \right] \tag{1}$$

The subscripts *f* and *s* denote the fluid and solid domain respectively and the (^) notation indicates the sought values. $Q$ is the interfacial heat flux (W.m$^{-2}$) and $T$ is the interface temperature. The general Robin condition (1) introduces the numerical coupling parameter $\alpha_f$ (W.m$^{-2}$.K$^{-1}$) the choice of which directly influences the stability of the coupling process.

## 2.3 Interface conditions on the fluid side

A similar expression could be used on the fluid side

$$\left[\hat{Q}_f + \alpha_s \hat{T}_f \right] = \left[Q_s + \alpha_s T_s \right] \tag{2}$$

Condition (2) introduces another coefficient $\alpha_s$ computed in the solid side and implemented in the fluid side. The simultaneous consideration of two coefficients could be complicated and thus it has been decided in this paper to focus on two complementary treatments :

A Dirichlet condition on the fluid side, obtained easily by setting $\alpha_s = \infty$ in (2)

$$\hat{T}_f = T_s \tag{3}$$

and a Neumann condition on the fluid side, obtained by setting $\alpha_s = 0$ in (2). We readily obtain

$$\hat{Q}_f = Q_s \tag{4}$$

### 2.4 Equations in the fluid and solid domains

The CHT strategy used to obtain rapidly a fluid-structure steady solution relies on the significant discrepancy of the characteristic physical times of the two domains, namely a fast transient process in the fluid, a very slow one in the structure. In the fluid subsystem, the Navier-Stokes (NS) equations are solved generally to steady-state by a time marching scheme. As a consequence, a time marching scheme will be employed here in the fluid domain in the CHT model.

On the contrary, in the solid domain, if we are only interested in computing the steady-state solution, then a second order ordinary differential equation can be solved directly and coupled to the unsteady fluid domain. Thus, only a Laplace equation is considered in the solid.

### 2.5 Characteristic factors

Three dimensionless numbers play a major role in CHT problems and appear in the stability analysis performed in this paper.

The first one is the mesh Fourier number defined as follows

$$D_f = a_f \Delta t_f \big/ \Delta x_f^2 \tag{5}$$

where $a_f$ is the fluid diffusivity ($m^2.s^{-1}$). $D_f$ is the mesh Fourier number that characterizes transient heat diffusion in the fluid domain.

The second one is a conventional Biot number defined by

$$Bi = h/K_s \tag{6}$$

where $h$ ($W.m^{-2}.K^{-1}$) is the heat transfer coefficient and $K_s$ is the thermal conductance of the solid. $Bi$ determines whether or not the temperatures inside a solid body vary significantly in space, while the body heats or cools over time, from a thermal gradient applied to its surface.

The third number, a local Biot number, naturally appears in the stability analysis performed hereafter and takes the following form

$$Bi^{(\Delta)} = K_f / K_s \tag{7}$$

$Bi^{(\Delta)}$ is a "new" parameter that can be regarded as the ratio of the thermal conductance of the 1st fluid cell over the thermal conductance of the whole solid domain. This number represents the balance at each coupling time between the solid resistance of the entire solid body (only a steady-state is considered in this domain) and the resistance of the transient fluid domain represented by the diffusion in the first near wall interface.

It is also interesting to introduce a normalized Fourier number $\overline{D}_f$ (see [6] for more details) defined by

$$\overline{D}_f = \frac{D_f}{1 + D_f + \sqrt{1 + 2D_f}} \tag{8}$$

The domain of $\overline{D}_f$ is $D_f \geq 0$ and the range is $0 \leq \overline{D}_f < 1$.

### 2.6 Normal mode analysis

We have performed a stability analysis by considering normal mode solutions in the form

$$T_j^n = \begin{cases} z^{n-1}\kappa_f^j, & j > 0 \\ z^n \kappa_s^j, & j \le 0 \end{cases} \tag{9}$$

where $z$ and $\kappa$ are two complex functions. $z$ is the "temporal amplification factor" and $\kappa$ is the "spatial amplification factor". The stability analysis is very similar to the standard Fourier stability method except that the Fourier analysis ignores boundary conditions and as these may affect stability, the theory of Godunov & Ryabenkii [2][3] is preferable. After elementary transformations not reported in this paper (see details in [1]), we find that the temporal amplification factor $z$ (i.e. each mode increases in amplitude by a ratio $|z|$) takes the following forms in the two interface conditions considered in this paper :

(1) Dirichlet-Robin interface condition : $(\alpha_f, \alpha_s) = (\alpha_f \ge 0, \infty)$

The temporal amplification factor can be written as

$$z = g(z, \alpha_f) = \frac{1}{K_s + \alpha_f}\left[\kappa_f(z, \alpha_f) - (K_f - \alpha_f)\right] \tag{10}$$

(2) Neumann-Robin interface condition : $(\alpha_f, \alpha_s) = (\alpha_f > 0, 0)$

The temporal amplification factor can be written as

$$z = g(z, \alpha_f) = \frac{1}{K_s + \alpha_f}\left[\alpha_f . \kappa_f(z, \alpha_f) + \frac{K_f - \alpha_f}{Bi^{(\Delta)}}\right] \tag{11}$$

### 2.7 A fundamental transition

The complex function $g$ is a rather complicated nonlinear equation for $z$. After the change of variable $z \to 1/z$, this function becomes holomorphic on the open set $|z| < 1$. As a result, on the basis of the maximum modulus principle in complex analysis, the maximum value of $|g|$ is achieved on the boundary, i.e., at some point on the unit circle $|z| = 1$.

Actually, the maximum is obtained either at $z = +1$ or $z = -1$. But, under certain conditions, there is a sudden transition from one point to another resulting in an amplification factor composed of two half-lines with a singular point at the intersection of these two lines. At this intersection the amplification factor turns back and attains its absolute minimum always located in the stable zone $|z| < 1$.

The point where the maximum is transferred from $z = -1$ to $z = +1$ is a fundamental transition in the aerothermal coupling.

## 3 DIRICHLET-ROBIN INTERFACE CONDITIONS

### 3.1 Stability bounds

The stability condition $\left|g(z,\alpha_f)\right|<1$ applied to Eq.(10) leads, after some basic calculus manipulations, to a lower stability bound $\alpha_f^{\min}$

$$\alpha_f > \alpha_f^{\min} = \frac{K_f}{2}(1-\overline{D}_f) - \frac{K_s}{2} \tag{12}$$

### 3.2 A dynamical Biot number

Substituting the definition of $Bi^{(\Delta)}$ (Eq.(7)), $\alpha_f^{\min}$ becomes

$$\alpha_f^{\min} = \frac{K_s}{2}\left[Bi^{(\Delta)}(1-\overline{D}_f)-1\right] \tag{13}$$

As the coupling coefficient is always positive, two zones are clearly identified :

- $Bi^{(\Delta)}(1-\overline{D}_f)\le 1$ : the coupling process is stable $\forall \alpha_f \ge 0$
- $Bi^{(\Delta)}(1-\overline{D}_f)\ge 1$ : the coupling procedure exhibits a lower stability bound $\alpha_f^{\min}$

This demonstrates how stability depends mainly on the ratio of thermal resistances, but also on the dynamics of the transient fluid system. The higher the local Biot number, the more difficult it will be to stabilize the coupling. Of course, the transient effects of the fluid system also play a major role and can stabilize any process via the term $(1-\overline{D}_f)$, but at a significant cost.

### 3.3 Optimal procedure

It is very noteworthy that the modulus of the amplification factor does not have a monotonic variation (in terms of $\alpha_f$) but goes through an absolute minimum, denoted $\alpha_f^{opt}$. In other words, the existence of a transition value for $\alpha_f$ can be identified. At this transition value, the shape of the curve of the amplification factor switches and turns back as can be seen in Figure 1. In this Figure, the two curves have been plotted using the following characteristics :

- Local Biot number $Bi^{(\Delta)} = 1.25\ \ 10^2$
- Mesh Fourier numbers : (1) $D_f = 40 \Rightarrow\ \overline{D}_f = 0.8$ (2) $D_f = 40{,}000 \Rightarrow\ \overline{D}_f = 0.992$

Thus, the stabilizing effect of the term $(1-\overline{D}_f)$ is noticeable. The most striking point is that at the remarkable coefficient $\alpha_f = \alpha_f^{opt}$, the modulus of the amplification attains its absolute minimum value, always less than one.

Low values of $\alpha_f$ ($\alpha_f < \alpha_f^{opt}$) will result in a rapid convergence, but in this case however $\overline{D}_f$ must be sufficiently large to allow heat diffusion on the fluid side. Otherwise, a low Fourier number will soak up a lot of heat. It will be then necessary to enhance stability by increasing the coupling parameter.

But, likewise, large values of $\alpha_f$ ($\alpha_f >> \alpha_f^{opt}$) will always lead to an extremely slow convergence since this corresponds to a slow diffusion of heat through the fluid subsystem. But it should be pointed out that relatively large Fourier numbers ($\overline{D}_f \approx 1$) indicate fast propagation and energy in this case will be unnecessarily frozen by $\alpha_f$.

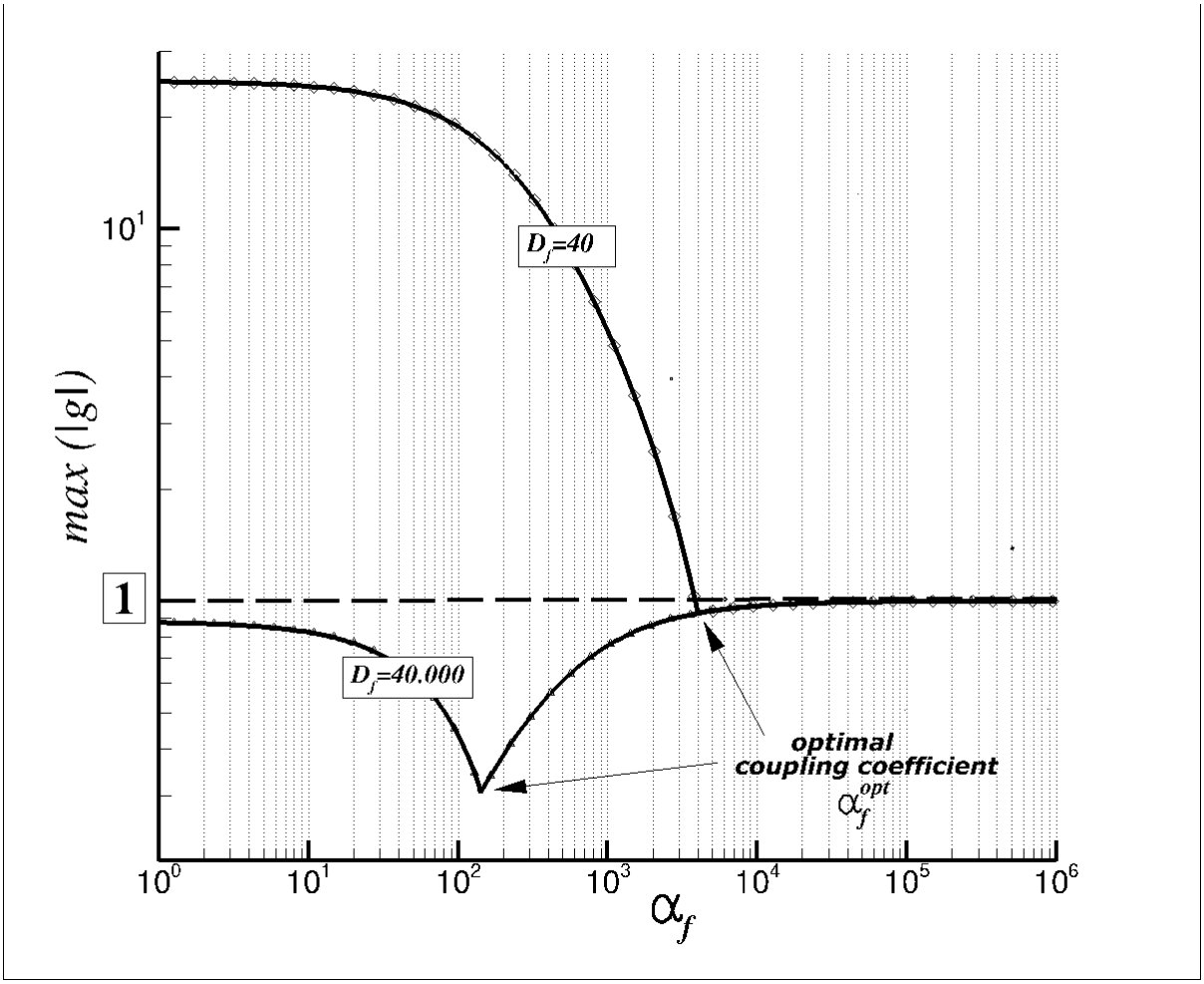



Figure 1: Amplification factor for $D_f = 40$ & $D_f = 40.000$
Dirichlet-Robin procedure

It should be noted that these situations might occur in the same coupled computation or even in the same fluid-structure interface. It is the reason why it is crucial to use local coupling coefficients and it has been shown that $\alpha_f = \alpha_f^{opt}$ is the optimal choice in the case of the model equation adopted herein.

### 3.4 Summary

The general behavior of the Dirichlet-Robin coupling procedure for steady CHT, in terms of $\alpha_f$ is illustrated in Table 1.

| $\alpha_f$ | 0 | | $\alpha_f^{\min}$ | | $\alpha_f^{opt}$ | | $\infty$ | **condition** |
|---|---|---|---|---|---|---|---|---|
| $\lvert g \rvert$ | | **UNSTABLE** | 1 | ↘ | $g^{opt}$ | ↗ | 1 | $Bi^{(\Delta)}(1-\overline{D}_f) \geq 1$ |
| $\lvert g \rvert$ | $g^0$ | ↘ | | | $g^{opt}$ | ↗ | 1 | $Bi^{(\Delta)}(1-\overline{D}_f) \leq 1$ |

Table 1: Numerical behavior of Dirichlet-Robin procedure vs $\alpha_f$ ( $\alpha_s = \infty ; \alpha_f \geq 0$)

## 4 NEUMANN-ROBIN INTERFACE CONDITIONS

### 4.1 Stability bounds

The stability condition $\left|g(z,\alpha_f)\right| < 1$ applied to Eq.(10) leads, after some basic calculus manipulations, to an upper stability bound $\alpha_f^{\max}$

$$\alpha_f < \alpha_f^{\max} = \frac{2K_s K_f}{K_s - K_f(1+\overline{D}_f)} \tag{14}$$

Substituting the definition of $Bi^{(\Delta)}$ (Eq.(7)), $\alpha_f^{\max}$ becomes

$$\alpha_f^{\max} = \frac{2K_f}{1 - Bi^{(\Delta)}(1+\overline{D}_f)} \tag{15}$$

As the coupling coefficient is always positive, two stability regions are highlighted :

- $Bi^{(\Delta)}(1+\overline{D}_f) \geq 1$ : the coupling process is stable $\forall \alpha_f \geq 0$
- $Bi^{(\Delta)}(1+\overline{D}_f) \leq 1$ : the coupling procedure exhibits an upper stability bound $\alpha_f^{\max}$

### 4.2 Optimal procedure

This time, in the framework of the Neumann-Robin procedure, three zones can be identified. The first zone exhibits an upper stability limit as just mentioned. There is also in this zone an optimal coefficient for which the modulus of the amplification factor attains an absolute minimum.

The second zone is very narrow. It also presents an optimal coefficient, but in contrast to the previous case, there is no stability bound.

The third zone is also unconditionally stable but no optimal coefficient can be defined. That is to say the minimum of the amplification factor is obtained for $\alpha_f = \infty$, i.e. for the Neumann-Dirichlet procedure (heat flux imposed on the fluid side & temperature imposed on the solid side) with no relaxation. Another way of explaining it, is that the amplification factor is a monotonic function.

### 4.3 Summary

The general behavior of the Neumann-Robin coupling procedure for steady CHT, in terms of $\alpha_f$ is illustrated in Table 2.

| $\alpha_f$ | 0 | | $\alpha_f^{opt}$ | | $\alpha_f^{\max}$ | | $\infty$ | **condition** |
|---|---|---|---|---|---|---|---|---|
| $\lvert g\rvert$ | 1 | ↘ | $g^{opt}$ | ↗ | 1 | **UNSTABLE** | | $Bi^{(\Delta)}(1+\overline{D}_f)\le 1$ |
| $\lvert g\rvert$ | 1 | ↘ | $g^{opt}$ | → | | | $\lvert g(\infty)\rvert$ | $1\le Bi^{(\Delta)}(1+\overline{D}_f)\le 2$ |
| $\lvert g\rvert$ | 1 | → | | | | | $\lvert g(\infty)\rvert$ | $Bi^{(\Delta)}(1+\overline{D}_f)\ge 2$ |

Table 2: Numerical behavior of Neumann-Robin procedure vs $\alpha_f$ ($\alpha_s = 0$ ; $\alpha_f \ge 0$)

## 5 COMPARISON OF THE TWO INTERFACE CONDITIONS

### 5.1 Stabilizing effect of the Fourier number

As just stated, the two interface conditions considered in this paper may be regarded as two complementary conditions, and it seems that we can go as far as to say that they are almost opposite. The first condition is unconditionally stable at small Biot number while the second one is unconditionally stable at large Biot numbers. But there is one major difference. In the case of the Dirichlet-Robin procedure, the term $(1-\overline{D}_f)$ may become as small as necessary and as a result the unconditional stability condition $Bi^{(\Delta)}(1-\overline{D}_f)\le 1$ ($2^{\text{nd}}$ row of Table 1) can always be satisfied. In other words, the lower stability bound can always be removed by an appropriate choice of $\overline{D}_f$. Conversely, in the condition $Bi^{(\Delta)}(1+\overline{D}_f)\le 1$ ($1^{\text{st}}$ row of Table 2), the term $(1+\overline{D}_f)$ has no stabilizing effect.

### 5.2 Overlapping zone

It is not quite accurate to say that the two interface conditions are opposite. Actually they exhibit an overlapping zone where both may be considered as stable. On the basis of the stability limits presented in the preceding sections, it is seen that there is an overlapping zone where both procedures are unconditionally stable and defined by

$$\frac{1}{1+\overline{D}_f} \le Bi^{(\Delta)} \le \frac{1}{1-\overline{D}_f} \tag{16}$$

This overlapping zone gets narrower as the normalized Fourier number $\overline{D}_f$ gets smaller. On the contrary, this zone is significantly extended for large Fourier numbers.

### 5.3 Existence and non-existence of optimal procedures

The two interface conditions presented in this paper exhibit 5 different zones (2 in Table 1 & 3 in Table 2). In one of them ($3^{\text{rd}}$ row of Table 2), there is no optimal coefficient. This does however not mean that the coupling procedure will converge slowly but simply that the condi-

tion $\alpha_f = \infty$ is a relevant choice. In other words the Neumann-Dirichlet transmission is recommended. In the first procedure, the optimal choice requires a certain amount of relaxation and this time a Dirichlet-Robin procedure is recommended in conjunction with the use of the optimal coefficient.

Actually, there exist two types of optimal procedures. The first one is not subject to any condition and always exhibits an optimal coefficient. The second one presents a conditional optimal coefficient. Clearly, the transient effects of the fluid system play a major role in the stability analysis.

### 5.4 Criterion for a relevant interface treatment

We have just seen how to optimize a specific transmission condition procedure. But this study suggests that it can be interesting to select adequately the most relevant procedure and not necessarily to optimize a particular one, more or less arbitrarily chosen. Two options might be considered :

(1) A "dynamic" option

The dynamic option is supported by the fact that the optimal coefficient of the Dirichlet-Robin condition is able to stabilize any procedure.

- $Bi^{(\Delta)}(1-\overline{D}_f) \le 1$ **:** Dirichlet-Robin procedure with $\alpha_f = \alpha_f^{opt}$
- $Bi^{(\Delta)}(1-\overline{D}_f) > 1$ :

  if $(\alpha_f^{opt} >> K_s)$ **:** Neumann-Robin procedure with $\alpha_f$ sufficiently great
  else
  Dirichlet-Robin procedure with $\alpha_f = \alpha_f^{opt}$
  endif

Note that in the overlapping zone, the procedure with the term $(1-\overline{D}_f)$ has been retained.

(2) A "static" option

The other option does not take into consideration the dynamics of the fluid system ($\overline{D}_f = 0$). The preceding conditions thus become :

- $Bi^{(\Delta)} \le 1$ : Dirichlet-Robin procedure with $\alpha_f = \alpha_f^{opt}$
- $Bi^{(\Delta)} > 1$ :
  if $(K_f >> K_s)$ : Neumann-Robin procedure with $\alpha_f$ sufficiently great
  else
  Dirichlet-Robin procedure with $\alpha_f = \alpha_f^{opt}$
  endif

The first option takes advantage of the stabilizing effect of the mesh Fourier number in the transient fluid domain. The second option is more secure especially when the ratio $Bi^{(\Delta)}$ is very high, for instance in CFD computations in which a high wall resolution is adopted. In

this case, it is always possible to exploit to advantage the term $(1-\overline{D}_f)$ such that $Bi^{(\Delta)}(1-\overline{D}_f) \leq 1$ but this can have a substantial cost.

### 5.5 Link between the local and the conventional Biot numbers

This local Biot number introduced by Eq. (7) takes directly into account the thermal and dynamic response of the boundary layer and directly participates in the stability of the coupled process. If a transient fluid flow is employed during the CHT computation it is this number that will drive and control the stability of the CHT problem.

On the other hand, the conventional Biot defined by Eq.(6) is a measure of the resistance to heat flow within the solid relative to the resistance presented by the convection processes at the surface. This parameter is also a key parameter, but only at steady state since it determines the stability of the fluid-solid equilibrium. However, it is clear that this conventional Biot number, not easily defined during the transients, cannot be used to set up a numerical CHT procedure, as long as a transient fluid state is involved in this procedure.

At steady state, the thermal fluid resistance is equal to 1/*h* and thus the various stability criteria and stability bounds presented in this paper may easily be extended to steady state by simply imposing $\overline{D}_f = 0$ and $K_f = h$.

### 5.6 Numerical example

A complex CHT computation has been performed in a 3D test case representing an industrial reheat furnace with heteroclite fluid-structure interface (details can be found in [7]). As a first step, the Dirichlet-Robin condition has been used and a systematic and comparative study of various coupling coefficients has been undertaken. This first study showed that the optimal coupling coefficient presented in this paper was by far, the best-performing procedure.

But by looking more closely at the fluid-solid interfaces of the furnace, high local Biot number can be found, for instance

$$Bi^{(\Delta)} = \frac{K_f}{K_s} = 1.25\ 10^2 \tag{17}$$

This ratio strongly suggests the use of a Neumann-Dirichlet procedure as mentioned in § 5.4. Figure 2 presents the convergence history of these CHT tests. All the tests performed are not indicated in this figure. It is sufficient to mention that the so-called, optimal Dirichlet-Robin procedure is always stable and oscillation-free and clearly the most efficient coupling method.

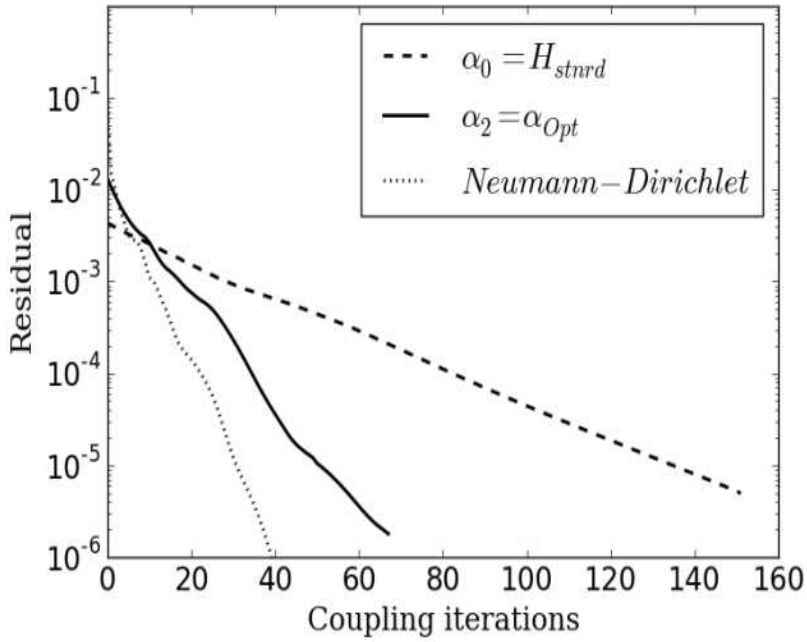



Figure 2 : Convergence history and comparison between Dirichlet and Neumann procedures

But, as suggested by the criterion proposed in this study, it is instructive to see that a relevant procedure, even with no relaxation parameter can be even more performing. This means that when the local Biot number $Bi^{(\Delta)}$ is high, stability can be obtained with the Dirichlet-Robin condition but at the expense of computational efficiency (CPU time). In this case, a natural physics-based approach is to impose a Neumann condition.

## 6 CONCLUSIONS

On the basis of a simplified model problem, the general expression of the amplification factor, the stability bounds, the optimal coefficient and the general numerical behavior have been presented for two different and complementary interface treatments. It was shown that the numerical properties depend on the ratio of fluid and thermal resistances and that the mesh Fourier number plays a crucial role. Furthermore, a local Biot number has been introduced and it was shown that this number "drives" the overall coupling process.

The two interface treatments have been compared and it has been highlighted that these interface treatments are almost opposite and complementary. When the first method is unconditionally stable, the second one exhibits an upper stability bound. When the second method is unconditionally stable, the first one exhibits a lower stability bound. These two interface schemes present an overlapping area where both of them are stable. It is also shown that the so-called optimal coefficient provides the best results in terms of stability and convergence in the Dirichlet-Robin procedure. The numerical criteria establishing the nature and character of the most relevant interface treatment have been identified and expressed. A comparison in a CHT test case has emphasized the importance of a physics-based numerical approach.

## REFERENCES

[1] Errera, M-P and Chemin, S. Optimal solutions of numerical interface conditions in fluid-structure thermal analysis, *Journal of Computational Physics*, **245**, pp 431-455, 2013.

[2] Godunov, SK. Ryabenkii, VS. *The theory of difference schemes*. An introduction. North-Holland: Amsterdam, 1964.

[3] Gustafsson, B. The Godunov-Ryabenkii condition : the beginning of a new stability theory. *Technical report 1999-014*. Department of Information Technology, 1999.

[4] Felippa, C.A. and Park, K.C. Staggered transient analysis procedures for coupled mechanical systems: Formulation. *Computer Methods in Applied Mechanics and Engineering*, **24**:61–111, 1980.

[5] Felippa, C.A. Park, K.C. and Farhat, C. Partitioned analysis of coupled mechanical systems. *Computer Methods in Applied Mechanics and Engineering*, **190**:3247–3270, 2001.

[6] Errera, M-P. and Duchaine, F. Comparative study of Coupling Coefficients in Dirichlet-Robin Procedure for Fluid-Structure AeroThermal Simulations, *Journal of Computational Physics*, *http://dx.doi.org/10.1016/j.jcp.2016.02.022,* 2016.

[7] El Khoury, R. R. Development and validation of coupling methods for the modelling of thermal systems. *PhD Thesis*, Mines Paris Tech, 21 december 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFECT OF DISC GEOMETRY ON THE DYNAMIC STABILITY OF DIRECT SPRING OPERATED PRESSURE RELIEF VALVES (ECCOMAS CONGRESS 2016)

**István Erdődi**[1], **Csaba Hős**[1]

[1]Budapest University of Technology and Economics
Department of Hydrodynamic Systems
1111 Budapest, Műegyetem rkp. 7-9.
e-mail: ierdodi@hds.bme.hu, cshos@hds.bme.hu

**Keywords:** pressure relief valve, dynamic stability, computational fluid dynamics, reduced order modelling

**Abstract.** *The most important parameter of a direct spring operated pressure relief valve is its capacity, which is the rated flow through the valve under conditions given by the corresponding industrial standard. There are several phenomena due to which dynamic instabilities may arise in the system, leading to dangerous oscillations and reduced flow rate. One of the causes of these instabilities is the acoustic coupling of the valve with its upstream piping, the mathematical background of which has already been thoroughly investigated by the researchers at our department. As a continuation of that work, this paper focuses on the engineering applications by proposing various valve disc geometries based on preliminary measurement results, and evaluating their dynamic stability performance in a wide range of parameters. Steady-state CFD computations were performed to determine the mass flow rate and force characteristics of the various valve discs. Through these quantities, the behaviour of each geometry was implemented into our one-dimensional coupled gas dynamical solver, which resolves the pipe dynamics using the one-dimensional continuity-, momentum- and energy equations. The valve itself is modelled as a one degree-of-freedom oscillator. Finally, the stability maps of each geometries were calculated using the gas dynamical model and it was shown that the shape of the fluid force function does indeed have a significant effect on the stable operating range.*

## 1 INTRODUCTION

The main purpose of pressure relief valves is to guarantee that the pressure in the protected system does not exceed a prescribed level, which is the *set pressure* of the valve. The other important parameter is the so-called *capacity*, that is the rated flow through the valve under stated conditions prescribed by the corresponding standard [1]. As this determines how quick the relief process is, it is imperative from the point of safety that this mass flow rate is guaranteed. However, certain instabilities may arise, leading to harmful vibrations and a reduction in the flow rate.

The literature distinguishes between static and dynamic instabilities. The former involves the study of the governing equations of the valve — usually in a linearised form — around its equilibrium positions [2], while the latter deals with more complicated, generally system-level issues [3]. In this paper a dynamic instability caused by the acoustic coupling between the valve and the straight upstream piping is investigated. Previous measurements and simulation results have shown that the stability in this case depends on the driving mass flow rate and the length of the upstream piping for a given valve geometry [4, 5]. The goal of this paper is to provide a numerical method for stability analysis, and also to get an insight into the geometry-dependence of this instability by qualitatively comparing the stability maps in the mass flow rate and pipe length plane for three different valve disc geometries.

## 2 THEORETICAL BACKGROUND

The valve disc can be modelled as a one degree-of-freedom oscillator, for which the equation of motion is

$$m\ddot{x}_\mathrm{v} + k\dot{x}_\mathrm{v} + s(x_\mathrm{v} + x_\mathrm{p}) = F_\mathrm{total}, \tag{1}$$

where $m$ is the reduced mass of the moving parts, $x_\mathrm{v}$ is the valve lift, $k$ is the damping coefficient, $s$ is the spring stiffness, $x_\mathrm{p}$ is the pre-compression of the spring and $F_\mathrm{total}$ is the total force acting on the valve disc. The latter is the sum of the pressure, momentum and viscous forces — its modelling poses a significant challenge, as the momentum and viscous components cannot be accurately approximated by analytical means without a detailed knowledge of the flow-field. To circumvent this, the so-called effective area was employed during our calculations [6, 7, 8], that is

$$F_\mathrm{total} = A_\mathrm{eff}(x_\mathrm{v})(p_\mathrm{v} - p_\mathrm{b}), \tag{2}$$

where $A_\mathrm{eff}(x_\mathrm{v})$ is the effective area function, $p_\mathrm{v}$ is the pressure upstream of the valve and $p_\mathrm{b}$ is the back pressure. This assumes that the total force can be expressed by multiplying the pressure drop on the valve by an equivalent area, which depends only on the valve lift. Note that this approach neglects any unsteady effects, such as those from a non-zero valve velocity. The effective area functions differ for all valve geometries, and as such these are unknown as of now. However, since both the momentum and viscous forces equal zero if the valve is in a closed position, the effective area at that point equals the cross-sectional area of the seat.

During dynamically unstable cases, it is possible that the valve hits the seat. The resultant impact is treated as a mapping in the model with

$$\dot{x}_\mathrm{v}{}^{+} = -r\dot{x}_\mathrm{v}{}^{-}, \tag{3}$$

where $\dot{x}_\mathrm{v}{}^{-}$ and $\dot{x}_\mathrm{v}{}^{+}$ are the valve velocities before and after the impact and $r$ is the coefficient of restitution. For convenience, let the seat be located at $x_\mathrm{v} = 0$. Bouncing occurs if the valve velocity is above a set threshold ($\dot{x}_\mathrm{v}^{-} \geq \dot{x}_\mathrm{th}$), or the velocity is lower than the threshold and the

sum of all forces pushes the valve away from the seat ($\dot{x}_\mathrm{v}^- < \dot{x}_\mathrm{th}$ and $\ddot{x}_\mathrm{v}^- > 0$). Otherwise, the valve sticks to the seat.

The one-dimensional continuity-, momentum-, and energy equations were written for the straight pipe section upstream of the valve, which are

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial \xi} = \mathbf{Q}, \tag{4}$$

$$\mathbf{U} = \begin{pmatrix} \rho \\ \rho v \\ \rho e \end{pmatrix}, \quad \mathbf{F} = \begin{pmatrix} \rho v \\ \rho v^2 + p \\ \rho v e + p v \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} 0 \\ -\frac{f(v)\rho}{2D_\mathrm{pipe}} v|v| \\ 0 \end{pmatrix},$$

where $t$ is the time, $\xi$ is the spatial coordinate, $\rho$ is the density, $v$ is the velocity, $e$ is the specific energy, $f$ is the Darcy friction factor and $D_\mathrm{pipe}$ is the diameter of the pipe. Preliminary tests have shown that treating the friction factor as a constant led to the erroneous estimation of the pressure drop due to the wide Reynolds number range in the pipe, and as such Blasius' correlation for turbulent flows was implemented — this means that the friction factor is a function of the velocity.

At the $\xi = 0$ end an infinitely large reservoir boundary condition was set, which is realized by prescribing the total pressure in the case of inflow to the pipe, and the static pressure otherwise. The valve boundary condition is implemented at $\xi = L$. The basic principle is that the mass flow rate leaving the pipe must also satisfy the discharge equation for the valve. If the flow is chocked, the equation to be solved is

$$\rho(L,t) A_\mathrm{pipe} v(L,t) = C_\mathrm{D} A_\mathrm{ref}(x_\mathrm{v}) \sqrt{\gamma \rho(L,t) p(L,t) \left(\frac{2}{\gamma+1}\right)^{\frac{\gamma+1}{\gamma-1}}}, \tag{5}$$

where $C_\mathrm{D}$ is the discharge coefficient, $A_\mathrm{ref}$ is the reference flow-through area and $\gamma$ is the ratio of specific heats. Similarly, for the non-chocked case we have

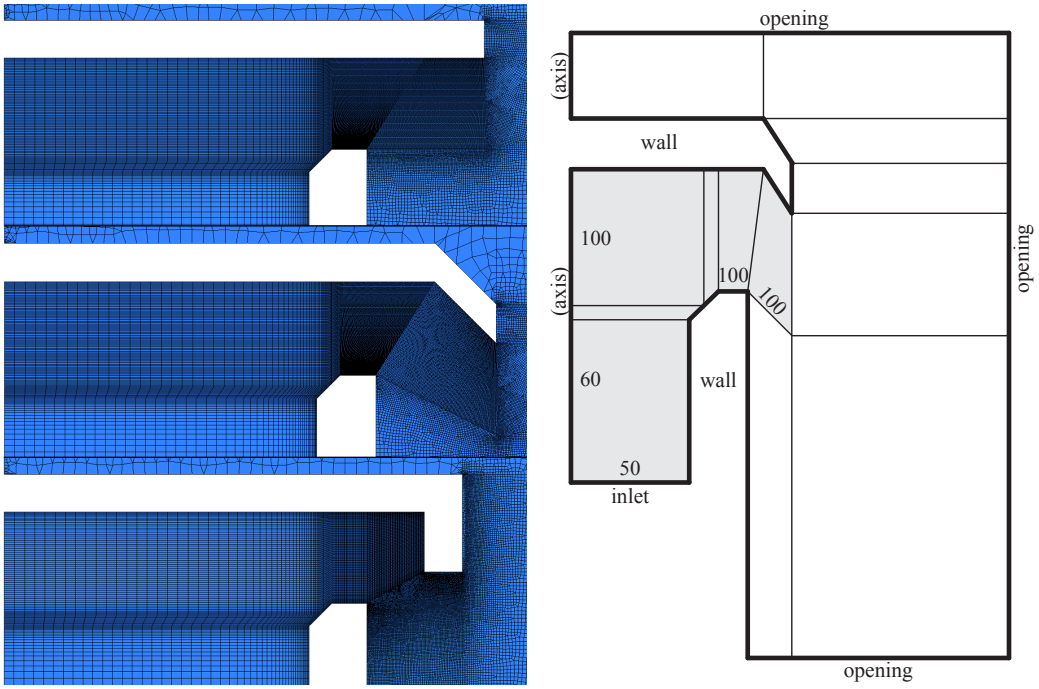
$$\rho(L,t) A_\mathrm{pipe} v(L,t) = C_\mathrm{D} A_\mathrm{ref}(x_\mathrm{v}) \sqrt{2\rho(L,t) p(L,t) \frac{\gamma}{\gamma-1} \left( \left(\frac{p_\mathrm{b}}{p(L,t)}\right)^{\frac{2}{\gamma}} - \left(\frac{p_\mathrm{b}}{p(L,t)}\right)^{\frac{\gamma+1}{\gamma}} \right)}. \tag{6}$$

The discharge coefficient is unknown as of yet, while the cylindrical surface between the seat and the valve disc is taken as the reference area is, that is

$$A_\mathrm{ref}(x_\mathrm{v}) = D_\mathrm{seat} \pi x_\mathrm{v}, \tag{7}$$

where $D_\mathrm{seat}$ is the diameter at the seat. In our case the seat was the chamfered pipe end, therefore $D_\mathrm{seat} = D_\mathrm{pipe}$ (see Figure 1(a)).

Equation (1) is solved using standard Runge-Kutta, while the Lax-Wendroff method is used for the discretization of equation (4). The boundary conditions were implemented with the isentropic method of characteristics.




(a) (b)

Figure 1: (a) The three geometries ($0°$, $45°$ and $90°$ from top to bottom) with the mesh lines at a valve lift of $0.15D_{\text{pipe}}$. (b) The number of nodes and the boundary conditions. The filled parts indicate the structured part of the mesh. (The figure is not proportional for illustration purposes.)

## 3 CFD SETUP

As mentioned in the previous section, two necessary parameters are unknown due to the lack of accurate analytical methods: the effective area function and the discharge coefficient. Their values were determined by means of CFD.

Simulations were conducted for three different geometries, which were generated by varying the collar angle (Figure 1(a)). Axisymmetric treatment of the flow field was possible due to the simplicity of the geometry, effectively rendering the problem two-dimensional and as such significantly reducing the needed computational time. This resulted in a wedge-shaped domain with a central angle of $5°$, meaning that the extensive quantities — namely the mass flow rate and the total force acting on the valve — had to be rescaled for the full-sized valve.

The meshes were automatically generated for all geometries and lifts, with the number of nodes varying between $80\,000$ and $85\,000$. They consist of both structured and unstructured regions according to Figure 1(b) (the modelled domain is similar for the $0°$ and $90°$ cases). The meshes were created in ICEM. The inner diameter of the pipe was $40.2\,\text{mm}$, the upstream piping had a length of $10D_{\text{pipe}}$ and the valve lift was varied between $0.05D_{\text{pipe}}$ and $0.30D_{\text{pipe}}$.

Total pressures between $2$ and $6\,\text{bar}$ were prescribed at the inlet boundary condition with a static temperature of $293\,\text{K}$. The pressure was set to $1\,\text{bar}$ at the opening boundary condition, which corresponds to static or total pressure for outflow and inflow, respectively. The temperature was also $293\,\text{K}$. The medium was air with the ideal gas law and the k-$\epsilon$ turbulence model. The simulations were done in CFX environment. The High Resolution scheme was employed for the advection terms and the turbulence model, and the Second Order Backward Euler for

| | | |
|---|---|---|
| Pipe diameter | $D_{\text{pipe}}$ | 40.2 mm |
| Relative pipe length | $\frac{L}{D_{\text{pipe}}}$ | varied |
| Reduced mass | $m$ | 1 kg |
| Spring stiffness | $s$ | 21000 N/m |
| Discharge coefficient | $C_{\text{D}}$ | geometry-dependent |
| Gas constant | $R$ | 287 J/(kgK) |
| Ratio of specific heats | $\gamma$ | 1.4 |
| Kinematic viscosity | $\nu$ | $1.511 \times 10^{-5}\text{m}^2/\text{s}$ |
| Set pressure | $p_{\text{set}}$ | 3 bar |
| Inlet total pressure | $p_{\text{tot,r}}$ | varied |
| Inlet static temperature | $T_{\text{r}}$ | 293 K |
| Back pressure | $p_{\text{b}}$ | 1 bar |
| Environmental temperature | $T_0$ | 293 K |
| Initial relative valve lift | $\bar{x}(0)$ | 20 |
| Initial valve velocity | $\dot{x}(0)$ | 0 $\frac{\text{m}}{\text{s}}$ |
| Spatial resolution | $n_x$ | 10 |

Table 1: Parameter values

the time stepping. A mesh dependency study was concluded at the lowest and largest valve lifts with the highest inlet total pressure by doubling the mesh resolution on all edges, i.e. quadrupling the number of elements. The relative differences for both the fluid forces and the mass flow rates were below one percent, therefore the results on the original mesh can be deemed as numerically accurate.

The mass flow rates, the pressures at the seat and the total forces acting on the valve disc were extracted from the simulation results. With them, both the discharge coefficients and the effective area functions were determined for all three geometries using Equations (2) and (5).

## 4 RESULTS

The gas dynamical model consisting of Equations (1) and (4) was solved with the parameters listed in Table 1. To create the stability maps, the inlet total pressure was varied between $3.1$ and $3.4$ bar with increments of $1250$ Pa, and the upstream pipe length was varied between $20D_{\text{pipe}}$ and $40D_{\text{pipe}}$ with increments of $0.5D_{\text{pipe}}$. The effective area functions and the discharge coefficients were set based on the fixed-lift CFD simulation results. The former are illustrated in Figure 2, which also shows the functions for incompressible simulations on the same geometries. The error bars represent the dependence on the pressure at the pipe end. The effective area values are the lowest for the $0^\circ$ geometry, which is expected since it has the lowest jet deflection angle — however, the $90^\circ$ results are only higher than the $45^\circ$ ones for lower lifts. It can also be seen that the pressure-dependence is much lower in the incompressible and that compressibility has a strong effect both qualitatively and quantitatively. Note that the incompressible effective areas served only illustrational purposes, all calculations were made with the compressible results. The discharge coefficients were

$$C_{\text{D},0^\circ} = 0.8545, \quad C_{\text{D},45^\circ} = 0.8515, \quad C_{\text{D},90^\circ} = 0.7787. \tag{8}$$

These values indicate that the $90^\circ$ geometry had a significant choking effect, reducing the mass flow rate under similar conditions by around $10\%$.

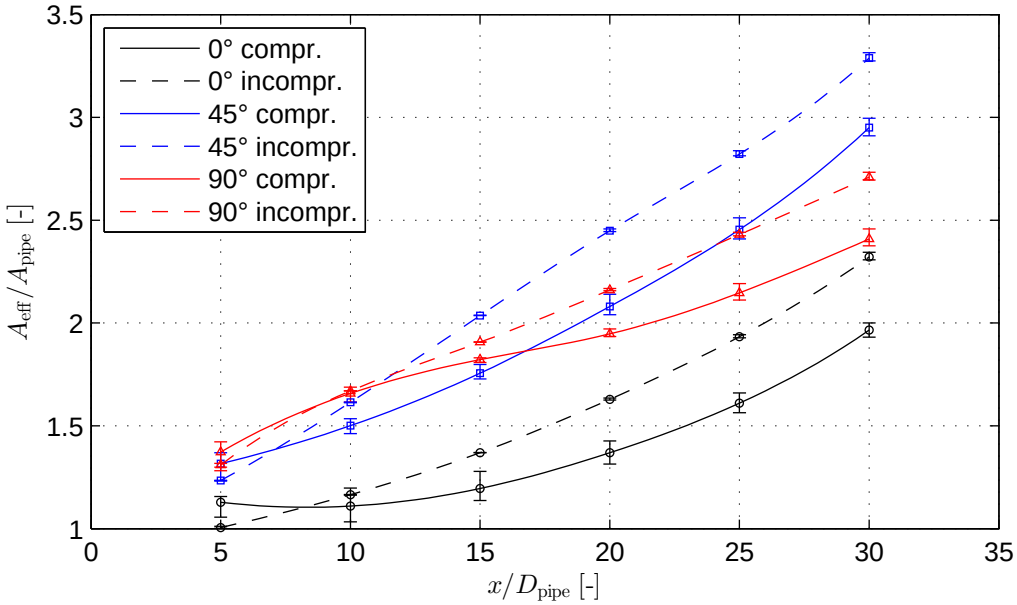


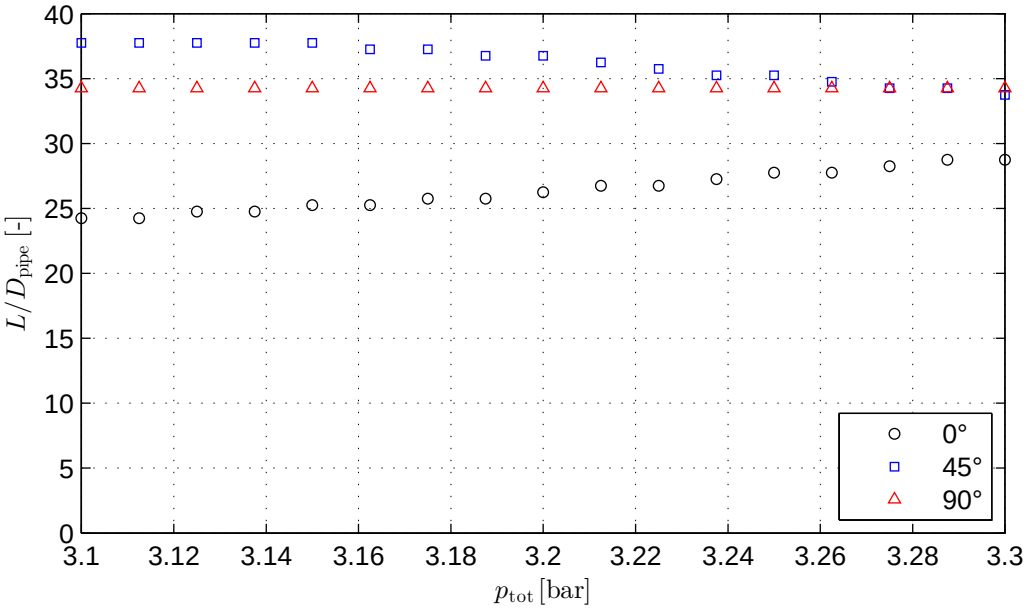

Figure 2: The effective area results for both the compressible (continuous lines) and incompressible (dashed lines) cases.

The results were said to be stable if the valve lift reached the equilibrium position in a decaying manner, and unstable otherwise, i.e. when periodic non-decaying oscillations were produced. First, the stability map was created in the inlet total pressure and upstream pipe length plane, as these were the two varied parameters (Figure 3(a)). The boundary curves differ in the three cases: for the $0°$ geometry, increasing the inlet total pressure extends the stable operating range, while the system behaves the opposite way for the $45°$ disc. In the case of the $90\,\text{deg}$ geometry, the boundary of stability is seemingly independent of the inlet total pressure. This shows that the accurate calculation of the fluid forces introduces not just qualitative, but also quantitative differences in regarding the boundary of stability.

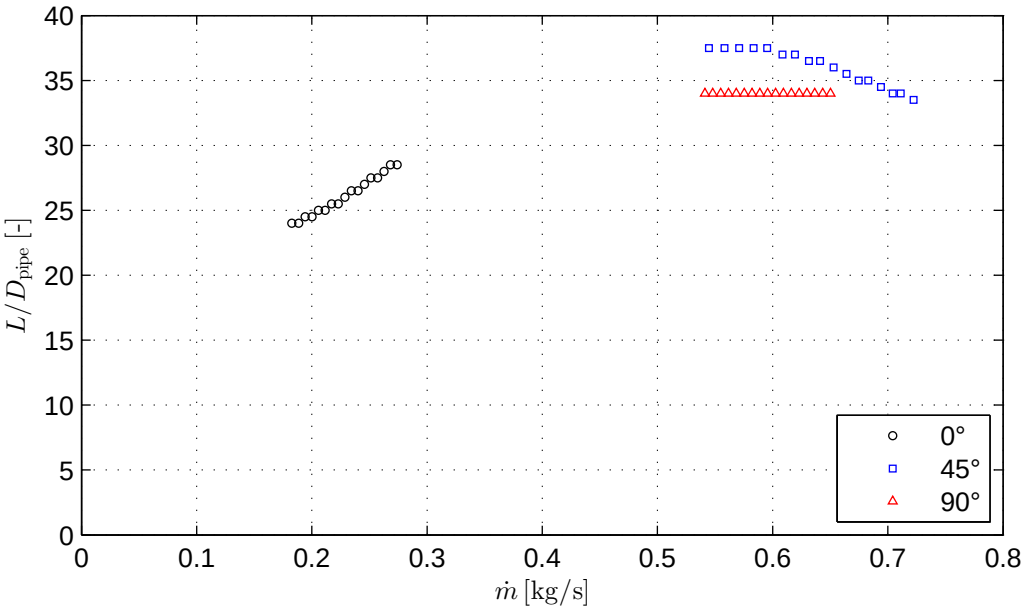
The corresponding industrial standard [1] and the literature [9] usually discusses stability as a function the driving mass flow rate and pipe length. This map can be seen in Figure 3(b). Similar tendencies can be observed to those mentioned above regarding the shape of the curves, however, the difference in the mass flow rates makes the comparison ambiguous — it is possible that at higher flow rates the $0°$ curve becomes flatter (or even decreasing). The lower flow rates can be traced back to the lower effective area values (see Figure 2) as lower fluid forces do not produce as much valve lift.

## 5 CONCLUSIONS

This study employed a relatively simple, one-dimensional model for computing the valve response for a given inlet total pressure. Steady-state CFD simulations were conducted for three different valve disc geometries in order to obtain their force characteristics and discharge coefficients. With these, their specifics were implemented into the reduced model and a parameter study was conducted in the mass flow rate and the inlet total pressure plane. It was shown that even slight changes in the disc geometry can have a significant effect on both the stable operating range and the resulting mass flow rate.




(a)




(b)

Figure 3: Boundaries of stability in the total pressure and pipe length (a) and mass flow rate and pipe length (b) planes. In both cases instability was observed above the boundaries.

## 6 ACKNOWLEDGEMENTS

This project was supported by the OTKA Grant K116549 of Csaba Hős. The authors would also like to express their thanks to Zalán Búcsú and Szabolcs Sánta for providing their incompressible simulation results and numerical meshes.

## REFERENCES

[1] American Petroleum Institute, "API Standard 520: Sizing, Selection, and Installation of Pressure-relieving Devices, Part II—Installation," 2015.

[2] C. Bazsó and C. Hős, "On the static instability of liquid poppet valves," *Periodica Polytechnica Mechanical Engineering*, vol. 59, no. 1, pp. 1–7, 2015.

[3] C. B. Domnick, F.-K. Benra, D. Brillert, and C. Musch, "Modification of a steam valve diffuser for enhanced full load and part load operation using numerical methods," in *Proceedings of CMFF'15*, (Budapest), 2015.

[4] A. Beune, J. Kuerten, and M. van Heumen, "CFD analysis with fluid-structure interaction of opening high-pressure safety valves," *Computers and Fluids*, vol. 64, pp. 108–116, 2012.

[5] O. Frommann and L. Friedel, "Analysis of safety relief valve chatter induced by pressure waves in gas flow," *Journal of Loss Prevention in the Process Industries*, vol. 11, no. 4, pp. 279–290, 1998.

[6] V. Dossena, F. Marinoni, F. Bassi, N. Franchina, and M. Savini, "Numerical and experimental investigation on the performance of safety valves operating with different gases," *International Journal of Pressure Vessels and Piping*, vol. 104, pp. 21–29, 2013.

[7] C. J. Hős, A. R. Champneys, K. Paul, and M. McNeely, "Dynamic behavior of direct spring loaded pressure relief valves in gas service: Model development, measurements and instability mechanisms," *Journal of Loss Prevention in the Process Industries*, vol. 31, no. 1, pp. 70–81, 2014.

[8] P. Moussou, R. J. Gibert, G. Brasseur, C. Teygeman, J. Ferrari, and J. F. Rit, "Instability of pressure relief valves in water pipes," *Journal of Pressure Vessel Technology*, vol. 132, no. 4, p. 041308, 2010.

[9] C. Hős, A. Champneys, K. Paul, and M. McNeely, "Dynamic behaviour of direct spring loaded pressure relief valves in gas service: II reduced order modelling," *Journal of Loss Prevention in the Process Industries*, vol. 36, pp. 1–12, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# CFD SIMULATION OF THERMO- AERODYNAMIC INTERACTION IN A SYSTEM HUMAN- CLOTH- ENVIRONMENT UNDER VERY LOW TEMPERATURE AND WIND CONDITIONS

**Irina Cherunova[1], Sina Samarbakhsh[2] and Nikolai Kornev[2]**

[1]Don State Technical University
Shevchenko str. 147, 346500 Shakhty, Russia
e-mail: i_sch@mail.ru

[2] University of Rostock
A. Einstein Str.2, 18059 Rostock, Germany
e-mail: nikolai.kornev@uni-rostock.de

**Keywords:** Human Thermodynamics, Conjugate Heat Transfer, Computational Fluid Dynamics, Clothes Design.

**Abstract.** *The paper presents study of coupled aerodynamic and thermodynamic interaction between human body and environment under low temperature and wind conditions. The passive heating or the prescribed heat flux model is used to model the metabolic heat release by human body. The heat flux is selected from the condition that the surface temperatures take experimental values for certain environmental conditions. Heat transport inside of the clothes, radiation, natural and forced convection are simulated using OpenFOAM toolkit. Results are obtained for the schematic human body in form of a cylinder and real body. It was shown that only forced convection and heat conduction play an essential role at wind conditions and low temperatures whereas the radiation and natural convection can be neglected. To keep the human body temperature on the acceptable level, the local heating elements can be embedded into the clothes textile. The power of this heating was calculated using the models of aero-thermodynamic coupled interaction including heat radiation. Under strong wind conditions the heat transfer from the human body can sufficiently be increased due to change of the thermal conductivity caused by cloth deformation under wind induced pressures. The results of the work are used for the design of real protection clothes for the work under low temperatures and wind conditions in oil gas industry.*

## 1 INTRODUCTION

Modeling the interaction in the system human- clothes - environment is a very complex interdisciplinary research subject requiring different competencies in the fields of physiology, thermodynamics, fluid mechanics and textile sciences. The simulation in the field of human thermodynamics can be done using either physical human body models (thermal manikins), which are manufactured in full size, or mathematical models. The manikins are equipped with heat wires and temperature sensors to imitate heat transfer over the entire body . Among the limitations of the experimental studies with manikins one can mention high costs in manufacturing. Technically, it's very difficult to divide a manikin in many partitions with different heating power and study the temperature distribution over the entire body surface. Moreover, it is impossible to separate the effects of convection and radiation, since various sensors are required for these phenomena. Even if all these difficulties can be overcome, the manikin remains a model and relies on many assumptions.

With a rapid development of numerical methods the focus of the investigation of the interaction between human and the environment is increasingly being shifted in the direction of the computer simulations. Development of the theoretical basis for modeling interactions in the human-clothing-environment and the design of intelligent smart clothing requires the development of the following three sub-models:

• mathematical model for the human thermodynamics with consideration of thermal conduction, heat transport by the blood, radiation, respiration, heat release by internal organs and active response of the body.

• mathematical model for heat transport in clothes and in the under garment space with taking into account humidity, permeability of clothing, anisotropy of heat propagation within the cloth, non-uniformity of the air layers between the body and clothes as well as between different layers of cloth. For the intelligent or smart clothes the internal heating inside the clothes should be modeled.

• mathematical model for the transport of the heat from the human body to the environment through the radiation, thermal conduction, natural and forced convections.

Short description of the mathematical model used in this paper is presented in the next section.

## 2 MATHEMATICAL MODEL

The first models for internal human thermodynamics were proposed by Fanger [1] and Stolwijk [2]. One of the most successful and widely used models of human thermoregulation, which also represents the state of the art in this field, has been developed by Fiala [3]. The group of models proposed by Stolwijk [2] and Fiala [3] are based on the partial differential equations describing the heat transport in cross sections of body which are coupled through the blood transport along the body. Since such models are not quite reliable we use a simplified model specifying the heat flux on the body surface $\dot{q}$ (prescribed wall heat flux). In this case, the body heat production is taken into account by a heat flux $\dot{q}$ distributed over the body either homogeneously as $\dot{q} = Q_{total}/A$, where $Q_{total}$ is the total heat released by the body and $A$ is the surface of the human body, or non homogeneously with use of empirical data on skin temperature.

The heat transport within the clothes occurs mainly due to the thermal diffusion, although the effects of natural and forced convection caused by the movement of people (bellows effect) are also available. Strictly speaking, the heat propagation in clothes is anisotropic because of the specific textile structure. The application of simple temperature equations with an isotropic

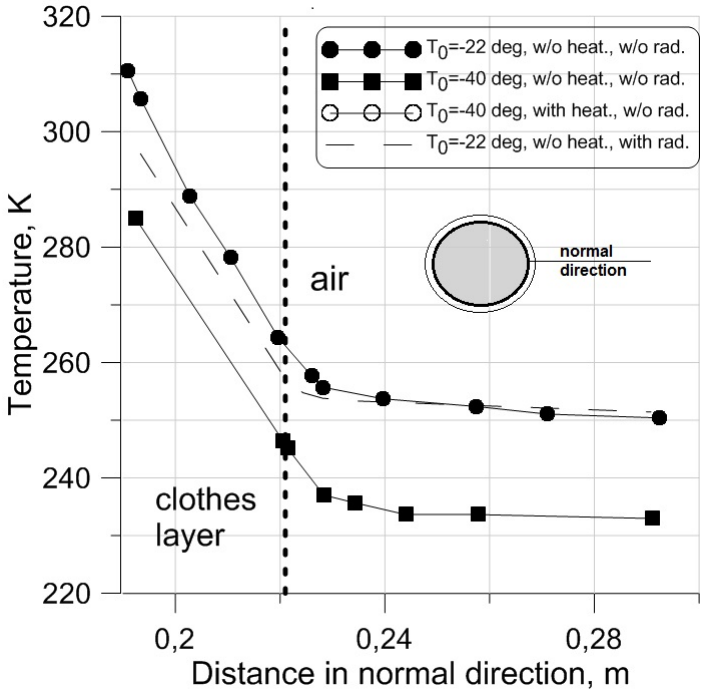



Figure 1: Distribution of the temperature in the normal direction to cylinder with clothes layer. Cases 1-1, 1-2, 2-1, 2-4.

scalar heat conduction coefficient is not correct. However, since the clothes is relatively thin, one can neglect the anisotropy and determine an integral value for the coefficient of the conduction from the measurements. In the present paper, the clothes is considered to be a solid body with the thermodynamic properties determined from experiments. The folding of the cloth, the bellows effect and irregularity of the air insulation layer are not considered.

Forced convection is calculated using the incompressible flow model with the temperature considered as a passive scalar determined from the temperature transport equation. The turbulent flow was calculated using steady RANS (Reynolds Averaged Navier- Stokes Equations) with the $k-\varepsilon$ (sec. 3.1) and $k-\omega$ SST models (sec. 3.2).

The framework OpenFOAM [4] was utilized for the numerical solution of the total problem. Four different models are available in OpenFOAM for calculating the radiation : fvDOM, P1, opaque solid and view factor models.

## 3 RESULTS

Before performing calculations for the real human body we studied a more simple case in which the body is schematically represented by a cylinder. This is done to avoid serious problems arising due to irregularity of the real human body surface and complex grids. Using a simplified body model we determined the heat transport phenomena which can be neglected when considering the real body model.

### 3.1 Preliminary study using cylinder model with cloth

The study was performed for a cylinder with the height of $1500$ mm and a diameter of $382$ mm, which mimics the human body approximately. The surface of the cylinder is $1.9m^2$ which corresponds to the human body surface. The cylinder case seems to be relevant since the heat conduction takes place mainly in the normal direction similar to the body case. The study was

performed at low and very low temperatures with consideration of the radiation, natural and forced convections. To model the protecting influence of cloth, an outer layer with the thickness of 30 mm was inserted as a cover of the cylinder. The coefficient of thermal conductivity of the layer is equal to this of the real clothes (s. subsec. 3.2.2).

Table 1: **Summary of cases**

| Case | Air temperature $^0C$ | Wind speed m/s | Radiation | Metabolic heat release W | Internal heating W | Body surface temperature $^0C$ | clothes surface temperature $^0C$ |
|---|---|---|---|---|---|---|---|
| 1-1 | -22 | 0 | No | 85 | 0 | 30.5 | -8.3 |
| 1-2 | -22 | 0 | Yes | 85 | 0 | 23.9 | -15.3 |
| 2-1 | -40 | 0 | No | 85 | 0 | 12.0 | -27.0 |
| 2-2 | -40 | 0 | No | 85 | 55 | 29.5 | -21.0 |
| 3-1 | -40 | 5 | No | 85 | 0 | 3.2 | NA |
| 3-2 | -40 | 5 | No | 85 | 105 | 29.5 | -31.0 |
| 3-3 | -40 | 5 | Yes | 85 | 105 | 28.2 | -31.8 |
| 3-4 | -40 | 10 | No | 85 | 118.4 | 29.0 | -33.2 |

The calculation cases are summarized in table 1. Two different ambient air temperatures were examined: $-22^0C$ and $-40^0C$. The boundary layer created by the natural convection can be either laminar or turbulent depending on the ratio between the buoyancy and the viscous forces, which is characterized by the Rayleigh number $Ra = GrPr$, where $Gr$ and $Pr$ are the Grashof and Prandtl numbers respectively. Since for the cases under consideration the Rayleigh number is $2.31 \times 10^9$ the flow caused by natural convection is turbulent. The standard $k - \varepsilon$ is applied. The degree of turbulence at the inlet amounts to $10\%$ what at the speed of $u = 0.5m/s$ corresponds to the r.m.s of $u^{/} = 0.05m/s$, while the turbulent kinetic energy is $k = 3/2u^{/2} = 0.00375m^2/s^2$. The dissipation rate $\varepsilon$ calculated using the integral length of $l = 5cm$ is $\varepsilon = C_\mu k^{3/2}/l = 4.13 \cdot 10^{-4}m^2/s^3$. The prescribed heat flux of $85$ Watts distributed uniformly over the cylinder surface simulates the metabolism heat release. This value allows one to obtain the temperature of $30^0C$ on the body surface. Calculations were performed with the code OpenFoam [4]. To couple the air environment with the solid bodies including the cloth the turbulent TemperatureCoupledBaffleMixed routine is used. A special interface has been introduced in order to thermally couple the air environment and solid body. The available in OpenFoam routine turbulentTemperatureRadCoupledMixed condition was applied to model the radiation.

### 3.1.1 Effect of natural convection

The natural convection is caused by the density differences that occur in proximity to the body due to the heating of the air. Heat transfer from the body is caused in this case by the two following mechanisms: molecular thermal conduction (thermal diffusivity) and the heat transport by the air movement. Four calculation cases 1-1, 1-2, 2-1 an 2-2 were investigated with account for the natural convection. The temperature distribution along the line, which is perpendicular to the cylinder surface, is shown in Fig. 1. The abscissa $0.191m$ corresponds to the body surface, while $0.221m$ represents the outer boundary of the clothes. Through the

metabolic heat release of $85W$ the body temperature (Case 1-1) reaches the value of $30.5^0C$ when the air temperature is $-22^0C$. This result is consistent with the experimental observations performed by the first author (IC) in the climate chamber with real test person. The insulating protective clothes allows to hold the body temperature at a comfortable level of $30^0C$ without additional heating. The temperature of the upper boundary of the clothes is negative and amounts to $-8.31^0C$ (see case 1-1 in the table 1). If the ambient air temperature decreases to $-40^0C$, the body temperature reduces to $12^0C$, while the temperature of the outer clothes boundary sinks to $-27^0C$ (s. case 2-1). Although a simplified geometry is considered, the results are more plausible and physically than our previous results for the realistic human body [5]. The temperature at the upper limit of the clothes is negative. This was achieved by reducing the numerical error caused by irregular body surface and grid problems. To attain a comfortable temperature at $T_0 = -40^0C$ we use the smart clothes with the internal heating placed at the interface surface between the body and the cloth. From calculations it was found that the heating power of $55W$ is required to bring the body temperature at $29.50^0C$ ( s . Fig. 1 and case 2-2 in Table 1). The temperature of the clothes boundary is $-21^0C$ for this case.

### 3.1.2 Effect of radiation

In this paper the opaque solid model was used which takes the radiation within the clothes into account. The emissivity was set equal to the value of $0.9$. The results are presented in Fig. 1 and case 1-2 in table 1. As can be seen , the body temperature drops by radiation from $30.5^0C$ to $23.9^0C$. Therefore the comfort conditions documented in the case 1-1 are disturbed by the radiation .

### 3.1.3 Effect of forced convection

Forced convection has a crucial effect on the heat interaction between human body and environment especially at low temperatures. In this work, four cases listed in the table 1 as cases 3 are studied. The ambient temperature was in all cases $-40^0C$. The results show that the forced convection caused by the wind of $5m/s$ results in the drop of the body temperature to $3^0C$ without heating (s. case 3-1). In order to achieve the comfortable body temperature of around $30^0C$, the internal heating of $105W$ is required (see case 3-2 in table 1). It is noteworthy that the radiation at the wind conditions is completely meaningless (s. Fig. 2 and compare cases 3-2 and 3-3). When the wind has the speed of $10m/s$ the required internal heating should have the power of $118.4W$ (see case 3-4). Concluding, when the heat exchange between the human body and environment is calculated at low temperatures only forced convection and thermal diffusion can be taken into account. Other effects like the radiation and natural convection are negligible.

## 3.2 Results for the human body model with cloth.

### 3.2.1 Human body model, computational domain and grid.

The finite volume method is implemented on an unstructured grid with approximately $4.3M$ of cells in the computational domain $15m \times 5.5m \times 5.5m$. Elongation of the computational domain in the longitudinal direction is necessary to calculate the wake effect. The grid has different resolutions in different areas. The finest unstructured grid is located in the block of $3m \times 1.8m \times 1.8m$ closest to the body. Away from the body the grid resolution is gradually reduced having relatively high resolution in the block $7.5m \times 3.6m \times 3.6m$. The real human body geometry of height of $1.8m$ designed at the Hohenstein Institute on the basis of the detailed

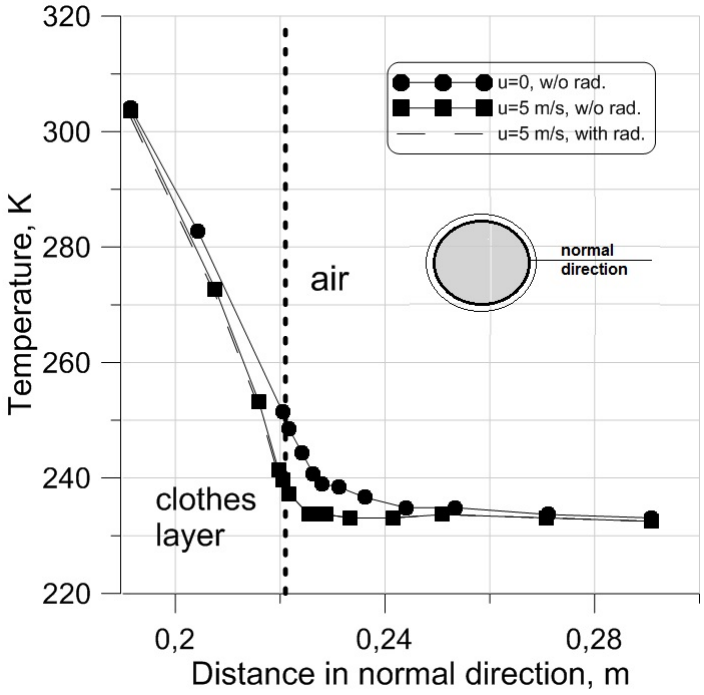



Figure 2: Distribution of the temperature in the normal direction to cylinder with clothes layer. Cases 2-2, 3-2, 3-3.

anthropological study of different categories of peoples is located in the symmetry plane at 5m from the inlet of the computational domain. For reliable resolution of the temperature boundary layers six thin prism layers are added on the inner and the outer surfaces of the cloth. A thin prism layer of the thickness of $1mm$ adjacent to the inner air layer between the clothes and the body is served as the heating layer with heat release source $f$.

### 3.2.2 Clothes construction

A clothes of thickness 30 mm comprises a savior, a flamestat cotton (upper sheet), an insulation thinsulate, and a taffeta as a lining. This clothes is used in oil industry for work at very low temperatures under oil contamination conditions. The air layer of the thickness of $10mm$ is located between the clothes and the body. The clothes textile has the following properties: the heat capacity is $C_p = 1800J/kgK$, the density is $8.57kg/m^3$, the thermal conductivity is $\lambda = 0.04W/mK$.

### 3.2.3 Prescribed heat flux model of the human body

The heat flux $\dot{q}$ is calculated from the condition that the resulting temperatures from simulations are equal to these measured on the real human body clothed in the protection coverall. Temperature measurements were performed at $-22$ deg of the ambient air at $18$ points directly on the skin of the human body. The body surface is subdivided into $34$ elements. The heat flux on each element was varied as long as the temperature on the skin from simulations is approximately equal to the measured ones. Details of this procedure are presented in [5].

### 3.2.4 Results for internal heating

The aim of the calculation is the determination of the heat release $f$ which allows one to maintain the temperature on the body skin in the comfortable range at the temperature of the ambient air of $-40^{\circ}$ without wind. The comfortable body skin temperatures are taken from the measurements performed at $-22^{\circ}$ with the real human clothed in the protection coverall. In numerical simulations the area and the power were varied as long as the simulated temperature at $18$ points is close to the comfortable one. The heating elements were selected under the following conditions:

• Heating elements can not be located close to the sensible inner organs because it can have negative impact on health,
• Heating elements can not be located in zones of high sweating,
• The necessary area of the heating elements and heating power should be minimal,
• Both overheating and under heating of the body should be avoided.

Distribution of heating elements with necessary area and heat flux as well as the temperature distributions on the body skin are given in Fig. 3. The total heat flux necessary to maintain the comfortable temperatures is estimated as $60$ Watt. This result obtained for the real human body is very close to this for the cylinder. Therefore, we can expect the same result for the case with wind speed. i.e. the power necessary to attain the comfortable temperature on the real body for the wind speed of $5m/s$ is around $105W$.

### 3.2.5 Clothes deformation due to wind induced pressure.

Wind causes big areas of the overpressure on the clothes surface which results in deformations and local change of the clothes thickness. In its turn, change of the local thickness leads to an alteration of heat conduction properties of the cloth. This means that the heat exchange between body and air is changed not only by intensification of convective heat transfer but also due to change of thermodynamic properties of clothes caused by wind induced deformations. The second effect has still not been discussed thoroughly in the literature. Under strong wind conditions the heat transfer from the human body can sufficiently be increased due to change of the thermal conductivity caused by cloth deformation under wind induced pressures. For instance, at 10 m/s the heat increase could be of ten percent. For the details of this study the reader is referred to the work [6].

## 4 CONCLUSIONS

The paper presents an application of numerical simulations for design of a special clothes for work under extremely low temperatures. The heating elements inside of the clothes are applied to maintain the comfortable temperatures on the body skin. The distribution of the temperature on the body, in the cloth, in the air layer between the body and the clothes and in the ambient air is calculated from the transport equations using the OpenFoam toolkit. Results are obtained for the schematic human body in form of a cylinder and real body. It was shown that only forced convection and heat conduction play an essential role at wind conditions and low temperatures whereas the radiation and natural convection can be neglected. Sizes and the power of heating elements are selected from numerical simulations. At present present numerical results are used to construct the protecting coverall of the new generation for employees of the gas and oil industry.

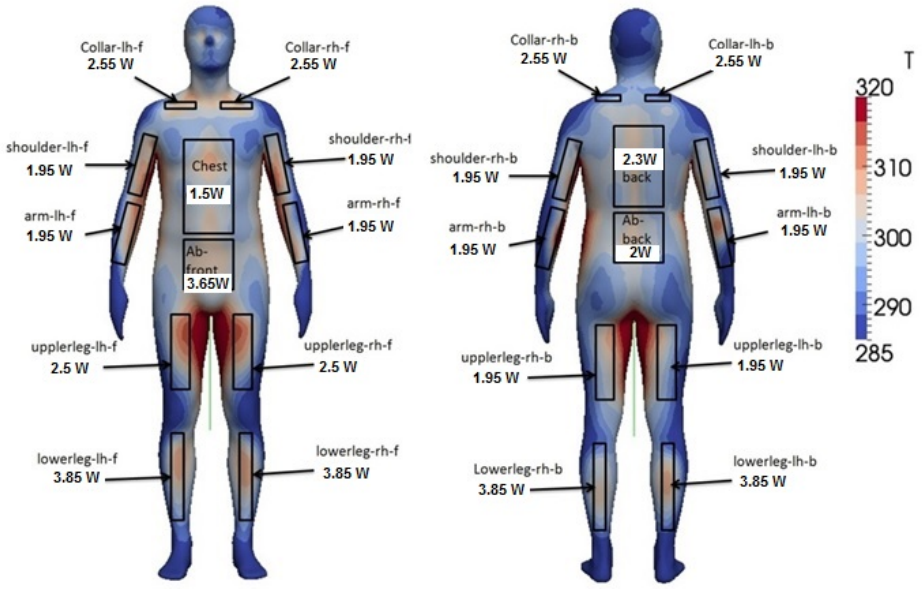



Figure 3: Power of the heating elements and the temperature distribution at the human body at $-40^0C$ of the surrounding air. CFD calculations with OpenFoam program.

**Acknowledgments.** This work was carried out with the financial support within the framework of the joint project 11032p/20253 of the German Federal Ministry of Education and Research (BMBF) and the Russian Foundation for Assistance to Small Innovative Enterprises (FASIE).

## REFERENCES

[1] Fanger, P.O. (1970) Thermal Comfort. *Danish Technical Press*, Copenhagen.

[2] Stolwijk, J.A.J. (1971) A mathematical model of physiological temperature regulation in man. *NASA Contractor Report 1855*.

[3] Fiala, D., Lomas K.J., Stohrer M.(2001) Computer prediction of human thermoregulatory and temperature responses to a wide range of environmental conditions. *Int. J. Biometeorol.*, 45, 143–159.

[4] Weller, H. G., Tabor, G., Jasak, H. and Fureby, C. (1998) A tensorial approach to computational continuum mechanics using object-oriented techniques. *Computers in Physics*, 12(6), 620–631.

[5] Cherunova I., Dhone M. and Kornev N. (2015). Coupled thermo-aerodynamical problems in design of protection cloth. *Proc. VI International Conference on Computational Methods for Coupled Problems in Science and Engineering COUPLED PROBLEMS 2015*, Venice, May 2015, 1303–1311.

[6] Cherunova I., Kornev N., Jacobi G., Treshchun I., Gro A., Turnow J., Schreier S. and Paschen M. (2014). Application of computational fluid mechanics for protection clothes design. *Proceedings of the 11th World Congress on Computational Mechanics (WCCM XI)*, 22.-25.07.2014, Barcelona, Spain, 7262–7273.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE INFLUENCE OF THE INCLUSIONS' MORPHOLOGY ON THE ACCURACY OF THE PREDICTION OF REINFORCED COMPOSITES MECHANICAL BEHAVIOUR

**Viwanou Hounkpati**[1,2,3,4,5], **Vladimir Salnikov**[6], **Philippe Karamian-Surville**[7,8], **Alexandre Vivet**[1,2,3,4,5]

[1] Normandie Univ., France; [2] UNICAEN, CIMAP, F-14032 Caen, France; [3] ENSICAEN, CIMAP, F-14050 Caen, France; [4] CNRS, UMR 6252 CIMAP, F-14032 Caen, France; [5] CEA, UMR 6252 CIMAP, F-14032 Caen, France
{viwanou.hounkpati,alexandre.vivet}@unicaen.fr

[6] University of Luxembourg, Mathematics Research Unit
6, rue Richard Coudenhove-Kalergi, L-1359 Luxembourg
vladimir.salnikov,@uni.lu

[7] UNICAEN, LMNO, F-14032 Caen, France; [8] CNRS, UMR 6139 CIMAP, F-14032 Caen, France
philippe.karamian@unicaen.fr

**Keywords:** Mean-field approaches, Fast Fourier Transform, Mechanical behaviour, Composites Materials, ellipsoidal and spherical reinforcements.

**Abstract.** *Nowadays, the prediction of mechanical behaviour of reinforced composites is a paramount problem in computational materials science. The macroscale effective coefficients are obtained from the microscale information also known as the so-called corrector problem. A numerical strategy based on integral equation known as the periodic Lippmann-Schwinger equation is used in the present study. For that purpose, several Representative Volume Elements (RVE) are generated with a mixture of spherical and ellipsoidal inclusions, by the Random Sequential Adsorption scheme or the time driven Molecular Dynamics. In this work, the influence of the inclusions morphology on the accuracy of some classical mean-field approximation methods and a full-field computational method based on numerical homogenization techniques to predict the mechanical properties of these materials is studied. For low volume fractions of inclusions, the results of the mean-field approximations and those of the Fast Fourier Transform-based (FFTb) full-field computation are very close, whatever the inclusions' morphology. For RVEs consisting of ellipsoidal or a mixture of ellipsoidal and spherical inclusions, when the inclusions volume fraction becomes higher, one observes that Lielens' model and the FFTb full-field computation give almost similar estimates. The accuracy of the computational methods depends on the shape of the inclusions and their volume fraction.*

## 1 INTRODUCTION

During the last decades, the micromechanics of heterogeneous materials, which is an area of very fertile researches at the boundary between physics and mechanics of materials, evolved from the so-called mean-field approaches to full-field schemes such as Finite Element Method (FEM) or more recently Fast Fourier Transforms (FFT) methods initiated by Moulinec and Suquet [1]. The aim of these approaches is to predict the effective behaviour of reinforced composite materials which is strongly influenced by the microstructure (e.g., constituents' properties, volume fractions, shapes, orientations, etc.). In such materials, inclusions can have several morphologies with different geometrical orientations. To exactly describe the effective behaviour of reinforced composite materials, a highly accurate description of the microstructures is required.

The full-field approach used in this paper is the Fast Fourier Transform-based (FFTb) homogenization technique [2] where an iterative scheme is used for computing effective properties of each given RVE (Representative Volume Element). One of the main advantages of this approach is its low time and memory consumption in comparison for example with finite elements methods. The application of the FFTb methods to the study of the overall behaviour of materials involves a preliminary step of 3D RVE generation. Another important feature of this method is that the generated RVE well approximates, at least in certain aspects, the real microstructure of the material. Mean-field homogenization methods rely on a simple statistical analysis of the microstructure.

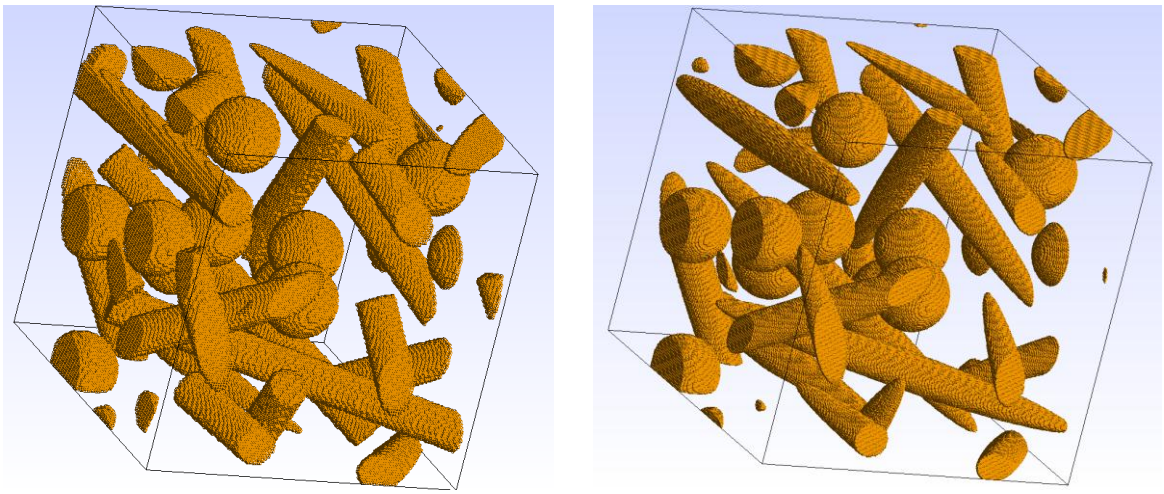
Several works were conducted to highlight the validity domains of these mean-field micromechanical models (by comparing them to full-field approaches) when predicting the elastic properties of various RVEs with different microstructures. Most of these studies have been focused on spherical particles reinforced composites, aligned fibers reinforced composites, randomly oriented fiber reinforced composites or microstructures with aligned or randomly oriented clay platelets.

In this work, the effective material properties obtained using the numerical FFTb homogenization techniques were compared to three different analytical methods: the normalized self-consistent scheme (NSC) [3,4], the Mori–Tanaka model (MT) [5] and Lielens' model [6]. The aim of this paper is to exhibit the influence of the inclusions morphology on the accuracy of these homogenization techniques to predict the mechanical properties of reinforced composites consisting not only for spherical or ellipsoidal inclusions but also for a mixture of spherical and ellipsoidal inclusions.

## 2 RVE GENERATION AND FFT-BASED HOMOGENISATION

In this paper, we focus on the influence of the inclusions morphology on reinforced composite materials. In order to study the heterogeneous medium, one needs to have access to RVEs. Basically there are two main families of algorithm to design an RVE with complex microstructures. The First technique is based on the RSA method (Random Sequential Adsorption) whereas the second technique is based on MD (Molecular Dynamics). In this paper our choice goes to the latter one. Actually this approach is richer, more fruitful to design RVEs. Moreover this approach is reliable and fast enough to build RVEs. Indeed, in this technique, one can play with different kind of parameters which allow one to design RVEs with various inclusions such as spheres, cylinders, ellipsoids or a mixture of different kind of these simple geometries. The MD-based method is also powerful for higher volume fractions (of order $50-60\%$): MD generates a configuration in about a second whereas the RSA can get stuck. The classical RSA and the time-driven version of MD applied to generate the mixture of inclusions of spherical and cylindrical shapes will not be presented is this paper. A detailed

description can be found in [7]. An example of a sample with a mixture of nonintersecting spherical and cylindrical inclusions is presented on the figure 1a. Because mean-field approaches usually describe inclusions shape by the means of ellipsoids, we introduced imperfections (see [2] for more details) on the cylinder inclusions without spoiling the efficiency of the generation algorithm, in order to obtain ellipsoidal inclusions instead of cylinder. The figure 1b shows the same RVE as in figure 1a where the extremities of cylinder inclusions are rounded in order to obtain ellipsoidal inclusions.



(a) (b)

Figure 1: Examples of 3D view of generated RVEs: (a) spherical and cylindrical inclusions, (b) spherical and ellipsoidal inclusions.

Once the RVE are generated, one seeks to find out the effective mechanical properties. For that purpose we have developed a homogenization technique based on the Lippmann-Schwinger equation, since the RVE are periodic. Actually, a mechanical homogenization problem for periodic heterogeneous media, leads to solve a microlocal problem, also known as the corrector problem, in which the corrector term is the periodic unknown. The corrector term is then a solution of an elliptic partial differential equations (PDE). From a mathematical point of view, solving a PDE returns to determine the Green function associated with the PDE. Hence, the above problem is a Fredholm integral equation leading to the Lippmann-Schwinger equation (see Moulinec and Suquet [1]) which can be solved using the FFT techniques. In the sequel, this approach has been privileged and is called the FFT-based homogenization (FFTb).

## 3 MEAN-FIELD HOMOGENIZATION MODELS

Because the inclusions frame of reference *Ri* attached to its main axes are not always identical to the RVE coordinate system *RI*, subscripts "*Ri*" and "*RI*" are added, for the sake of clarity, to each tensor to indicate the coordinate system in which it is expressed. One can note that the subscripts "*Ri*" and "*RI*" are not necessary when the inclusions are spherical or all aligned, because in this case *Ri* and *RI* are the same. For randomly oriented inclusions reinforced composites, the geometrical orientation of each inclusion is described by three Euler angles $(\varphi_1,\phi,\varphi_2)$. The transition between the local coordinate system *Ri* of the inclusion and the RVE system *RI* is made by these three Euler angles in the Bunge convention [8] where the transformation matrix is given by the following equation (1). This matrix is applied, for a given rank four tensor K, as showed in equation (2).

$$m(\varphi_1,\phi,\varphi_2)=\begin{bmatrix} \cos\varphi_1\cos\varphi_2-\sin\varphi_1\sin\varphi_2\cos\phi & -\cos\varphi_1\sin\varphi_2-\sin\varphi_1\cos\varphi_2\cos\phi & \sin\varphi_1\sin\phi \\ \sin\varphi_1\cos\varphi_2+\cos\varphi_1\sin\varphi_2\cos\phi & -\sin\varphi_1\sin\varphi_2+\cos\varphi_1\cos\varphi_2\cos\phi & -\cos\varphi_1\sin\phi \\ \sin\varphi_2\sin\phi & \cos\varphi_2\sin\phi & \cos\phi \end{bmatrix} \quad (1)$$

$$_{RI}K_{ijkl} = m_{im}m_{jn}m_{ko}m_{lp}\ {}_{Ri}K_{mnop} \quad (2)$$

The homogenized stiffness tensor $C$ of a composite consisting of $n$ different types of inclusions (in terms of shape and geometrical orientation) in the RVE coordinate system $RI$ is:

$$C = \left\langle {}_{RI}C : {}_{RI}A \right\rangle_{RVE} = f_m\ {}_{RI}C_m : {}_{RI}A_m + \sum_{i=1}^{n} f_i\ {}_{RI}C_i : {}_{RI}A_i \quad (3)$$

where $f_i$, $C_i$ and $A_i$ denote respectively the volume fraction, the stiffness tensor and the strain-localization tensor of inclusions exhibiting the same shape and the same geometrical orientation. $f_m$, $C_m$ and $A_m$ denote respectively the volume fraction, the stiffness tensor and the strain-localization tensor of the matrix. $A:B$ denotes the double scalar product using the Einstein summation convention. $\langle\ \rangle_{RVE}$ stands for the volume average over the whole RVE (matrix + inclusions). From the average strain theorem, one can obtain:

$$C = {}_{RI}C_m + \sum_{i=1}^{n} f_i\ \left({}_{RI}C_i - {}_{RI}C_m\right) : {}_{RI}A_i \quad (4)$$

This is the main relation used to calculate the homogenized stiffness tensor. The strain-localization tensor $A_i$ in this relation differs from one model to another. In the normalized self-consistent [3,4], Mori-Tanaka [5] and Lielens' [6] models, the strain-localization tensors are expressed respectively as follows:

$${}_{Ri}A_i^{NSC} = \left[I + {}_{Ri}E : \left({}_{Ri}C_i - {}_{Ri}C\right)\right]^{-1} : \left\langle \left[I + {}_{Ri}E : \left({}_{Ri}C_i - {}_{Ri}C\right)\right]^{-1} \right\rangle_{RVE}^{-1} \quad (5)$$

$${}_{Ri}A_i^{MT} = \left[I + {}_{Ri}E_m : \left({}_{Ri}C_i - {}_{Ri}C_m\right)\right]^{-1} : \left\langle \left[I + {}_{Ri}E_m : \left({}_{Ri}C_i - {}_{Ri}C_m\right)\right]^{-1} \right\rangle_{RVE}^{-1} \quad (6)$$

$${}_{Ri}A_i^{Lielens} = \left\{ (1-f)\left[{}_{Ri}A_i^{lower}\right]^{-1} + f\left[{}_{Ri}A_i^{upper}\right]^{-1} \right\}^{-1} : \left\langle \left\{ (1-f)\left[{}_{Ri}A_i^{lower}\right]^{-1} + f\left[{}_{Ri}A_i^{upper}\right]^{-1} \right\}^{-1} \right\rangle_{RVE}^{-1} \quad (7)$$

$E$ is the Morris' tensor [9], which represents the interaction between an inclusion with a given morphology and the homogeneous equivalent medium. In the case of an ellipsoidal inclusion whose principal axes lengths are $\{2a_1, 2a_2, 2a_3\}$, it is written in the coordinate system of the inclusion $Ri$ as follows:

$${}_{Ri}E_{ijkl} = \frac{1}{4\pi}\int_0^{\pi} \sin\theta d\theta \int_0^{2\pi} \gamma_{ikjl}\, d\phi = {}_{Ri}S_{ijkl}^{Esh}\ {}_{Ri}C_{ijkl}^{-1} \quad (8)$$

where $\gamma_{ikjl} = K_{ik}^{-1}(\xi)\xi_j\xi_l$, $K_{jp}(\xi) = {}_{Ri}C_{ijpl}\xi_i\xi_l$, $\xi_1 = \dfrac{\sin\theta\cos\phi}{a_1}$, $\xi_2 = \dfrac{\sin\theta\sin\phi}{a_2}$, $\xi_3 = \dfrac{\cos\theta}{a_3}$.

$S^{Esh}$ is the Eshelby tensor; $a_1$, $a_2$, $a_3$ are used to describe the inclusion shape. In the Mori-

Tanaka model (equation (6)), the Morris' tensor $E_m$ is computed by using the matrix properties $C_m$ as the infinite media, instead of $C$. In equation (7) related to Lielens' model, $f$ is an interpolating factor which depends on the inclusions volume fraction [6,10]. For a set of inclusions with the same shape (ellipsoid, sphere,…) having a volume fraction $f_{shape}$, the interpolating factor is given by:

$$f = \frac{1}{2} f_{shape}\left(1 + f_{shape}\right) \tag{9}$$

In our study, for the RVEs exhibiting inclusions with two different shapes (ellipsoid and sphere), an interpolating factor is considered for each shape according to their volume fraction. $A_i^{lower}$ and $A_i^{upper}$ in equation (7) are given by:

$$_{Ri}A_i^{lower} = \left[I + {}_{Ri}E_m : \left({}_{Ri}C_i - {}_{Ri}C_m\right)\right]^{-1} \tag{10}$$

$$_{Ri}A_i^{upper} = \left[I + {}_{Ri}E_i : \left({}_{Ri}C_m - {}_{Ri}C_i\right)\right]^{-1} \tag{11}$$

Note that in these relations, the Morris' tensors $E_m$ and $E_i$ are computed by using the matrix and the inclusion properties respectively as the infinite media.

## 4 RESULTS AND DISCUSSION

In this part, several RVEs exhibiting different morphologies have been studied. The accuracy of homogenization models was evaluated for composites made of an isotropic matrix reinforced with isotropic spherical and/or ellipsoidal particles. Effective properties predicted by mean-field homogenization analytical models have then been compared to those obtained numerically (by FFT-based homogenization technique) in order to rigorously validate the investigated models and to highlight the influence of the inclusion volume fraction on their accuracy and their sensitivity to the inclusions morphology. The effective bulk, shear and Young's moduli were computed for each RVE using the corresponding homogenized stiffness tensor $C$. For the sake of simplicity, only the effective shear $\mu_{eff}$ and bulk $\kappa_{eff}$ moduli were presented in this paper. In the following, the Poisson ratio is assumed to be $\nu$ = 1/3 for all phases (inclusions and matrix).

### 4.1 Influence of spherical inclusions on the accuracy of the models

Many works were conducted to highlight the validity domains of these mean-field micromechanical models (by comparing them to full-field approaches) when predicting the elastic properties of composites reinforced with spherical particles [11,12]. Consequently, our study is not concerned with this kind of RVEs. However, some simulations were made in order to confirm if our results are in accordance with theirs. For inclusion volume fraction of about 10%, predictions of $\mu_{eff}$ and $\kappa_{eff}$ by the mean-field approaches are almost satisfactory. Around 30% of inclusions, mean-field models are accurate for low contrast and deviate from the FFTb solution for high contrasts. When the inclusions volume fraction reaches 50%, the same behaviour was observed. Figure 2 shows the normalized effective shear $\mu_{eff}$ and bulk $\kappa_{eff}$ moduli evolutions as function of the Young's modulus contrast $E_i/E_m$ (ratio between the Young's modulus $E_i$ of the inclusion and $E_m$ of the matrix) in this case. The predictions of Mori-Tanaka and normalized self-consistent models diverge very rapidly when the contrasts

increase, as observed in [11]. The Mori-Tanaka model underestimates the accurate solution (supposed to be FFTb results) while the normalized self-consistent model overestimates it. For low contrasts, (up to 50 for the bulk modulus and up to 20 for the shear modulus), Lielens' model is close to the FFTb, with a relative deviation from the FFTb less than 10%.

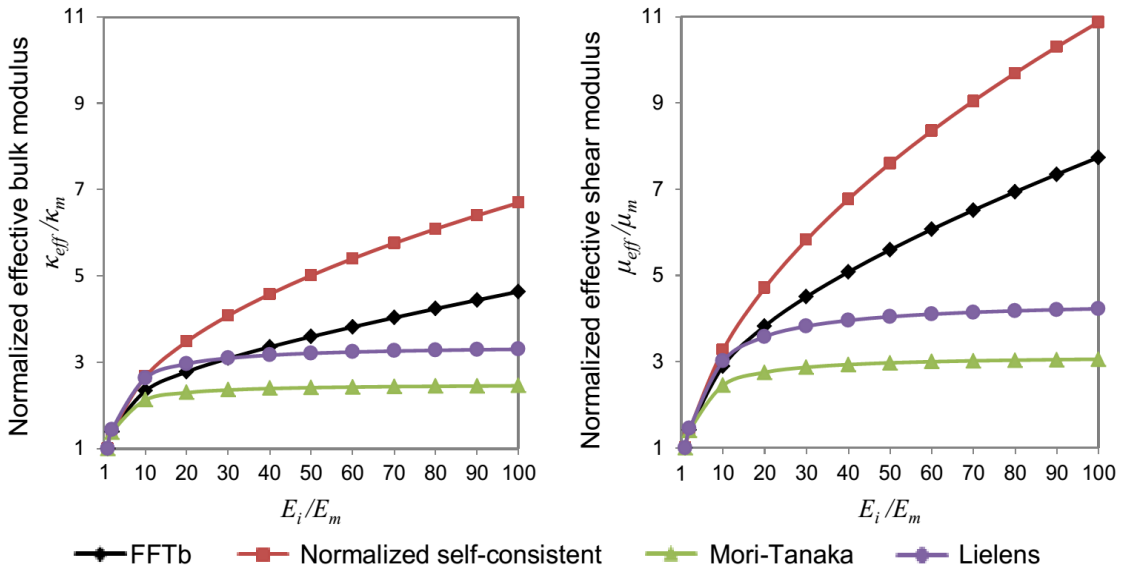



Figure 2: Normalized effective shear and bulk moduli as function of the Young's modulus contrast for a RVE consisting of spherical inclusions with a volume fraction of 50%.

### 4.2 Influence of ellipsoidal inclusions on the accuracy of the models

The aim of this part is to highlight the accuracy of the mean-field approaches while predicting the mechanical behaviour of these kinds of composites made of randomly oriented isotropic ellipsoidal inclusions distributed into an isotropic matrix. To do this, six RVEs containing respectively 4, 8, 12, 16, 20 and 30% (in volume) of ellipsoidal inclusions have been investigated. Each inclusion contains 10 ellipsoidal inclusions randomly distributed. This distribution differs from one RVE to another. For this study, the aspect ratio of the ellipsoids is taken equal to 5 ($a_1/a_2 = 5$, $a_2 = a_3$) for all inclusions.

Using the RVEs' parameters given by the RSA and MD-based generation algorithms, we computed the homogenized properties of each RVE. The figure 3 shows the evolution of the normalized effective shear modulus as function of the Young's modulus contrast for some of the RVEs investigated. The effective bulk modulus results are not showed for the sake of clarity and especially because the relative deviations from the FFTb observed in this case are less than those obtained for the effective shear modulus. In the case of high volume fraction of ellipsoidal inclusions, the normalized self-consistent model cannot be reasonably used. Actually, in that case, the computation time increases drastically and the algorithm involves a high memory consumption due to the iterative resolution. For all these reasons, this model was not presented in this part. Figure 3 shows that Lielens' model is close to FFTb. The mean-field analytical models are more sensitive for ellipsoidal inclusions' volume fraction. Up to about 20% of inclusions, all the models deliver accurate estimates of the homogenized properties regardless of contrasts. The relative deviations from the FFTb model observed in this case do not exceed 4% for Lielens' model and 6% for Mori-Tanaka model. For a volume fraction of 30%, the models deliver accurate estimates only for low contrasts ($E_i / E_m \leq 20$).

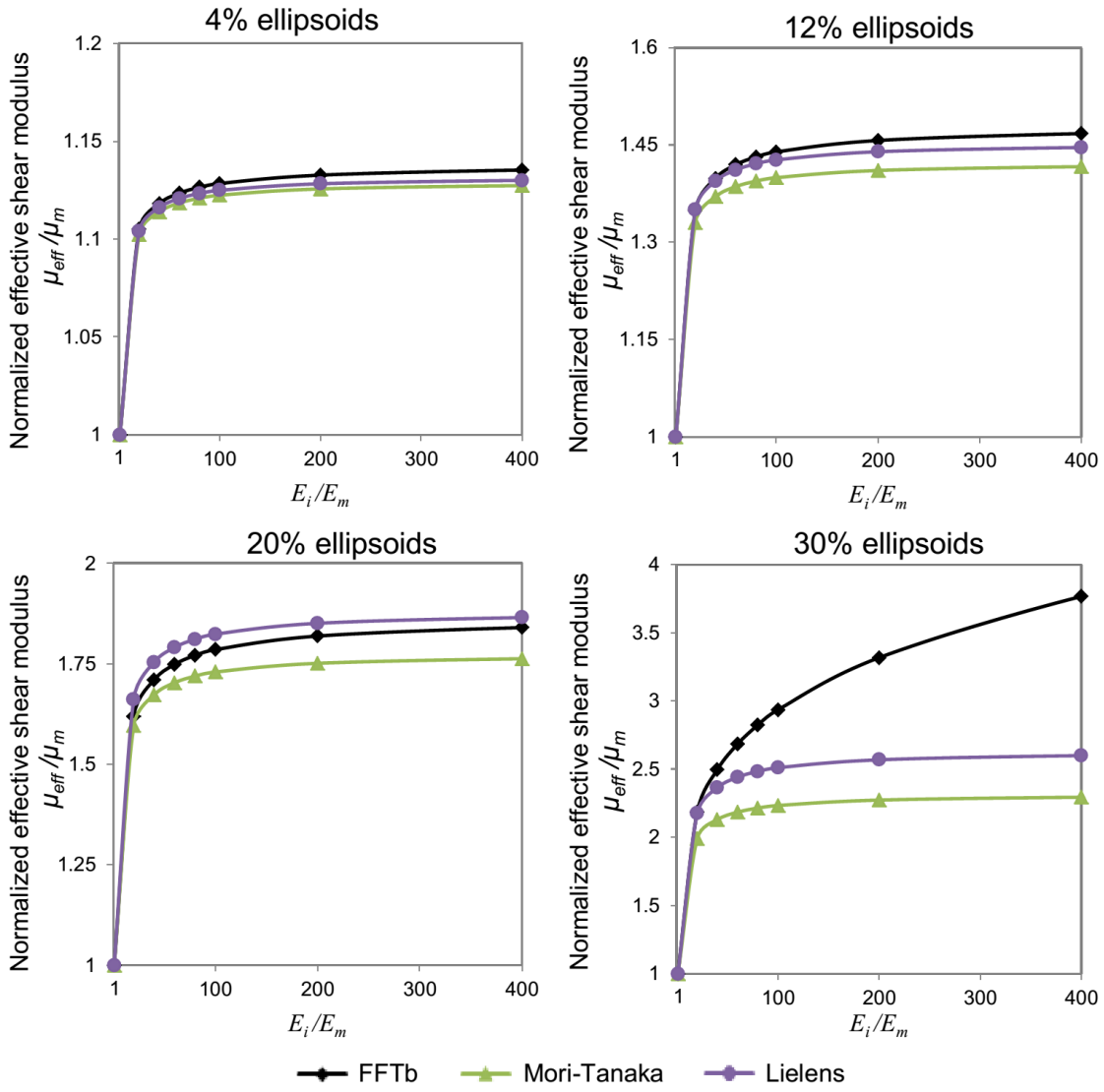



Figure 3: Normalized effective shear modulus as function of the Young's modulus contrast for RVEs consisting of ellipsoidal inclusions with a volume fraction of 4, 12, 20 and 30%, respectively. In each RVE, 10 ellipsoidal inclusions exhibiting different orientations were considered. The aspect ratio of the ellipsoids is $a_1/a_2$=5.

### 4.3 Influence of a mixture of spherical an ellipsoidal inclusions on the accuracy of the models

Now, we consider two RVEs containing a mixture of spherical and ellipsoidal inclusions. For the two RVEs, the volume fraction of the spherical and ellipsoidal inclusions is set at 5% and 6.7% respectively. For one RVE, the ellipsoids were oriented in only 2 directions different from the main directions of the RVE. It is clear that the macroscopic behaviour will be strictly anisotropic in this case. The aspect ratio of the ellipsoids is fixed to 5. For the other RVE, 10 different geometrical orientations were considered for the ellipsoidal inclusions. The aspect ratio of the ellipsoids for this RVE is set at 10. Figure 4 shows the normalized effective shear and bulk moduli evolutions related to the contrast of these RVEs. The models are in good agreement for the shear and bulk moduli in each case. Figure 4 shows that Lielens and Mori-Tanaka models deliver, for the volume fractions investigated, the most accurate predictions when the RVE contains a mixture of spherical and ellipsoidal inclusions, regardless of the macroscopic behaviour being isotropic or not. Although the normalized self-consistent

model deviates from the other models, the maximal relative deviation from the FFTb results observed for this later do not exceed 5%. One can also notice that the mean-field models are sensitive to the ellipsoids aspect ratio because the discrepancies observed are more important in the RVE where the ellipsoids exhibits the great aspect ratio.

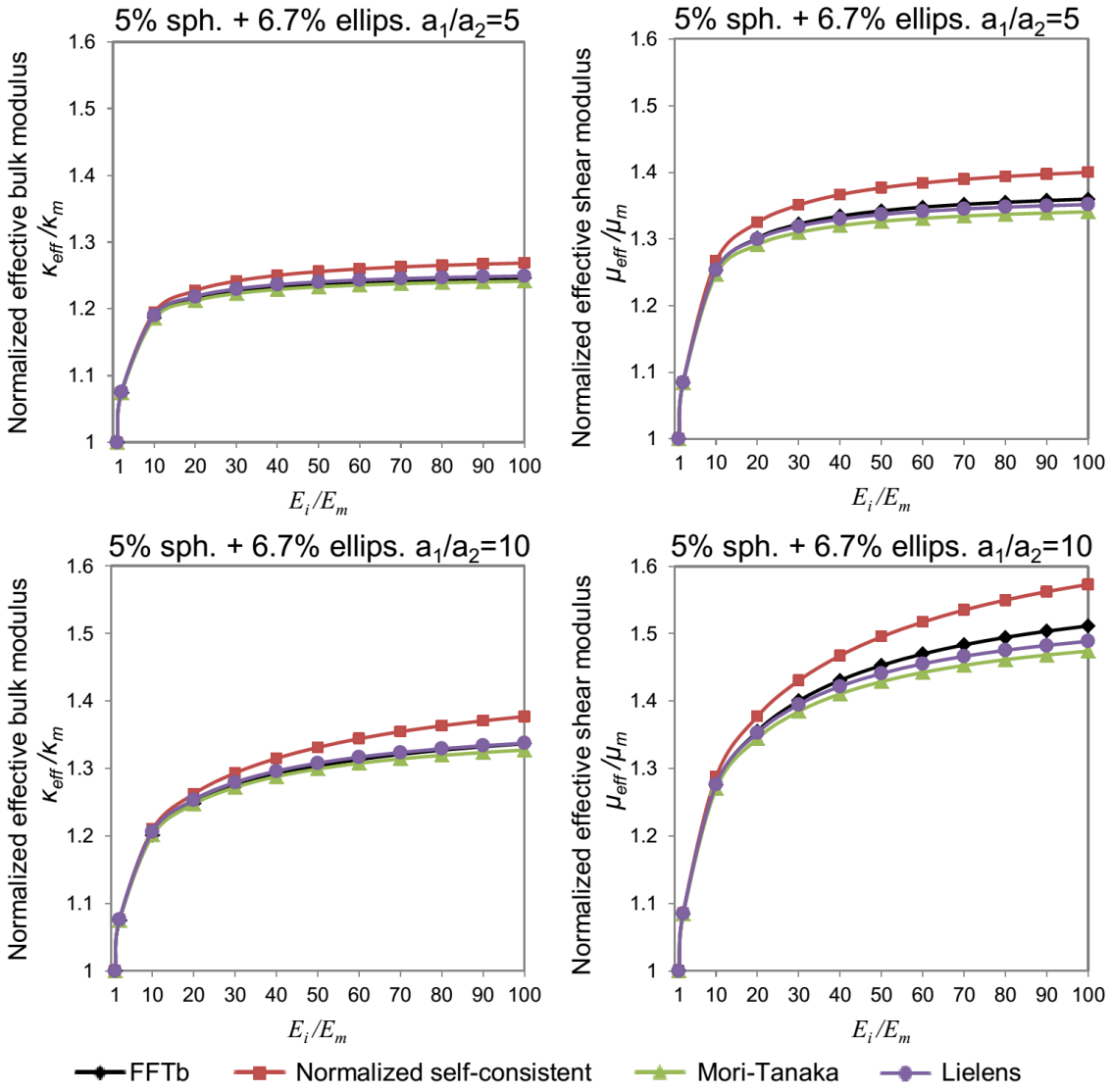



Figure 4: Normalized effective shear and bulk moduli as function of the Young's modulus contrast for 2 RVEs consisting of a mixture of spherical (5%) and ellipsoidal (6.7%) inclusions. The ellipsoidal inclusions ($a_1/a_2$=5) of the first RVE exhibit 2 different geometrical orientations. In the second RVE, the ellipsoidal inclusions ($a_1/a_2$=10) exhibit 10 different geometrical orientations.

## 5 CONCLUSIONS

The study highlights the influence of the inclusions morphology on the accuracy and discrepancy of some classical mean-field approximation methods and a full-field computational method based on numerical homogenization techniques to predict the mechanical properties of these materials. Several Representative Volume Elements containing spherical, ellipsoidal inclusions and a mixture of both were studied. For low volume fractions of inclusions, the results of the mean-field approximations and those obtained with the Fast Fourier Transform-based (FFTb) full-field computation are very close regardless to the inclusions morphology.

RVEs consisting of ellipsoidal or a mixture of ellipsoidal and spherical inclusions, when the inclusions volume fraction becomes higher, one observes that Lielens' model and the FFTb full-field computation give almost similar estimates. The sensitivity of the computational methods, consequently their accuracy depends on the shape of the inclusions and their volume fraction. The ellipsoids aspect ratio has also some influence on the accuracy of the models but this one is less than the influence of the volume fraction. For microstructures with a mixture of ellipsoidal and spherical inclusions, Lielens' and Mori-Tanaka models could be a good alternative to the FFTb model when the total inclusions vol-ume fraction is about 12%. In this case, the normalized self-consistent model could also be an alternative to the FFTb model with an error less than 5%.

## REFERENCES

[1] H. Moulinec, P. Suquet, A numerical method for computing the overall response of nonlinear composites with complex microstructure, *Computer Methods in Applied Mechanics and Engineering*. **157** (1998) 69–94.

[2] V. Salnikov, S. Lemaitre, D. Choï, P. Karamian-Surville, Measure of combined effects of morphological parameters of inclusions within composite materials via stochastic homogenization to determine effective mechanical properties, *Composite Structures*. **129** (2015) 122–131.

[3] J.Y. Li, On micromechanics approximation for the effective thermoelastic moduli of multi-phase composite materials, *Mechanics of Materials*. **31** (1999) 149–159.

[4] V. Hounkpati, S. Fréour, D. Gloaguen, V. Legrand, Accounting for a distribution of morphologies and orientations on stresses analysis by X-ray and neutron diffraction: Normalized Self-Consistent modeling, *Advanced Materials Research*. **996** (2014) 82–87.

[5] T. Mori, K. Tanaka, Average stress in matrix and average elastic energy of materials with misfitting inclusions, *Acta Metallurgica*. **21** (1973) 571–574.

[6] G. Lielens, P. Pirotte, A. Couniot, F. Dupret, R. Keunings, Prediction of thermo-mechanical properties for compression moulded composites, *Composites Part A: Applied Science and Manufacturing*. **29** (1998) 63–70.

[7] V. Salnikov, D. Choï, P. Karamian-Surville, On efficient and reliable stochastic generation of RVEs for analysis of composites within the framework of homogenization, *Computational Mechanics*. **55** (2015) 127–144.

[8] H.-J. Bunge, *Texture Analysis in Materials Science*, Butterwort, London, 1982.

[9] P.R. Morris, Elastic Constants of Polycrystals, *International Journal of Engineering Science*. **8** (1970) 49–61.

[10] C.L. Tucker, E. Liang, Stiffness predictions for unidirectional short- fiber composites: Review and evaluation, *Composites Science and Technology*. **59** (1999) 655–671.

[11] E. Ghossein, M. Lévesque, A fully automated numerical tool for a comprehensive validation of homogenization models and its application to spherical particles reinforced composites, *International Journal of Solids and Structures*. **49** (2012) 1387–1398.

[12] S. Kari, H. Berger, R. Rodriguez-Ramos, U. Gabbert, Computational evaluation of effective material properties of composites reinforced by randomly distributed spherical particles, *Composite Structures*. **77** (2007) 223–231.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF ADVANCED PULTRUSION PROCESSES WITH MICROWAVE HEATING

**Evgeny Barkanov[1], Pavel Akishin[1], Rudolf Emmerich[2], and Matthias Graf[2]**

[1] Institute of Materials and Structures, Riga Technical University
Kalku St. 1, LV-1658, Riga, Latvia
barkanov@latnet.lv

[2] Fraunhofer Institute for Chemical Technology ICT
Joseph-von-Fraunhofer St. 7, 76327 Pfinztal, Germany
rudolf.emmerich@ict.fraunhofer.de

**Keywords:** Pultrusion, Microwave Heating, Finite Element Method, Electromagnetic Model, Thermo-Chemical Model, Cylindrical Rod.

**Abstract.** *Pultrusion is a continuous and cost-effective process for a production of composite structural components with a constant cross-sectional area. This technological process could be made more effective applying instead of conventional heaters a high frequency electromagnetic energy source characterized by the fast, instantaneous, non-contact and volumetric heating. To provide better understanding of the microwave assisted pultrusion process, to support the pultrusion tooling design and process control, new simulation methodology consisting of two sub-models has been developed. Each of them is constructed by using the general-purpose FE software that results in considerable savings in development time and costs, and also makes available various modeling features of the FE package. In the first step the electromagnetic sub-model is used to evaluate the electric field distribution by solving the Maxwell's equations with the COMSOL Multiphysics. It is necessary to note that in this sub-model the transport phenomena is neglected since an influence of the pull speed of the processing material on the electric field distribution is negligible. The objective of simulations is to find the microwave field as homogeneous as possible inside the cured composite profile in the ceramic inlet. In the second step an absorption energy field in the composite material determined with the electromagnetic sub-model is used as a heating source in the pultrusion process modeled with the thermo-chemical sub-model. This simulation procedure is developed in ANSYS Mechanical environment and based on the mixed time integration scheme and nodal control volumes method to decouple the coupled energy and species equations. To demonstrate an application of the developed methodology for the design of technological process, the microwave assisted pultrusion of the cylindrical rod made of glass fibers Unifilo 4800 tex and polyester resin POLRES 305BV has been investigated.*

## 1 INTRODUCTION

Pultrusion is a continuous and cost-effective process for a production of composite structural components with a constant cross-sectional profile (Fig. 1). During pultrusion the fiber reinforcements are saturated with the resin in a resin tank and then continuously pulled through a heated die by a puller. Inside the die the resin gradually cures and solidifies to form a composite part with the same cross-section profile as in the die. At the final stage a traveling cut-off saw cuts the composite profile into desired lengths.

This technological process could be made more effective applying instead of conventional heaters a high frequency electromagnetic energy source characterized by the fast, instantaneous, non-contact and volumetric heating. It is necessary to note that at present time microwave heating [1] is successfully and widely used in different industrial curing processes [2, 3]. Due to the complexity of the multiphysical problem to be solved for a microwave assisted pultrusion process, initially these processes have been analyzed experimentally [4, 5]. A brief review of the potential for microwave heating in the pultrusion manufacturing processes of fiber reinforced composites is presented in paper [6]. Recently, due to the rapid development of the finite element software for a solution of multiphysical problems, a computational modeling of the microwave assisted pultrusion processes has started. In paper [7] the finite element model and corresponding analysis have been developed for the microwave assisted pultrusion process of a fiberglass-epoxy cylindrical profile with the diameter of 5 mm. Several die external radius and two different sets of material die properties have been investigated to obtain an effective manufacturing process.

To provide better understanding of the microwave assisted pultrusion process, to support the pultrusion tooling design and process control, new simulation methodology consisting of two sub-models are developed. Each of them is constructed by using the general-purpose FE software that results in considerable savings in development time and costs, and also makes available various modeling features of the FE package. In the first step the electromagnetic sub-model is used to evaluate the electric field distribution by using the COMSOL Multiphysics software. In the second step an absorption energy field obtained in the composite material is applied as a heating source in the thermo-chemical sub-model developed in ANSYS Mechanical environment to determine temperature and degree of cure fields in the pultruded composite. To demonstrate an application of the developed methodology for the design of technological process, the microwave assisted pultrusion of the cylindrical rod is investigated. Conditions of the advanced pultrusion process, and the temperature and degree of cure fields in the composite material, are determined for the design of modular pultrusion die consisting of microwave block with the ceramic inlet located at the entrance of the steel mould. To obtain desired characteristics of the investigated pultrusion process, an influence of the capacity of the absorption energy field, thermal insulation, preheating of steel mould, heating of resin bath and reduction of the steel mold length on the parameters of pultrusion process are studied.

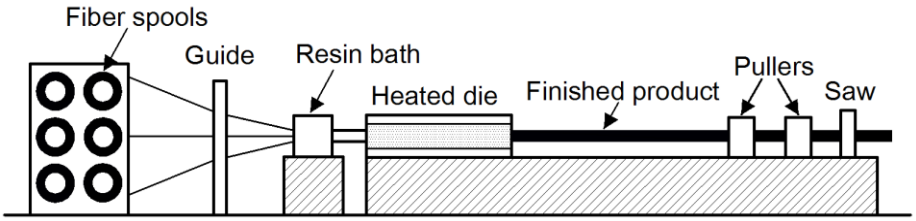



Figure 1: Pultrusion process.

## 2 STATEMENT OF THE PROBLEM

For a numerical simulation of the microwave assisted pultrusion process it is necessary to solve two coupled multiphysical problems - electromagnetic and thermo-chemical. The electromagnetic problem is solved with the purpose to determine the electric field distribution and as the result to obtain the absorbed energy field in the composite material which will be used later as a heating source in the subsequent thermo-chemical problem. Solving the last problem, the temperature and degree of cure fields in the pultruded composite could be estimated.

Using the common approach of a harmonic oscillating electric field $\vec{E}$

$$\vec{E}(\vec{r},t) = \vec{E}(\vec{r}) \cdot e^{2\pi i f} \tag{1}$$

where $\vec{r}$ is the location vector, $t$ is a time and $f$ is the microwave frequency, Maxwell's equations could be written as following

$$\nabla \times \nabla \times \vec{E}(\vec{r}) - \varepsilon_0 \mu_0 (2\pi f)^2 \varepsilon_r \vec{E}(\vec{r}) = \vec{0} \tag{2}$$

where $\varepsilon_0$ is the vacuum permittivity and $\mu_0$ is the magnetic constant. This complex valued equation is solved numerically for the amplitudes of the electric field $\vec{E}(\vec{r})$ with respect to the relative permittivity $\varepsilon_r$ which for loss dielectric materials like mixtures of glass fibers and polyester resin is a complex function of frequency $f$, temperature $T$ and degree of cure $\alpha$

$$\varepsilon_r(f,T,\alpha) = \varepsilon'(f,T,\alpha) - i \cdot \varepsilon''(f,T,\alpha) \tag{3}$$

It can be obtained by the cavity perturbation method [8] at fixed frequencies or Corbino probe measurements [9, 10] with variable frequency. After numerical solution of Equation (2), the absorbed microwave energy generated by dielectric losses could be obtained like in paper [11]:

$$Q(\vec{r}) = 2\pi f \varepsilon_0 \varepsilon'' \left| \vec{E}(\vec{r}) \right|^2 \tag{4}$$

Knowing the absorbed energy field $Q$ in the composite material, the following thermo-chemical problem consisting of three governing equations should be solved in the second step:

$$\begin{cases} \rho c_p \dfrac{\partial T}{\partial t} - \dfrac{\partial}{\partial x}\left( k_x \dfrac{\partial T}{\partial x} \right) - \dfrac{\partial}{\partial y}\left( k_y \dfrac{\partial T}{\partial y} \right) - \dfrac{\partial}{\partial z}\left( k_z \dfrac{\partial T}{\partial z} \right) - q_b = 0 \\ \bar{\rho}\, \bar{c}_p \left( \dfrac{\partial T}{\partial t} + u \dfrac{\partial T}{\partial x} \right) - \dfrac{\partial}{\partial x}\left( \bar{k}_x \dfrac{\partial T}{\partial x} \right) - \dfrac{\partial}{\partial y}\left( \bar{k}_y \dfrac{\partial T}{\partial y} \right) - \dfrac{\partial}{\partial z}\left( \bar{k}_z \dfrac{\partial T}{\partial z} \right) - q - Q = 0 \\ \left( \dfrac{\partial \alpha}{\partial t} + u \dfrac{\partial \alpha}{\partial x} \right) = R_r \end{cases} \tag{5}$$

where $T$ is the temperature, $\rho$ and $c_p$ are the density and specific heat of the tooling material, $k_x$, $k_y$, $k_z$ are the thermal conductivities in $x$, $y$, $z$ directions, $q_b$ is the rate of energy exchange at the boundary, $u$ is the pull speed, $\bar{\rho}$ and $\bar{c}_p$ are the lumped density and specific heat for the composite material, $\bar{k}_x$, $\bar{k}_y$, $\bar{k}_z$ are the lumped thermal conductivities in $x$, $y$, $z$ directions, $q$ is the generative term related to the internal heat generation due to the exo-

thermic resin reaction, $\alpha = H(t)/H_{tr}$ is the degree of cure and $H(t)$ is the amount of heat evolved during the curing up to time $t$. It is necessary to note that the first equation in the system (5) presents the energy equation for the tool, second – the energy equation for the composite moving in the pull direction and third – the species equation (transport equation) for the resin.

Heat transfer in the composite occurs as a result of the conversion of electromagnetic energy into heat, according to specific loss mechanisms, conduction and the generation of heat resulting from the exothermic chemical reaction initiated by the microwave heating. The generative term related to the internal heat generation due to the exothermic resin reaction could be written as:

$$q = V_r \rho_r H_{tr} R_r \tag{6}$$

where $V_r$ is the resin volume fraction, $\rho_r$ is the resin density, $H_{tr}$ is the total heat of reaction and $R_r$ is the rate of resin reaction determined as:

$$R_r(\alpha, T) = \frac{\partial \alpha}{\partial t} = \frac{1}{H_{tr}} \frac{dH(t)}{dt} = K(T) \cdot f(\alpha) \tag{7}$$

where $f(\alpha)$ depends on the resin properties and varies with the applied resin reaction model, and $K(T)$ is defined by the Arrhenius relationship:

$$K(T) = K_0 \exp\left(-\frac{E}{RT}\right) \tag{8}$$

where $R = 8.314$ J/mol·K is the universal gas constant. It is necessary to note that coefficients of the Arrhenius relationship: activation energy $E$ and frequency factor $K_0$ are the physical values and are determined by the Kissinger method or ASTM E 698 standard methodology from the DSC tests. Coefficients of the selected function $f(\alpha)$ could be obtained in a simple way by a fitting of the experimental heat flow curves applying the least squares method.

The reinforcement is saturated with the resin before entering the heated die in pultrusion process. Therefore, it is reasonable to assume that the resin does not flow. In most cases the continuous model with lumped material properties are used for a simulation of the pultrusion processes. These properties are evaluated by the rule of mixture:

$$\bar{\rho} = (1 - V_r)\rho_f + V_r \rho_r$$

$$\bar{c}_p = \frac{(1 - V_r)\rho_f c_{pf} + V_r \rho_r c_{pr}}{\bar{\rho}} \tag{9}$$

$$\bar{k} = \frac{k_f k_r \bar{\rho}}{(1 - V_r)\rho_f k_r + V_r \rho_r k_f}$$

where indexes $f$ and $r$ relates to the fibers and resin, respectively.

The above system (5) of three governing equations could be solved by using the prescribed initial and boundary conditions. It is assumed that at time $t = 0$, for a curing composite the temperature is $T = T^0$ and the degree of cure is $\alpha = \alpha^0$, where index "0" denotes the initial values. The temperature of the composite at the die entrance and everywhere in the composite

at the first step of numerical algorithm is prescribed as the pre-heat resin temperature. The value of the degree of cure at the die entrance and everywhere in the composite at the first step of numerical algorithm is taken as zero or very small value depending on the resin curing kinetic model. In some pultrusion processes this value could be much higher in relation to the ambient room temperature which is used also as the boundary conditions for a die to describe the convection effect. It is important to note that the power flux at the entrance and exit of the composite in the die is specified as zero. The insulated boundary conditions could be applied also to reduce unnecessary heat losses and conserve energy in pultrusion processes.

## 3 NUMERICAL SIMULATION PROCEDURES

The developed algorithms simulate the phenomenon of microwave assisted pultrusion by solving the set of coupled electromagnetic, energy and species equations.

### 3.1 Electromagnetic sub-model

The targeted application of an industrial microwave pultrusion process comes along with two major requirements: the usage of a cost-effective microwave source with a power output in the range of some kW and the capability to generate a homogeneous microwave field distribution for the complete products cross section. As magnetrons provide sufficient power output at fixed frequencies very economically it is obvious to use a magnetron as microwave source working at one of the ISM-frequency-bands (Industrial Scientific and Medical use).

The cross-section of the products is predetermined and the first step of the problem is to design the base model of a microwave set-up with an analytically known field distribution most suitable for the requested cross-section. In a second step, the model is improved numerically towards a homogeneous field distribution and high efficiency of the provided energy. This is performed adjusting the models dimensions and variation of some material properties until the target specifications are reached (Fig. 2). For the electromagnetic sub-model the FEM software package COMSOL Multiphysics is used with the optional radiofrequency (RF) module.

Against standard design drawings, for the simulation of microwave phenomena it is necessary to generate a design defining a closed (e.g. metallic) structure from electric point of view and allocate material properties in every region inside – especially the “air” typically left out as logical region in design drawings – is almost important for wave propagation. On the other hand, the thickness of solid metallic machinery parts is rarely important, because their surfaces usually are treated as perfect electric conductor. The microwave field is reflected at the surface and does not penetrate into the metallic regions – so they can be left out in the drawing.

According to the Nyquist-Shannon sampling theorem [12] for the discrete sampling of a wave function, the maximum element size $x_{\max}$ of the mesh applied to the geometry has to fulfill $x_{\max} < \lambda/2$ with $\lambda$ - local wavelength. Local wavelength means that $\lambda$ might be shorten by dielectric materials or influenced by conductive parts so that a good value for the maximum mesh element size is given by $x_{\max} < \lambda_0/20\sqrt{\varepsilon_r}$ with $\lambda_0$ - the wavelength in free space and $\varepsilon_r$ - the relative permittivity.

For this mesh Equation (2) can be solved for the amplitude of the electric field $\vec{E}(\vec{r})$ with respect to the local material properties from Equation (3) by iterative or direct solvers in the frequency domain with the fixed frequency $f$ of the magnetron. The wave excitation and thereby the power input is applied by a so called “port” boundary condition provided by the

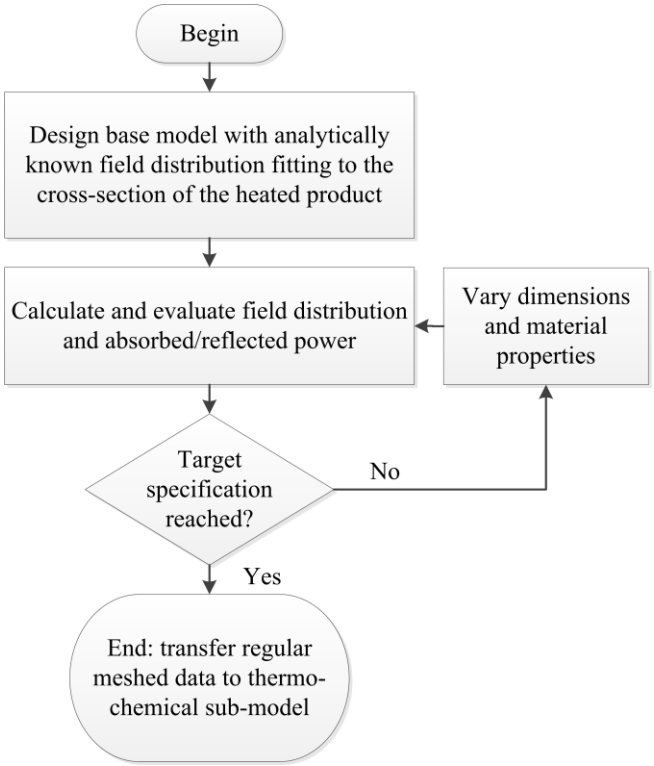



Figure 2: Flowchart of the numerical procedure based on COMSOL Multiphysics.

RF – module. Then the heating rate $Q$ can be calculated using Equation (4) and evaluated regarding homogeneity and efficiency.

### 3.2 Thermo-chemical sub-model

This simulation procedure is developed in ANSYS Mechanical environment and is based on the mixed time integration scheme and nodal control volumes method to solve simultaneously the coupled temperature state and degree of cure by using an iteration technique. A uniform finite element discretization is applied in the pull direction. The nodal control volumes are constructed based on the finite element mesh as presented in Figure 3. Centers of the control volumes coincide with the nodal points of the finite element. In the control volume the distribution of a field variable is assumed constant and its value is defined by the field variable calculated at the representative finite element node.

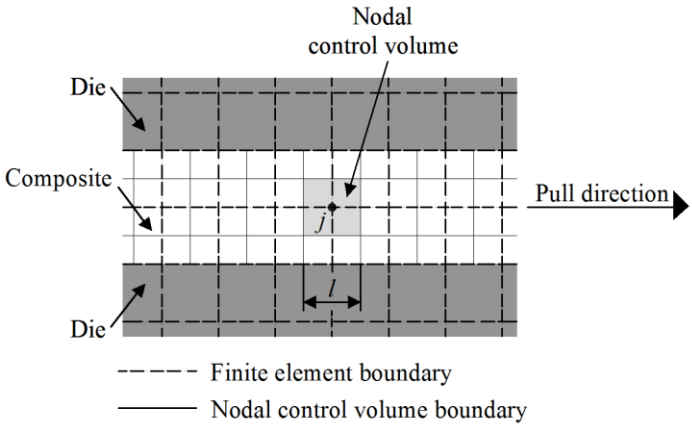



Figure 3: Finite element and nodal control volume meshes.

At the beginning it is assumed that the degree of cure has the same value $\alpha^0$ in each nodal control volume of the composite. In most cases it equals zero. Then the transient thermal finite element analysis is performed to obtain an initial state of temperatures for each element. From the temperature field, the rate of cure for the nodal control volume $j$ at any time step $i$ is calculated outside the finite element software:

$$\frac{\partial \alpha_j^i}{\partial t} = \left[\frac{\Delta \alpha_j^i}{\Delta t}\right] = K_0 \exp\left(-\frac{E}{RT_j^i}\right) f(\alpha_j^{i-1}) \tag{10}$$

For $t > 0$, the degree of cure can be obtained continuously by using the following relation:

$$\alpha_j^i = \alpha_j^{i-1} + \left[\frac{\Delta \alpha_j^i}{\Delta t}\right] \Delta t \tag{11}$$

where $\Delta t$ is the time step determined as:

$$\Delta t = \frac{1}{p} \cdot \frac{l}{u} \tag{12}$$

where $l$ is the length of nodal control volume in the pulling direction, $u$ is the pulling speed and $p$ is the number of sub-steps. If the procedure of sub-stepping is not applied, $p = 1$. The exothermic effects of cure reaction are evaluated as the equivalent nodal heat power for a nodal control volume or node $j$ by the following relation:

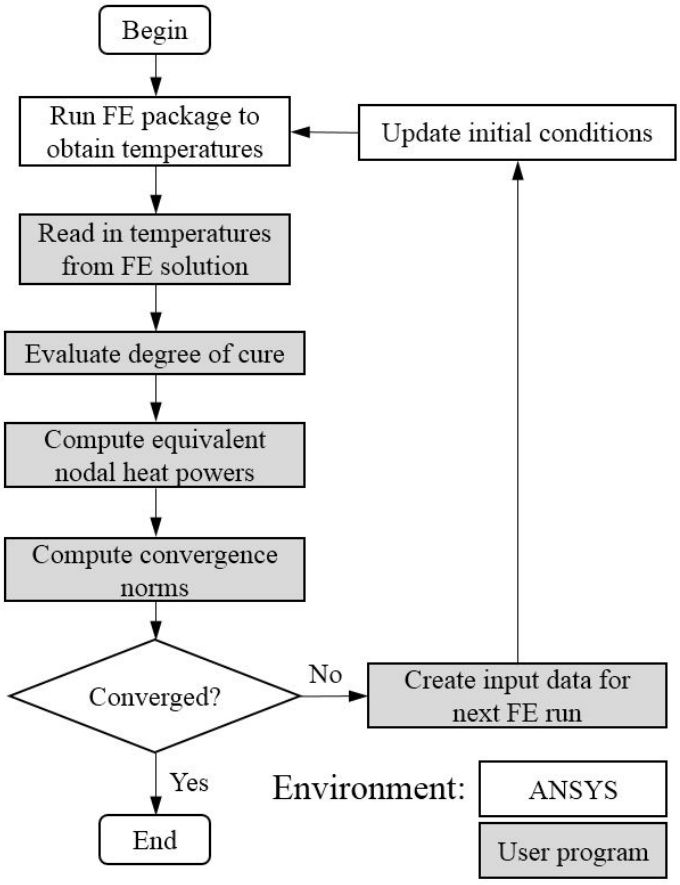



Figure 4: Flowchart of the numerical procedure based on ANSYS Mechanical.

$$q = V_r \rho_r H_{tr} R_r = V_r \rho_r H_{tr} \left[ \frac{\Delta \alpha_j^i}{\Delta t} \right] \tag{13}$$

These values will be applied to calculate the temperature field for a new step of iteration. Thereby, a movement of the resin-saturated composite is simulated by shifting the temperature and degree of cure fields after each calculation step. It is necessary to note that at the entrance of the die, the degree of cure remains unchanged and equals to $\alpha^0$ at any step of iterations. In general, the algorithm can be summarized as a flowchart presented in Figure 4.

## 4 MICROWAVE ASSISTED PULTRUSION OF CYLINDRICAL ROD

To demonstrate an application of the developed methodology for the design of technological process, the microwave assisted pultrusion of the cylindrical rod with the diameter of 16 mm and made of glass fibers Unifilo 4800 tex and polyester resin POLRES 305BV has been investigated.

### 4.1 Analysis with electromagnetic sub-model

In the present case, the base model has to provide a symmetric cylindrical field distribution with an extended maximum of electrical field covering the products diameter. As starting point, an electric maximum of a standing wave pattern in a shorted rectangular waveguide section is used. To minimize reflection losses the waveguide has been tapered to the dimensions of the pultrusion die. The centerpiece of the microwave system consists out of a ceramic part ($ZrO_2$) with a 16 mm hole, which has been inserted as a section into the metal pultrusion die. The objective was that the microwave field inside the hole and therefore inside the rod running through that hole is as homogenous as possible. The Zirconium oxide has a very high real part of the permittivity $\varepsilon'$, which shortens the wave length of the microwaves. Therefore a magnetron of frequency 915 MHz and a waveguide WR 975 have been chosen.

Figure 5 shows the scheme of the electromagnetic sub-model for microwave pultrusion with the ceramic insert and coupling of the microwave to the ceramic part with a waveguide. The different color has indicated the different parts and the materials. Dark blue indicates steel (pultrusion die), light blue - air (waveguide), red - ceramic (insert) and green - product (rod). The dielectric properties $\varepsilon_r$ of the glass fiber/polyester resin composition have been measured before with a cavity perturbation method versus temperature and mean values taken at a temperature of $T$=60 °C have been used for the calculation of the electric field distribution.

Running through the optimization loop is shown in Figure 2 several times, the shape of the ceramic inlay has been adjusted to achieve a homogeneous field distribution inside the rod and a minimized power reflection. The materials and properties applied to the domains are specified in Table 1 and the boundaries - in Table 2.

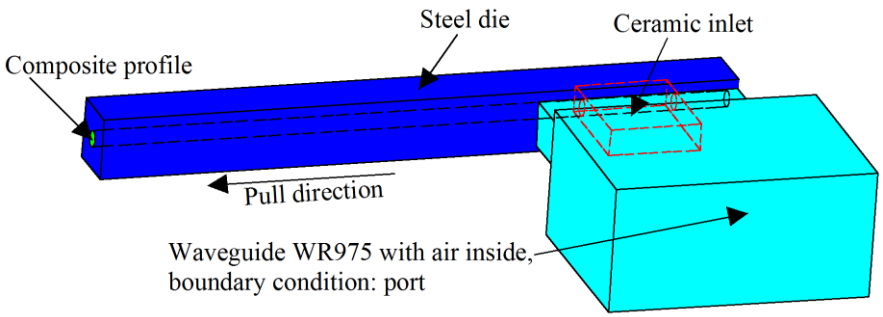



Figure 5: Scheme of the microwave model.

| Property | Symbol | POLRES 305BV + Unifilo 4800 tex | Zirconia ceramic | Steel 40Cr | Air (volume in waveguide) |
|---|---|---|---|---|---|
| Dielectric constant | $\varepsilon'$ | 5.7 | 29 | Perfect electric conductor | 1 |
| Dielectric loss | $\varepsilon''$ | 0.32 | 0.2 | | 0 |

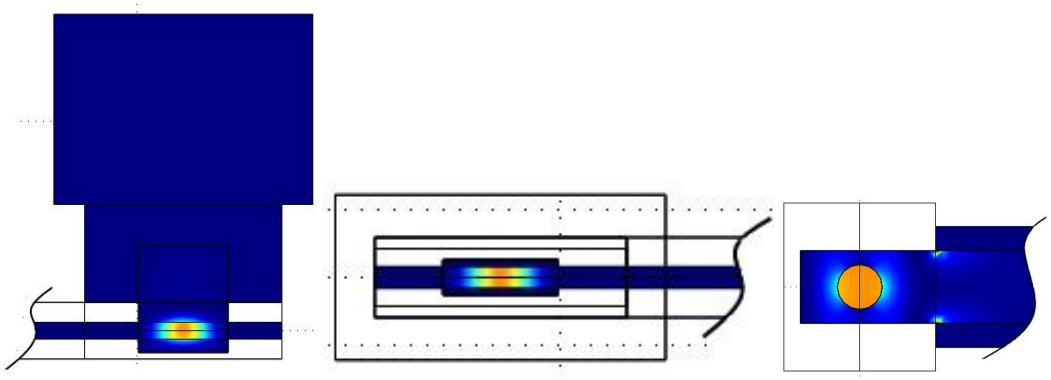
Table 1: Parameters of the electromagnetic sub-model.

| Boundary | Material | Property |
|---|---|---|
| Port | Air | $TE_{10}$ mode for rectangular waveguide power $P$ = 1 W (scalable) |
| Surrounding areas (die, waveguide) | Steel 40Cr | perfect electric conductor |
| Rod | POLRES 305BV + Unifilo 4800 tex | open circular ports |

Table 2: Boundaries for the electromagnetic sub-model.

Figure 6 shows the distribution of the resulting heating rate $Q$ in different 3D- and cross-sectional views for the optimized geometry of the ceramic inlay. In Figures 3a and 3b it is obviously, that only the rod inside - the ceramic insert is heated. The microwave coupling and the pultrusion die are not heated by the microwave. Outside of the ceramic insert the microwave field is close to zero. Figure 3c shows the cross-section of the heating rate in the die and the waveguide. The heating rate is homogeneously concentrated in the rod. Additionally a small heating effect inside the ceramic material itself gets visible.

In detail the power density in the cross-section of the rod is about 7000 W/m$^3$ inside the ceramic insert but outside the rod it is between 200 and 800 W/m$^3$ (Fig. 7). The deviation of the power density is low, therefore the heating is homogenous. The power density is concentrated to the cross-section of the rod. Figure 7 shows the power density in the cross-section of the ceramic pultrusion die. The power is concentrated to about 1.6 cm along the rod with a variation of about 7%.



(a) (b) (c)

Figure 6: Distribution of the heating rate $Q$: a) top-view, b) side-view, c) cross-section.

## 4.2 Analysis with thermo-chemical sub-model

The thermo-chemical analysis means determination of the temperature and degree of cure fields in the composite material travelling in the pultrusion die. In our case the absorption energy field in the composite material determined with COMSOL Multiphysics is used as a heating source in the pultrusion process modeled with ANSYS Mechanical.

The advanced pultrusion die consists of microwave block with the ceramic inlet located at the entrance of the steel die (Fig. 8). The thermal properties of steel material used for an advanced die production is given in Table 3. Table 4 includes the thermal properties of ceramic material.

Heating of the composite profile is realized by using an absorption energy field obtained by COMSOL Multiphysics. It is simulated only in the composite material within boundaries of the ceramic inlet. Distribution of the absorption energy in the cross-section of the profile is constant (Fig. 7) but in the longitudinal direction varies accordingly to the law demonstrated in Figure 9. Integrating the curve presented in Figure 9 and taking into account area of the profile cross-section, it is possible to obtain the total absorption energy which is equal to 100.2 W. The simulated absorption energy field with the total value of 95.8 W is shown in Figure 10. It is necessary to note that temperature control is not applied in this case, and energy source is working continuously. Conditions of the pultrusion process are presented in Table 5.

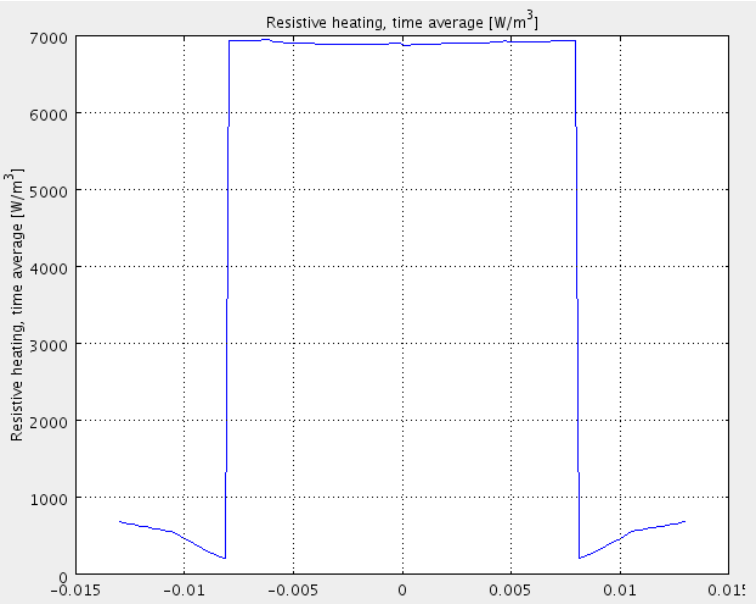



Figure 7: Power density in the cross-section of the ceramic pultrusion die.

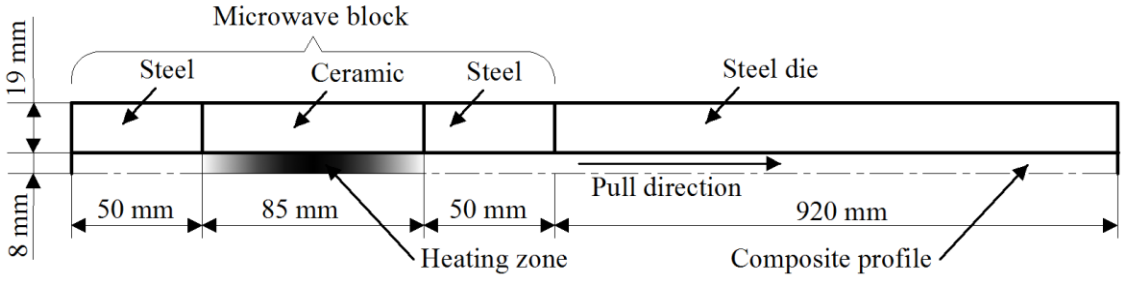



Figure 8: Scheme of the advanced pultrusion die.

The materials used for a production of the cylindrical rod are glass fibers Unifilo 4800 tex and polyester resin POLRES 305BV. The resin volume content in the pultruded material is 40% (resin mass fraction 22.6%). The thermal properties of the applied materials are given in Table 6. The rate of resin reaction is described by using the Kamal-Sourour curing kinetic model:

$$\frac{\partial \alpha}{\partial t} = \left( K_1 \exp\left( -\frac{E_1}{RT} \right) + K_2 \exp\left( -\frac{E_2}{RT} \right) \cdot \alpha^m \right) \cdot (1-\alpha)^n \qquad (14)$$

with the parameters presented in Table 7.

| Property | Symbol (unit) | Steel 40Cr |
|---|---|---|
| Density | $\rho$ (kg/m$^3$) | 7820 |
| Specific heat | $C$ (J/(kg·K)) | 460 |
| Thermal conductivity | $k_x$ (W/(m·K)) | 66.6 |

Table 3: Thermal properties of die material.

| Property | Symbol (unit) | Circonia ceramic |
|---|---|---|
| Density | $\rho$ (kg/m$^3$) | 6000 |
| Specific heat | $C$ (J/(kg·K)) | 418 |
| Thermal conductivities | $k_x$ (W/(m·K)) | 2 |

Table 4: Thermal properties of ceramic inlet.

| Name | Value | Unit |
|---|---|---|
| Room and resin temperature in the bath | 17 or 27 or 37 | $^0$C |
| Pulling speed | 0.18 or 0.12 or 0.06 | m/min |

Table 5: Conditions of pultrusion process.

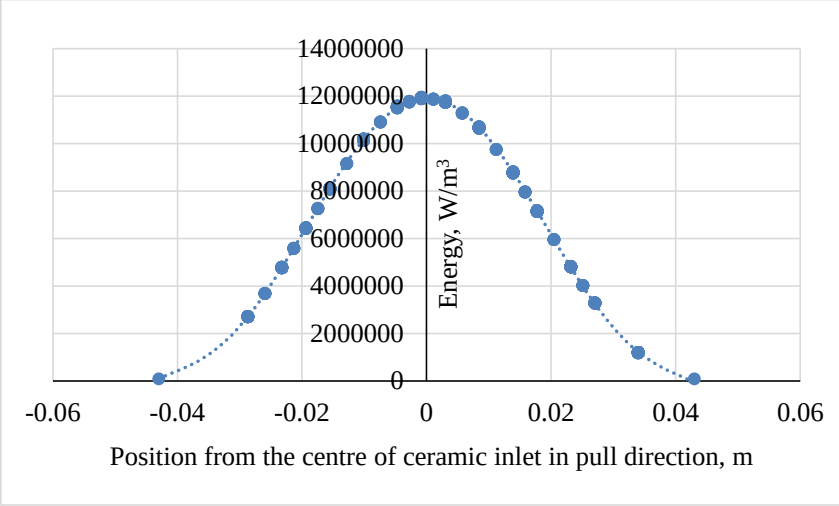



Figure 9: Distribution of the absorption energy in the longitudinal direction of the pultruded profile.

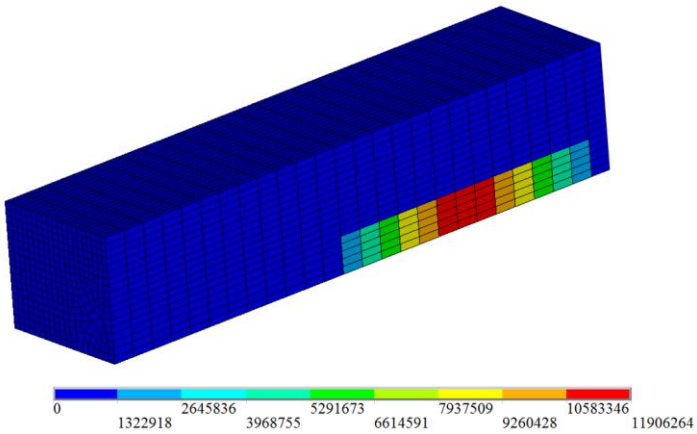



Figure 10: Simulation of the absorption energy in the pultruded profile.

The finite element model for a simulation of the cylindrical rod pultrusion has been created in ANSYS Mechanical by using 3-D thermal solid finite elements Solid 70. Using symmetry of the simulated domain, only a quarter of advanced pultrusion die is modeled. A fragment of the finite element model is presented in Figure 11. The finite element mesh is regular in pull direction and has 221 elements along the profile. The total number of finite elements is 74258 (3978 for the composite, 58512 for the die, 9728 for the steel part of microwave block and 2040 for the ceramic part). The time step of the solution depends on the pull speed. It is equal to 1.67 s for the speed of 18 cm/min, 2.5 s for the speed of 12 cm/min and 5 s for the speed of 6 cm/min. During this time the composite travels in the pull direction on the distance equals to the dimension of one finite element (5 mm). It is necessary to note that iterations (sub-steps) are not performed in the finite element analysis. Initial conditions are applied at time $t = 0$ when all nodal points of the composite have the room temperature and degree of cure equals to $\alpha = 10^{-10}$.

| Property | Symbol (unit) | Unifilo 4800 tex | POLRES 305BV |
|---|---|---|---|
| Density | $\rho$ (kg/m$^3$) | 2500 | 1100 |
| Specific heat | $C$ (J/(kg·K)) | 1235 | 1360 |
| Thermal conductivity | $k_x$ (W/(m·K)) | 11 | 0.209 |
| Thermal conductivity | $k_y$ (W/(m·K)) | 1 | |

Table 6: Thermal properties of pultruded material.

| Parameter | Symbol (unit) | POLRES 305BV |
|---|---|---|
| Heat reaction | $H_{tr}$, (J/kg) | 323074 |
| Frequency factor | $K_1$ (s$^{-1}$) | 14289310986 |
| Frequency factor | $K_2$ (s$^{-1}$) | 285.870 |
| Activation energy | $E_1$ (J/mol) | 85573 |
| Activation energy | $E_2$ (J/mol) | 33141 |
| Order of the reaction | $n$ | 2.342 |
| Order of the reaction | $m$ | 0.519 |

Table 7: Parameters of curing kinetic model.

Results of simulation, temperature and degree of cure obtained in the composite at the die exit in dependence on time, are presented in Figure 12 for room temperature 27°C and pull speed 12 cm/min. It is necessary to note that process is not steady-state after 55 min from the beginning of the numerical experiment in all tests for the degree of cure at the profile surface.

Results of simulation, temperature and degree of cure at the profile centerline and surface, are given for the time 50 min from starting of the pultrusion process in Figure 13 for room temperature 27°C and pull speed 12 cm/min. It is seen that, as opposed to conventional pultrusion process, now the curing comes considerably faster in the profile centerline than on the surface. So high degree of cure in the profile centerline is reached at the beginning of steel die already for the pultrusion process with pull speed equals to 18 cm/min and room temperature equals to 37°C. Moreover, full curing in the profile centerline happens already in the ceramic inlet for the pultrusion processes with pull speed equals to 6 cm/min and all examined room temperatures while the degree of cure at the profile surface has not reached high values.

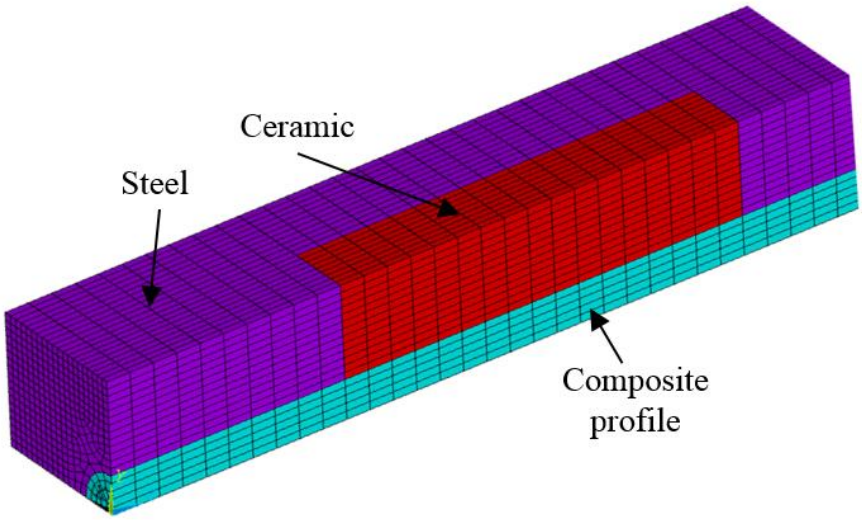



Figure 11: Fragment of the finite element model.

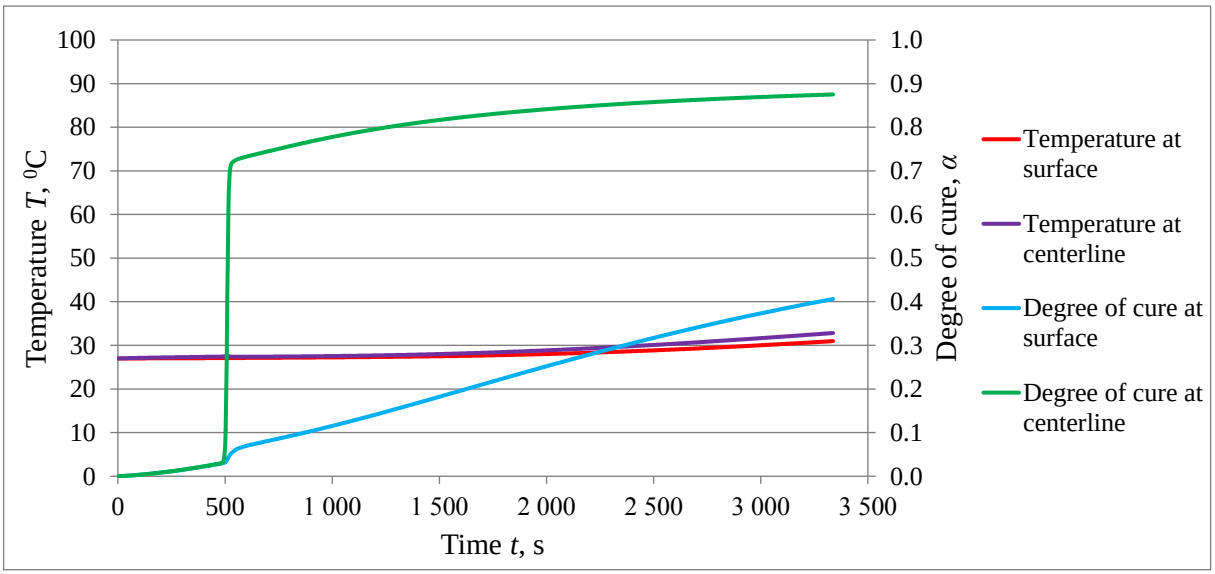



Figure 12: Dependence of temperature and degree of cure in the composite at the die exit on time for room temperature 27°C and pull speed 12 cm/min.

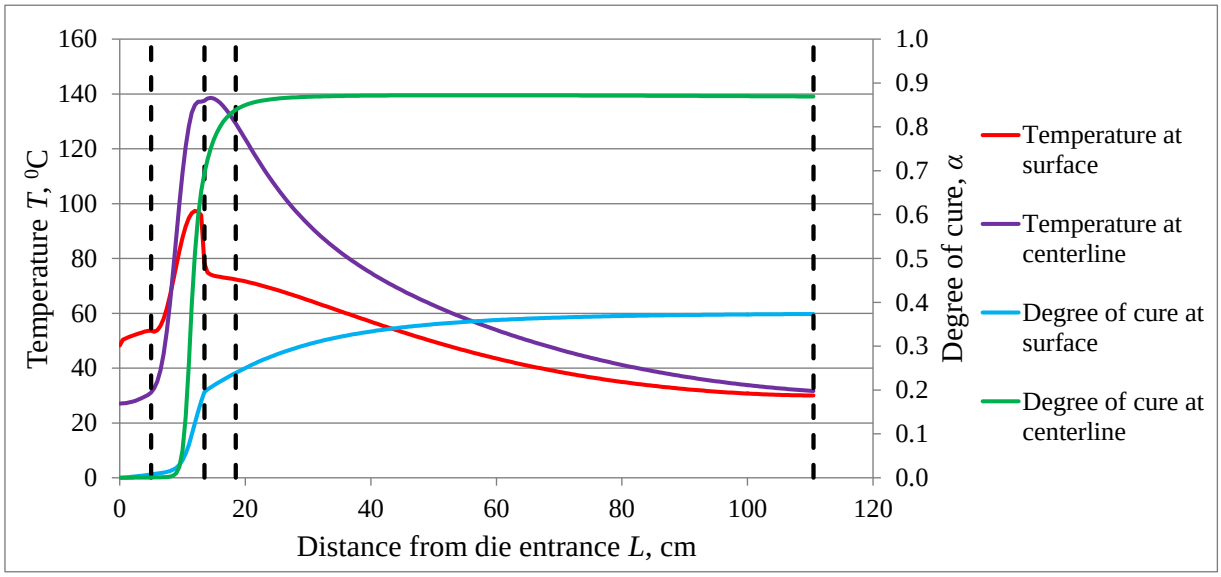



Figure 13: Temperature and degree of cure at the profile centerline and surface for room temperature 27°C and pull speed 12 cm/min.

To obtain the high value of the degree of cure at the profile surface in the advanced pultrusion die and to increase the pull speed of the process, improving by this way effectiveness of the pultrusion process, an influence of the following parameters of pultrusion process has been investigated:

- reduction in twice of steel die length,
- application of thermal insulation at all external surfaces of advanced pultrusion die (both steel die and microwave block),
- preheating of steel die, excluding microwave block, with the uniform temperature of 60°C before the beginning of pultrusion process,
- heating of resin bath (composite) with the temperature of 50°C for the full duration of pultrusion process.

An influence of the examined factors has been studied for the pultrusion process with the room temperature of 27°C and pulling speed of 12 cm/min. Simulation results are presented in Figures 14-17.

Figure 14 shows that reduction of steel die length has no effect on the degree of cure at the profile surface. However, any reduction of all important parameters has not been observed too. This fact could be used in the design of advanced pultrusion die for a saving of expensive die material. Figure 15 demonstrates that an application of thermal insulation at all external surfaces of advanced pultrusion die has increased considerably by 36% the value of the degree of cure at the profile surface. Figure 16 shows that only by preheating of steel die with the temperature of 60°C before the beginning of pultrusion process, it is possible to increase considerably by 29% the value of the degree of cure at the profile surface. These effects could be effectively used in the design of advanced pultrusion die. Figure 17 demonstrates that heating of resin bath has increased considerably by 44% the value of the degree of cure at the profile surface that could be used also in the design of advanced pultrusion die. However, it is necessary to note that an effectiveness of the pultrusion process could be not so high since the heating of resin bath is realized for the full duration of pultrusion process.

Joint influence of two factors having large effect on the degree of cure at the profile surface and on the process effectiveness, namely an application of thermal insulation and preheating of steel die, has been investigated additionally. Figure 18 shows that by preheating of

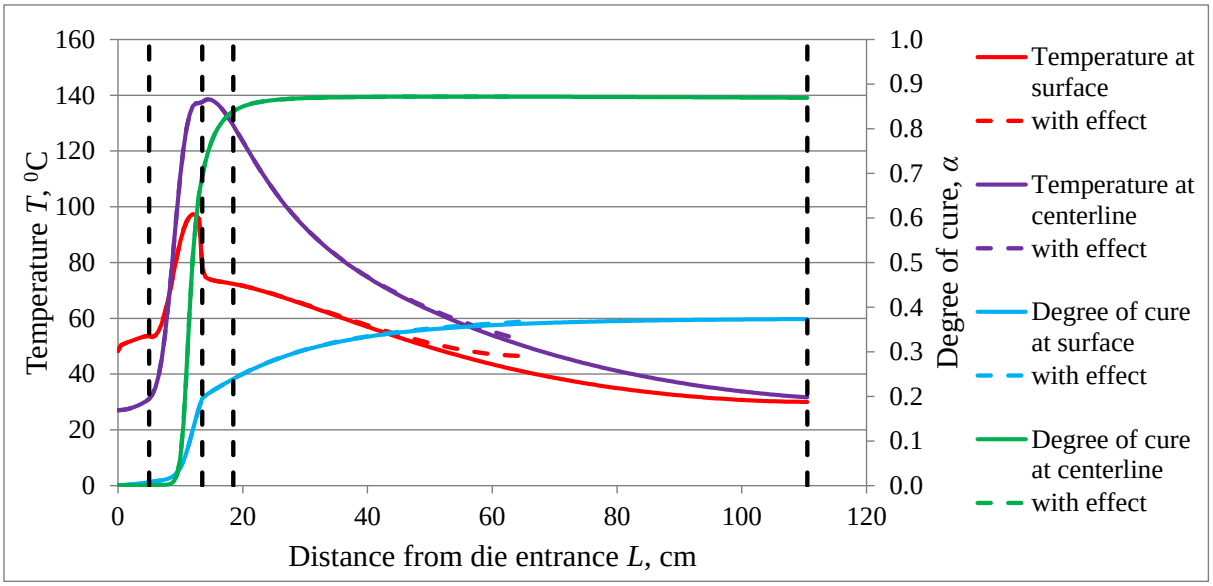



Figure 14: Temperature and degree of cure at the profile centerline and surface taking into account the effect of reduction of steel die length.

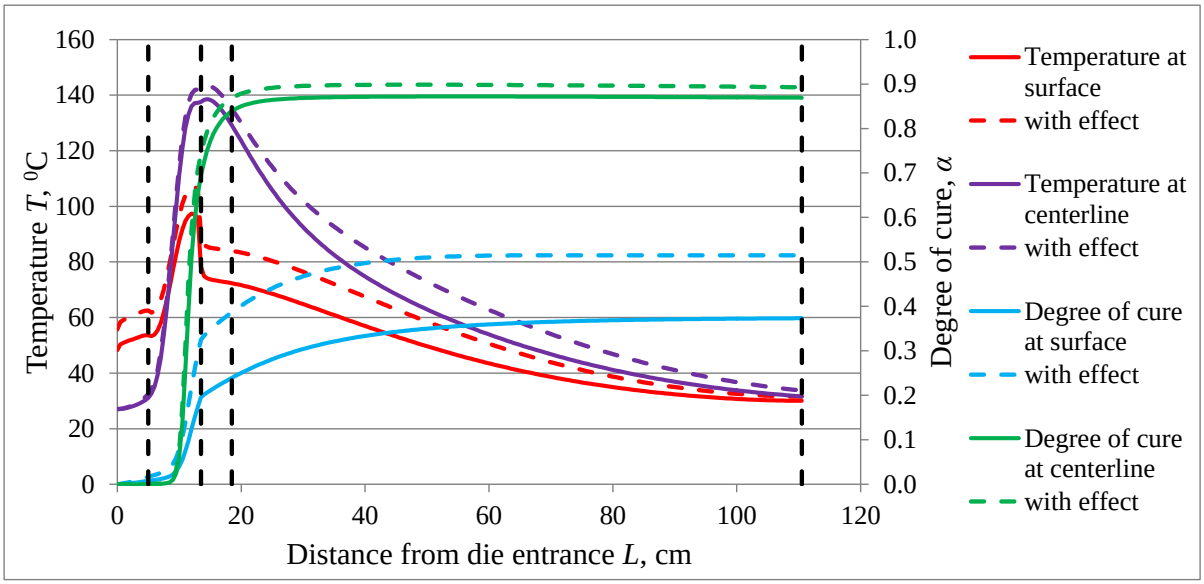



Figure 15: Temperature and degree of cure at the profile centerline and surface taking into account the effect of thermal insulation.

steel die with the temperature of 60°C before the beginning of pultrusion process and applying the thermal insulation, it is possible to increase considerably by 79% the value of the degree of cure at the profile surface. It is necessary to note that joint influence of two examined factors (79%) is higher than their separate influence summarized together (65%).

To obtain highest values of the degree of cure at the profile surface in the advanced pultrusion die, an influence of the capacity of the absorption energy field on the parameters of pultrusion process has been investigated. It is necessary to note that a distribution of the absorbed energy in the composite profile is not changed while its intensity has been increased proportionally with the coefficients: 1.5, 2.0 and 2.5. Figure 19 shows that the high value of the degree of cure at the profile surface could be reached already in the ceramic inlet of the microwave block. However, it is necessary to note that the material temperature in this case

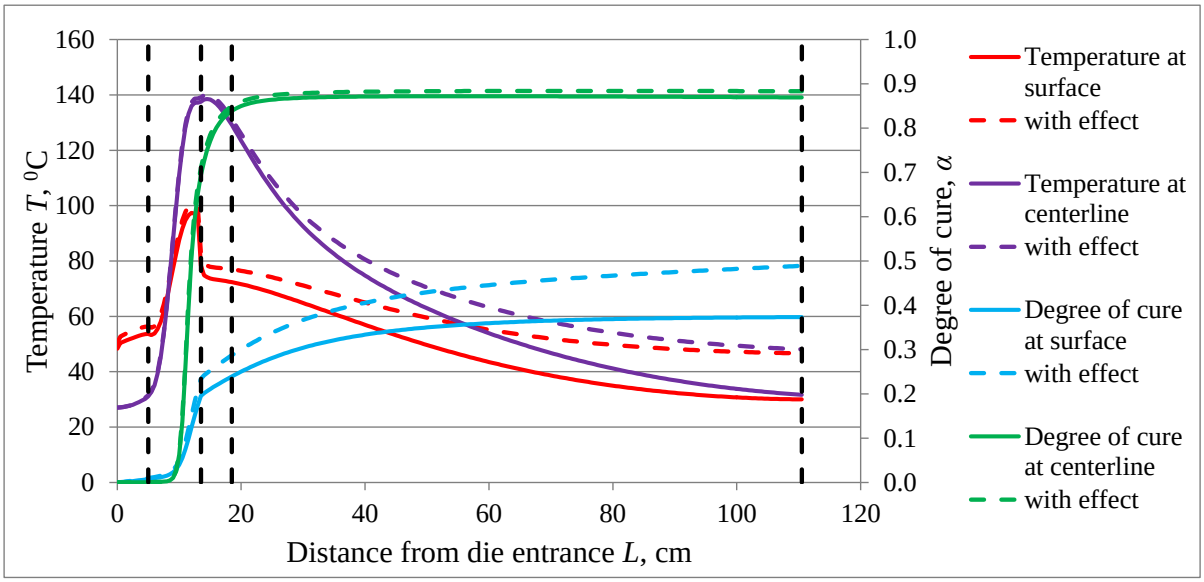



Figure 16: Temperature and degree of cure at the profile centerline and surface taking into account the effect of preheating of steel die.

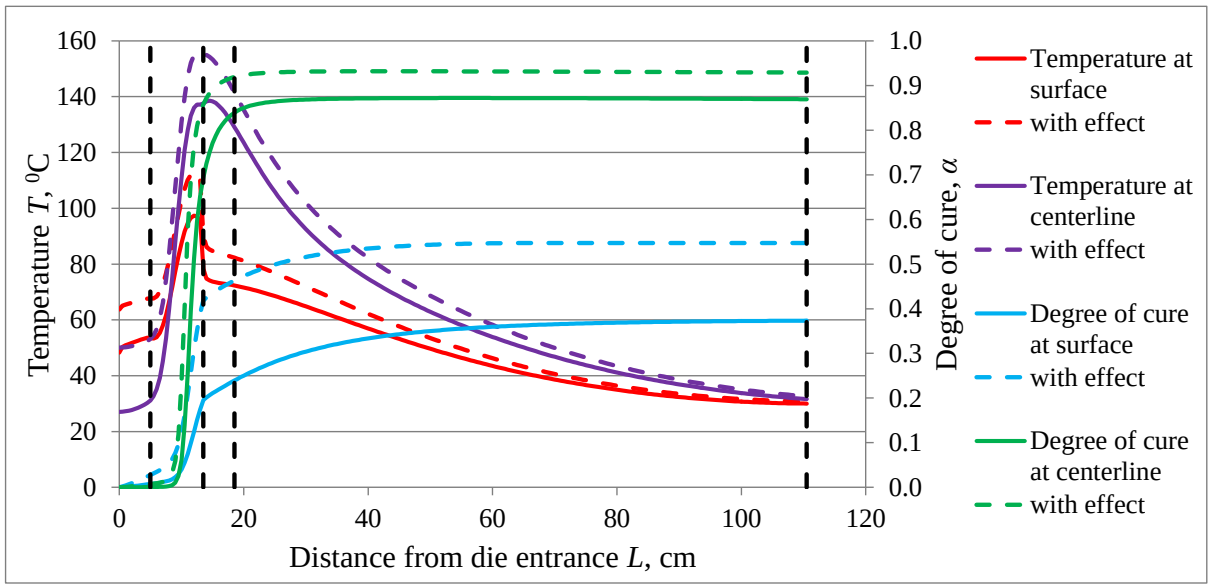



Figure 17: Temperature and degree of cure at the profile centerline and surface taking into account the effect of heating of resin bath.

could be higher than allowable resin temperature. To avoid overheating of the resin, the pulling speed should be increased in the same time that will contribute to the considerable increase of an effectiveness of the microwave assisted pultrusion process.

## 5 CONCLUSIONS

To understand the microwave assisted pultrusion process, to support the pultrusion tooling design and process control, new simulation methodology based on COMSOL Multiphysics and ANSYS Mechanical has been developed. This approach has been constructed by using the general-purpose FE software that results in considerable savings in development time and costs, and also makes available various modeling features of the FE packages.

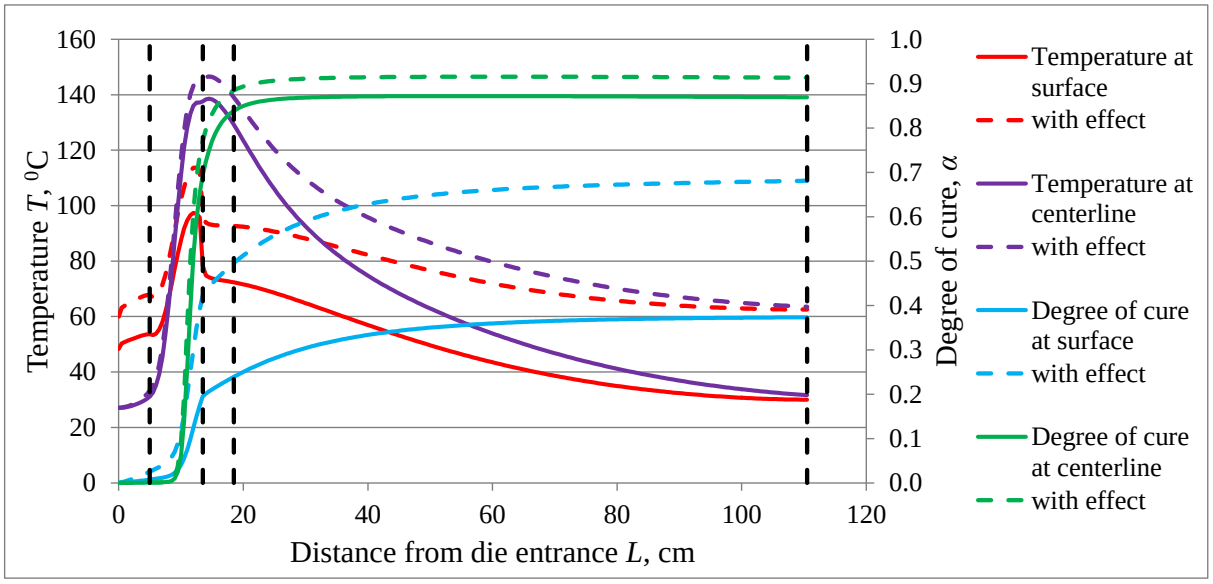



Figure 18: Temperature and degree of cure at the profile centerline and surface taking into account the effects of thermal insulation and preheating of steel die.

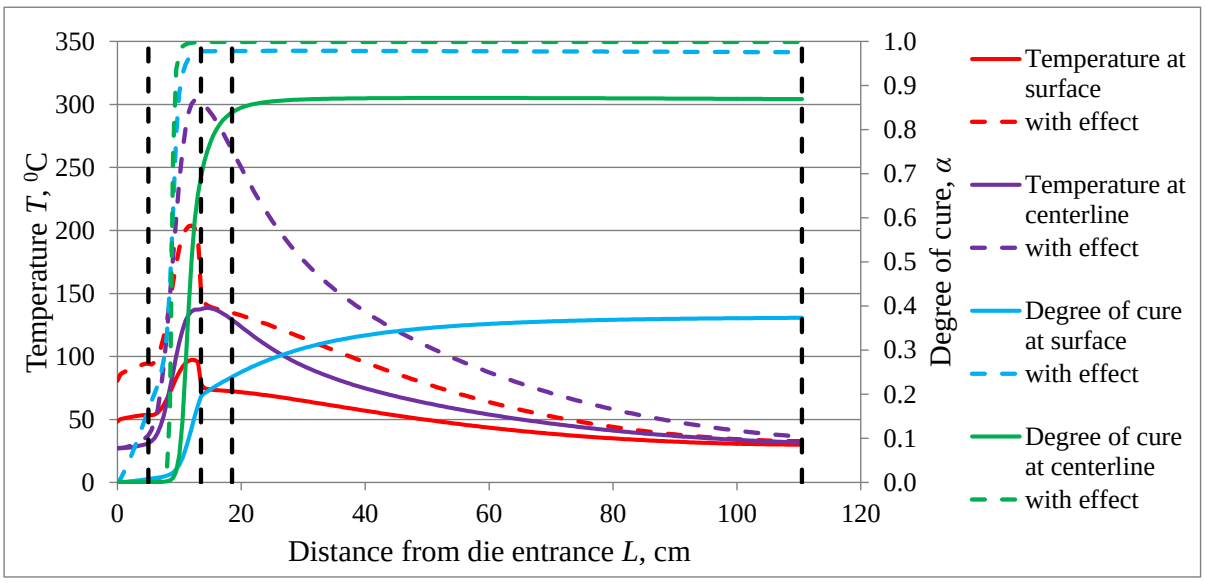



Figure 19: Temperature and degree of cure at the profile centerline and surface with the intensity coefficient of the absorption energy field equals to 2.5 (pull speed: 12 cm/min).

Conditions of the microwave assisted pultrusion process, and the temperature and degree of cure fields in the composite material, have been determined for the design of advanced pultrusion die. Simulation results show that the developed technology gives the possibility to ensure a homogeneous curing of the composite profile and to obtain high values of the degree of cure in the cured composite.

To improve an effectiveness of the microwave assisted pultrusion process, an influence of different process parameters has been additionally investigated. Based on the obtained results, the following recommendations could be done for technologists designing advanced pultrusion dies:

- More effectively is to apply the thermal insulation at all external surfaces of advanced pultrusion die or to preheat the steel die before the beginning of pultrusion process if the highest value of the degree of cure at the profile surface is required.

- To obtain largest effect on the degree of cure at the profile surface and on the process effectiveness, the joint application of both factors, namely an application of thermal insulation and preheating of steel die, should be realized.
- The heating of resin bath could be used also to obtain the highest values of the degree of cure at the profile surface. However, an effectiveness of the pultrusion process in this case is not so high since the heating of resin bath should be organized for the full duration of pultrusion process.
- The reduction of steel die length has no effect on the degree of cure at the profile surface. However, this effect could be used in the design of advanced pultrusion die for a saving of expensive die material.

## ACKNOWLEDGEMENTS

This work has been supported by the European Commission under FRAMEWORK7 program, contract No. NMP2-SL-2013-609149, project "Development of an Innovative Manufacturing Process for the In-line Coating of Pultruded Composites (COALINE)".

## REFERENCES

[1] A.C. Metaxas, R.J. Meredith, *Industrial microwave heating, IEEE Power Engineering Series 4, Peter Peregrinus Limited*. London, 1983.

[2] E.T. Thostenson, T.-W. Chou, Microwave and conventional curing of thick-section thermoset composite laminates: experiment and simulation. *Polymer Composites*, **22** (2), 197-212, 2001.

[3] L. Liu, S. Yi, L.S. Ong, K.S. Chian, S. Osiyemi, S.H. Lim, F. Su, Chemothermal modeling and finite-element analysis for microwave cure process of underfill in flip-chip packaging. *IEEE Transactions on Electronics Packaging Manufacturing*, **28** (4), 355-363, 2005.

[4] M. Lin, M.C. Hawley, Preliminary tests of the application of continuous microwave technique to pultrusion. *Proceedings of the 38 International SAMPE Symposium and Exhibition*, Anaheim, USA, 1993.

[5] A.C. Smith, A.M. McMillan, M.C. Hawley, Microwave pultrusion of graphite/epoxy composites using single frequency and variable frequency processing techniques. *Proceedings of the 30 International SAMPE Technical Conference*, San Antonio, USA, 1998.

[6] J.M. Methven, S.R. Ghaffariyan, A.Z. Abidin, Manufacture of fiber-reinforced composites by microwave assisted pultrusion. *Polymer Composites*, **21** (4), 586-594, 2000.

[7] P. Carlone, G.S. Palazzo, Numerical modeling of the microwave assisted pultrusion process. *International Journal of Material Forming*, **1** (1), 1323-1326, 2008.

[8] M. Sucher, J. Fox, *Handbook of microwave measurements, Vol. II, 3rd Edition, Polytechnic Press*. New York, London 1963.

[9] *Hewlett-Packard: Basics of measuring the dielectric properties of materials, Application Note 1217-1, Hewlett-Packard Co.*, 1992.

[10] *Hewlett-Packard: HP-Network Analyzer 8720 D + Dielectric Probe Kit 85070B: User´s Manual*, 1993.

[11] Püschner, *Wärme durch Mikrowellen*. Eindhoven: Philips Technische Bibliothek, 1964.

[12] C.E. Shannon, Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers*, **37** (1), 10-21, 1949.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFICIENT NUMERICAL SIMULATION OF INDUSTRIAL SHEET METAL BENDING PROCESSES

**Ch. Zehetner[1], P. Reimer[1], F. Hammelmüller[1], H. Irschik[2] and W. Kunze[3]**

[1] Linz Center of Mechatronics GmbH
Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: {christian.zehetner, paula.reimer, franz.hammelmueller}@lcm.at

[2] Johannes Kepler University of Linz
Altenbergerstrasse 69, 4040 Linz, Austria
hans.irschik@jku.at

[3] Salvagnini Maschinenbau GmbH
Dr. Guido Salvagnini-Straße 1, 4482 Ennsdorf, Austria
wolfgang.kunze@salvagninigroup.com

**Keywords:** plasticity, sheet metal bending, shell theory

**Abstract.** *In industrial production processes focus is given to high precision, quality, resource efficiency and productivity. In order to achieve these goals, efficient numerical simulation models are required. In the following, we consider an industrial sheet metal bending process, in which the sheet is fixed on one side and formed by a moving tool. On the one hand, there is a very large number of process parameters. On the other hand, the production sites are complex and have to be modelled in detail. Parameter studies are very extensive and take a large numerical effort. Therefore high efficient simulation tools are necessary to handle this challenge. In this paper two strategies for increasing the efficiency of modelling are investigated. First, the main focus is set to an efficient Finite Element model for sheet bending. Two Finite Element formulations are compared based on 3D-continuum elements and continuum shell elements. Secondly, a proper normalized formulation (similarity solution) of the bending process is derived starting from a reduced-order model, which subsequently is successfully applied to the complex bending process. Utilizing the concept of similarity, the number of cases in parameter studies can be reduced significantly.*

## 1 INTRODUCTION

In industrial metal forming processes, high versatility is claimed with respect to material types, geometric dimensions and process parameters. In order to control and optimize the production process, powerful simulation models are required. Usually, metal forming simulation models are very complex and highly non-linear. Thus, parameter studies are accompanied by a large computational effort such that an efficient simulation environment is required.

In sheet bending processes we deal with thin structures, which can be efficiently modelled by plate elements. In section 2, appropriate formulations are investigated, based on the bending of a one-side clamped plate. Using the powerful code ABAQUS, the results for continuum shell elements and three dimensional solid elements are compared. The goal is to find suitable configurations (mesh, element type, etc.) and the range of applicability of the specific elements.

A second important aspect is that the number of cases in parameter studies can be reduced significantly by using the concept of similarity, cf. [1] and [2] for the theory of similarity. Several combinations of parameters behave in a similar way. It is the challenge to find the appropriate parameter combinations, i.e. a reduced set of non-dimensional parameters describing the physical process without loss of information. In [3] this has been demonstrated for the bending of a cantilever beam subject to inertial loads arising in a drop test. In section 3, the similarity concept is applied to the one-side clamped plate and it is demonstrated how the number of dimensional parameters can be reduced by appropriate non-dimensional ones. This is first investigated analytically in the framework of a geometrically linearized elasto-plastic theory. A similarity formulation is derived which is successfully applied to the real large-deformation bending process afterwards.

## 2 FINITE ELEMENT MODEL OF SHEET BENDING

In the following a thin sheet of metal with length $L$, width $B$ and thickness $H$ is clamped on the left side and subject to a distributed force $f$ per unit width at the right side as shown in Fig. 1. The total bending force is $F = fB$.

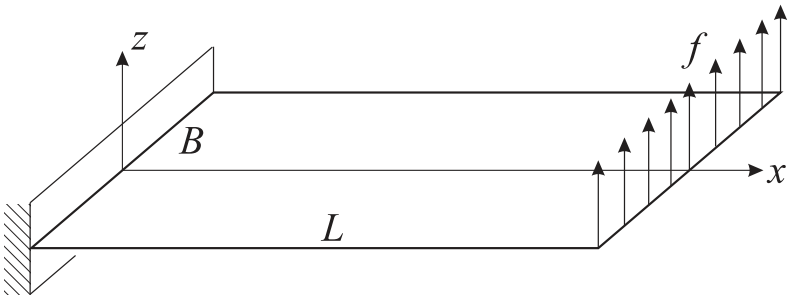



Figure 1: Bending of a plate

The material behavior is assumed to be elasto-plastic with exponential stress-strain relation of Ludwik's type, e.g. [3] and [4]. For a uni-axial state of stress the constitutive relations can be expressed as

$$\sigma = E\varepsilon \quad \text{for } \sigma \le \sigma_Y, \qquad \sigma = a\varepsilon^b \quad \text{for } \sigma \ge \sigma_Y, \tag{1}$$

where $\sigma$ is the axial stress, $\varepsilon$ the axial strain, $E$ Young's modulus, $a$ and $b$ specify the exponential stress-strain relation, and $\sigma_Y$ is the yield stress, following from solving Eq. (1) as

$$\sigma_Y = E\left(E/a\right)^{1/(b-1)}. \tag{2}$$

For this bending process, a Finite Element model has been implemented in ABAQUS, see [5] considering large strains. A quasi-static solution is computed with the implicit solver (ABAQUS standard).

## 2.1 Finite Element Discretization and Element Types

Figure 2 shows two discretizations of the plate. For numerical reasons, the clamping is here approximated by two elastic bodies which are fixed on the outer surfaces. The distributed force is applied by a moving elastic cylindrical body. Between the bodies contact is defined. Figure 2a shows a fine mesh with three-dimensional solid elements (C3D8R). On the other hand, the plate is discretized by continuum shell elements (SC8R) as defined in the ABAQUS documentation [5]. The continuum shell elements consider transverse shear stiffness as well as thickness contraction. They are capable for multi-layered shells. In the present context, two equal layers are used, where the continuum shell nodes work as super-elements, see Figure 2.

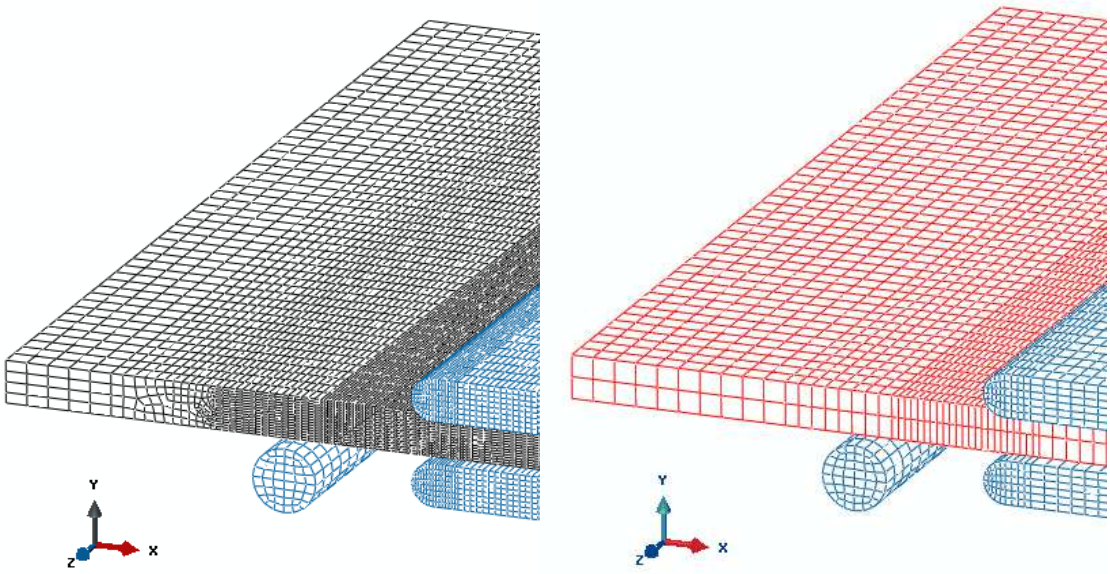



Figure 2: Finite Element models (a) 3D-continuum elements, (b) continuum shell elements

## 2.2 Results

Figure 3 shows a comparison of the deformed configuration of the two element types for a bending angle of about 90°. Near the clamping there are slight differences in the displacement field since the cross-sectional warping is restricted in the continuum shells. Note that even for large bending angles the cross-sections remain approximately plane, such that it is possible to obtain suitable results by plate theory.

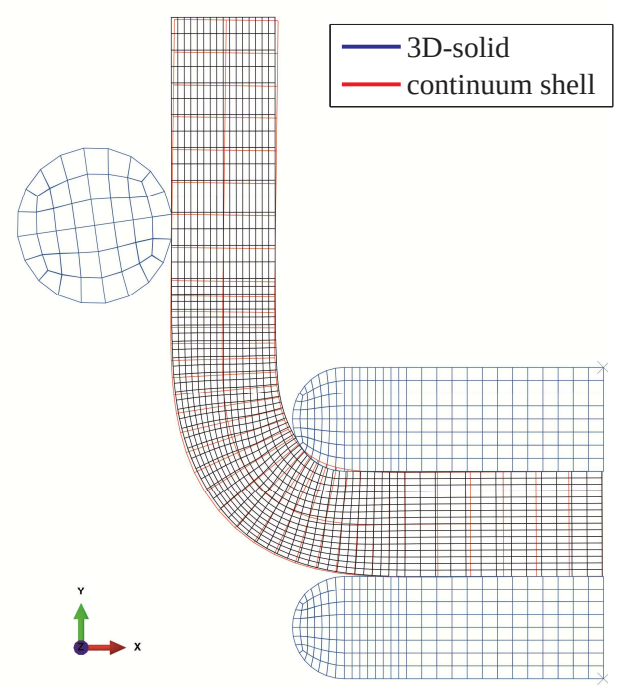



Figure 3: Deformed shape

Secondly, the contact force between cylinder and plate is compared, also denoted as bending force. The results are shown in Fig. 4, where a normalized resulting contact force is shown as a function of the bending angle. The results show a good correspondence for all bending angles.

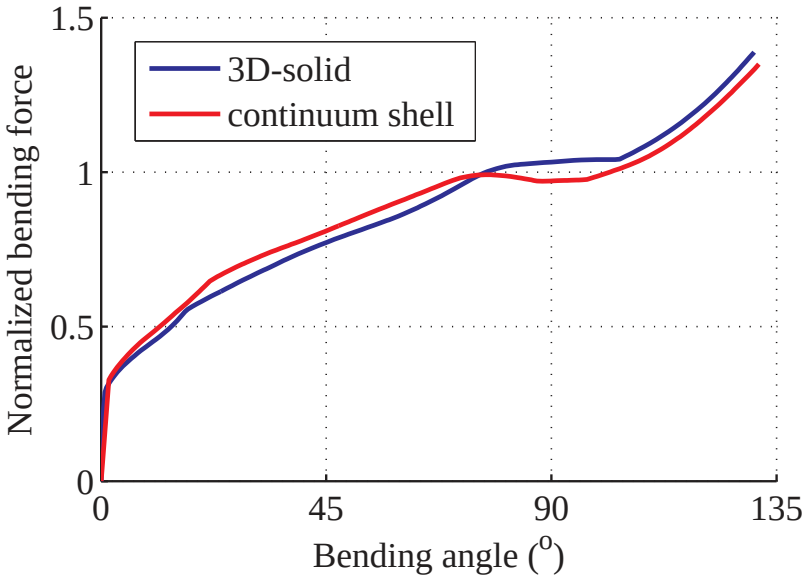



Figure 4: Comparison of the contact force

Finally the distribution of stress in the plate is compared for an angle of 90°. Figure 5 shows the result of a normalized v.Mises stress along the normalized coordinate *x/L* (see Fig. 1) on the top (a) and on the bottom (b) of the plate. The results for 3D-solids and continuum shells show a good coincidence especially for the maximum values. In this range the normal

strains are dominating, i.e. bending and longitudinal extension. In the surrounding of the clamping and near the contact with the cylinder, there are some differences resulting from a more general state of stress which is not represented in such detail by the continuum shell elements. Nevertheless, the result of the contact force is reliable as shown in Figure 4.

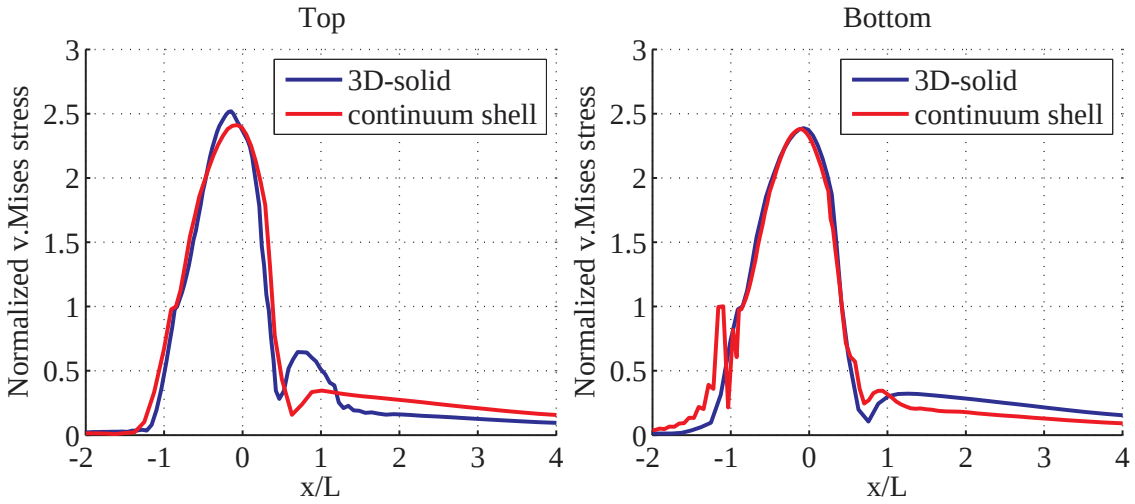



Figure 5: v.Mises stress

Finally the computation time is compared in Table 1 for the two element types. With the continuum shells, a speed-up factor of about 40 has been obtained.

| Element type | Total CPU time (sec) |
|---|---|
| 3D-solid | 140356 |
| Continuum shell | 3672 |

Table 1: Computation time

### 2.3 Discussion

In sections 2.1-2.2 the results of plate bending obtained with 3D-solid and continuum shell elements have been compared. Displacement and stress fields as well as the resulting bending force show a good coincidence even for large deformations. The advantage of continuum shells is a significant reduction of computation time.

Since the bending process can be modelled with shell elements with satisfactory accuracy, it appears to be feasible to derive a non-dimensional (similarity) formulation of the complex bending process by using a strongly reduced modelling, as shown in the following section.

## 3 NON-DIMENSIONAL FORMULATION OF SHEET BENDING

Figure 6 shows the bending of a cantilever beam, i.e. a projection of the plate in Fig 1. The flexible transversal displacement is denoted as $w$. For the sake of deriving a proper similarity formulation, we perform model reduction and use the geometrically linearized (small strain and displacement) elasto-plastic theory of beams.

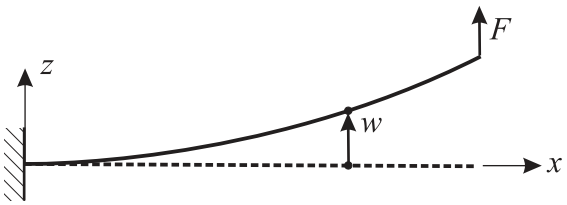



Figure 6: Bending of a beam/plate

In the framework of a quasi-static modeling, the bending moment follows to

$$M = -F(L-x). \tag{3}$$

### 3.1 Plastic zones

The bending moment as a resultant of the stress is written as

$$M = B\int \sigma(z) z dz. \tag{4}$$

In the elastic range, the distribution of the stress is linear over the cross-section. The yield moment corresponds to the state where the yield stress is reached at $z=H/2$, i.e. $M_Y = \frac{1}{6}\sigma_Y BH^2$. In the quasi-static case, the yield load follows from Eq. (3) by substituting $x=0$,

$$F_Y = \frac{BH^2\sigma_Y}{6L}. \tag{5}$$

If the load is higher than the yield load, the domain is divided into elastic and plastic zones as shown in Figure 7.

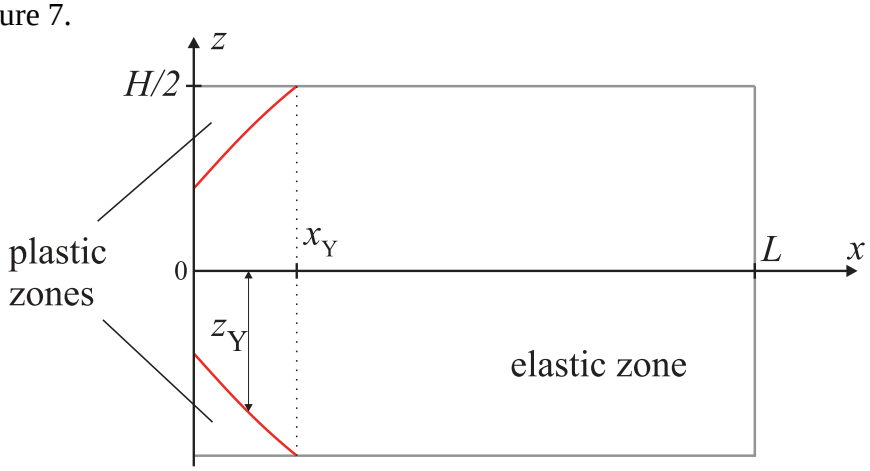



Figure 7: Elastic and plastic zones

Assuming that the deformations remain small and that the Bernoulli-Euler hypothesis for beams does hold, the axial strain is expressed as

$$\varepsilon = -zw''. \tag{6}$$

A prime denotes differentiation with respect to x. Thus, the stress in the elastic and plastic zones can be expressed by

$$\sigma = \sigma_Y \frac{z}{z_Y} \qquad \text{for } z \le z_Y, \tag{7}$$

$$\sigma = a\left(\frac{\sigma_Y}{E}\frac{z}{z_Y}\right)^b \qquad \text{for } z \ge z_Y.$$

The boundary $z_Y$ of the plastic zone follows from equating Eqs. (3) and (4), using Eq. (7):

$$2B\left(\int_0^{z_Y} \sigma_Y \frac{z}{z_Y} z\,dz + \int_{z_Y}^{H/2} a\left(\frac{\sigma_Y}{E}\frac{z}{z_Y}\right)^b z\,dz\right) = -F(L-x). \tag{8}$$

Integration yields

$$\frac{\sigma_Y z_Y^2}{3} + a\left(\frac{\sigma_Y}{E z_Y}\right)^b \left(\frac{H^{b+2}}{2^b(4b+8)} - \frac{z_Y^{b+2}}{b+2}\right) = -\frac{F}{2B}(L-x). \tag{9}$$

From Eq. (9) the length $x_Y$ of the plastic zone is obtained by substituting $z_Y$=$H$/2. An analytical solution is possible. The height $z_Y$ of the plastic zone is obtained as a function of $x$. Due to the nonlinearity of the equation, $z_Y$ has to be solved numerically.

### 3.2 Curvature

From Eqs. (1) and (6) follows that

$$\sigma_Y = -E z_Y w''. \tag{10}$$

Inserting Eq. (10) into Eq. (9) yields an equation for the curvature in the elasto-plastic case:

$$-\frac{w'' E z_Y^3}{3} + a\left(-w''\right)^b \left(\frac{H^{b+2}}{2^b(4b+8)} - \frac{z_Y^{b+2}}{b+2}\right) = -\frac{F}{2B}(L-x) \tag{11}$$

Due to the nonlinearity with respect to $w''$, only a numerical solution is possible. For the elastic range, $x > x_Y$, Eq. (11) simplifies to the well-known form

$$w'' = \frac{12F}{BH^3E}(L-x). \tag{12}$$

### 3.3 Normalization

In order to obtain a generalized representation, a normalization of the above equations has been performed as follows: By introducing the non-dimensional quantities

$$\xi = \frac{x}{L}, \quad \zeta = \frac{z}{L}, \quad H^* = \frac{H}{L}, \quad a^* = \frac{a}{E}, \quad F^* = \frac{F}{aLB}, \quad \sigma^* = \frac{\sigma}{E}, \tag{13}$$

Eq. (8) can be rewritten as

$$2a^{**}\zeta_Y^2(2+b-3(a^{**})^b)+3\left(\frac{1}{2}\right)^{1+b}a^*H^{*2}(a^{**}H^*\zeta_Y^{-1})^b=\frac{3}{2}F^*a^*(\xi-1)(2+b) \tag{14}$$

with

$$a^{**}=\left(a^*\right)^{\frac{1}{1-b}}. \tag{15}$$

Substituting $\zeta_Y = H^*/2$ and solving Eq. (14) yields the normalized extension $\xi_Y$ of the plastic zone. On the other hand side, $\zeta_Y$ is obtained as a function of $\xi$. The normalized curvature and deflection are defined by

$$w^*=\frac{w}{L},\quad (w^*)''=Lw'' \tag{16}$$

and the following convention is used:

$$(w^*)''=\frac{d^2w^*}{d\xi^2}. \tag{17}$$

Substituting Eq. (13) into Eq. (11) yields the equation for the normalized curvature for the elasto-plastic range ($\xi \le \xi_Y$):

$$4(w^*)''\zeta_Y^3\left(2+b\right)+3a^*\left(-(w^*)''\right)^b\left(4\zeta_Y^2\left(\zeta_Y\right)^b-H^{*2}\left(H^*\right)^b\right)=3F^*a^*(2+b)\left(1-\xi\right) \tag{18}$$

Solving Eq. (18) numerically we obtain the curvature as a function of $\xi$. By numerical integration of Eq. (18) the deflection can be computed. Figure 8a shows the plastic tip displacement $w_{tip,plast}(L)$ as a function of the load $F$ related to the yield force $F_Y$. In this dimensional representation, a separate curve is obtained for each configuration of $H$ and $L$. However, in the non-dimensional representation in Figure 8b, in which non-dimensional quantities according to Eq. (13) are used, only one curve is obtained for the three cases.

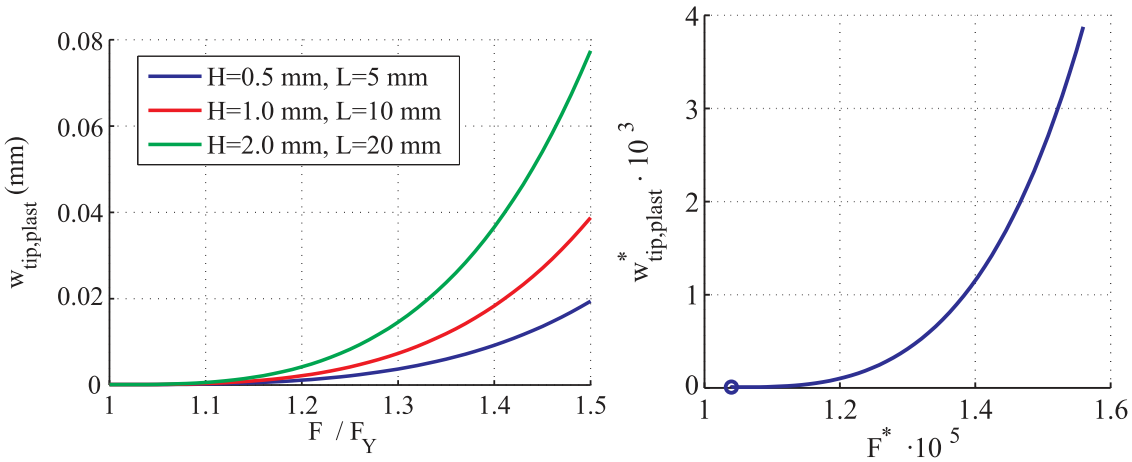



Figure 8: Plastic tip displacement (a) dimensional and (b) non-dimensional representation

### 3.4 Discussion of the non-dimensional formulation

In the dimensional representation with the coordinates $x$, the bending process is specified by eight physical quantities:

$$w, F, E, a, b, H, L, B. \tag{19}$$

On the other hand, in the dimensionless formulation derived in section 3.3 with the normalized coordinate $\xi = x / \mathrm{L}$ the number of essential parameters has been reduced to the following five non-dimensional quantities

$$\frac{w}{L}, \frac{F}{aBL}, \frac{a}{E}, b, \frac{H}{L}. \tag{20}$$

The results show that for the simple model of an elasto-plastic beam bending process the same normalized quantities have been derived as have been obtained by Baker et al. [1] using the Buckingham Pi theorem. For the large-deformation elasto-plastic bending problem with distributed loading, see also [3].

Recall that all the results in this section are based on the geometrically linear Bernoulli-Euler beam theory with exponential stress-strain relation. Thus the results are valid for small deformations. It has been shown in [3] that the non-dimensional parameters according to Eq. (20) are also suitable for a non-dimensional formulation in case of large strains. This gives evidence for using it by analogy also for the complex industrial sheet metal bending process.

### 3.5 Industrial application

The results of sections 3.1-3.4 have been applied to an industrial bending process. On a Salvagnini automatic panel bender, the bending forces for six bending processes (90° bending angle) have been measured for three sheet thicknesses ($H_1$, $H_2$, $H_3$) with two values for length ($L_1$, $L_2$) respectively. By measuring the forces of the drives, the resulting bending force has been computed. In Figure 9 the measured force for the six cases is shown as a function of the bending angle. In this representation the force is related to the maximum admissible force on the machine.

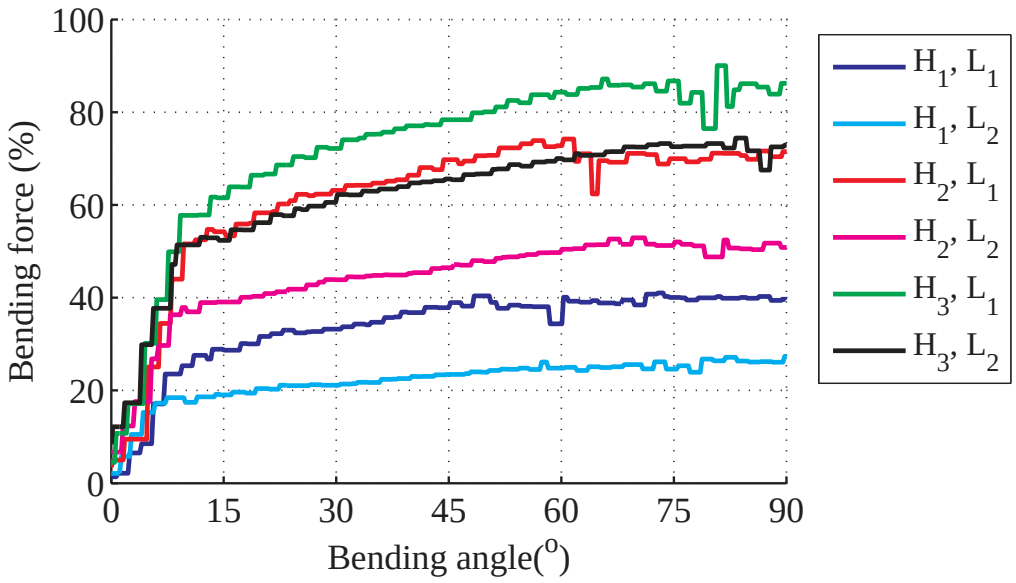



Figure 9: Measurement of bending force, dimensional representation

The normalized bending force according to Eq. (20) is shown in Figure 10. The normalized representation demonstrates the reduction of complexity by utilizing similarity of several parameter combinations. The concept of similarity is a powerful tool for analyzing complex physical processes.

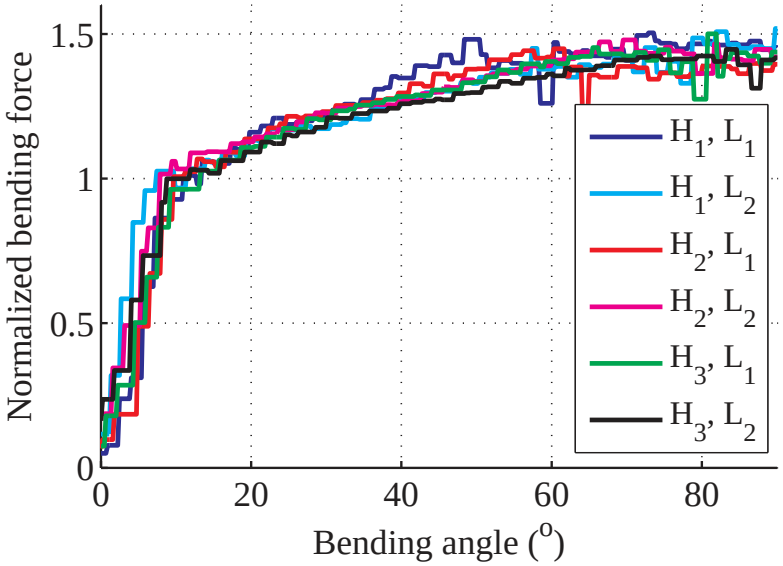



Figure 10: Measurement of bending force, normalized representation

## 4 CONCLUSIONS

For an industrial bending process two strategies for reducing the numerical effort of parameter studies have been investigated. First, an efficient and precise Finite Element formulation based on continuum shell elements has been investigated. Secondly it has been shown how it is possible to find similar configurations by an appropriate non-dimensional formulation. The method has been verified by measurements on an industrial bending machine. These two strategies are very powerful tools for optimizing the bending process.

## ACKNOWLEDGMENT

This work has been supported by the Austrian COMETK2 program of the Linz Center of Mechatronics (LCM), and was funded by the Austrian federal government and the federal state of Upper Austria.

## REFERENCES

[1] W.E. Baker, P.S. Westine, F.T. Dodge, *Similarity Methods in Engineering Dynamics, Theory and Practice of Scale Modeling*, Elsevier, 1991.

[2] L.I. Sedov, *Similarity and Dimensional Methods in Mechanics*, CRC Press, 1993.

[3] Ch. Zehetner, F. Hammelmüller, H. Irschik, W. Kunze, Similarity Methods in Elasto-Plastic Beam Bending. E. Oñate, D.R.J. Owen, D. Peric and M. Chiumenti eds. *XIII International Conference on Computational Plasticity (COMPLAS XIII)*. Barcelona, Spain, September 1-3, 2015.

[4] W.F. Chen, D.J. Han, *Plasticity for Structural Engineers*, J.Ross Publishing (2007)

[5] Abaqus 6.14 Documentation, Dassault Systèmes, 2015

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL MODELLING AND SIMULATION OF SHEET METAL CUTTING PROCESSES

**P. Reimer[1], Ch. Zehetner[1], F. Hammelmüller[1] and W.Kunze[2]**

[1] Linz Center of Mechatronics GmbH
Altenbergerstrasse 69, 4040 Linz, Austria
e-mail: {paula.reimer, christian.zehetner, franz.hammelmueller}@lcm.at

[2] Salvagnini Maschinenbau GmbH
Dr. Guido Salvagnini-Straße 1, 4482 Ennsdorf, Austria
wolfgang.kunze@salvagninigroup.com

**Keywords:** sheet metal cutting, coupled Eulerian-Lagrange, Smoothed Particle Hydrodynamics

**Abstract.** *Material processing is a very important industrial sector. In order to guarantee high precision and quality of the products, reliable numerical simulation models are required. This contribution concerns the simulation of material cutting. As a benchmark example, a sheet metal is fixed by two clamping tools and the cutting process is controlled by a moving blade. The tools are modelled as linear elastic materials, and the sheet metal as elasto-plastic material with hardening. Main scope is the comparison of two simulation methods with respect to industrial application: (i) Coupled Eulerian-Lagrangian Finite Element Method and (ii) Coupled Lagrangian Finite Element and Smoothed Particle Analysis. Numerical simulation models for the two variants of the above benchmark example are implemented in the commercial code Abaqus. The numerical results of the two models are compared with respect to accuracy and numerical effort, and the advantages and disadvantages of the two methods are investigated. The implemented models for the cutting process and the materials can be applied to several other kinds of industrial material processing like stamping or punching.*

## 1 INTRODUCTION

Material processing is a very important industrial sector. In order to guarantee high precision and quality of the products, reliable numerical simulation models are required. This contribution concerns the simulation of cutting processes as part of a mechatronic system.

An example of a complex mechatronic system is an automatic panel bender in which complete products of sheet metal are fabricated. During production it is necessary to cut the panels causing high loads on the machine. To optimize the cutting process itself as well as the lifetime of the machine refined and reliable simulation models are required. It is the scope of this work to develop a simulation model for the cutting of the panel which can be adapted by measurement results.

Another example is the cutting of different kinds of material with a chain saw. On the one hand the dynamic behavior of the chain has influence on quality of the process, lifetime of the machine and noise emission. The load of the chain due to the cutting forces is a very important parameter in the system, but it is not possible to measure these forces directly, so that appropriate simulation models are required.

It is the scope of this work to investigate and compare two simulation methods for cutting. As a simple benchmark example, a sheet metal is fixed between two clamping tools and the cutting process is controlled by a moving blade. Two simulation strategies are compared:

First, the coupled Eulerian-Lagrangian Finite Element Method [1] is considered: The clamping and cutting tools are modelled by three-dimensional Lagrangian Finite Elements, and the sheet metal by three-dimensional Eulerian elements. The coupling of the elements is achieved by an appropriate contact formulation.

Secondly, the sheet is modeled by Smoothed Particles [2], a mesh-free discretization with Lagrangian formulation. Again, the tools are modelled by three-dimensional Lagrangian Finite Elements, and the coupling of the two models is realized by contact interaction.

Numerical simulation models for these two variants of the benchmark example are implemented in the powerful commercial code Abaqus [3]. The numerical results of the two methods are compared, and the advantages and disadvantages of the respective methods are investigated.

Finally, the results are compared to measurement on the automatic panel bender mentioned above. Because of differing dimensions of the sheet in the simulation model and on the real machine, only a qualitative comparison is possible which show a better correspondence of the measurements with the Smoothed Particle model than with the Coupled Euler-Lagrangian model.

## 2 SIMULATION METHODS

In this work two simulation methods are compared for modeling a cutting process. In both cases the clamping tools and the cutting blade are linear elastic bodies modeled by Lagrangian Finite Elements. During cutting large deformations arise in the sheet, and the material is divided. There are several strategies to handle this. One possibility is to use Lagrangian Finite elements for the sheet and an appropriate damage model. In this paper, two other methods are applied to model the sheet without definition of a damage model.

First, a coupled Eulerian-Lagrangian analysis (CEL) is studied, cf [1]. The sheet is modeled by Eulerian elements. In contrast to Lagrangian elements, the material flows through the mesh, and the mesh is fixed in space. This strategy is capable for modelling very large deformations. However, the shape of the deformed material cannot be resolved as good as with a Lagrangian formulation. The coupling of Lagrangian and Eulerian Elements has to be done by an appropriate contact formulation.

As a second strategy, the sheet is discretized by Smoothed Particles. Smothed Particles Hydrodynamics (SPH) is a meshfree Lagrangian method, [2]. With this method fluids as well as solids can be modelled. The material is divided into discrete particles which are interacting to each other within a defined area.

As mentioned before, it is not necessary to define a damage model for CEL and SPH. It is the scope of this work to compare the influence of the formulations on the behavior during cutting, and to find out the applicability for industrial purposes.

## 3 SIMULATION MODELS OF THE CUTTING PROCESS

The set-up of the simulation model is shown in Figure 1. A sheet metal of 4 mm thickness and 6 mm width is fixed by an upper and a lower clamping tool. The cutting process is controlled by a blade with 6° inclination. It moves downwards with a cutting speed of 0.5 m/s, which is a common average cutting velocity in blanking processes due to [4]. The cutting clearance is 0.1 mm. In both simulation models the clamping tools and the blade are discretized by Lagrangian elements with an edge length of approximately 0.5 mm.

Figure 1a) shows the SPH model. The sheet is discretized by Smoothed Particles. In the unloaded, initial state the particles are evenly spread at intervals of 0.25 mm in each spatial direction.

The CEL model is presented in Figure 1b) and 1c). Subplot c) shows the clamping tools, the blade and the complete Eulerian mesh with an element size of 0.25 mm. The material of the sheet metal can move and deform only within the region that is covered by the Eulerian mesh [3]. Material that leaves the mesh is no longer taken into account within the simulation. In the initial state most of the Eulerian mesh is void. Only the red region shown in Figure 1b is occupied by material. During the cutting process a chip is generated and displaced by the blade. Because of the relatively large volume needed to preserve the chip from leaving the model space the number of Eulerian Finite elements is much higher than the number of Smoothed Particles.

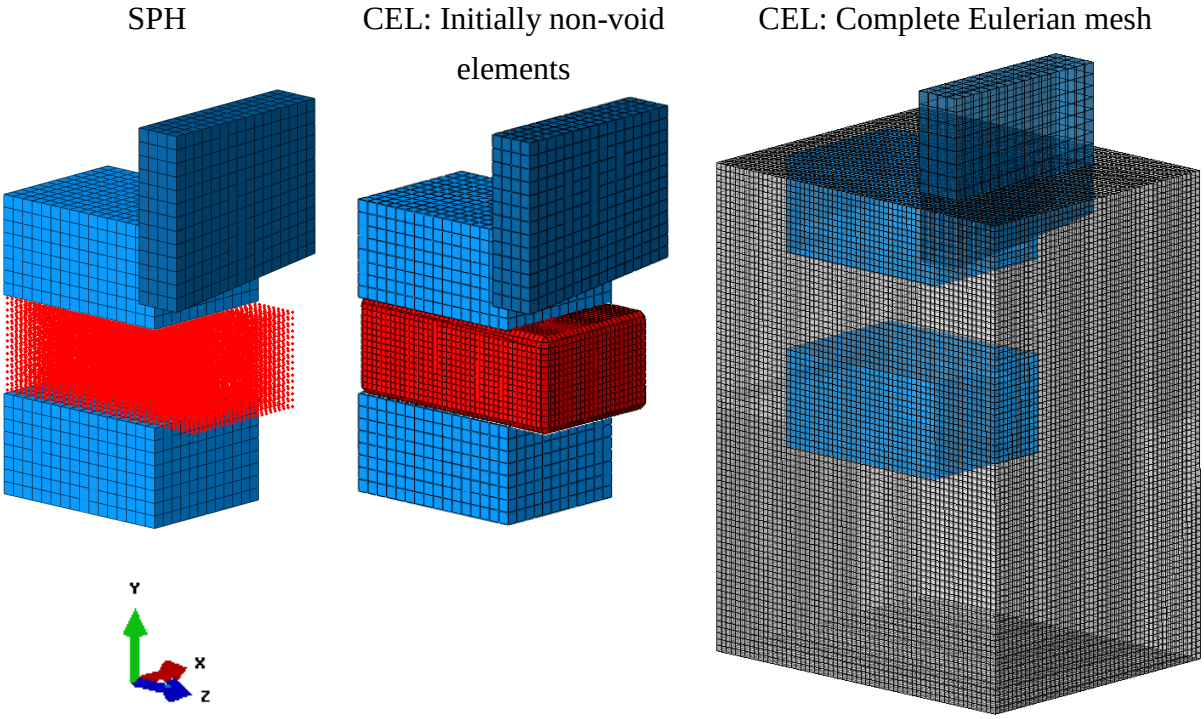



Figure 1: Simulation models: a) SPH, b) CEL: non-void region, b) CEL: Complete mesh

Both models contain 3904 linear Lagrangian Elements of type C3D8R. In the first model the metal is modelled by 15360 Smoothed Particles whereas in the second model the Eulerian domain consists of 230400 elements of type EC3D8R.

Starting with a convergence study with different mesh refinements, it turned out that a reliable distribution of contact forces between sheet metal and cutting blade are obtained with a ratio of 1:2 concerning the element size of the cutting tool and the sheet metal.

It is assumed that the clamping and cutting tools are much harder that the work piece and perform only small deformations during the cutting process. Therefore the tools are modeled as linear elastic material, characterized by Young's modulus $E$=210000 N/mm² and Poisson's ratio ν = 0.3. The same model is used to model the sheet's material in the linear elastic range. After exceeding the yield stress, plastic strains occur. In the following we assume isotropic hardening according to v.Mises yield function. In the first step, a very simple flow curve as shown in **Figure 2** is taken into account. The figure shows the Cauchy stress as a function of the plastic strain (Hencky strain).

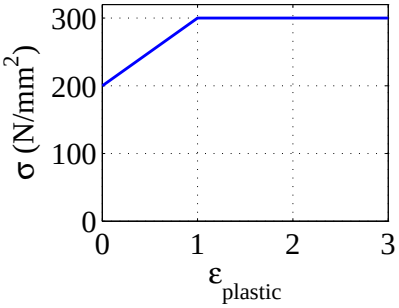



Figure 2: Stress-strain relationship in the applied model of isotropic hardening

For the chosen mesh refinement the simulation time for both models is approximatively equal. However, in case of a 20 % finer mesh of all components, the computation time of the CEL model becomes more than twice compared to the SPH model.

## 4 EVALUATION OF THE SIMULATION RESULTS

First the stresses in the chip as well as the shape of the chips will be regarded. Afterwards the stress distribution in the cutting tools will be examined. At last the cutting forces are discussed and qualitatively compared to experimental results.

When comparing the distribution of the v.Mises stresses in the sheet metal during the cutting process there is a high level of consistency between the two models. This is exemplified for the beginning of the cutting at the time 3.95 ms in Figure 3.

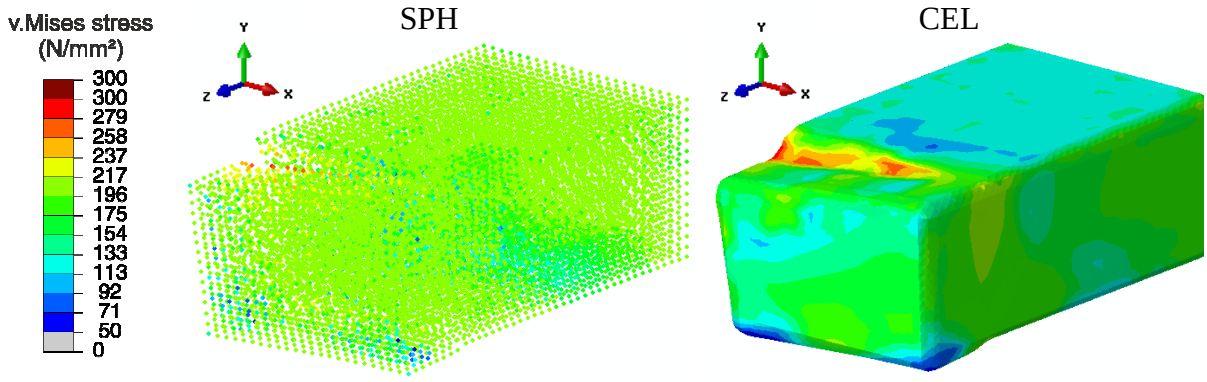



Figure 3: Distribution of v.Mises stresses in the sheet metal at the beginning of the cutting process

Figure 4 shows the shape of the chip at time 9.79 ms. At that time the sheet metal has already been highly deformed, i.e. almost cut off. On the one hand there is a high resemblance concerning the shape as well as the size of the flakes in the two models. On the other hand there are qualitative differences in the region of the cutting surface. In the model that uses SPH there is contact between the sheet metal and the blade almost across the whole cutting surface. In contrast in the CEL model the cutting surface touches the blade only in a small region at the bottom. This effect is a result of the Eulerian formulation: The Eulerian elements interacting with the blade are partially void.

It should be mentioned that in both models the contour plots of the metal sheets provide less significant results than the ones of the cutting tools which are modeled by Lagrangian Finite elements. Because of the discrete arrangement of the Smoothed Particles they do not provide exact information of the free boundaries of the material.

Within the CEL method the free boundaries of the material-filled regions in the Eulerian mesh are estimated based on the material volume fractions [3], which is the ratio of void and filled material within an element.

In the next step the distribution of the v.Mises stresses in the cutting tools is compared. Figure 4 shows a high level of consistency in the distribution of the stresses. However there are higher magnitudes in the v.Mises stresses in the model with CEL elements.

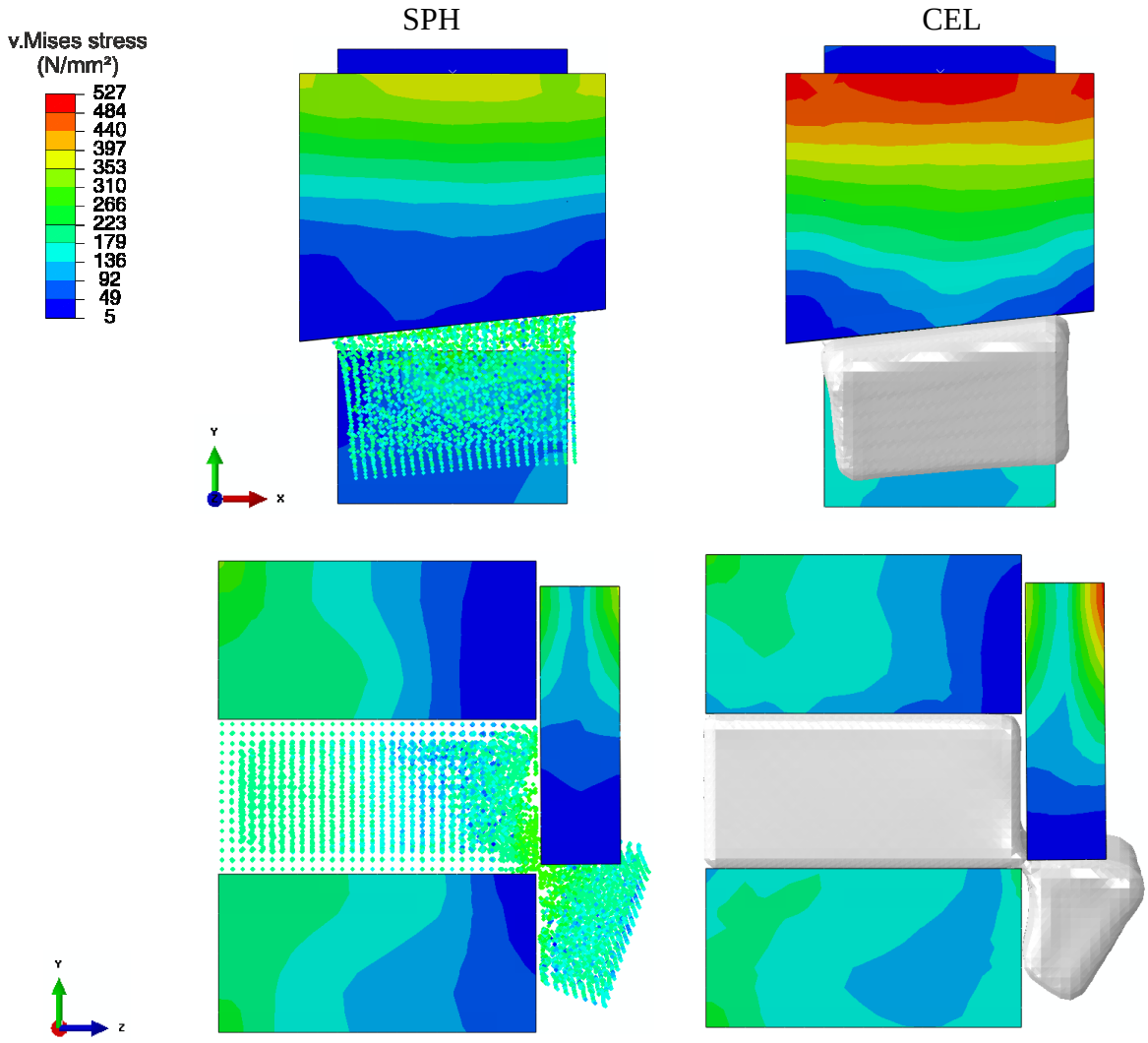



Figure 4: Distribution of the v.Mises stresses in the cutting tools and shape of the chip

This is also reflected by the magnitude of the contact pressure on the blade shown in Figure 5 at time 3.08 ms. It is higher in the CEL model than in the SPH model. The distribution of the contact pressure especially on the upper clamping tool shows high similarity for both methods.

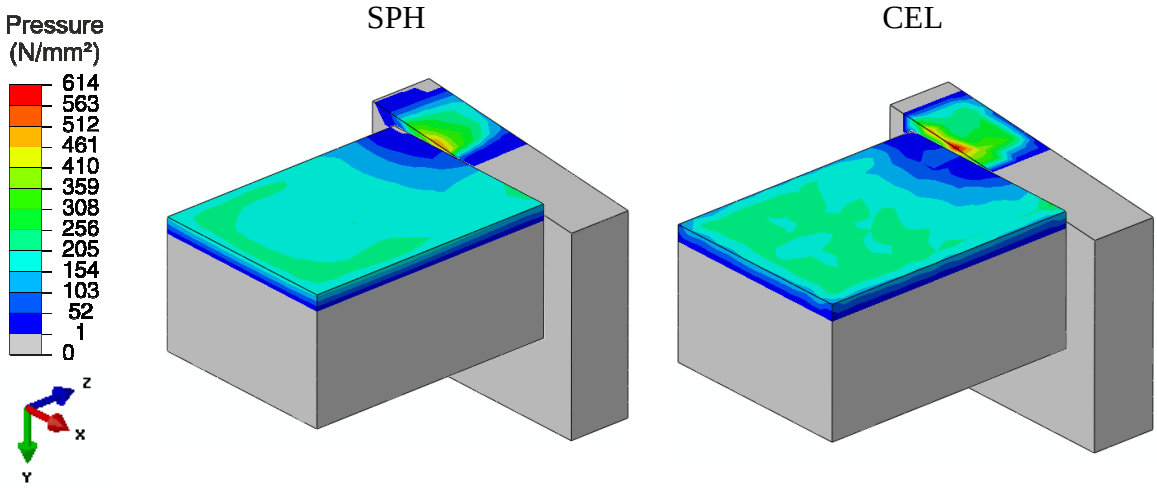



Figure 5: Contact pressure on blade and upper tool

Figure 6 shows the resulting forces that occur at the clamping of the blade in the cutting direction. For both methods the beginning of the cut is characterized by a steep rise of the blade forces. The maximum force in the CEL solution is approximatively 25 % higher than the one of the SPH method. The material is cut off after a time of 12.4 ms. After that point of time higher friction forces at the cutting surface can be found in the case of SPH than in the case of CEL which is resembled in the blade forces.

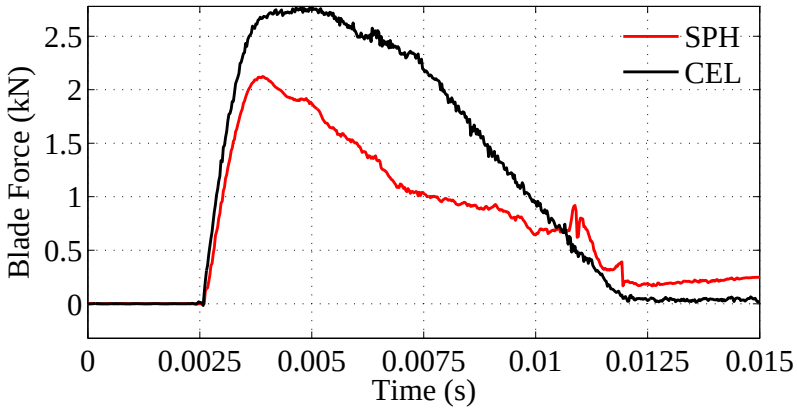



Figure 6: Blade forces in vertical direction

## 4.1 Comparison with measurements

Finally, the simulation results are compared to measurements performed on a Salvagnini automatic panel bender. A direct comparison of forces is not possible yet because of differing configurations in the simulation model and the experiment concerning dimensions of the sheet, tools and trajectories. It is the planned next step to adapt the simulation model to the configuration of the real machine. However, a qualitative comparison is already possible.

Figure 7 shows the measurement result for the resulting vertical force related to the maxi-

mum admissible force of the actuators. Three materials are investigated: two kinds of stainless steel with thickness 1.5 mm and almost the same nominal tensile strength $R_{m,}$ as well as mild steel with thickness 2 mm. The length of the sheets has been 260 mm. Figure 7 shows an increase of the force up to a maximum value. After reaching the first cut through the sheet on the first boundary, the force starts to decrease. Finally, the force drops rapidly when the cut is finished. It has turned out that for the two sheets of stainless steel with almost the same nominal tensile strength the measured force differs of about 15%.

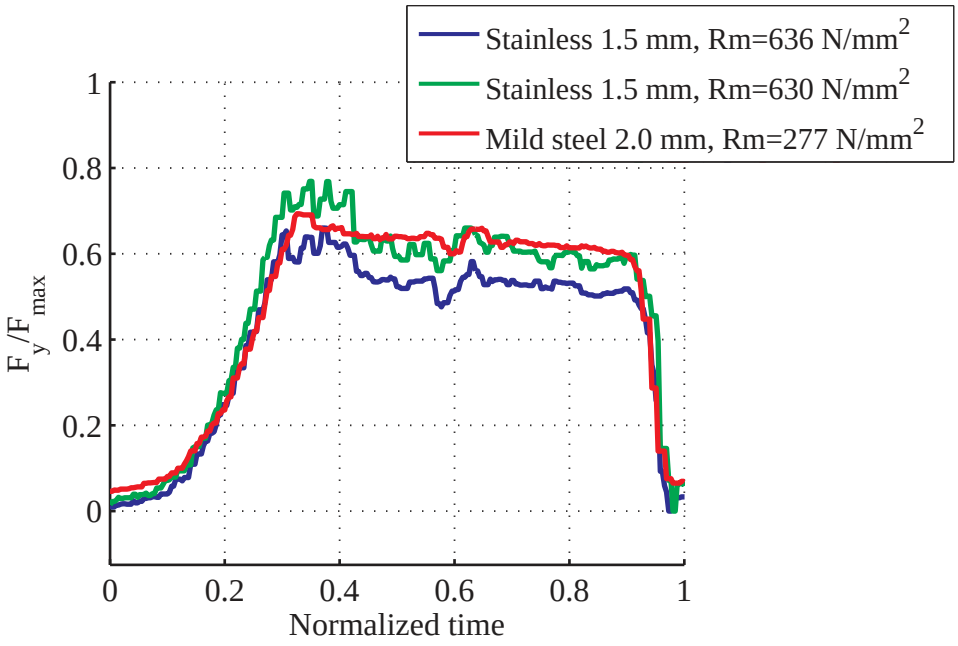



Figure 7: Measurement of resulting force in vertical direction

As mentioned above, only a qualitative comparison with the simulation results in Figure 6 is possible, because of different configurations of simulation model and experimental setup:

- Geometrical dimensions: Differing sheet length and thickness, differing tool geometries. Moreover the construction of the machine is far more complex than the simple simulation model. Elastic deformations have an influence on the measured forces,
- Different constitutive behavior: In the simulation models very a simple material behavior has been assumed as shown in Figure 2. On the other hand on the machine real materials with exponential flow curves have been cut. Thus, the decrease of the force after reaching the maximum value is less in the experiment than in the simulation.

The comparison shows the following qualitative coincidence of simulation and measurement:

- An increase of the force up to a maximum value
- A moderate decrease of the force after the first cut through the sheet.
- A rapid decrease at the end of the cut: This behavior is shown only in the SPH solution.
- The maximum forces in the two simulation models are differing of about 25%. In contrast on the machine differences of 15% for almost equal materials have been measured. Therefore the differences of the two simulation methods are in the same order of magnitude.

## 5 CONCLUSIONS

Two simulation methods have been compared for modelling cutting processes, i.e. CEL and SPH. Both methods are suitable to represent the main effects. Concerning the geometry of the cutting surface and the stress distribution in the cut material the methods show a good correspondence. Due to the differences of the formulations, the resulting forces do differ of about 25%. On the other hand, deviations in the same order of magnitude have been noticed in the experiment, comparing two kinds of stainless steel with almost the same tensile strength. The effect of rapid decrease of the resulting force at the end of the cut is only obtained in the SPH solution. The next planned steps are to adapt the simulation model to the experiment (geometry and constitutive relations) and to calibrate the parameters by measurements.

## ACKNOWLEDGMENT

This work has been supported by the Austrian COMETK2 program of the Linz Center of Mechatronics (LCM), and was funded by the Austrian federal government and the federal state of Upper Austria.

## REFERENCES

[1] D.J. Benson, Computational Methods in Lagrangian and Eulerian Hydrocodes, *Comput. Methods Appl. Mech. Engrg.*, **99,** 235–394, 1992.

[2] J.J. Monaghan, Smoothed particle hydrodynamics, *Rep. Prog. Phys.* **68,** 1703–1759, 2005.

[3] Abaqus 6.14 Documentation, Dassault Systèmes, 2015

[4] M. Kolbe, W. Hellwig, *Spanlose Fertigung Stanzen*, 11. Auflage, Springer-Vieweg Wiesbaden 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SIMULATION OF PRECIPITATION IN V-CONTAINING HSLA STEEL FOR THE STRENGTHENING ENHANCEMENT

**Piyada Suwanpinij*[1], Paolo Massaro[2], Annalisa Pola[2], and Prasonk Sricharoenchai[3]**

[1] The Sirindhorn International Thai-German Graduate School of Engineering (TGGS), King Mongkut's University of Technology North Bangkok (KMUTNB), 1518 Pracharat Sai 1, Bangkok, 10800
e-mail: piyada.s.mme@tggs-bangkok.org

[2] Department of Mechanical and Industrial Engineering, University of Brescia, Brescia, Italy
e-mail: p.massaro@studenti.unibs.it
e-mail: annalisa.pola@unibs.it

[3] Department of Metallurgical Engineering, Chulalongkorn University, 254 Phayathai Road, Bangkok 10330
e-mail: Prasonk.S@chula.ac.th

**Keywords:** Precipitation, High Strength Low Alloyed Steel, Hot Rolling, Vanadium

**Abstract.** *The precipitation of the microalloying elements in high strength low alloyed (HSLA) steel controls the strength of the steel greatly through grain refinement and particle hardening mechanisms. The current work simulates the precipitation of vanadium a hot rolling process for the optimized strengthening effect in a microalloyed steel. Taking into account the effect of deformation, cause of the drastic increase in the dislocation density, namely higher nucleation site density, it can be clearly seen that the precipitation of all species at dislocations, dominate the precipitation kinetics. The diffuse interface effect on the interfacial energy as well as a volumetric misfit of AlN at dislocations is also taken into account. The latter is because of its significant difference in the lattice parameter from the matrix. The presence of AlN at dislocations does not override that of V(C,N) as found in other cases with low density of dislocation. Slow cooling rate in the process ensures the consumption of the microalloying elements which in turn strengthen the product and minimise the production cost. The experimental verifications for the precipitates are performed by scanning transmission electron microscopy (STEM) as well as X-ray absorption spectroscopy (XAS) from synchrotron radiation.*

# 1 INTRODUCTION

This work focuses on the application of physically-based simulation approaches for a thermomechanical process of high strength low alloyed (HSLA) steel. The approaches are based on the classical nucleation and thermodynamic extremal principal or SFFK model for the growth and coarsening of precipitates, with thermodynamic and diffusion databases calculated by CALPHAD method.

To study the optimisation of the hot rolling process of V-microalloyed steel according to the precipitation hardening by vanadium, the effect of deformation and cooling rate before the coiling process in hot strip mill were studied by thermokinetics approach of precipitationcontained in the software MatCalc. It enables the possibilities to simulate the precipitation kinetics as a result of different processing parameters, i.e, thermomechanical cycles as well as the microstructure parameters such as grain size and equilibrium dislocation density. All changes in the thermomechanical cycle are able to be compiled in separated steps in the software.

Precipitates nucleate both at dislocations and grain boundaries. The increase in the dislocation density by deformation provokes much larger number of nuclei and the diffusion is immensely accelerated through pipe diffusion at the dislocation core. Long interpass times were allowed in the experiments to ensure high degree of recrystallisation of austenite after each rolling pass to allow repeated recrystallisation and consequently austenite grain refinement. The density of dislocations is updated along the simulated process.

Large retained strain together with slow cooling rate is expected to facilitate the precipitation of fine V(C,N) in ferrite and lead to higher strength by the advantage of particle hardening, which is strictly called dispersion hardening [1-2]. They impede the movement of dislocations by the well-known Orowan-Ashby looping mechanism.

# 2 THE RELATIONSHIP OF PLASTIC DEFORMATION AND DISLOCATION DENSITY

## 2.1 Balance of generation and annihilation of dislocations

The increase in the dislocation density under deformation can be described by the so-called one parameter model [3] as

$$\frac{d\rho}{dt} = \left( \frac{M\dot{\varepsilon}}{bL} \right), \tag{1}$$

where $M$ is the Taylor factor, which is 3.06 for FCC crystals [4], $\dot{\varepsilon}$ is the strain rate and $b$ is Buger's vector. The travelling distance $L$ is limited by the average spacing between dislocations as shown in

$$L = \frac{A}{\sqrt{\rho}}, \tag{2}$$

where $A$ is a materials constant in a range of 50 to 100.

The opposing dislocations with antiparallel Burger's vectors annihilate the formed dislocation and can be calculated by

$$\frac{d\rho}{dt} = -B \cdot 2 \frac{d_{ann}}{b} \rho M \dot{\varepsilon}\,, \tag{3}$$

where *B* is a constant related to the number of activated slip planes. The term $d_{ann}$ represents the critical distance, controlled by shear modulus of the matrix, Poisson's ratio, vacancy formation energy. Also, dislocation density is reduced by thermally activated dislocation climb [5], which is controlled by self-diffusion coefficient along dislocations or the so-called pipe-diffusion, $D_d$, [6] as represented in Eq.4.

$$\frac{d\rho}{dt} = -C \cdot 2 \cdot D_d \frac{Gb^3}{k_B T}\left(\rho^2 - \rho_{eq}^2\right), \tag{4}$$

where $k_B$ is the Bolzmann constant, T is temperature in Kevin, $\rho_{eq}$ is the equilibrium dislocation density and C is a constant, related to the solute trapping effect.

## 3 SIMULATION SETUP

The precipitation simulation was performed in MatCalc version 5.60. The chemical composition of the investigated steel is shown in Table 1. The content of titanium did not involve the simulation calculation but some amount of nitrogen was calculated according to mass balance of the formation of TiN, which usually forms during the solidification of steel melt and is generally not dissolved during the reheating. The thermomechanical cycle simulating a hot strip rolling is represented in Figure 1. It can be seen that the varying parameters in this work are the degree of deformation in the last rolling pass, $\varepsilon_4$, and the final cooling rate before the coiling process, $\dot{T}_{coil}$. They are taken into account in the simulation. The high strain rate helps produce a large number of dislocations. The experimental work is described in details earlier [7].

| Material | C | S | M | P | S | C | Mo | Al | T | V | N |
|---|---|---|---|---|---|---|---|---|---|---|---|
| V- HSLA | 0. | 0. | 0. | 0.0 | 0.0 | 0 | 0.0 | 0.0 | 0. | 0 | 0.0 |

Table 1: The chemical composition of the simulated steel, in mass%.

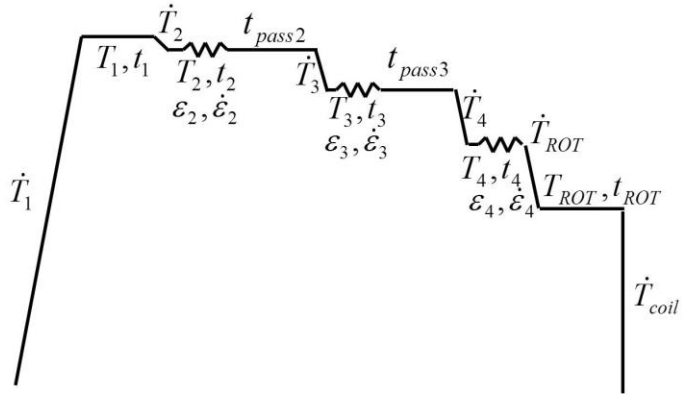



Figure 1: The thermomechanical cycles adopted in the simulation to simulate the precipitation.

| | |
|---|---|
| $\dot{T}_1, T_1, t_1$ | 200 °C/minute, 1250 °C, 15 minute |
| $\dot{T}_2, T_2, t_2$ | 5 °C/s, 1200 °C, 3 s |
| $\varepsilon_2, \dot{\varepsilon}_2, t_{pass2}$ | 0.2, 12 $s^{-1}$, 10 s |
| $\dot{T}_3, T_3, t_3$ | 60 °C/s, 1150 °C, 3 s |
| $\varepsilon_3, \dot{\varepsilon}_3, t_{pass3}$ | 0.4, 12 $s^{-1}$, 10 s |
| $\dot{T}_4, T_4, t_4$ | 60 °C/s, 900 °C, 3 s |
| $\varepsilon_4, \dot{\varepsilon}_4, t_{pass4}$ | 0.0/0.3/0.6, 12 $s^{-1}$, 0 s |
| $\dot{T}_{ROT}, T_{ROT}, t_{ROT}$ | 60 °C/s, 680 °C, 6 s |
| $\dot{T}_{coil}$ | 0.1/ 1/ 5/ 10/ 60 °C/s |

Table 2: The parameters in the thermomechanical cycle in Figure 1.

Only sharp transformation temperature can be set up in the calculation. Hence, the simulated temperature of Run Out Table (ROT), $T_{ROT}$, of 680 °C was chosen for the transformation from austenite into ferrite, as ferrite was only little to observed before reaching this isothermal period. The precipitation of AlN and V(C,N) is both at grain boundaries and dislocations in both austenite and ferrite parent phases, while that of cementite is at dislocations only. The grain sizes of 100 µm and 10 µm were selected for austenite and ferrite, respectively. The equilibrium dislocation density in austenite was assigned as $10^{11}$ $m^{-2}$ while that in ferrite as $10^{12}$ $m^{-2}$. The grain refinement in austenite due to recrystallisation during the hot rolling process was not considered. But the generation and annihilation of dislocation were considered by choosing one parameter model from Sherstnev and Kozesnik with the parameter A in Eq.2 of 50, B in Eq.3 of 2 and C in Eq.4 of $1\times10^{-4}$. The volumetric misfit of only AlN at dislocations was considered, which is 0.27. As the critical temperature is recommended to take into account for the calculation of interfacial energy due to the diffuse interface, a critical temperature of 1727 °C was selected for V(C,N) as an estimation from those reported for VC and VN in both austenite and ferrite [8].

## 4 MATERIALS CHARACTERISATION

Samples under different process parameters were measured for the amount of vanadium in different states by means of synchrotron X-ray Absorption Spectroscopy (XAS), combined with Linear Combination Fitting (LCF) as published earlier in [9]. Also, selected sample were investigated for different morphology and size of precipitates by Scanning Transmission Electron Microscopy (STEM).

## 5 SIMULATION RESULTS AND DISCUSSION

The multipass rolling thermomechanical cycles result in high dislocation density as shown in Figure 2. The first two peaks of dislocation density gradually drop after long interpass time, which allows the recovery during the long interpass time. One can see that without the retained strain in the last deformation step the dislocation density is annihilated to $3.79\times10^{12}$ $m^{-2}$. With the logarithm strain of 0.3 and 0.6, the dislocation density reaches $3.62\times10^{14}$ $m^{-2}$ and $4.59\times10^{14}$ $m^{-2}$ respectively,

interpreting that the retained strain therefore increases the dislocation density with 2 order of magnitude.

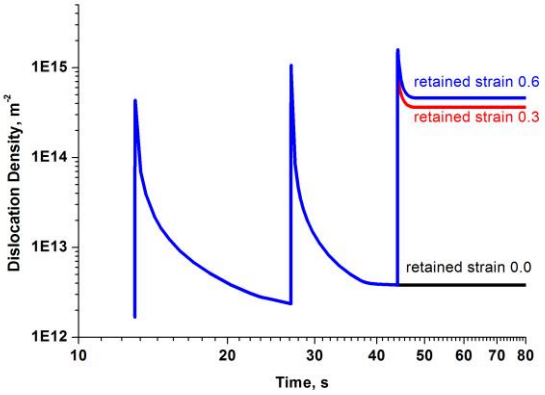
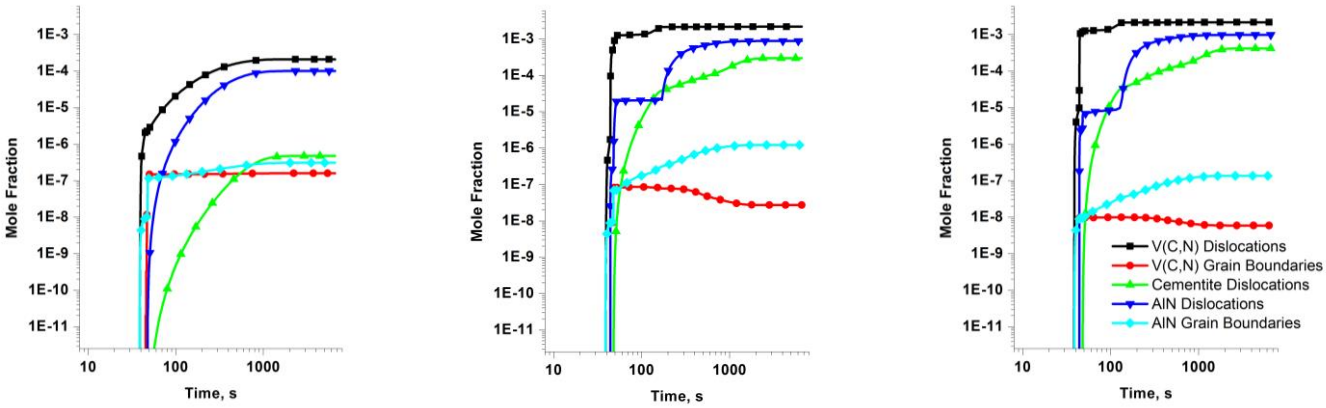


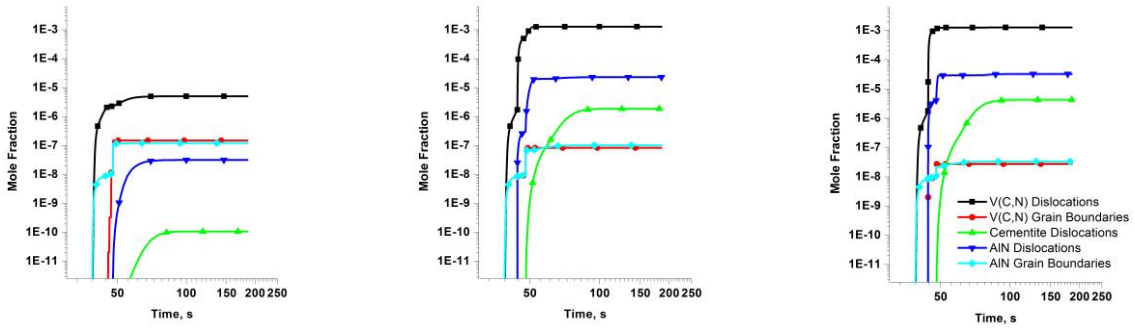

Figure 2: The change in dislocation density as a result of generation and annihilation after deformation.




a) b) c)

Figure 3: The mole fraction of different precipitates, both at dislocations and grain boundaries under a slow cooling rate of 0.1 °C/s with a retained strain of a) 0.0 b) 0.3 and c) 0.6.

Figure 3 sums and compares different precipitate phase fractions under different degree of deformation at a slow cooling rate of 0.1 °C/s. The high density of dislocations, main nucleation site, results in up to 3 order of magnitude higher fraction of the precipitates at this site. Also, higher dislocations by retained strain in the last deformation step results in the consumption of V(C,N) at the grain boundaries. This is difference from what that found by Radis et al. [10], whose case shows that the predominant site of AlN is only grain boundaries and the precipitation of AlN at dislocations overrides that of V(C,N). The first short period showing constant fraction of V(C,N) and AlN at dislocations was the precipitation in austenite matrix, before transformed into ferrite. It can be clearly seen that the higher diffusivity in ferrite allows abrupt increase in the precipitation fraction.

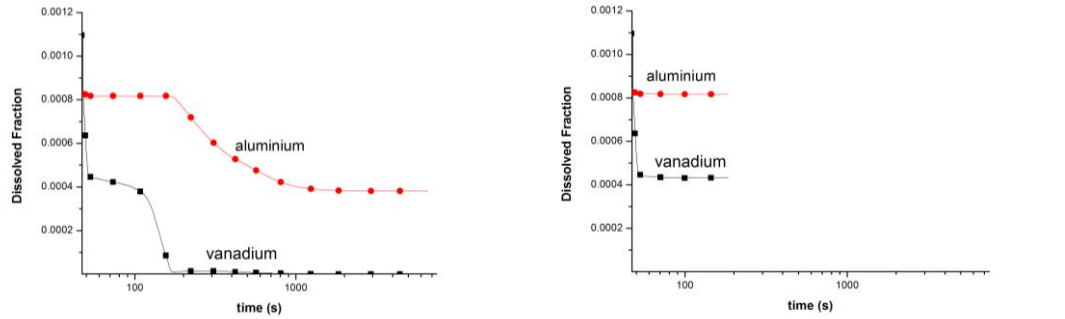



a) b) c)

Figure 4: The mole fraction of different precipitates, both at dislocations and grain boundaries under a moderate cooling rate of 5 °C/s with a retained strain of a) 0.0 b) 0.3 and c) 0.6.

When the cooling rate is faster such as 5 °C/s as shown in Figure 4, the pronounced precipitation due to deformation is more obvious, up to 4 orders of magnitude for cementite. Due to shorter precipitation time at higher cooling rates, deformation obviously has a vital role in accelerating the precipitation. Comparing V(C,N) at dislocations between the two cooling rates, under a fixed retained strain of 0.3, its fraction shows double amount at slower cooling rate, i.e., 0.00219, compared with 0.00128.

To illustrate the consumption of the microalloying elements in from the matrix into the precipitates, their concentrations are plotted in Figure 5. It is quite clear that higher cooling rate does not complete the available amount of vanadium and aluminium. Therefore, significant amount of these elements is still dissolved in ferrite matrix at the end. This implies ineffective process control of microalloyed steel.




a) b)

Figure 5: Decrease in the microalloying elements dissolved in ferrite matrix in case of a retained strain of 0.3 and a) a cooling rate of 0.1 °C/s and b) 5 °C /s. The plot starts from just before ferrite has formed at 680 °C in the simulation.

However, small inaccuracy might exist in these simulations due to the fact that only sharp transformation temperature from austenite to ferrite is able to assign. However, the trend of the influence of deformation and cooling rate from the simulation are in good agreement with the results from XAS as published in [9].

## 6 CONCLUSIONS

The density of dislocation is immensely multiplied by deformation to several orders of magnitude. This results in the increase in the number of nuclei of V(C,N) and AlN that their precipitation at dislocations dominates the kinetics. Slower cooling rate allows enough time for the growth of precipitation and consumption of the available elements. This is beneficial for the coiling process as the cooling rate is very slow. This work is an example of how a physically based simulation of precipitation can help the optimisation of thermomechanical process parameters, which is advantage for the strength improvement of the products.

## 7 ACKNOWLEDGEMENTS

The authors are grateful to the support of MatCalc Engineering GmbH team, Dr. Piotr Warczok, who gives always fast response and helpful suggestions during our long term working. Also, the financial support from project No. TRG5780214 granted by The Thailand Research Fund (TRF) is acknowledged.

## REFERENCES

[1] T. N. Baker: Processes, microstructure and properties of vanadium microalloyed steels, Materials Science and Technology 25 (2009), No. 9, pp. 1083 – 1107.

[2] A. J. Ardell: Metall. Trans. A, 1985, 16A, 2131 – 2165.

[3] U.F. Kocks: Laws of work-hardening and low-temperature creep. ASME Journal of Engineering Materials and Technology, 98 (1976), 76 - 83

[4] Y. Bergström: The plastic deformation of metals – a dislocation model and its applicability, Div. of Physical Metallurgy, Royal Institute of Technology, Stockholm, 1983.

[5] L.E. Lindgren, K. Domkin, S. Hansson: Dislocation, vacancies and solute diffusion in physical based model for AISI 316L, Mechanics of Materials, 40 (2008) 907 – 919.

[6] M.S. Soliman, E.A. El-Danaf, A.A. Almajid: Effect of heat treatment conditions on the high temperature deformation of 6082-Al alloy, Advanced Materials Research, 83 - 86 (2009) 407 – 414.

[7] P. Suwanpinij and H.H. Dickert: Application of Vanadium Precipitation for Lower Rolling Force and Enhanced Strength of Hot Strip Steels, Archives of Materials Science and Engineering, 70 (2014), No.1, pp.39 – 44.

[8] P. Warczok: Evaluation of Interfacial Engergies, Technical Paper on MatCalc #2014001.

[9] P. Suwanpinij, H.H. Dickert, N. Thammajak and P. Srichareonchai: Effect of Cooling Rate during Coiling in Strengthening Mechanisms of Vanadium HSLA Steel and Its Quantitative Measurement of Precipitation by Synchrotron XAS, Materials Testing 58 (2016), No.1, pp.5 - 11.

[10] Radis, R. and Kozeschnik, E. (2010), Concurrent Precipitation of AlN and VN in Microalloyed Steel. steel research int., 81: 681 – 685.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL ANALYSIS OF SANDWICH PANELS SUBJECTED TO POINT LOADS

**Zbigniew Pozorski**[1], **Jolanta Pozorska**[2]

[1] Institute of Structural Engineering, Poznan University of Technology
Piotrowo 5, 60-965 Poznań, Poland
e-mail: zbigniew.pozorski@put.poznan.pl

[2] Institute of Mathematics, Czestochowa University of Technology
Armii Krajowej 21, 42-200 Częstochowa, Poland
e-mail: jolanta.pozorska@im.pcz.pl

**Keywords:** Sandwich Panels, Soft Core, Boundary Conditions, Local Instability, Indentation.

**Abstract.** *The paper concerns the problem of static structural behavior of sandwich panels. The panels are made of thin steel sheets and a thick, but soft core. The effects of the impact of concentrated loads on the structural behavior of the panels are discussed. Several different cases of concentrated loads were considered: load acting in the plane of the panel, load perpendicular to the panel and the additional action of concentrated bending moment. The loads are transferred to the panel by a small steel slab. The slab is attached to the facing of the panel. Described load cases correspond to situations encountered in engineering practice. In order to analyze the issue, a 3-D numerical model was used. Between the facing and the deformable core an interface layer was introduced, wherein the criterion of damage initiation and damage propagation were defined. Different failure modes were taken into account. Due to the character of the problem, the geometrically nonlinear analysis and Riks' method were applied. For each load case a failure load and extreme stresses leading to failure were determined. The differences in the obtained results were explained and the appropriate failure mechanisms were discussed. Analyzed sandwich panel has fixed geometries, however, made observations are universal. The obtained results tend to look more carefully at the problem of the impact of concentrated forces. This problem is particularly important in the case of sandwich panels, which due to its construction are very susceptible to local influence.*

## 1 INTRODUCTION

The paper concerns sandwich panels made of thin steel sheets and a thick, but soft core. The most commonly used are polyurethane foam cores. The sandwich elements are very attractive because of high load-bearing capacity at low self-weight and excellent thermal insulation. Therefore, these elements are used in the space industry, aerospace, shipbuilding, and automotive industries. Currently, it is also difficult to imagine civil engineering without the sandwich panels that are used as a building envelope (walls, roofs), partition walls and suspended ceilings.

The structural behavior of sandwich structural elements with uniform support and load conditions has been repeatedly studied and described [1, 2]. The analysis of systems subjected to a concentrated load is much more difficult. The existence of deformable core leads to sensitivity of sandwich panels with respect to localized effects [3]. One of these effects is wrinkling, which is a form of local instability of the sandwich facing. Wrinkling failures of sandwich columns under compression, beams in three- and four-point bending and cantilever beams under end loading were investigated in [4]. Analytical, numerical and experimental approaches to the problem of wrinkling in the case of nonlinear materials were presented in [5, 6]. Another form of damage is debonding. It occurs for example in the case of tensile forces between a core and a facing. The inter-layer failure modes are very often initiated by an impact [7, 8, 9]. An important failure mode is indentation. The failure is caused by the crushing of the core under a localized force. The problem of indentation loads was investigated in [10]. Recently, in order to better describe this phenomenon, sophisticated core models are applied [11]. Another, experimental approach was presented in [12].

In practice, the most common are three cases of concentrated actions: the impact of devices on the roof panel, interactions at the points of attachment of a sandwich panel to the supporting structure and the impact of equipment (e.g. advertising banners) on the wall panel. In the latter case, the load in the plane of the sandwich plate usually dominates, although it is almost always in combination with loads perpendicular to the plate. It should be noted that in the case of concentrated loads, we have to deal with the simultaneous occurrence of global and local phenomena (debonding, indentation, wrinkling). The difficulty in analyzing systems subjected to point loads stems not only from the nature of the load, but also its location. Structural analysis, in the case of asymmetric load conditions, requires the use of at least 2-D models [13]. This is very important because the sandwich systems are characterized by a material anisotropy deepened by different geometries of each layer. In this paper the described phenomena are analyzed using a 3-D model [14]. This article is focused on the case of point loads applied to the wall panel. Most attention was paid to the identification of failure mechanisms.

## 2 DESCRIPTION OF THE PROBLEM

Consider a wall sandwich panel that can be installed vertically or horizontally. For the brevity of presentation, a one-span panel is analyzed. The supporting structure is located inside the building. One can say that the panel is simply supported on two opposite edges, although a 3-D model will be considered, in which the depth of the panel has an impact on the description of the boundary conditions. Assume that the external loads are localized in the middle of the sandwich panel and are applied to the external facing of the panel (Fig. 1). The load conditions should simulate the impact of equipment (e.g. advertising banners) on the wall panel. Taking into account the expectations of the market, with full awareness it is assumed that the load will be applied only to the external facing, and there will be no additional elements to ensure cooperation between the two facings of the panel. This corresponds to the situation when a piece of equipment is attached to the external facing by glue or screws only.

In fact, there are not strictly 'point' forces, but rather distributed on a small surface. Therefore, it was assumed that the defined forces will be transferred to the sandwich panel by a square steel slab with a side length $a$ and a thickness $t$ ($a$=0.08 m, $t$=0.008 m).

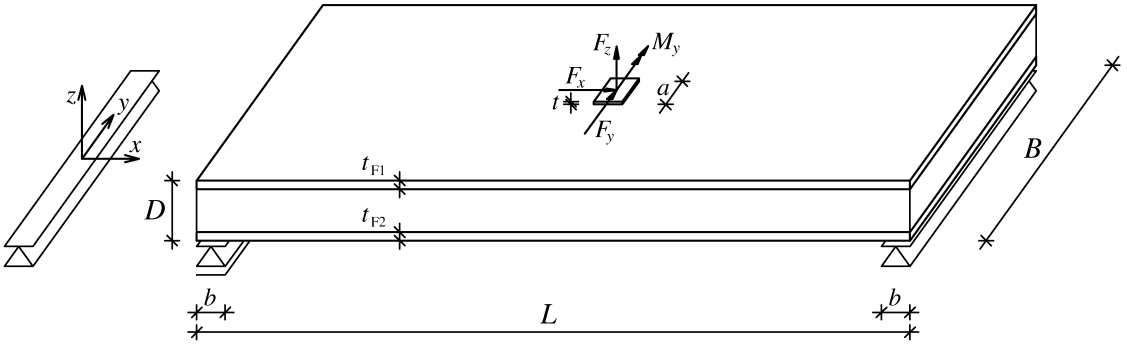



Figure 1: The 3-D model for the analysis of sandwich panels subjected to concentrated loads.

In general, on the upper surface of the steel slab, at its center, there may occur 3 forces and 3 concentrated moments. The paper examined the 6 load cases:
a) there is only $F_x$ (the name of the model: *eco_21*),
b) there is only $F_y$ (*eco_21h*),
c) the increasing of $F_x$ accompanied by the constant compressive load $F_z$ = –1 kN (*eco_21n*),
d) the increasing of $F_x$ accompanied by the increasing of $M_y = 0.1F_x$ (*eco_22*),
e) there is only the tensile force $F_z$ (*eco_24_ten*),
f) there is only the compressive force $F_z$ (*eco_25_com*).
The above load cases correspond to the direct attachment of the loading element to the sandwich panel arranged vertically (*eco_21*) and horizontally (*eco_21h*). The additional localized effect of wind pressure was considered in model *eco_21n*. Model *eco_22* corresponds to the case of weight mounted on the arm 0.1 m. Cases of the load acting perpendicular to the sandwich panel (*eco_24_ten, eco_25_com*) were prepared for comparison with the other cases. Loads perpendicular to the sandwich panel simulate (to some extent) concentrated impacts on the ceiling. In all cases they were considered to be only static loads.

The aim of the paper is to identify the phenomena occurring during the action of concentrated loads on the sandwich panel. The key issue is to determine the failure mechanisms and estimate the values of load capacity. The problem of distribution of applied load is equally interesting.

## 3 THE NUMERICAL MODEL

The 3-D numerical model was prepared in the ABAQUS system. The parameters of the system correspond to the values determined in laboratory tests. A sandwich panel with a length 5.00 m is located on two supports with a width of 0.10 m. The width of the panel is $B$ = 1.0 m. The total depth of the panel is 98.43 mm. The thickness of each of the faces is $t_{F1} = t_{F2}$ = 0.471 mm. The modulus of elasticity of the facing material is $E_F$ = 195 GPa and Poisson's ratio equals $\nu_F$ = 0.3. The actual relationship between stress and strain was introduced. The yield strength was 360 MPa, and the ultimate strength reached 436 MPa. Facings were modeled using a four node doubly curved, thin or thick shell, finite membrane strain elements S4. At this stage of the study, the core was considered as isotropic material with a modulus of elasticity and $E_C$ = 8.61 MPa and Poisson's ratio $\nu_C$ = 0.02. As a result, shear modulus is $G_C$ = 4.22 MPa. The core was modeled using eight node brick elements C3D8.

To assess the cause of the failure of the sandwich panel, between the steel facing and the soft core, an interface layer of a thickness of 0.5 mm was introduced. The interface was modeled using COH3D8, 8-node 3-D cohesive elements. The following uncoupled elasticity law for cohesive material was used:

$$\begin{bmatrix} t_n \\ t_s \\ t_t \end{bmatrix} = \begin{bmatrix} K_{nn} & 0 & 0 \\ 0 & K_{ss} & 0 \\ 0 & 0 & K_{tt} \end{bmatrix} \begin{bmatrix} \varepsilon_n \\ \varepsilon_s \\ \varepsilon_t \end{bmatrix} \tag{1}$$

where $t_n$ is normal traction (stress) and $t_s$, $t_t$ are shear tractions. Corresponding nominal strains are defined as $\varepsilon_n = \delta_n/T_0$, $\varepsilon_s = \delta_s/T_0$, $\varepsilon_t = \delta_t/T_0$ using separation $\delta$ and constitutive thickness of cohesive element $T_0$. The failure initiation was conditioned by a stress state. The quadratic nominal stress criteria of damage initiation and linear softening damage evolution were used. The damage initiation criterion has the form of a quadratic nominal stress function. The damage is initiated when the function reaches a value of one:

$$\left\{\frac{\langle t_n \rangle}{t_n^0}\right\}^2 + \left\{\frac{t_s}{t_s^0}\right\}^2 + \left\{\frac{t_t}{t_t^0}\right\}^2 = 1 \tag{2}$$

where the typical notation of Macaulay brackets is used. The following parameters of the interface were used: $K_{nn}$ = 8.61 MPa, $K_{ss} = K_{ss}$ = 4.22 MPa, $t_n^0$ = 123 kPa, $t_s^0 = t_t^0$ = 112 kPa.

The point loads were applied in the middle of the external surface of the transferring steel slab. The mesh size was equal to 0.02 m over almost the entire FEM model. In the vicinity of the force application, the mesh was condensed to 0.0067 m. Interaction between all parts was assumed as the TIE type, which makes equal displacements of nodes. It is certainly not a perfect type of connection, however, it is sufficient to evaluate the most important phenomena. The supports were modeled using analytical rigid elements. Boundary conditions at reference points ensure the freedom of rotation around the *y*-axis. One of the supports is free to shift along the *x*-axis. Because the problem was expected to exhibit significant local effects, a geometrically nonlinear static analysis and the Riks' method were applied.

## 4 DISCUSSION OF THE RESULTS

For each static scheme was obtained the force causing local damage to the interface. In every case, the failure results from the tensile stress $\sigma_{zz}$ in combination with shear stresses $\tau_{xz}$ i $\tau_{yz}$. The respective values are given in Table 1.

| Model | Failure load | Extreme normal stress $\sigma_{zz}$ [kPa] | Extreme shear stress $\tau_{xz}$ [kPa] | Extreme shear stress $\tau_{yz}$ [kPa] |
|---|---|---|---|---|
| *eco_21* | $F_x$ = 2.22 kN | +119.9 / –138.2 | +20.66 / –24.93 | +27.13 / –27.13 |
| *eco_21h* | $F_x$ = 2.21 kN | +119.0 / –136.5 | +20.71 / –20.71 | +20.32 / –24.81 |
| *eco_21n* | $F_x$ = 4.28 kN with $F_z$ = –1 kN | +81.19 / –530.1 | +42.32 / –106.2 | +96.67 / –96.67 |
| *eco_22* | $F_x$ = 0.164 kN with $M_y$ = 0.0164 kNm | +119.9 / –135.6 | +21.43 / –22.96 | +26.85 / –26.85 |
| *eco_24_ten* | $F_z$ = 0.670 kN | +115.3 / –0.934 | +33.99 / –33.99 | +34.46 / –34.46 |
| *eco_25_com* | $F_z$ = –1.95 kN | +17.86 / –431.6 | +106.0 / –106.0 | +96.03 / –96.03 |

Table 1: Failure load and extreme stresses in the interface layer.

Failure load is identical in the case of a load applied along the panel (*eco_21*) and crosswise the panel (*eco_21h*). This is despite the fact that in the latter case, the panel rotates about the axis *x*. The displacements are small because they do not exceed 1 mm. Applying an excitation force to the steel slab causes that the force acts on an arm in relation to the plane of the supports. As a result, the steel slab indents the sandwich core on the one edge and detaches the facing from the core on the other edge. This leads to the destruction of the interface. Interestingly, these phenomena occur only locally, i.e. in the immediate vicinity of the applied load (Fig. 2).

Also note that the small additional load compressing the panel (*eco_21n*) increases the load capacity, since it counteracts the detachment of the facing from the core. In the model *eco_22* even a slight bending moment results in a drastic decrease of the maximum force $F_x$. This leads to the conclusion that the attachment of any elements to sandwich panels should be avoided, if the attachment is realized only to one facing.

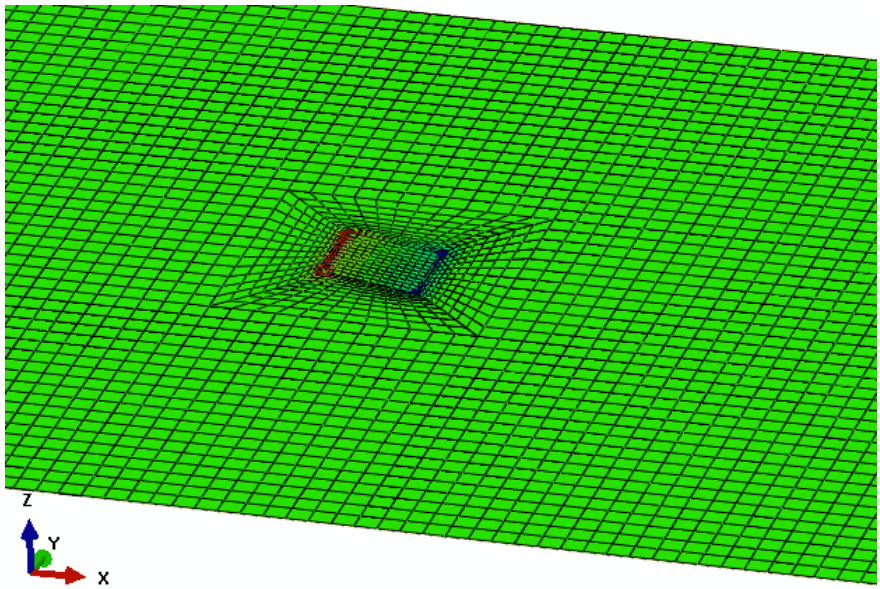



Figure 2: Distribution of the stress $\sigma_{zz}$ in the interface layer of the model *eco_21*. Blue indicates compressive stress and red indicates tensile stress.

In the case of tension load directed perpendicularly to the facing (*eco_24_ten*), the maximum force reached the expected level, although somewhat surprising is the low redistribution of the applied load. In the case of compressive load (*eco_25_com*) there first appears the effect of plasticity of the steel facing, and later shear of the interface. Interestingly, there was no wrinkling of the steel facing, although the numerical model is suitable for the analysis of such phenomena.

## 5 CONCLUSIONS

Conducted numerical analysis shows that the problem of concentrated loads is very important. For each load case, local damage occurred between the facing and the core. Failure loads were low in value. A particularly unfavorable case is the load acting on an arm. Even a small bending moment causes locally high stresses with different signs. According to the authors of the paper, in the case of the bending moment, it is necessary to ensure close cooperation between the two facings of the sandwich panel. In the case of concentrated forces acting in the plane of one facing or perpendicular to the facing, the small load-bearing capacity of the panel should be considered. Local interface failure can lead to further damage propagation.

The discussed phenomenon, in particular the lack of significant redistribution of applied load, puts a new light on the classical approach to the problem of concentrated loads. Assessment of the effects of concentrated loads cannot be limited to the global effects, but must also take into account phenomena occurring locally.

## REFERENCES


[1] J.M. Davies, *Lightweight sandwich construction*, Blackwell Science, Oxford, 2001.

[2] D. Zenkert, *An introduction to sandwich construction*, Engineering Materials Advisory Services, 1995.

[3] O.T. Thomsen, Theoretical and experimental investigation of local bending effects in sandwich plates. *Composite Structures*, 30(1), 85-101, 1995.

[4] E.E. Gdoutos, I.M. Daniel, K.-A. Wang, Compression facing wrinkling of composite sandwich structures, *Mechanics of Materials*, 35(3-6), 511-522, 2003.

[5] V. Koissin, A. Shipsha, V. Skvortsov, Effect of physical nonlinearity on local buckling in sandwich beams, *Journal of Sandwich Structures and Materials*, 12, 477-494, 2010.

[6] M.A. Stiftinger, F.G. Rammerstorfer, Face layer wrinkling in sandwich shells – theoretical and experimental investigations, *Thin-walled Structures*, 29(1-4), 113-127, 1997.

[7] T.C. Triantafillou, L.J. Gibson, Debonding in foam core sandwich panels, *Materials and Structures*, 15, 266-276, 1989.

[8] L.R. Xu, A.J. Rosakis, Impact failure characteristics in sandwich structures, Part I: Basic failure mode selection, *International Journal of Solids and Structures*, 39, 4215-4235, 2002.

[9] D. Zenkert, Strength of sandwich beams with interfacial debondings, *Journal of Composite Structures*, 17, 331-350, 1991.

[10] E. Lolive, J-M. Berthelot, Non-linear behaviour of foam cores and sandwich materials, Part 2: indentation and three-point bending, *Journal of Sandwich Structures and Materials*, 4(4), 297-352, 2002.

[11] C.A. Steeves, N.A. Fleck, Collapse mechanisms of sandwich beams with composite faces and a foam core, loaded in three-point bending, Part I: analytical models and minimum weight design, *International Journal of Mechanical Sciences*, 46, 561-583, 2004.

[12] E.A. Flores-Johnson, Q.M. Li, Experimental study of the indentation of sandwich panels with carbon fibre-reinforced polymer face sheets and polymeric foam core, *Composites Part B: Engineering*, 42(5), 1212-1219, 2011.

[13] R. Studziński, Z. Pozorski, A. Garstecki, Structural behavior of sandwich panels with asymmetrical boundary conditions, *Journal of Constructional Steel Research*, 104, 227-234, 2015.

[14] J. Pozorska, Z. Pozorski, The numerical model of sandwich panels used for specifying wrinkling stress, *Journal of Applied Mathematics and Computational Mechanics*, 13(3), 187-194, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# BOND OF REINFORCEMENT IN REACTIVE POWDER CONCRETE: EXPERIMENTAL STUDY

**Mingde Sun[1], Ri Gao[1], Aili Li [1] and Yongjing Wang[1]**

[1] School of Civil Engineering, Beijing Jiaotong University, Beijing 100044, China
E-mail: mingdesun@163.com , rigao@bjtu.edu.cn, liaili2016@outlook.com, wyjbjtu@163.com

**Keywords:** Reactive Powder Concrete; Steel rebar; Bond performance; Anchorage length;

**Abstract:** *This study investigates the bond performance of steel rebar embedded in reactive powder concrete(RPC). The assessment of the bond behavior at concrete ages of 7, 14 and 28 days using pull-out specimens was the subject of the investigation. Varying parameters in this research program were embedment length, rebar diameter, concrete age and size of the concrete cover. The concrete compressive and flexural strength were measured at the same test ages. The test results show that with the increase of rebar embedment length and rebar diameter, the ultimate bond stress decreases. This decrease was caused by the nonlinear distribution of the bond stress, voids on the contact surface and the Poisson's ratio effect. When protection layer thickness and standard curing age of reactive powder concrete increases, the ultimate bond stress increase. The bond of reinforcement in RPC is characterized by very high maximum bond stresses, even at early ages. This paper presents the details of the investigation and discusses the results obtained.*

## 1. INTRODUCTION

Reactive powder concrete (RPC) is one of the newest development in concrete building materials. By adding steel fibers to RPC and its homogenized microstructure, it has many demonstrated advantages such as outstanding mechanical properties, ductility, and durability [1, 2, 3, 4]. In particular, its unique tensile performance, along with its high strength and strain capacity make RPC is very useful in bending structural applications [5, 6].

One of the important aspects of the structural behavior of reinforced reactive powder concrete elements is the development of an appropriate bond capacity. The structural will premature failure for the insufficient bond resistance results. So it is very useful to accurate predict the relative slip between the rebar and concrete in the design and analysis work.

Experimental work had been performed in order to examine the bond of steel bars embedded in RPC [5, 6 ,7, 8, 9]. The main objective was to investigate the bond strength activated in pullout tests and to determine the relationship of bond-slip. Pullout tests were performed to examine the effect of varying parameters such as embedment length, bar diameter, concrete age and concrete cover.

## 2. EXPERIMENTAL PROGRAM

### 2.1. Material properties and mix proportion

The mix proportion of RPC used in this study is summarized by the weight in Table 1. The water-to-binder ratio (W/B) used in the test was 0.16. The micro steel fibers having a length of 12 mm to 15mm, a diameter of 0.22mm and the tensile strength is greater than 2800MPa were incorporated by 2% of the total volume. Portland cement and silica fume were used as cementitious materials. Sand with a grain size of 0.16mm to 1.25mm was used as fine aggregate. In addition, the polycarboxylate superplasticizer was added to satisfy the workability, the flow was measured by 270 mm from the flow test in accordance with GB/T 14902-2012 [10].

After casting, the test specimens were placed in the standard curing room and they were cured for 24h before demolding. After demolding, the test specimens were placed in the standard curing room again until reaching the testing concrete age.

Table 1: Mixing ratio of RPC ($kg/m^3$)

| Cement | Silica fume | Sand | Steel fiber | Superplasticizer | Water |
|---|---|---|---|---|---|
| 920 | 170 | 1080 | 120 | 22 | 174 |

### 2.2. Mechanical tests

The compressive strength and flexural strength of RPC is the basic mechanical parameters of the analysis and design of RPC structure. The flexure, compression specimens were cast and cured under the same conditions as the Pull-out specimens. Three prism specimens with a dimension of 100mm×100mm×300mm were fabricated and tested for evaluating the compressive behavior in accordance with GB/T 50081-2002 [11]. Compressive strength was measured using universal testing machine (WAW) with a maximum load capacity of 2000kN. To obtain the strain along with the stress, three electronic strain gauges were also installed on the prism specimens. According to the Chinese standard GB/T 50081-2002 [11], a four-point bending test was performed to investigate the flexural behavior. The compressive and flexural strength were measured after 7, 14 and 28 days. The properties of the steel rebars are summarized in Table 2.

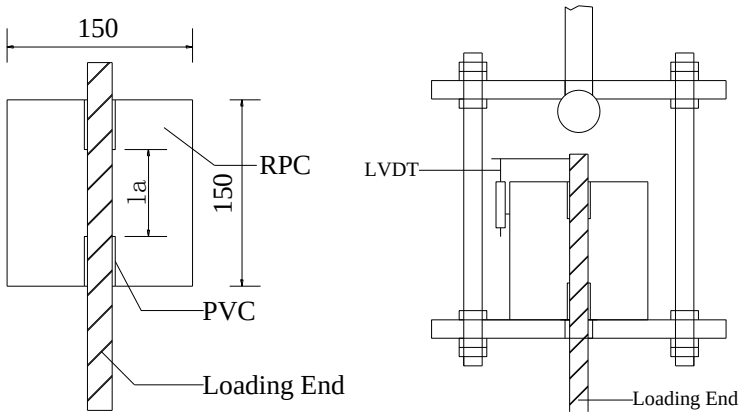
Table 2: Measured strength of rebar

| Diameter(mm) | Elastic modulus (GPa) | Yield Strength (MPa) | Ultimate strength (MPa) |
|---|---|---|---|
| 16 | 200 | 413 | 587 |
| 20 | 200 | 430 | 613 |

### 2.3. Pull-out test

To evaluate the bond behaviors of steel rebar embedded in RPC, the pull-out test was carried out in accordance with GB/T 50081-2002[11], the direct pull-out test method was preferred because it is simple and leaves the free end accessible for the slip to be measured.

The nominal diameters of the steel rebar were 16mm and 20mm, the embedded length were 3, 4, 5 and 6 times the bar diameter, the concrete cover were 1, 1.6, 2.5cm and the concrete age were 7, 14 and 28 days to investigate the bond properties. The dimension of Pull-out specimens is 150mm×150mm×150mm. The detailed geometry and test setup are shown in Fig. 1(a). The single bar placed horizontally along the central axis. The unbonded region of the rebar was sheathed with PVC pipe. A pull-out load was applied using a WAW with a maximum load capacity of 1000 kN through the pulling force, with the rate of load increase was 0.1 kN/s during testing. The LVDTs were used to measure the bond slip between the rebar and the concrete at free end, as shown in Fig.1(b).




(a)Dimensions of specimen (b) Test setup

Fig.1 Pull-out test

## 3. EXPERIMENTAL RESULTS AND DISCUSSION

The effects of rebar embedment length, rebar diameter, concrete age and size of the concrete cover on the bond behavior is analyzed in this section. The bond-slip curves for the steel rebar embedded in the RPC were plotted using the experimental results obtained directly from the slip measurements at the free end, the bond properties are summarized in Table 3. The bond stress distribution between the rebar and RPC is not constant along the embedment length. For simplification, the average bond stress $\tau$is calculated using Eq. (1).

$$\tau = \frac{F}{\pi d l_a} \tag{1}$$

where $F$ is the pull-out load, $l_a$ is the embedment length of rebar, and $d$ is the diameter of rebar.

The notation adopted for the specimens in Table 3 is as follows: the first number indicates the bar diameter, the next number is the embedment length, the third number

is the curing age, and the last number is size of the concrete cover. For example, 16-48-28d-1cm refers to a 16 mm bar with an embedment length 48cm, the concrete cover 1cm, tested after 28 days.

For all test specimens, the whole process of pull-out test is as follows: the bond stress-slip relationship is characterized by an initial increase in the pullout stress with slight slip, followed by softening after the bond strength. As the pull-out load increased, the slip increased and became measurable with crack formation, resulting in the shearing off and crushing of concrete at the bonded region[12]. However, high residual bond strength remains during the bar pullout phase because of friction.

Table 3: Pull-out test results

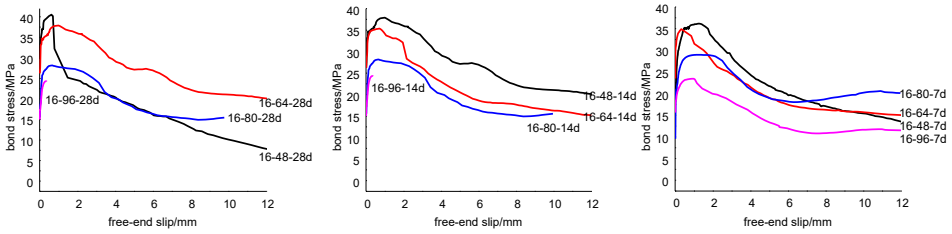
| Specimens | Embedment length(mm) | Concrete age(days) | Maximum average bond strength (MPa) | Slip at maximum average bond strength (mm) |
|---|---|---|---|---|
| 16-48-7d | 48 | 7 | 35.55 | 0.55 |
| 16-48-14d | 48 | 14 | 37.89 | 1.00 |
| 16-48-28d | 48 | 28 | 40.53 | 0.64 |
| 16-64-7d | 64 | 7 | 33.18 | 1.25 |
| 16-64-14d | 64 | 14 | 35.36 | 0.72 |
| 16-64-28d | 64 | 28 | 36.57 | 1.00 |
| 16-80-7d | 80 | 7 | 27.64 | 0.40 |
| 16-80-14d | 80 | 14 | 28.54 | 0.22 |
| 16-80-28d | 80 | 28 | 29.26 | 0.29 |
| 16-96-7d | 96 | 7 | 20.04 | 0.99 |
| 16-96-14d | 96 | 14 | 24.11 | 0.51 |
| 16-96-28d | 96 | 28 | 24.07 | 0.36 |
| 20-80-7d | 80 | 7 | 26.59 | 0.27 |
| 20-80-14d | 80 | 14 | 27.72 | 0.18 |
| 20-80-28d | 80 | 28 | 28.85 | 0.22 |
| 16-48-28d-1cm | 48 | 28 | 22.45 | 0.19 |
| 16-48-28d-1.6cm | 48 | 28 | 24.69 | 0.22 |
| 16-48-28d-2.5cm | 48 | 28 | 33.39 | 0.23 |
| 16-64-28d-1cm | 64 | 28 | 21.18 | 0.19 |
| 16-64-28d-1.6cm | 64 | 28 | 25.29 | 0.06 |
| 16-64-28d-2.5cm | 64 | 28 | 30.64 | 0.64 |

### 3.1. Effect of the embedment length

Fig.2 shows the bond stress and slip relations of different embedment lengths (3d, 4d, 5d and 6d) for steel rebars embedded in RPC. The main findings were as follows:

The results show that with the increase of rebar embedment length, the ultimate bond stress decreases pseudo-linearly[13-14]. The specimens with longer embedment length developed lower bond strength. One reason of this decrease is caused by the nonlinear distribution of bond stresses along the steel bar embedded in concrete. The longer of the rebar embedment length, the shorter and narrower of the high stress

region, so specimens with longer embedment length had low average bond stress. The other reason of this decrease is caused by Poisson's ratio effect where the substantial elongation of the bar throughout the embedment length leads to reduction in friction. The bond stress decreased rapid and largely(about 40%) when the local bond stress reach the ultimate bond stress.




(a) 16-28d (b) 16-14d (c) 16-7d

Fig.2 Embedment length effect on the bond-slip relationship

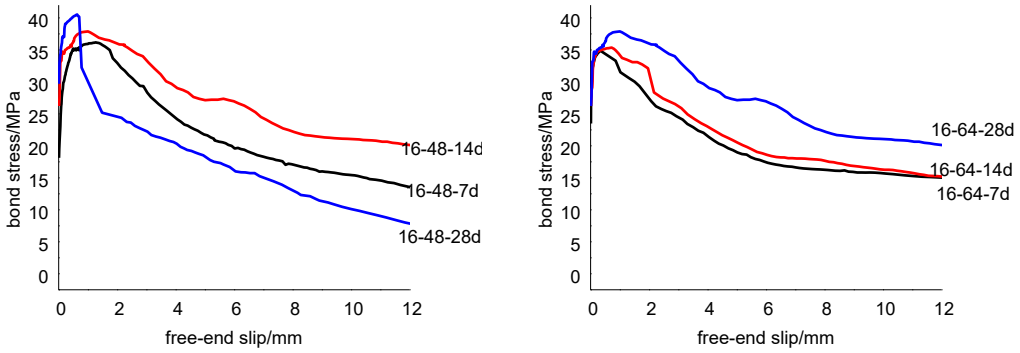
### 3. 2. Effect of bar diameter

The test results in Table.3 show that with the increase of the diameter of the bars, the ultimate average bond strength decreased slightly. For example, at the embedment length of 80cm, the bond strength of the specimens with a diameter of 16mm were 27.64MPa(7d), 28.54MPa(14d), 29.26MPa(28d), which are about 1.4%-3.9% higher than those of the specimens with a diameter of 20 mm. This decrease is caused by the bleeding of the water in the concrete. The larger the diameter, the more bleed water is trapped beneath the bar, which creates larger voids in the interface [14, 15]. The voids reduce the bond strength on the contact surface between the bar and the concrete. The Poisson's ratio effect also can allow a slight reduction in bar diameter as a result of longitudinal stress. It's more easily to brittle failure in concrete surface when the bars were pulled out as the diameter of the bar increases.

### 3. 3. Effect of concrete age (confinement effect)

With the growth of the concrete age, the compressive strength and flexural strength of RPC increase. Studying the effect of concrete age amounts to evaluating the effect of the confinement pressure caused by concrete shrinkage on the bond strength[16]. The results obtained from pullout tests performed after 7, 14and 28 days with smooth bars (d=16mm) are displayed in Figs. 3 and Table 3. The following remarks can be made:

- The bond-slip relationship curves were the same shape for all the specimens. The bond-slip relationship become full with the increase of concrete age.
- With the growth of the concrete age, the ultimate bond strength of specimens increase. That because with the growth of the concrete age, the radial confined pressure on the bond strength increased, too.



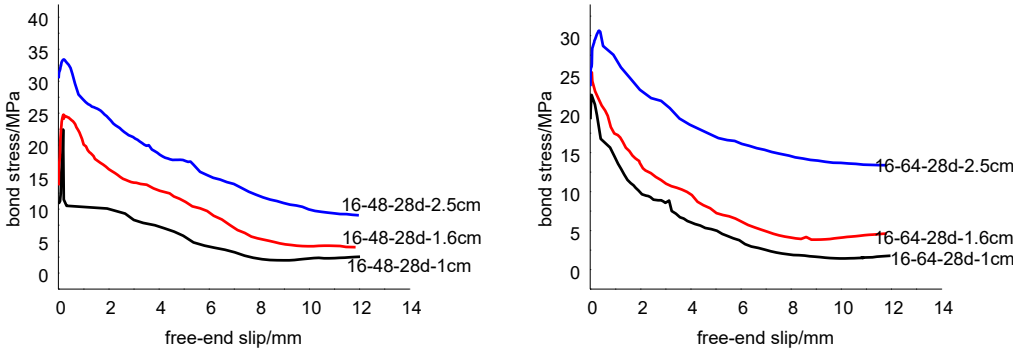

(a) 16-48 (b) 16-64

Fig.3 Effect of concrete age on the bond-slip relationship

### 3. 4. Effect of concrete cover

Studies have been performed to evaluate the effect on the bond strength of the concrete cover. The results obtained from pullout tests performed after 28 days with the concrete cover (1cm, 1.6cm and 2.5cm) are displayed in Figs.4. The main findings were as follows:

- The bond performance curves were the same shape for all the specimens. The bond-slip relationship become very full with the increase of concrete cover.
- With the growth of the concrete cover, the ultimate bond strength of specimens increased. That is caused as the growth of concrete cover increases the contact surface confinement pressure between the bar and the concrete.




(a) 16-48-28d (b) 16-64-28d

Fig.4 Effect of concrete cover on the bond-slip relationship

### 3.5 Anchorage length

The force applied in the rebar is resisted by the bond stress acting on the surface of the rebar based on the equilibrium condition. The equilibrium of the forces is expressed by Eq. (2).

$$f_y A_s = \pi d l_a \tau \tag{2}$$

$$l_a = \frac{f_y d}{4\tau} \tag{3}$$

where $f_y$ is steel yield strength, $A_s$ is the area of one rebar, $d$ is the diameter of the steel rebar, $l_a$ is the critical anchorage length, $\tau$ is bond strength. By this formula, the adequate anchorage length for bars with the diameters of 16mm is worked out to be 5.5 times of the bar diameters.

## 4 CONCLUSIONS

- All test specimens failed by pull-out of the rebar with the shearing off and crushing of concrete between the lugs of the rebar. The bond of reinforcement in RPC is characterized by very high maximum bond stresses, even at early ages. High residual bond strength remains during the bar pullout phase because of friction.
- The bond strength was slightly decreased with increasing embedment length due to the nonlinear distribution of the bond stress and the Poisson's ratio effect. The specimens with longer embedment length developed lower bond strength.
- The average maximum bond strength decreased when the diameter of the bars increased. This decrease is caused by the voids in the interface and the voids reduce bond strength on the contact surface between the bar and the concrete.
- With the growth of the concrete age and the concrete cover, the ultimate bond strength of specimens increase. The reason of this increase is caused by the radial confined pressure on the bond strength increased, too.
- The adequate anchorage length for bars with the diameters of 16mm is worked out to be 5.5 times of the bar diameters.

## ACKNOWLEDGMENTS

This research was supported by a grant from China railway company science and technology research and development program (2014G010-D).

## REFERENCES

[1] Dugat J, Roux N, Bernier G. Mechanical properties of reactive powder concretes[J]. Materials and structures, Vol. 29, May 1996, 233-240.

[2] Richard P, Cheyrezy M. Composition of reactive powder concretes. Cem Concrres 1995;25(7):1501-11.

[3]Bonneau O, Pouhn C. Reactive powder concrete from theory to practice[J]. Concrete International.1996, 18(4):47-49.

[4]Dallaire S, Korb J P. Analysis of microporsity and setting of reactive powder concrete by proton nuclear relaxation[J]. Magnetic Resonance Imaging. 1998, 16:515-518.

[5] An Mingzhe, Zhang Meng. Experimental research of bond capability between deformed bars and reactive powder concrete[J]. China Railway Science, 2007, 28(2):50-54(in Chinese)

[6]Deng Zongcai, Yuan Changxing. Experimental study on bond capability between high strength rebar and reactive powder concrete[J].China civil engineering journal, 2014, 47(3)：69-78(in Chinese)

[7] Si Jinyan, Fan Xiaoning, Zhou Licheng, Xiao Rui, Yuan Changxing. Research on Bond Behavior of Reactive Powder Concrete and High-strength Reinforcement[J]. Construction Technology, 2013(12)：60-62 (in Chinese)

[8] Tang Hao, Wang Jian. Relationship of bar diameter and bonding performance of reactive powder concrete[J]. Journal of Beijing University of Civil Engineering and Architecture, 2012, 28(3):6-9, 18(in Chinese)

[9]Cosenza E, Manfredi G, Realfonzo R. Analytical modeling of bond between FRP reinforcing bars and concrete. In: Taerwe L, editor. Proceedings of second.

[10]GB/T14902-2012 Code for the ready-mixed concrete[S]. Beijing: China Architecture Building Press, 2012( in Chinese)

[11]GB 50010-2002 Standard for test method of mechanical properties on ordinary concrete[S]. Beijing: China Architecture Building Press, 2002( in Chinese)

[12]Lee JY, Kim TY, Kim TJ, Yi CK, Park JS, You YC, et al. Interfacial bond strength of glass fiber reinforced polymer bars in high-strength concrete. Compos Part B: Eng 2008;39(2):258-70.

[13] Cosenza E, Manfredi G, Realfonzo R. Behaviour and modelling of bond of FRPrebars to concrete. ASCE J Compos Constr 1997;1(2):40-51.

[14] Nanni A, Bakis CE, Boothby TE. Test methods for FRP‐concrete systems subjected to mechanical loads: state of the art review. Reinf Plast Compos J 1995;14(6):524-58.

[15] Benmokrane B, Tighiouart B, Chaallal O. Bond strength and load distribution of composite GFRP reinforcing bars in concrete. ACI Mater J 1996;93(3):254-9.

[16] Sayed Ahmad Firas, Foret Gilles *, Le Roy Robert. Bond between carbon fibre-reinforced polymer (CFRP) bars and ultra-high performance fiber reinforced concrete (UHPFRC): Experimental study. Construction and Building Materials 25 (2011) 479-485.

[17] GB 50010-2010 Code for design of concrete structures[S]. Beijing: China Architecture Building Press, 2010( in Chinese)

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MODELLING SOIL-FIBRE COMPOSITE BEHAVIOUR USING A MICROMECHANICAL APPROACH

**Thomas A. Bower[1], Anthony D. Jefferson[2], Peter J. Cleall[2] and Paul Lyons[3]**

[1]School of Engineering, Cardiff University
Queen's Buildings, The Parade, Cardiff, CF24 3AA, UK
e-mail: BowerTA@cardiff.ac.uk

[2] School of Engineering, Cardiff Univerisity
Queen's Buildings, The Parade, Cardiff, CF24 3AA, UK
e-mail: {JeffersonAD,Cleall}@cardiff.ac.uk

[3] LUSAS
Forge House, 66 High Street, Kingston Upon Thames, Surrey, KT1 1HN, UK
e-mail: paul.lyons@lusas.com

**Keywords:** composite, soil, fibre, constitutive model, shear lag, debonding

**Abstract.** *Experimental data has shown that the addition of manufactured fibres to a soil can significantly increase its strength. The mechanism of strength increase is due to the frictional interaction between the soil particles and each fibre. This has a global effect of increasing the shear strength of the soil. This paper presents the initial development of a micro-mechanics based soil-fibre model. This model considers the micromechanical soil-fibre interaction and in particular utilises the shear lag model which is extended to include fibre de-bonding. The effects of a single fibre are then homogenised using a numerical spherical integration technique to take into account the distribution of fibre orientations. The fibre model is used in conjunction with a hardening soil plasticity model. The proposed soil-fibre model is then compared against experimental data for triaxial compression of fibres mixed with a well graded sand and different fibre concentrations. The volumetric response shows the correct trend with different fibre contents however the actual values are somewhat inaccurate, The shear stress response is very well captured. It is concluded that the results could be improved by addressing the known limitations in the base soil model and the fibre model. Future work will involve the improvement of both.*

## 1 INTRODUCTION

The requirements for engineering soils are becoming more and more demanding, with larger structures being built and less choice in where to put them, engineers have fewer options with site selection and soils often need to be improved. This can be achieved using methods such as vibro-compaction and drop weight compaction to increase the soil density, or by importing stronger soils. Geosynthetics is another option to improve soil strength which can be used in conjunction with other methods. These include geogrids, geomats, geomembranes and fibre reinforcement. One advantage of using fibre reinforced soils over planar grids / mats is that they do not introduce any particular planes of weakness.

A variety of fibre materials are available for soil reinforcement. Possibly the original application of fibre reinforcement is the use of plant roots which have been used historically for the stabilisation of slopes and embankments. More recently, investigations have been undertaken into the behaviour of manufactured fibres in soils. Such studies have looked at fibres constructed from plastics such as polypropylene [5, 9, 11] and polyamide [15], or natural fibres such as coir [13] and oil palm [1]. Figure 1 depicts the crimped polypropylene fibres used in this study.

The primary mechanism for strength increase is due to the interfacial friction between the soil particles and the fibre: isotropic pressure compacts the composite material, causing soil particles to partially embed in the fibre. The presence of the fibre resists shear displacements hence increasing the shear strength of the soil.

Fibres used for reinforcing soils must be significantly stronger than the soil mass. It would be futile adding fibres which would break before soil failure. Manufactured fibre materials such as polypropylene and polyethylene generally do not reach the yield stress within the strain range observed in soils; also axial fibre strains dominate radial fibre strains by far; thus fibres are typically considered as one-dimensional axially loaded elastic elements in fibre models.

Several soil-fibre models have been developed in recent years. Namely, Maher and Gray [12] created a micro-mechanical soil-fibre interaction model whereby fibres crossing potential shear planes increased the shear strength across that plane. The effects on a single direction were then homogenised using a spherical integration technique. Michalowski and Čermák [15] developed a failure criterion for fibre reinforced soils whereby a Mohr-Coulomb type cone model was expanded due to the presence of fibres. Diambra et al. [8] used the rule of mixtures to create an equivalent fibre stiffness matrix which was superimposed onto the soil stiffness matrix. A relationship for fibre debonding was included in this model and later improved by Diambra and Ibraim [10].

## 2 FIBRE MODEL

The first step in modelling soil containing randomly distributed fibres is to analyse the effect a single fibre has on the soil surrounding it; an idealised schematic of the soil and fibre is shown in Figure 2. By taking an infinitesimal slice of fibre / soil, and balancing the forces due to the applied matrix strain $\varepsilon_{ma}$ leads to the well known shear lag equation [7].

$$\frac{\mathrm{d}\sigma_{fa}}{\mathrm{d}x} = \frac{-2\tau}{r_f} \tag{1}$$

where $\sigma_{fa}$ is the fibre tensile stress, $x$ is the distance from the fibre centre, $\tau$ is the interface shear stress, $r_f$ is the fibre radius. Through manipulation of Eq. 1, using stress/strain relationships in the fibre and soil, the slip between the matrix and fibre at any point along the fibre is found to




Figure 1: Polypropylene Loksand fibres used in experimental study [6]

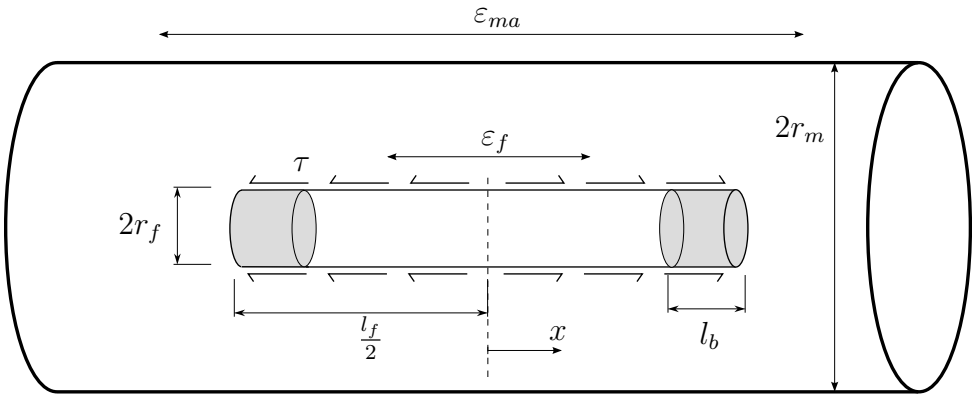



Figure 2: Single fibre and surrounding matrix material, showing debonding regions

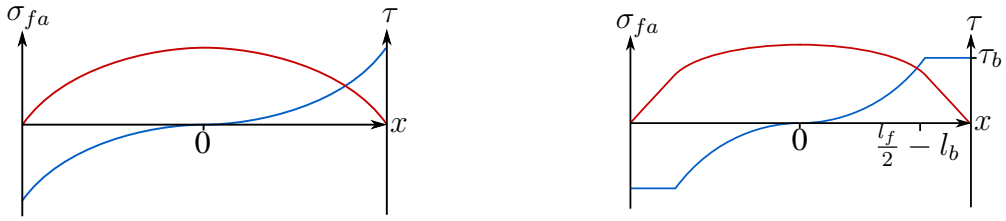
be

$$S(x) = \frac{\varepsilon_{ma}}{\beta \cosh(\beta l_f/2)} \sinh(\beta x) \tag{2}$$

where $\beta$ is a constant relating various properties of the soil, the fibre and the interface. As the matrix strain $\varepsilon_{ma}$ increases, the interface shear stress also increases, up to a limit $\tau_b$. With further loading, the limiting shear stress will expand towards the fibre centre. The length of this debonded region is given by $l_b$ which is calculated using the applied matrix strain.

$$S(x) = \begin{cases} \dfrac{\tau_b}{k_s} \dfrac{\sinh(\beta x)}{\sinh(\beta(l_f/2 - l_b))} & \text{if } 0 < x \leq l_f/2 - l_b \\ \dfrac{\tau_b}{k_s} & \text{if } l_f/2 - l_b < x \leq l_f/2 \end{cases} \tag{3}$$

The fibre tensile stress is found by integrating the slip relations from the fibre centre to the desired position. This applies to both the bonded case 2 and the debonding case 3. The axial stress and shear stress distribution functions are visualised in Figure 3.




(a) Fully bonded

(b) Partially debonded

Figure 3: Tensile fibre stress $\sigma_{fa}$ (red) and interface shear stress $\tau$ (blue) distributions for bonded and partially debonded fibre along the length of the fibre [4]

$$\sigma_{fa} = \frac{-2k_s}{r_f} \int_0^x S(x)dx \tag{4}$$

## 3 SPHERICAL INTEGRATION

To translate individual fibre stresses to global stresses suitable for use in finite element analysis, the behaviour of the fibres must be homogenised. Another important factor at this stage is the distribution of fibre orientation. The sample preparation method has a significant impact on the fibre orientations [11, 17]. In a triaxial compression test, the vertical fibres have little impact on the results, whereas fibres on the horizontal plane resist the tensile strains and affect the results significantly. The orientation distribution function used in this study [14] is given in Eq. (5) with the parameters $A = 0$, $B = 0.57$, $n = 7$ based on results reported by Wang [17].

$$\rho_\theta = A + B|\sin^n \theta| \tag{5}$$

The distribution is then used in a numerical spherical integration method [2]. This approach is similar in form to Gauss-Legendre integration used in finite element formulations, summating the fibre effects over a number of sampling directions and applying sampling weights to each direction.

$$\sigma_f = v_f \sum_{i_d=1}^{n_d} \rho(\theta_{id}) \cdot \chi(\theta_{id})^T \int_0^{l_f/2} \frac{2}{l_f} \sigma_{fa}(x, \varepsilon_{ma})\, \mathrm{d}x \cdot w_{id} \tag{6}$$

The integral term in Equation (6) takes into account the different positions of the fibres. $n_d$ is the number of sampling directions, $\sigma_{fa}$ is the fibre axial stress for either fully or partially bonded fibres and $w_{id}$ is the current sample weight. $\chi(\theta_{id})$ forms a direction cosine vector for each sampling direction. This vector translates the fibre stresses into the global stress space (Cartesian, axisymmetric etc.). The same vector is also used to translate the global strains into the applied axial matrix strain $\varepsilon_{ma}$.

## 4 THE HARDENING SOIL MODEL

The hardening soil model, initially developed by Schanz [16] and later improved by Benz [3], is a constitutive soil model capable of capturing non-linear behaviour exhibited by many soils including stress dependent stiffness, dilatancy and hardening from initial loading. The model performs well for simulating both laboratory tests and more complex boundary value

problems. The HS model relies on a set of yield surfaces which define the boundary of elastic and plastic strains. The shear surface (7) is a curved cone in principal stress space (see [16]).

$$f_s = \frac{q_a}{E_{50}} \frac{q}{q_a - q} - \frac{2q}{E_{ur}} - \gamma_p \tag{7}$$

where $q_a$ is the failure shear stress (governed by a Mohr-Coulomb criterion), $E_{50}$ and $E_{ur}$ are the current secant and unload-reload stiffnesses, $q$ is the shear stress and $\gamma_p$ is the current value of the plastic shear strain and controls the size of the shear surface. This is a non-associated surface and the plastic potential in shearing is the Drucker-Prager cone with the dilatancy angle $\psi$ controlling the cone steepness. The cap surface takes the form of an ellipse in p-q stress space.

$$f_c = p^2 + \frac{q^2}{\alpha^2} - p_p^2 \tag{8}$$

where $p$ is the mean stress, $\alpha$ is a parameter which controls the cap steepness and $p_p$ is the equivalent pre-consolidation stress and controls the size of the cap surface. The cap surface uses an associated flow rule therefore the plastic potential function is equal to the yield function. During drained triaxial testing, both surfaces are active, therefore a double hardening closest point projection method is employed.

## 5 RESULTS

The predicted results were generated by a single point constitutive driver in Cartesian stress space using the fibre and soil properties listed in Table 1 and Table 2. An additional subset of soil properties were used in the hardening soil model, however they are not listed here. The triaxial test simulation was undertaken by supplying an initial isotropic stress of 0.3 N/mm$^2$, an axial strain was applied at a rate of $\Delta\varepsilon_y = 0.6\%$. Shear stress and volumetric strain curves are compared with experimental data [6] in Figure 4. A quadratic rate of convergence was observed throughout the simulations.

Table 1: Properties of the Loksand fibre

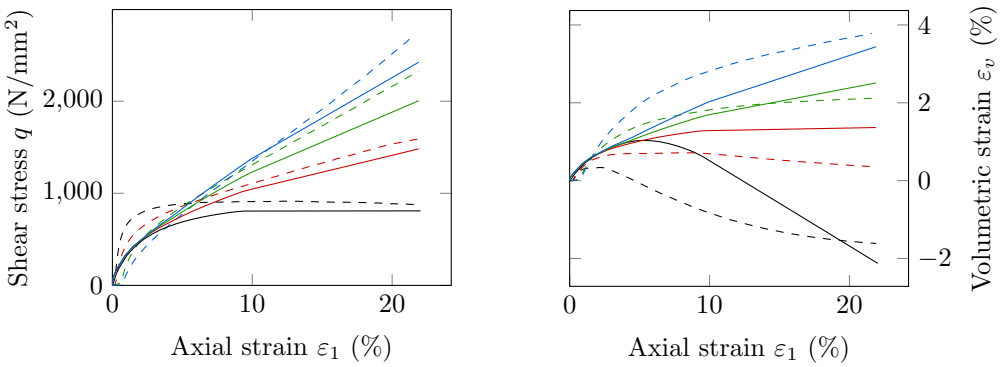
| Young's modulus $E_f$ | radius $r_f$ | length $l_f$ |
|---|---|---|
| 900 N/mm$^2$ | 0.044 mm | 35 mm |

Table 2: Strength properties of the sand

| Young's modulus $E_m$ | Poisson's ratio $\nu$ | Friction angle $\varphi$ | Dilatancy angle $\psi$ | Cohesion $c$ |
|---|---|---|---|---|
| 40.6 N/mm$^2$ | 0.3 | 35° | 2° | 0 |

## 6 DISCUSSION AND CONCLUSIONS

The shearing response of an unreinforced soil starts with a high stiffness which reduces with increasing axial load until it reaches failure where no further shear stress can develop. For the reinforced soil, this failure limit is not reached (Figure 4a), and the shear stress is allowed to




(a) Shear response

(b) Volumetric response

Figure 4: Simulated (solid) and experimental (dashed) drained triaxial results for fibre contents by weight of 0.0% (black), 0.3% (red), 0.6% (green) and 0.9% (blue); experimental results by Chatzopoulos [6]

increase with axial load indefinitely within the strain range considered. With increasing fibre content, this effect is increased. This behaviour is captured well in the soil-fibre model with only small differences in the shear response past 10% axial strain.

The initial response (lower than 10% axial strain) is somewhat mis-represented in shearing. The initial stiffness of the unreinforced sample is under-predicted; suggesting that there are some limitations in the base soil model. Also, the initial stiffness of the reinforced samples is shown to decrease with increasing fibre content; the fibre model does not reproduce this behaviour.

The unreinforced sand undergoes an initial compressive stage, then after 3% axial strain (Figure 4b), begins to exhibit dilatancy. The predictions of this do show dilatancy but at a later strain of 6%. This is most likely due to the simplified dilatancy rule used in the soil model. The experimental result also shows a dilatancy cut-off at $\varepsilon_v = -1.8\%$, the model does not predict this.

As the fibre content is increased, the experimental results show a decrease in dilatancy. This trend is captured well however the values are less accurate than the shear response. The effects of the simplified dilatancy rule in the soil model may be dominating over the influence of the fibres, so it is difficult to comment on the accuracy of this element of the fibre model.

In summary, the predictions match the experimental results somewhat better than in previous work in this ongoing project [4]. In particular, the volumetric response is much better predicted. Further work will focus on the debonding behaviour of the fibres and the dilatant response of the soil.

## 7 Acknowledgements

The work presented here was supported by LUSAS / Finite Element Analysis ltd. and Innovate UK / TSB as part of KTP project number 9082. We would also like to thank Drake Extrusion Inc. for supplying the fibres used in the experimental testing.

## REFERENCES

[1] F. Ahmad, F. Bateni, M. Azmi, Performance evaluation of silty sand reinforced with fibres. *Geotextiles and Geomembranes*, **28.1**, 93–99, 2010.

[2] Z.P. Bažant and B.H. Oh., Efficient numerical integration on the surface of a sphere. *ZAMM - Journal of Applied Mathematics and Mechanics*, **66.1**, 37–49, 1986.

[3] T. Benz, *Small-strain stiffness of soils and its numerical consequences*. Univ. Stuttgart, Inst. f. Geotechnik, 2007.

[4] T.A. Bower, A.D. Jefferson, P.J. Cleall, P. Lyons A micro-mechanics based soil-fibre composite model for use with finite element analysis. *submitted to Association of Computational Mechanics in Engineering*, 2016.

[5] Y. Cai, B. Shi, C.W.W. Ng, C.S. Tang, Effect of polypropylene fibre and lime admixture on engineering properties of clayey soil. *Engineering Geology*, **87.3**, 230–240, 2006.

[6] S. Chatzolpoulos, *Soil reinforced with discrete polypropylene fibers*. MSc Thesis, Cardiff University, 2015.

[7] H.L. Cox, The elasticity and strength of paper and other fibrous materials. *British Journal of Applied Physics*, **3.3**, 72–79, 1952.

[8] A. Diambra, E. Ibraim, D. Muir Wood, A.R. Russell, Fibre-reinforced sands: experiments and modelling. *Geotextiles and Geomembranes*, **28.3**, 238–250, 2010.

[9] A. Diambra, E. Ibraim, A.R. Russell, D. Muir Wood, Fibre reinforced sands: from experiments to modelling and beyond. *International Journal for Numerical and Analytical Methods in Geomechanics*, **37.15**, 2427–2455, 2013.

[10] A. Diambra, E. Ibraim, Fibre reinforced sands: interaction at the fibre and grain scale. *Géotechnique*, **65.4**, 296–308, 2015.

[11] E. Ibraim, A. Diambra, A.R. Russell, D. Muir Wood, Assessment of laboratory sample preparation for fibre reinforced sands. *Geotextiles and Geomembranes*, **34**, 69–79, 2012.

[12] M.H. Maher, D.H. Gray, Static response of sands reinforced with randomly distributed fibers. *Journal of Geotechnical Engineering*, **116.11**, 1661–1677, 1990.

[13] T. Maliakal, S. Thiyyakkandi, Influence of randomly distributed coir fibers on shear strength of clay. *Geotechnical and Geological Engineering*, **31.2**, 425–433, 2013.

[14] R.L. Michalowski, J. Čermák, Strength anisotropy of fiber-reinforced sand. *Computers and Geotechnics*, **29.4**, 279–299, 2002.

[15] R.L. Michalowski, J. Čermák, Triaxial compression of sand reinforced with fibers. *Journal of Geotechnical and Geoenvironmental Engineering*, **129.2**, 125–136, 2003.

[16] T. Schanz, P.A. Vermeer, P.G. Bonnier, The hardening soil model: formulation and verification. *Beyond 2000 in computational geotechnics*, 281–296, 1999.

[17] Z. Wang, *Fibre reinforced sand: from sample preparation to triaxial tests and orientation investigation*. MSc Thesis, Cardiff University, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF MECHANICAL BEHAVIOR OF WOVEN COMPOSITE AT DIFFERENT STRAIN RATE BY A COLLABORATIVE ELASTO-PLASTO-DAMAGE MODEL WITH FRACTIONAL DERIVATIVES

**Alina Krasnobrizha**[1], **Patrick Rozycki**[1], **Laurent Gornet**[1], **Pascal Cosson**[1]

[1] Institut de recherche en Génie Civil et Mécanique (GeM) - UMR CNRS 6183, Ecole Centrale de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3, France
e-mail: alina.krasnobrizha@ec-nantes.fr, patrick.rozycki@ec-nantes.fr, laurent.gornet@ec-nantes.fr, pascal.cosson@ec-nantes.fr

**Keywords:** Composite material, Behavior law, Fractional derivative, Hysteresis, Strain-rate sensitivity.

**Abstract.** *This paper deals with a collaborative model to represent hysteresis behavior at different strain rates. The model consists of two sub-models. The first one treats the behavior during loading path. The elastic and in-elastic strains are computed as well as the in-ply damages. The strain-rate sensitivity is also taken into account. The second sub-model involves a fractional derivate approach to describe viscoelastic material response during unloading path. The hysteresis loops and strain rate sensitivity are taken into account by fractional constitutive law. Fractional model involves a few parameters which are easily identified through an optimization procedure from the experimental data. The model is validated for thermoset and thermoplastic composite materials at different strain rates.*

## 1 INTRODUCTION

The extensive use of composite materials in industrial applications requires a better understanding of their mechanical behavior. Composites materials are anisotropic and heterogeneous. Complex models are required to adequately describe their behaviors. There are a lot of works concerning the behavior of unidirectional or woven composites with thermosetting matrix under quasi-static loadings [1, 2, 3]. These models take into account the damage propagation and the in-elastic strains appearing commonly in the shear and transverse directions. The strain rate dependent models were developed for unidirectional composites [4, 5] and for woven fabrics [6].

The previously developed models describe different physical phenomena, such as, the elastic and in-elastic strains, the damages propagation in orthotropic directions and the strain rate sensitivity very well. However, they cannot represent the hysteretic behavior of the material during its unloading path. The fractional derivative approach is a good technique to model the viscoelastic behavior for some natural structures and modern heterogenic materials such as elastomers and polymers [7]. A significant number of works deals with fractional viscoelastic constitutive equations for different materials under various types of loading. Caputo applied fractional Zener model to represent the behavior of glass and a few metals [8]. Bagley and Torvik proposed a fractional law in frequency domain to describe the response of certain polymers and elastomers [9]. The physical sense of fractional operators is given in the works [10], [11] for different types of polymers by using Rouse molecular theory [12]. Rabotanov presented a generalized rheological model to describe the behavior of hereditary medium [13], [14]. The hysteresis cycles for a few metals under fatigue loading, were represented by Caputo, in the frequency domain [15]. Mateos [16] proposed a constitutive model for composite materials under cyclic quasi-static shear loading and this was based on the fractional derivative approach. This model also includes strain-rate dependence and thus can be applied for dynamic loading [17]. The model is also able to represent hysteresis composite behavior using a few material parameters. So, the elastic and irreversible strains, damage and strain rate effects are taken into account but not material hardening. To fill this gap, a collaborative model is developed [18] which includes the elastoplastic damage behavior law [2] with strain-rate sensitivity [4] and a fractional derivative approach. In this paper the simulation is made by the collaborative model at different strain rates.

## 2 THEORETICAL MODEL FOR COMPOSITE PLY

The constitutive model is developed within the framework of thermodynamics isothermal irreversible processes for a woven composite elementary ply under a state of plane stress. Subscripts 1 and 2 represent the warp and the weft directions, respectively. The composite is considered to be perfectly balanced and thus the longitudinal and transverse behaviors are considered to be equivalent. The continuum damage mechanics theory is used to describe material degradation such as the matrix micro-cracking and the fiber/matrix debonding. Also an isotropic hardening is assumed and the viscoelastic effects are expressed by the fractional derivatives.

The material parameters are obtained from the experimental campaign based on the cyclic tensile tests. The tensile test performed on the [0°/90°] composite allows us to characterize the longitudinal (transverse) material behavior. The shear response is obtained from the tensile cyclic test on $[\pm 45°]$ composite. Commonly, woven composite material has a brittle elastic response in the fiber direction. The elastic and the in-elastic strain, damage propagation and hysteresis loops, as well as the strain-rate sensitivity are present during shear cyclic test. The

typical material response is illustrated for thermoset woven composite (carbon/epoxy) in the fiber direction in the Figure 1 and in the shear direction, for two different strain rates, in the Figure 2.

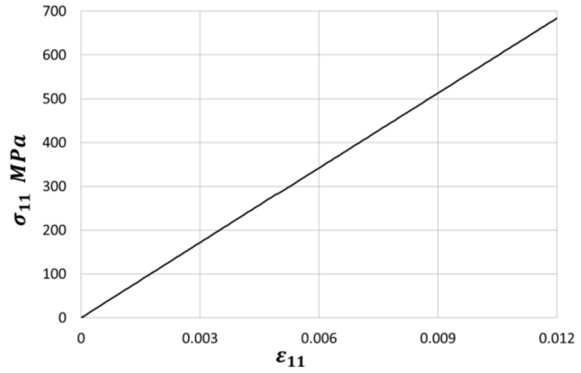



Figure 1. Longitudinal (transverse) traction curve for the carbon/epoxy composite

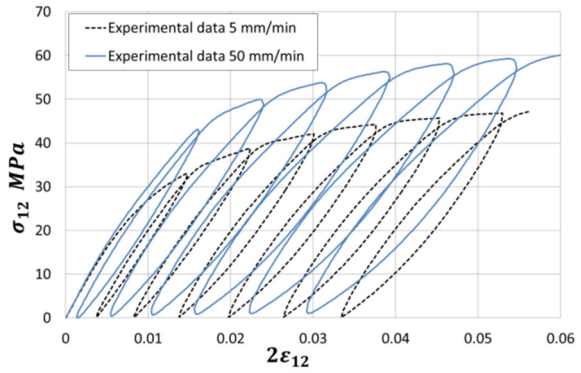



Figure 2. Shear stress/strain curve for the carbon/epoxy composite for two different strain-rates

The developed constitutive model is composed of two sub-models. The first one deals with the elastoplastic damage behavior during material loading and the second sub-model describes the hysteresis behavior using a fractional derivative law. Both sub-models are strain-rate dependent and can be applied for the dynamic problems.

Within the framework of thermodynamic irreversible theory, we choose Helmholtz potential depending on the internal variables:

$$\rho\psi = \rho\psi(\boldsymbol{\varepsilon}^e, d_i, p) \tag{1}$$

where $\boldsymbol{\varepsilon}^e$, $d_i$, $p$ are internal variables associated with elastic strain, damage in the orthotropic directions and cumulated plasticity respectively.

## 2.1 Damage model

Following the second principal of thermodynamic, constitutive equations are deduced from the elastic strain energy of the damaged material (equivalent to the Helmholtz potential $\psi$) which has a following form:

$$\begin{aligned} W_e^d = \frac{1}{2}\{&C_{11}^0(1-d_{11})(\varepsilon_{11}^e)^2 + C_{22}^0(1-d_{22})(\varepsilon_{22}^e)^2 + 2\nu_{21}^0 C_{11}^0 \varepsilon_{11}^e \varepsilon_{22}^e \\ &+ G_{12}^0(1-d_{12})(2\varepsilon_{12}^e)^2\} \end{aligned} \tag{2}$$

where $\nu_{12}^0$ and $\nu_{21}^0$ are the Poisson's ratios, $C_{11}^0$, $C_{22}^0$ and $G_{12}^0$ are components of stiffness matrix (3).

$$\boldsymbol{C^0} = \begin{pmatrix} C_{11}^0 & \nu_{21}^0 C_{11}^0 & 0 \\ \nu_{12}^0 C_{22}^0 & C_{22}^0 & 0 \\ 0 & 0 & G_{12}^0 \end{pmatrix} \quad \text{with} \quad \begin{cases} C_{11}^0 = \dfrac{E_{11}^0}{1-\nu_{12}^0\nu_{21}^0}, C_{22}^0 = \dfrac{E_{22}^0}{1-\nu_{12}^0\nu_{21}^0} \\ \nu_{21}^0 E_{11}^0 = \nu_{12}^0 E_{22}^0 \end{cases} \tag{3}$$

The stress-strain relation is:

$$\boldsymbol{\sigma} = \frac{\partial W_e^d}{\partial \boldsymbol{\varepsilon}^e} \implies$$

$$\begin{pmatrix} \sigma_{11} \\ \sigma_{22} \\ \sqrt{2}\sigma_{12} \end{pmatrix} = \begin{pmatrix} (1-d_{11})C_{11}^0 & \nu_{21}^0 C_{11}^0 & 0 \\ \nu_{12}^0 C_{22}^0 & (1-d_{22})C_{22}^0 & 0 \\ 0 & 0 & 2(1-d_{12})G_{12}^0 \end{pmatrix} \begin{pmatrix} \varepsilon_{11}^e \\ \varepsilon_{22}^e \\ \sqrt{2}\varepsilon_{12}^e \end{pmatrix} \tag{4}$$

Thermodynamic forces associated with internal variable $d_{ij}$ are defined as following:

$$Y_{ij} = -\frac{\partial W_e^d}{\partial d_{ij}} \implies$$

$$Y_{11} = \frac{1}{2}C_{11}^0(\varepsilon_{11}^e)^2 \quad ; \quad Y_{22} = \frac{1}{2}C_{22}^0(\varepsilon_{22}^e)^2 \quad ; \quad Y_{12} = \frac{1}{2}G_{12}^0(2\varepsilon_{12}^e)^2 \tag{5}$$

These associated thermodynamic forces characterize the damage propagation. The state of damage can only grow [1, 2] and therefore, the threshold of undamaged zone is defined as a maximal thermodynamic force for all previous time ($\tau$) up to the current time ($t$) [5]:

$$\overline{Y}_{ij} = sup_{\tau \le t}\left(Y_{ij}(t)\right), \{i,j = 1,2\} \tag{6}$$

The damage variables $d_{ij}, \{i,j = 1,2\}$ represent a loss of material stiffness in different orthotropic directions. In shear the damage variable is defined from the shear modulus diminution during experiment:

$$d_{12} = 1 - \frac{G_{12}^i}{G_{12}^0} \tag{7}$$

where $G_{12}^i$ is the current shear modulus associated to each unloading-loading.

Damage evaluation law is chosen as the best approximation of experimental data. Different types of functions can be used such as linear, polynomial, logarithmic, Heaviside function.

## 2.2 Plasticity modelling and damage-plasticity coupling

The experimental data shows irreversible strains appearance mainly in shear [3, 4, 5, 6]. Thus, plastic flow is considered to be blocked in fiber directions:

$$\varepsilon_{11}^p = \varepsilon_{22}^p = 0 \quad ; \quad \varepsilon_{12}^p \neq 0 \tag{8}$$

The damage and plasticity coupling is made using the effective stress notation (9).

$$\tilde{\sigma}_{12} = \frac{\sigma_{12}}{(1-d_{12})} \tag{9}$$

The isotropic strain hardening is assumed. The elastic domain is defined by the yield function $f$:

$$f = \frac{|\sigma_{12}|}{(1 - d_{12})} - R(p) - R_0 \tag{10}$$

where $R_0$ is a yield stress and the function $R(p)$ is a material characteristic function of the cumulative plastic strain $p$. Generally, the hardening function $R(p)$ is approximated by a power law:

$$R = \beta p^k \text{ with } p = \int_0^{\varepsilon_{12}^p} (1 - d_{12}) d\varepsilon_{12}^p \tag{11}$$

where $\beta$ and $k$ are material parameters identified from the experimental data.

### 2.3 Strain-rate sensitivity

Polymer-matrix composite materials have strain rate dependence especially in the shear direction [5, 19]. As the strain rate increases it becomes important to modify the shear modulus and the yield stress values accordingly. To describe the material response on dynamic loading the model developed by [4] is used. The material parameters are modified using following relations:

$$G_{12}^{0\ dyn} = G_{12}^{0\ static}(1 + F^G(\dot{\varepsilon}, \dot{\varepsilon}_0)) \tag{12}$$

$$R_0^{dyn} = R_0^{static}(1 + F^R(\dot{\varepsilon}, \dot{\varepsilon}_0)) \tag{13}$$

where $F^G(\dot{\varepsilon}, \dot{\varepsilon}_0)$, $F^R(\dot{\varepsilon}, \dot{\varepsilon}_0)$ functions are taken into account for the modification of shear modulus and the yield stress. These functions are dependent on the current strain rate $\dot{\varepsilon}$ and on the "threshold" strain rate $\dot{\varepsilon}_0$ from which the strain rate affects the composite behavior. Functions $F^G$, $F^R$ are determined from the experimental data using traction tests at different strain rates. They can be expressed using different forms such as linear, polynomial, power, etc.

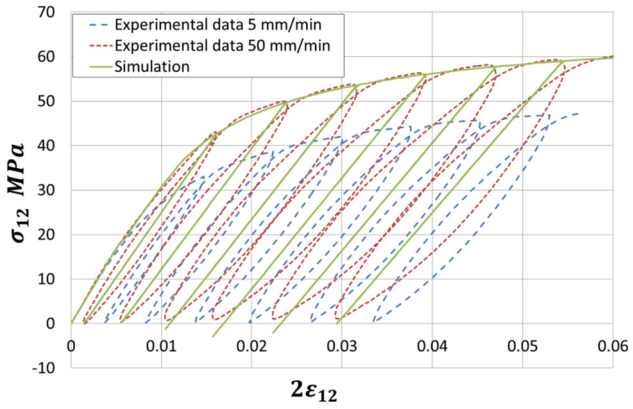



Figure 3. Numerical and experimental stress-strain curve comparison at dynamic loading

Damage parameters should now be recalculated from a new shear modulus value. The material parameter identification procedure is described in the next sections. Thus we can obtain the stress-strain curve (Figure 3) for the woven carbon/epoxy composite for a strain rate of 50 mm/min using this method by using the experimental data for a strain rate of 5 mm/min. The first sub-model represents elastoplastic damage behavior of composite for different strain rates. The simulation results are found to be in good agreement with the experimental data. The next step is to adjoin the hysteresis loops modelling to the current model.

## 2.4 Fractional derivative model

The hysteresis loops are associated with energy dissipation of composite under cyclic loading. Hysteresis is a hereditary phenomenon, i.e. the previous loading history has to be taken into account. To describe this viscoelastic behavior, fractional derivatives are introduced in the constitutive equation. The Riemann-Liouville fractional derivative [20] is defined as following:

$$D^{\alpha}f(x) = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_{0}^{x}\frac{f(t)}{(x-t)^{\alpha}}dt, \quad 0 < \alpha < 1 \tag{14}$$

where $D^{\alpha}$ is a fractional derivative of order $\alpha$ and $\Gamma$ is the Gamma-function defined by:

$$\Gamma(z) = \int_{0}^{+\infty} e^{-x}x^{z-1}dx, \quad z \in \mathbb{R}_{+}^{*} \tag{15}$$

According to elastoplastic damage model during unloading the plastic strain stays constant and the elastic strain is a linear function of time. However, a non-linearity of strain is observed in the experimental curve in the unloading/reloading path (Figure 4).

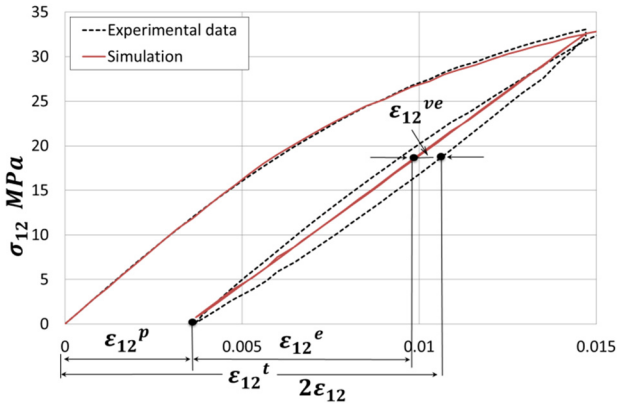



Figure 4. Total strain composition

In order to represent strain non-linearity, fractional derivatives are introduced in the governing law in the second sub-model. The total strain within the hysteresis loop $\varepsilon_{12}^{FD}$ is defined as following:

$$\varepsilon_{12}^{FD} = A + 2BD^{\alpha}\varepsilon_{12}^{e}(t) \tag{16}$$

where $\varepsilon_{12}^{e}$ is the elastic strain determined by the elastoplastic damaged model, $D^{\alpha}$ is the Riemann-Liouville fractional derivative (14) and $A$, $B$ and $\alpha$ are fractional model parameters.

As the plastic flow stays constant, the stress is expressed by the elastic law:

$$\sigma_{12}(t) = G_{12}^{0}(1-d_{12})\,\varepsilon_{12}^{FD}(t) \tag{17}$$

By substitution the equation (16) in the (17), the constative law is:

$$\sigma_{12}(t) = G_{12}^{0}(1-d_{12})A + 2G_{12}^{1}D^{\alpha}\varepsilon_{12}^{e}(t) \tag{18}$$

with $G_{12}^{1} = G_{12}^{0}(1-d_{12})B$.

### 2.5 Collaboration of two sub-models

The collaboration between the models is performed automatically, depending on the sign of the yield function (10) and its derivative. If $f = 0$ and $\dot{f} = 0$, the elastoplastic damage model is used. During unloading if $f < 0$ or $f = 0$ and $\dot{f} < 0$ and during reloading if $f < 0$ or $f = 0$ and $\dot{f} > 0$, we consider that the damage and plastic strain stays constant, and thus we switch to the fractional derivative model.

## 3 PARAMETERS IDENTIFICATION

In the following section, the identification procedure of material parameters is described. The suggested methodology of experimental identification is applied to unidirectional carbon/epoxy composite under different strain rates of loading.

### 3.1 Material characterization in fiber directions

In the fiber direction composite material has a linear brittle response. Damage of material is instantaneous and is described by the Heaviside function (19) (Figure 5). The parameters are hence easily identified and their values are presented in the Table 1 and Table 2.

$$d_{ii} = \mathcal{H}\left(\sqrt{\bar{Y}_{ii}} - \sqrt{\bar{Y}_{ii}^R}\right) \text{ if } d_{ii} < 1 \text{ and } \bar{Y}_{ii} < \bar{Y}_{ii}^R; \text{ otherwise } d_{ii} = 1, \{i = 1,2\} \qquad (19)$$

where $\bar{Y}_{ii}^R$ is a failure-damage threshold.

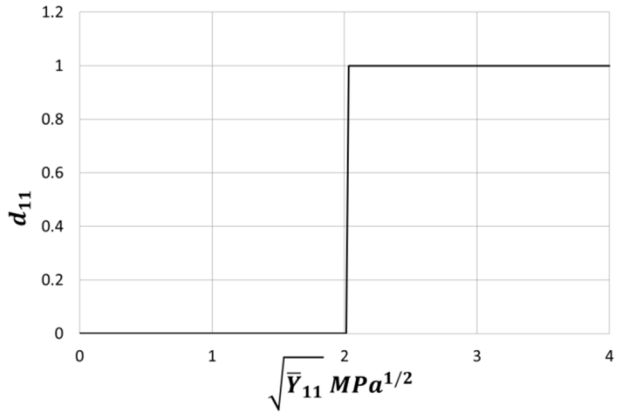



Figure 5. Evaluation of the longitudinal damage function for the carbon/epoxy composite

| Parameter | Value |
|---|---|
| $E_{11}^0 = E_{22}^0$ | 57.3 GPa |
| $\nu_{12}^0 = \nu_{21}^0$ | 0.07 |
| $\sigma_{11}^R = \sigma_{22}^R$ | 687 MPa |

Table 1. Longitudinal (transverse) identification for the carbon/epoxy composite

| Parameter | Value |
|---|---|
| $\sqrt{\bar{Y}_{11}^0} = \sqrt{\bar{Y}_{11}^R}$ | 2 $MPa^{1/2}$ |

Table 2. Longitudinal (transverse) damage parameters for the carbon/epoxy composite

### 3.2 Material characterization in shear

From the shear test (Figure 2) we had observed the strain rate sensitivity even for low values: 5 mm/min and 50 mm/min. The identified elastic parameters are presented in the Table 3. Significant difference between yield stresses for two different strain-rates can be observed.

| Parameter | Value for 5 mm/min | Value for 50 mm/min |
|---|---|---|
| $G_{12}^0$ | 3.36 GPa | 3.37 GPa |
| $R_0$ | 12 MPa | 20.2 MPa |
| $\sigma_{12}^R$ | 80 MPa | 94 MPa |

Table 3. Elastic parameters in shear for the carbon/epoxy composite

The damage propagation is identified, using the linear law (20) (Figure 6), for the low strain-rate of 5 mm/min. Using these damage parameters which are identified, the cumulative plastic strain $p$ is obtained. The strain hardening function $R(p)$ is fitted by the power law (11) on the cumulative plastic strain $p$ (Figure 7). The material parameters are presented in the Table 4.

$$d_{12} = \frac{\sqrt{\bar{Y}_{12}} - \sqrt{\bar{Y}_{12}^0}}{\sqrt{\bar{Y}_{12}^c}} \text{ if } d_{12} < 1 \text{ and } \bar{Y}_{12} < \bar{Y}_{12}^R; \text{ otherwise } d_{12} = 1 \tag{20}$$

where $\bar{Y}_{12}^0$ is the initial damage threshold, $\bar{Y}_{12}^R$ is the failure-damage threshold and $\bar{Y}_{12}^c$ is the speed of damage propagation.

| Parameter | Value for 5 mm/min |
|---|---|
| $\sqrt{\bar{Y}_{12}^C}$ | 2.25 $MPa^{1/2}$ |
| $\sqrt{\bar{Y}_{12}^0}$ | 0.15 $MPa^{1/2}$ |
| $\sqrt{\bar{Y}_{12}^R}$ | 0.8 $MPa^{1/2}$ |
| $\beta$ | 266.6 MPa |
| $k$ | 0.36 |

Table 4. Shear damage and plasticity parameters for the carbon/epoxy composite

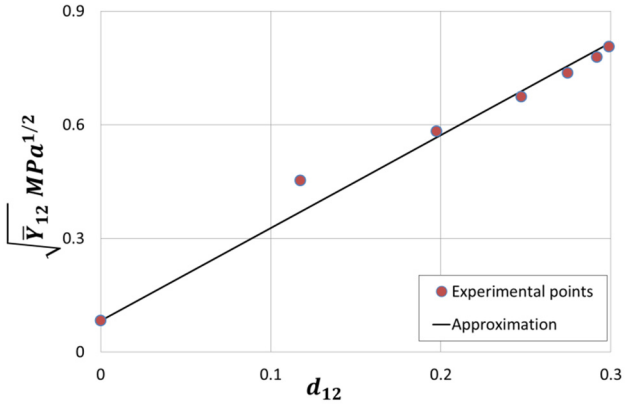



Figure 6. Shear damage function of the carbon/epoxy composite

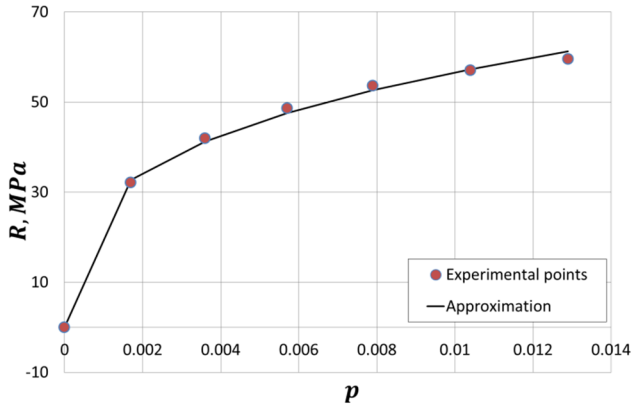



Figure 7. Hardening function $R(p)$ for the carbon/epoxy composite

### 3.3 Identification of fractional derivative model parameters

In order to determine fractional derivative model parameters, an optimization problem has been resolved. Due to specification of fractional operators, the differentiable function must be vanished at the first point of calculus. Otherwise, the fractional operator will tend to infinity. To avoid this singularity, zero initial conditions are required for the elastic strain function $\varepsilon_{12}^{e}$. That's why we include in the computational interval the "preloading" path where $\varepsilon_{12}^{e} = 0$. Once the fractional derivatives are calculated on "preload-unload-reload" time-interval, the optimal solution is found within the "unloading-reloading" time-interval or within the hysteresis loop (

Figure 8). The objective function is expressed by relative error $\delta$:

$$\delta = \frac{\sqrt{\sum_{i=1}^{N}(\varepsilon_{12}^{test}(t_i) - \varepsilon_{12}^{FD}(t_i))^2}}{N} \tag{21}$$

where $\varepsilon_{12}^{test}$ is the experimental strain determined by the equation (15), $\varepsilon_{12}^{FD}$ is the strain calculated by the fractional model (22) and $N$ is the number of time-points inside the considering interval.

$$\varepsilon_{12}^{test} = \frac{\sigma_{12}^{test}}{G_{12}^{0}(1 - d_{12})} \tag{22}$$

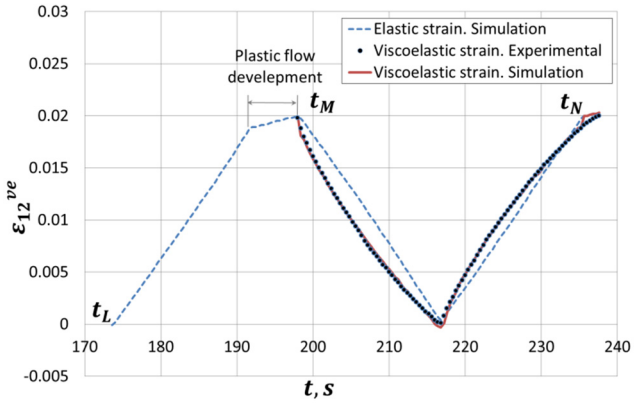



Figure 8. Elastic strain referring to a hysteresis loop

To implement the fractional derivative in a numerical code the *M1*-method [18] is used. The Riemann-Liouville fractional integral of elastic deformation $\varepsilon_{12}^e$ can be expressed in alternative form as:

$$(I_0^{1-\alpha}\varepsilon_{12}^e(t))_{M1} = \frac{t^{1-\alpha}}{\Gamma(\alpha)}\int_0^1 \varepsilon_{12}^e\left(t\left(1-v^{\frac{1}{1-\alpha}}\right)\right)dv \tag{23}$$

The fractional derivative is calculated using central difference scheme:

$$(D_0^{\alpha}\varepsilon_{12}^e(t))_{M1} = \frac{(I_0^{1-\alpha}\varepsilon_{12}^e(t+\Delta t)_{M1} - (I_0^{1-\alpha}\varepsilon_{12}^e(t-\Delta t))_{M1}}{2\Delta t} \tag{24}$$

The *M1*-method provides numerical error less than 1%. It can be easily implemented in the numerical code if analytical expression of function $\varepsilon_{12}^e$ is known. Thus the elastic strain $\varepsilon_{12}^e$ is approximated by a piecewise function in further calculations.

The fractional model parameters *A, B,* and $\alpha$ are determined by resolving an optimization problem for each loop. Their values stay constant within one hysteresis loop but they are varied loop by loop. We consider *A, B,* and $\alpha$ as a function of damage as the damage is constant within one hysteresis loop. The parameters *A, B,* and $\alpha$ can be approximated by linear function of damage, for the carbon/epoxy composite during a shear test with a strain rate of 5 mm/min, as the following:

$$A = m_A d_{12} + b_A \tag{25}$$

$$B = m_B d_{12} + b_B \tag{26}$$

$$\alpha = m_\alpha d_{12} + b_\alpha \tag{27}$$

Approximations are illustrated on the Figure 9, Figure 10 and Figure 11. Coefficients of fittings (25), (26) and (27) are presented in the Table 5.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| $m_A$ | 0.0189 | $b_A$ | -0.0013 |
| $m_B$ | 1.795 | $b_B$ | 0.8605 |
| $m_\alpha$ | 0.6944 | $b_\alpha$ | -0.0035 |

Table 5. Coefficients of fractional parameters fitting for the carbon/epoxy composite

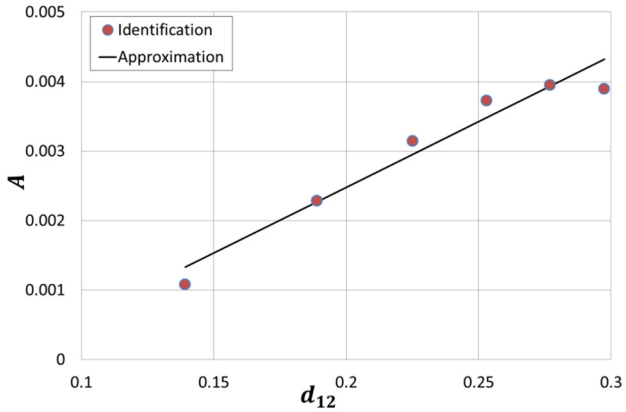



Figure 9. Fractional parameter *A* evaluation with damage for the carbon/epoxy composite

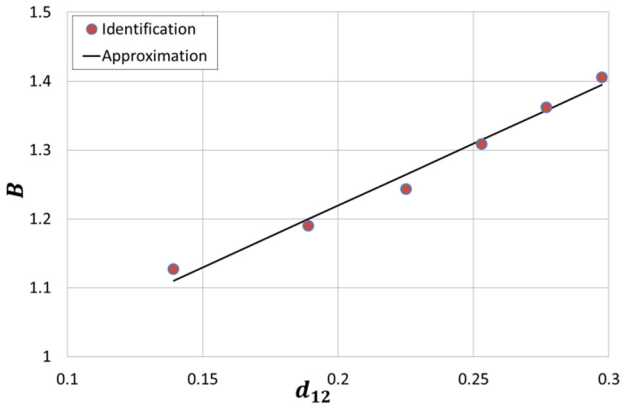



Figure 10. Fractional parameter $B$ evaluation with damage for the carbon/epoxy composite

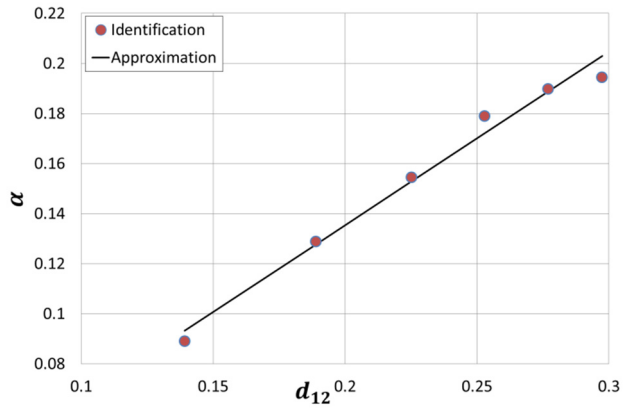



Figure 11. Fractional parameter $\alpha$ evaluation with damage for the carbon/epoxy composite

## 4 RESULTS

Taking into account the previous assumptions and the identified material parameters, the simulation of stress-strain curves is performed. The brittle elastic behavior in fiber directions is represented exactly in the Figure 12. The shear curve at 5 mm/min is represented by collaborative model in the Figure 13. The numerical simulation is in a good agreement with the experimental data. The in-elastic strains, damage and hysteresis loops are taken into account during the numerical simulation.

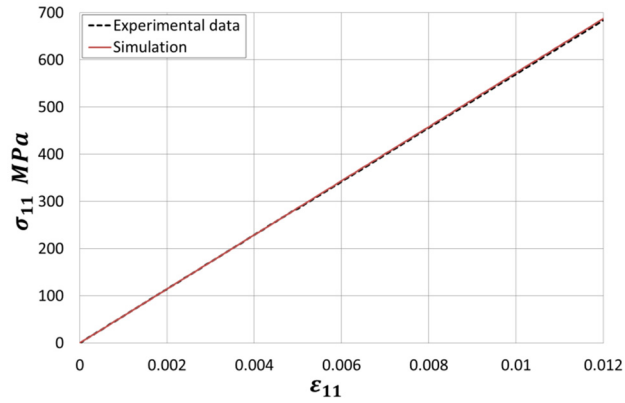



Figure 12. Experimental and numerical behavior comparison in longitudinal direction for thermoset composite

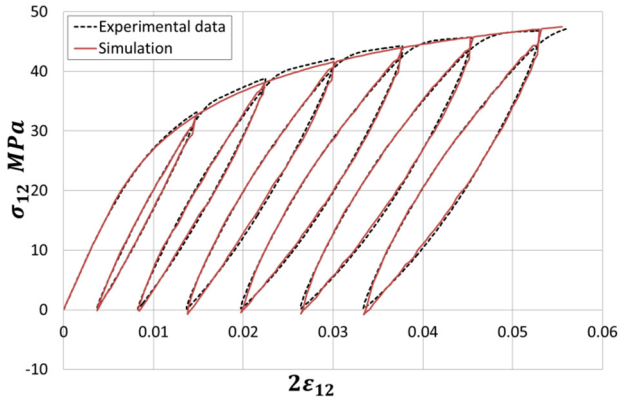



Figure 13. Experimental and numerical behavior comparison in shear for carbon/epoxy composite for the strain-rate of 5 mm/min

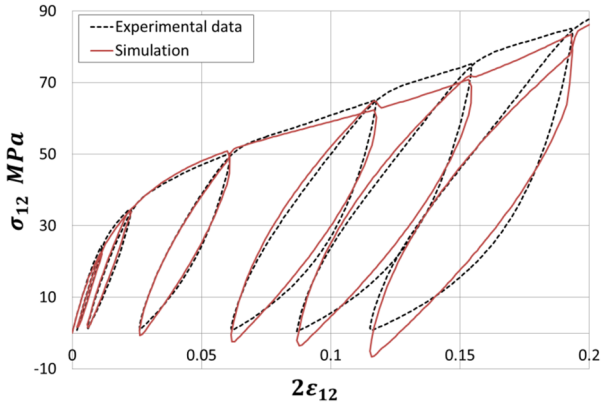



Figure 14. Experimental and numerical behavior comparison in shear for thermoplastic composite

The next simulation concerns the strain-rate sensitivity of carbon epoxy woven composite which is observed from the experimental data (Figure 2). To take this into account, the material parameters are identified for strain rates of 5 mm/min and 50 mm/min. The evolutions of the parameters are represented by the linear laws. The material response in shear for a strain rate of 50 mm/min is simulated by the collaborative model from the 5 mm/min shear test. The model is able to reproduce hysteresis loops from the quasi-static test at low strain rates. It is a promising approach to model the hysteresis loops at different strain rates and to signify damage propagation.

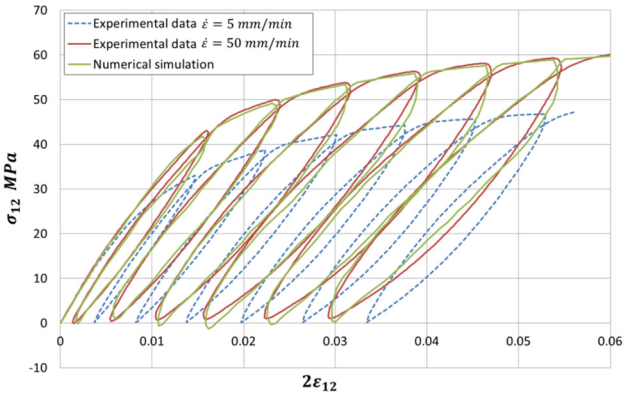



Figure 15. Stress-strain dynamic response for a woven thermoset composite

## 5 CONCLUSIONS

The proposed collaborative model takes into account the damage, plastic strain and viscoelastic effects such as hysteresis loops and strain rate sensitivity. The hysteresis behavior is modeled by a fractional derivative approach. Few parameters are required to represent the hysteresis loops. These parameters are determined by resolving an optimization problem. The simple implementation of *M1*-method is proposed for fractional derivatives. The model is validated for thermoset and thermoplastic carbon fiber woven composite materials. The strain rate sensitivity is demonstrated on the example of carbon/epoxy composite material at low strain rates.

This constitutive model is able to describe the behavior of woven composite materials under cyclic quasi-static and dynamic loading. One of the main advantages of the proposed model is that the elastoplastic damage model [2] and strain rate dependent model [4] are classical and are completed by the fractional derivative model. The numerical implementation is simple and the computational cost is low. The collaborative model is a promising approach to quantify and correlate the material damage evolution in respect of the strain rate.

## REFERENCES


[1] P. Ladevèze, Sur la mécanique de l'endommagement des composites. *Journées Nationales sur les Composites (JNC5)*, Paris, France, 1986.

[2] P. Ladevèze, E. Le Dantec, Damage modelling of the elementary ply for laminated composites. *Composites Science and Technology,* **43**, 257-267, 1992.

[3] C. Hochard, J. Payan, O. Montagnier, Design and computation of laminated composite structures. *Composites Science and Technology,* **65**, 467–474, 2005.

[4] P. Rozycki, *Contribution au développement de lois de comportement pour matériaux composites soumis à l'impact*. PhD thesis, Valenciennes , 2000.

[5] A. Johnson, A. Pickett, P. Rozycki, Computational methods for predicting impact damage in composite structures. *Composites Science and Technology,* **61** (15), 2183-2192, 2001.

[6] S. Marguet, P. Rozycki, L. Gornet, A rate dependence constitutive model for carbon-fiber reinforced plastic woven fabrics. *Mechanics of Advanced Materials and Structures,* **14** (8), 619-631, 2007.

[7] V. V. Uchaikin, *Metod Drobnyh Proizvodnyh*. Ulianovsk: Artishok, 2008.

[8] M. Caputo, F. Mainardi, A new dissipation model based on memory mechanism. *Pure and Applied Geographic,* **91** (134), 134-147, 1971.

[9] R. Bagley, P. Torvik, On the fractional calculus model of viscoelastic behavior. *Rheology,* **30** (11), 133-155, 1986.

[10] R. Bagley, P. Torvik, A theoretical basis for the application of fractional calculus to viscoelasticity. *Rheology,* **27**, 1–10, 1983.

[11] J. D. Ferry, R. F. Landel, M. Williams, Extensions of the Rouse theory of viscoelastic properties to undiluted linear polymers. *Applied Physics,* **26** (4), 359-362, 1955.

[12] R. E. Rouse, The theory of the linear viscoelastic properties of dilute solutions of coiling polymers. *Chemical Physics,* **21**, 1272–1280, 1953.

[13] Y. N. Rabotnov, Ravnovesie uprugoj sredi s posledstviem. *Prikladnaya Matematika i Mechanika,* **12**, 53-62, 1948.

[14] Y. N. Rabotnov, *Elementy Naledstvennoj Mehaniki Tverdyh Tel*. Moscow: Nauka, 1977.

[15] M. Caputo, J. Carcione, Hysteresis cycles and fatige criteria using anelastic models based on fractional derivatives. *Rheologica Acta,* **50**, 107-115, 2011.
[16] M. Mateos, L. Gornet, H. Zabala, L. Aretxabaleta, P. Rozycki, Hysteretic behaviour of fibre-reinforced composites. *15th European Conference on Composite Materials (ECCM15)*, Venice, Italy, 201.
[17] M. H. Mateos, *Hysteretic and damage behaviour modelling of composite materials by fractional operators*. PhD thesis, Nantes, 2014.
[18] A. Krasnobrizha, *Modélisation des mécanismes d'hystérésis des composites tissésà l'aide d'un modèle collaboratif élasto-plastique endommageable à dérivées fractionnaires*. PhD thesis, Nantes, 2015.
[19] A. Gilat, R. K. Goldberg, G. R. Roberts, Experimental study of strain-rate-dependent behavior of carbon/epoxy composite. *Composite Science and Technology,* **62**, 1469-1476, 2002.
[20] S. G. Samko, A. A. Kilbas, O. I. Marichev, *Fractional Integrals and Derivatives: Theory and Applications*. Amsterdam: Gordon and Breach Science, 1993.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NUMERICAL SIMULATION OF MECHANICAL BEHAVIOR OF COMPOSITE SANDWICH PANELS WITH DEFECTS

**A.N. Anoshkin[1], V.Yu. Zuiko[2], M.A. Alikin[2] and A.V. Tchugaynova[2]**

[1]Department of Mechanics of Composite Materials and Constructions, Perm National Research Polytechnic University, 29, Komsomolski Ave., Perm, 614990, Russian Federation
anoshkin@pstu.ru

[2]Scientific & Educational Center for Aviation Composite Technologies, Perm National Research Polytechnic University, 29, Komsomolski Ave., Perm, 614990, Russian Federation
zuyko81@gmail.com

**Keywords:** Composite Sandwich Panel, Defect, Finite Element Analysis, Carbon Fiber Reinforced Plastic (CFRP), Fiberglass, ANSYS.

**Abstract.** *This work presents a methodology based on the numerical simulation for assessment of the effects of defects in the form of delamination on the mechanical properties of composite sandwich panels. Various types of core are used in these panels depending on their purpose - tubular, cellular or honeycomb, arranged in one, two or even three layers. High-quality bonding of core and skins is one of the key factors for providing of overall stiffness and structural strength. Often constructions with multilayer sandwich panels may work under the complex stress condition and the local areas of disbond between core and skins can considerably reduce bearing strength of the product. Also, defects in the form of delamination can be observed in multi-ply skins of sandwich panels. In addition, some skins have perforation which occupies up to 10% of the entire area, and perforation holes diameter is correlated with the ply thickness. This factor should also be considered in assessing of performance of the sandwich panels.*

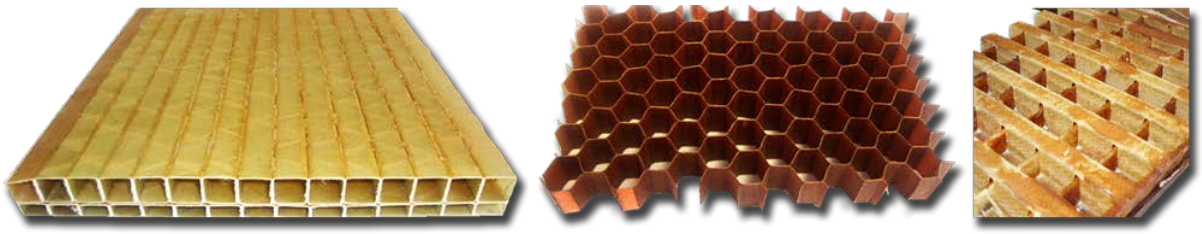
*The three-dimensional parametric models of panels with tubular and honeycomb cores were created. Both solid and shell models were used. Simulation of delamination was performed by contact parameters changing in local areas of predetermined shape and dimensions. Such materials as carbon fiber reinforced plastic (CFRP) and fiberglass were examined. The numerical stress-strain analysis of panels under various loading condition was obtained with ANSYS Workbench software.*

*The developed numerical models allow to study the effects of delamination defects on the stress and strain fields in multilayer composite sandwich panels according to the size and location of the defect.*

## 1. INTRODUCTION

Today sandwich panels made of fibrous polymer composite materials are vital elements of many modern aircrafts and turbojet engines. Sandwich structures consist of a lightweight thick core and two thin strong skins bonded by thin adhesive films. These structures have high specific stiffness, sound and thermal insulating properties [1].

Various types of core are used in these panels depending on their purpose - tubular, cellular or honeycomb, arranged in one, two or even three layers (Fig.1). High-quality bonding of core and skins is one of the key factors for providing of overall stiffness and structural strength. Often constructions with multilayer sandwich panels may work under the complex stress condition and the local areas of disbond between core and skins can considerably reduce bearing strength of the product. Also, defects in the form of delamination can be observed in multi-ply skins of sandwich panels.



a) b) c)

Figure1: Types of core: tubular (a), honeycomb (b) and cellular (c).

The most researches that are devoted to mechanical behavior of sandwich panels describe the honeycomb core structures, for example [2-7]. A comprehensive review of the computational models on honeycomb sandwiches was given by Noor et al. [8], where numerous references were cited.

Present work continues the previous study of the mechanical properties of composite sandwich panels with various core types [9-13]. The goal of this research is the development of methodology for analysis of effects of delamination on the mechanical properties of composite sandwich panels.

It may be noted that the created numerical models and the results of this work will be used for fundamental investigation devoted to the development of experimental and theoretical principles of the mechanical analysis of smart materials with embedded fiber Bragg grating (FBG) sensors for structural health monitoring (SHM) [14-16].

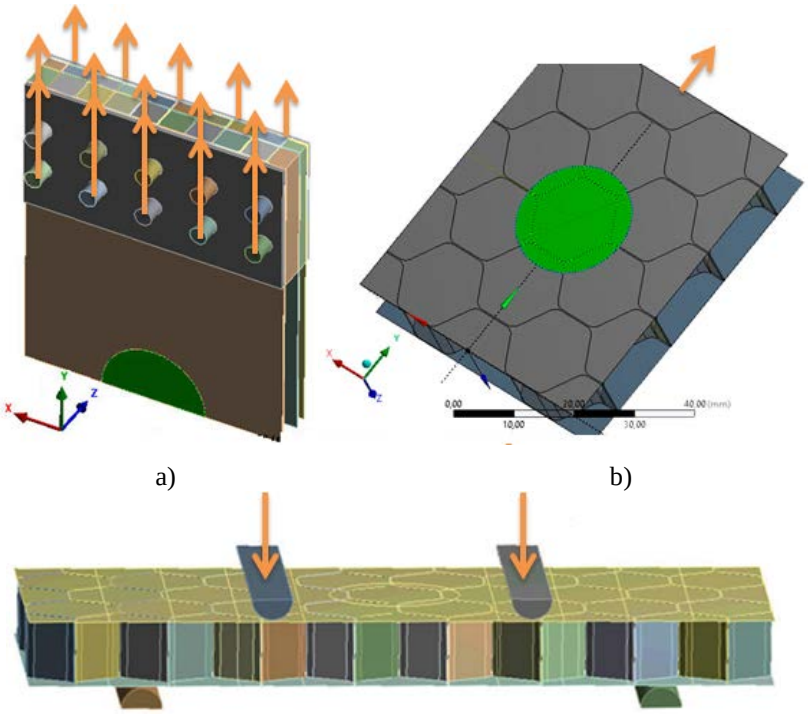
## 2. STRUCTURES AND MODELS OF SANDWICH PANELS.

The numerical analysis of stress-strain state of sandwich panel under various conditions of loading was obtained with ANSYS Workbench software. For this simulation the 3D parametric models of sandwich panels corresponding to full-scale samples used for mechanical tests were created [9-11]. The describing of test technique and results can be found in [12, 13]. Initial mechanical properties and strength characteristics of the materials used in numerical studies were obtained in in-house tests.

The following simulation cases were examined: 1 – the longitudinal tension of sandwich panel with tubular core made of fiberglass, 2 – the longitudinal tension and 4-point bending of sandwich panel with carbon skins and fiberglass honeycomb core.

Simulation of delamination between the core and skins was performed by contact parameters changing in local areas of predetermined shape and dimensions. Thus the areas in form of circle with radius of 10 mm, 15 mm and 30 mm were searched.

Fig.2 shows the 3D models of panels with delamination area located in the center. For model of panel with tubular core (Fig.2, a) thanks to symmetry of the modelled specimens, simulations can be reduced by investigation of their halves only. Gripping fixtures, used for installing a sample in a testing machine, were also modelled in order to reproduce a character of load application in real experiments. The fixture system consisted of steel pads, a gasket, a seal and bolts.

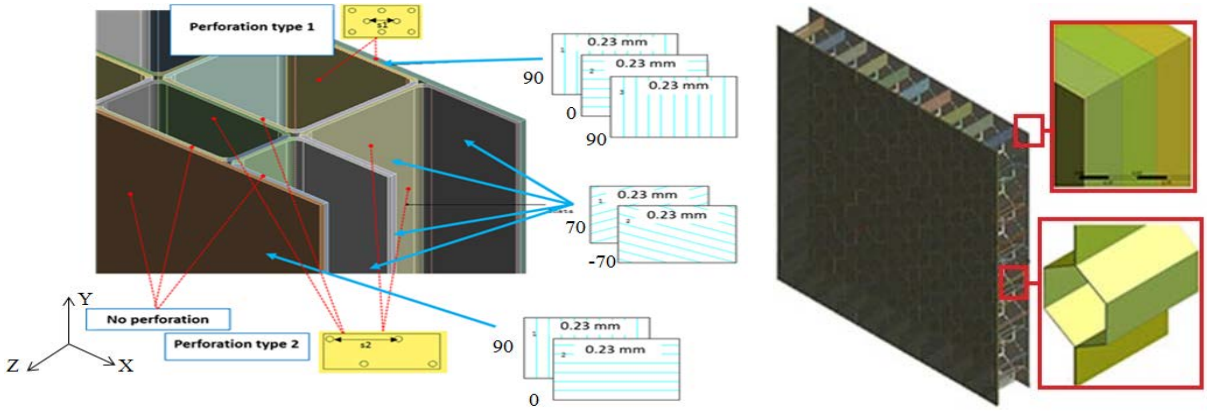


a) b)

c)

Figure 2: 3D models of sandwich panels with delamination area: longitudinal tension of sandwich panel with tubular core (a), longitudinal tension (b) and 4-point bending (c) of sandwich panel with honeycomb core

The layouts of internal structure of sandwich panels are shown on Fig.3. For example, the studied sandwich panel with tubular core consists of three-ply 90/0/90 and two-ply 90/0 laminate skins and tubes with 70/-70 laminate faces.




a) b)

Figure 3. Layout of internal structure of sandwich panel: with tubular core (a), with honeycomb core (b)

Fig.4 shows the FE meshes of sandwich panels for various simulation cases. Both solid and shell models were used for panel with honeycomb core to analyze the possible influence of

different approaches of FEM technique on the results. Meshing was automatically implemented in ANSYS Workbench using tetrahedral Solid186 or Shell281 finite elements. The mesh density was considered optimal when the difference between the results of successive calculations with a refined mesh did not exceed 5 - 10%.

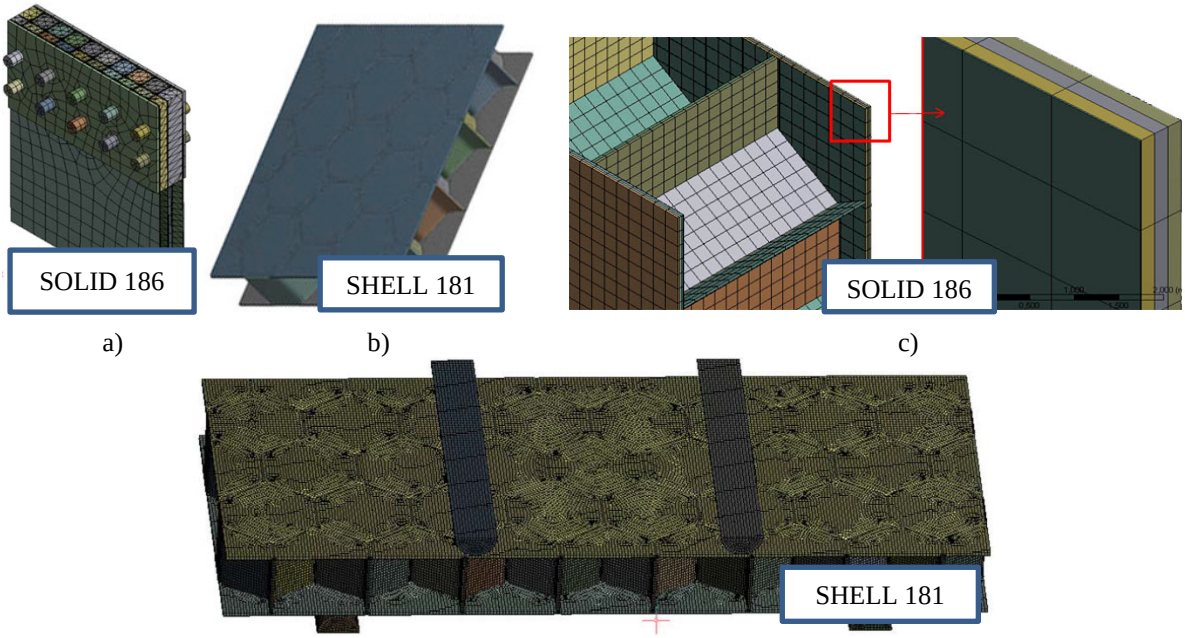



Figure 4: FE meshes of sandwich panels for various simulation cases

Some skins of panels have perforation which occupies up to 10% of the entire area, and perforation holes diameter is correlated with the ply thickness. The direct modelling of such structural elements is connected with computational difficulties due to a large number of small-sized curved areas (holes) that require a fine FE mesh. Thus it was necessary to obtain the effective properties of each perforated laminate. Various reinforcement schemes and two types of perforation were considered. The first type of perforation is dense perforation with a distance between holes is about 9 mm. Perforation of the second type is sparse, with the same diameter of holes and distance of about 18 mm between them.

As perforation was periodic, representative volume elements (RVE) of the laminates could be considered for assessment of their effective properties. Their dimensions were chosen as 60 mm × 60 mm to include a significant number of perforation holes (Fig.5).

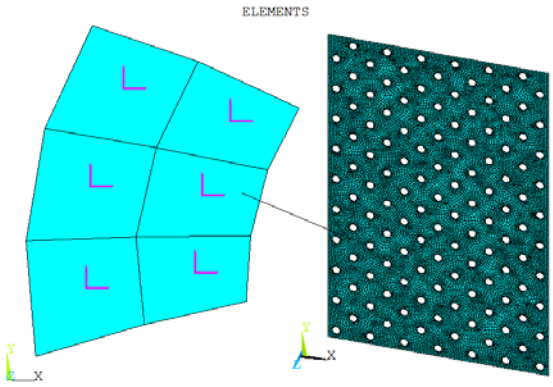



Figure 5: FE mesh of representative volume elements (RVE) of the laminates.

The considered laminates were orthotropic, characterized by nine independent elastic constants: the Young's moduli $E_X$, $E_Y$, $E_Z$, Poisson's ratios $\nu_{XY}$, $\nu_{YZ}$, $\nu_{XZ}$, and shear moduli $G_{XY}$, $G_{YZ}$, $G_{XZ}$. Three types of numerical experiments were implemented for each RVE of the perforated laminates: tension along axes X and Y to determine the effective Young's moduli and Poisson's ratios, and pure shear in the XY plane to calculate the shear modulus. In each case the laminate RVE model was subjected to a kinematic loading condition of 1 mm displacement on the corresponding border.

The calculations of effective elastic and strength characteristics of perforated layered shells were obtained with ANSYS Mechanical software employing Shell281 finite elements.

The homogenized constants of the perforated laminates – the elastic moduli and Poisson's coefficient – were obtained with the following equations:

$$\begin{gathered} E_x^* = \frac{\langle \sigma_x \rangle}{\varepsilon_x^*}, \quad E_y^* = \frac{\langle \sigma_y \rangle}{\varepsilon_y^*}, \quad E_z^* = \frac{\langle \sigma_z \rangle}{\varepsilon_z^*} \\ G_{xy}^* = \frac{\langle \tau_{xy} \rangle}{\gamma_{xy}^*}, \quad G_{yz}^* = \frac{\langle \tau_{yz} \rangle}{\gamma_{yz}^*}, \quad G_{xz}^* = \frac{\langle \tau_{xz} \rangle}{\gamma_{xz}^*} \\ \nu_{xy}^* = \left| \frac{\varepsilon_y^*}{\varepsilon_x^*} \right|, \quad \nu_{yz}^* = \left| \frac{\varepsilon_z^*}{\varepsilon_y^*} \right|, \quad \nu_{xz}^* = \left| \frac{\varepsilon_z^*}{\varepsilon_x^*} \right| \end{gathered} \tag{1}$$

where strains $\varepsilon_x^*$, $\varepsilon_y^*$, $\varepsilon_z^*$ and $\gamma_{ij}^*$ were predetermined in the numerical experiments.

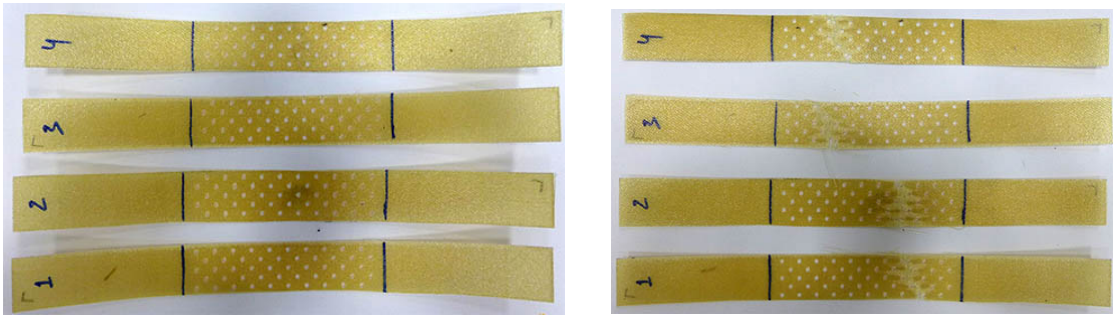
The mean stress values were calculated using the following relation:

$$\langle \sigma_i \rangle = \frac{\sum_{i=1}^{n} \sigma_i V_i}{V_M} \tag{2}$$

where $n$ is number of finite elements in the RVE, $\sigma_i$ is stress in the $i^{th}$ finite element, $V_i$ is its volume, $V_M$ is the representative volume.

It was shown that the perforation leads to decrease of material stiffness up to 30%, and strength – up to 40-60%.

Verification of the proposed numerical algorithm for calculation of effective properties of the perforated laminates and obtained results were carried out with the help of tensile tests of special specimens of 2-plies laminates (Fig.6) according to ASTM D 882 [17].




a) b)

Figure 6: Perforated specimens of 2-plies laminate before (a) and after (b) tensile testing.

Static tests were conducted on Zwick/Roell Pro Line Z100 testing machine with contact-type long-travel extensometer. Tensile stress-strain curves for tested specimens are given in Fig.7. The results of numerical simulations were compared with the test data. Good correlation was found.

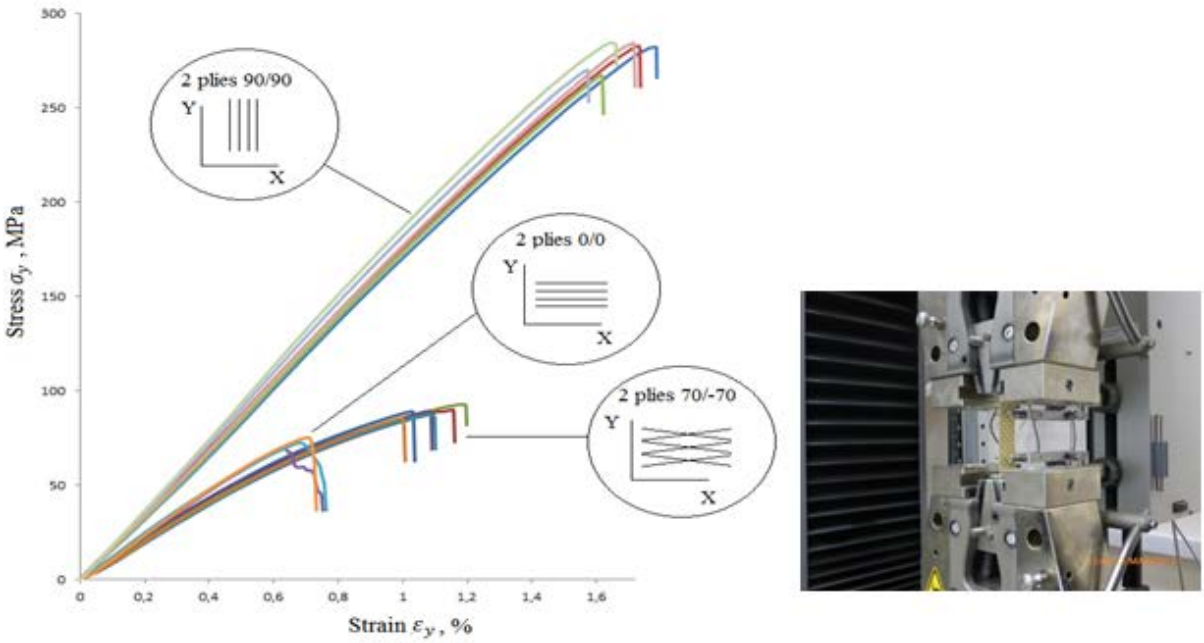



Figure7: Tensile stress-strain curves for perforated specimens with various reinforcement schemes and testing fixture with contact-type extensometer.

## 3. NUMERICAL MODELLING OF MECHANICAL BEHAVIOUR OF SANDWICH PANELS.

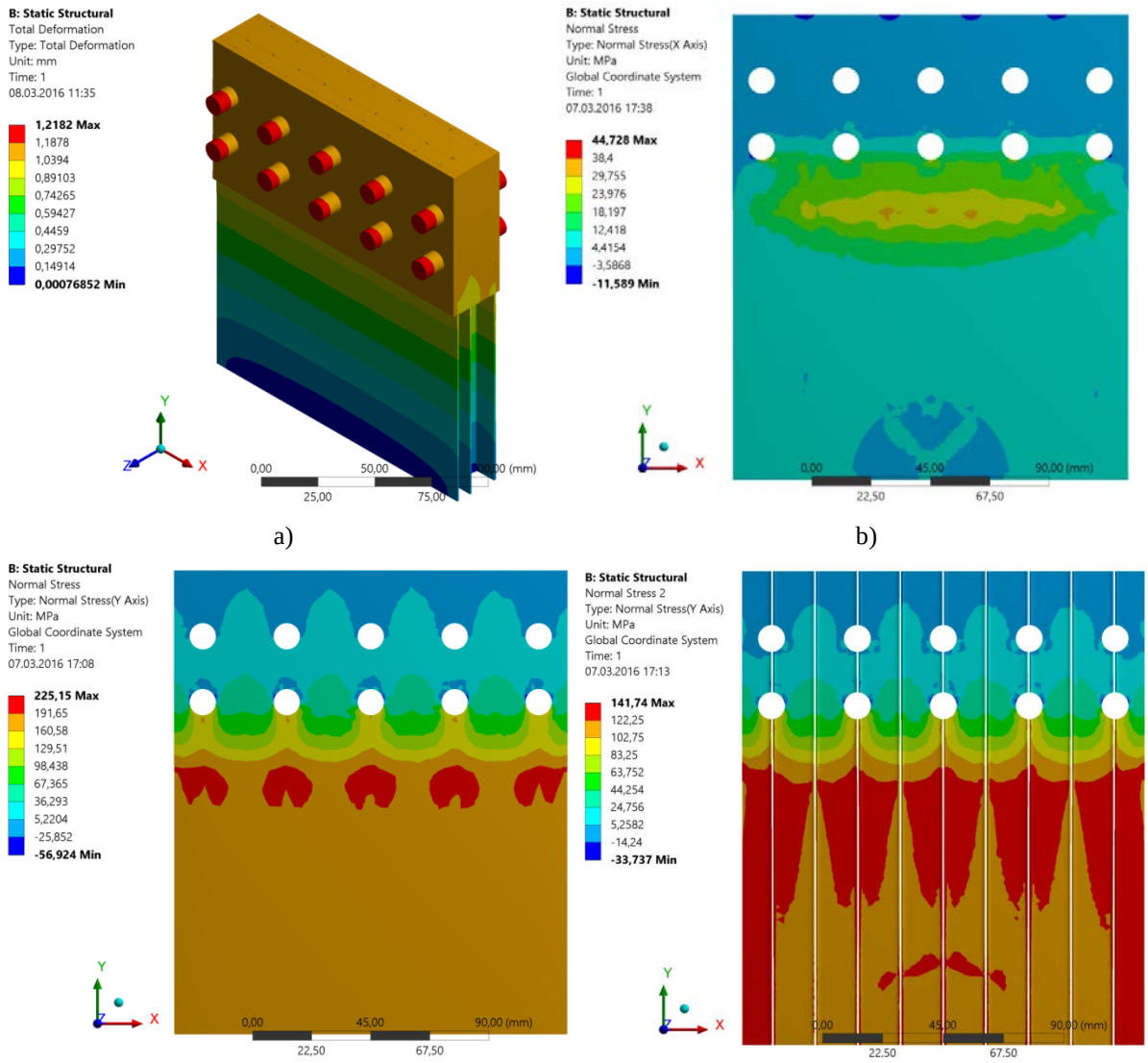
The numerical studies included development of finite-elements models of the considered sandwich panel specimens under tensile loading and 4-point bending. The problems were solved in a general statement for an anisotropic elastic body. A mathematical formulation was given in [10, 11].

As a result of the numerical simulation, stress and strain fields were obtained for each structure component of panels. For example, Fig.8 shows the calculated total displacements in the sample of panel with delamination defect under the load of 100 kN and fields of stress in skins and tubular core.

Similarly the stress analysis for sandwich panel with honeycomb core can be realized (Fig.9). The effective approach for research of mechanical behavior is the analysis of stress distribution along the different paths. Thus Fig.10 shows various paths on the top three-ply laminate skin of sandwich panel under 4-point bending load.

A comparison of obtained solution data for various delamination defects with the results for a defect-free panel allows to estimate the impact of defect on the static strength and stress-strain state of panel (Fig.11).

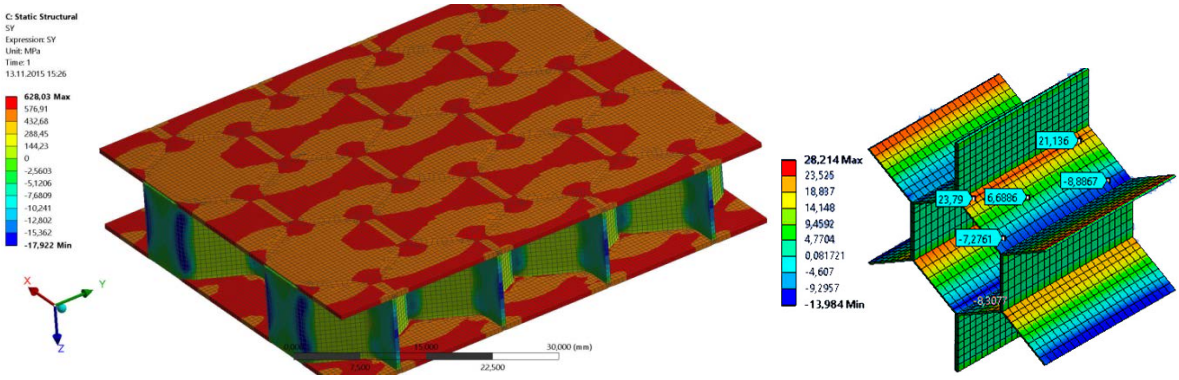
The results analysis shows that delamination of the core from facesheet invokes the stress redistribution near the defect area. The several sizes of delamination were considered and it was shown that in general there was no effect on static strength of sandwich panel with tubular core under tensile loading condition. The local changes of stress components are more noticeably for sandwich panels with honeycomb core due to specific type of contact boundary. Nevertheless the difference between stress values for defect-free and defective panel does not exceed 5%.




a) b)

c) d)

Figure 8: Calculated total displacements [mm] (a) and fields of stress [MPa] in skins (b, c) and tubular core (d) of sandwich panel with delamination defect.




a) b)

Figure 9: Calculated fields of SY stress [MPa] in sandwich panel with honeycomb core with delamination defect (a) and honeycomb cell (b).

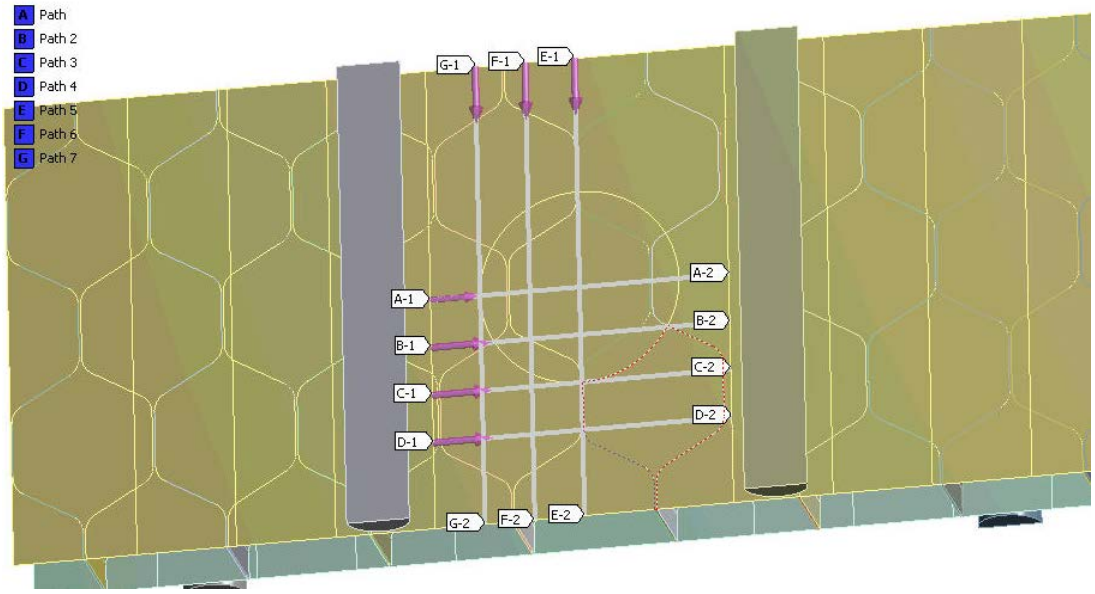



Figure 10: Paths for stress distribution analysis on the top three-ply laminate skin of sandwich panel under 4-point bending load.

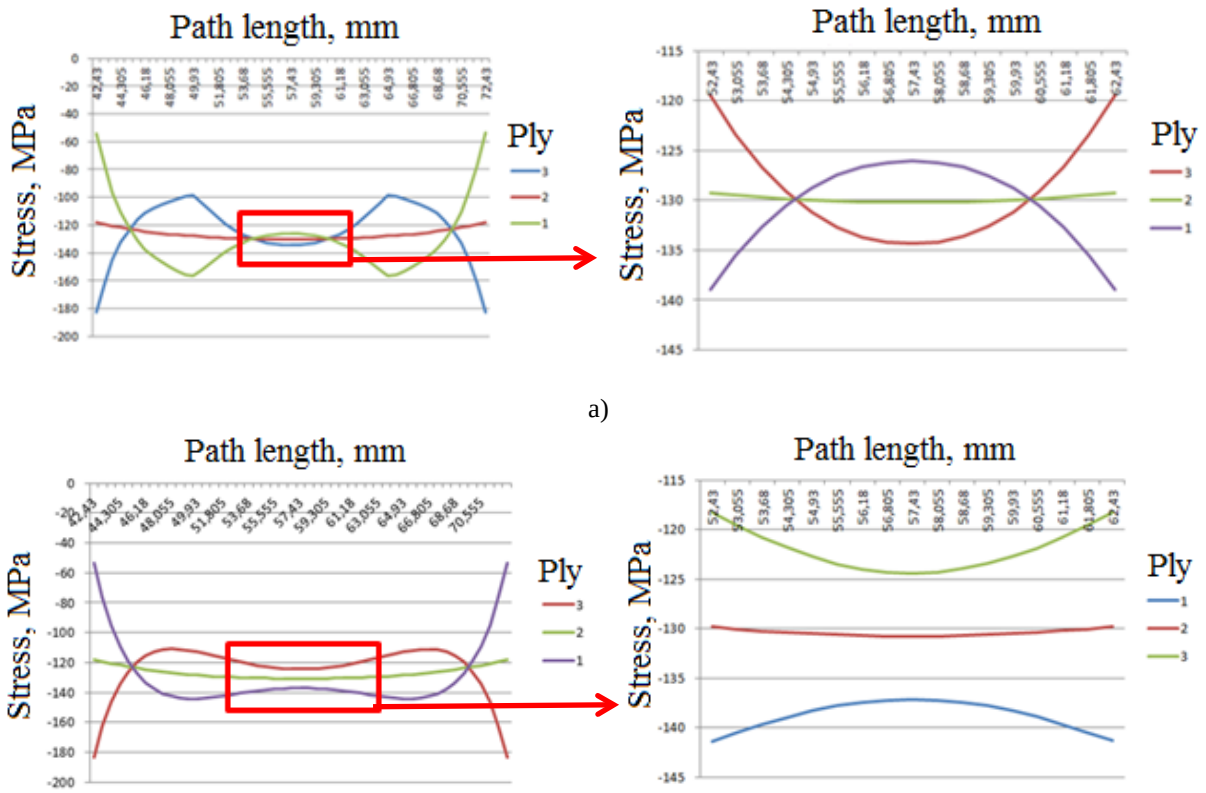



Figure 11: Bending stress distribution along the A1-A2 path: a) – defect-free panel, b) – panel with delamination defect (radius 10 mm).

## 4. CONCLUSION

This work is devoted to development of a methodology based on the numerical simulation for assessment of the effects of defects in the form of delamination on the mechanical properties of composite sandwich panels. Various types of core are used in these panels depending on their purpose - tubular, cellular or honeycomb. Several studies have shown that

delamination of the core from facesheet is a typical failure mode for sandwich panels. In addition, some skins have perforation which occupies up to 10% of the entire area, and perforation holes diameter is correlated with the ply thickness. This factor should also be considered in assessing of performance of the sandwich panels.

The three-dimensional parametric models of panels with tubular and honeycomb cores were created. Both solid and shell FE models were used. The numerical stress-strain analysis of panels under various loading condition was obtained with ANSYS Workbench software. The following simulation cases were examined: 1 – the longitudinal tension of sandwich panel with tubular core made of fiberglass, 2 – the longitudinal tension and 4-point bending of sandwich panel with carbon skins and fiberglass honeycomb core. Simulation of delamination was performed by contact parameters changing in local areas of predetermined shape and dimensions. A comparison of obtained solution data for various delamination defects with the results for a defect-free panel allows to estimate the impact of defect on the static strength and stress-strain state of panel.

The prior analysis has shown that in general there was no effect on static strength of sandwich panel with tubular core under tensile loading condition. The local changes of stress components depending on size of delamination are more noticeably for sandwich panels with honeycomb core due to specific type of contact boundary. Nevertheless the difference between stress values for defect-free and defective panel does not exceed 5%. However it is significant to note that the problem of delamination in composite sandwich panels is extremely important for fatigue life prediction for such structures. The solution of this problem will be the aim of the further researches.

The analysis of effect of perforation on mechanical properties of fiberglass sheets was also performed in this work. It was shown that the perforation leads to decrease of material stiffness up to 30%, and strength – up to 40-60%. Verification of the proposed numerical algorithm and obtained results was carried out with the help of tensile tests of special specimens according to ASTM D 882.

It can be concluded that the developed numerical models allow to study the effects of delamination defects on the stress and strain fields in multilayer composite sandwich panels according to the size and location of the defect. The results of this work can be used for structural health monitoring (SHM) of these structures with embedded fiber Bragg grating (FBG) sensors. Thus the numerically obtained stress distribution near the possible defect can help to determine the dimension of effective sensitive zone for embedding FBG sensors.

The study was performed in Perm National Research Polytechnic University with support of the Russian Science Foundation (project №15-19-00243).

## REFERENCES

[1] Vinson J.R., The behavior of sandwich structures of isotropic and composite materials. *Technomic Publishing Company, Inc*., Lancaster, Pennsylvania, 1999.

[2] Venugopal M.M., Maharana S.K., Badarinarayan K.S., Finite element evaluation of composite sandwich panel under static four point bending load. *Journal of Engineering, Science and Technology Management*, **2**, 2013.

[3] Sayyidmousavi A., Malekzadeh K., Bougharar H., Element buckling analysis of laminated composite sandwich panels with transversely flexible core containing a face/core debond. *Journal of Composite materials*, **46**, 193-202, 2011.

[4] Chen D.H., Ozaki S., Analysis of in-plane elastic modulus for a hexagonal honeycomb core: effect of core height and proposed analytical method. *Composite Structures*, **88**, 17–25, 2009.

[5] Aktaya L., Johnson A.F., Numerical modeling of honeycomb core crush behavior. *Engineering Fracture Mechanics*, **75**, 2616–2630, 2008.
[6] Meraghni F., Desrumaux F., Benzeggagh M.L., Mechanical behaviour of cellular core for structural sandwich panels. *Composites Part A*, **30**, 767–779, 1999.
[7] Masters I.G., Evans K.E., Models for the elastic deformation of honeycombs, *Composite Structures*, **35**, 403–422, 1996.
[8] A.K. Noor, W.S. Burton, C.W. Bert, Computational models for sandwich panels and shells. *Applied Mechanics Review ASME*, **49 (3)**, 155–199, 1996.
[9] Anoshkin A.N., Zuiko V.Yu., Glezman A.V., Analysis of mechanical properties of composite sandwich panels with fillers. *16th European Conference for Composite Materials (ECCM16)*, Seville, Spain, June 22-26, 2014.
[10] Anoshkin A.N., Zuiko V.Yu., Tashkinov M.A., Silberschmidt V.V., Repair of damage in aircraft composite sound-absorbing panels. *Composite Structures*, **120**, 153–166, 2015.
[11] A.N. Anoshkin, V.Yu. Zuiko, A.V. Tchugaynova, E.N. Shustova, Experimental-theoretical research of mechanical properties of perforated composite sandwich panels. *Solid State Phenomena*, **243**, 1–10, 2016.
[12] Zuiko V.Yu., Lobanov D.S., Anoshkin A.N., Technique of experimental determination of ultimate strength of composite sandwich panels in tension, compression and shear static tests (in Russian). *PNRPU Mechanics Bulletin*, **2**, 99–111, 2012.
[13] Anoshkin A.N., Vildeman V.E., Lobanov D.S., Chikhachev A.I., Evaluation of repair efficiency in structures made of fibrous polymer composite materials. *Mechanics of Composite Materials*, **50**, 311-316, 2014.
[14] Minakuchi S., Okabe Y., Takeda N., Real-time detection of debonding between honeycomb core and facesheet using a small-diameter FBG sensor embedded in adhesive layer. *Journal of Sandwich Structures and Materials*, **9**, 9-33, 2007.
[15] Okabe Y., Minakuchi S., Shiraishi N., Murakami K., Takeda N., Smart honeycomb sandwich panels with damage detection and shape recovery functions. *Advanced Composite Materials*, **17**, 41-56, 2008.
[16] Ramlya R., Embedded FBG sensor in aircraft smart composite materials for structural monitoring. *Applied Mechanics and Materials*, **393**, 311-316, 2013.
[17] ASTM D 882: Standard Test Method for Tensile Properties of Thin Plastic Sheeting, 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# TIMOSHENKO BEAM ELEMENT WITH ANISOTROPIC CROSS-SECTIONAL PROPERTIES

**Alexander R. Stäblein and Morten H. Hansen**

[1]Technical University of Denmark, Department of Wind Energy,
Frederiksborgvej 399, DK-4000 Roskilde
e-mail: {alsta,mhha}@dtu.dk

**Keywords:** Anisotropic Beam Element, Composite Beams

**Abstract.** *Beam models are used for the aeroelastic time and frequency domain analysis of wind turbines due to their computational efficiency. Many current aeroelastic tools for the analysis of wind turbines rely on Timoshenko beam elements with classical cross-sectional properties (EA, EI, etc.). Those cross-sectional properties do not reflect the various couplings arising from the anisotropic behaviour of the blade material. A two-noded, three-dimensional Timoshenko beam element was therefore extended to allow for anisotropic cross-sectional properties. For an uncoupled beam, the resulting shape functions are identical to the original formulation. The new element was implemented into a co-rotational formulation and validated against natural frequencies and several static load cases of previous works.*

## 1 INTRODUCTION

Beam models are used for the aeroelastic time and frequency domain analysis of wind turbines due to their computational efficiency. Many current aeroelastic tools for the analysis of wind turbines rely on Timoshenko beam elements with classical cross-sectional properties (EA, EI, etc.). Those beam properties do not reflect the various couplings arising from the anisotropic behaviour of the blade material. The cross-sectional properties of anisotropic beams are commonly expressed in a $6 \times 6$ cross-section stiffness matrix. Theories for determining the cross-section stiffness matrix have been presented by e.g. Giavotto et al. [1] and Yu et al. [2]. The method by Giavotto et al. invokes the virtual work per unit beam length to obtain a linear system of second-order differential equations with constant coefficients that have a homogeneous and particular solution. The particular solution is used to determine the $6 \times 6$ cross-sectional stiffness matrix. The homogeneous solution is related to warping and is generally ignored. The method by Yu et al. is based on the variational-asymptotic method by Berdichevskii [3].

The anisotropic cross-sectional properties require a suitable beam element for the analysis. Ghiringhelli's [4] element formulation uses the cross-section compliance matrix and beam forces, which vary linearly along length, to obtain the element stiffness by principle of virtual forces. The two-noded element of Kim et al. [5] assumes polynomial shape functions of arbitrary order. The shape function coefficients are determined by minimizing the elastic energy of the beam while satisfying the boundary conditions. Both elements assume small nodal displacements and require a co-rotational or multi-body formulation for geometric nonlinear analysis. A beam element that directly permits large displacement analysis is the mixed variational formulation of Hodges [6].

This paper extends the two-noded, three-dimensional Timoshenko beam element by Bazoune et al. [7]. A cross-section constitutive relationship with a $6 \times 6$ stiffness matrix was introduced and the 14 coefficients of the polynomial shape functions were eliminated by two equilibrium equations of the shear force and bending moment relationship and 12 compatibility conditions of the nodal displacements at the element boundaries. With the displacements and rotations known along the beam, the element stiffness matrix was obtained by numerical integration along the element. For an uncoupled beam, where the cross-section stiffness matrix is diagonal, the present formulation is identical to the original formulation. The new element is implemented into a co-rotational formulation by Battini and Pacoste [8] to allow for large displacements and rotations. The anisotropic, co-rotational Timoshenko beam element is validated against natural frequencies and several static cases of previous works. A new test case with a coupled 45-degree bend cantilever is also proposed and compared to results obtained with the beam element by Kim et al. [5].

## 2 METHODS

In this section a Timoshenko beam formulation for the analysis of anisotropic beams is derived. The present element is an extension of the two-noded, three-dimensional Timoshenko formulation by Bazoune et al. [7] to allow for fully populated $6 \times 6$ cross-section stiffness matrices.

### 2.1 Kinematic Assumptions

The element coordinate system has its origin at the first node of the element. The beam axis $x$ is along the length of the beam, pointing towards the second node. Axes $y$

and $z$ define the cross-sectional plane of the beam. The lateral displacements $u, v$ and $w$ of the beam axis are expressed as a function of the cross-sectional coordinate $x$ along the element length $L$. A first order polynomial is assumed for displacement $u$ along the beam axis and third order polynomials are assumed for displacements $v$ and $w$ in the cross-sectional plane.

$$u(x) = c_1 x + c_2 \tag{1}$$
$$v(x) = c_3 x^3 + c_4 x^2 + c_5 x + c_6 \tag{2}$$
$$w(x) = c_7 x^3 + c_8 x^2 + c_9 x + c_{10} \tag{3}$$

For torsional displacements along the beam a first order polynomial is assumed

$$\theta_x(x) = c_{11} x + c_{12} \tag{4}$$

The rotational displacements $\theta_y$ and $\theta_z$ around the beam cross section axes follow from Timoshenko's assumption that the curvature of the beam equals the slope plus a contribution from shear deformation

$$\theta_y(x) = -\frac{\partial w}{\partial x} + c_{13} \tag{5}$$
$$\theta_z(x) = \frac{\partial v}{\partial x} - c_{14} \tag{6}$$

To express the displacements and rotations along the beam, the shape function coefficients $c_k$ for $k \in \{1, \ldots, 14\}$ in the equations above have to be determined.

### 2.2 Constitutive Relations

By introducing the beam strain vector

$$\boldsymbol{\varepsilon} = \left\{ \frac{\partial u}{\partial x}, \frac{\partial v}{\partial x} - \theta_z, \frac{\partial w}{\partial x} + \theta_y, \frac{\partial \theta_x}{\partial x}, \frac{\partial \theta_y}{\partial x}, \frac{\partial \theta_z}{\partial x} \right\}^T \tag{7}$$

and the $6 \times 6$ cross-section stiffness matrix

$$\boldsymbol{K}_{cs} = \begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} & K_{15} & K_{16} \\ & K_{22} & K_{23} & K_{24} & K_{25} & K_{26} \\ & & K_{33} & K_{34} & K_{35} & K_{36} \\ & & & K_{44} & K_{45} & K_{46} \\ & \text{sym.} & & & K_{55} & K_{56} \\ & & & & & K_{66} \end{bmatrix} \tag{8}$$

the cross-section constitutive relation is

$$\boldsymbol{F} = \boldsymbol{K}_{cs}\, \boldsymbol{\varepsilon} \tag{9}$$

where $\boldsymbol{F} = \{F_x, F_y, F_z, M_x, M_y, M_z\}^T$ are the beam forces and moments in the cross-section.

### 2.3 Equilibrium and Compatibility

The 14 shape function coefficients $c_k$ are eliminated by introducing two equilibrium equations of the shear force and bending moment relationship

$$\frac{\partial M_y}{\partial x} - F_z = 0, \qquad \frac{\partial M_z}{\partial x} + F_y = 0 \tag{10}$$

and 12 compatibility conditions (6 nodal displacements + 6 nodal rotations) at the element boundaries $x = 0, L$

$$\begin{array}{rclrcl} u(0) &=& u_1 & u(L) &=& u_2 \\ v(0) &=& v_1 & v(L) &=& v_2 \\ w(0) &=& w_1 & w(L) &=& w_2 \\ \theta_x(0) &=& \theta_{x1} & \theta_x(L) &=& \theta_{x2} \\ \theta_y(0) &=& \theta_{y1} & \theta_y(L) &=& \theta_{y2} \\ \theta_z(0) &=& \theta_{z1} & \theta_z(L) &=& \theta_{z2} \end{array} \tag{11}$$

where $u_n, v_n, w_n$ and $\theta_{xn}, \theta_{yn}, \theta_{zn}$ for $n = 1, 2$ are the nodal displacements and rotations at the first and second node of the element. With the displacements known along the beam, the elastic energy is

$$V = \frac{1}{2} \int_0^L \boldsymbol{\varepsilon}^T \boldsymbol{K}_{cs} \boldsymbol{\varepsilon} \, dx \tag{12}$$

The element stiffness $\boldsymbol{K}_{el}$ is obtained by creating the Hessian of the elastic energy $V$ with respect to the nodal degrees of freedom.

### 2.4 Implementation

For implementation in a finite element code the beam element derived above is rewritten in matrix notation. The beam displacements and rotations $\boldsymbol{u}(x) = \{u, v, w, \theta_x, \theta_y, \theta_z\}^T$ of Equations (1) – (6) can be expressed as

$$\boldsymbol{u}(x) = \boldsymbol{A}(x)\boldsymbol{c} \tag{13}$$

where $\boldsymbol{A}(x)$ is the coefficient matrix of the displacements and rotations with respect to the shape function coefficient vector $\boldsymbol{c} = \{c_1, \ldots, c_{14}\}^T$. And similarly for their derivative with respect to the beam axis $\boldsymbol{du}(x) = \{\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}, \frac{\partial w}{\partial x}, \frac{\partial \theta_x}{\partial x}, \frac{\partial \theta_y}{\partial x}, \frac{\partial \theta_z}{\partial x}\}^T$

$$\boldsymbol{du}(x) = \boldsymbol{dA}(x)\boldsymbol{c} \tag{14}$$

The equilibrium and compatibility Equations (10) and (11) can be written as

$$\boldsymbol{E}(x)\boldsymbol{c} = \boldsymbol{T}\boldsymbol{d} \tag{15}$$

where $\boldsymbol{E}(x)$ is the coefficient matrix of the equilibrium and compatibility equations with respect to the shape function coefficient vector and

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{0}_{2\times 12} \\ \boldsymbol{I}_{12\times 12} \end{bmatrix} \tag{16}$$

is a transformation matrix in which $\underset{2\times12}{\mathbf{0}}$ is a $2 \times 12$ zero matrix and $\underset{12\times12}{\boldsymbol{I}}$ is a $12 \times 12$ identity matrix. The nodal displacements are expressed in the vector

$$\boldsymbol{d} = \{u_1, v_1, w_1, \theta_{x1}, \theta_{y1}, \theta_{z1}, u_2, v_2, w_2, \theta_{x2}, \theta_{y2}, \theta_{z2}\}^T \tag{17}$$

Substituting Equation (15) into (13) and (14), the beam displacements and their derivatives can be expressed as

$$\boldsymbol{u}(x) = \boldsymbol{N}(x)\boldsymbol{d} \tag{18}$$
$$\boldsymbol{du}(x) = \boldsymbol{dN}(x)\boldsymbol{d} \tag{19}$$

where $\boldsymbol{N}(x) = \boldsymbol{A}(x)\boldsymbol{E}(x)^{-1}\boldsymbol{T}$ and $\boldsymbol{dN}(x) = \boldsymbol{dA}(x)\boldsymbol{E}(x)^{-1}\boldsymbol{T}$. Finally, the beam strains $\boldsymbol{\varepsilon}$ are expressed in terms of nodal displacements and rotations

$$\boldsymbol{\varepsilon} = \boldsymbol{B}(x)\boldsymbol{d} = \left[\boldsymbol{dN}(x) + \boldsymbol{T}_N\boldsymbol{N}(x)\right]\boldsymbol{d} \tag{20}$$

where $\boldsymbol{B}(x) = \boldsymbol{dN}(x) + \boldsymbol{T}_N\boldsymbol{N}(x)$ is the strain displacement matrix and

$$\boldsymbol{T}_N = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{21}$$

a transformation matrix. The element stiffness matrix $\boldsymbol{K}_{el}$ of the anisotropic Timoshenko beam element is obtained by numerical integration over the beam length $L$

$$\boldsymbol{K}_{el} = \int_0^L \boldsymbol{B}(x)^T\boldsymbol{K}_{cs}\boldsymbol{B}(x)\, dx \tag{22}$$

A consistent mass matrix for the above element is obtained from

$$\boldsymbol{M}_{el} = \int_0^L \boldsymbol{N}(x)^T\boldsymbol{M}_{cs}\boldsymbol{N}(x)\, dx \tag{23}$$

where $\boldsymbol{M}_{cs}$ is the cross-section mass matrix containing mass and inertia of the cross-section with respect to the beam displacements and rotations.

## 3 RESULTS

The proposed element is validated by comparing the results of different test cases with those of previous publications. To allow for large displacements and rotations, the beam element above was combined with an implementation of the co-rotational formulation proposed by Battini and Pacoste [8]. The system was solved using a Newton-Raphson procedure.

### 3.1 Eigenfrequencies of a coupled cantilever

In the first example, the natural frequencies of a coupled cantilever box beam proposed by Hodges et al. [9] were investigated. The beam is 2.54 m long, has a height of 16.76 mm (0.66 in) and a width of 33.53 mm (1.32 in). The wall thickness is 0.84 mm (0.033 in) with six layers of unidirectional lamina stacked $(20/-70/20/-70/-70/20)$ from outside

to inside. The material is T 300 / 5208 Graphite / Epoxy with properties provided by Stemple and Lee [10]. The material density is given by Hodges et al. as 1604 kg/m$^3$ ($1.501 \cdot 10^{-4}$ lbsec$^2$/in$^4$). The cross-section stiffness matrix was taken from Hodges et al. and converted to SI units

$$\boldsymbol{K}_{cs} = \begin{bmatrix} 5.0576 \cdot 10^6 & 0 & 0 & -1.7196 \cdot 10^4 & 0 & 0 \\ & 7.7444 \cdot 10^5 & 0 & 0 & 8.3270 \cdot 10^3 & 0 \\ & & 2.9558 \cdot 10^5 & 0 & 0 & 9.0670 \cdot 10^3 \\ & & & 1.5041 \cdot 10^2 & 0 & 0 \\ & \text{sym.} & & & 2.4577 \cdot 10^2 & 0 \\ & & & & & 7.4529 \cdot 10^2 \end{bmatrix} \tag{24}$$

The cantilever was discretised with 16 elements. In Table 1 the results of the present model are compared with beam models by Hodges et al. [9] and Armanios and Badir [11] as well as a finite element shell model by Kim et al. [5].

### 3.2 Tip displacements and rotations of a coupled cantilever

A test case for the static analysis of a coupled cantilever was taken from Wang et al. [12]. The stiffness matrix is provided in the original study as

$$\boldsymbol{K}_{cs} = \begin{bmatrix} 1368.17 & 0 & 0 & 0 & 0 & 0 \\ & 88.56 & 0 & 0 & 0 & 0 \\ & & 38.78 & 0 & 0 & 0 \\ & & & 16.96 & 17.61 & -0.351 \\ & \text{sym.} & & & 59.12 & -0.370 \\ & & & & & 141.47 \end{bmatrix} \cdot 10^3 \tag{25}$$

The beam has a length of 10 m and was discretised by 10 elements. A tip load of 150 N was applied to the cantilever. The tip displacements and rotations (in Wiener-Milenkovic Parameter) are shown in Table 2.

### 3.3 Curvature and twist of a coupled cantilever

Chandra et al. [13] conduct experiments on box beams with different layups. The cross-section has a dimension of $13.6 \times 24.2$ mm ($0.537 \times 0.953$ inch) with a wall thickness of (6 plies) 7.6 mm (0.03 inch). The length of the cantilever is 76.2 cm (30 inch). The symmetric layup with $(45)^6$ in the flanges and $(45/-45)^3$ in the webs under a tip load of 4.448 N (1 lb.) was chosen for comparison. As the material properties of the AS4/3501-6 Unidirectional Graphite/Epoxy are incomplete in the original publication, the following

| Mode | Freq. [Hz] | | | | Rel. Diff. [%] | | |
|---|---|---|---|---|---|---|---|
| | Present | Hodges | Armanios | Kim | Hodges | Armanios | Kim |
| 1 vert. | 2.94 | 3.00 | 2.96 | 2.98 | 2.0 | 0.7 | 1.4 |
| 1 horiz. | 5.07 | 5.19 | 5.10 | 5.12 | 2.4 | 0.6 | 1.0 |
| 2 vert. | 18.38 | 19.04 | 18.54 | 18.65 | 3.6 | 0.9 | 1.5 |
| 2 horiz. | 31.72 | 32.88 | 31.98 | 32.02 | 3.7 | 0.8 | 0.9 |
| 3 vert. | 51.37 | 54.65 | 51.92 | 52.17 | 6.4 | 1.1 | 1.6 |
| 3 horiz. | 88.43 | 93.39 | 89.55 | 93.39 | 5.6 | 1.3 | 5.6 |
| 1 tors. | 180.10 | 180.32 | 177.05 | - | 0.1 | -1.7 | - |
| 2 tors. | 542.04 | 544.47 | 531.15 | - | 0.4 | -2.0 | - |

Table 1: Eigenfrequencies of a coupled cantilever obtained with the present model compared to results by Hodges et al. [9], Armanios and Badir [11] (both beam models) and Kim et al. [5] (FEM model).

| | $u_1$ | $u_2$ | $u_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ |
|---|---|---|---|---|---|---|
| Present | -0.09013 | -0.06320 | 1.22950 | 0.18447 | -0.17987 | 0.00523 |
| Wang | -0.09064 | -0.06484 | 1.22998 | 0.18445 | -0.17985 | 0.00488 |
| Rel. Diff. [%] | 0.57 | 2.59 | 0.04 | -0.01 | -0.01 | -6.77 |

Table 2: Tip displacements and rotations (in Wiener-Milenkovic Parameter) of a coupled cantilever obtained with the present model compared to results by Wang et al. [12].

properties were assumed for this study

$$\begin{aligned} E_{11} &= 142 \text{ GPa} & E_{22} &= 9.81 \text{ GPa} \\ G_{12} &= 6.00 \text{ GPa} & G_{23} &= 3.77 \text{ GPa} \\ \nu_{12} &= 0.30 & \nu_{23} &= 0.42 \end{aligned} \tag{26}$$

The cross-sectional properties were determined using BECAS, an implementation of the theory by Giavotto et al. [1], as

$$\boldsymbol{K}_{cs} = \begin{bmatrix} 11.387 \cdot 10^5 & 2.909 \cdot 10^5 & 0 & -30.458 & 12.674 & 0 \\ & 4.189 \cdot 10^5 & 0 & -11.932 & 8.689 & 0 \\ & & 3.122 \cdot 10^5 & 0 & 0 & 12.302 \\ & & & 62.692 & -21.741 & 0 \\ & \text{sym.} & & & 35.146 & 0 \\ & & & & & 80.594 \end{bmatrix} \tag{27}$$

Smith and Chopra [14] compare the experimental results by Chandra et al. to an analytical formulation and a finite element beam model. Figures 1 and 2 show the slope and twist along the beam obtained with the present model, the experimental results by Chandra et al., and the beam model used by Smith and Chopra. The experimental and beam model data was obtained by digitizing the plots of Smith and Chopra.

### 3.4 Pre-bend cantilever

This example illustrates a truly three-dimensional response and was initially presented by Bathe and Bolourchi [15]. It comprises a 45° bend cantilever with a radius of 100 m as shown in Figure 3. A square unit cross section with a modulus of elasticity of $10^7$ N/m$^2$ was used. Bend-twist coupling was introduced by setting $K_{45} = -0.3\sqrt{K_{44}K_{55}}$ of the cross-section stiffness matrix. With a tip load of 300 N, the solution converges with 5 iterations on average. Table 3 shows the tip displacement of the uncoupled beam compared to results by Simo & Vu-Quoc [16]. And the coupled beam compared to results obtained with the element proposed by Kim et al. which was implemented in the co-rotational formulation used in the present study.

## 4 DISCUSSION

A linear Timoshenko beam element with anisotropic cross-sectional properties was derived by extending an existing formulation. The beam model was implemented into a co-rotational formulation to allow for geometric nonlinear analysis and several test cases were analysed. A Python implementation of the proposed beam element, together with the element by Kim et al. and the co-rotational formulation by Battini and Pacoste is available on GitHub [1].

[1] https://github.com/alxrs/eccomas_2016.git

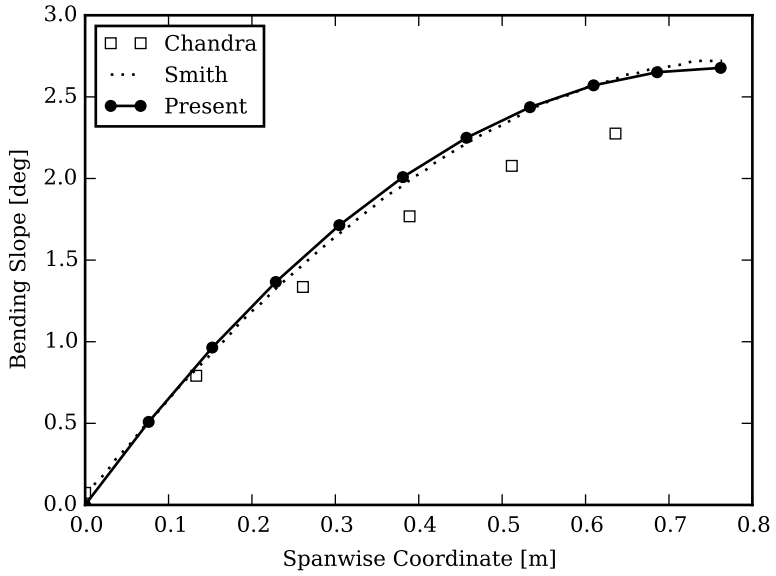



Figure 1: Comparison of slope along the beam obtained from experiments by Chandra et al. [13], beam model by Smith and Chopra [14], and present formulation.

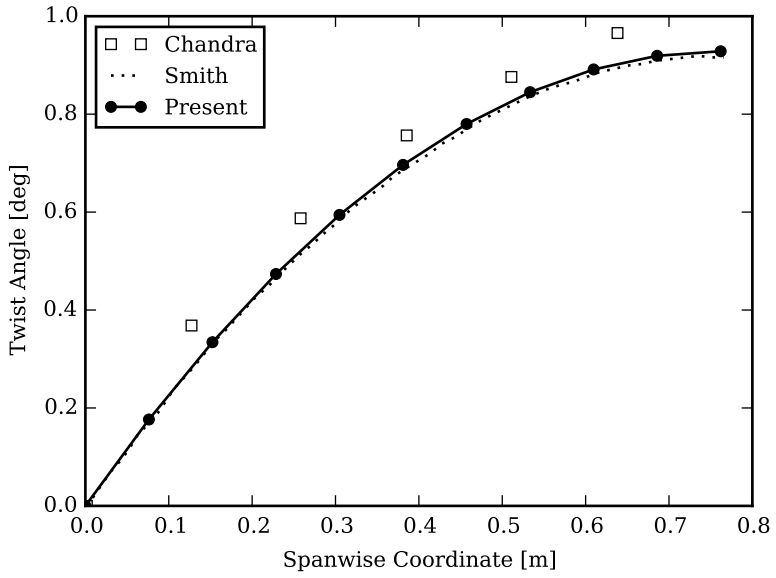



Figure 2: Comparison of twist along the beam obtained from experiments by Chandra et al. [13], beam model by Smith and Chopra [14], and present formulation.

| | Displacement [m] | | | Rel. Diff. [%] | | |
|---|---|---|---|---|---|---|
| | x | y | z | x | y | z |
| Simo & Vu-Quoc | -11.87 | -6.96 | 40.08 | - | - | - |
| Present uncpl. | -12.15 | -7.15 | 40.49 | 2.3 | 2.7 | 1.0 |
| Present cpl. | -10.66 | -6.53 | 38.68 | - | - | - |
| Kim et al. cpl. | -10.66 | -6.53 | 38.70 | 0.1 | 0.0 | 0.0 |

Table 3: Comparison of pre-bend cantilever tip displacements original and bend-twist coupled beam.

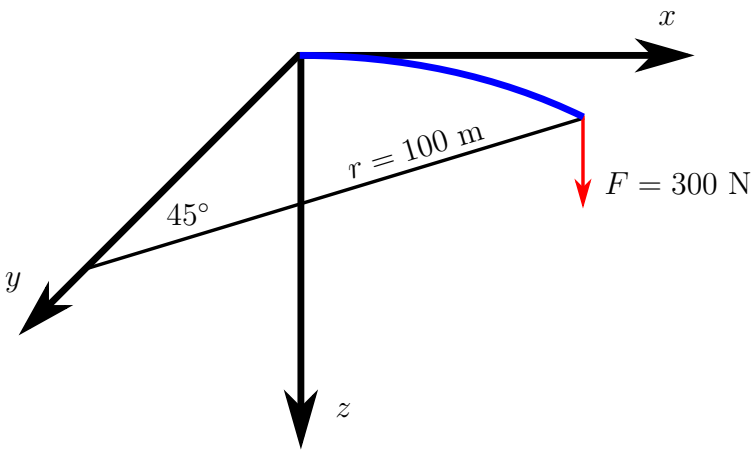



Figure 3: Prebend cantilever.

The present beam element is in good agreement with other formulations. The eigenvalues of a coupled cantilever beam of the first three vertical and horizontal modes and the first two torsional modes are within 2% of another beam formulation and within 6% of a finite element shell model. The tip displacements and rotations of a different coupled cantilever with a tip load are within 3% and 7% of the original study. The curvature and twist along a third cantilever is in good agreement with previous beam model results but deviates somewhat from experiments. The tip displacements of an uncoupled prebend cantilever are within 3% of previous studies. The tip displacements of a coupled prebend cantilever a nearly identical with the results obtained from a different beam element that has been implemented.

## ACKNOWLEDGEMENT

The present work is funded by the European Commission under the programme 'FP7-PEOPLE-2012-ITN Marie Curie Initial Training Networks' through the project 'MARE-WINT - new MAterials and REliability in offshore WINd Turbines technology', grant agreement no. 309395.

## REFERENCES

[1] V. Giavotto, M. Borri, P. Mantegazza, G. L. Ghiringhelli, V. Carmaschi, G. C. Maffioli, and F. Mussi. Anisotropic beam theory and applications. *Computers & Structures*, 16(1):403–413, 1983, DOI: 10.1016/0045-7949(83)90179-7.

[2] W. Yu, D. H. Hodges, V. Volovoi, and C. E. Cesnik. On timoshenko-like modeling of initially curved and twisted composite beams. *International Journal of Solids and Structures*, 39(19):5101–5121, 2002, DOI: 10.1016/S0020-7683(02)00399-2.

[3] V. L. Berdichevskii. Variational-asymptotic method of constructing a theory of shells. *Journal of Applied Mathematics and Mechanics*, 43(4):711–736, 1979, DOI: 10.1016/0021-8928(79)90157-6.

[4] G. L. Ghiringhelli. On the linear three-dimensional behaviour of composite beams. *Composites Part B: Engineering*, 28(5):613–626, 1997, DOI: 10.1016/S1359-8368(96)00075-3.

[5] T. Kim, A. M. Hansen, and K. Branner. Development of an anisotropic beam finite element for composite wind turbine blades in multibody system. *Renewable Energy*, 59:172–183, Nov. 2013, DOI: 10.1016/j.renene.2013.03.033.

[6] D. H. Hodges. A mixed variational formulation based on exact intrinsic equations for dynamics of moving beams. *International journal of solids and structures*, 26(11):1253–1273, 1990, DOI: 10.1016/0020-7683(90)90060-9.

[7] A. Bazoune, Y. A. Khulief, and N. G. Stephen. Shape functions of three-dimensional Timoshenko beam element. *Journal of Sound and Vibration*, 259(2):473–480, 2003, DOI: 10.1006/jsvi.2002.5122.

[8] J.-M. Battini and C. Pacoste. Co-rotational beam elements with warping effects in instability problems. *Computer Methods in Applied Mechanics and Engineering*, 191(17):1755–1789, 2002, DOI: 10.1016/S0045-7825(01)00352-8.

[9] D. H. Hodges, A. R. Atilgan, M. V. Fulton, and L. W. Rehfield. Free-vibration analysis of composite beams. *Journal of the American Helicopter Society*, 36(3):36–47, 1991, DOI: 10.4050/jahs.36.36.

[10] A. D. Stemple and S. W. Lee. Finite-element model for composite beams with arbitrary cross-sectional warping. *AIAA journal*, 26(12):1512–1520, 1988, DOI: 10.2514/3.48797.

[11] E. A. Armanios and A. M. Badir. Free vibration analysis of anisotropic thin-walled closed-section beams. *AIAA journal*, 33(10):1905–1910, 1995, DOI: 10.2514/3.12744.

[12] Q. Wang, M. Sprague, J. Jonkman, and N. Johnson. Nonlinear legendre spectral finite elements for wind turbine blade dynamics. In *Proceedings of the 32nd ASME Wind Energy Symposium, National Harbor, Maryland*, 2014, DOI: 10.2514/6.2014-1224.

[13] R. Chandra, A. Stemple, and I. Chopra. Thin-walled composite beams under bending, torsional, and extensional loads. *Journal of Aircraft*, 27(7):619–626, July 1990. WOS:A1990DU94000007, DOI: 10.2514/3.25331.

[14] E. C. Smith and I. Chopra. Formulation and evaluation of an analytical model for composite box-beams. *J. Am. Helicopter Soc.*, 36(3):23–35, 1991, DOI: 10.2514/6.1990-962.

[15] K.-J. Bathe and S. Bolourchi. Large displacement analysis of three-dimensional beam structures. *International Journal for Numerical Methods in Engineering*, 14(7):961–986, 1979, DOI: 10.1002/nme.1620140703.

[16] J.C. Simo and L. Vu_Quoc. A three-dimensional finite-strain rod model. Part II: Computational aspects. *Computer Methods in Applied Mechanics and Engineering*, 58:79–116, 1986, DOI: 10.1016/0045-7825(86)90079-4.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# STABILIZATION OF THE NUMERICAL PRIMING TO STABILIZATION PHENOMENON IN THE CRUSHING OF A COMPOSITE PANEL USING LOCAL AND DIFFUSE DAMAGE LAWS

**Christine Espinosa**[1*], **Floran Tostain**[1], **Samuel Rivallant**[1]

[1] Université de Toulouse – ISAE-SUPAERO - ICA/MS2M-MSC UMR CNRS 5312
3 Rue Caroline Aigle – 31400 Toulouse - France
e-mail: Christine.Espinosa@isae-supaero.fr*;

Floran.Tostain@isae.fr; Samuel.Rivallant@isae-supaero.fr

**Keywords:** Composite, Carbon Fibers, Crush, Crash, Edge Compression, Damage, Fragmentation, Splaying, Debris, Energy Absorption.

**Abstract.** *This study lies within the general scope of damage and rupture prediction during transverse crash (edge crushing) of a beveled composite plate using numerical simulation. Structures under consideration are aeronautical composite structures in purpose of design and certification. The present study proposes as an enhancement of an existing crushing model the introduction of a continuous softening law to represent the highly complex local fragmentation process at the vicinity of the crushing process on the ground. It is shown how the added softening stress strain law helps the numerical stability and the reproduction of all the processes in the different phases of the crushing process.*

# 1 INTRODUCTION

This work is devoted to the analysis of the link between the initiation phase of the crushing of a chamfered composite plate up to the peak force, and the stability of the crushing process in a steady phase of the process of energy absorption. The analysis aims to evaluate the effect of interlaminar strength resistance and the predominant ruin modes corresponding to different combinations of strength values. More precisely, we propose a strategy of modelling the structure that uses two behavior laws, in order to represent both the creation of new surfaces by delamination and rupture through plies induced by diffuse damage, and delamination induced by specific dust and chip formation at the vicinity to the ground. We are concerned here with CFRP composite plates crushed at about 9m/s, loading a heavy carriage at the ear face to represent the inertial effects of the structure to protect. Numerical simulations are compared with experimental results.

## 1.1 General purpose and position of the work

The Specific Energy Absorption (SEA) is the characteristics of a structure that represents its scalar capacity to absorb the of energy during either quasi-static or dynamic crushing processes. The SEA is defined as the ratio between the mean crushing force $F_{moy}$ by the specific mass of matter that is involved and in the crushing process. This mass is obtained by the product of the specific mass of the material $\rho_s$ and the involved area $A_s$. The energy is given by (1):

$$SEA = \frac{F_{moy}}{\rho_s \times A_s} \tag{1}$$

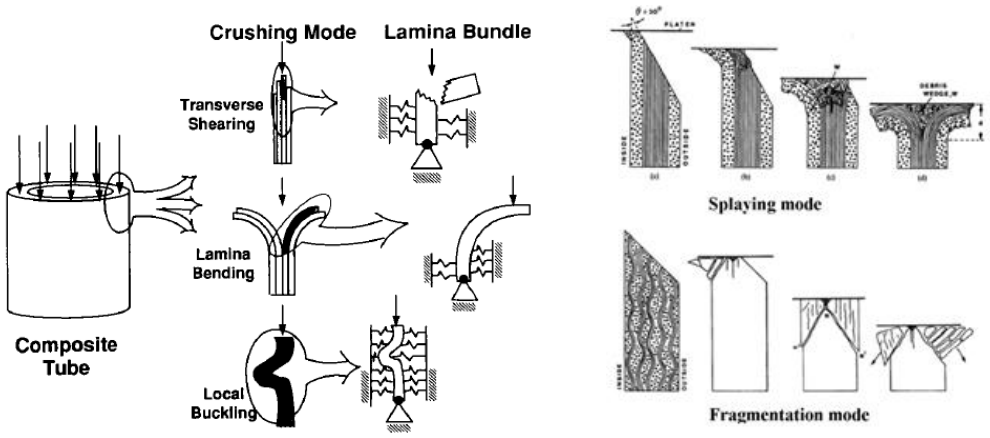



Figure 1: Left: Crushing modes of a composite tube [1]; Right: splaying and fragmentation crushing modes for chamfered plates [2]

The process is controlled by the ruin modes of the material, and constrained by the structure geometry. The SEA is conditioned by the co-existence and the competition of both ruin modes. The material ruin modes (fiber breakage, fiber/matrix decohesion, matrix cracking) can furthermore depend upon strain rate effects and distributed defects which are responsible of uncertainty on their relative contribution. The structure ruin modes are mainly described by geometrical aspects of the contact surface to the ground, and of the integrity (buckling, delamination and lamina bending, lamina bundle). Theses modes are illustrated on Figure 1. In particular, the open or closed form of the structure is a crucial point [1, 2].

For certification to crash of new generations of aircraft structures without landing gears, it is necessary to design energy-absorbing structures that are linked to the fuselage both maximizing the absorbed energy and minimizing the total mass [3, 4]. The "quality" of the optimum design of energy absorption structure in driven by two criteria: the structure should stay in a stable energy absorption process, and a maximum volume of material should be damaged during this process. The design must take into account the complexity of loading conditions and the competition of structural long-dimensions behaviours with the local structural thickness-dimensions damages [5, 7]. Even if efforts have been made to represent material properties degradations, debris wedge generation, and residual shape of structures, initiation is the most critical phase [2, 7, 8] and the part of the process where fragmentation is the more active.

Understanding the interactions between the structural geometry evolution and the degradation of the material properties is crucial. Well known for metallic materials, it is still a challenge to determine for composite materials for which the link between initiation in the first beginning of the crushing process up to the peak force, and stabilization of the process to reach a 'stable' or 'steady' crushing force. It is the main interest of this study to contribute to this understanding, and to propose a numerical model that controls all these phenomena.

### 1.2 Structure effects on initiation and stabilization of the process

Structural effects are associated to boundary evolutions, since boundaries define the geometries either of the whole structure or the assemble of substructures. It is proposed here to define essentially two categories of boundaries for the crash/crush absorbers. First, boundaries through which the compression is applied (extremities), in particular the boundary which is in contact with the ground/wall which represent indeed the obstacle to the structure global displacement. This boundary is responsible of the load increase, and is the place where the internal energy is released through decohesions of internal link, and where the three modes of energy absorption appear. Boundaries at the vicinity of the obstacle evolve as failures and ruptures appear and fragments are evacuated all the process long. Second, boundaries where internal energy is not increased but is only released because the fragments or parts separate from each other. The resulting SEA describes the balance between these entering and exiting energies through the evolution of the boundaries, or solely exiting energies through the second kind. Following this boundaries decomposition, it is proposed to analyze and model separately the phenomena appearing at the vicinity of the obstacle, and those appearing in the rest of the structure.

Studies of the literature devoted to phenomena at the vicinity of the obstacle are much more numerous for closed profiles because these structures are historically used by transportation companies (cars for example). For closed structures it is interesting to focus on the effects of contact shape and geometry between the structure and the obstacle on the mean crushing force and stabilization of the crushing process. Interesting works can be found using variable thickness in the structure height (distance to the obstacle) to increase progressively the contact surface, thus avoiding any loss of contact or dynamic buckling or bundling effects [9-11]. The shape of the structure (polygonal) and thus of the contact also has an effect on the number of fragments and chips and on the global structure strength [12]. Some profiles are demonstrated to activate more progressive tear related by the authors to the stable distribution of the contact force between the rigid obstacle and the structure, for different composite structures [12, 3-4]. It is interesting to link this observation with the effect of the contact angle (20°) on the sudden and catastrophic breakdown of the structure for unidirectional CFRP [13], which is not the case for pultruded composite [14]. Pultruded composites appear to be more stable and the

structural effects are much more similar to metallic structure for which the shape of the initiator as a low effect on the energy absorption performance [15]. As far as we are concerned, for unidirectional CFRP laminates, it has been proved that the trigger and the angles of contact have a great effect on the initiation of the crushing process. Furthermore, the mean crushing force depends considerably to what happens in the few instants of the contact between the chamfered structure and the obstacle [4, 6, 16-17]. This influence is due to the kind of material and the evolution of the contact conditions between the structure boundaries and the obstacle [18-19]. As a consequence, it is of primary importance to represent the specificity of the material behavior and the contact non-linearities between all the boundaries as they are more and more created forward and interacting with the obstacle and the created fragments. This is the object of the first part of this work.

### 1.3 Material effects on initiation and stabilization of the process

In parallel with these local effects, it can be noticed that drops down and increase of resisting forces after the peak force and in the mean force plateau, are conditioned by the capacity of the plies to continuously transmit the loading in the height of the structure (distance to the obstacle) or to break the loading path through local buckling kink band formation and ruptures (confined plies) or outward global buckling of plies [20]. This phenomenon has been identified by Lavoie [21] whose fundamental work shows that the plateau value and the oscillations around the mean value for APC2 laminates is twice that of CFRP or GR-Kv epoxy laminates. Indeed, the PEEK matrix exhibits a high interlaminar fracture toughness which is suspected to be at the origin of these observations. The link between the frequency of the mean force oscillations in the plateau region and the width of the kink band can be due to the relative value of the interlaminar strength and the longitudinal strength and critical value of strength in compression of laminates [22]. The more the fiber proportion, the higher the SEA and the amplitude of the mean force in the plateau region [20]. As for the case of the influence of the structure profile, these result must be put into perspective with the kind of fibers [23]. Nevertheless, it is proven that the stability of the crushing process in the plateau region, and the amplitude and frequency of the mean force oscillations, are both related to damage modes of the material in the core of the laminate but are also influenced by the waves of loading that come directly from the contact vicinity. This loading transfer is continuous up to the limit load of the material. It is thus crucial to represent in a proper way the damage modes and the relative strength in the core and between the plies of the laminate. This interaction of the interlaminar strength and the pies damage is the point of interest of the second part of this work.

## 2 COMPOSITE LAMINATE CRUSHING PROCESS: MODELLING STRATEGY

Based on observations made from experimental testing campaigns, different mechanical models were proposed in previous studies [8, 16-17]. The basic model used in this work is the nonlinear elastic anisotropic homogenized continuum damage mechanics model developed in the frame of the irreversible thermodynamics [17]. In order to circumvent the problems coming from the local compression at the contact on the ground, the elements that face the ground are changed the material model into the MCS and FFC models to follow the strategy proposed by [16]. These elements are eroded when the maximum load bearing criterion is reached. This simulation strategy is illustrated on Figure 2. Damage is represented at the mesoscale (ply size) which is the scale under interest.

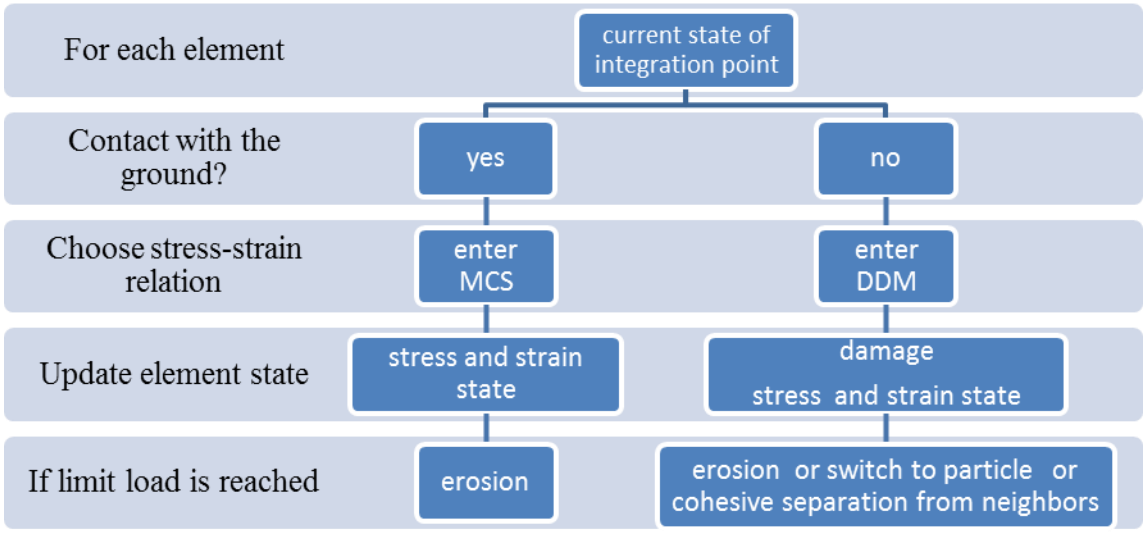



Figure 2: Deep ply damage and localized fragmentations models switch.

The continuous diffuse damage model is available for 3D material, whatever the fibre orientations are, and can manage strain rate effects (not considered here, despite the observation of Lavoie [21]). Modes of ruin are represented by criteria obtained from experiments and are directly related to the deformation energy. The MCS and FFC models have been characterized by other experiments.

## 2.1 Continuous diffuse damage model

The continuous Diffuse Damage Model has been developed on the basis of the previous work of Ylias [17]. Enhancements have been proposed to adapt the model to the strategy described on Figure 2. The DDM emphasizes that anisotropy is unchanged after damage and the material remains elastic in each ply. Thus, as done "classically", damage is modelled through its effect on the elastic rigidity loss in further loading or unloading without a change in the anisotropy directions until it reaches the value of 1: di is the damage of $E_{ii}$ for i=1,3, d4 is the damage of $G_{12}$, d5 of $G_{23}$ and d6 of $G_{13}$.

The model distinguishes 6 damage variables {di}, i=1,6 affecting the stiffness of the ply, and 6 failure modes {fj}, j=1,6 (eq. 2-7), where:

<.> defines the Macauley brackets,

$X_T$ and $X_C$ are respectively tensile/compressive failure stresses in fibres direction,

$Y_T$ and $Y_C$ are respectively tensile/compressive damage threshold stresses in direction 2,

$Z_T$ and $Z_C$ are respectively tensile/compressive damage threshold stresses in direction 3,

$S_{12}$, $S_{23}$ and $S_{13}$ are respectively shear damage threshold stresses,

$r_i \in [1,\infty]$ is called the limit load ratio.

$$f_1(\varepsilon, r_1) = \left[\frac{E_1\langle\varepsilon_1\rangle}{X_T}\right]^2 + \left[\frac{G_{12}\varepsilon_{12}^2 + G_{13}\varepsilon_{13}^2}{S_{fs}^2}\right] - r_1^2 = 0 \qquad (2)$$

*tension fiber direction*

$$f_2(\varepsilon, r_2) = \left[\frac{E_1\langle -2\varepsilon_1\rangle + \langle -E_2\varepsilon_2 - E_2\varepsilon_3\rangle}{2X_C}\right]^2 - r_2^2 = 0 \qquad (3)$$

*compression fiber direction*

$$f_3(\varepsilon, r_3) = \left[\frac{E_3\langle -\varepsilon_3\rangle}{Z_C}\right]^2 - r_3^2 = 0 \quad \text{out-of plane tension} \tag{4}$$

$$f_4(\varepsilon, r_4) = \left[\frac{E_2\langle -\varepsilon_2\rangle}{Y_C}\right]^2 - r_4^2 = 0 \quad \textit{tension perp. fiber dir} \tag{5}$$

$$f_5(\varepsilon, r_5) = \left[\frac{E_2\langle \varepsilon_2\rangle}{Y_T}\right]^2 + \left[\frac{G_{12}\varepsilon_{12}}{S_{12^\circ} + S_{SRB}}\right]^2 + \left[\frac{G_{23}\varepsilon_{23}}{S_{23^\circ} + S_{SRB}}\right]^2 - r_5^2 = 0 \quad \textit{matrix cracking} \tag{6}$$

$$f_6(\varepsilon, r_6) = \left[\frac{E_3\langle \varepsilon_3\rangle}{Z_T}\right]^2 + \left[\frac{G_{13}\varepsilon_{13}}{S_{13^\circ} + S_{SRB}}\right]^2 + \left[\frac{G_{23}\varepsilon_{23}}{S_{23^\circ} + S_{SRB}}\right]^2 - r_6^2 = 0 \quad \textit{delamination} \tag{7}$$

At the undamaged state, limit load ratios ri are set to 1. They evolve following the current state of damage using the first thermodynamics principle and following (eq. 8) [15]:

$$\dot{d}_i(\sigma, \omega, \dot{\varepsilon}) = \sum_{j=1}^{5} q_{ij}.\dot{\phi}_j(m, r_j), \quad \phi_j = 1 - e^{\frac{1}{m}\left(1 - r_j^m\right)}, \quad r_j \geq 1 \tag{8}$$

## 2.2 Mean Crushing Stress and Free Face Crushing laws for local fragmentation

The stress-strain curve of elements at the tip of the plies that are in contact with the ground is similar to an elastic perfectly plastic behaviour in compression (Figure 3).

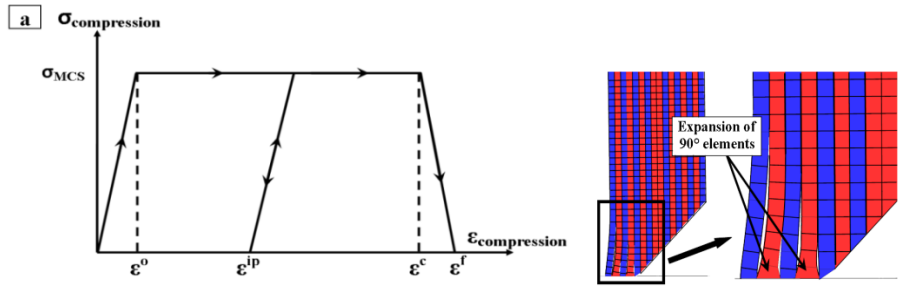



Figure 3: Left: MCS law for both 0° and 90° plies; Right: expansion of 90° elements to allow initiation of delamination using cohesive elements.

Data used in this study have been characterized by Israr & al. [16]. The elastic deformations $\varepsilon_{0^\circ}^0$ and $\varepsilon_{90^\circ}^0$ are given by equations (9) and (10), where $H_{11c}$ = 101 GPa, and $H_{22c}$ = 9GPa. The maximum compression strain $\varepsilon_c$ and failure strain $\varepsilon_f$ are the same for plies at 0° and 90° and are respectively -3 (95%) and -4.6 (99%).

$$\varepsilon_{0^\circ}^0 = \frac{\sigma_{MCS\,0^\circ}}{H11c} \text{ with } H_{11c} = 101\text{GPa and } \sigma_{MCS\,0^\circ} = 277\ MPa \tag{9}$$

$$\varepsilon_{90^\circ}^0 = \frac{\sigma_{MCS\,90^\circ}}{H22c} \text{ with } H_{22c} = 9\text{GPa and } \sigma_{MCS\,90^\circ} = 270\ MPa \tag{10}$$

In order to treat differently the elements where the continuous diffuse damage law is used or the Mean Crushing Sress law is used, and to reduce spurious oscillations in the contact force, the FFC is implemented in a user program and called as a sub-routine in the user material of Impetus Afea. The principle of the Free Face Crushing is illustrated on Figure 4.

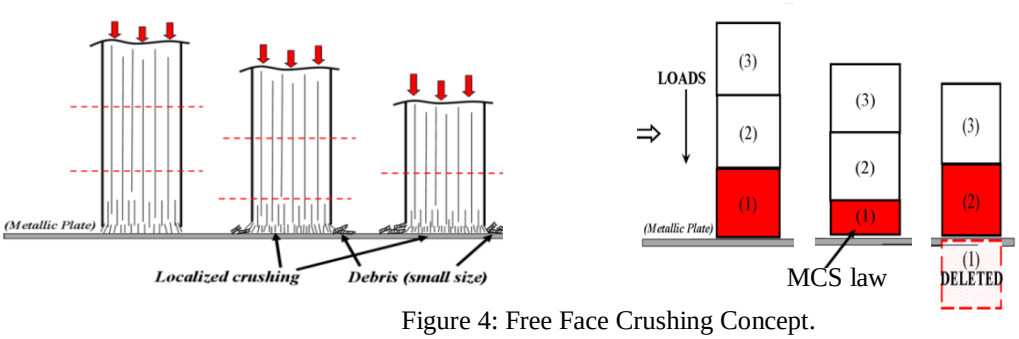



Figure 4: Free Face Crushing Concept.

## 2.3 Cohesive failure

The cohesive failure is generated through a tied contact with failure and propagated using a quadratic failure criterion using mode I and mode II strength related to $G_{Ic}$ and $G_{IIC}$ energy release rates. Energy release follows a linear tension to opening relationship. In order to limit the mesh size effect in the opening crack propagation, a scale factor $\zeta$ is introduced [17].

## 2.4 Experimental case under interest

$[(0°/90°)_4]$s samples are made of 268g/m² T700GC/M21e prepreg (35% of resin), so that the global density is 1535 kg/m$^3$. Sample plates are 160mm (0°_fiber_direction) x 60mm (90°_perp.plan_direction) chamfered rectangular plates so that initiation is conditioned by the wedge angle, here 45°. Each ply is 0,25mm thick. A heavy carriage is fixed at the upper boundary of the plate to adjust the desired initial kinetic energy. The mass chosen here is 36kg, and the initial velocity is 9 m/s. The free height for crushing is 20 mm. The experimental set-up is described in detail in references [7-8, 16, 17].

## 2.5 Models inputs

The finite element model is described Figure 5 (shortened in order to show details; the real model has the same height as in the experiment). Finite elements for the composite structure are eight nodes degree 1 fully integrated isogeometric elements, with an edge length of 0,25mm. To generate the chamfer, lines of elements in the stacking plies are 1 element shorter from top to bottom along the (z) direction of the frame. In order to stabilize the ignition of the crushing process, the mesh of the extremity of the plate and the contact parameters has been smoothed to get a strait surface for the chamfer.

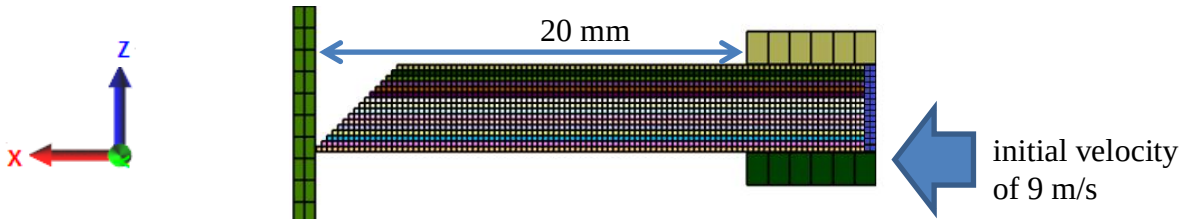



Figure 5: One element thick finite element model for 2D simulation

The heavy carriage is non deformable and perfectly tied to the composite structure's end opposite to the contact extremity. This system has an initial velocity of 9m/s. The vertical guides are rigid and fixed in space. Penalty contacts are defined between the composite structure and the rigid guides, and between the structure and the base (with a friction of 15%

with the ground). Faces of elements perpendicular to the (y) direction are constrained with kinematic symmetry conditions to obtain a 2D plane strain model.

### 2.6 Effects of damage modelling on the force

Typical force-displacement curves obtained in experiments (grey plain and dashed lines) are compared on Figure 6 with the numerical result (orange line) of a previous model with the scale shape mesh at the extremity [8]. The contact force is derived from the total amount of internal energy in the simulation. Numerical and experimental frequencies higher than 1500Hz are filtered (moving average). The displacement is that of the heavy carriage centre of mass.

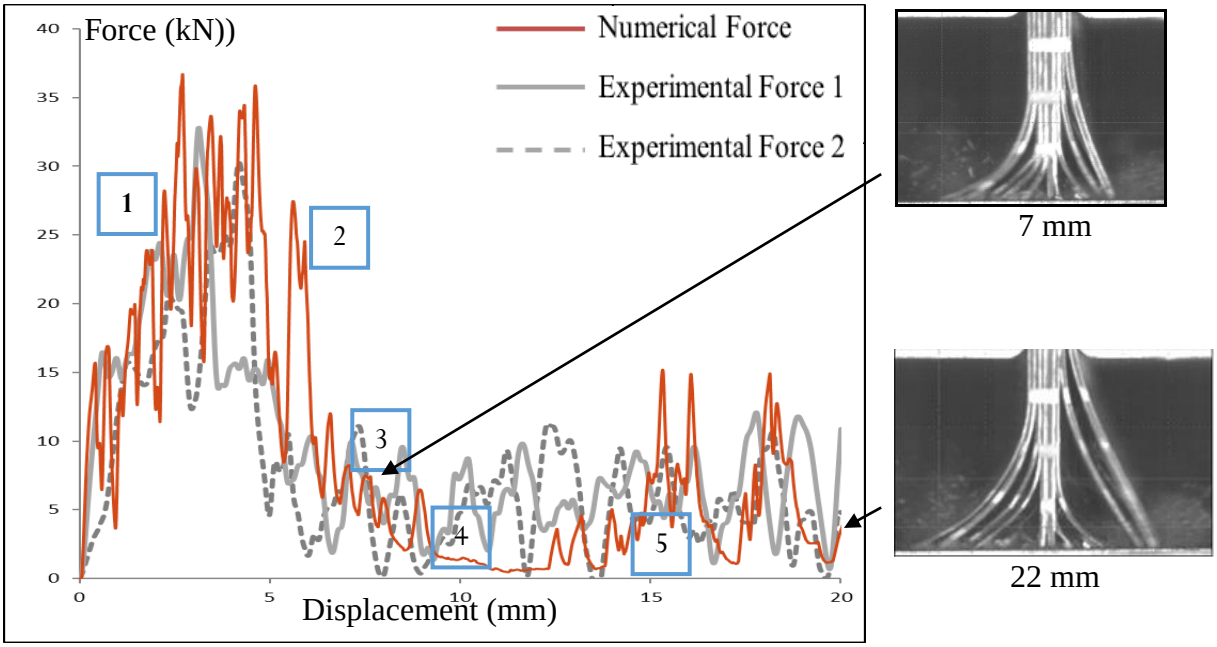



Figure 6: Numerical force – displacement curve versus experiments [8]

As can be seen on Figure 6, the two experimental curves have similar peak force around 32kN. They illustrate the variability in the experiments, but still the global phases in the crushing process remain. We have selected five phases in the force-displacement curve [8, 16-17]. Phase 1 corresponds to the initiation of the crushing from the first contact (t=0) to the peak of force. Oscillations in this phase correspond to the propagation of the acoustic wave generated by the contact, up and down to in the height (160mm). The height of the plate damaged during phase 1 corresponds to about one time the thickness for local fragmentation, and three times the thickness for diffuse core damage. Phases 2 and 3 correspond to the global breakdown of the structure. In these phases, the height of the plate that is damaged is the full free height, essentially in splaying (delamination of outer plies). Phase 3 is a transition phase during which the global damage and the local fragmentation modes interact to set the steady process in. Phases 4 to 5 illustrate the beginning of the stabilization process for which the mean crushing force is reached, and its shape as an oscillatory function in time is set.

In a previous work [8], it was demonstrated that the proposed simulation strategy that represents in a different way the local fragmentation (dust) and the global core damage (diffuse) allows to catch the global behaviour in phases 1 3 and 5. This strategy uses erosion technique to represent the elimination of matter a yield of volume reduction [26]. For the present study,

element are eroded when they reach 8% of longitudinal crushing for 0° plies, and 10% for 90° plies. Because of this matter sudden suppression, the model is not precise enough to represent in one hand the peak force because of local instabilities at the vicinity of the ground, and in another hand phase 3-4 of stabilization because of losses of contact when a group of elements were eroded or when a group of bended plies slide on the ground. For that purpose, two phases are particularly examined here, phase 1 and phase 3-4 and the effect of interlaminar fracture toughness on the force and the crushing modes in these phases and in phase 5.

## 3 STABILIZATION OF THE CRUSHING PROCESS

Following the observation done by Lavoie [21], it is aimed here to investigate the effect on the different phases of the crushing process illustrated on figure 6, and the effect of the interlaminar toughness on the interaction between stress localization in the core of the laminate and the global breakdown and delamination. These interactions are suspected to be at the origin of the zero forces obtained in the simulations because of sudden erosion of groups of elements or loss of contact due to sudden sliding of elements done up to now with the finite elements models that do not use particles [8].

A simulation plan is defined and investigated, using different values of the interlaminar fracture toughness. The maximum normal and tangential stresses are kept constant (except one data set), and energy release rates are modified (see Table 1).

| Case # | S (GPa) | T (GPa) | GI (N/mm) | GII (N/mm) | Comments |
|---|---|---|---|---|---|
| 1 | 0.075 | 0.15 | 0.6 | 2.1 | T800s/M21e [8] |
| 2 | | | 0.54 | 1.89 | T700s/M21e [16] |
| 3 | | | 0.54 | 2.1 | Mixture of #1 and #2 |
| 4 | | | 0.6 | 1.89 | Mixture of #2 and #1 |
| 5 | | | 0.6 | 1.6 | GII*75% for dynamic effects |
| 6 | | | 0.75 | 2.1 | GI*1.25% for dynamic effects |
| 7 | | | 0.75 | 1.6 | GI*1.25% and GII*75% |
| 8 | | | 0.545 | 1.387 | T700s/M21e [24] |
| 9 | 0.05 | 0.09 | 0.75 | 1.2 | T800s/M21e [25] |

Table 1: Example of the construction of one table.

All force-displacements curves obtained from the numerical simulations of cases #1 to #9 are compared with experimental force 1 (grey line) presented in Figure 6. Numerical curves obtained for a thickness of 0.25mm in the frame of a 2D plane strain hypothesis are scaled in order to be compared with experimental curves obtained on the complete structure of 60mm thick. Cases #1 and #2 correspond to data sets of different materials obtained from experimental characterization in the laboratory. Cases #3 and #4 are combinations of GI and GII from #1and #2. Cases #5 #6 and #7 are adjustments of the values of GI and GII for dynamic effects, that give values not far from data sets of cases #8 and #9. Cases #8 and #9 are obtained from characterizations tests done in the laboratory, and #9 was the set used for the model that used the FE-to-particle switch technique [25]. Different comparisons are made to evaluate: the effect of GI value at the fixed values of GII for the T700 and the T800 composites; the effect of GII at the fixed values of GI for the T700 and the T800 composites. Discussions are proposed on the role of the stabilization in the initiation phase 1 on the transition phase 3 and the steady compression process in phases 4-5.

### 3.1 Effect of extreme of GII at different values of GI

Figure 7, Figure 8, Figure 12, and Figure 12 present the effect of different values of GI at fixed values of GII, on the force-displacement curve. From the comparison of all these curves, it can be noticed that data sets of #8 and #9 for GII around 1.3N/mm gives really good correlations of the numerical curves compared to experiment 1, both for the peak force and for the amplitude of the mean force in phases 4-5. But as reminded earlier, these data set exhibit long zones of almost zero force. For #8, this is due to breakdown of groups of plies that create very big fragments in an unsymmetrical mode of crushing while one half of the structure remains in contact. For #9, this is due to erosion of elements that are in contact with the ground, , or sliding or groups of elements associated with splaying of outer plies.

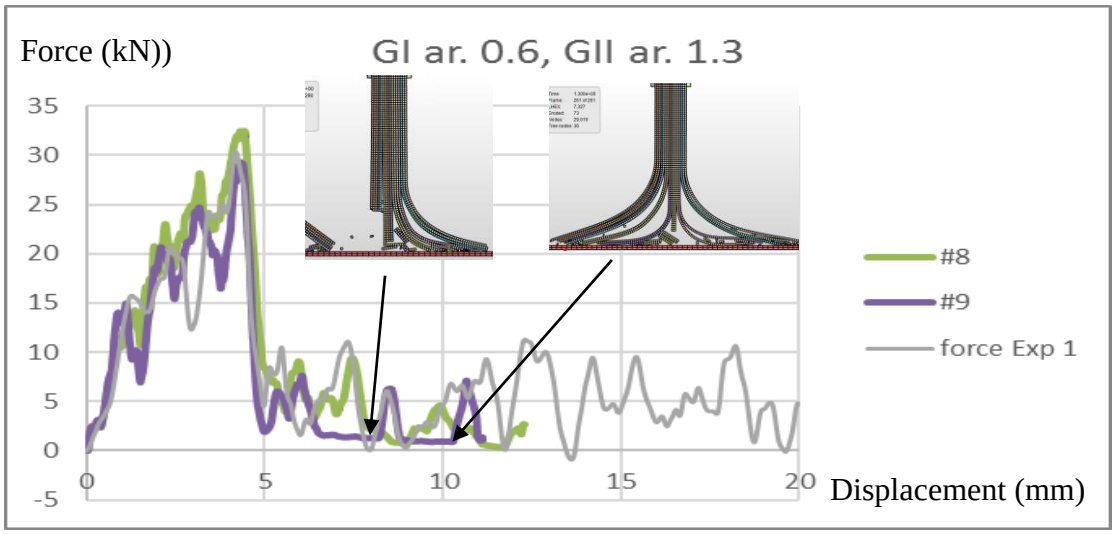



Figure 7: Numerical force – displacement curve at fixed GII around 1.3N/mm versus experiment 1

Comparing the three figures, we can see that increasing GII has two effects. In the initiation phase 1, the force exhibits more oscillations, as observed in the experiment. For and more for #6 the force-displacement curve is very similar to the experiment up to 8mm (phase 3).

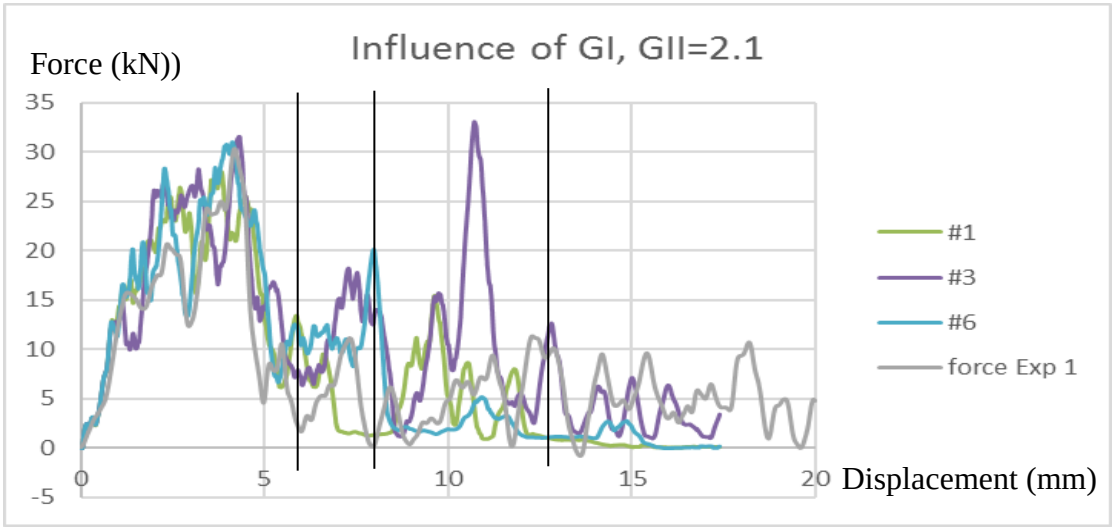



Figure 8: Numerical force – displacement curve at fixed GII=2.1N/mm versus experiment 1

But for phases 4-5, comparing the curves of Figure 8, we can see that for the same high value of GII, the higher GI the earlier zero forces arise. For #3, where GI has the lowest value, the frequency and amplitude of the force-displacement curve are very similar to the experimental one in phase 5 (>12mm).

Crushing profiles of #3, #1, and #6, are presented at three stages of displacement, 6mm, 8mm, and 12.7mm (see Figure 8). The high values of force in phase 5 of case #3 is explained by the accumulation of debris and the angle of contact between the outer plies and the ground. The angle of contact is explained by debris accumulation between groups of two plies at most (top right in Figure 9). The angle of contact is also quite high which explains the high resistance of #6 in phase 3 (8mm).

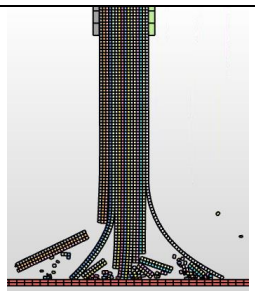
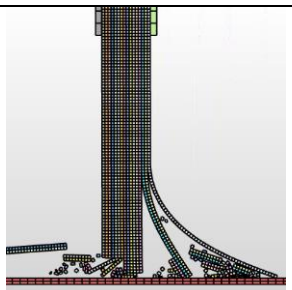
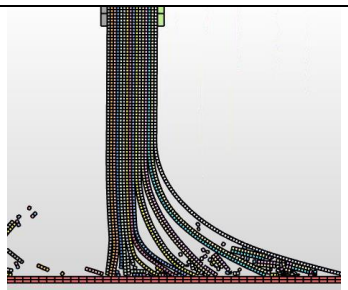
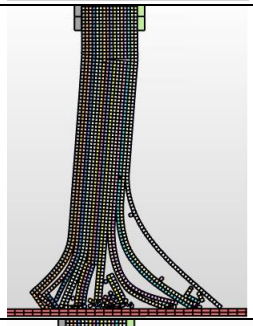
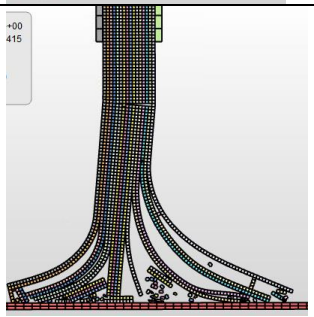
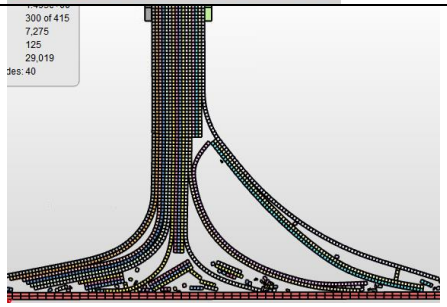
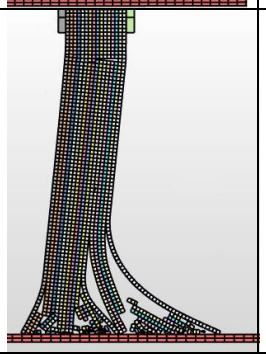
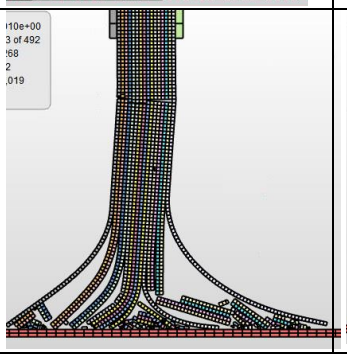
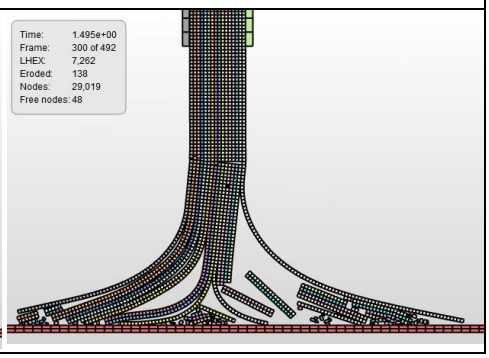
| 2.1 N/mm | d=6mm | d=8mm | d=12.7mm |
|---|---|---|---|
| #3 ; GI= 0.54 N/mm |  |  |  |
| #1 ; GI=0.6 N/mm |  |  |  |
| #6 ; GI= 0.75 N/mm |  |  |  |

Figure 9: Crushing modes for GII=2.1N/mm for #3, #1, and #6

A good compromise for the force curve is obtained for #1 for which the outer plies are bended and slide on each other and on the ground with help of debris on which they ‘roll’ on the left part. On the right part very big debris ensure a non-zero effort for a long period. The angle of contact of the plies on the ground are nevertheless not well reproduced. This explains why the contact force decrease. For #3, the low value of GI coupled with a high value of GII is responsible of a rapid breakage of groups of plies subjected to bending. Debris are created by the brittle fracture of plies which explains the high amplitude of oscillations of the re-

sistance force in phase 4. Even if the mean force is very similar to the experimental one, the crushing process is too brittle and not representative of the tests.

### 3.2 Effect of relative variations of GII and GI

As for GII=2.1N/mm on Figure 8, it can be observed on Figure 10 and Figure 12 that a lower value of GI generates higher amplitudes of oscillations in phase 1, more comparable to the experimental curve. Furthermore, a higher value of GI generates and more losses of contact in phase 3. But on the crushing modes, a higher value of GI helps to stabilize the process and the symmetry of energy dissipation (Figure 11).

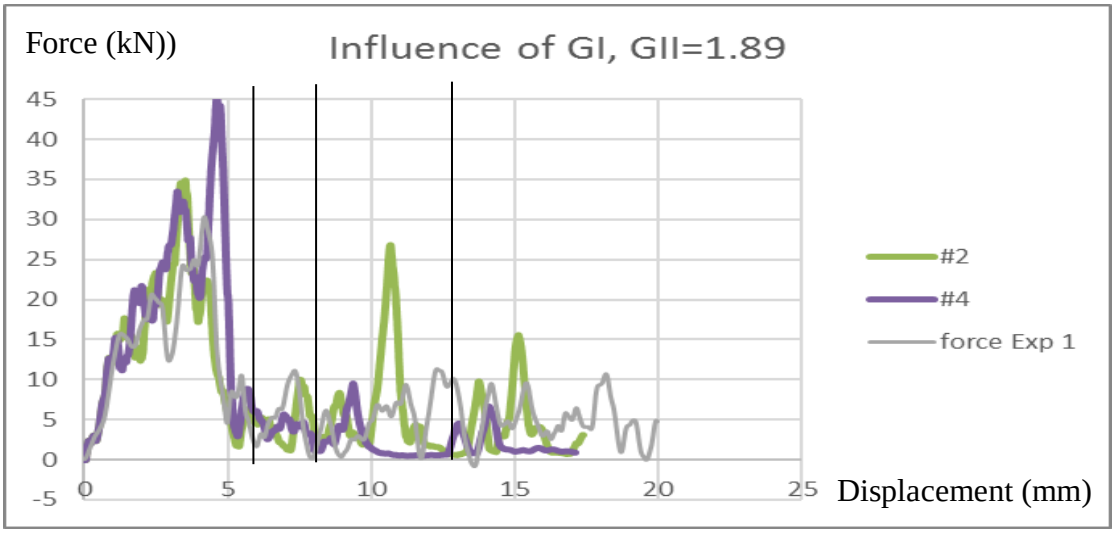



Figure 10: Numerical force – displacement curve at fixed GII=1.89N/mm versus experiment 1

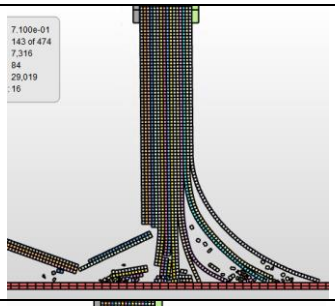
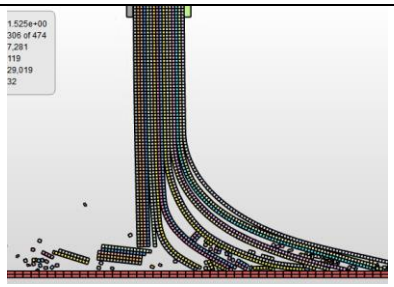
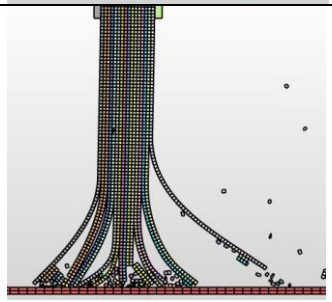
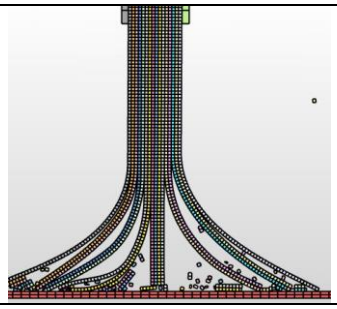
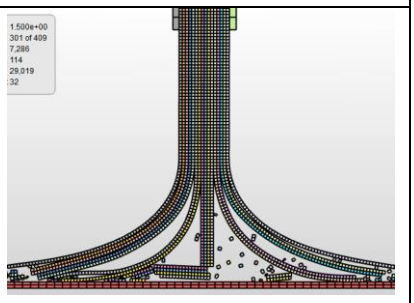
| 1.89 N/mm | d=6mm | d=8mm | d=12.7mm |
|---|---|---|---|
| #2 ; GI= 0.54 N/mm |  |  |  |
| #4 ; GI= 0.6 N/mm |  |  |  |

Figure 11: Crushing modes for GII=1.89N/mm for #2, and #4

For #2, which have the same GI=0.54N/mm than #3, an artificial peak force appears in phase 4 as well. This peak force is due to the bad configuration of the crushing process (big

fragments broken by bending. It does not appear in #8. It is concluded that the peak appears only for higher values of GII in that transition phase. Drop down of the force in phase 3 is realistic for both cases #5 and #7.

The crushing profile of cases #2 and #3 are very similar. Case #2 generates two sizes of fragments. Clouds of single elements are confined between outer plies on the right part as in #3. This causes as in #3 an angle of contact that is responsible of a non-zero force during almost all the process except during short periods. Peaks of axial force are due to the breakdown of bigger fragments of about 3.5mm length and 1.5 mm width. Should these fragments not completely break but have a smoother behavior, case #2 would be very close to the test.

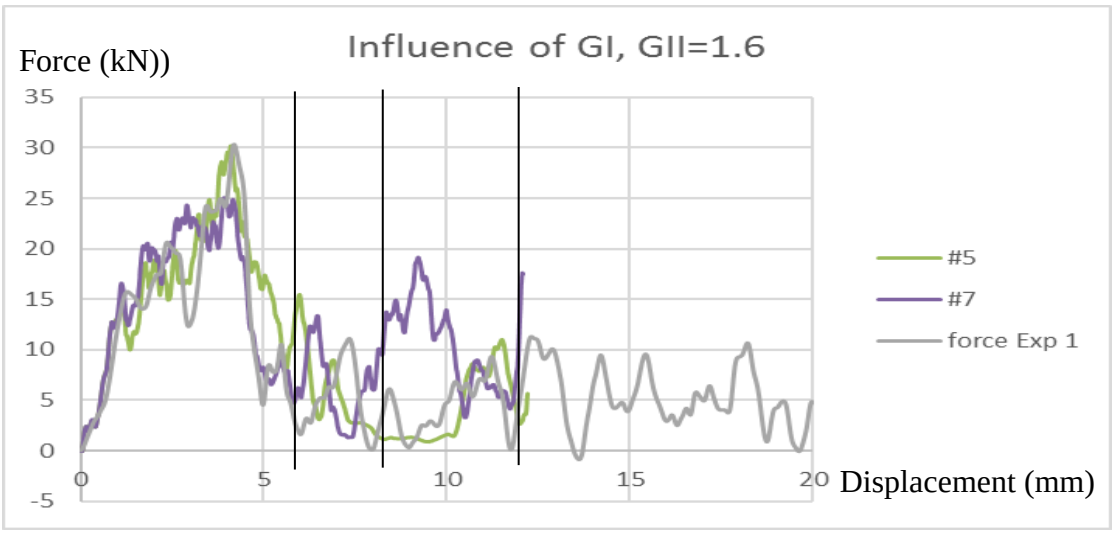



Figure 12: Numerical force – displacement curve at fixed GII=1.6N/mm versus experiment 1

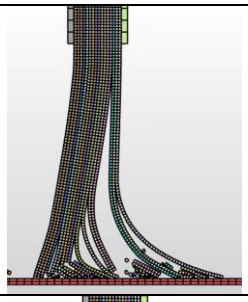
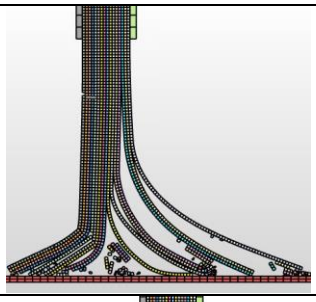
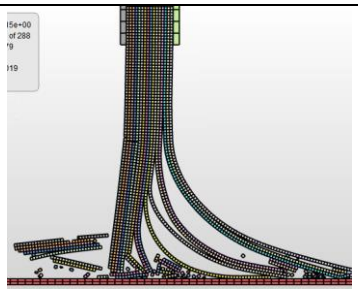
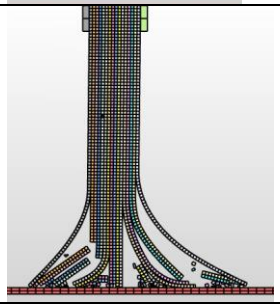
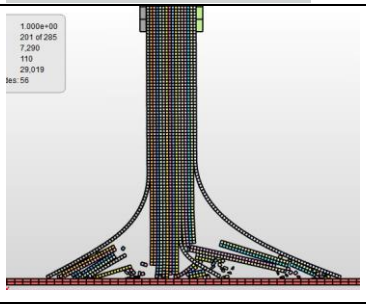
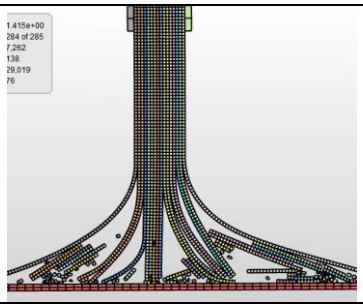
| 1.6 N/mm | d=6mm | d=8mm | d=12mm |
|---|---|---|---|
| #5 ; GI= 0.6 N/mm |  |  |  |
| #7 ; GI= 0.75 N/mm |  |  |  |

Figure 13: Crushing modes for GII=1.6N/mm for #5, and #7

The same analysis can be conducted for #5 and #7. Phase 1 of #5 is very close to the experiment. Oscillations are higher than for GII=1.3N/mm, but smaller than GII=2.1N/mm. This lead to the general conclusion that a high value of GII is responsible of the amplitude of oscillations of the force-displacement curves, and of the size distribution of the fragments.

This observation is also comforted by the observation of the crushing modes of Figure 13. From the comparison of cases #6 and #7, it can also be noticed that a higher GI helps the local fragmentation at the vicinity with the ground, and thus the symmetry of the process.

### 3.3 Comparison with experiments

Figure 14 below shows the evolution of the crushing profile in the different phases of the experiment, same phases used here to analyze the different cases of simulation.

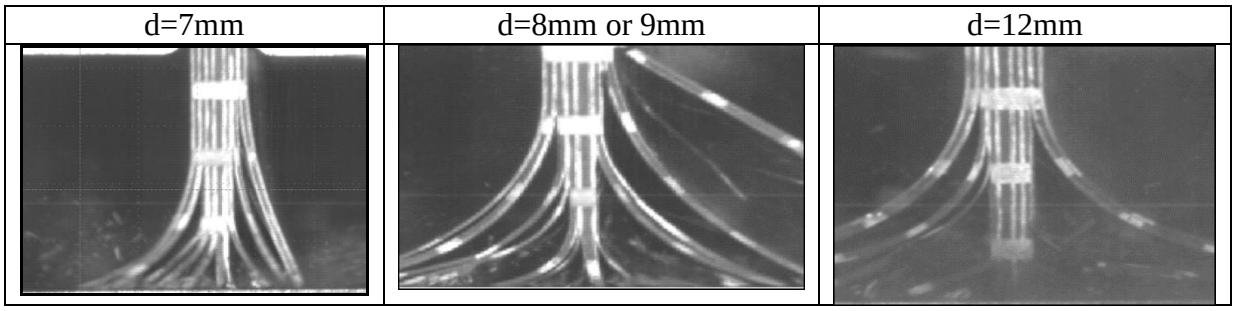



Figure 14: Crushing modes in the experiment

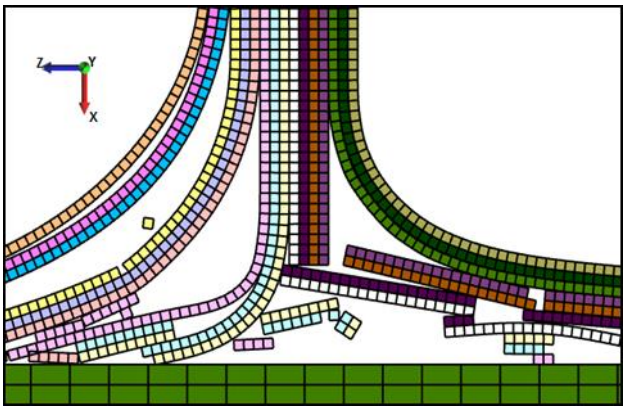



Figure 15: Crushing modes in the previous model : loss of contact due to breakdown of group of elements, erosion, or sliding at a flat angle

Cases #5 and #7 give the best comparison with the experiment for the crushing force, as well for the crushing process. Values of GI and GII are quite close from values of #8 and #9. GII has a lower value allowing the stabilization of the process. In #5, the lower value of GI is responsible of the decrease in the resistance force, as it is in the real test. The final decrease in phase 3 is due to creation of big fragments which are not suspected to come from the GII effect, but rather from the erosion technique. Indeed, erosion creates unstable breakdown. It is necessary for us to improve the diffuse damage model in order to stabilize the crushing inside plies, and the formation of kink bands or shear local resistance instead of deleting the matter.

## 4 DISCUSSION AND CONCLUSIONS

The proposed approach is an enhancement of previous developed numerical models that proved that it is necessary to represent the subtle coupling between the ruin modes of the material and the geometric versus the continuous representation of them to capture the initiation phase of the crushing process and the energy absorption characteristics of the structure. The first model showed that the representation of each size of fragment was necessary and this was done by a FE-particle switching. But this first model also suffered from instabilities at the contact vicinity between the composite sample and the ground. To reproduce in a stable manner, the peak and plateau forces at the macro scale, and the different fragmentation processes at the meso scale, it was necessary to deactivate the local crushing ruin mode in axial compression in the stress-strain material model and to deactivate criteria that were responsible of spurious interactions between out of plane continuous damage inside plies and the cohesive failure between the plies. The present study has proposed an enhancement of this previous model through the introduction of a continuous damage law dedicated to the numerical representation of the highly complex local fragmentation process at the vicinity of the crushing process on the ground in a soft manner. Since oscillations and sudden drops still appear during the crushing process, it was suspected that the interlaminar fracture toughness was not correctly reproduced, leading to an early breakdown of groups of plies because of a dynamic unsteady bending of plies.

The study presented here has investigated the influence of the interlaminar fracture toughness on two measurements, the force-displacement curve, and the crushing process. The different combinations of GI and GII give contrary results for our models, whether the force-displacement or the crushing process is analyzed, and two cases could be selected as the best compromise. It has been concluded from the comparison with the experiments that the interlaminar values are not solely responsible of the remaining instabilities of the model. The erosion technique is suspected to be at the origin of the breakdown of groups of plies and the creation of very big fragments in the best configurations of GI and GII. This remain a perspective of our work to enhance the numerical representation of fractures and ruptures.

In a more general analysis, it has been shown that high values of GII are responsible of gig fragments generation, and can either cause the complete rupture of the laminate near the rigid guides. GII is responsible also and for the same reason of the high amplitude of oscillations and disturbing peaks on the resistance force in the force-displacement curve during all the process. This result is concordant with results of the literature that have used other couples of matrix and fibers. For a given value of GII, it has been shown that the higher the GI, the higher also the amplitudes of oscillations in the initiation phase 1, leading also to a more symmetrical crushing mode from the stabilization phase 3 to the steady phase 5. This stabilization of the crushing process is a good thing up to a certain proportion indeed, because if the stable splaying mode is favored, plies only separate through delamination and slide on each other ‘rolling’ on debris that are confined between them and the ground, and making a flat angle of contact with the ground. This rapidly leads to decreasing a lot the absorbed energy down to a quasi-null energy dissipation. Before this inefficient splaying mode appears, it has been shown that a right choice of GII and GI combination allows the separation of plies by groups of two or four (for the chosen layup) and at the same time the creation of small to intermediate debris confined between them, forcing the angle of contact and so the bending in the configuration where the accumulated internal energy before rupture is maximized.

It is also shown with these results, that a good design of architecture for a crushing structure is certainly a structure that is not uniform in the height, so that crushing modes are activated at different phases during the process. It is suggested here that the configuration of #6

should be used for the initiation phase because it is a good way to release a higher amount of energy and to initiate a stable crushing, then the structure should be designed to have a higher GI compared to GII for outer plies (like #8, #9, #5 or #7) in order to create small to intermediate debris confined between adjacent plies, forcing them to keep an optimized angle of contact with the ground and help the steady process in a higher level of energy absorption.

## 5 ACKNOWLEDGEMENTS

This work has been funded by the French ANR Program, and led in the frame of the ANR VULOMP2 project. Jerôme Limido of Impetus Afea is sincerely thanked for his great help in user's material modelling and development, and general use of Impetus Code. All the technical staffs of ICA and ISAE-SUPAERO is thanked for their help.

## REFERENCES

[1] G.L. Farley, R. Jones, Crushing characteristics of continuous fiber-reinforced composite tubes, *Journal of Composite Materials*, **26/1**, 37-50, 1992.

[2] D. Hull, A unified approach to progressive crushing of fiber-reinforced composite tubes, *Composite Science and Technology*, **40**, 377-421, 1991.

[3] J.F.M. Wiggenraad, A.L.P.J. Michielsen, D. Santoro, F. Lepage, C. Kindervater, F. Beltran, M. Al-Khalil, Finite element methodologies development to simulate the behaviour of composite fuselage structure and correlation with drop test. *Air & Space Europe*, **3/3-4**, 228-233, 2001.

[4] M.A. McCarthy, J.F.M. Wiggenraad, Numerical investigation of a crash test of a composite helicopter subfloor structure. *Composite Structures*, **51**, 345-359-256, 2001.

[5] P. Qiao, J. Davalos, E.J. Barbero, Design optimization of fiber reinforced plastic composite shapes. *Journal of Composite Materials*, **32/2**, 177-196, 1998.

[6] P. Feraboli, Development of a Corrugated Test Specimen for Composite Materials Energy Absorption. *Journal of Composite Materials*, **42/3**, 229-256, 2008.

[7] D. Guillon, Etude des mécanismes d'absorption d'énergie lors de l'écrasement progressif de structures composites à base de fibre de carbone. *Ph D. Thesis*, Université de Toulouse-ICA-ISAE, France, 2008.

[8] F. Tostain, C. Espinosa, S. Rivallant, Crushing of aeronautical composite absorbers using canonical and phenomenological ruin criteria. V *ECCOMAS Thematic Conference on the Mechanical Response of Composites*, Bristol, United Kingdom, September 7-11, 2015.

[9] G. Sun, F. Xu, G. Li, Q. Li, Crashing analysis and multiobjective optimization for thin-walled structures with functionally graded thickness. *International Journal of Impact Engineering*, **64**, 62-74, 2014.

[10] P. Xu, C. Yang, Y. Peng, S. Yao, D. Zhang, B. Li, Crash performance and multi-objective optimization of a gradual energy-absorbing structure for subway vehicles, *International Journal of Mechanical Sciences*, **107**, 1-12, 2016.

[11] X. Zhang, H. Zhang, Z. Wang, Bending collapse of square tubes with variable thickness. *International Journal of Mechanical Sciences*, **106**, 107-116, 2016.

[12] H. Assaee, J. Rouzegar, M.S. Saeedi Fkher, A. Niknejad, Axial splitting of composite columns with different cross sections, *Thin-walled Structures*, **99**, 109-118, 2016.

[13] M. Morthorst, P. Horst, Crushing of conical composite shells: a numerical analysis of the governing factors, *Aerospace Science and Technology*, **10**, 127-135, 2006.

[14] M.A. Jimenez, A. Miravete, E. Larrodé, D. Revuelta, Effect of trigger geometry on energy absorption incomposite pofiles, *Composite Structures*, **48**, 107-111, 2000.

[15] A.T. Beyene, E.G. Koricho, G. Belingardi, B. Martorana, Design and manufacturing issues in the development of lightweight solution for a vehicle frontal bumper, *International Symposium on Dynamic Response and Failure of Composite Materials (DRaF2014), Procedia Engineering*, **88**, 77-84, 2014.

[16] H. A. Israr, S. Rivallant, C. Bouvet, J.J. Barrau, Finite element simulation of 0/90 CFRP laminated plates subjected to crushing using a free-face-crushing concept. *Composites Part A: Applied Science and Manufacturing*, **62**, 16-25, 2014.

[17] C. Espinosa, F. Lachaud, J. Limido, J.-L. Lacome, A. Bisson, M. Charlotte, Coupling continuous damage and debris fragmentation for energy absorption prediction by CFRP structures during crushing, *Computational Particle Mechanics*, **2/1**, 1-17, 2015.

[18] L. N.S. Chiu, B. G. Falzon, D. Ruan, S. Xu, R. S. Thomson, B. Chen, W. Yan, Crush responses of composite cylinder under quasi-static and dynamic loading, *Composite Structures*, **131**, 90-98, 2015.

[19] F. Tostain, P. G. Ulecia, L. F.r Alfonso, S. Rivallant, C. Espinosa, Crushing of composite plates: an experimental comparison between laminate and interlock behaviour under static and dynamic loads, *1st International Conference on Impact Loading of Structures and Materials (ICILSM 2016)*, Turin, Italy, May 22-26, 2016.

[20] A. Esnaola, I. Tena, J. Aurrekoetxea, I. Gallego, I. Ulacia, Effect of fibre volume fraction on energy absorption capabilities of E-glass/polyester automotive crash structures, *Composite Part B*, **85**, 1-7, 2016.

[21] J. A. Lavoie, S. Kellas, Dynamic crush tests of energy-absorbing laminated composite plates, *Composites Part A*, **27A**, 467-475, 1996.

[22] T.J. Vogler, S. Kyriakides, Initiation and axial propagation of kik bands in fiber composites, *Acta Mater.*, **45/6**, 2443-2454, 1997.

[23] R.A. Eshkoor, A.U. Ude, S.A. Oshkovr, A.B. Sulong, R. Zulkifli, A.K. Ariffin, Failure mechanism of woven natural silk/epoxy rectangular composite tubes under axial quasi-static crushing test using trigger mechanism, *International Journal of Impact engineering*, **64**,53-61, 2014.

[24] M. Ilyas: Damage modeling of carbon epoxy laminated composites submitted to impact loading, PhD thesis, Université de Toulouse- ICA-ISAE, France, 2010.

[25] C. Espinosa, F. Lachaud, J. Limido, J.-L. Lacome, Combining a CDM model and a FE-Particle method to analyse the ruin modes of a composite structure during a crash. *European Congress on Computational Methods in Applied Sciences and Engineering, (ECCOMAS 2012)*, Vienna, Austria, September 10-14, 2012.

[26] Material Sciences Corporation (AMT162 of LS-DYNA): http://www.materials-sciences.com/dyna.aspx

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# STRUCTURAL ANALYSIS OF ADHESIVE BONDING FOR THICK-WALLED TUBULAR COMPOSITE PROFILES

**Geminiano Mancusi**[1], **Agostina Orefice**[1], **Luciano Feo**[1], **Fernando Fraternali**[1]

[1] Department of Civil Engineering, University of Salerno
Via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy
e-mail: {g.mancusi,aorefice,l.feo,f.fraternali}@unisa.it

**Keywords:** Tubular profiles, GFRP, Space frame truss, Adhesive bonding.

**Abstract** A *mechanical model for predicting the response of a GFRP tubular truss member glued to terminal* (*nodal*) *devices is presented. The given model captures shear deformation and simulates adhesion between the composite tube and the nodal devices via an appropriate cohesive interface. A comparison between theoretical and experimental results highlights a very good theory-experiment matching in terms of failure load and global displacements.*

## 1. INTRODUCTION

Thick-walled tubular composite profiles with annular cross-section represent the optimal solution for 3-D large space frame trusses.

Within this context, glass fibre-reinforced polymers (GFRPs) are usually considered, due to the best match between expected costs and benefits, in combination with adhesive bonding solutions [1-4] between the composite member and standard co-axial nodal devices, generally made of stainless stell parts (Fig. 1).



Figure 1: Example of nodal device (spherical hinge).

Although internal stresses essentially accord to the axial regime, shear stresses originated by the interfacial interactions over the adhesive bonding zones can be present within the composite tube. As a consequence, the local behaviour of the composite profile is affected by shear stresses and, of course, the failure criterion should account for this aspect.

Considering the usually low values of shear moduli for GFRP, there is a great interest in investigating the complex interplay between I/II interfacial failure modes [5-7] and local shear deformations [8-10] which occur within the thickness of the composite member.

In this view, a mechanical model based on an appropriate kinematics has been proposed in combination with a mixed mode cohesive failure criterion. The aim is to predict both the structural response at the service conditions and the ultimate strength of the joint, thus allowing a refined structural analysis of the whole structure accounting for the local shear strains too.

In Figure 2 the equilibrium problem under discussion is exemplified with respect to a single composite tube bonded to apposite hinge nodal devices able to guarantee the internal connections among many parts of a generic space frame truss. Orthonormal unit vectors $\mathbf{i}_1$, $\mathbf{i}_2$ and $\mathbf{k}$ are introduced, with $\mathbf{k}$ aligned to the z-axis, as well as the point $\mathbf{O}$ as an origin. An axisymmetric normal stress distribution $p_z(r)$ is considered over the end cross-sections of the system, being $r$ the radial coordinate, as indicated in the following.

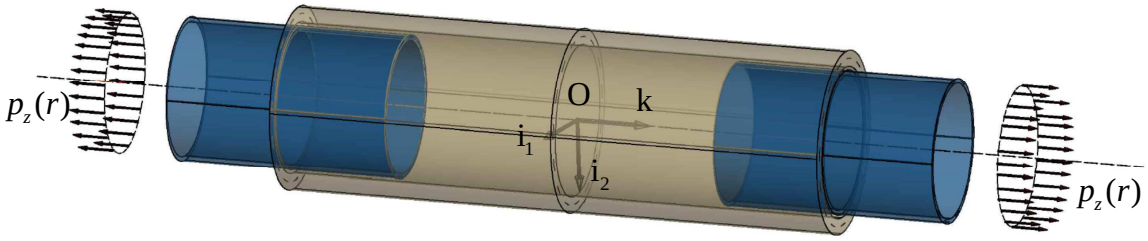



Figure 2: Equilibrium problem under consideration.

## 2. THE MECHANICAL MODEL

The mechanical model proposed for studying the adhesive bonding of tubular profiles is based on appropriate kinematic hypotheses, briefly summarised in the Appendix.

Three coaxial elements are considered: the GFRP annular pultruded profile and the nodal devices bonded at the ends of the tube. Symbols are as indicated in Figs. 3.a-b.

Orthotropic stress-strain relationships with transverse isotropy are chosen to simulate the constitutive response of the composite (see Appendix), while isotropic equations are used for what concerns the constitutive behaviour of the stainless steel.

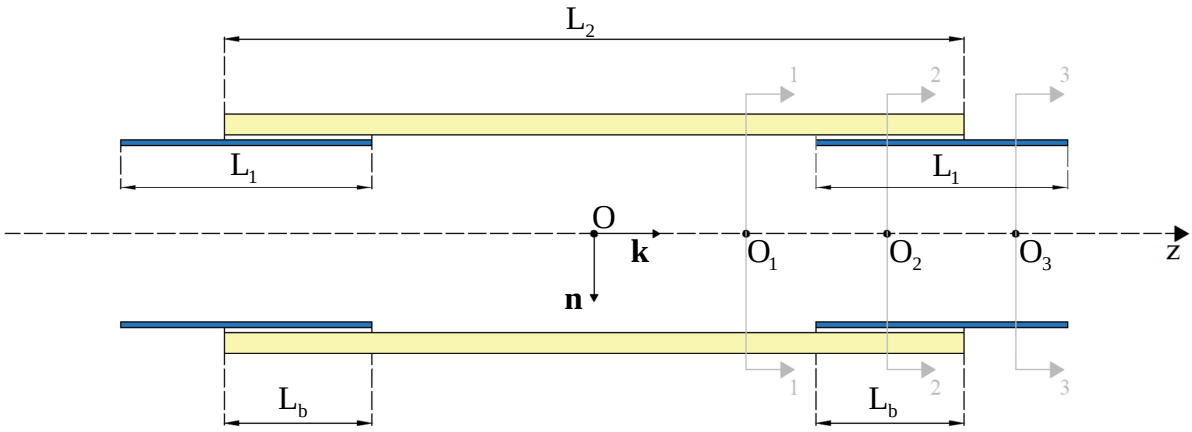



Figure 3.a : Geometric configuration (longitudinal section).

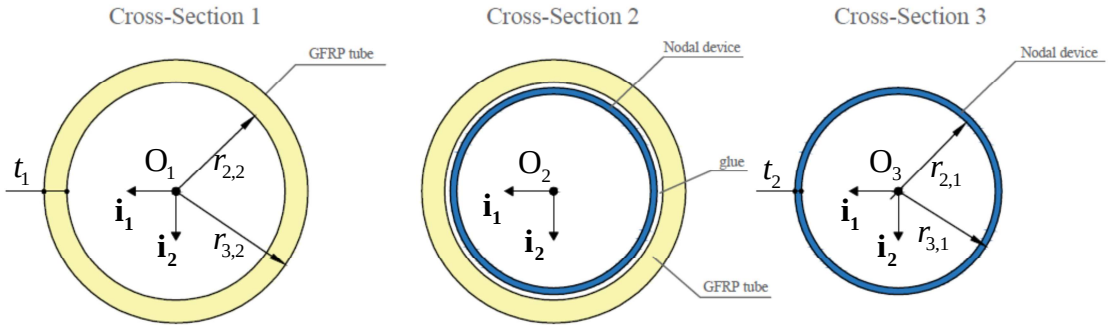



Figure 3.b : Geometric configuration (relevant cross-sections).

Due to axisymmetric condition, the adhesive bonding is modelled by a 2-D continuous distribution of cohesive forces acting over the bonding interfaces along both the normal ($\mathbf{n}$, radial) and the tangential ($\mathbf{k}$, axial) directions; these forces work to contrast the interfacial displacements between the composite tube and the nodal devices. The following additional symbols are introduced with this regard: $d_r$ and $t_{rr}$ denoting the interfacial normal displacement (i.e. discontinuous) and the interfacial normal traction (per unit surface) as well as $d_k$ and $t_{rk}$ indicating the analogous quantities with respect to the axial direction (Fig. 4).

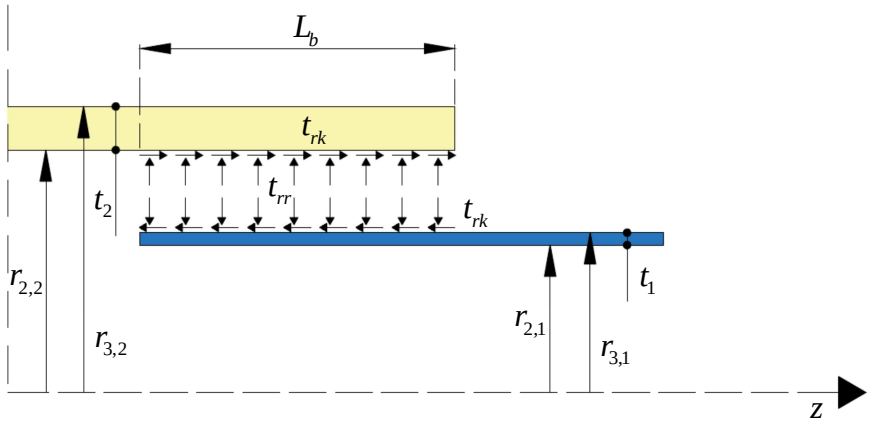



Figure 4 : Interfacial cohesive forces $t_{rr}$ and $t_{rk}$ (per unit surface).

An exponential cohesive interface model has been derived from that one proposed in [11-12] in order to include possible coupling between normal and tangential discontinuous displacements. The following potential has been considered:

$$F(h) = \Phi_{\mathrm{U}} \left[ 1 - \left( 1 + \frac{h}{h_c} \right) \mathrm{e}^{-(h/h_c)} \right] \tag{1}$$

In eqn. (1) the symbol $h$ indicates the norm of the vector $\mathbf{h}$:

$$\mathbf{h} = \lambda_{\mathrm{I}}\, d_r\, \mathbf{n} + \lambda_{\mathrm{II}}\, d_k\, \mathbf{k} \tag{2}$$

with $\lambda_{\mathrm{I}}$ and $\lambda_{\mathrm{II}}$ accounting for the coupling between the normal and tangential interfacial displacements. The corresponding interaction, $\mathbf{p}$, is assumed as follows:

$$\mathbf{p} = \frac{1}{\lambda_{\mathrm{I}}} t_{rr}\, \mathbf{n} + \frac{1}{\lambda_{\mathrm{II}}} t_{rk}\, \mathbf{k} = \mathrm{p} \frac{\mathbf{h}}{h} \tag{3}$$

where:

$$p = \frac{dF}{dh} = p_{\mathrm{c}} \frac{h}{h_{\mathrm{c}}} \mathrm{e}^{(1-h/h_{\mathrm{c}})} \tag{4}$$

The quantities $h_{\mathrm{c}}$ and $p_{\mathrm{c}}$ are parameters of the interface model, while $\Phi_{\mathrm{U}}$ represents the fracture energy per unit surface. It results:

$$h = \sqrt{\left(\lambda_{\mathrm{I}}\, d_r\right)^2 + \left(\lambda_{\mathrm{II}}\, d_k\right)^2} \qquad p = \sqrt{\left(\frac{t_{rr}}{\lambda_{\mathrm{I}}}\right)^2 + \left(\frac{t_{rk}}{\lambda_{\mathrm{II}}}\right)^2} \qquad \Phi_{\mathrm{U}} = \mathrm{e}\, p_{\mathrm{c}}\, h_{\mathrm{c}} \tag{5.a-c}$$

The cohesive model simulates the softening effect for $h > h_c$. Moreover, from a rigorous point of view, the full degradation is only asymptotic (for $h \to \infty$).

## 3. NUMERICAL ANALYSES

The mechanical model presented in the previous Section 2 has been approximated via a finite element technique in order to provide many numerical results of major relevance concerning the behaviour of a tubular composite profile bonded at the ends to apposite nodal devices.

The equilibrium problem indicated in Fig. 2 has been approached numerically by means of a finite element approximation. The mesh proposed is composed of 2 node finite elements characterized by 24 degrees of freedom, including 12 d.o.f. per each node.

According to the discussion summarized in the Appendix, let $w_{i,j}(z)$ be the axial displacement field of the j-part at $r = r_i$ (with j=1 indicating the nodal device while j=2 denoting the GFRP tube). The approximation of $w_{i,j}$ is obtained by cubic interpolating polynomials as follows:

$$w_{i,j}(\zeta) = \mathrm{h}_{10}(\zeta)\, w_{i,j}^{(1)} + \mathrm{h}_{11}(\zeta)\, w_{i,j}'^{(1)} + \mathrm{h}_{20}(\zeta)\, w_{i,j}^{(2)} + \mathrm{h}_{21}(\zeta)\, w_{i,j}'^{(2)} \tag{6}$$

where:

$$\mathrm{h}_{10} = \frac{1}{4}\left(2 - 3\zeta + \zeta^3\right) \qquad \mathrm{h}_{11} = \frac{l_\mathrm{e}}{8}\left(1 - \zeta - \zeta^2 + \zeta^3\right) \tag{7.a-b}$$

$$\mathrm{h}_{20} = \frac{1}{4}\left(2 + 3\zeta - \zeta^3\right) \qquad \mathrm{h}_{21} = \frac{l_\mathrm{e}}{8}\left(-1 - \zeta + \zeta^2 + \zeta^3\right) \tag{7.c-d}$$

being $l_e$ the length of the finite element; $w_{i,j}^{(1)}$ and $w_{i,j}^{(2)}$ the nodal values of the kinematic unknown under consideration while $w_{i,j}'^{(1)}$ and $w_{i,j}'^{(2)}$ the nodal values of the derivative of $w_{i,j}$ with respect to the axial coordinate z.

Using standard notations, the following generalized displacements vector $\mathbf{w}_{(e)}$, with dimensions 6×1, is introduced as a function of the nodal unknowns $\mathbf{U}_{(e)}$ over the generic finite element:

$$\mathbf{w}_{(e)} = \left[\left(w_{1,1}, w_{2,1}, w_{3,1}\right), \left(w_{1,2}, w_{2,2}, w_{3,2}\right)\right]^T = \mathbf{N}\mathbf{U}_{(e)} \tag{8}$$

In eqn. (8) the symbol $\mathbf{N}$ denotes the following four-block matrix, with dimensions 6×24:

$$\mathbf{N} = \left[\mathbf{N}_{10}, \mathbf{N}_{11}, \mathbf{N}_{20}, \mathbf{N}_{21}\right] \tag{9}$$

where:

$$\mathbf{N}_{\mathrm{pq}} = \mathrm{diag}\left(\mathrm{h}_{\mathrm{pq}}, \mathrm{h}_{\mathrm{pq}}, \mathrm{h}_{\mathrm{pq}}, \mathrm{h}_{\mathrm{pq}}, \mathrm{h}_{\mathrm{pq}}, \mathrm{h}_{\mathrm{pq}}\right) \qquad (\mathrm{p}=1,\ 2 \text{ and } \mathrm{q}=0,\ 1) \tag{10}$$

Furthermore, $\mathbf{U}_{(e)}$ is a numeric vector, with dimensions 24×1, which collects the values of the kinematic unknowns attained at both the nodes of the finite element:

$$\mathbf{U}_{(e)} = \left[\mathbf{U}_{(e,1)}^T, \mathbf{U}_{(e,2)}^T\right]^T \tag{11}$$

with $\mathbf{U}_{(e,n)}$ denoting the kinematic unknowns at the n node (n=1, 2):

$$\mathbf{U}_{(e,n)} = \left[ \left( w_{1,1}^{(n)}, w_{2,1}^{(n)}, w_{3,1}^{(n)}, w_{1,2}^{(n)}, w_{2,2}^{(n)}, w_{3,2}^{(n)} \right), \left( w_{1,1}^{\prime(n)}, w_{2,1}^{\prime(n)}, w_{3,1}^{\prime(n)}, w_{1,2}^{\prime(n)}, w_{2,2}^{\prime(n)}, w_{3,2}^{\prime(n)} \right) \right]^T \tag{12}$$

In Figure 5 the whole list of the nodal unknowns is identified.

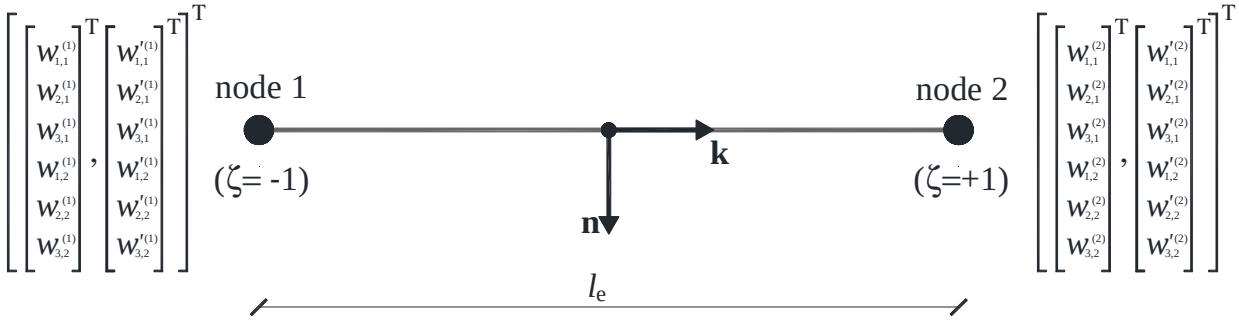



Figure 5 : Finite element.

A more accurate discussion deals with the approximation of the non-linear interfacial interactions. With reference to the generic finite element, the interfacial displacements along both the n-axis and the z-axis, $d_r$ and $d_k$, can be expressed as a function of the nodal unknowns $\mathbf{U}_{(e)}$. If Mode II only is present ($d_r = 0$), then $\mathbf{h} = \lambda_{II}\, d_k\, \mathbf{k}$. Under this assumption, the interfacial behaviour is represented by a unique cohesive relationship between $t_{rk}$ and $d_k$. It results:

$$d_k = \mathbf{DNU}_{(e)} \tag{13}$$

with $\mathbf{D}$ the following numeric vector, with dimensions 1×6:

$$\mathbf{D} = [0, 0, -1, 0, 1, 0] \tag{14}$$

Moreover, the interfacial interaction along the z-axis, $t_{rk}$, is:

$$t_{rk} = \kappa d_k \tag{15}$$

where the symbol $\kappa$ denotes the secant slope of the cohesive law:

$$\kappa = \lambda_{II} \frac{p_c}{h_c} e^{(1-\lambda_{II} d_k / h_c)} \tag{16}$$

A standard iterative procedure can thereby be adopted for searching the numerical solution. The convergence of the numerical solution to the continuous solution is discussed in [13].

As a case study, a numerical simulation has been carried out in order to underline the power of the proposed model. The FE analysis deals with a configuration made of two stainless steel terminal devices bonded at the ends of a GFRP tube. The following geometric assumptions are considered.

| | |
|---|---|
| inner radius of nodal device | ($r_{2,1}$ = 59.5 mm) |
| outer radius of nodal device | ($r_{3,1}$ = 63.5 mm) |
| thickness of nodal device wall | ($t_1 = r_{3,1} - r_{2,1}$ = 4.0 mm) |
| inner radius of GFRP tube | ($r_{2,2}$ = 66.0 mm) |
| outer radius of GFRP tube | ($r_{3,2}$ = 80.0 mm) |
| thickness of GFRP tube wall | ($t_2 = r_{3,2} - r_{2,2}$ = 14.0 mm) |
| length of the nodal device (coaxial part) | ($L_1$ =120.0 mm) |
| length of the GFRP tube | ($L_2$ = 500.0 mm) |
| length of the bonding zones | ($L_b$ = 100.0 mm) |
| thickness of the glue layer | ($t_a = r_{2,2} - r_{3,1}$ = 2.5 mm) |

Table 1: Geometric properties.

| | |
|---|---|
| transverse normal modulus | ($E_T$ =3700 N/mm$^2$) |
| longitudinal normal modulus | ($E_L$ =37000 N/mm$^2$) |
| shear modulus (**n** - **k** plane) | ($G_{TL}$ =1850 N/mm$^2$) |
| Poisson coefficients | ($\nu_{TT} = \nu_{TL} = \nu_{LT} = 0$) |

Table 2: GFRP constitutive properties.

| | |
|---|---|
| normal modulus | ($E$ =196000 N/mm$^2$) |
| shear modulus | ($G$ =98000 N/mm$^2$) |
| Poisson coefficient | ($\nu = 0$) |

Table 3: Stainless steel constitutive properties.

| | |
|---|---|
| fracture energy (per unit surface) | ($\Phi_U = 3.65 \times 10^{-2}$ Nmm/mm$^2$) |
| characteristic value of $h$ | ($h_c$ = 0.001 mm) |
| Mode I / Mode II interaction | ($\lambda_I = \lambda_{II} = 1$) |

Table 4: Parameters of the cohesive potential.

The loading condition considered is a resultant traction force equal to T=106000 N applied at the right end of the model, while the left end of the model is constrained to be fixed. From a mathematical point of view, the traction force has been simulated by means of two concentrated forces T/2 applied in a dual manner with respect to the following degrees of freedom, both relative to the steel device: $w_{2,1}$ and $w_{3,1}$. As discussed in [13], this is not rigorous in an absolute manner, but it can be acceptable for technical purposes when dealing with a moderately thick tubular wall. On the opposite end, the following degrees of freedom have been constrained to be equal to zero : $w_{1,1}$, $w_{2,1}$ and $w_{3,1}$.

The load intensity has been chosen in consideration that a preliminary experimental test has been recently performed, indicating the failure point of the bonded system exactly for T=106000 N, for a global elongation equal to 0.225 mm. Thereby, a numerical comparison in terms of interfacial energy absorption and global stiffness seems to be appropriate, in addition to a brief comment on the shear deformability of the composite tube.

The FE simulation has been carried out by means of a mesh composed of 540 finite elements with a constant length $l_e$ =1.0 mm. The convergence finally achieved corresponds to a residual vector with a norm equal to $9.82\times10^{-5}$.

The following Figure 6.a-d show the effects of the shear strains as a function of $r$ within the system at fixed values of the axial coordinate z. As it can be seen, these effects become relevant over the bonding zones ( $z_I \le |z| \le z_{III}$ with $z_I = 150$ mm and $z_{III} = 250$ mm ).

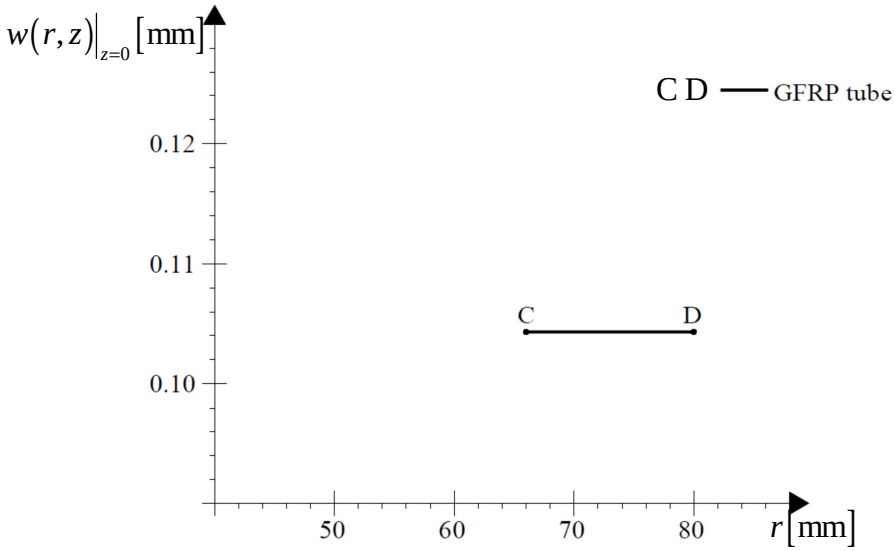



Figure 6.a : Axial displacement versus radial coordinate at $z = 0$ mm

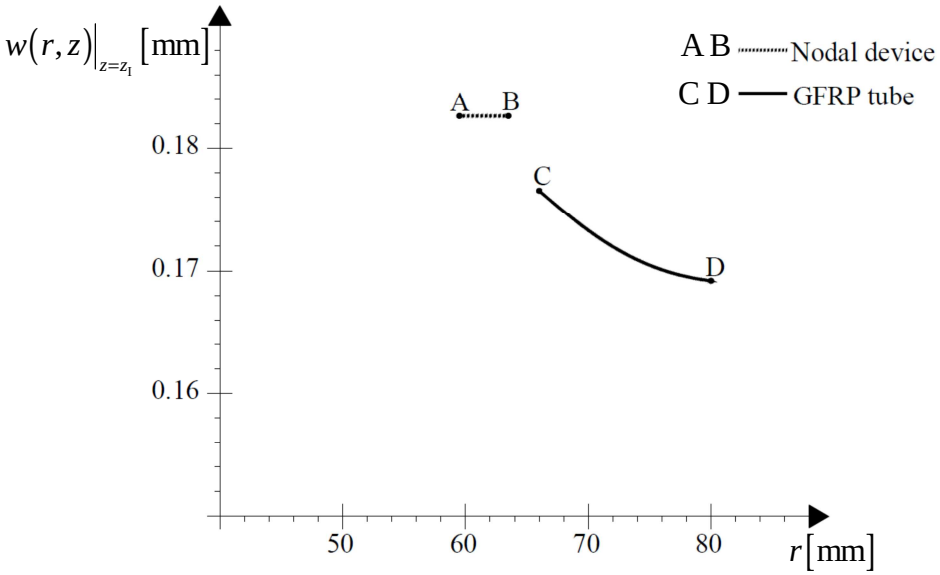



Figure 6.b : Axial displacement versus radial coordinate at $z = z_I = 150$ mm

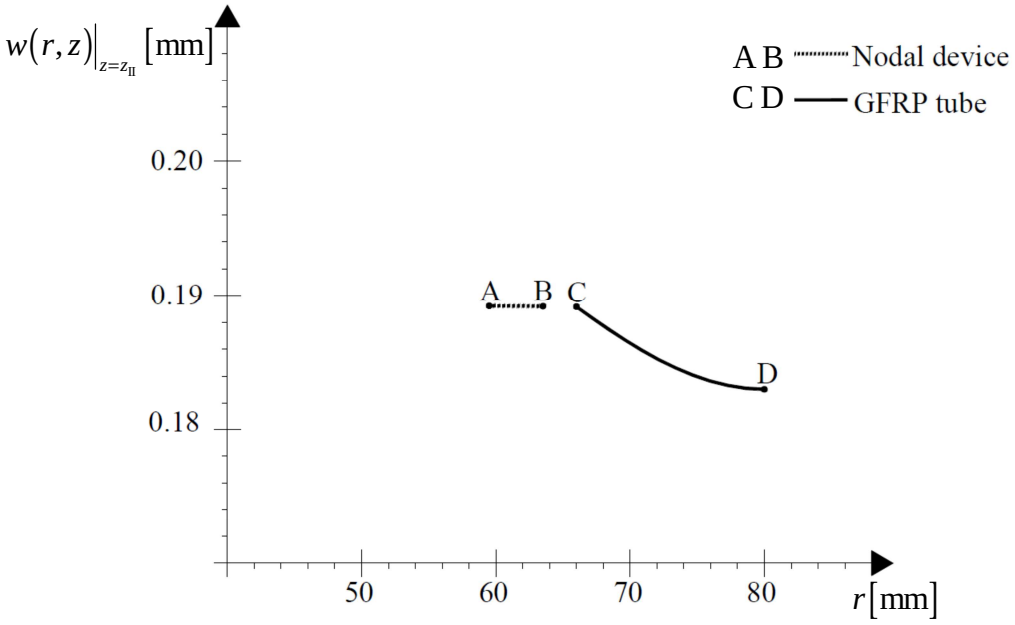



Figure 6.c : Axial displacement versus radial coordinate at $z = z_{II} = 200$ mm

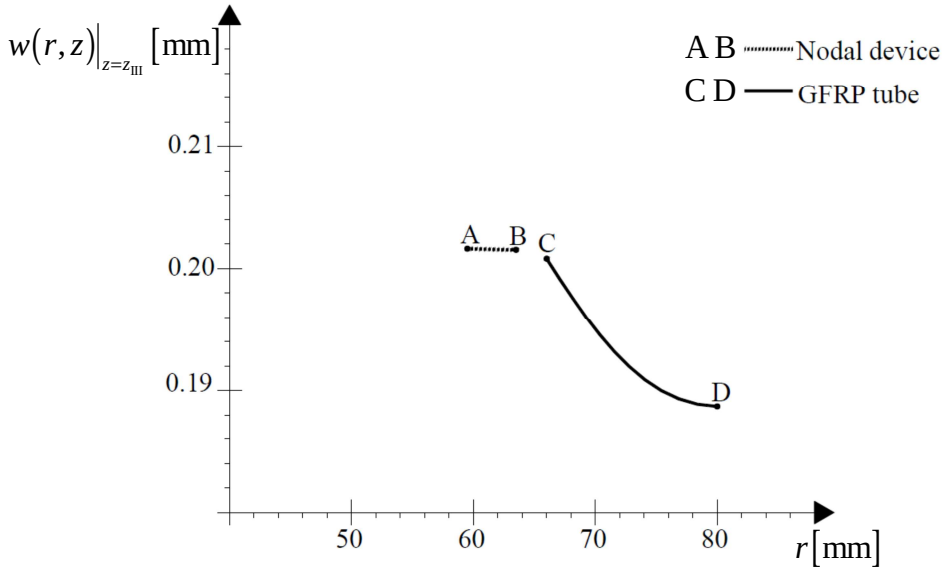



Figure 6.d : Axial displacement versus radial coordinate at $z = z_{III} = 250$ mm

In Figure 7, the interfacial cohesive force $t_{rk}$ (per unit surface) has been plotted over the bonding length. Due to the symmetric behaviour of both the bonding zones, only the interface at the right end has been considered ( $0 \le z^* \le L_b$ with $z^* = z - z_I$ , $z_I = 150$ mm and $L_b = 100$ mm ). The numerical analysis indicates that under the applied resultant traction force T=106000 N, the full degradation of the interface is substantially achieved at the axial coordinate $z_I = 150$ mm ( $z^* = 0$ mm ), as observed for the experimental test.

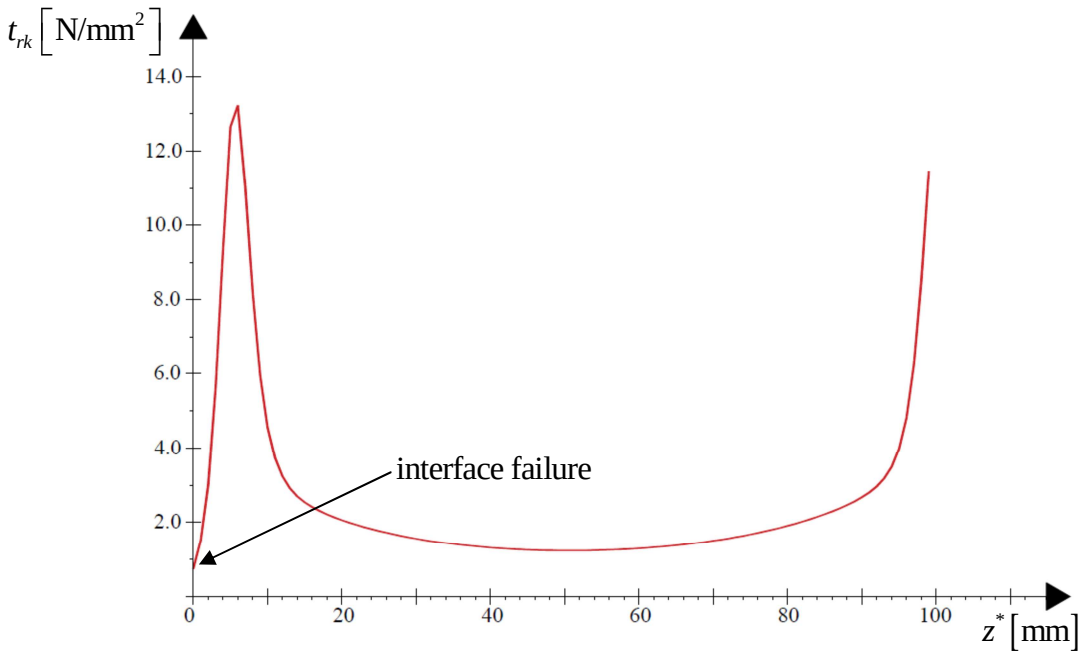



Figure 7: Interfacial cohesive forces $t_{rk}$ (per unit length) versus $z^* = z - z_1$.

Finally, the numerical prediction of the global elongation of the system is equal to 0.209 mm, substantially the same that the experimental value (0.225 mm).

The previous example has been presented for the pure scope of highlighting the main features of the proposed model. Further developments will concern more sophisticated approach to the interface behaviour, on line with what is proposed in [14-16], as well as the optimization strategy with regard to the design of complex spatial bridge following recent trends as discussed in [17-19].

## 4. CONCLUDING REMARKS

In this paper the authors have presented a mechanical model for the 1-D study of a composite tubular profile loaded in traction/compression and bonded to coaxial terminal devices able to guarantee the connection of the member within a general structure. This problem is highly relevant in practice, considering the context of large space frame trusses, where the advantage in using GFRP members with annular cross-section is universally acknowledged.

The proposed formulation is based on a cohesive approach for the modeling of the interface as well as on appropriate assumptions for the kinematics of the composite member, which accounts for possible shear strains within the thickness of the tubular wall. This aspect assumes a relevant role in the structural analysis of the member from both a point of view concerning the accurate prediction of the global displacements and with reference to the local behavior of the tube. Over the bonding length, in fact, shear strains appear on the lateral surface of the tube as an effect of the interface cohesive interactions between the tube and the nodal device. As a consequence, the interface behavior is affected from shear deformability of the tubular profile as well as the failure load of the system.

## 5. APPENDIX – KINEMATICS OF ORTHOTROPIC COMPOSITE TUBES

### 5.1. Global reference frame

According to the global reference frame shown in Fig. 8, the position of a generic point **P** is given by:

$$\mathbf{x} = r\cos\vartheta\mathbf{i}_1 + r\sin\vartheta\mathbf{i}_2 + z\mathbf{k} \tag{A.1}$$

where $r$, $\vartheta$ and $z$ are assumed as spatial variables. Due to axisymmetric geometry, the displacement field depends on $r$ and $z$ only. Moreover, the displacement field components are: $u_r$ (radial), $u_\vartheta = 0$ (circumferential, null), $w$ (axial).

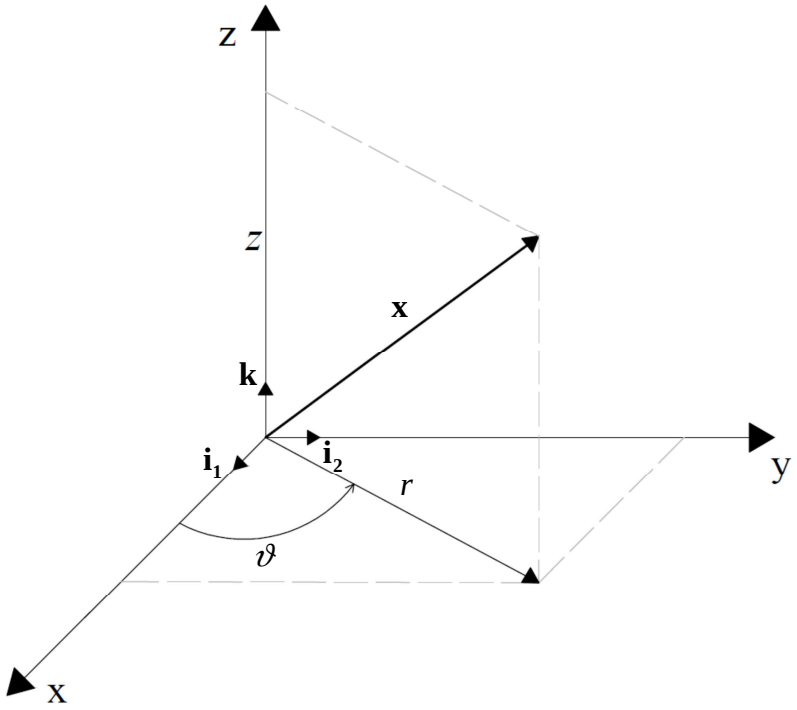



Figure 8 : Global reference frame.

### 5.2. Axial displacements

The axial displacement field is proposed in the following form:

$$w(r,z) = w_{\mathrm{i}}(z)\, f_{\mathrm{i}}(r) \quad (\mathrm{i} = 1,\ 2,\ 3) \tag{A.2}$$

where $w_{\mathrm{i}}(z)$ denotes the axial displacement field at a defined radial coordinate $r_{\mathrm{i}}$, while $f_{\mathrm{i}}(r)$ is a polynomial truncated at the second order terms.

The following assumptions are introduced:

$$r_1 = 0\,, \qquad r_2 = R_i\,, \qquad r_3 = R_e\,. \tag{A.3.a-c}$$

being $R_i$ and $R_e$ the inner and the outer radius of the annular cross-section (Fig.9).

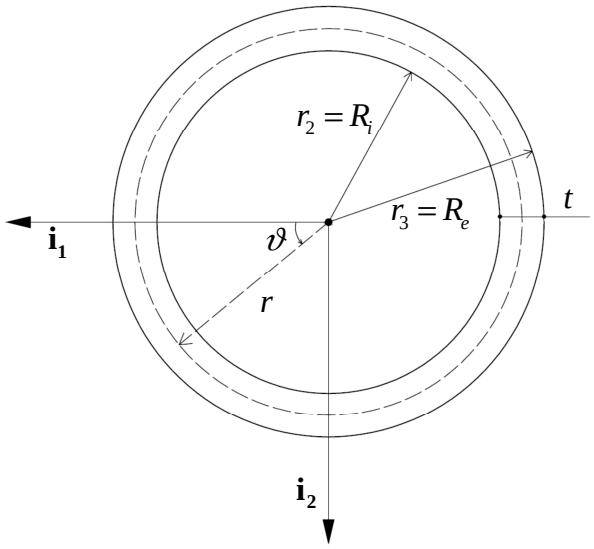



Figure 9 : Annular cross-section.

The polynomials $f_i(r)$ can be specified as indicated below:

$$f_1(r) = 1 - \left(\frac{1}{R_i} + \frac{1}{R_e}\right) r + \left(\frac{1}{R_i R_e}\right) r^2, \tag{A.4.a}$$

$$f_2(r) = \left(\frac{R_e}{R_i t}\right) r - \left(\frac{1}{R_i t}\right) r^2, \tag{A.4.b}$$

$$f_3(r) = -\left(\frac{R_i}{R_e t}\right) r + \left(\frac{1}{R_e t}\right) r^2, \tag{A.4.c}$$

being $t = R_e - R_i$ the thickness of the annular wall. It results:

$$\lim_{t\to 0} f_1 = 0, \qquad \lim_{t\to 0} f_2 = \frac{1}{2}, \qquad \lim_{t\to 0} f_3 = \frac{1}{2} \tag{A.5.a-c}$$

### 5.3. Strain and stress components

Due to the condition $u_\vartheta = 0$, the infinitesimal strain tensor assumes the following form:

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & 0 & \varepsilon_{13} \\ 0 & \varepsilon_2 & 0 \\ \varepsilon_{31} & 0 & \varepsilon_3 \end{bmatrix} \tag{A.6}$$

where:

$$\varepsilon_1 = \varepsilon_r = \frac{\partial u_r}{\partial r}, \qquad \varepsilon_2 = \varepsilon_\vartheta = \frac{u_r}{r}, \tag{A.7.a-b}$$

$$\varepsilon_3 = \varepsilon_z = \frac{\partial w}{\partial z}, \qquad \varepsilon_{13} = \varepsilon_{31} = \varepsilon_{rz} = \frac{1}{2}\left(\frac{\partial u_r}{\partial z} + \frac{\partial w}{\partial r}\right). \tag{A.7.c-d}$$

For orthotropic behaviour, it also results:

$$\begin{bmatrix} \varepsilon_r \\ \varepsilon_\vartheta \\ \varepsilon_z \\ \varepsilon_{rz} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & S_{13} & 0 \\ S_{21} & S_{22} & S_{23} & 0 \\ S_{31} & S_{32} & S_{33} & 0 \\ 0 & 0 & 0 & S_{44} \end{bmatrix} \begin{bmatrix} \sigma_r \\ \sigma_\vartheta \\ \sigma_z \\ \tau_{rz} \end{bmatrix}, \quad \begin{bmatrix} \sigma_r \\ \sigma_\vartheta \\ \sigma_z \\ \tau_{rz} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 \\ C_{21} & C_{22} & C_{23} & 0 \\ C_{31} & C_{32} & C_{33} & 0 \\ 0 & 0 & 0 & C_{44} \end{bmatrix} \begin{bmatrix} \varepsilon_r \\ \varepsilon_\vartheta \\ \varepsilon_z \\ \varepsilon_{rz} \end{bmatrix} \tag{A.8.a-b}$$

where:

$$S_{11} = \frac{1}{E_T} \qquad S_{12} = -\frac{\nu_{TT}}{E_T} \qquad S_{13} = -\frac{\nu_{LT}}{E_L} \tag{A.9.a-c}$$

$$S_{21} = -\frac{\nu_{TT}}{E_T} \qquad S_{22} = \frac{1}{E_T} \qquad S_{23} = -\frac{\nu_{LT}}{E_L} \tag{A.9.d-f}$$

$$S_{31} = -\frac{\nu_{TL}}{E_T} \qquad S_{32} = -\frac{\nu_{TL}}{E_T} \qquad S_{33} = \frac{1}{E_L} \tag{A.9.g-i}$$

$$S_{44} = \frac{1}{2G_{TL}} \tag{A.9.j}$$

$$C_{11} = \frac{1-\nu_{TL}\nu_{LT}}{E_T E_L \Delta}, \qquad C_{12} = \frac{\nu_{TT}+\nu_{LT}\nu_{TL}}{E_T E_L \Delta} \qquad C_{13} = \frac{\nu_{LT}+\nu_{TT}\nu_{LT}}{E_T E_L \Delta} \tag{A.10.a-c}$$

$$C_{21} = \frac{\nu_{TT}+\nu_{LT}\nu_{TL}}{E_T E_L \Delta} \qquad C_{22} = \frac{1-\nu_{TL}\nu_{LT}}{E_T E_L \Delta} \qquad C_{23} = \frac{\nu_{LT}+\nu_{TT}\nu_{LT}}{E_T E_L \Delta} \tag{A.10.d-f}$$

$$C_{31} = \frac{\nu_{LT}+\nu_{TT}\nu_{LT}}{E_T E_L \Delta} \qquad C_{32} = \frac{\nu_{LT}+\nu_{TT}\nu_{LT}}{E_T E_L \Delta} \qquad C_{33} = \frac{1-\nu_{TT}\nu_{TT}}{E_T E_T \Delta} \tag{A.10.g-i}$$

$$C_{44} = 2G_{TL} \tag{A.10.j}$$

with $\frac{\nu_{LT}}{E_L} = \frac{\nu_{TL}}{E_T}$, $\Delta = \frac{1-\nu_{TT}\nu_{TT}-2\nu_{TL}\nu_{LT}-2\nu_{TT}\nu_{TL}\nu_{LT}}{E_T E_T E_L}$, while $\mathbf{C} = \mathbf{S}^{-1}$. Moreover, the symbols "L" and "T" denote the axial and transverse directions.

### 5.4. Radial displacements

By appropriate steps:

$$\frac{\partial u_r}{\partial r} + a\frac{u_r}{r} + b w_i' f_i = 0 \tag{A.11}$$

where the symbol (.)' denotes the partial derivative with respect to z and the symbols “a” and “b” denote the following quantities:

$$a = \left(S_{1i}C_{i2}\right)/\left(1-S_{1i}C_{i1}\right) \qquad b = \left(S_{1i}C_{i3}\right)/\left(1-S_{1i}C_{i1}\right) \tag{A.13.a-b}$$

In eqs. (A.13) the Einstein summation has been considered.

It is worth noting that for thin or moderately thick wall annular cross-section, the assumption $\sigma_r \cong 0 \;\; \forall r \in \left[R_i, R_e\right]$ can be accepted if both the inner and the outer lateral surface are free of forces. This allows to specialize the values of a and b:

$$a = \frac{\nu_{TT} + \nu_{TL}\nu_{LT}}{1-\nu_{TL}\nu_{LT}} \qquad b = \frac{\nu_{LT} + \nu_{TT}\nu_{LT}}{1-\nu_{TL}\nu_{LT}} \tag{A.14.a-b}$$

A solution of eqn. (A.11) is searched according to the following form:

$$u_r(r,z) = \mathrm{q_i} w_i(z) r^{-a} + w_i'(z) g_i(r) \tag{A.15}$$

where $\mathrm{q_i} = 1$ while $g_i(r)$ (i=1, 2 , 3) are still unspecified functions of $r$.

Now eqn. (A.11) can be rewritten accounting for eqn. (A.15), thus giving:

$$w_i'(z)\left[\frac{dg_i(r)}{dr} + a\frac{g_i(r)}{r} + bf_i(r)\right] = 0 \tag{A.16}$$

By integrating the first order ordinary differential equations $\frac{dg_i(r)}{dr} + a\frac{g_i(r)}{r} + bf_i(r) = 0$ (i=1, 2, 3), the final expressions of $g_i(r)$ are obtained:

$$g_i(r) = \frac{1}{r^a}\left[c_i - b\int f_i(r) r^a dr\right] \tag{A.17}$$

with $c_i$ (i=1, 2, 3) to be determined by means of appropriate boundary conditions, as discussed in [13].

### 5.5. Explicit strain components

From eqs. (A.7.a-d), considering eqn. (A.2) and eqn. (A.15), it is easy to obtain the explicit form of the local strains ( $a \neq 0$ ):

$$\varepsilon_r = \mathrm{q_i} w_i(z) r^{-(a+1)} + w_i'(z)\frac{\partial g_i(r)}{\partial r}, \qquad \varepsilon_\vartheta = \mathrm{q_i} w_i(z) r^{-(a+1)} + w_i'(z)\frac{g_i(r)}{r}, \tag{A.18.a-b}$$

$$\varepsilon_z = w_i'(z) f_i(r), \qquad \varepsilon_{rz} = \frac{1}{2}\left(\mathrm{q_i} w_i'(z) r^{-a} + w_i''(z) g_i(r) + w_i(z)\frac{\partial f_i(r)}{\partial r}\right) \tag{A.18.c-d}$$

If $a = 0$, then eqn. (A.15) becomes

$$u_r(r,z) = \mathrm{q_i} w_i(z) + w_i'(z) g_i(r) \tag{A.19}$$

As a consequence, eqs. (A.18.a-d) are replaced by:

$$\varepsilon_r = -b w_i'(z) f_i(r) \qquad \varepsilon_\vartheta = \mathrm{q_i} w_i(z)\frac{1}{r} + w_i'(z)\frac{g_i(r)}{r} \tag{A.20.a-b}$$

$$\varepsilon_z = w_{\mathrm{i}}'(z)\, f_{\mathrm{i}}(r) \qquad \varepsilon_{rz} = \frac{1}{2}\left( \mathrm{q}_{\mathrm{i}} w_{\mathrm{i}}'(z) + w_{\mathrm{i}}''(z) g_{\mathrm{i}}(r) + w_{\mathrm{i}}(z) \frac{\partial f_{\mathrm{i}}(r)}{\partial r} \right) \quad \text{(A.20.c-d)}$$

If $a = 0$ and $b = 0$, eqn. (A11) indicates that $u_r$ does not depend on the radial coordinate $r$. Considering the condition:

$$\lim_{r \to 0} u_r = 0 \quad \text{(A.21)}$$

it results:

$$u_r = 0 \quad \text{(A.22)}$$

Finally:

$$\varepsilon_r = 0 \qquad \varepsilon_{\vartheta} = 0 \quad \text{(A.23.a-b)}$$

$$\varepsilon_z = w_{\mathrm{i}}'(z)\, f_{\mathrm{i}}(r) \qquad \varepsilon_{rz} = \frac{1}{2}\left( w_{\mathrm{i}}(z) \frac{\partial f_{\mathrm{i}}(r)}{\partial r} \right) \quad \text{(A.23.c-d)}$$

More in detail, eqn. (A.23.d) indicates that the shear strain $\varepsilon_{rz}$ is a linear function of $r$. This last simplification has been considered with reference to the numerical example presented in Section 3.

## REFERENCES


[1] M.S. Kafkalidis, M.D. Thouless, The effects of geometry and material properties on the fracture of single lap-shear joints. *International Journal of Solids and Structures*, **31**, 2537-2563, 2002.

[2] A.T. Abdelaziz, R. Boukhili, S. Achiou, S. Gordon, H. Boukehili, Bonded joints with composite adherends. Part I. Effect of specimen configuration, adhesive thickness, spew fillet and adherent stiffness on fracture. *International Journal of Adhesion and Adhesives*, **26**, 226-236, 2006.

[3] A.T. Abdelaziz, R. Boukhili, S. Achiou, H. Boukehili, Bonded joints with composite adherents. Part II. Finite element analysis of joggle lap joints. *International Journal of Adhesion and Adhesives*, **26**, 237-248, 2006.

[4] L.F.M. da Silva, R.D. Adams, Joint strength predictions for adhesive joints to be used over a wide temperature range. *International Journal of Adhesion and Adhesives*, **27**, 362-379, 2007.

[5] F. Ascione, G. Mancusi, Axial/Bending coupled analysis for FRP adhesive joints. *Mechanics of Advanced Materials and Structures*, **17**, 1–14, 2010.

[6] F. Ascione, G. Mancusi, Failure criteria for FRP adhesive lap joints: a comparative anal ysis. *Mechanics of Advanced Materials and Structures*, **17**,157–164, 2010.

[7] G. Mancusi, F .Ascione, Performance at collapse of adhesive bonding. *Composite Structures*, **96** ,256–261, 2013.

[8] L. Ascione, L. Feo, G. Mancusi, On the statical behaviour of fibre-reinforced polymer thin-walled beams, *Composites Part B: Engineering*, **31** (8), 643-654, 2000.

[9] L. Feo, G. Mancusi, Modeling shear deformability of thin-walled composite beams with open cross-section, *Mechanics Research Communications*, **37** (3), 320-325, 2010.

[10] L. Feo, G. Mancusi, The influence of the shear deformations on the local stress state of pultruded composite profiles. *Mechanics Research Communications* **47,** 44– 49, 2013.

[11] J.H. Rose, J. Ferrante, J.R. Smith, Universal binding energy curves for metals and bimetallic interfaces. *Physical Review Letter*, **47**, 675–678, 1981.

[12] J.H. Rose, J.R. Smith, J. Ferrante, Universal features of bonding in metals. *Physical Review B*, **28**,1835–1845, 1983.

[13] G. Mancusi, Elastic Behaviour of Thick-Walled Composite Tubes. Submitted to *Composite Structures*.

[14] F. Fraternali, Free Discontinuity Finite Element Models in Two-Dimensions for In-Plane Crack Problems. *Theoretichal and applied fracture mechanics*, **47**, 274-282, 2007.

[15] F. Fraternali, M. Negri, M. Ortiz, On the convergence of 3D free discontinuity models in variational fracture. *International Journal of Fracture*, **166** (1-2), 3-11, 2010.

[16] M. Angelillo, E. Babilio, A. Fortunato, Numerical solutions for crack growth based on the variational theory of fracture. *Computational Mechanics*, **50** (3), 285-301, 2012.

[17] G. Carpentieri, M. Modano, F. Fabbrocino, F. Fraternali, Optimal design and dynamics of truss bridges. *5th Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering (COMPDYN-ECCOMAS2015)*, 1731-1740, 2015.

[18] G. Carpentieri, R.E. Skelton, F. Fraternali, Minimum mass and optimal complexity of planar tensegrity bridges. *International Journal of Space Structures*, **30** (3-4), 221-243, 2015.

[19] R.E. Skelton, F. Fraternali, G. Carpentieri, A. Micheletti, Minimum mass design of tensegrity bridges with parametric architecture and multiscale complexity. *Mechanics Research Communications*, **5**8, 124-132, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# TOWARDS REAL-TIME STRUCTURAL HEALTH MONITORING DAMAGE DETECTION WITHOUT USER INPUT

**M.S. Salmanpour*, Z. Sharif Khodaei, and M.H. Aliabadi**

Department of Aeronautics, Imperial College London
South Kensington Campus, SW72AZ, London, UK.
{m.salmanpour, z.sharif-khodaei, m.h.aliabadi}@imperial.ac.uk

**Keywords:** Lamb wave, Piezoelectric, Ultrasonic, Finite elements, Composite material, Vibration test.

**Abstract.** *In this work a real-time damage detection platform is presented, requiring little to no user input after initial installation and set-up. Diagnostic ultrasonic signals are generated using attached piezoelectric transducers, which also serve to capture the structural response. This paper shows real-time detection in a flat CFRP panel. The necessary data acquisition and signal processing is carried out in an automated manner. A visualization of the damage map is then given as the primary output highlighting the predicted damage location. The developed system is flexible in allowing scalable deployment to cater for an increased number of transducers. The detection platform is experimentally demonstrated by real-time localization of artificial damages at various locations on a CFRP panel. This was also done under operational environment vibration loading, yielding accurate damage localization.*

## 1 INTRODUCTION

There has been an increase in the use of Carbon fiber reinforced polymer (CFRP) composites for aerospace applications due to their superior properties in comparison to metallic structures. However, stringent requirements on damage tolerance are a major factor in preventing greater adoption of composites in this sector. Barely visible impact damage (BVID) is a damage type of huge concern. As the name suggests these damages may not be readily detectable with visual inspections. Non-destructive testing (NDT) methods can locate and characterize damage. However, these often require operational downtime, are labor intensive and performed by highly trained specialists. With a spars network of permanently attached transducers, guided wave structural health monitoring (SHM) can be used for effective inspection of large plate like structures. In contrast to other NDT methods SHM approaches offer in-situ diagnostics with minimal intrusion to the operation schedule.

With low mass and small overall size piezoelectric patches are often used both actuators and sensors in SHM. These allow for excitation of Lamb wave guided wave signal within the structure and measurement of the response. Propagation of Lamb waves are highly sensitive to geometric features. Thus through analysis of wave features it is possible to detect and localize the damages. Many researchers have demonstrated the effectiveness of the baseline comparison approaches for detection of damage presence [1-3]. Delay and sum methods utilize key wave features in diagnostic signals; including time of flight (TOF), velocity and damage scatter, to predict the location of damage. Information from all transducers within the transducer network is fused to assemble a damage prediction map highlighting the location of damage. The delay and sum detection method adopted centers on comparison of the current or damaged signal with those recorded in a pristine condition. Through this baseline comparison, damage scatter features can be isolated. Assuming any residual between the baseline and current signal are due damage, the TOF and envelope of the damage can be manipulated to localize the damage.

In previous work diagnostic signal acquisition has been accomplished with manual channel switching and also with automated switching systems. These have suffered from deteriorated signal quality including signal crosstalk, low signal to noise ratio or repeatability when a signal amplifier was not used [4, 5]. In this work a real-time automated damage detection platform is presented, requiring little to no user input after initial installation and set-up. Diagnostic ultrasonic signals are generated using attached piezoelectric transducers. The damage detection approach used has previously been shown to be effective in detecting BVID in curved stiffened CFRP panels [6]. This paper shows real-time damage detection in a flat CFRP panel. Using a signal switching system integrated into a high performance acquisition platform, all the necessary data acquisition and signal processing is carried out in an automated manner. The tone burst actuation is applied to the transducers in turn and pitch-catch signals are obtained with minimal interference, by exploiting the switch architecture. A visualization of the damage map is then given as the primary output highlighting the prediction of damage location. The developed system is flexible in allowing scalable deployment to cater for an increased number of transducers.

The SHM system developed must be demonstrated under the environmental and operational conditions, in ordered to be deployed on aircraft. The RTCA DO-160 and MIL-STD 810 [7, 8] provide testing conditions and procedures that must be tailored to the relevant operational environment. The operational environments that are relevant for SHM systems include humidity, temperature, lightning strike, altitude, shock, ice formation and vibration. The main novelty in this work is demonstration of the integrity and performance of the developed real-time damage detection system under operational vibration loading.

The detection platform is first verified using experimental results to detect a softening area in a CFRP panel. It is then experimentally demonstrated by real-time detection and localization of artificial damages at various locations on a CFRP panel. Finally, the vibration tests are performed and the results are reported.

## 2 FINITE ELEMENT SIMUALTION

Computational methods have been used for simulation of wave propagation and finite element (FE) methods have proven as an effective development tool [2, 9]. The Abaqus explicit FE package was used to simulate diagnostic signals in a composite plate shown in Figure 1. The layup was modeled as $[0,45,-45,90]_{2s}$ with an overall size of 300 x 225 x 2 mm. Only the PZT part of the six DurAct transducers were modeled using 3D solid elements. These were attached to the plate with surface tie constraints. The composite plate was modeled with S4R shell elements. It has been previously suggested that an element size of at least 20 nodes per wavelengths (NPW) are required to ensure a converged solutions [2, 9]. A NPW of 30 was used by setting an element sizes to 3.42 E-4 and 1.98 E-4 for the plate and PZTs respectively. Damage was modeled as a circular softening area of radius 5 mm with a 50% reduction in in in the local stiffness.

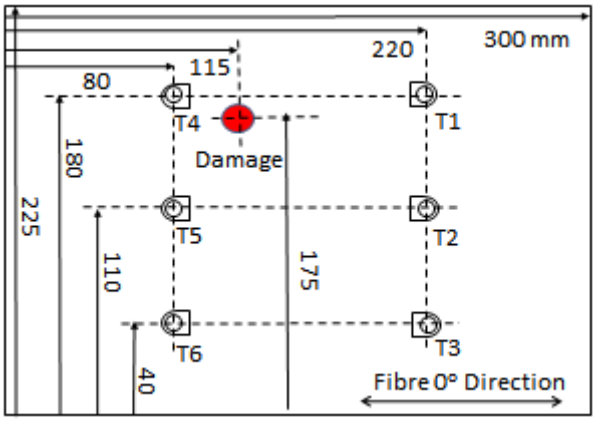



**Figure 1: Schematic of composite plate used for damage detection.**

A five cycled tone burst signal of central frequency 300 kHz was used as the actuation signal $V$ :

$$V = \left(1 - cos\left(2\frac{\pi ft}{n}\right)\right) sin(2\pi ft) H(\tau) \tag{1}$$

Where $f$ is the central frequency, $n$ is the number of cycles, $t$ is the time and $H$ is the Heaviside function.

The actuation was applied as a radial displacement to the top circumference of the PZTs, Figure 2. Each transducers was excited in turn as the actuator, while the response at the other transducers were recorded at a 1000 time points over the simulation time of 200 μs.

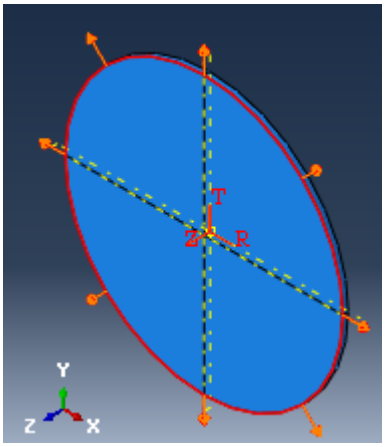



**Figure 2: Actuation of piezoelectric patches with radial displacement.**

## 3 EXPERIMENTAL SETUP

The developed real time detection system was built around a National Instrument (NI) instrumentation and software. Utilizing the available hardware, as outlined below, it was possible to collect all the necessary diagnostic signals from the structure under interrogation in under a minute. Immediately after the data was collected the damage detection algorithm was performed, providing a visualization of the predicted damage location.

### 3.1 Signal Generation and Acquisition

The data acquisition system was built around the NI PCI eXtensions for Instrumentation (PXI) system. For purposes of generating ultrasonic signals, a PXIe 5412 single channel arbitrary voltage generator card was used. This was operated, without the need for a standalone amplifier, at an amplitude of 12 volts. The voltage response of the PZTs were then captured and digitized using a PXIe 5105 digital oscilloscope card. To allow for time efficient and automated acquisition of the required signals, a Pickering 40-726A switch was used. This facilitated channeling of the actuation voltage to the correct transducer and capture of the voltage responses of the rest of the transducer network. Each transducer was excited with a tone burst signal of central frequency 300 kHz with voltage defined in equation 1. The response of the plate was recorded as the voltage output of the transducers.

### 3.2 Cross-Talk and Switch Architecture

The actuation voltage was in the order of tens of volts while the voltage response of the transducers in the order of tens to a hundreds of millivolts. This disparity in the voltage scales of the input and the output exacerbates the effects of electromagnetic interference between these two sets of signals. This so called crosstalk can be mitigated by wire shielding, but typically the close physical proximity of the circuitry within switch card brings about significant signal crosstalk as shown in Figure 3a.

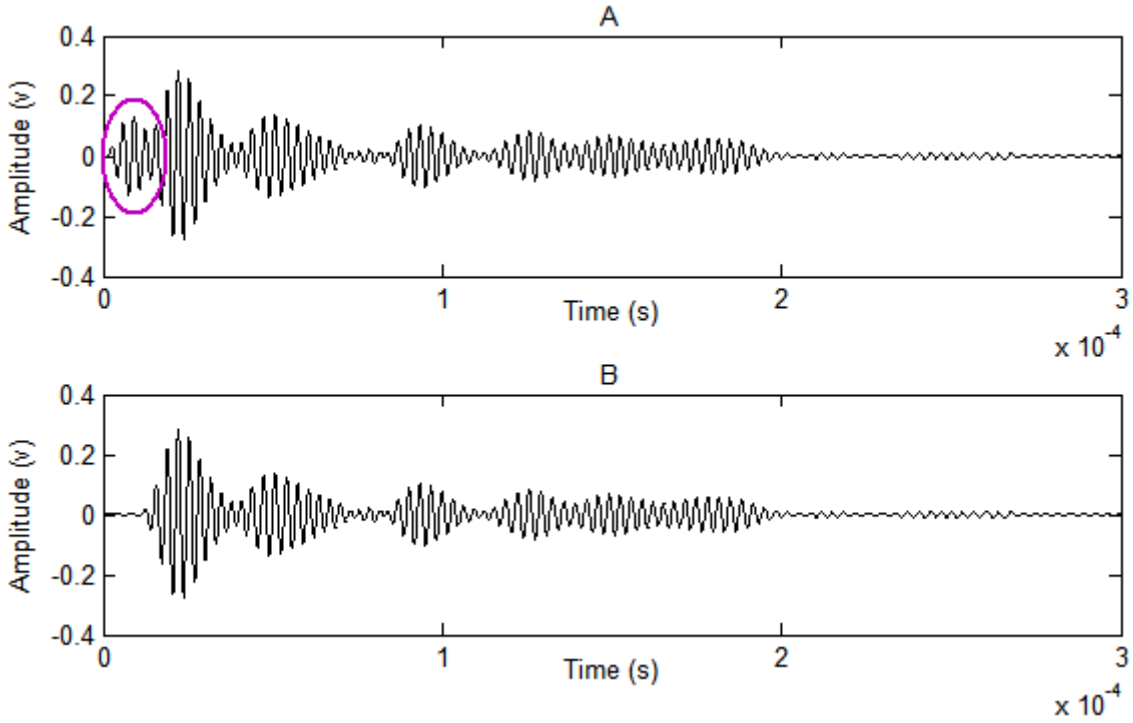



**Figure 3: (A) Sensor voltage response with significant crosstalk (highlighted in purple ellipse), (B) signal without crosstalk obtained by exploiting switch architechture.**

The Pickering 12 x 8 RF switch caters for the interconnection of 8 Y (instrument) channels with 12 X (PZT) channels as illustrated in Figure 4. The internal architecture actually consisted of four smaller sub-switches, allowing for decoupling of the relatively high voltage input from weaker output signal. Thus sub-switch two was used exclusively for sensing, with sub-switch three reserved for actuation. This interference free signal could be obtained at the cost of a reduced number of instrument channels, meaning that four oscilloscope channels can be used at any one time. This would not result in a reduced overall PZT count, but rather an increase in the overall time to record all the voltages, as PZTs signals can be recorded in sets of four. It thould be noted that separate switching cards can be interconnected. Hence the number of PZTs in the network is limited by the the avaliable number of switch cards.

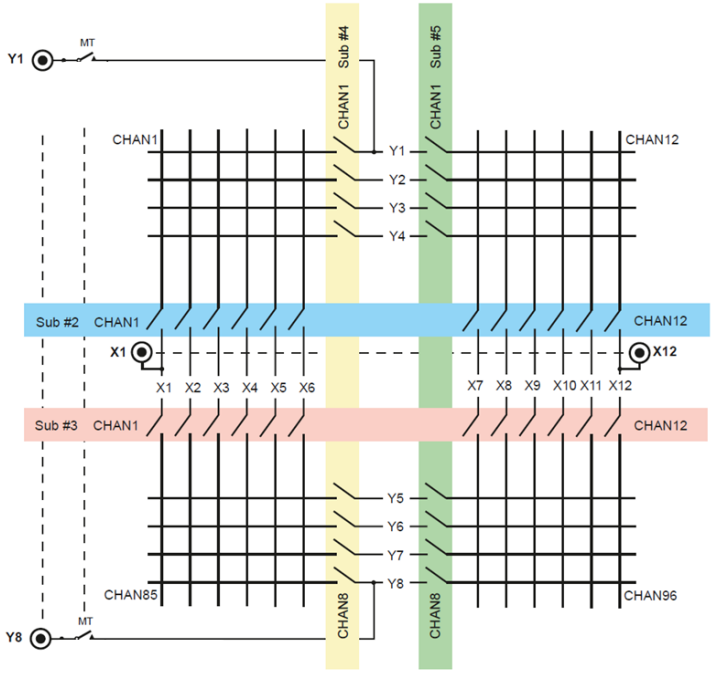



**Figure 4: Schematic of switch architecture, adapted from [10].**

### 3.3 Plate Definitions

The structure under interrogation was a composite panel of size 300 x 225 mm shown in Figure 1. The CFRP panel consisted of 16 unidirectional Fiberdux 914 TS 5 plies with layup $[0,45,-45,90]_{2s}$ and cured thickness of 2 mm. A total of six DurAct transducers were permanently bonded to the plate with Hexcel Redux film adhesive.

Researchers [11, 12] have used intuitive methods of simulating damage, including mastic. Though artificial damage does not fully represent actual damage, it provides a means by which to cause scattering of waves at a particular location. This allows for development of methods for finding the location of the source of this additional wave scatter. In this work BluTak mastic with a nominal diameter of 10 mm was used as artificial damage.

### 3.4 Operational Environment: Vibration Test Setup

The vibration profile applied during this test was selected from the RTCA DO-160 certification requirements. This profile corresponded to that of the operational environment of an electronic components attached to the fuselage of a fixed wing aircraft with a turbofan propulsion system. The setup consisted of a TMS shaker with power amplifier, ICP high sensitivity accelerometer used as control accelerometer mounted on the fixture and a four channel control unit as shown in Figure 5.

The vibration acceleration power spectral density (APSD) was set as per the requirements with slopes of 6 dB/octave and a grms value of 4.12. The plate was mounted into a fixture representative of a fuselage panel spaced within structural ribs. The tests were performed for one hour in the z-axis. The tests were set to abort if the APSD deviated from the set profile by more than 3 dB. Diagnostic signal was recorded with under no vibration before and after the tests. While the plate was subjected to vibrational loading signals were recorded with and without the simulated damage.

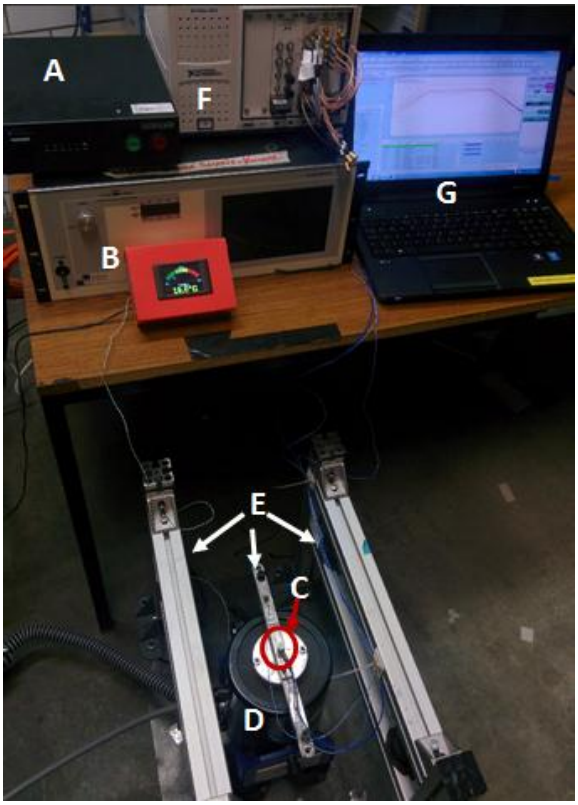



**Figure 5: Setup used for vibration tests, (A) controller, (B) shaker power amplifier, (C) control accelerometer, (D) shaker, (E) fixture, (F) real-time detection platform and (G) workstation.**

## 4 RESULTS

### 4.1 Computational Results

The obtained FE results were used as inputs to the detection system, mimicking the data acquisition system for the initial verification. The location of the 5 mm radius softening damage was accurately predicted with a localization error of less than 9 mm. The damage prediction map has been normalized to its peak value with a range of 0 to 1. It should be noted that the region with peak index is localized to an area around the actual damage location.

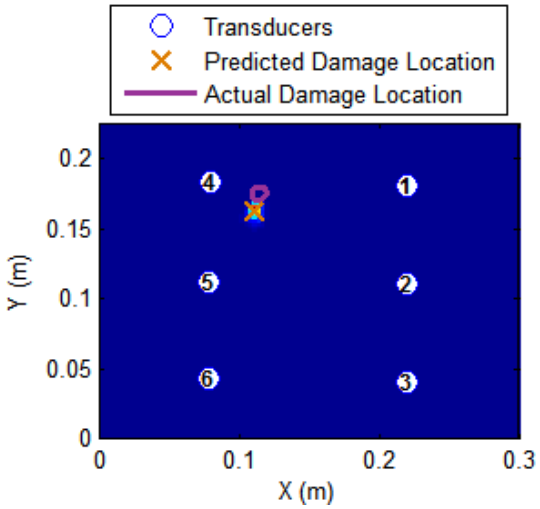



**Figure 6: Damage location prediction of a softening region with FE simulated results.**

## 4.2 Experimental Results

The artificial damage was located at different locations and localization predictions shown below. For damage located outside the transducer area the location prediction as not accurate. It should be added that the peak index values did provide a coarse indication of the general location of the simulated damage.

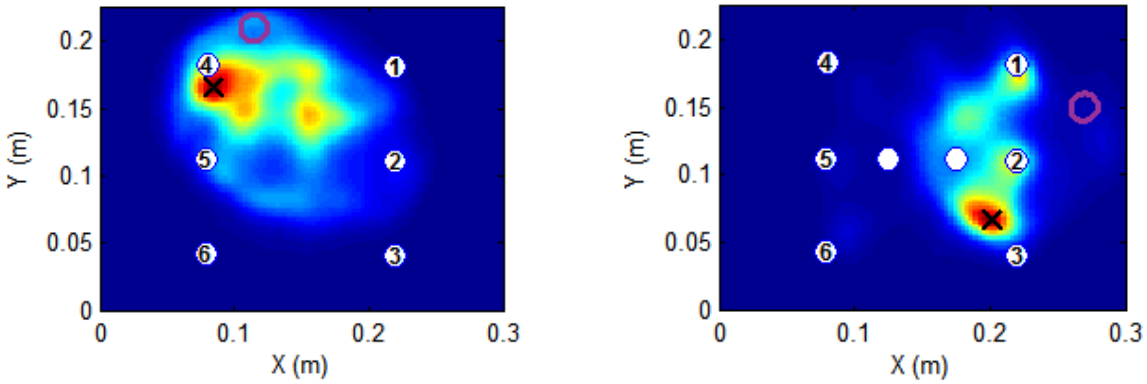



**Figure 7: Real time damage location prediction maps for simulated damage outside the transducer area on a composite panel.**

It was found that for the positions within the enclosed transducer area the localization was very accurate as shown in Figure 8. Furthermore, areas with highest index were confined to locations near the actual damage position.

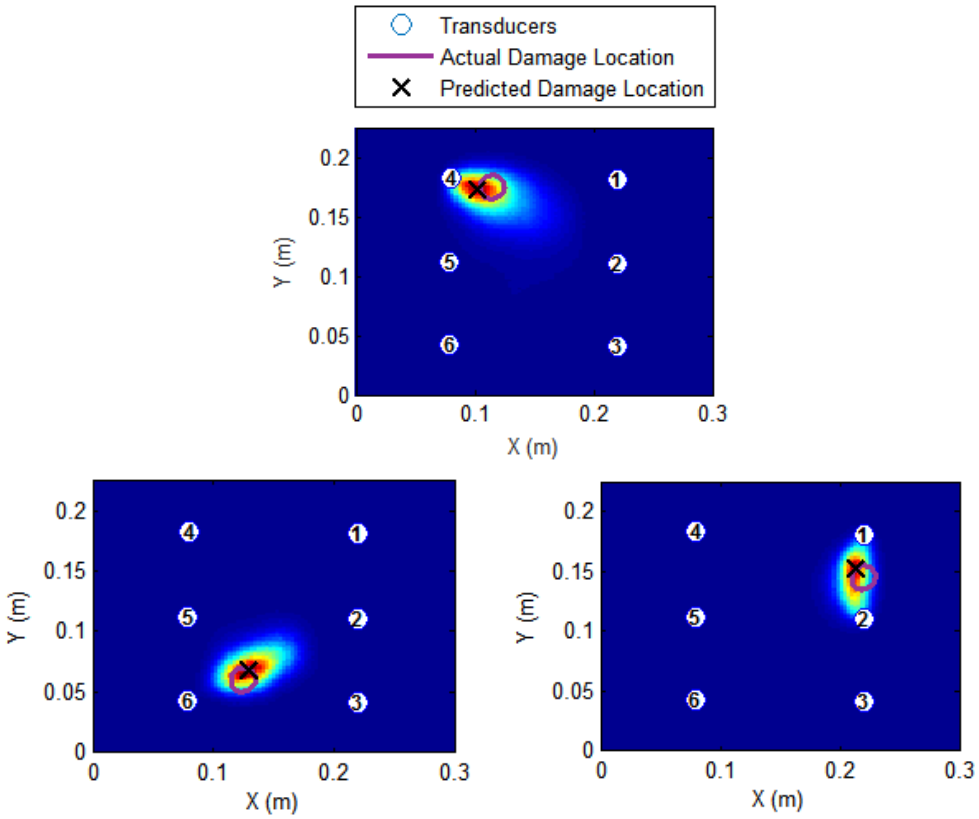



**Figure 8: Real time damage location prediction maps for simulated damage within the transducer area on a composite panel.**

### 4.3 Vibration Tests

The plate was subjected to the vibration profile for the full duration without the APSD abort limit being exceed. Real-time damage detection was performed at various intervals during the vibration test. The final results obtained at nearly 59 minutes of vibration, highlighted accurate damage location prediction with localization of error of 7 mm, as depicted in Figure 9. It should be noted that after the vibration test a similar level of detection accuracy could be attained. Also, visual and electro-mechanical impedance inspections did not highlight any loss in integrity of the transducers or wiring.

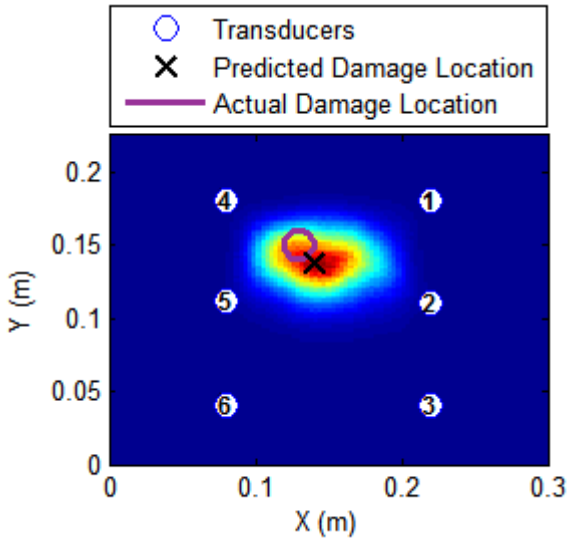



**Figure 9: Real-time damage detection under vibration loading**

## 5 CONCLUSIONS

It was shown that the developed real-time damage detection system is able to record the all the diagnostic signal required and perform the necessary calculation for damage detection in a matter of minutes. The system was first verified using FE simulated results to localize a softening region in CFRP plate. The real-rime detection system was applied to six transducer plate to localize artificial damage. It was able to accurate predict the location of artificial damages within the area enclosed by the transducers.

It was also shown that the permanently attached transducer network was able to maintain its integrity throughout the vibration tests specified for a fixed wing aircraft with a turbofan propulsion system. There was also no significant reduction in performance of the damage location prediction during or after the test. Additional tests will be performed with actual BVID cases under other operation and environmental conditions including temperature and humidity, temperature shock, altitude and icing.

## REFERENCES

[1] Michaels, J.E. and T.E. Michaels, *Guided wave signal processing and image fusion for in situ damage localization in plates.* Wave Motion, 2007. **44**(6): p. 482-492.

[2] Sharif-Khodaei, Z. and M.H. Aliabadi, *Assessment of delay-and-sum algorithms for damage detection in aluminium and composite plates.* Smart Materials and Structures, 2014. **23**(7): p. 075007.

[3] Flynn, E.B. *A Bayesian experimental design approach to structural health monitoring with application to ultrasonic guided waves*. Ph. D thesis, UC San Diego. 2010. California.

[4] Meng, S.Y. and M. Aliabadi. *Automation Using Matrix Switch for Piezoelectric Actuator/Sensor Based Structural Health Monitoring*. in *Key Engineering Materials*. 2014. Trans Tech Publ.

[5] Wandowski, T., P.H. Malinowski, and W.M. Ostachowicz, *Circular sensing networks for guided waves based structural health monitoring.* Mechanical Systems and Signal Processing, 2016. **66-67**: p. 248-267.

[6] Sharif-Khodaei, Z., M. Thiene, and M.H. Aliabadi, *Structural Health Monitoring Platform for Sensorised Composite Structures*. 2015, DESTECH PUBLICATIONS, INC. p. 1981-1987.

[7] SC-135, R., *DO-160C Environmental Conditions and Test Procedures for Airborne Equipment*. 1989, Washington DC, USA.

[8] 810G, M.-S., *ENVIRONMENTAL ENGINEERING CONSIDERATIONS AND LABORATORY TESTS*. DEPARTMENT OF DEFENSE TEST METHOD STANDARD. 2008.

[9] Su, Z. and L. Ye, *Identification of damage using Lamb waves : from fundamentals to applications*. Vol. 48. 2009: Springer Science & Business Media.

[10] *Pickering Interfaces*. http://www.pickeringtest.com/products/pxi-switching/rf-microwave-switch-modules/rf-matrix [cited 2015 July].

[11] Anton, S.R., D.J. Inman, and G. Park, *Reference-Free Damage Detection Using Instantaneous Baseline Measurements.* AIAA Journal, 2009. **47**(8): p. 1952-1964.

[12] De Marchi, L., A. Perelli, and A. Marzani, *A signal processing approach to exploit chirp excitation in Lamb wave defect detection and localization procedures.* Mechanical Systems and Signal Processing, 2013. **39**(1-2): p. 20-31.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FIBROUS TECTONICS: A RETHINKING OF COMPOSITE PRODUCTION THROUGH INNOVATION AND EXPLORATION OF MOLDING TECHNIQUES AND METHODOLOGIES (ECCOMAS CONGRESS 2016)

**David N. Costanza**[1,2]
[1] Rice University,
6100 Main St Houston, TX 77005,
e-mail: Costanza@rice.edu
[2] Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139
e-email: Davidnc@mit.edu

**Keywords:** Vacuum Assisted Resin Transfer Molding; Bio-Composites; Inflatables.

**Abstract.** *The inherent possibilities of composites present an exciting frontier in architecture that has remained largely untapped. In light of the current computational capacities and new digital tools in manufacturing, composites are just beginning to re-situate themselves in the field of architecture. Efficiency and durability coupled with a load bearing capacity make a strong case for the use of composites as a primary building material.*

*We now possess the computational and digital manufacturing tools that make the development of a composite building viable. On a holistic level, the research has concerned itself with an overarching focus on developing a composite building which minimizes the required costs and labor while simultaneously creating the potential for customized forms. Based on the concepts of mass customization, when the workflow from digital conception to digital production is seamless, a variety of composite structures can be produced at no greater expense. This potential for an efficient "one off" composite architecture empowered by digital manufacturing and computation, is where the research is positioned.*

*At present, the research has been focused on exploring surface composite structures through a reinvention of the 'mold'. This approach has involved using inflated bladders, rather than traditional molds of milled foam or aluminum in order to produce composite structures. In doing so, the benefits of inflatables are all encompassing. Not only do they allow for inexpensive transportation and rapid deployment, but they also lend themselves to the production of large scale structures through the simple use of air and pressure, thus minimizing both material and effort. This lies in stark contrast to traditional composite manufacturing techniques which require molds to be milled out of solid aluminum blocks or high density foam volumes, whereas inflatable molds are easily heat sealed and inflated. When considering issues of scalability, traditional molding techniques demand significantly more labor, material, and with that, overarching costs. Inflatable molds however, require only more air.*

# 1. BACKGROUND

## 1.1. Monsanto House

The Monsanto House of the Future, designed by Marvin Goody & Richard Hamilton, was the first of very few precedents for high performance, composite buildings. Sponsored by the Monsanto petrochemical company, the Monsanto House was built in 1958 from fiberglass composites and challenged the accepted notions of building materials and construction. Due to the one off nature of the house, and the manually intensive and expensive construction, it was not a viable housing concept at the time. However, nearly sixty years later with digital manufacturing, we now possess the technology to make the development of the Monsanto House feasible.

## 1.2. Composite Constraints

Many limitations that have prevented composites from becoming a viable building material stem from the numerous misconceptions that have been harbored through decades of previously unsolved problems. In the past, composites were stereotyped and stigmatized by early concerns originating from high toxicity, consistent off gassing, and poor UV ray resistance. However, in recent years composites have evolved quickly and through consistent use and research, many of the previous downfalls have become obsolete. New and innovative solutions are constantly being developed today to address composites' shortcomings, such as the use of low VOC resins, UV resistant resins as well as fire rated resins.

Because the materials involved in composites are critical to many industries outside of architecture such as the aerospace, automotive, and marine industries, many large international organizations and companies are quite invested in the success and performance of these materials. NASA, Boeing, Airbus, and BMW, just to name a few, all have 'Research and Development' departments aiming to extend the capacities of composite materials and resolve any shortcomings. The innovations and findings developed by these satellite industries, in particular with fiber and matrix systems, can be utilized and perhaps even further developed, in the field of architecture. Of particular interest is the ongoing research on methods and alternatives for more environmentally friendly materials. There has, for instance, been significant development in the utilization of natural fibers such as flax, rather than synthetic fibers like glass fiber. In addition to this, matrices have evolved from using thermosets to thermoplastics, which can be recycled to emerging natural bio resins. In the next five to ten years the possibility to produce natural composites will be readily available as they will not only be easier and safer to build with, but they will also become cheaper as synthetic, petro fibers become more and more expensive.

## 1.3. Composite Advantages

Composite materials present many advantages and possibilities within the field of design that make them quite appealing. However, it is important to note that being a material which has been adopted from other fields, composites require a disparate set of criteria for their application within architecture. Rather than imitating methods of evaluation from other industries, architects and designers must form their own. For example, in the field of high end sailboats where composites have been used for many years, companies spend millions of dollars in order to achieve a lighter product by only a few hundred pounds. However in architecture, there is the chance to negotiate a wide variety of tradeoffs in order to efficiently and economically incorporate these new materials specifically to the needs and interests of the field of architecture. In relation to traditional building materials, composites transcend many of the typical concerns. The most widely known advantage is the strength to weight ratio whereby resulting composite structures are able to achieve supreme thinness and unparalleled lightness. This ultralight advantage simplifies transportation of materials to site and their

subsequent handling. Meaning, an entire building can potentially be brought to site at one time and then erected into place with minimal equipment. Architects have been limited by the capacity of traditional building materials, but composite materials extend these limitations. Perhaps less often considered, is the advantage of consolidation. Composite materials also present the opportunity to collapse many building systems into a single dynamically changing wall section. Much like a traditional wall section, a composite structure is made of many layers. Building systems, such as electrical and plumbing, can be integrated into this multi layered system. Composite parts are molded, allowing for extremely complex geometries to be produced in a single part. This reduces the number of required parts, greatly simplifying and expediting the assembly; assembly can be completed with a handful of individuals, rather than a large construction crew.

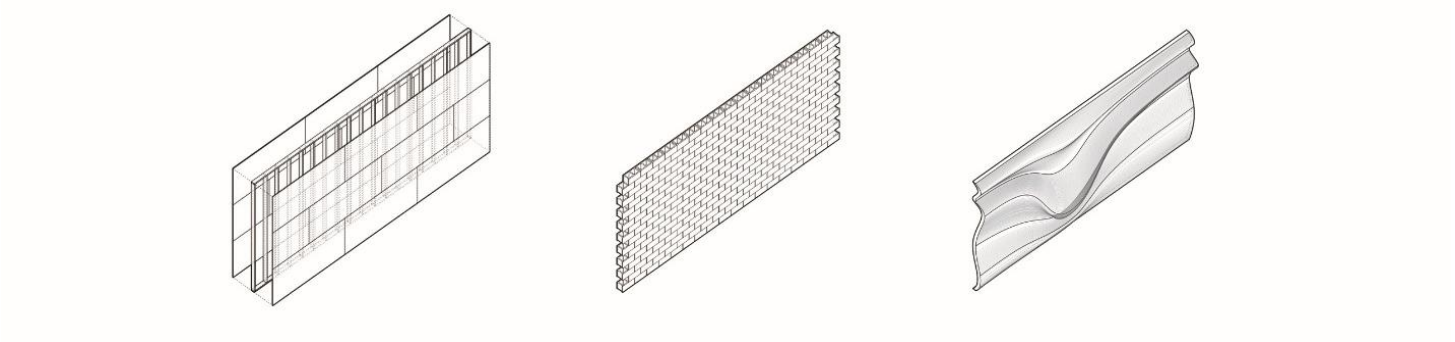


Fig. 1. (a) Wood Stud Wall 840 lbs; (b) CMU Wall 5,800 lbs; (c) Composite Wall 230 lbs.

### 1.4 Composite Manufacturing

Architecture presents an interesting challenge in that nearly every built project is unique. When we discuss automotive, like the BMW i3 or Aerospace's Boeing 787 Dreamliner, an economy of scale exists. Every BMW i3 chassis is exactly the same just as every Boeing 787 Dreamliner is exactly the same. In architecture, a building must resolve the requirements of a specific site, specific program, and specific client, resulting in a custom product. Thus far, molds have been a large inhibitor to the integration and deployment of composites in architecture, particularly at a large scale, due to high costs and slow production times. However, as a result of digital manufacturing, molds are increasingly becoming cheaper and faster to produce. Innovation in mold production will be the deciding force on whether composites can become mainstream within architecture. The possibility of a composite building has only become feasible within the past five years as digital manufacturing and ideas of mass customization have become integrated into everyday practice. For example, automated tape placement has allowed for the first fully composite fuselage, eliminating the more typical, manually intensive processes. The benefit of this transition from manual to automated is clearly demonstrated by automobile assembly lines, which are continuously becoming more autonomous. The digital manufacturing paradigm has allowed the efficiency and accuracy of robotics to accelerate production in many fields, and will similarly allow composites to become active participants within the field of architecture.

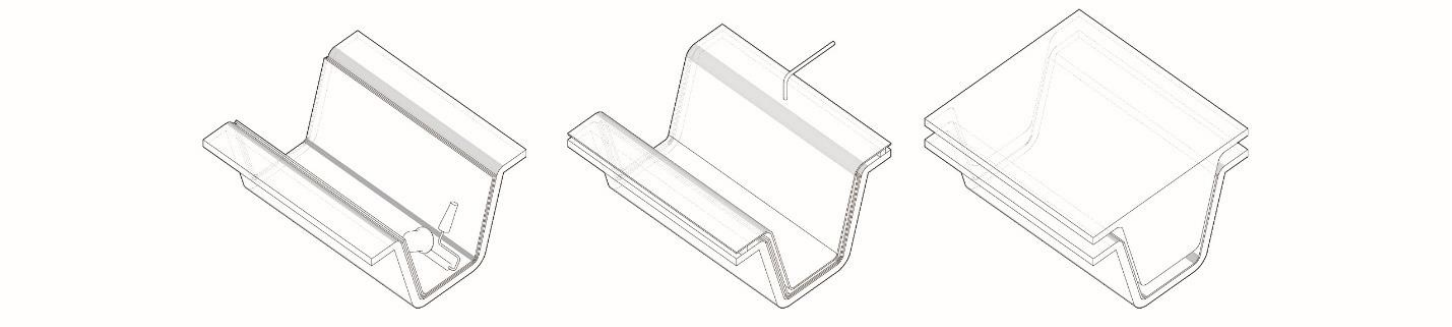


Fig. 2. (a) Hand Wet Layup; (b) Vacuum Bagging Wet Lay Up; (c) Wet Lay Up Compression Molding.

## 2. METHODOLOGY

More and more I have become fascinated by the potential possibilities present in composite materials. As part of a disaster relief effort after the 2010 earthquake in Haiti I designed a modular composite house for rapid deployment in the countryside. For my MIT architecture master's thesis, I designed a composite house for southern Iraq as part of an effort to rebuild after the war which considered petroleum as a building material. Consistently however, I found that the limits of the design process were reached, since the ability to produce such designs, do not in fact exist within architecture today. Composite manufacture thus far, has largely been tailored to high tolerances which entail extreme conditions, such as those found in aviation and aerospace.

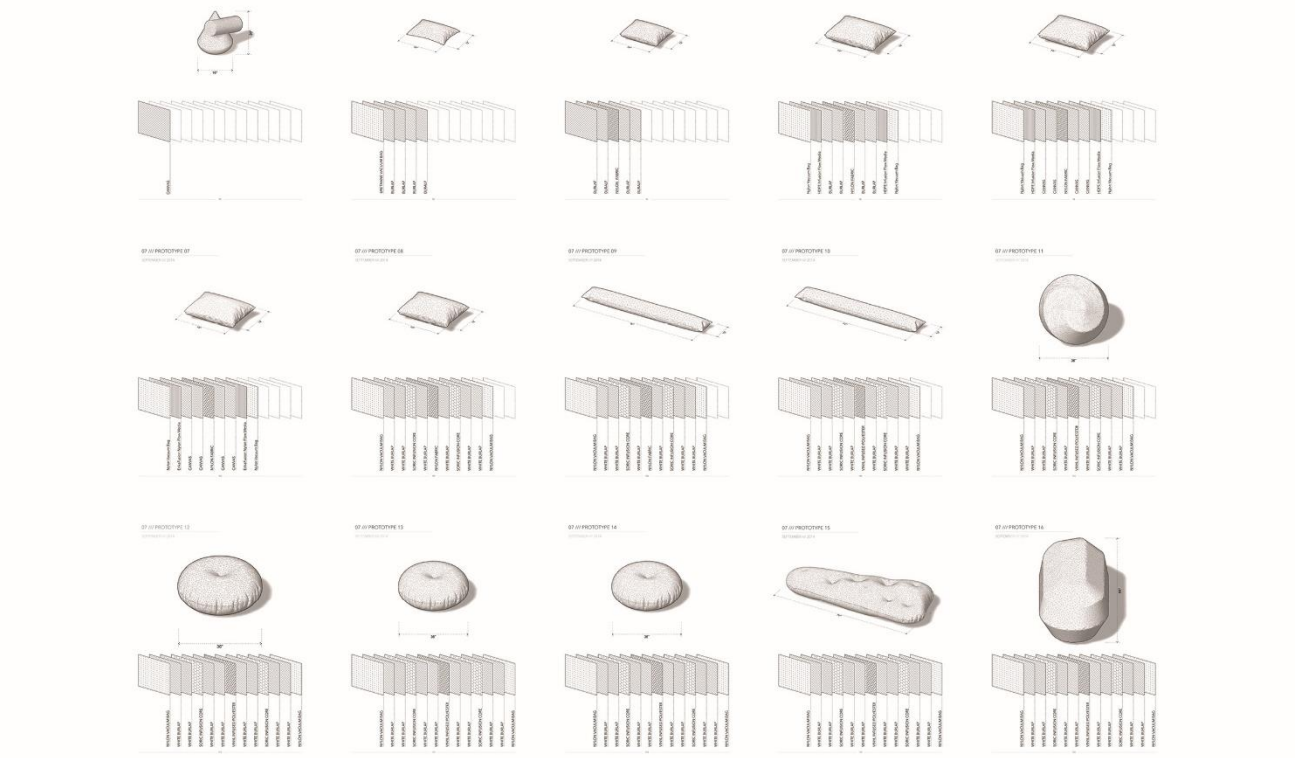


Fig. 3. (a) Prototypes 1 – 16, simulation and build-up.

Embedded in the notion of architectural innovation is an intimate consideration of materiality. In the capacity to understand material systems and manufacturing techniques, the discovery for variable forms of architecture may unfold and present themselves as what was, previously unforeseen possibilities. In this unfolding, therein lies the necessity for full scale investigation and exploration in order to develop comprehensively the relationships and synthesis between materiality and constructability. This direct method of working enables the designer to more accurately and pointedly evaluate a material's strengths and weaknesses, its advantages and disadvantages. In this process and exploratory development, composite material systems are no exception.

Rooted in hands on material testing and making, the current tests have been modest in size partly due to funding and material costs. The opportunity to fully engage in one to one investigations, to confront and interface with composite materials on a full scale inquiry, has been instrumental in the enrichment of the research. Using hands on methodologies, and one to one scale investigations as a starting point has helped to identify and establish an effective composites workflow. The final goal of the research is to produce a full scale, one to one proto-architecture.

## 3. PROCESS

### 3.1 Inflated Molds

Pneumatic structures have a long history in proto-architecture offering the promise of inexpensive transportation and rapid deployment and scalability. By capitalizing on the simple use of air and pressure, with minimum effort and materials, inflated bladders can produce architectural scale spaces. Typically, pneumatic structures require constant pressure but the use of inflated bladders as molds in a composites workflow capitalizes on the advantages of pneumatic structures while coupling it with the rigidity and structural capacity of composite materials.

### 3.2 Workflow

The work flow begins with the design of an equal pressure geometry such as spheres and cylinders with spherical ends and conics with spherical ends. The combination of these geometries can produce a wide variety of inflatable forms. These inflated geometries are then discretized and decomposed into flat unrolled shapes. Once the geometry has been broken down and laid flat onto the cut sheets the details of the part can be added to help form the bladder. This includes extending the edges to create lap joints or flange joints, adding registration lines and cut lines as well as annotating the disparate parts. The flat nested cut sheets are imported into a CAM (computer aided manufacturing) software where the part manufacturing is programed. This includes all annotation with a fabric pen, as well as cut paths with a free wheel drag knife.

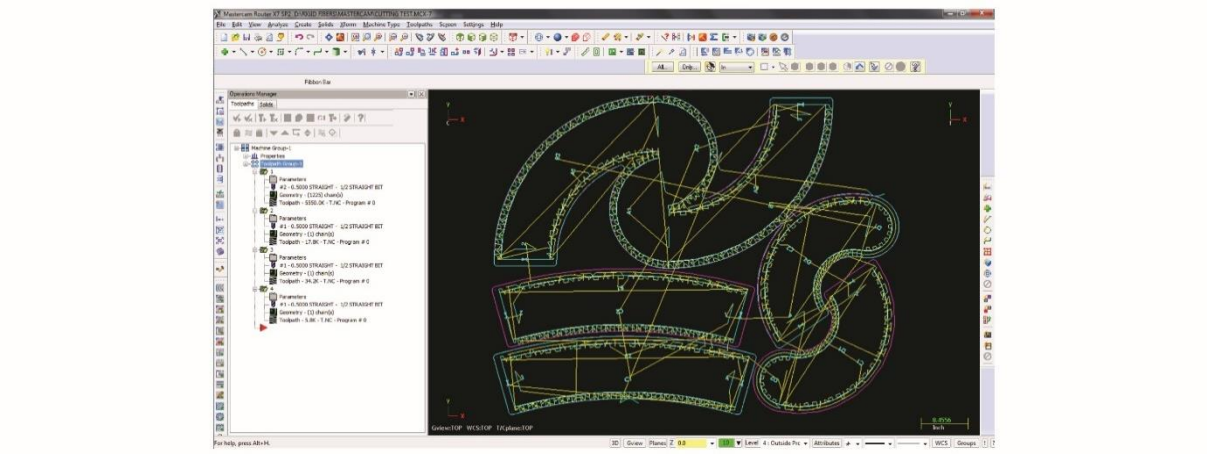


Fig. 4. (a) Nested cut sheets imported into a CAM (computer aided manufacturing) software.

The material is assembled into the anticipated composite before being placed on the vacuum table of the CNC machine. The layers of the assembly include three alternating directions of natural fiber fabric on both sides of a 3mm Soric infusion core. The bladder material, 18oz vinyl infused polyester fabric, is placed on top of the composite assembly and is used to tightly hold all the layers to the cnc vacuum bed. Once the parts are cut, the vinyl infused polyester fabric is stitched and chemically welded to produce the airtight seams of the bladder. The thread provides the strength and the chemical weld provides the air tightness. The two part chemical weld breaks down the polymers in the vinyl, fusing the material and creating a resilient airtight seal. The bladder itself is constructed of both lap and flange joints, which are designed to project into the inflated bladder to allow for a good infusion surface. During the stitching and welding process the through bag valve that is used to inflate the bladder is added.

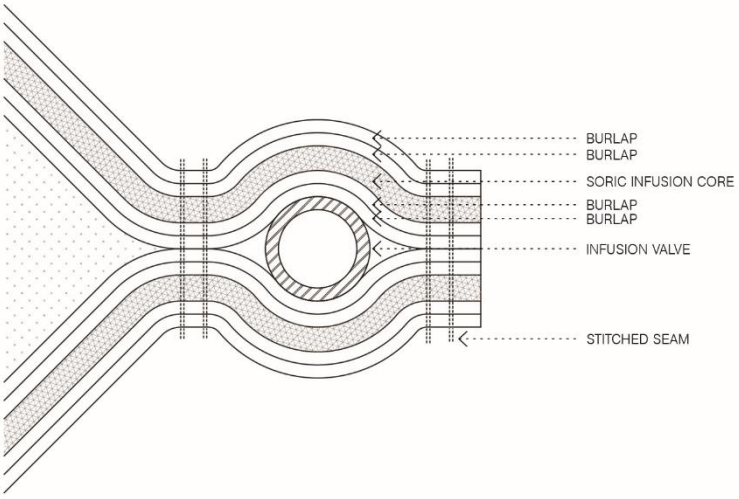



Fig. 5. Welded seam detail.

The preform (composite assembly) is also sewn together using a heavy weight polyester thread. VARTM (vacuum assisted resin transfer molding) requires inlet (resin) and outlet (vacuum) valves which are calibrated according to the geometry of the part. Inlets and outlets can either be point valves or distributed valves depending on the geometry and the infusion layout. Linear distribution uses a ⅜” polyethylene spiral tubing sewn into the preform. The spiral tubing allows the resin (intake) or vacuum (outlet) to rapidly travel the length of the

tube before penetrating the infusion core. This is an effective way to create an even distribution infusion. Point distribution uses a PLA 3d printed valve designed to allow rapid and even distribution of both resin and vacuum pressure. The 3d printed valve is sewn into the preform at the same layer in the composite as the infusion core. The integration of the inlets and outlets significantly reduces the setup time.

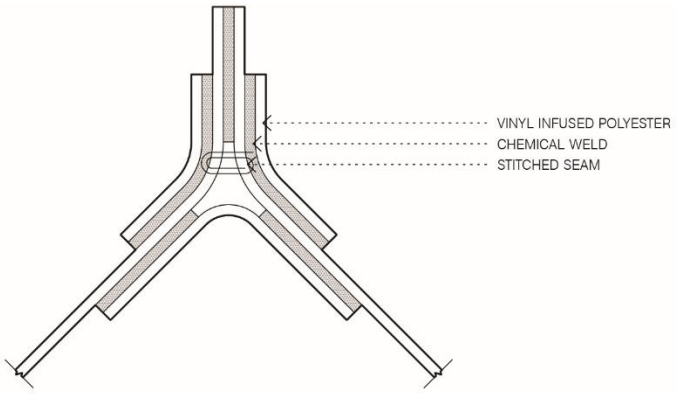



Fig. 6. (a) Infusion seam detail.

The integration and robustness of the preform reinforces the research ambition for a compactable and shippable composite. By integrating all the complexity and intelligence into the perform the entire assembly can be vacuum compressed, packaged and shipped.

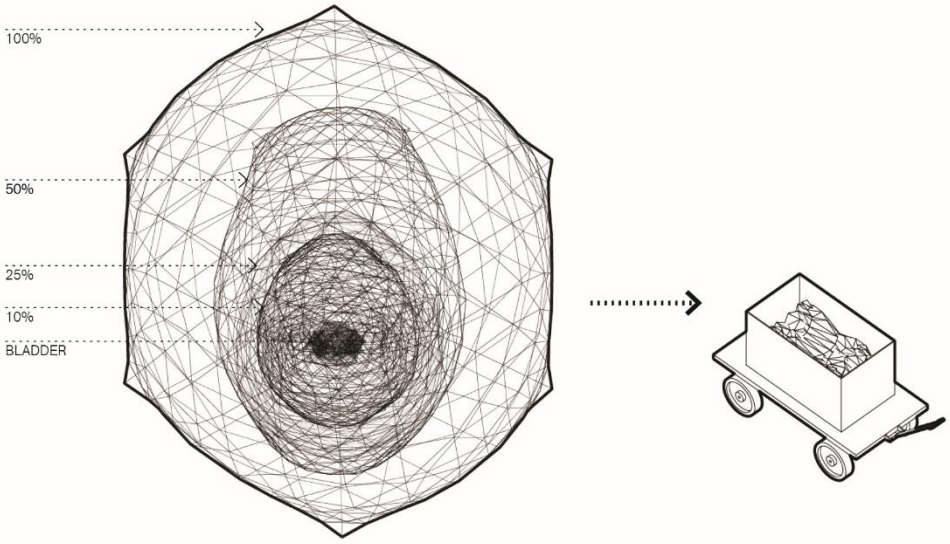



Fig. 7. (a) Inflation Process; (b) Transportation of deflated bladder.

The most conventional aspect of the process is bagging the assembly. Bagging the composite puts the entire assembly under compaction in order to produce a high performance composite structure. Because a high performance part must be bagged for compaction the required elements for vacuum assisted resin transfer molding already exist. Using 300f nylon bagging

film and sealant tape an oversized bag is constructed, non specific to the geometry of the part. Once the entire assembly is constructed the bladder is inflated to 5 psi and a vacuum is pulled on the sealed assembly at 25 inHg. The assembly is checked for leaks to guarantee successful compaction and infusion. Once the assembly is surveyed, the inlet line is submerged in a low viscosity infusion epoxy resin. When the hose clamp is removed the vacuum pulls the resin through the linear / point valve and into the infusion core which wets out the fibers as it is pulled across the part. The infusion process is regulated by the viscosity and pot life of the resin and can be carefully calibrated as needed. Once the part is fully infused the vacuum continues to compress the part extracting excess resin into a resin trap in order to remove all voids and create a uniform, lightweight composite.

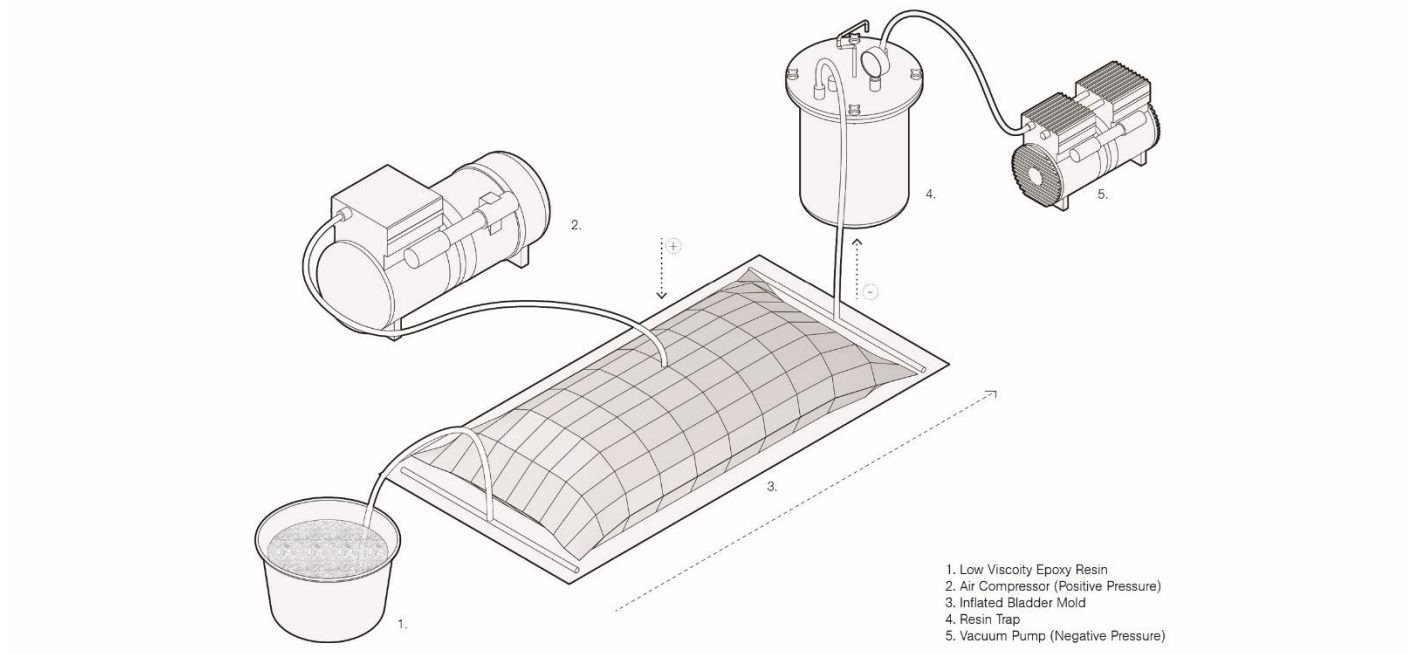



Fig. 8. (a) Intlated and infused workflow using VARTM.

## 4. CONCLUSION

The workflow presented here has evolved over the last year by working directly with the materials. Initial challenges existed in producing the inflated mold in which a number of different materials were tested. After a number of mockups, the addition of sewing the heat sealed edge as a reinforcement and prevention of catastrophic failure was included. Another primary material that was tested was the infusion mesh. Most infusion meshes are applied to the exterior of the part and allow resin flow between the part and the nylon vacuum bag. However, the welcomed discovery of an infusion mesh core that is situated between the fabric plies, has significantly reduced the waste and helped simplify the process while adding notable structural properties.

The process is just beginning to prove to be reliable. Future works is based on the desire to scale up, having thus far tested primarily in the object furniture scale. Following the success of the initial prototypes and the established understanding of the material systems, rules, and constraints, I would like to continue by designing a small proto-architecture, with the ultimate goal of designing a full scale house that evolves from this material and construction system.

## REFERENCES

[1] Campbell, F. C. Manufacturing processes for advanced composites. New York: Elsevier, 2004. Print.
[2] Strong, A B. Fundamentals of composites manufacturing materials, methods and applications. Dearborn, Mich: Society of Manufacturing Engineers, 2008. Print.
[3] Bank, Lawrence C. Composites for construction: structural design with FRP materials. Hoboken, N.J: John Wiley & Sons, 2006. Print.
[4] Topham, Sean. Blowup: inflatable art, architecture, and design. Munich New York: Prestel, 2002. Print.
[5] Herzog, Thomas, Gernot Minke, and Hans Eggers. Pneumatic structures: a handbook of inflatable architecture. New York: Oxford University Press, 1976. Print.
[6] Clausen, Barbara, and Carin Kuoni. Thin skin: the fickle nature of bubbles, spheres and inflatable structures. New York: Independent Curators International, 2002. Print.
[7] Jenkins, C. H. Gossamer spacecraft: membrane and inflatable structures technology for space applications. Reston, Va: American Institute of Aeronautics and Astronautics, 2001. Print.
[8] Krauel, Jacobo. Inflatable art, architecture & design. S.l: Watson-Guptill, 2013. Print.
[9] Oñate, E, and Bern Kröplin. Textile composites and inflatable structures II. Dordrecht: Springer, 2008. Print.
[10] Ballast, David K. Inflatable structures and air pressure supported roofs. Monticello, Ill., USA: Vance Bibliographies, 1988. Print.
[11] Dessauce, Marc. The inflatable moment: pneumatics and protest in ‘68. New York: Princeton Architectural Press, 1999. Print.
[12] Mallick, P. K. Fiber-reinforced composites materials, manufacturing, and design. New York, N.Y: M. Dekker, 1993. Print.

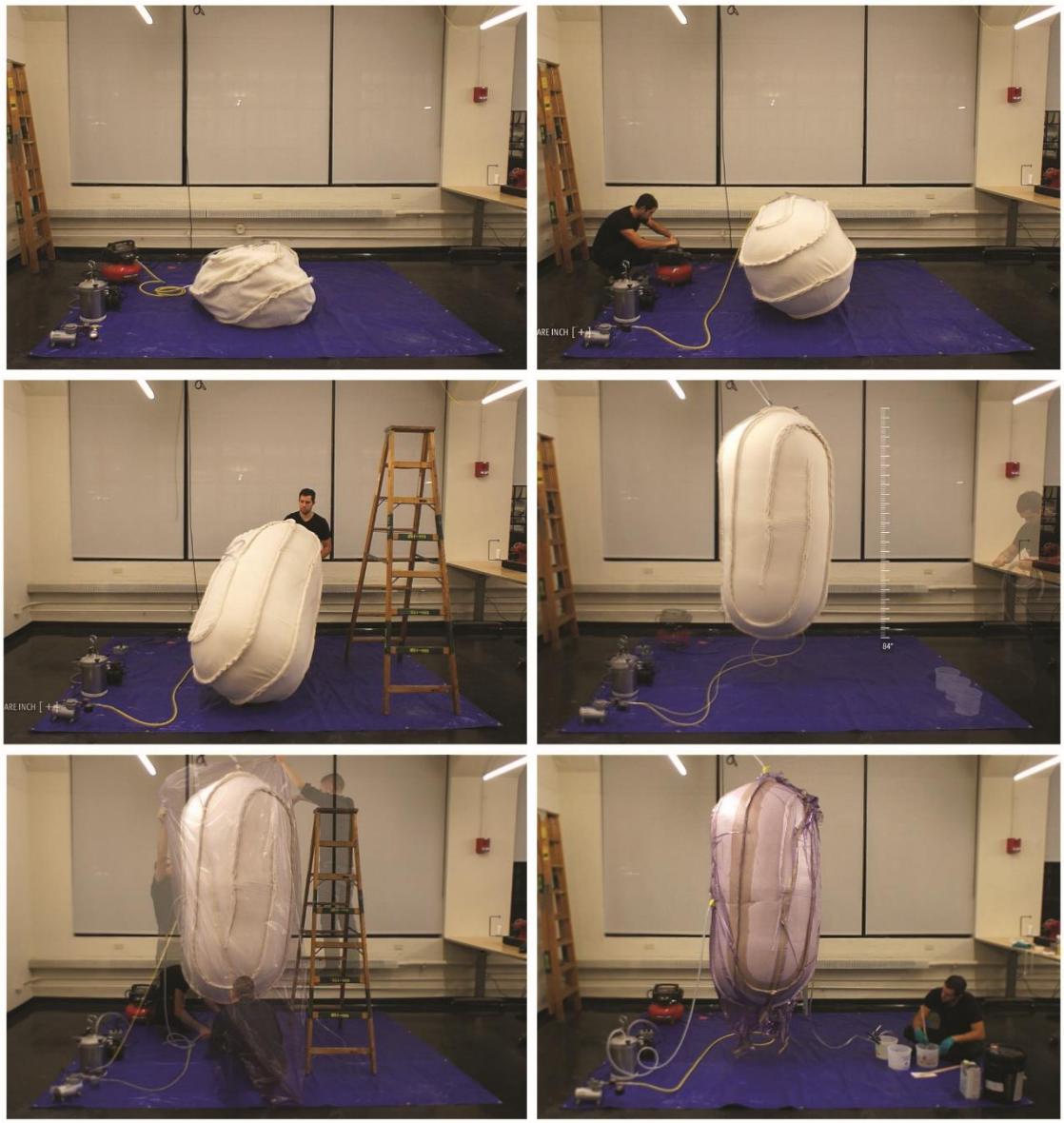


Fig. 9. (a) Workflow for prototype number 16.

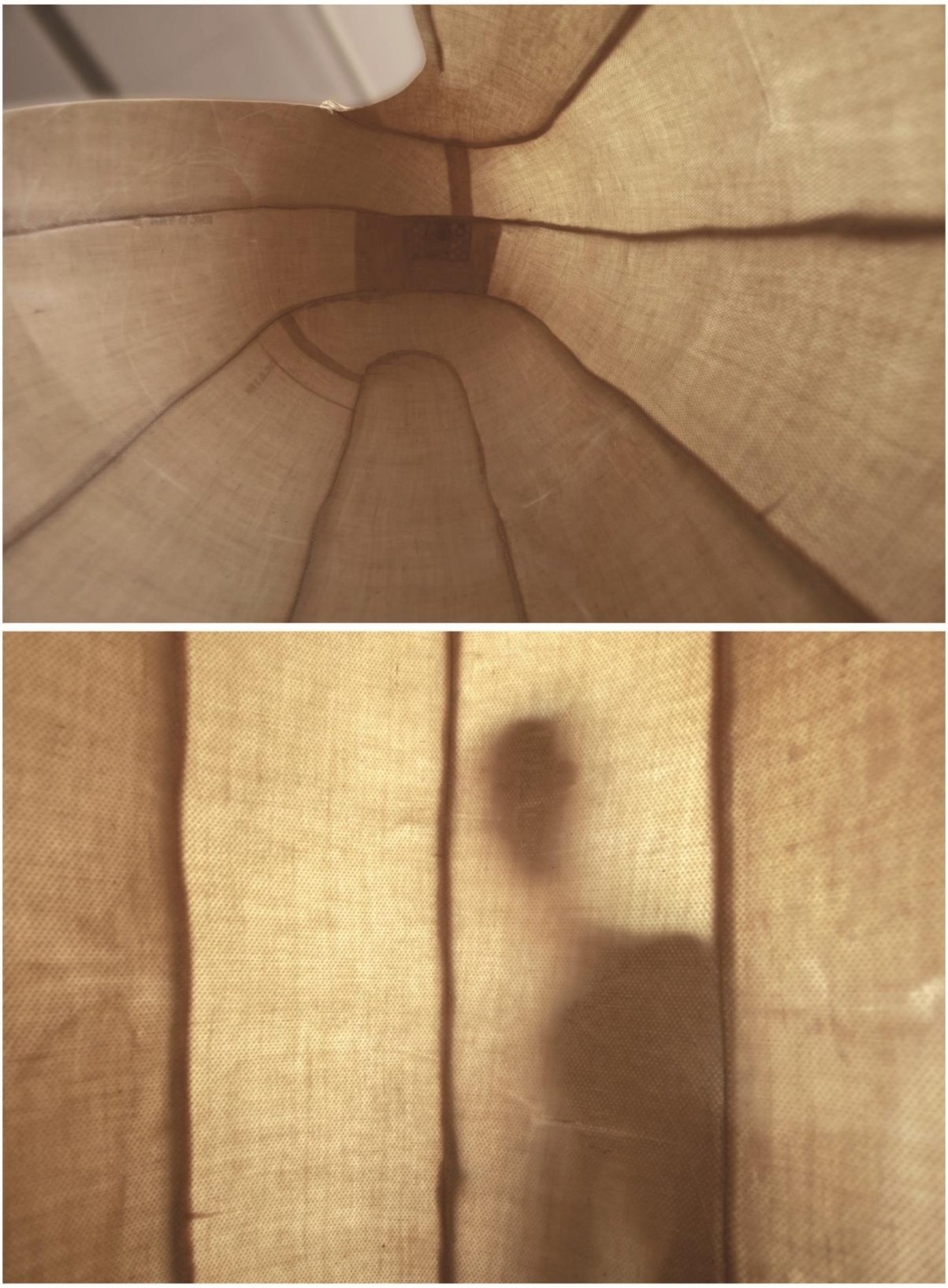


Fig. 10. (a) Final images after infusion.



Fig. 11. (a) Final images after infusion.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFECT OF NANOSIZED ASPERITIES AT THE SURFACE OF A NANOHOLE

**Mikhail A. Grekov**[1], **Aleksandra B. Vakaeva**[1]

[1]St. Petersburg State University
7/9, Universitetskaya nab., 199034, St. Petersburg, Russia
e-mail: m.grekov@spbu.ru, a.vakaeva@spbu.ru

**Keywords:** Nanosized Asperities, Nanohole, Surface Stress, Perturbation Method, Integral Equation, Stress Concentration.

**Abstract.** *The two-dimensional problem on a curvilinear nanohole in an infinite elastic body under arbitrary remote loading is solved. The shape of the hole is assumed to be weakly deviated from the circular one and the complementary surface stresses are acting at the boundary. The boundary conditions are formulated according to the generalized Laplace – Young law. The study is based on Gurtin – Murdoch surface elasticity model. Using Goursat – Kolosov complex potentials and the boundary perturbation technique, the solution of the problem is reduced to a singular integro-differential equation for any-order approximation. The algorithm of solving this integral equation is constructed in the form of a power series. Solutions of the integral equation and corresponding complex potentials are obtained for zero-order and first-order approximations. The size effect in the form of the dependence of the stress distribution at the surface on the size of the hole is demonstrated.*

## 1 INTRODUCTION

Material properties are determined by its chemical composition and structure. The study of surface phenomena is of great interest as it gives an opportunity to obtain an information about the physical and mechanical behavior of a whole material as, for example, in the case of mechanochemical corrosion [1, 2]. In particular, it concerns nanosized materials and structures widely used for developing functional devices in the semiconductor industry, nanoelectromechanical systems, gas and chemical sensors or even biomedical equipment [3].

During the past half-century, much effort to study the mechanical behavior of nanomaterials (beams, plates, wires, films, etc.) and composites containing nanosized inhomogeneities (inclusions, voids and holes) has been made in materials science, solid state science and nanomechanics (see reviews [3, 4, 5]). Investigation of material properties and elastic fields of nanostructures and nanocomposites is important for further development of nanotechnology involving analysis, design and fabrication of devises and structures at the nanoscale. At least one of the overall dimensions of nanomaterials and nano-inhomogeneities is in the nanometer range. This implies that the ratio of the volume occupied by the atoms at and near the corresponding surface/interface to the volume of the bulk or nano-inhomogeneity becomes significant. As a consequence, the energy of the atoms adjoining the surface/interface, called the surface free energy [6], and the surface stress relating the variation of the surface free energy to the variation of the surface strain [7] highly influence on material properties and elastic fields of nanomaterials and nanocomposites. The size dependency of an elastic state at the nanoscale is one of the corroborations of this influence [8, 9, 10].

The basic concepts of surface/interface free energy and surface/interface stress in solids were first formulated by Gibbs [6] and developed later by a lot of researchers. Gurtin and Murdoch [11, 12] and Gurtin et al. [13] elaborated the mathematical framework incorporating surface stress into continuum mechanics. Miller and Shenoy [8] performed atomistic simulations with the embedded atom method of nanosized plates and beams subjected to uniaxial loading and bending and found that their results were in excellent agreements with those obtained by means of the Gurtin – Murdoch continuum model.

Within the Gurtin – Murdoch model of the surface/interface elasticity, a number of classical problems related to elastic phenomena at the nanoscale have been studied and some size effects have been found for relevant nanoscale materials. For example, thin film, inclusion, inhomogeneity and surface defect formation problems were resolved in the works [14]–[26].

The influence of surface stress on the elastic field was reported in [21] for the case when the external forces applied to a flat surface have changed within a nanosized region (size effect). Incorporating surface stresses, size effect is appeared as the dependence of an elastic field near an elliptical nanohole [22] and nanoinclusion [16, 17] on their sizes, and the local instability of a plate with a circular nanohole on the hole size under uniaxial tension [23, 24]. Recently, analysis of elastic fields at the nanosized surface defects arising on a film coating due to diffusion has been carried out in [25, 26] using the surface elasticity theory [11, 12] coupled with the universal boundary perturbation technique.

In the previous paper [27], using the universal boundary perturbation method, we have solved the 2-D elasticity problem on a nearly circular hole in an infinite plane at the macrolevel. The results allowed us to evaluate the effect of deviation of the hole boundary from the circular one on the stress concentration and the stress-strain state near the curvilinear macrohole. In connection with the intensive development of nanotechnology and the use of nanomaterials and nanostructures in a variety of optical, electronic and other devices, it is important to study the

problem of changing the stress state near nanohole due to incorporating surface stress. So, the purpose of the present work is to study the effect of the surface stresses on an elastic field of a body containing a nearly circular nanohole. In order to solve the corresponding problem, we use Gurtin – Murdoch surface elasticity theory [11, 12] and the boundary perturbation technique developed in [25]–[29]. As a result, we come to the singular integro-differential equation in any order approximation of the perturbation method. The analytical solution of this equation and numerical results are given for the first-order approximation.

## 2 PROBLEM FORMULATION

We consider an infinite elastic plane of complex variable $z = x_1 + ix_2$ ($i$ is the imaginary unit) with a nanohole the shape of which is weakly deviated from the circle of radius $a$ with the center in the origin of Cartesian coordinates $x_1$, $x_2$. The plane is under arbitrary remote loading and extra surface stresses at the boundary. The plane strain conditions are supposed to be satisfied. The boundary of the hole $\Gamma$ is determined by the equation:

$$z \equiv \zeta = \rho e^{i\theta} = a\left(1 + \varepsilon f(s)\right) s, \tag{1}$$

where $s = e^{i\theta}$, $f(s)$ is the continuous function and $|f| \leq 1$, $\varepsilon$ is the small parameter which is equal to the maximum deviation of the hole boundary from the circular one of radius $a$, $\varepsilon > 0$, $\varepsilon \ll 1$.

The boundaries of the hole determined by the equation (1) are shown in Fig. 1 for $f(s) = \cos 2\theta$ and $a = 1$, and different values of the parameter $\varepsilon$. These different forms of the hole are used in our work to get numerical results.

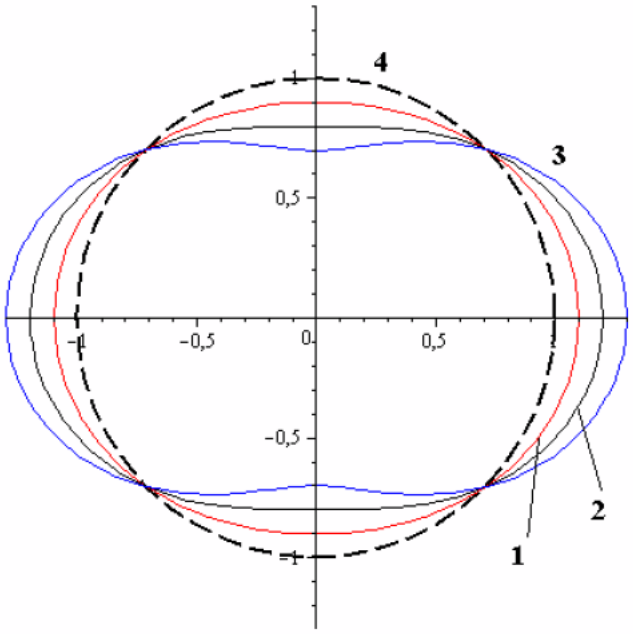



Figure 1: Boundaries of the hole described by the function $f(s) = \cos 2\theta$ in equation (1) for $a = 1$, $\varepsilon = 0,1; 0,2; 0,3$ (curves 1, 2, 3) and $\varepsilon = 0$ ( curve 4).

In the case of the 2-D problem, the condition at the boundary is described by the generalized

Laplace – Young law [14, 18] and has the form [22, 25]:

$$\sigma_n(\zeta) = \sigma_{nn} + i\sigma_{nt} = \frac{\sigma_{tt}^s}{R} - i\frac{1}{h}\frac{d\sigma_{tt}^s}{d\theta} + p(\zeta) \equiv t^s(\zeta) + p(\zeta), \quad \zeta \in \Gamma, \tag{2}$$

where $\sigma_{nn}, \sigma_{nt}$ are normal and tangential stresses (in equation (2), the unit vectors **n** and **t** are the basis vectors of the local Cartesian coordinates $n, t$); $\sigma_{tt}^s = \tau$ is the surface stress, $p$ is the external force. Metric coefficient $h$ [30] and the curvature radius of the boundary $R$ are evaluated in terms of the radius $a$ by the formulas:

$$L = \frac{1}{h} = \frac{1}{\sqrt{{x_1'}^2 + {x_2'}^2}}, \quad K = \frac{1}{R} = \frac{x_1' x_2'' - x_2' x_1''}{h^3}$$

and can be written as

$$L = \frac{1}{\sqrt{(\delta')^2 + (1+\delta)^2}}, \quad K = \frac{2(\delta')^2 - (1+\delta)\delta'' + (1+\delta)^2}{h^3}, \tag{3}$$

where $\delta = \varepsilon f(s)$.

At infinity, the stresses $\sigma_{ij}$ $(i, j = 1, 2)$ and rotation angle $\omega$ are specified as

$$\lim_{z\to\infty} \sigma_{ij} = \sigma_{ij}^\infty = s_{ij}, \quad \lim_{z\to\infty} \omega = 0. \tag{4}$$

In the case of the plane strain, the constitutive equations of the surface [11, 12] and volume [31] linear elasticities are respectively reduced to the following [21, 22]

$$\sigma_{tt}^s = \gamma_0 + (\lambda_s + 2\mu_s)\varepsilon_{tt}^s, \quad \sigma_{33}^s = \gamma_0 + \lambda_s \varepsilon_{tt}^s \tag{5}$$

and

$$\begin{gathered} \sigma_{nn} = (\lambda + 2\mu)\varepsilon_{nn} + \lambda\varepsilon_{tt}, \quad \sigma_{tt} = (\lambda + 2\mu)\varepsilon_{tt} + \lambda\varepsilon_{nn}, \\ \sigma_{nt} = 2\mu\varepsilon_{nt}, \quad \sigma_{33} = \lambda(\varepsilon_{tt} + \varepsilon_{nn}). \end{gathered} \tag{6}$$

In equations (5), (6), $\gamma_0$ is the residual surface stress; $\varepsilon_{nn}$, $\varepsilon_{nt}$, $\varepsilon_{tt}$ are the components of the volume strain tensor; $\varepsilon_{tt}^s$ is the surface strain; $\lambda_s, \mu_s$ are the surface elastic constants similar to Lame constants $\lambda, \mu$.

To find the surface stress $\tau$ and solve the boundary value problem, we use the relations (2)–(6) and the inseparability condition of the surface and bulk, expressed in terms of hoop strains:

$$\varepsilon_{tt}^s(\zeta) = \varepsilon_{tt}(\zeta), \quad \zeta \in \Gamma. \tag{7}$$

## 3 BOUNDARY EQUATION FOR COMPLEX POTENTIALS

Goursat – Kolosov complex potentials and Muskhelishvili's method [31] are used in order to solve the problem. According to [27, 31], the vector of stresses $\sigma_n = \sigma_{nn} + i\sigma_{nt}$ at the area with the normal **n** can be expressed via two functions $\Phi$ and $\Psi$ holomorphic outside of their boundaries as

$$\sigma_n(z) = \Phi(z) + \overline{\Phi(z)} + \left[z\Phi'(z) + \overline{\Psi(z)}\right]\frac{d\bar{z}}{dz}, \quad z \in \Omega,$$

where $dz = |dz|e^{i\alpha}$, $d\bar{z} = \overline{dz}$, $\alpha$ is the angle between the axes $\mathbf{t}$ and $x_1$.

Following [31], we introduce the new function $\Upsilon(z)$ holomorphic in the finite region $D = \{z : \bar{z}^{-1} \in \Omega\}$ with the boundary $\widetilde{\Gamma}$ which is symmetrical to the boundary $\Gamma$ with respect to the unit circle:

$$\Upsilon(z) = -\overline{\Phi(\bar{z}^{-1})} + z^{-1}\overline{\Phi'(\bar{z}^{-1})} + z^{-2}\overline{\Psi(\bar{z}^{-1})}, \quad \bar{z}^{-1} \in \Omega.$$

Passing to the limit for $z \to \zeta \in \Gamma, z \in \Omega$ [27] and taking into account (2), we get the following boundary equation for complex potentials $\Phi$ and $\Upsilon$:

$$\sigma_n(\zeta) = \Phi(\zeta) + \overline{\Phi(\zeta)} + \frac{\rho' - i\rho}{\rho' + i\rho}\left[\frac{1}{\overline{\zeta^2}}\left(\overline{\Phi(\zeta)} + \Upsilon\left(\frac{1}{\bar{\zeta}}\right)\right) + \left(\zeta - \frac{1}{\bar{\zeta}}\right)\overline{\Phi'(\zeta)}\right]\bar{s}^2, \tag{8}$$

where $\Phi(\zeta) = \lim\limits_{z\to\zeta}\Phi(z)$ when $z \in \Omega$ and $\Upsilon(\zeta) = \lim\limits_{z\to\zeta}\Upsilon(z)$ when $z \in D$.

## 4 BOUNDARY PERTURBATION PROCEDURE

Following perturbation method [25]–[29], we represent unknown functions $\Phi$, $\Upsilon$ and the surface stress $\tau$ as power series in the small parameter $\varepsilon$:

$$\Phi(z) = \sum_{n=0}^{\infty}\frac{\varepsilon^n}{n!}\Phi_n(z), \quad \Upsilon(z) = \sum_{n=0}^{\infty}\frac{\varepsilon^n}{n!}\Upsilon_n(z), \quad \tau(\zeta) = \sum_{n=0}^{\infty}\frac{\varepsilon^n}{n!}\tau_n(\zeta). \tag{9}$$

We also derive the expressions for all functions in (8). Taking into account expansions

$$K = \sum_{n=0}^{\infty}\frac{\varepsilon^n}{n!}K_n(s), \quad L = \sum_{n=0}^{\infty}\frac{\varepsilon^n}{n!}L_n(s), \tag{10}$$

we obtain for the zero-order and first-order approximations:

$$K_0 = 1, \quad L_0 = 1, \quad K_1 = -f''(s) - f(s), \quad L_1 = -f(s). \tag{11}$$

Substituting series (9), (10) into equation (8), we equate the sum of coefficients of the same power $\varepsilon^n$ $(n = 0, 1, \ldots)$ to zero. Then we arrive at the Rimann – Gilbert boundary value problem on the jump of the holomorphic function $\Xi_n(z)$ for each-order approximation:

$$\Xi_n^+(s) - \Xi_n^-(s) = -\left(\tau_n(s) + s\tau_n^{'}(s)\right) + F_n(s). \tag{12}$$

Here,

$$\Xi_n^{\pm} = \lim_{|z|\to 1\mp 0}\Xi_n(z), \qquad \Xi_n(z) = \begin{cases} \Phi_n(z), & |z| > 1, \\ \Upsilon_n(z), & |z| < 1. \end{cases}$$

According to Muskhelishvili [31], solution of the problem (12) can be written in terms of Cauchy type integrals

$$\Xi_n(z) = -I_n(z) + J_n(z) + S_n(z), \tag{13}$$

where

$$I_n(z) = \frac{1}{2\pi i}\int\limits_{|\zeta|=1}\frac{\tau_n(\zeta) + \zeta\tau_n^{'}(\zeta)}{\zeta - z}d\zeta, \quad J_n(z) = \frac{1}{2\pi i}\int\limits_{|\zeta|=1}\frac{F_n(\zeta)}{\zeta - z}d\zeta, \tag{14}$$

and $S_0 = C + C_2 z^{-2}$, $S_n = 0$ $(n = 1, 2, \ldots)$, $4C = s_{11} + s_{22}$, $2C_2 = s_{22} - s_{11} - 2is_{12}$.

The functions $F_n$ $(n = 1, 2, \ldots)$ in (13) are the known functions depending only on all previous approximations. So, in the zero-order and first-order approximations, we obtain:

$$F_0(s) = -p(s),$$
$$F_1(s) = -p'(s) - \bar{s}f(\bar{s})\frac{d\Upsilon_0\left(\bar{s}^{-1}\right)}{d\bar{s}} + sf(s)\Phi_0'(s) + 2(f(s) - if_\theta'(s))\left(\overline{\Phi_0(s)} + \Upsilon_0\left(\bar{s}^{-1}\right)\right) -$$
$$-2\bar{s}f(s)\overline{\Phi_0'(s)} - sf(s)\frac{d\tau_0(s)}{d\theta} + (f_\theta''(s) + f(s))\tau_0(s) + i\left(sf(s)\frac{d^2\tau_0(s)}{d\theta^2} - f(s)\frac{d\tau_0(s)}{d\theta}\right).$$

It is easy to see that the integral $J_n$ in the equation (14) is a known function.

## 5 INTEGRAL EQUATION OF N-ORDER APPROXIMATION. FIRST-ORDER APPROXIMATION

As a result, we use the constitutive equations (5), (6) by Gurtin – Murdoch surface elasticity theory and inseparability condition (7) and obtain the following singular integro-differential equation in the unknown functions $\tau_n$:

$$\tau_n(s) + \frac{M(\varkappa+1)}{2a - M(\varkappa-1)}\mathrm{Re}\left(\frac{1}{2\pi i}\int\limits_{|\eta|=1}\frac{\tau_n(\eta) + \eta\tau_n'(\eta)}{\eta - s}d\eta\right) = G_n(s), \tag{15}$$

where $M = (\lambda_s + 2\mu_s)/2\mu$; $\varkappa = (\lambda + 3\mu)/(\lambda + \mu)$.

The function $G_0$ depends from the load $p(s)$ ($G_0 = 0$ when $p(s) = 0$), and functions $G_n$ ($n > 0$) are expressed through all the previous approximations.

For any-order approximation, the solution of the integral equation (15) can be found in the form of a power series $\tau_n(\zeta) = \sum\limits_{k=-\infty}^{\infty} a_{nk}\zeta^k$ for any external forces $p(s)$ and function $f(s)$. In the case of a hole free from external forces ($p(s) = 0$), the complex potentials in the zero-order approximation, which correspond to the solution of the appropriate boundary value problem for the circular hole, are determined as

$$\Phi_0(z) = C + C_2 z^{-2} - \overline{a_{02}}z^{-2}, \quad \Upsilon_0(z) = C + C_2 z^{-2} - a_{00} - 3a_{02}z^2, \tag{16}$$

where $a_{00} = 4H_1 C - \sigma_0$, $a_{02} = H_2\overline{C_2}$ and

$$\sigma_0 = -\frac{\gamma_0}{a + M}, \quad H_1 = \frac{M(\varkappa+1)}{4(a+M)}, \quad H_2 = \frac{M(\varkappa+1)}{2a + M(\varkappa+3)}. \tag{17}$$

So, surface stress is written in the form:

$$\tau_0 = a_{00} + a_{02}\zeta^2 + \overline{a_{02}}\zeta^{-2} \tag{18}$$

and hoop stress for the circular hole ($|z| = a$) is defined by the equality:

$$\sigma_{tt} = \sigma_0 + (1 - H_1)(s_{11} + s_{22}) + (2 - 3H_2)(s_{22} - s_{11})\cos 2\theta - 2(2 - 3H_2)s_{12}\sin 2\theta. \tag{19}$$

This solution is identical to the solution obtained in [22] by the another way.

Consider a curvilinear hole the boundary of which is described by the relation (1) for the function $f(s) = \cos 2\theta$. The integral equation in the first-order approximation is

$$\tau_1(s) + \frac{M(\varkappa+1)}{2a - M(\varkappa-1)} \mathrm{Re}\, I_1(s) = G_1(s), \tag{20}$$

where

$$G_1(s) = -\frac{M(\varkappa+1)}{2a - M(\varkappa-1)} \mathrm{Re}\, J_1(s).$$

To find the surface stress $\tau_1$ and therefore complex potentials $\Phi_1$ and $\Upsilon_1$, represent $\tau_1$ in the form of the power series. Following the algorithm as shown for any-order approximation, one can solve the boundary value problem in the first-order approximation.

To find the function $G_1$, it is necessary to substitute equations (1), (11) and (16)-(18) in the equation (14). After that, one can obtain the solution of the equation (20) in the form:

$$\tau_1(\zeta) = \sum_{k=-5}^{5} a_{1k}\zeta^k, \quad a_{1-k} = \overline{a_{1k}}, \quad k = \overline{0,5} \tag{21}$$

for the boundary of the hole given by the function $f(s) = \cos 2\theta$.

Coefficients $a_{1k}$ in the equation (21) depend on coefficients $a_{00}$ and $a_{02}$ of the first-order approximation and aren't given here because of bulky expressions. We substitute (21) into (12), (13) and obtain the complex potentials as

$$\Phi_1(z) = -I_1^-(z) + J_1^-(z), \quad \Upsilon_1(z) = -I_1^+(z) + J_1^+(z), \tag{22}$$

where $I_1^{\pm}(z) = I_1(z)$ and $J_1^{\pm}(z) = J_1(z)$ when $\pm|z| < \pm 1$.

Using expressions (9), (10) and complex potentials (16), (22), we obtain $\sigma_{tt}$ for the hole determined by the function $f(s) = (s^2 + s^{-2})/2 = \cos 2\theta$:

$$\begin{aligned}\sigma_{tt} = \varepsilon \left( \Phi_1(s) + 2\overline{\Phi_1(s)} + \Upsilon_1\left(\bar{s}^{-1}\right) + sf(s)\Phi_0'(s) + \bar{s}f(\bar{s}) \left( 2\overline{\Phi_0'(s)} + \frac{d\Upsilon_0\left(\bar{s}^{-1}\right)}{d\bar{s}} \right) \right) + \\ +2\varepsilon \left( \left(if_\theta'(s) - f(s)\right) \left( \overline{\Phi_0(s)} + \Upsilon_0\left(\bar{s}^{-1}\right) \right) + \bar{s}f(s)\overline{\Phi_0'(s)} \right) + \\ +\sigma_0 + (1 - H_1)(s_{11} + s_{22}) + (2 - 3H_2)(s_{22} - s_{11})\cos 2\theta - 2(2 - 3H_2)s_{12}\sin 2\theta.\end{aligned} \tag{23}$$

## 6 NUMERICAL RESULTS

According to the equation (23), the hoop stress $\sigma_{tt}$ depends on the radius $a$ of the basic circular hole (size effect). The dependence of maximum values of $\sigma_{tt}$ (when $\theta = 0$) in the case of the uniaxial tension $s_{22}$ along the axis $x_2$, and $p = \gamma_0 = 0$, is shown in Fig. 2. The plots are constructed for aluminium [22]. Surface elastic modulus [8, 15] lead to $M = 0,1$ nm (continuous red lines in Fig. 2) or $M = -0,152$ nm (dashed green lines in Fig. 2). Dotted lines correspond to the classical solution ($M = 0$).

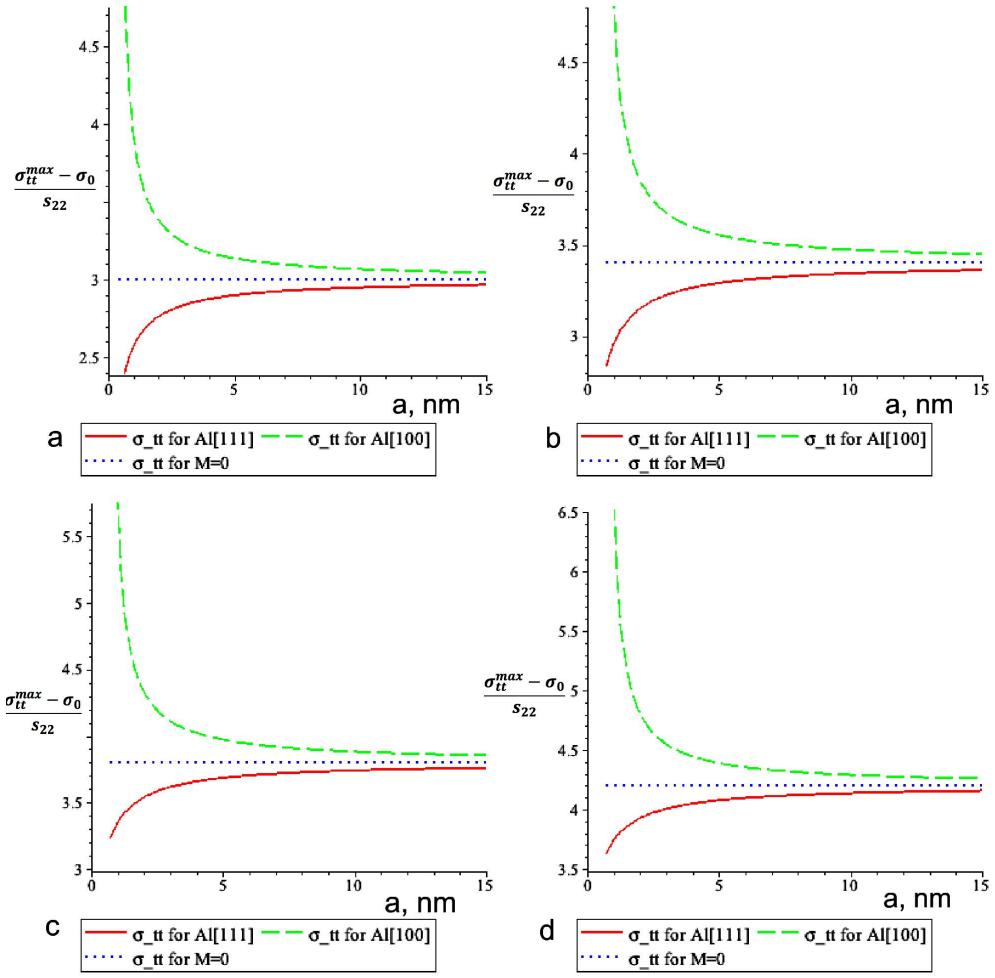



Figure 2: Dependence of maximum hoop stresses upon the radius $a$ of the basic circular hole when $\varepsilon = = 0; 0,1; 0,2; 0,3$ (a, b, c, d).

## 7 CONCLUSIONS

In the paper, we have presented the solution of the two-dimensional boundary value problem on a nanohole slightly deviated from the circular one in an infinite elastic solid. In particular:

- The analytical approximate solution of the problem on a stress-strain state of an elastic solid containing nanosized asperities at the surface is constructed by means of the perturbation method developed in [27].
- Different forms of surface asperities have been considered.
- The rigorous mathematical algorithm of solving the problem to any-order approximation of the perturbation technique has been developed.
- The solution of the integral equation is evaluated analytically in terms of a power series.
- The expressions for the hoop stress in the zero-order and first-order approximations were obtained. In contrast with [22], the solution is built without the use of conformal mapping.

- The size effect as a dependence of the stress state on the size of the hole has been demonstrated.

## ACKNOWLEDGEMENT

The work was supported by the Russian Foundation for Basic Research under the grant 14-01-00260.

## REFERENCES

[1] Y.G. Pronina, Analytical solution for decelerated mechanochemical corrosion of pressurized elastic-perfectly plastic thick-walled spheres. *Corrosion Science*, **90**, 161–167, 2015.

[2] O. Sedova, Y. Pronina, Generalization of the Lame problem for three-stage decelerated corrosion process of an elastic hollow sphere. *Mechanics Research Communications*, **65**, 30–34, 2015.

[3] F. Rosei, Nanostructured surfaces: challenges and frontiers in nanotechnology. *Journal of Physics: Condensed Matter*, **16**, S1373–S1436, 2004.

[4] P. Moriarty, Nanostrutured materials. *Reports on Progress in Physics*, **64**, 297–381, 2001.

[5] J. Wang, H. Huang, Yu.S. Duan, X. Feng, G. Wang, W. Zhang, T. Wang, Surface stress effect in mechanics of nanostructured materials. *Acta Mechanica Solida Sinica*, **24**, 52–82, 2011.

[6] J.W. Gibbs. *The Scientific Papers of J. Willard Gibbs, Vol. 1*. Longmans Green, London, 1906.

[7] R.C. Cammarata, Surface and interface stress effects on interfacial and nanostructured materials. *Materials Science and Engineering*, **A237**, 180–184, 1997.

[8] R.E. Miller, V.B. Shenoy, Size-dependent elastic properties of nanosized structural elements. *Nanotechnology*, **11**, 139–147, 2000.

[9] V.B. Shenoy, Atomic calculations of elastic properties of metallic fcc crystal surfaces. *Physical Review*, **B71**, 94–104, 2005.

[10] R. Dingrevill, J. Qu, M. Cherkaoui, Surface free energy and its effect on the elastic behavior of nanosized particles, wires and films. *Journal of the Mechanics and Physics of Solids*, **53**, 1827–1854, 2005.

[11] M.E. Gurtin, A.I. Murdoch, A continuum theory of elastic material surfaces. *Archive of Rational Mechanics and Analysis*, **57**, 291–323, 1975.

[12] M.E. Gurtin, A.I. Murdoch, Surface stress in solids. *International Journal of Solids and Structures*, **14**, 431–440, 1978.

[13] M.E. Gurtin, J. Weissmuller, F. Larche, A general theory of curved deformable interfaces in solids at equilibrium. *Philosophical Magazine*, **A 78**, 1093–1109, 1998.

[14] Yu.Z. Povstenko, Theoretical investigation of phenomena caused by heterogeneous surface tension in solids. *Journal of the Mechanics and Physics of Solids*, **41**, 1499–1514, 1993.

[15] P. Sharma, S. Ganti, N. Bhate, Effect of surfaces on size-dependent elastic state of nano-inhomogeneities. *Applied Physics Letter*, **82**, 535–537, 2003.

[16] L. Tian, R.K.N.D. Rajapakse, Elastic field of an isotropic matrix with nanoscale elliptical inhomogeneity. *International Journal of Solids and Structures*, **44**, 7988–8005, 2007.

[17] S.G. Mogilevsaya, S.I. Crouch, H.K. Stolarski, Multiple interacting circular nano-inhomogeneities with surface/interface efects. *Journal of the Mechanics and Physics of Solids*, **56**, 2298–2327, 2008.

[18] H.L. Duan, J. Wang, B.L. Karihaloo, Theory of elasticity at the nanoscale. *Advances in Applied Mechanics*, **42**, 1–68, 2009.

[19] H. Altenbach, V.A. Eremeyev, N.F. Morozov, On equations of the linear theory of shells with surface stresses taken into account. *Mechanics of Solids*, **45**, 331–342, 2010.

[20] R.V. Goldstein, V.A. Gorodtsov, K.V. Ustinov, Effect of residual stress and surface elasticity on deformation of nanometer spherical inclusions in an elastic matrix. *Physical Mesomechanics*, **13**, 318–328, 2010.

[21] Yu.I. Vikulina, M.A. Grekov, The stress state of planar surface of a nanormeter-sized elastic body under periodic loading. *Vestnik St. Petersburg University. Mathematics*, **45**, 174-180, 2012.

[22] M.A. Grekov, A.A. Yazovskaya, Effect of surface elasticity and residual surface stress in an elastic body weakened by an elliptic hole of a nanometer size. *Journal of Applied Mathematics and Mechanics*, **78**, 172–180, 2014.

[23] A.O. Bochkarev, M.A. Grekov, Local instability of a plate with circular nanohole under uniaxial tension. *Doclady Physics*, **59**, 330–334, 2014.

[24] A.O. Bochkarev, M.A. Grekov, The influence of the surface stress on the local buckling of a plate with a circular nanohole. *"Stability and Control Processes" in Memory of V.I. Zubov (SCP), 2015 International Conference*. IEEE, 367–370, 2015. DOI: 10.1109/SCP.2015.7342150.

[25] M.A. Grekov, S.A. Kostyrko, Surface defect formation in nanosized film coatings due to diffusion. *2015 International Conference on Mechanics – Seventh Polyakhov's Reading*. IEEE, 1–4, 2015. DOI: 10.1109/POLYAKHOV.2015.7106732.

[26] S.A. Kostyrko, G.M. Shuvalov, Morphological stability of multilayer film surface during diffusion processes. *"Stability and Control Processes" in Memory of V.I. Zubov (SCP), 2015 International Conference*. IEEE, 392–395, 2015. DOI: 10.1109/SCP.2015.7342172.

[27] E.A. Bashkankova, A.B. Vakaeva, M.A. Grekov, Perturbation method in the problem on a nearly circular hole in an elastic plane. *Mechanics of Solids*, **50**, 198–207, 2015.

[28] M.A. Grekov, The perturbation approach for a two-component composite with a slightly curved interface. *Vestnik St-Petersburg University Ser 1. Matematika Mekhanika Astronomiya*, 81–88, 2004.

[29] A.B. Vakaeva, M.A. Grekov, Effect of surface stresses in an elastic body with a curvilinear nanohole. *"Stability and Control Processes" in Memory of V.I. Zubov (SCP), 2015 International Conference*. IEEE, 440–443, 2015. DOI: 10.1109/SCP.2015.7342166.

[30] V.V. Novozhilov, *Theory of elasticity*. Oxford: Pergamon Press, 1961.

[31] N.I. Muskhelishvili, *Some Basic Problems of the Mathematical Theory of Elasticity*. Groningen: Noordhoff, 1963.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE GEMA FRAMEWORK – AN INNOVATIVE FRAMEWORK FOR THE DEVELOPMENT OF MULTIPHYSICS AND MULTISCALE SIMULATIONS

**Carlos A. T. Mendes**[1], **Marcelo Gattass**[1], **and Deane Roehl**[1]

[1]Tecgraf / PUC-Rio
Rua Marques de Sao Vicente, 225. Rio de Janeiro-RJ. Brazil.
e-mail: {cmendes, mgattass, deane}@tecgraf.puc-rio.br

**Keywords:** Multiphysics Simulation, Configurable orchestration, Extensible environments, Framework, Basin modeling

**Abstract.** *This paper summarizes the GeMA (Geo Modelling Analysis) framework, a library intended to support the development of new multiphysics simulators and its integration with existing ones. GeMA uses software engineering techniques to allow engineers to focus on the programming of the physical simulation and letting the framework take care of data management and other necessary support functions to develop efficient, professional programs. These programs are capable of simulating complex problems that may involve multiple physics that interact in different spatial and time scales. GeMA architecture supports multiple simulation and coupling paradigms, with particular emphasis given to finite element methods. GeMA support includes the coupling of different physics, each one with possibly different spatial domain discretizations (meshes). It has functions to support efficient transfer of state variable values from one discretization to another. The framework also implements some important concepts of extensibility, through the combined use of plugins and abstract interfaces, configurable orchestration and fast prototyping through the use of the Lua language. In this paper, we also present some results of a 2D basin modeling test case that couples FEM non-linear temperature calculations, compaction and kinetic oil maturation and generation algorithms. This scenario includes a time evolving mesh and different time scales among physics.*

## 1 INTRODUCTION

The building of a new physical simulator comprises three major steps. In the first one, the problem at hand must be studied so that a physical representation can be created for it. This representation is then mapped to a set of equations that define the mathematical model of the problem, usually composed of ordinary or partial differential equations. In the second step, these equations are are then discretized to allow the problem to be solved. Finally, the third stage represents the construction of a simulation system that implements the models defined in the previous steps [1].

Clearly, the knowledge and skills necessary to determine the appropriate physical representation and to create the mathematical model of the problem, are different from those needed to transform this same model into a computer system capable of simulating the desired phenomena. This dichotomy is even more pronounced when the objective is not the simulation of a particular scenario, but the simulation of a class of scenarios supporting user given parameters, in an efficient and robust way.

Ideally, simulators should be developed by multidisciplinary teams. Nevertheless, it is common that these are implemented by researchers with extensive knowledge of the problem physics, but limited software engineering skills, resulting in systems that are more error-prone and difficult to maintain/extend, especially if the resulting system extrapolates the research environment and is used in a production environment by other users.

Physics and software engineering are different worlds. However, a framework supporting the development of multiphysics simulations can create a "bridge" between these two worlds, allowing the engineer to focus on the construction of the physical representation and on the mathematical model, being supported and guided by the framework to follow best programming practices. As a result, new simulators can be created in less time and with better quality, compatible with production environments.

The idea behind the GeMA (Geo Modelling Analysis) framework is to bring state of the art software engineering techniques such as extensibility, reusability, modularity and portability to engineering physical modeling, leaving engineers free to focus on the mathematical formulation of the physics of the problem. The framework takes care of all the data management and required support functions, speeding up code development. A central point is that the framework does not dictate how physics are simulated or how multiple physics are coupled. Those decisions can be made by engineers according to the scenario at hand.

## 2 THE GEMA FRAMEWORK

Executing a simulation in the GeMA environment is a two-step process. In the first step, the model data and the solution method description are loaded, defining, respectively, what will be simulated and how. In the second step, an orchestration script is executed to do the required calculations (Figure 1).

The orchestration script is the central object of the solution method. Its role is to allow the user to describe the sequence of processes that should be applied to the model, so that the desired results are calculated, providing the main simulation loop. This script is written by the user using the Lua [2] language.

The Lua language is an interpreted language specially built to be embedded into applications, allowing them to dynamically run user provided programs. It is considered a simple, easy to learn language that includes some powerful concepts, such as dynamic typing, garbage collection, functions as first class objects and the use of associative maps for building data structures.

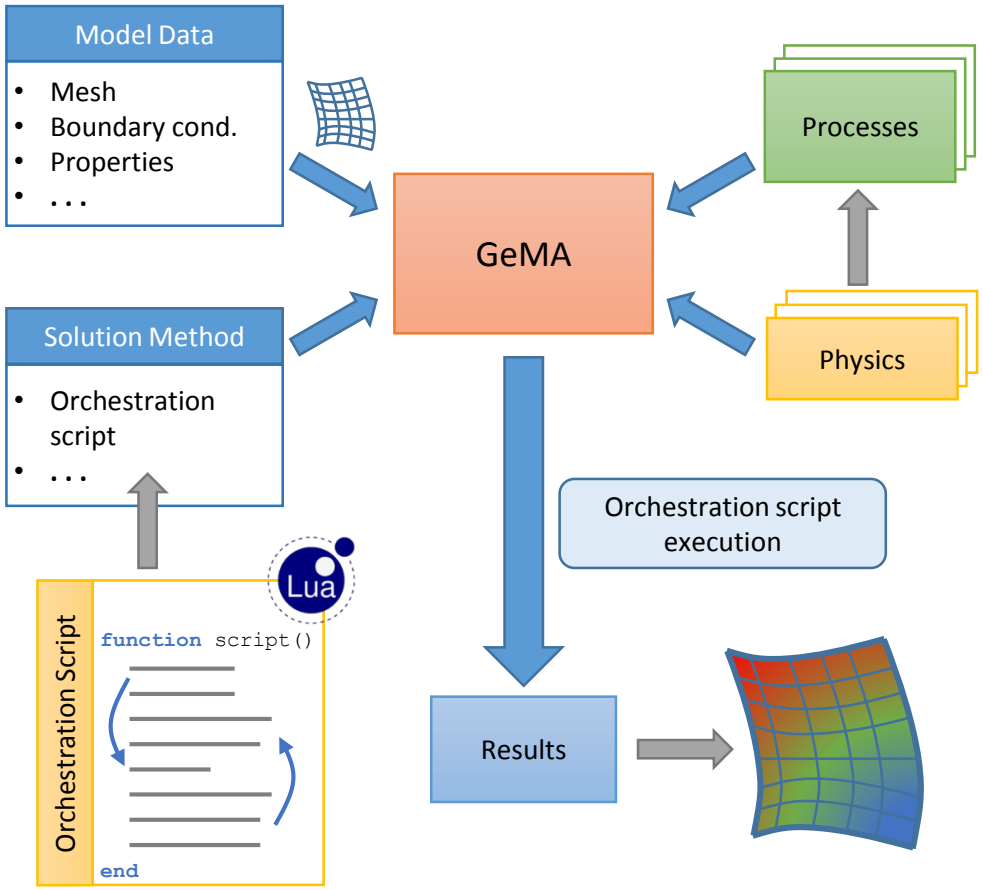



Figure 1: Main steps in a GeMA simulation.

It is an extensible language that allows the creation of domain-specific languages. Widely used as a language for integrating and extending applications, it is also considered as one of the fastest available interpreted languages[3].

By adopting a complete programming language which includes, among others, flow control, functions and support for data structures, instead of a simplified solution that only identifies the process execution order, the framework gives the orchestration script the freedom to run complex operations as needed. The GeMA orchestrator has a similar role as the orchestrator used by the Rocstar framework[4], but the flexibility given by the use of a script, instead of a C++ API as in Rocstar, allows for users with minimum computer language skills to be able to create their own orchestration models.

Processes are the basic unit used to describe the solution method and can be written in C++ or in the Lua language. In general, they are high-level primitives that describe a complete action, such as running a finite element analysis, transferring data between meshes, adaptively refining a mesh or saving a set of results.

Inside the framework, processes are abstract interfaces having concrete implementations given by plugins. Meshes, linear system solvers and other relevant entities are treated the same way. The use of abstract interfaces to model the main entities of the framework promotes its extensibility. The use of plugins, among other advantages, forces the existence of a clear separation of concepts and interdependencies, ensuring code modularity.

In this way, the combination of extensibility through abstract interfaces and plugins, with

the flexibility introduced by the orchestration script, provides the modeler with all the needed freedom to define how the simulation will be structured.

The possibility of implementing processes and physics in Lua, instead of in C++, allows for the framework to be used for rapid prototyping of new ideas, which, once tested and validated, can be converted into C++ code for greater efficiency, if at all needed.

The current framework implementation includes processes to support solutions based on the finite element method. These processes, in turn, define abstract physics interfaces, also implemented by plugins, which are responsible for providing the FEM process with the specific discrete mathematical formulation for a problem. Other discretization methods are supported by creating new processes. The framework also supports processes for integration with pre-existing simulators.

For several reasons, multiphysics models can contain more than one spatial domain discretization. To support those situations, where there is need to work with multiple meshes, at multiple scales, including heterogeneous types of elements and possibly representing different partitions of the spatial domain, the GeMA framework allows the simulation model to contain a set of meshes and includes processes for implementing data transfer between them, supported by spatial index data structures.

## 3 TEST CASES

Several test simulations were implemented to assess the framework correctness and expressiveness, including basic tests using the finite element method for stress and temperature calculation. Stress simulations were carried out with linear and non-linear trusses and with elements under plane stress state. Temperature simulations were based on heat conduction in steady and transient states, with several types of boundary conditions and the possibility of using Lua user defined functions to create temperature dependent material properties, such as thermal conductivity, making the problem nonlinear. Phase change scenarios were explored by the effective heat capacity method [5]. Results were compared with analytical models and/or with literature results.

Multiphysics simulations were tested by coupled stress-temperature models and by a considerably more complex scenario of a 2D sedimentary basin model, described below. This model includes treatment for several physical phenomena, such as geological layer sedimentation with mechanical compaction, thermal history and hydrocarbon maturation and generation. The basin evolution over time requires the use of a dynamic mesh to follow the deposition of sedimentary layers and igneous intrusions, whose presence makes it necessary to use adaptive time steps in the simulation. Temperature calculations were made based on the finite element method process, implemented in C++. Other calculations, including the mesh evolution over time, were implemented in Lua to evaluate the environment potential.

Basin modeling consists of a set of techniques designed to study the formation and the evolution of sedimentary basins. Through the use of physical simulation, basin models are used to characterize the petroleum system and to quantify potential hydrocarbon accumulations, clarifying the risks involved in exploration processes. More details can be found in [6].

The basin model under study is provided to the simulator through a mesh representing the current layer geometry, associated with a table that provides, for each layer, ages for its deposition start and end. Figure 2 shows the mesh used in the example simulation. Due to confidentiality issues, this model is not a real case, having been built to illustrate the simulator features. It contains two intrusive layers represented by the “diabase” layers.

Layer compaction and decompaction calculations are done in 1D through equation 1, given

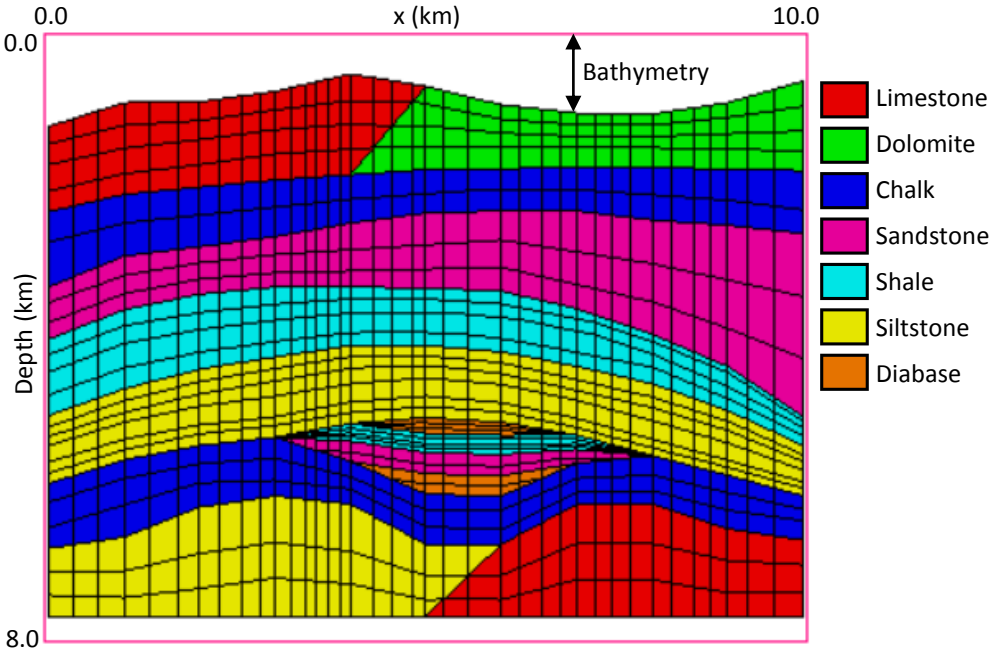



Figure 2: Sample basin model for the test simulation.

by [7], where $\phi_0$ and $c$ represent, respectively, the material initial porosity and compression rate, $z_1$ and $z_2$ are today's layer top and bottom depths and $z'_1$ and $z'_2$ are the layer top and bottom depths at the compaction/decompaction moment.

$$z'_2 = z'_1 + z_2 - z_1 - \frac{\phi_0}{c}(e^{-cz_1} - e^{-cz_2}) + \frac{\phi_0}{c}(e^{-cz'_1} - e^{-cz'_2}) \tag{1}$$

Some restrictions were imposed on the current mesh layout to ease simulation mesh evolution over time and compaction calculations. The mesh should be structured and composed of strips of vertically aligned quadrilaterals and/or triangles. Each layer should also consist of a single mesh row. If a material change is necessary within a layer (facies change), that must occur along a triangle diagonal, so that any vertical mesh edge always have the same material on both sides, as proposed in [8]. This allows for 1D compaction calculations along vertical mesh lines, without concerns about which material parameters should be used.

Temperature calculations over time are based on equation 2, where $\rho$ is the material density, $c_p$ is its specific heat capacity at constant pressure, $T$ is the temperature, $\lambda$ is the thermal conductivity and $G$ is the rate of internal heat generation. This equation is discretized in space with the finite element method and in time by an implicit method based on finite differences. Applied boundary conditions are given by the surface temperature, $T_s(x,t)$, and by the heat flow at the base of the basin, $q(x,t)$, both varying over geological time.

$$\rho c_p \frac{\partial T}{\partial t} = \nabla \cdot (\lambda \nabla T) + G \tag{2}$$

Sedimentary layers are porous media, and, therefore, material properties must reflect the grain-fluid mixture in each layer. To do so, the model considers that density and internal heat generation rates are porosity functions, $\rho(\phi)$ and $G(\phi)$. Specific heat capacity and thermal conductivity are also functions of the temperature, $c_p(\phi, T)$ and $\lambda(\phi, T)$, turning the problem into a non-linear one. If desired, the model can also consider that the thermal conductivity

is a function of the oil and gas saturation in the layer $\lambda(\phi, T, S_g, S_o)$. Mixture models and temperature dependency models used in the simulator can be found in [6].

Kinetic models are used to quantify the conversion rate of a compound into its derivatives, and can be used to quantify how the organic matter present in the source rock reacts to changes in temperature over time, reducing the initial quantity of kerogen and forming hydrocarbons. They can also be used to predict vitrinite reflectance (Ro), an important maturation indicator. Equation 3 calculates the fraction of the converted material, depending on the thermal history $T(t)$ and on the organic matter kinetic properties, given by its activation energy $E$ and frequency factor $A$. Assuming that the basins thermal history can be decomposed into multiple periods with constant heating rate, equation 3 can be solved analytically as shown in [9].

$$F(t) = 1 - e^{-\int_0^t A e^{-\frac{E}{RT(t)}}\, dt} \tag{3}$$

Figure 3 presents a schematic view of the simulation orchestration script, presenting the coupling model between the required physics and the main processes called in each time step.

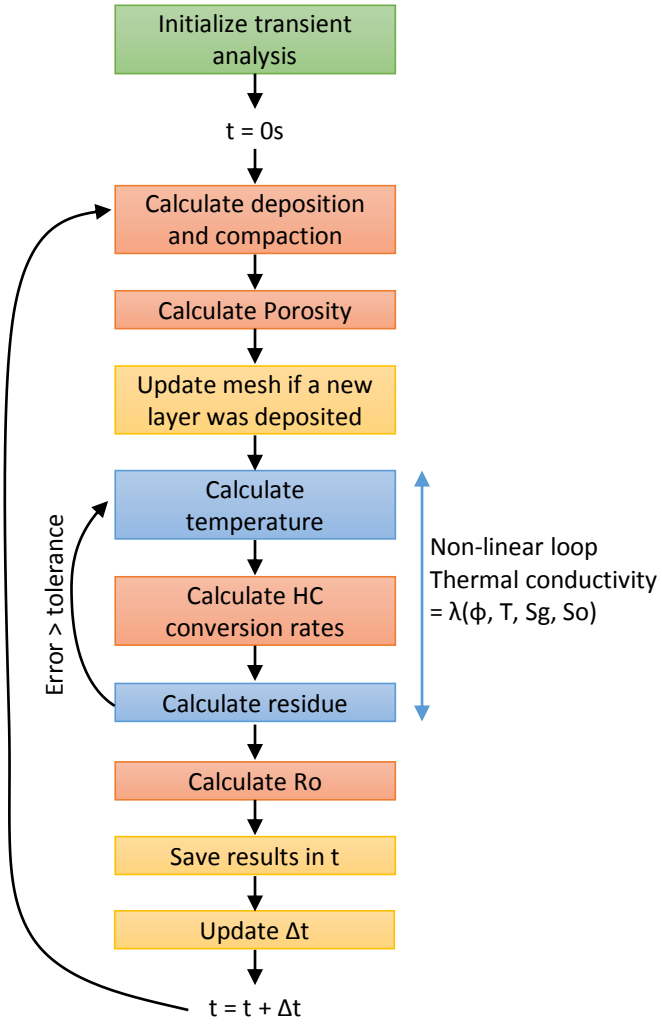



Figure 3: Schematic view of the basin simulation orchestration script.

The simulation uses a second mesh, initially empty, that is constructed and updated as layers are deposited to follow the progress of layer deposition and compaction. At each timestep, the

simulator calculates the fraction of the layer that will be deposited, ensuring a constant layer material deposition rate through time. This fraction is decompacted based on present thickness to calculate the size of the deposited layer. After deposition, older layers are compacted due to the weight of the new sediments and mesh node depths are adjusted accordingly.

If a time step involves the beginning of the deposition of a new layer, new elements are included in the mesh, and their initial temperature adjusted according to the surface temperature. If the new layer is presently over an intrusion that has not yet occurred so far in the simulation history, the new layer is "sewn" over the layer below the intrusion. Intrusion layers are always included instantaneously in a time step, being inserted between two layers, forcing the reorganization of the mesh elements in the upper one.

After compaction and porosity determination, temperature and hydrocarbon generation calculations are executed and repeated in a non-linear loop until convergence is achieved. The last process step consists in determining the time step to be used in the next iteration. After an intrusion, adaptive time steps are used to capture the cooling of the intruded body.

Figure 4 shows porosity, temperature, Ro and converted fraction values calculated at selected time steps. It is interesting to notice the influence of intrusions in Ro and converted fraction results. Temperature evolution in time steps following an intrusion event can be observed in Figure 5, which shows how the influence of the intrusion in the basin temperature history is localized and dissipates relatively quickly. In 100,000 years the effect has nearly vanished. Its impact on the maturation and generation of hydrocarbons in nearby rocks are, however, permanent.

## 4 CONCLUSIONS

The GeMA framework is, naturally, an extensible environment. The use of an architecture based on abstract interfaces and plugins allows new functionality to be easily added by third party users, without the need to modify or recompile existing code. It also promotes a separation of concepts and the decoupling between system modules, easing their development and maintenance.

Despite its current focus on simulations using the finite element method, the entire GeMA environment is prepared to include new process types, implementing support for other discretization methods. In particular, the given basin modeling example includes new processes based on analytical models for compaction, kinetic analysis and mesh evolution that were integrated with the finite element method used for temperature calculations.

The use of an orchestration script based on the Lua language as a central simulation component is largely responsible for the flexibility offered by the framework in the creation of multiphysics simulations, allowing the user to easily define coupling strategies and the actions to be taken at each time step.

The presented 2D basin modeling example illustrates some of the GeMA framework potential and flexibility. In its first version, the framework is currently being explored for the integration of hydro-mechanical simulators and their coupling with new chemical processes physics. It is being actively developed and will be further extended to deal with parallel simulations.

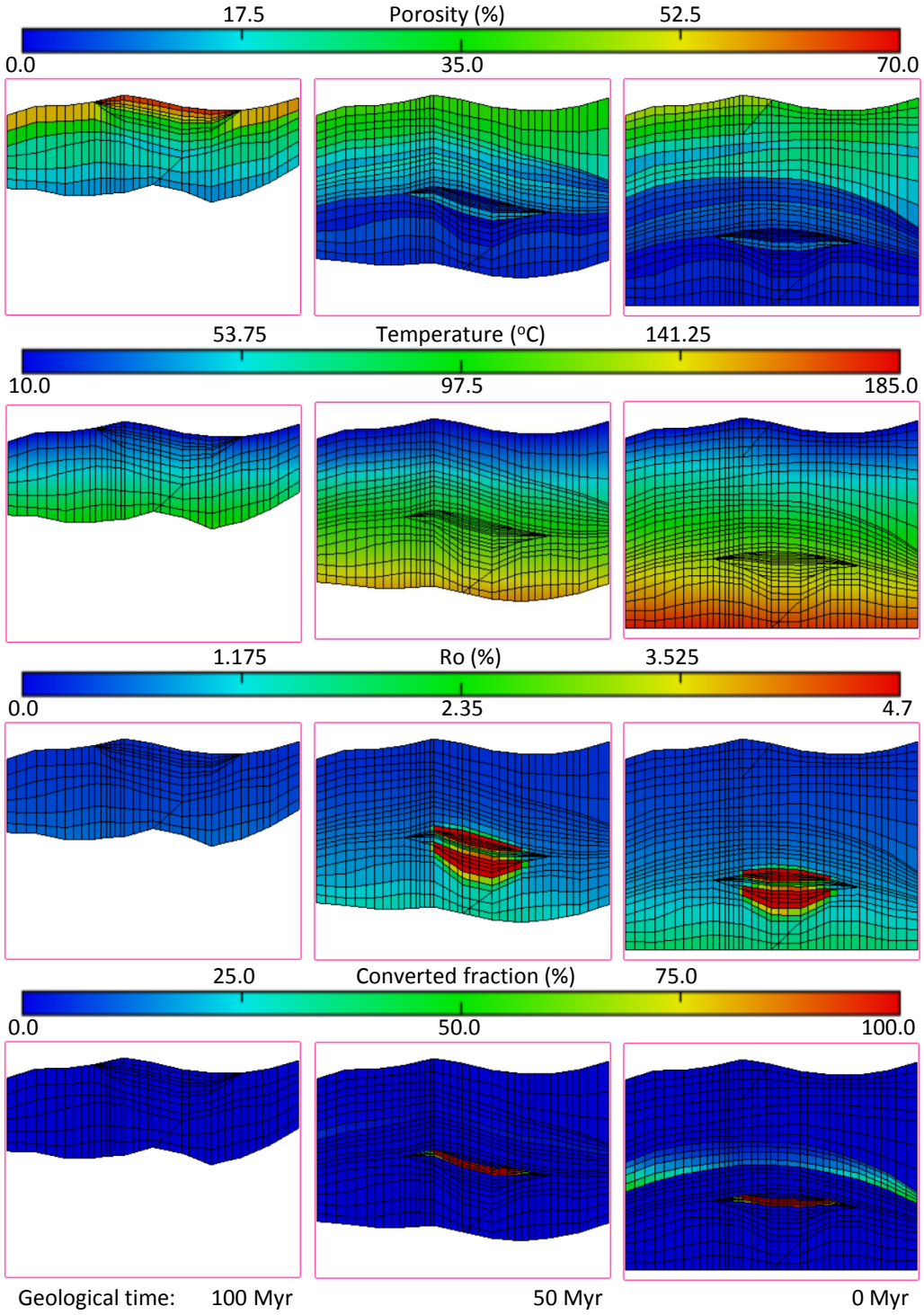



Figure 4: Basin analysis results on selected time steps.

## ACKNOWLEDGMENT

The authors gratefully acknowledge support from BG E&P Brasil through the "Coupled Geomechanics" project at TecGraf Institute (PUC-Rio) and the strategic importance of the support given by ANP (Brazils National Oil, Natural Gas and Biofuels Agency) through the R&D levy regulation.

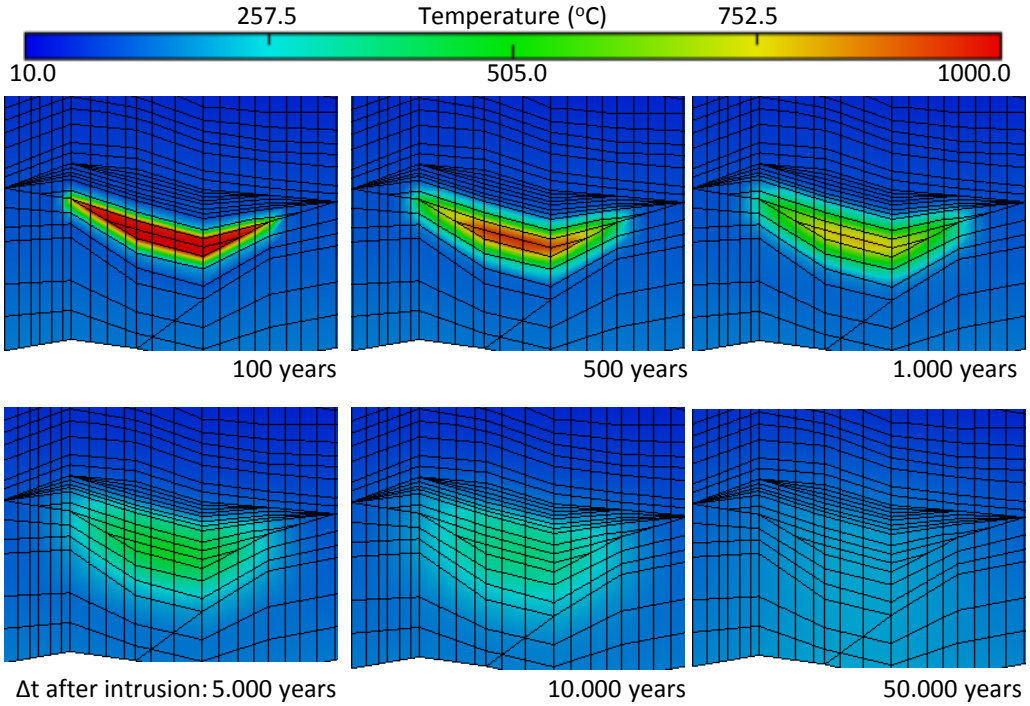



Figure 5: Temperature evolution in time steps following an intrusion event.

## REFERENCES

[1] J. Peir, S. Sherwin, Finite Difference, Finite Element and Finite Volume Methods for Partial Differential Equations. *Handbook of Materials Modeling*, 2415–2446, 2005.

[2] R. Ierusalimschy, L. Figueiredo, W. Celes, LuaAn Extensible Extension Language. *Software: Practice and Experience*, Vol. 26, Num. 6, 635–652, 1996.

[3] R. Ierusalimschy, *Programming in Lua*. 2003.

[4] X. Jiao, G. Zheng, P. A. Alexander, M. T. Campbell, O. S. Lawlor, J. Norris, A. Haselbacher, M. T. Heath, A system integration framework for coupled multiphysics simulations. *Engineering with Computers*, Vol. 23, Num. 3-4, 293–309, 2006.

[5] R. W. Lewis, P. Nithiarasu, K. N. Seetharamu, *Fundamentals of the Finite Element Method for Heat and Fluid Flow*. John Wiley & Sons, 2004.

[6] T. Hantschel, A. I. Kauerauf, *Fundamentals of Basin and Petroleum Systems Modeling*. Springer, 2009.

[7] J. G. Sclater, P. A. F. Christie, Continental stretching: An explanation of the post-mid-cretaceous subsidence of the central north sea basin. *Journal of Geophysical Research*, Vol. 85, Num. B7, 3711–3739, 1980.

[8] M. Wangen, Modelling heat and fluid flow in sedimentary basins by the finite element method. *Int. J. Numer. Anal. Methods Geomech.*, Vol. 15, Num. 10, 705–733, 1991.

[9] R. L. Braun, A. K. Burnham, Analysis of chemical reaction kinetics using a distribution of activation energies and simpler models. *Energy & Fuels*, Vol. 1, Num. 2, 153–161, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NONLINEAR PATCH METHOD AND APPLICATION

**Antoine Brunet[1], Pierre Sarrailh[1], François Rogier[2], Jean-François Roussel[1], and Denis Payan[3]**

[1]ONERA, The French Aerospace Lab, Department of Space Environment
2 Avenue Edouard Belin, Toulouse Cedex 4 31055, France
e-mail: {antoine.brunet,pierre.sarrailh,jean-francois.roussel}@onera.fr

[2] ONERA, Department of Information Processing and Modelisation
2 Avenue Edouard Belin, Toulouse Cedex 4 31055, France
e-mail: francois.rogier@onera.fr

[3] CNES, Centre spatial de Toulouse
18 avenue Edouard Belin, 31401 Toulouse Cedex 9, France
e-mail: denis.payan@cnes.fr

**Keywords:** finite elements, patch, multiscale, plasma simulation

**Abstract.** *We have extended the iterative multiscale patch method to a class of nonlinear problems using an adaptative linearization strategy. Theoretical convergence results are provided in this paper, and we showcase this nonlinear patch method in a simulation of a space plasma around a negatively charged array of solar generator interconnect.*

# 1 INTRODUCTION

Since 2005, the French aerospace laboratory (ONERA) together with the French and European space agencies (CNES and ESA) have been developing an open-source plasma-satellite simulation code called *Spacecraft Plasma Interaction System* (SPIS) [16]. Modern spacecraft technological challenges, such as the use of plasma thrusters and bigger solar generators, require simulation codes to resolve small-scale structures in relatively big computational domains. In order to solve these multiscale problems of interest to the industry, modern numerical methods have to be integrated in plasma simulation codes.

Several numerical methods have been proposed for the simulation of multiscale plasma effects. One important category of methods uses analytical models for the small scales, as is done in the *Variational Multiscale method* [13, 14]. Notably, it has been used to simulate the dynamics of an electric arc in a hot plasma coupled with a flowing gaz [18], and to model the transport of charges in fluids [20, 8]. While these methods are showing promising results, they require the integration of analytical models in the simulation codes. One can use numerical models to solve these scales instead, such as what is done in the case of highly oscillatory problems with the *Finite Element Heterogeneous Multiscale Method* [2, 5, 1]. These methods are effective for solving multiscale problems with non-local small-scale dynamics, such as fast varying medium properties.

In our case however, the small-case effects are due to local singularities, for instance on a boundary condition. Facing this type of problems, *numerical zoom* and *finite element patches* have been used in various fields [10, 3, 15, 4]. We propose a linearization scheme that allows to use the patch method to solve a class of nonlinear problems. These methods are sufficiently flexible to be easily integrated in current plasma simulation codes such as SPIS, and allow the resolution of local singularities to simulate their impact on the global problem.

We showcase this numerical method in a simulation of a plasma around a negatively charged array of solar generator interconnects. These spacecraft components are typically highly interacting with the surrounding plasma, and are too small to be resolved with classical numerical schemes.

# 2 THE NONLINEAR PATCH METHOD

## 2.1 Overview

The patch method [19, 7] is a method to solve elliptic multiscale problems using multiple grids. It is a flexible domain decomposition method similar to the Chimera method used in computational fluid dynamics [17].

The algorithm used in this method is quite straightforward. The problem is successively solved on a coarse mesh and on a finer, local mesh called *patch* to compute corrections added the complete solution. An extensive analysis of this method has been carried out in [19], including the derivation of convergence properties. This algorithm can be seen as a multiplicative Schwarz domain decomposition method, without conformity between the meshes. Convergence has been proven, and links between the convergence rate and the meshes geometry have been established [19] in the linear case. A fast converging variant of the patch method using harmonic functions has been presented in [9].

We here introduce an extension of the patch method to solve a class of nonlinear problems. We propose a linearization strategy that provides good convergence while limiting the number of linearizations. The proposed method is similar to a patch method variant of Newton-MG [11], but does not require the full convergence of the multigrid solver loop between linearizations.

### 2.2 Definitions

Consider a nonlinear problem on the domain $\Omega \subset \mathrm{R}^n$, which weak formulation can be written

$$\langle a(\mathbf{u}) \,|\, \varphi \rangle = l(\varphi),\ \forall \varphi \in H_0^1(\Omega) \tag{1}$$

Here, $a \in \mathcal{C}^1(H_0^1(\Omega))$ is a nonlinear operator, and $l$ is a continuous form on $H_0^1(\Omega)$. For all $\mathbf{u} \in H_0^1(\Omega)$, the Jacobian $J_\mathbf{u}$ of $a$ near $\mathbf{u}$ is supposed continuous and coercive.

Suppose the solution $\mathbf{u}$ of the problem exists and varies rapidly in a small sub-domain $\Lambda \subset \Omega$, but varies slowly in $\Omega \backslash \Lambda$. We define a *coarse mesh* of $\Omega$, with associated function space $V_H \subset H_0^1(\Omega)$, and a *fine mesh* of $\Lambda$ called *patch*, associated to the function space $V_h \subset H_0^1(\Lambda)$. We set $V = V_H + V_h$ and approximate the solution $\mathbf{u}$ of the continuous problem with the solution $\mathbf{u}_{Hh}$ of

$$\langle a(\mathbf{u}_{Hh}) \,|\, \varphi \rangle = l(\varphi),\ \forall \varphi \in V \tag{2}$$

One should note that, even when $a$ is linear, this problem is not trivial because it is generally not possible to construct a finite element basis of the function space $V$. The patch method builds the solution $\mathbf{u}_{Hh}$ by solving iteratively on $V_H$ and $V_h$, adding a source term to the problem, corresponding to the coupling between the meshes.

The other difficulty arrising in this problem is the nonlinearity of the operator. Several linearization schemes have been studied and provide convergence of the method toward the solution of the nonlinear problem:

- Unconditionnal linearization: provided the initial error is not too big, it can be shown that linearizing the problem between each iteration will allow the solution to converge. Using this scheme, a high number of linearization are required, but numerical results show good stability.
- Linearization at convergence: similarly to the Newton-MG method, one can choose to linearize the problem only when the previous linearized problem has been solved to a given accuracy. This scheme minimizes the number of linearizations, but will require a high number of overall iterations. Numerically, we have found that the stability of this scheme is relatively poor.
- Conditionnal linearization: we propose a convergence criterion based on the relative linearization and solving errors. As we will show in this paper, this scheme requires a limited number of linearizations while ensuring good overall convergence and stability.

Let us introduce the following notation:

$$\forall (\varphi, \mathbf{u}_\ell, \delta\mathbf{u}) \in V^3, \langle R_{\mathbf{u}_\ell}(\delta\mathbf{u}) \,|\, \varphi \rangle = l(\varphi) - \langle a(\mathbf{u}_\ell) + J_{\mathbf{u}_\ell}\delta\mathbf{u} \,|\, \varphi \rangle \tag{3}$$

In equation 3, $R_{\mathbf{u}_\ell}(\delta\mathbf{u})$ can be seen as a residual in the linearized problem at $\mathbf{u}_\ell$, with approximate solution $\delta\mathbf{u}$. It is readily seen that equation 2 is equivalent to $R_{\mathbf{u}_{Hh}}(\mathbf{0}) = \mathbf{0}$.

### 2.3 Proposed algorithm

The proposed iteratif algorithm can be written as the following:

1. Find $\mathbf{u}_h \in V_h$ s.t. $\forall \varphi_h \in V_h$,

$$\langle J_{\mathbf{u}_\ell}\mathbf{u}_h \,|\, \varphi_h \rangle = \left\langle R_{\mathbf{u}_\ell}(\mathbf{u}_n \quad - \mathbf{u}_\ell) \,\middle|\, \varphi_h \right\rangle$$

2. Let $\mathbf{u}_{n+\frac{1}{2}} = \mathbf{u}_n \quad + \omega \mathbf{u}_h$
3. Find $\mathbf{u}_H \in V_H$ s.t. $\forall \varphi_H \in V_H$,
$$\langle J_{\mathbf{u}_\ell} \mathbf{u}_H \,|\, \varphi_H \rangle = \left\langle R_{\mathbf{u}_\ell}(\mathbf{u}_{n+\frac{1}{2}} - \mathbf{u}_\ell) \,\middle|\, \varphi_H \right\rangle$$
4. Let $\mathbf{u}_{n+1} = \mathbf{u}_{n+\frac{1}{2}} + \omega \mathbf{u}_H$
5. If $\|R_{\mathbf{u}_\ell}(\mathbf{u}_{n+1} - \mathbf{u}_\ell)\| < \epsilon \|R_{\mathbf{u}_\ell}(\mathbf{0})\|$, let $\mathbf{u}_\ell = \mathbf{u}_{n+1}$

Note that the steps 1 and 3 are resolutions of the linearized problem, respectively on the $V_h$ and $V_H$ subspaces. The relaxation parameter $\omega$ has been studied in [19], and an optimal value is computed from geometrical properties of the meshes used.

It is readily apparent that the biggest numerical challenge involved in this method is the computation of the mixed terms $a(u_H, \varphi_h)$ and $a(u_h, \varphi_H)$. In a FEM code, these terms can be approximated with a quadrature rule on the intersection of a coarse element and a patch element. The computation of such intersection is trivial when using 2D structured meshes, but become much more difficult when using 3D unstructured meshes. See [6] for an efficient algorithm solving this problem.

### 2.4 Convergence properties

Let us prove the convergence of the proposed method. Let $\mathbf{u}_\ell$ be an initial linearization and $\mathbf{u}'_\ell$ be the next computed linearization point. We note $\mathbf{u}^*_\ell$ the solution of the linearized problem around $\mathbf{u}_\ell$. It has been proven in [19] that the following holds:

**Lemma 1.** *Let $\mathbf{u}^*_\ell$ be the solution of the linearized problem. There is a constant $C \in ]0,1[$ independent of the initial approximation $\mathbf{u}_\ell$, such that*

$$\|\mathbf{u}'_\ell - \mathbf{u}^*_\ell\| \leq C \|\mathbf{u}_\ell - \mathbf{u}^*_\ell\|$$

The Taylor expansion of $a$ near $\mathbf{u}$ can be written:

$$a(\mathbf{u} + \delta\mathbf{u}) = a(\mathbf{u}) + J_{\mathbf{u}_\ell}\delta\mathbf{u} + h_{\mathbf{u}_\ell}(\delta\mathbf{u}) \tag{4}$$

where $J_{\mathbf{u}}$ is the Jacobian of $a$ in $\mathbf{u}$ and $h_{\mathbf{u}}$ is a quadratic residual verifying:

$$\lim_{\delta\mathbf{u}\to 0} \frac{\|h_{\mathbf{u}}(\delta\mathbf{u})\|}{\|\delta\mathbf{u}\|} = 0 \tag{5}$$

Moreover, since $a$ is of class $\mathcal{C}^1$, the Jacobian $J_{(\cdot)}$ is continue, so

$$\lim_{\delta\mathbf{u}\to 0} \|J_{\mathbf{u}} - J_{\mathbf{u}+\delta\mathbf{u}}\| = 0 \tag{6}$$

Using the Taylor expansion around $\mathbf{u}$ and $\mathbf{v}$ gives:

$$\begin{aligned} a(\mathbf{u}) - a(\mathbf{v}) &= J_{\mathbf{v}}(\mathbf{u} - \mathbf{v}) + h_{\mathbf{v}}(\mathbf{u} - \mathbf{v}) \\ &= J_{\mathbf{u}}(\mathbf{u} - \mathbf{v}) - h_{\mathbf{u}}(\mathbf{v} - \mathbf{u}) \end{aligned}$$

Hence we have the following equation:

$$\forall (\mathbf{u}, \mathbf{v}) \in V^2, h_{\mathbf{u}}(\mathbf{v} - \mathbf{u}) = (J_{\mathbf{u}} - J_{\mathbf{v}})(\mathbf{u} - \mathbf{v}) - h_{\mathbf{v}}(\mathbf{u} - \mathbf{v}) \tag{7}$$

**Lemma 2.** *Let $(\mathbf{u}_n)_{n\in\mathbb{N}} \in V^{\mathbb{N}}$ a convergent sequence and note $\mathbf{u}_\infty$ its limit. Then,*

$$\lim_{n\to+\infty} \frac{\|h_{\mathbf{u}_n}(\mathbf{u}_\infty - \mathbf{u}_n)\|}{\|\mathbf{u}_n - \mathbf{u}_\infty\|} = 0$$

*Proof.* From equation 7, we have

$$\forall(\mathbf{u},\mathbf{v}) \in V^2,\ \|h_{\mathbf{u}}(\mathbf{v}-\mathbf{u})\| \leq \|J_{\mathbf{v}} - J_{\mathbf{u}}\| \cdot \|\mathbf{u}-\mathbf{v}\| + \|h_{\mathbf{v}}(\mathbf{u}-\mathbf{v})\|$$

Hence, $\forall n \in \mathbb{N}$,

$$\frac{\|h_{\mathbf{u}_n}(\mathbf{u}_\infty - \mathbf{u}_n)\|}{\|\mathbf{u}_n - \mathbf{u}_\infty\|} \leq \|J_{\mathbf{u}_\infty} - J_{\mathbf{u}_n}\| + \frac{\|h_{\mathbf{u}_\infty}(\mathbf{u}_n - \mathbf{u}_\infty)\|}{\|\mathbf{u}_n - \mathbf{u}_\infty\|}$$

Using equations 5 and 6, this upper bound converges to zero when $\mathbf{u}_n \to \mathbf{u}_\infty$. □

**Lemma 3.** *Let $\alpha \in ]C, 1[$, where $C$ is the bound in lemma 1. Then*

$$\|\mathbf{u}_\ell^* - \mathbf{u}_{Hh}\| < \frac{\alpha - C}{1+\alpha}\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\| \Rightarrow \frac{\|\mathbf{u}_\ell' - \mathbf{u}_{Hh}\|}{\|\mathbf{u}_\ell - \mathbf{u}_{Hh}\|} < \alpha$$

*Proof.* Suppose $\|\mathbf{u}_\ell^* - \mathbf{u}_{Hh}\| < \frac{\alpha-C}{1+\alpha}\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\|$, then

$$\begin{aligned}
\|\mathbf{u}_\ell' - \mathbf{u}_{Hh}\| &= \|\mathbf{u}_\ell' - \mathbf{u}_\ell^* + \mathbf{u}_\ell^* - \mathbf{u}_{Hh}\| \\
&\leq \|\mathbf{u}_\ell' - \mathbf{u}_\ell^*\| + \|\mathbf{u}_\ell^* - \mathbf{u}_{Hh}\| \\
&< C\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\| + \frac{\alpha - C}{1+\alpha}\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\| \\
&= \alpha\left(1 - \frac{\alpha - C}{1+\alpha}\right)\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\| \\
&< \alpha\left(\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\| - \|\mathbf{u}_\ell^* - \mathbf{u}_{Hh}\|\right) \\
&\leq \alpha\|\mathbf{u}_\ell - \mathbf{u}_{Hh}\|
\end{aligned}$$

□

**Lemma 4.**

$$\forall \mathbf{u}_\ell \in V,\ J_{\mathbf{u}_\ell}(\mathbf{u}_\ell^* - \mathbf{u}_{Hh}) = h_{\mathbf{u}_\ell}(\mathbf{u}_{Hh} - \mathbf{u}_\ell)$$

*Proof.* $\forall \mathbf{v} \in V$,

$$\begin{aligned}
\langle J_{\mathbf{u}_\ell}(\mathbf{u}_\ell^* - \mathbf{u}_{Hh}) \,|\, \mathbf{v}\rangle &= \langle J_{\mathbf{u}_\ell}(\mathbf{u}_\ell - \mathbf{u}_{Hh}) \,|\, \mathbf{v}\rangle + \langle J_{\mathbf{u}_\ell}(\mathbf{u}_\ell^* - \mathbf{u}_\ell) \,|\, \mathbf{v}\rangle \\
&= \langle J_{\mathbf{u}_\ell}(\mathbf{u}_\ell - \mathbf{u}_{Hh}) \,|\, \mathbf{v}\rangle + l(v) - \langle a(\mathbf{u}_\ell) \,|\, \mathbf{v}\rangle \quad \mathbf{u}_\ell^* \text{ solution of linearized problem} \\
&= \langle J_{\mathbf{u}_\ell}(\mathbf{u}_\ell - \mathbf{u}_{Hh}) + a(\mathbf{u}_{Hh}) - a(\mathbf{u}_\ell) \,|\, \mathbf{v}\rangle \quad \mathbf{u}_{Hh} \text{ solution of initial problem} \\
&= \langle h_{\mathbf{u}_\ell}(\mathbf{u}_{Hh} - \mathbf{u}_\ell) \,|\, \mathbf{v}\rangle
\end{aligned}$$

□

**Theorem 5.** *Let $\alpha \in ]C, 1[$, where $C$ is the bound in lemma 1. Then*

$$\frac{\|h_{\mathbf{u}_\ell}(\mathbf{u}_{Hh} - \mathbf{u}_\ell)\|}{\mathcal{C}_{J_{\mathbf{u}_\ell}}\|\mathbf{u}_{Hh} - \mathbf{u}_\ell\|} < \frac{\alpha - C}{1 + 2\alpha - C} \ \Rightarrow\ \|\mathbf{u}_\ell' - \mathbf{u}_{Hh}\| < \alpha\|\mathbf{u}_\ell - \mathbf{u}_{Hh}\|$$

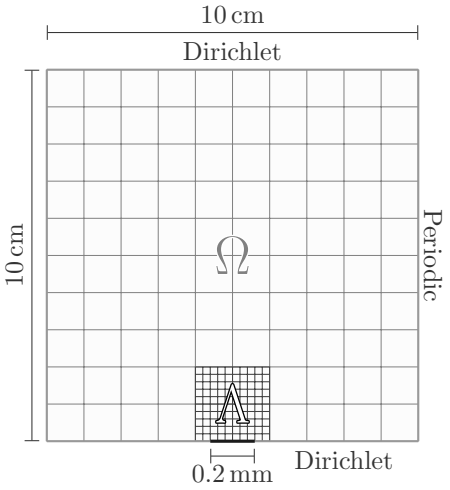



Figure 1: Computational domain and meshes

*Proof.* If $\mathbf{u}_\ell^* = \mathbf{u}_{Hh}$, lemma 3 gives the result. Suppose that $\mathbf{u}_\ell^* \neq \mathbf{u}_{Hh}$, then

$$\begin{aligned}
\frac{\|\mathbf{u}_\ell - \mathbf{u}_\ell^*\|}{\|\mathbf{u}_{Hh} - \mathbf{u}_\ell^*\|} &\geq \frac{\|\mathbf{u}_{Hh} - \mathbf{u}_\ell\| - \|\mathbf{u}_{Hh} - \mathbf{u}_\ell^*\|}{\|\mathbf{u}_{Hh} - \mathbf{u}_\ell^*\|} \\
&= \frac{\|\mathbf{u}_{Hh} - \mathbf{u}_\ell\|}{\|\mathbf{u}_{Hh} - \mathbf{u}_\ell^*\|} - 1 \\
&\geq \frac{\mathcal{C}_{J_{\mathbf{u}_\ell}} \|\mathbf{u}_{Hh} - \mathbf{u}_\ell\|}{\|J_{\mathbf{u}_\ell}(\mathbf{u}_\ell^* - \mathbf{u}_{Hh})\|} - 1 \\
&= \frac{\mathcal{C}_{J_{\mathbf{u}_\ell}} \|\mathbf{u}_{Hh} - \mathbf{u}_\ell\|}{\|h_{\mathbf{u}_\ell}(\mathbf{u}_{Hh} - \mathbf{u}_\ell)\|} - 1 \quad \text{from lemma 4}
\end{aligned}$$

Thus,

$$\frac{\|h_{\mathbf{u}_\ell}(\mathbf{u}_{Hh} - \mathbf{u}_\ell)\|}{\mathcal{C}_{J_{\mathbf{u}_\ell}} \|\mathbf{u}_{Hh} - \mathbf{u}_\ell\|} < \frac{\alpha - C}{1 + 2\alpha - C} \;\Rightarrow\; \|\mathbf{u}_\ell^* - \mathbf{u}_{Hh}\| < \frac{\alpha - C}{1 + \alpha} \|\mathbf{u}_\ell - \mathbf{u}_\ell^*\|$$

and lemma 3 gives the result. □

Using lemma 2, and since $\mathbf{u} \mapsto J_{\mathbf{u}}$ is continue, there exists a neighborhood $B$ of $\mathbf{u}_{Hh}$ verifying the left side expression in theorem 5, ensuring that, for every $\mathbf{u}_\ell \in B$, the proposed algorithm converges towards the solution $\mathbf{u}_{Hh}$.

## 3 NUMERICAL EXAMPLE

### 3.1 A nonlinear multiscale plasma problem

The simulation of multiscale plasmas can be done with a traditional FEM Poisson solver, using a nonuniform mesh with caracteristic sizes of elements spanning several orders of magnitude. This is the method currently used in the Spacecraft Plasma Interaction Software (SPIS) to solve local sub-centimeter scales in complete satellite simulations [16]. When reaching to sub-millimeter scales, the discrepancies on the size of the elements lead to an ill-conditioned problem, which can be difficult to solve accurately. It is of interest to investigate new numerical methods, such as the nonlinear patch method, which can avoid this issue by using multiple

meshes to solve each scale of the simulation. We will present here an application of the non-linear patch method to solve the Poisson-Boltzmann equation describing the electrostatic potential in a plasma at thermodynamic equilibrium[12]. This nonlinear equation is often used in spacecraft-plasma simulations with low or negative potentials, modelling both the electrostatic field and the electronic density distribution.

Here we present the simulation of a plasma around a negatively-biased array of solar cell interconnects, which are small conductors exposed to the spacecraft environment. The Poisson-Boltzmann equation can be written as:

$$-\Delta\varphi - \frac{q_e c_e^0}{\epsilon_0} \exp\left(-\frac{q_e \varphi}{k_B T}\right) = \frac{\rho_i}{\epsilon_0} \tag{8}$$

where $\epsilon_0$ is the medium permitivity, $\varphi$ is the electrostatic potential, $q_e$ is the electric charge of an electron, $c_e^0$ the electronic density at 0V, $\rho_i$ the ionic charge density, and $k_B T$ the electron temperature.

The computational domain and the used structured meshes are presented in figure 1. The spacecraft is modelled by a Dirichlet condition applied on the south boundary, were small local discontinuities represent the interconnects. On the north boundary, an homogenous Dirichlet condition models the space plasma. A patch is added locally, next to the interconnect.

The ion charge density will be computed using a Particle in Cell (PiC) solver, starting with an uniform density $c_i^0 = c_e^0$.

### 3.2 Numerical results

Using $(\varphi)_i$ and $(\psi)_i$ as basis of the subspaces $V_h$ and $V_H$, we note

$$\|R_{\mathbf{u}_\ell}(\mathbf{u})\|^2 = \sum_i \langle R_{\mathbf{u}_\ell}(\mathbf{u}) \,|\, \varphi_i \rangle^2 + \sum_i \langle R_{\mathbf{u}_\ell}(\mathbf{u}) \,|\, \psi_i \rangle^2 \tag{9}$$

Hence, at iteration $n$ using linearization at $\mathbf{u}_\ell$, $\|R_{\mathbf{u}_\ell}(\mathbf{0})\|$ is the norm of the linearization residual, and $\|R_{\mathbf{u}_\ell}(\mathbf{u}_n)\|$ is the norm of the residual in the linearized system.

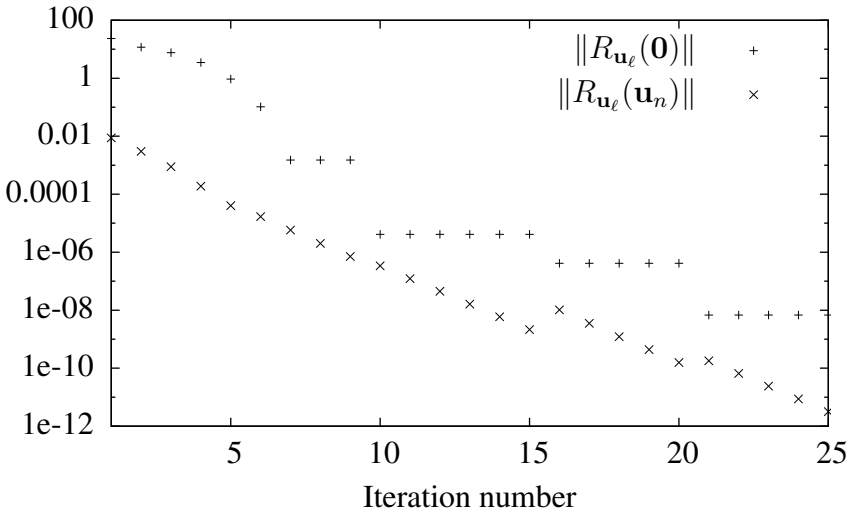



Figure 2: Convergence of the residuals

Figure 2 shows the convergence of these residuals using our method for a coarse mesh of 11x11 cells, and a refined patch of 120x120 cells covering a 3x3 coarse cells area near the interconnect. The relinearization criterion has been set to $\epsilon = 10^{-2}$.

On this figure, $\|R_{\mathbf{u}_\ell}(\mathbf{0})\|$ only changes on iterations involving linearization. Note that during the first 8 iterations in this example, the problem is systematicaly linearized, and the method is equivalent to a coarse Newton method. This is due to the fact that one patch iteration is enough to make the solving error small relative to the linearization error. During these initial iterations, the overall convergence is quadratic, as a typical Newton method convergence.

As soon as the linearization error reaches the same level as the solving error, our scheme allows for more patch iterations between linearizations. This can be seen in figure 2, where $\|R_{\mathbf{u}_\ell}(\mathbf{0})\|$ is levelling for several iterations, starting at iteration numbers 8, 10, 16 and 21.

Notice that linearizing may increase the patch residual $\|R_{\mathbf{u}_\ell}(\mathbf{u}_n)\|$, for instance at iteration number 16, but that the linearization residual $\|R_{\mathbf{u}_\ell}(\mathbf{0})\|$ is monotonically decreasing. Temporary increase of patch residual is due to a previous over-resolution of the system, and can be avoided by increasing the $\epsilon$ convergence parameter. Doing so will increase the frequency of linearization, and allows to weight the cost of linearizing against the cost of a patch iteration. A good value for the $\epsilon$ parameter would be the smallest one that ensures strict monotony for the patch residual.

Overall, the convergence is assymptotically linear, as with the linear patch method.

## 4 CONCLUSION

A linearization scheme has been proposed to solve nonlinear multiscale problems using the patch finite elements method, and a theoretical convergence result has been given. The resulting nonlinear patch method allows nonlinear multiscale, such as the simulation of small satellite elements in the space environment, to be solved efficiently with the required precision. This method is being implemented in the numerical engine of SPIS, and will provide a mean for industrial and scientific partners to simulate multiscale effects in the interaction between spacecrafts and their surrounding plasma.

### Acknowledgements

This work is supported by the French national space agency (CNES) and the French aerospace laboratory (ONERA).

## REFERENCES

[1] Assyr Abdulle, Yun Bai, and Gilles Vilmart. Reduced basis finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *Discrete and Continuous Dynamical Systems - Series S*, 8(1):91–118, July 2014.

[2] Assyr Abdulle, E Weinan, Björn Engquist, and Eric Vanden-Eijnden. The heterogeneous multiscale method. *Acta Numerica*, 21:1–87, May 2012.

[3] Jean-Baptiste Apoung Kamga and Olivier Pironneau. Numerical zoom for multiscale problems with an application to nuclear waste disposal. *Journal of Computational Physics*, 224(1):403–413, May 2007.

[4] Mathilde Chevreuil, Anthony Nouy, and Elias Safatly. A multiscale method with patch for the solution of stochastic partial differential equations with localized uncertainties. *Computer Methods in Applied Mechanics and Engineering*, 255:255–274, March 2013.

[5] Weinan E and Bjorn Engquist. The Heterognous Multiscale Methods. *Communications in Mathematical Sciences*, 1(1):87–132, 2003.

[6] Martin Jakob Gander and C. Japhet. An Algorithm for Non-Matching Grid Projections with Linear Complexity. 2009.

[7] Roland Glowinski, Jiwen He, Jacques Rappaz, and Joël Wagner. Approximation of multi-scale elliptic problems using patches of finite elements. *Comptes Rendus Mathematique*, 337(10):679–684, November 2003.

[8] Gabriel M. Guerra, Souleymane Zio, Jose J. Camata, Fernando A. Rochinha, Renato N. Elias, Paulo L.B. Paraizo, and Alvaro L.G.A. Coutinho. Numerical simulation of particle-laden flows by the residual-based variational multiscale method. *International Journal for Numerical Methods in Fluids*, 73(8):729–749, November 2013.

[9] Jiwen He, Alexei Lozinski, and Jacques Rappaz. Accelerating the method of finite element patches using approximately harmonic functions. *Comptes Rendus Mathematique*, 345(2):107–112, July 2007.

[10] Frédéric Hecht, Alexei Lozinski, and Olivier Pironneau. Numerical Zoom and the Schwarz Algorithm. In *Domain Decomposition Methods in Science and Engineering XVIII*, number 70 in Lecture Notes in Computational Science and Engineering, pages 63–73. Springer Berlin Heidelberg, January 2009.

[11] Van E. Henson. Multigrid methods nonlinear problems: an overview. volume 5016, pages 36–48, 2003.

[12] M. J. Holst. The Poisson-Boltzmann Equation. *An analysis and multilevel numerical solution*, 1994.

[13] Thomas J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127(1–4):387–401, November 1995.

[14] Thomas J. R. Hughes, Gonzalo R. Feijóo, Luca Mazzei, and Jean-Baptiste Quincy. The variational multiscale method - a paradigm for computational mechanics. *Computer Methods in Applied Mechanics and Engineering*, 166(1–2):3–24, November 1998.

[15] Marco Picasso, Jacques Rappaz, and Vittoria Rezzonico. Multiscale algorithm with patches of finite elements. *Communications in Numerical Methods in Engineering*, 24(6):477–491, June 2008.

[16] J. Roussel, F. Rogier, G. Dufour, J.-C. Mateo-Velez, J. Forest, A. Hilgers, D. Rodgers, L. Girard, and D. Payan. SPIS Open-Source Code: Methods, Capabilities, Achievements, and Prospects. *IEEE Transactions on Plasma Science*, 36(5):2360–2368, October 2008.

[17] Joseph L. Steger and John A. Benek. On the use of composite grid schemes in computational aerodynamics. *Computer Methods in Applied Mechanics and Engineering*, 64(1–3):301–320, October 1987.

[18] Juan Pablo Trelles, Emil Pfender, and Joachim Heberlein. Multiscale Finite Element Modeling of Arc Dynamics in a DC Plasma Torch. *Plasma Chemistry and Plasma Processing*, 26(6):557–575, December 2006.

[19] Joël Wagner. *Finite element methods with patches and applications*. PhD thesis, EPF, Lausanne, Swiss Confederation, 2006.

[20] Guo-Wei Wei, Qiong Zheng, Zhan Chen, and Kelin Xia. Variational Multiscale Models for Charge Transport. 2012.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ASSESSMENT OF A CONTINUUM MICROMECHANICS-BASED MULTISCALE MODEL FOR CONCRETE BY MEANS OF SENSITIVITY ANALYSIS AND UNCERTAINTY PROPAGATION

**Luise Göbel[1], Andrea Osburg[2], Tom Lahmer[3]**

[1]Research Training Group 1462, Bauhaus-Universität Weimar
Berkaer Strasse 9, 99423 Weimar, Germany
e-mail: luise.goebel@uni-weimar.de

[2] F. A. Finger Institute for Building Material Engineering, Bauhaus-Universität Weimar
Coudraystrasse 11A, 99423 Weimar, Germany
e-mail: andrea.osburg@uni-weimar.de

[3] Institute of Structural Mechanics, Bauhaus-Universität Weimar
Marienstrasse 15, 99423 Weimar, Germany
e-mail: tom.lahmer@uni-weimar.de

**Keywords:** Multiscale Modeling, Concrete, Continuum Micromechanics, Uncertainty Propagation, Sensitivity Analysis, Variance Decomposition

**Abstract.** *In structural engineering, the behavior of materials is often defined at the macroscopic level where experiments are suitable. Utilizing the advanced knowledge of the microstructure, multiscale modeling represents a beneficial tool for material properties estimation. During the last two decades, multiscale models based on continuum micromechanics for concrete have been developed. The implementation of these multiscale models is performed in a two-step manner: First, the volume fractions of the phases at the lower scale are calculated, followed by an upscaling process. These models require a high number of input parameters which partially exhibit high uncertainties. The paper discusses the question how the uncertainties of the input parameters are propagated through the different scales of concrete during the upscaling process. It is investigated whether the upscaled mechanical properties will show a high variability due to the diversifying properties of the initial components. Sensitivity analysis is applied to cement pastes, mortars, and concretes with water-to-cement ratios of 0.40, and 0.60. The total order sensitivity indices according to the variance decomposition method by Sobol are computed for all input parameters. By means of interpretation of sensitivity indices, the importance of parameters on the model output is quantified. A more accurate a priori knowledge of the input parameters can lead to a decreased uncertainty of the upscaled results. Furthermore, the uncertainty propagation starting at two commonly applied hydration kinetics models is compared. The results show that the scatter of the model responses increases significantly during the upscaling process; hence, it should be considered in the assessment of the multiscale modeling results.*

## 1 INTRODUCTION

In civil engineering, the behavior of materials is mostly defined at the macroscopic level. Macroscopic models for the prediction of the materials' behavior are often empirical; therefore they contain parameters without a specific physical meaning.
Concrete is commonly handled as a homogeneous material at the macroscopic scale in prediction models. However, it behaves highly heterogeneous at the microscale. Utilizing the additional knowledge of the microstructure and the materials' chemistry, multiscale modeling represents a beneficial tool for material properties estimation. Recently, multiscale models based on continuum micromechanics have been developed and successfully applied to cementitious materials. Following the modeling philosophy in continuum micromechanics, no fitting parameters are introduced; rather, a particular physical meaning is assigned to the model parameters.
By means of multiscale models, properties regarding the elasticity, strength, and creep of concrete can be predicted using intrinsic elastic properties of the constituents and their volume fractions (e.g. [1], [2], [3]). This paper focuses on the multiscale elastic analysis. The prediction of elasticity via semi-analytical homogenization methods in the framework of continuum micromechanics is performed in a two-step manner. The calculation of the volume fractions of the phases at each scale is followed by an upscaling process. The evolution of the microstructure of cement paste during the hydration process is computed using hydration kinetics models. In this study, two commonly utilized hydration kinetics models are employed: the Powers-Acker hydration model [4], [5] and the hydration model proposed by Bernard [6] and refined by Pichler et al. [7].

Commonly, multiscale analyses are performed using a deterministic approach including established values of the input parameters. Calculations comprising the stochastic input parameters result in probabilistic multiscale analyses where the upscaled mechanical property represents a stochastic variable which is a function of the parameter variation. This seems to be useful because multiscale models require a large number of input parameters that commonly exhibit a high uncertainty. The question arises whether the upscaled mechanical properties will show an increased variability due to the diversifying properties of the initial components. The paper aims to study the propagation of uncertainties of the input parameters through the different scales of concrete during the upscaling process.
Uncertainty quantification in multiscale modeling has been rarely investigated so far. Studies have been conducted for several types of materials and with different methodologies of uncertainty quantification. Clément et al. [8] considered the uncertain nature of hyperelastic heterogeneous materials at the microscopic scale and proposed a methodology for the uncertainty quantification based on polynomial chaos representation. Vu-Bac et al. [9] proposed a stochastic multiscale method for polymer nanocomposites across four length scales and quantified the uncertainties of several input parameters using different types of sensitivity analyses. Particularly, approaches for independent as well as for correlated input parameters were introduced. Uncertainties in the multiscale modeling of concrete were investigated by Berveiller [10] and Venkovic et al. [11]. Berveiller in [10] discussed the variability of the Young's modulus of cement paste using polynomial chaos expansions. Venkovic et al. [11] computed the uncertainty propagation of a multiscale poromechanics-hydration model for concrete by means of stochastic meta-models and polynomial chaos expansions.

However, a detailed uncertainty analysis at different scales of a continuum micromechanics-based multiscale model for concrete has not been performed yet to the best of the authors'

knowledge. This study aims to assess the influence of the variabilities and uncertainties of the initial composition of cement-based materials on elastic properties across the scales. Additionally, differences in the statistical variation of the model output of two different hydration kinetics models are appraised. A framework for the application of the sensitivity analysis on probabilistic multiscale modeling for concrete is proposed. Results of the sensitivity analysis applied to cement pastes, mortars, and concretes, respectively, with water-to-cement ratios of 0.40 and 0.60 are presented.
The paper is organized as follows: First, the fundamentals of continuum micromechanics are recalled and the deterministic multiscale model is introduced. Then, probabilistic multiscale modeling is performed considering uncertainties of all input parameters. The uncertainties of the model predictions at different stages of the hydration process are computed. The parameters influencing the uncertainty most are identified, and their probability density functions at different time steps are estimated. The paper is concluded with a discussion and further recommendations.

## 2 DETERMINISTIC MICROMECHANICAL MODEL OF CONCRETE

Theoretical fundamentals of modeling and homogenization methods within the realm of micromechanics have been presented in the works by [12], [13], [14], [15], [16]. Subsequent investigations by [6], [17], [18][19], [1] extended the principles towards the application on cementitious materials and the upscaling of elastic, strength, and creep properties from a microscopic to a macroscopic observation level.

### 2.1 Micromechanical representation

Within the framework of continuum micromechanics, the concept of representative volume elements (RVE) including the separation of scales requirement is followed [20]. The microstructure of concrete cannot be resolved and described in every detail. Thus, at each scale of the multiscale model, one RVE comprising quasi-homogeneous subdomains with known physical quantities is defined [2].

The morphological model for concrete adopted here is based on [7]. The model includes four scales of observation which are shown in Figure 1. For the sake of simplicity, the shapes of all embedded phases are assumed to be sphericular. It is known that the distinction of different phases morphologies increases the accuracy of the model prediction [1].

### 2.2 Hydration kinetics models

There are several models describing the kinetics and the evolution of volume fractions in cement pastes during the hydration process of cement particles. The two most important and principally used models are the Powers-Acker hydration model [4], [5] (”Powers model”) and the hydration model proposed by Bernard [6], later improved by Pichler et al. [7] (”Bernard model”).

#### 2.2.1 The Powers model

The Powers model, often denoted as the ”engineering model”, provides the volume fractions of unhydrated clinker ($f_{clin}$), water ($f_{H_2O}$), hydration products ($f_{hyd}$), and of air ($f_{air}$) at the cement paste scale as functions of the initial water-to-cement ratio ($w/c$) and of the degree of hydration ($\xi$). The latter value is defined as the mass of hydration products formed up to

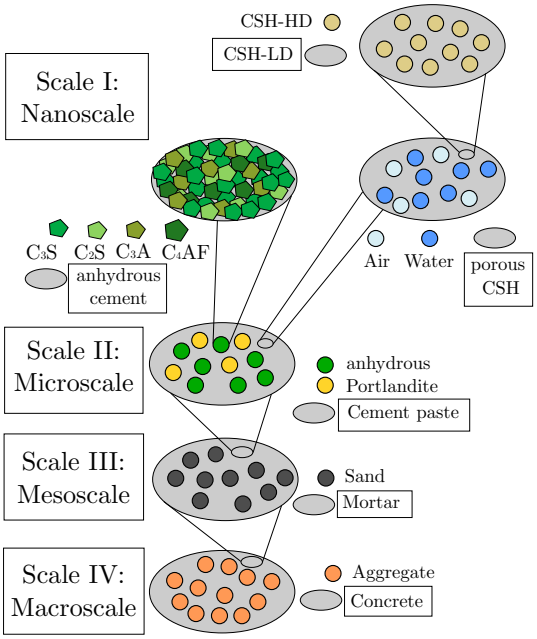



Figure 1: Scales of observation of the continuum micromechanics-based multiscale model (adopted from [7])

the current stage of the hydration process divided by the mass of hydrates formed in case of a completed hydration ($\xi = 1$). The calculation of the phase volume fractions is exemplary described in [2].

### 2.2.2 The Bernard model

Bernard et al. [6] proposed a hydration kinetics model which describes the evolution of relative volume properties of the elementary clinker phases. The model comprises stoichiometric and kinetic equations for the calculation of volume fractions of reactants and hydration products at different stages of the hydration. The underlying stoichiometric relations utilized in this model are adopted from Tennis and Jennings [21]. The kinetics of the hydration are described by kinetic laws that link the reaction rate $\mathrm{d}\xi/\mathrm{d}t$ to the affinity $A(\xi_x)$ as well as to kinetic constants that determine the characteristic time associated with the chemical reaction $\tau$. Detailed information can be found in [6].

## 2.3 Upscaling of the elastic properties

Once the morphological model is established, the mechanical behavior of the material can be estimated using homogenization schemes. The underlying fundamentals necessary for the upscaling procedure of elastic properties can be found e.g. in [6], [17], [18][19], [1].
The self-consistent (SC) scheme [14], [15] is appropriate for polycrystalline structures, i.e., for materials which phases are dispersed in the RVE. Then, the matrix medium coincides with the homogenized medium. For materials in which a continuous matrix can be distinguished from particular inclusions, the Mori-Tanaka (MT) scheme [22], [23] is suitable. In this work, the MT scheme was applied to all RVE, except to the anhydrous cement particles. There, the SC scheme was used.
The micromechanical model estimates the elastic properties of cement pastes, mortars, and

concretes as functions of the degree of hydration and of the composition. Figure 2 shows the homogenized Young's modulus derived from multiscale modeling using the two introduced hydration kinetics models at four scales of observation. The application of the homogenization schemes yields simplified equations of the effective shear $\mu_{hom}$ and bulk modulus $k_{hom}$ with which the Young's modulus can be computed. The respective equations can be found in [3].

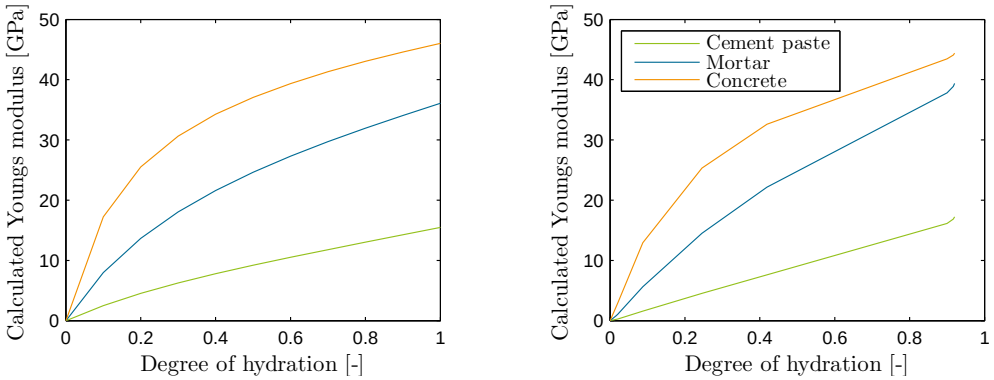



Figure 2: Deterministic Young's modulus for the Powers model (left) and the Bernard model (right) as functions of the degree of hydration calculated for a water-to-cement ratio of 0.60.

## 3 PROBABILISTIC MULTISCALE MODELING

Approaches for assessing the quality of the prognosis of engineering models are demanded. Methods based on uncertainty and sensitivity analyses represent suitable tools to rank input parameters according to their sensitivities on the output and to identify main contributors to the uncertainties of models. Utilizing results of these analysis helps to improve the accuracy of model predictions.

### 3.1 Assessment methodology

#### 3.1.1 Sensitivity Analysis

The sensitivity analysis determines how the uncertainty of the model output can be linked to the uncertainty of the model input parameters in a qualitative and a quantitative manner. The main objectives are the identification of input parameters which have the most significant influence on the model output and the quantification of their relative importance. In subsequent modeling steps, input parameters with the least influences can be considered as deterministic, i.e. fixing a non-influential parameter at one value of its range of variance.

In this study, the variance-based global sensitivity analysis according to Saltelli [24] was applied. In contrast to local sensitivity analyses, global ones consider the simultaneous variation of all stochastic input parameters.

For a model with a scalar output $Y$ as a function of $n$ random input parameter sets $X_i$ (i.e. $Y = f(X_1, X_2, ..., X_n)$), the first order sensitivity indices are calculated according to [25]:

$$S_i = \frac{V_{X_i}\left(E_{X_{\sim i}}(Y \mid X_i)\right)}{V(Y)}, \quad \sum_{i=1}^{k} S_i \leq 1, \tag{1}$$

where $V(Y)$ denotes the unconditional variance of the model output, $V_{X_i}\left(E_{X_{\sim i}}(Y \mid X_i)\right)$ is the variance of conditional expectation, and $X_{\sim i}$ identifies the matrix of all factors but $X_i$.

The first order sensitivity indices measure only the first order effect of $X_i$ on the model response, i.e. the influence of every single variable in a decoupled way. Hence, higher order terms have been introduced (total effect sensitivity indices) [24]:

$$S_{T_i} = 1 - \frac{V_{X_{\sim i}}\left(E_{X_i}(Y \mid X_{\sim i})\right)}{V(Y)}, \quad \sum_{i=1}^{k} S_{T_i} \geq 1, \tag{2}$$

where $Y \mid X_{\sim i}$ is the variance of the model response caused by all input parameters $X_i$.
The method of Saltelli can only be applied in case of uncorrelated the input parameters. Methods considering correlated input parameters are proposed by several researchers (e.g., [26]).

### 3.1.2 Uncertainty Analysis

Uncertainty describes the incomplete knowledge about models and parameters. By means of uncertainty analysis, the uncertainty of the model response as a result of the uncertainty in the input parameters is quantified. There are different sources of uncertainty in the modeling process due to measurement inaccuracies, natural variations in the data, and uncertain scientific backgrounds of models.

As a first step of the uncertainty analysis, all sources of uncertainties in the input parameters that affect the model output have to be identified. The specification of model probability distributions of all input parameters is required. These distributions are either reported in the literature or are based on empirical evidence. Subsequently, the probability densitiy functions (PDF) are used to generate sample sets independently for each of the input parameters. Then, the model is evaluated with the respective parameter sets and, finally, its output is analyzed using statistical methods.

### 3.1.3 Sampling Approaches

Several methods are used to generate samples from given probability density functions (PDF). Some of these include the Monte Carlo Simulation (MCS), Latin Hypercube Sampling (LHS) and Advanced Latin Hypercube Sampling (ALHS).
In this study, the method of LHS has been used, since it is independent of the number of random variables. Thus, the number of required samples is significantly reduced [27]. The underlying idea of this method is the allocation of the ranges of the input parameters into intervals that have equal marginal probabilities. In the presented probabilistic multiscale model, 10,000 samples for each input parameter have been generated.

## 3.2 Model input parameters

The two hydration kinetics models from section 2.2 use several input parameters. In total, there are 17 stochastic input parameters for the Powers model and 26 variables for the Bernard model. All parameters are assumed to be uncorrelated.

The stochastic properties and specific values for all of the input parameters are given in detail in Appendix A.1. Due to a lack of experimental data, some probability distributions were assured by the authors which introduces bias in the results of both the uncertainty and the sensitivity analysis [11]. This represents a future topic in research.

## 4 RESULTS

The elastic multiscale analysis allows the prediction of the Young's modulus of cement paste, mortar, and concrete for several water-to-cement ratios at different stages of hydration. In comparison to the deterministic approach, the probabilistic analysis allows the investigation of distributions of the model responses.

### 4.1 Prediction of a probabilistic Young's modulus

The multiscale model was evaluated for each of the sample sets. From the stochastic model responses, the PDF and the Coefficient of Variation (CoV) of the model responses at three scales of observation were determined.
Figure 3 shows the evolution of the CoV for the two hydration kinetic models, the Powers model and the Bernard model, as functions of the degree of hydration and for different water-to-cement ratios (w/c=0.40 and 0.60).

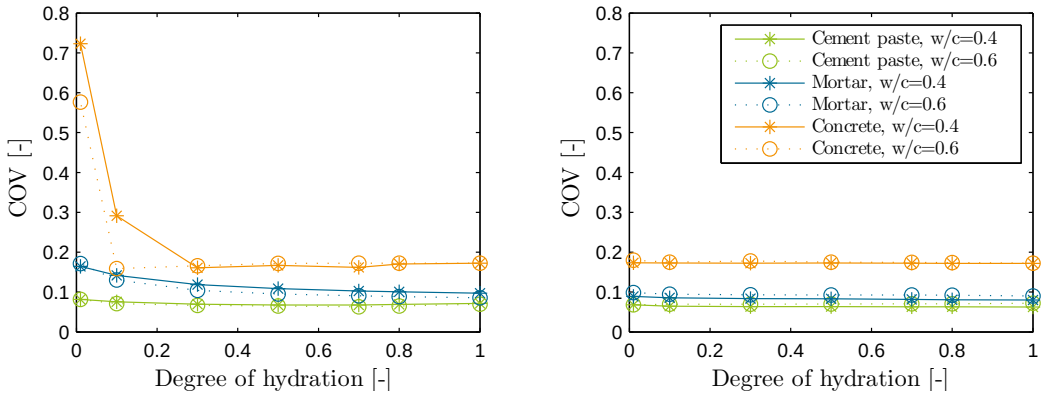



Figure 3: Evolution of the coefficient of variation of the predicted Young's modulus for the Powers model (left) and for the Bernard model (right) as functions of the degree of hydration and the w/c-ratio.

Apparently, the coefficient of variation increases during the upscaling process. At the cement paste scale, there is a fewer number of input parameters compared to the concrete scale. Thus, the CoV is the lowest at the cement paste scale. Both, the Powers model and the Bernard model, exhibit the same CoV at the fully hydrated stage. Major differences between the two hydration models occur at early hydration stages (Degree of hydration $< 0.4$). A higher variation in the Powers model is observed during the first stages of the hydration. In the Powers model, there are no kinetic calculations included; the volume fractions of the phases are solely functions of the w/c-ratio and the hydration degree. Since the reactions occuring in the early hydration are not described in the same detail as in the Bernard model, the model responses of Powers are constrained with larger CoV. With the advancement of the hydration process, the influence of the kinetic parameters on the Young's modulus is less significant. For the considered water-to-cement ratios, the evolutions of the CoV are almost similar to each other.

Figures 4 and 5 show the probability densities of the model responses of the two hydration kinetics model as functions of the degree of hydration. Using the 10,000 realizations of the models at different time steps, the mean, the standard deviation, and the probability distribution of the Young's modulus were estimated. The model responses of both hydration kinetic models were found to be normally distributed.

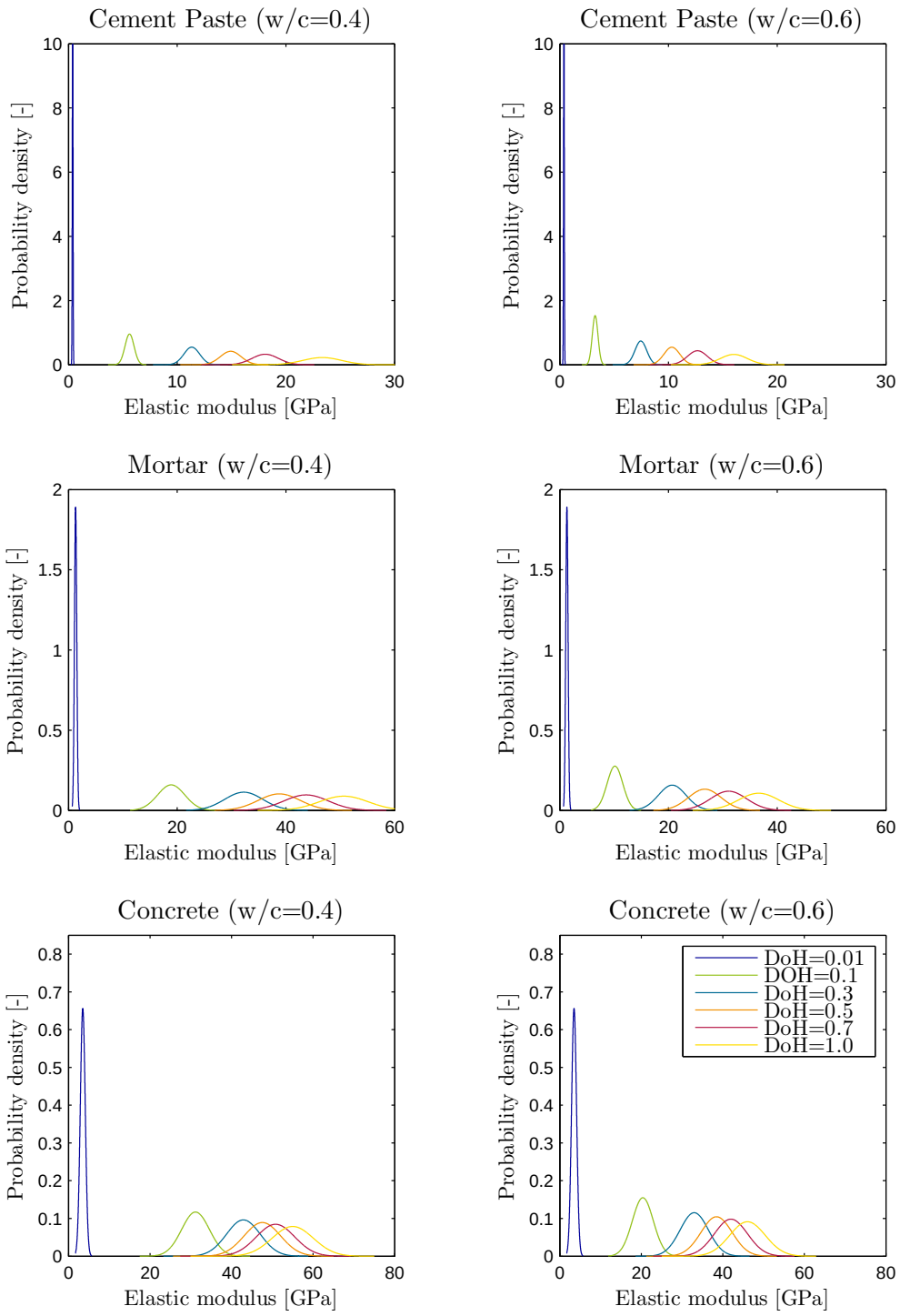



Figure 4: Probability densities of the predicted elastic modulus for the Powers model for different degrees of hydration (DoH) as well as for different w/c-ratios at the cement paste scale, the mortar scale, and the concrete scale.

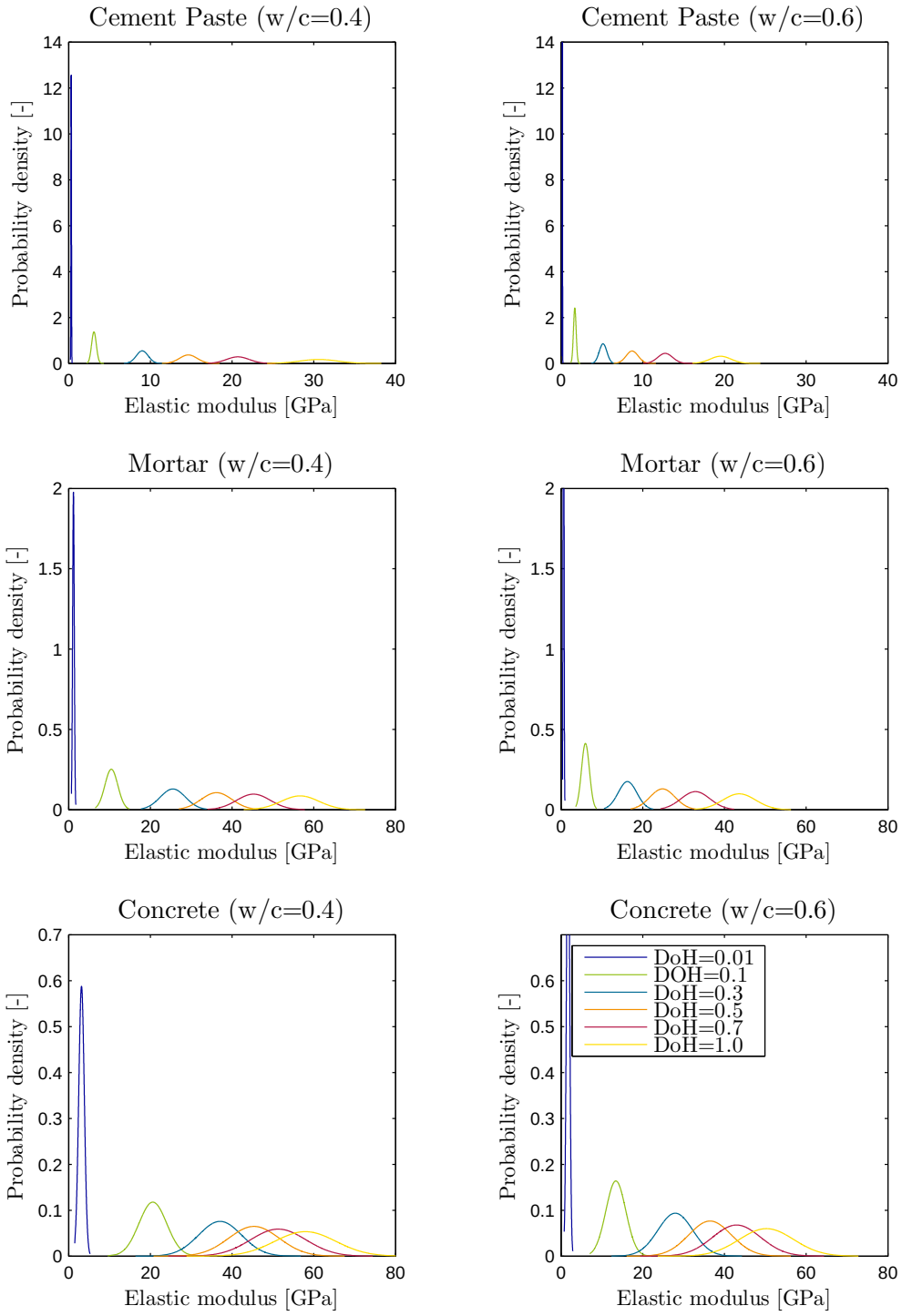



Figure 5: Probability densities of the predicted elastic modulus for the Bernard model for different degrees of hydration (DoH) as well as for different w/c-ratios at the cement paste scale, the mortar scale, and the concrete scale.

In Figure 4, the resulting probability densities of Powers' model are presented. The probability densities are shown separately for the three respective scales of observation. It is observed that the probability densities become lower during the upscaling modeling process. Furthermore, the shapes of the curves at the concrete scale are broader than at the cement paste scale. This is closely related to the evolution of the CoV as described before. Large CoV come along with broad proability density functions; both express a larger variability of the results. By showing the PDF, the differences between the results of the two different w/c-ratios are more visible than in the evolution of the CoV. The PDF shows slightly broader curves in case of a water-to-cement ratio of 0.4. One reason might be that in case of w/c=0.4 more unhydrated cement clinker particles remain in the material and thus contribute to the variation of the model results. For a water-to-cement ratio of 0.6 almost no clinker particles are left in the material. Hence, only the uncertainties of the hydration products influence the response variations.
The same conclusions can be drawn for the Bernard model. Compared to the distributions of the Powers model, the distributions of the Bernard model are slightly sharper which is related to a lower scatter in the model results. .

### 4.2 Sensitivity Analysis

The results of the global sensitivity analysis are exemplary discussed for the model responses using the Powers model. They are represented in Figure 6. The sum of the sensitivity indices is close to 1. Thus, interaction effects between the input parameters are negligible. The numerical effort necessary for execution of the sensitivity analysis is much higher for the Bernard model than for the Powers model due to the higher number of input parameters.

At the cement paste scale, the intrinsic elastic properties of the low-density C-S-H phase are mainly important for the model output. It is obvious that the prediction of the Young's modulus at this scale is closely related to this mechanical value since this phase represents the main component of cement paste. During the hydration process, Portlandite is formed which is linked to an increase of its sensitivity index. At the mortar scale, the volume fraction of sand is the most influencing factor whereat its influence on the model output decrease with ongoing hydration and the accompanied production of hydrates. At the concrete scale, the volume fraction of the coarse aggregates is determining at early-age hydration stages and the Young's modulus of the aggregates influences the model results in the later stages. Compared to the results for a w/c-ratio of 0.4, the results for w/c=0.6 are lower. This might be due to the fact that at the lower w/c-ratios there are some unhydrated clinker particles which also contribute to the Young‘s modulus.
The elastic properties as well as the volume fractions of the minor components do not influence the model responses much. Their sensitivity indices are mostly lower than 0.2.

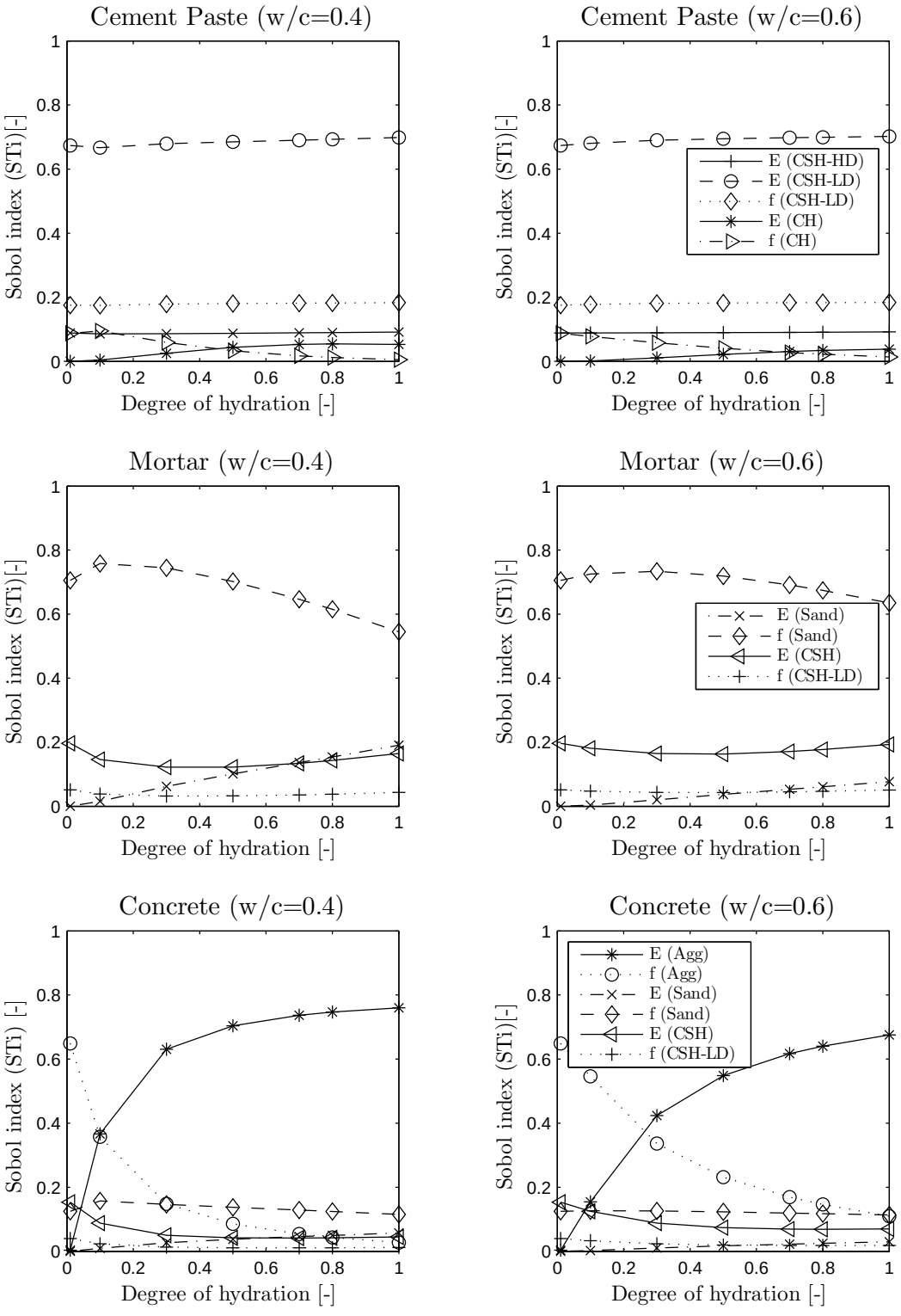



Figure 6: Evolution of total-order sensitivity indices of the Powers model as functions of the degree of hydration for two different w/c-ratios and at three scales of observation (*f – Volume fraction, E – Young's modulus, CSH-HD – C-S-H high density, CSH-LD – C-S-H low density, CH – Portlandite, Agg – Aggregate*).

## 5 CONCLUSION

A probabilistic approach was applied on a continuum-micromechanics based multiscale model to investigate the uncertainty propagation of input parameters across the length scales of concrete. The variability of the model responses was predicted and the parameters with the highest influence on the model outputs were identified. A probabilistic multiscale analysis through several scales of observation provides more insight on the nature of the upscaled elastic properties of cementitious materials than a deterministic analysis. Probability distributions of the model outputs and the intervals of the likelihood of the mechanical values were shown to be some of the possibilities of probabilistic model analyses.

By means of the sensitivity analysis, the significant influences of the elastic parameters at the cement paste scale and of the volume fractions at the larger scales were discussed. Furthermore, the increase of the model output uncertainties during the upscaling process as a result of the propagation of uncertainties in the input parameters was presented. In order to reduce the uncertainty of the upscaled mechanical properties of cement-based materials, the input parameters having a high sensitivity towards the Young's modulus should be determined precisely prior to the modeling process.

Comparison between the two well-known hydration kinetic models from Powers-Acker and Bernard revealed that the uncertainties and variations of both model responses are almost in the same range. The Bernard model only shows slightly higher uncertainties which seems to be surprising when the higher number of parameters in the Bernard model are considered.

This paper presented the application of a probabilistic assessment methodology on multiscale models. A further research topic could be the assessment of the accuracy of the two hydration kinetics models. It is worth to investigate whether the higher complexity of the Bernard model with its higher number of input parameter leads to an improved model prediction compared to the Powers model.

## 6 ACKNOWLEDGEMENTS

The research of the first author is supported by the German Research Foundation (DFG) via the Research Training Group "Evaluation of Coupled Numerical and Experimental Partial Models in Structural Engineering (GRK 1462)", which is gratefully acknowledged.

## REFERENCES

[1] J. Sanahuja, L. Dormieux, and G. Chanvillard. Modelling elasticity of a hydrating cement paste. *Cement and Concrete Research*, 37(10):1427–1439, 2007.

[2] B. Pichler and C. Hellmich. Upscaling quasi-brittle strength of cement paste and mortar: A multi-scale engineering mechanics model. *Cement and Concrete Research*, 41:467–476, 2011.

[3] Chr. Pichler and R. Lackner. A multiscale creep model as basis for simulation of early-age concrete behavior. *Computers and Concrete*, 5(4):295–328, 2008.

[4] T. C. Powers and T. L. Brownyard. Studies of the physical properties of hardened portland cement paste. *Research Laboratories of the Portland Cement Association Bulletin*, 22:101–992, 1948.

[5] P. Acker. Micromechanical analysis of creep and shrinkage mechanisms. In Franz-Josef Ulm, Z. P. Bažant, and F. H. Wittmann, editors, *Creep, Shrinkage and Durability Mechanics of Concrete and Other Quasi-brittle Materials: Proceedings of the Sixth International Conference CONCREEP-6@MIT*, pages 15–26. Elsevier, 2001.

[6] O. Bernard, F.-J. Ulm, and E. Lemarchand. A multiscale micromechanics-hydration model for the early-age elastic properties of cement-based materials. *Cement and Concrete Research*, 33(9):1293–1309, 2003.

[7] Chr. Pichler, R. Lackner, and H. A. Mang. A multiscale micromechanics model for the autogenous-shrinkage deformation of early-age cement-based materials. *Engineering Fracture Mechanics*, 74(1-2):34–58, 2007.

[8] A. Clement, C. Soize, and J. Yvonnet. Uncertainty quantification in computational stochastic multiscale analysis of nonlinear elastic materials. *Computational Methods in Applied Mechanics and Engineering*, 254:61–82, 2013.

[9] N. Vu-Bac, R. Rafiee, X. Zhuang, T. Lahmer, and T. Rabczuk. Uncertainty quantification for multiscale modeling of polymer nanocomposites with correlated parameters. *Composites: Part B*, 68:446–464, 2015.

[10] M. Berveiller, Y. Le Pape, Julien Sanahuja, and A. Giorla. Sensitivity analysis and uncertainty propagation in multiscale modeling of concrete. In *Poro-Mechanics IV 2009*.

[11] N. Venkovic, L. Sorelli, B. Sudret, T. Yalamas, and R. Gagné. Uncertainty propagation of a multiscale poromechanics-hydration model for poroelastic properties of cement paste at early-age. *Probabilistic Engineering Mechanics*, 32:5–20, 2013.

[12] A. Zaoui. Continuum mechanics: A survey. *Journal of Engineering Mechanics*, 128(8):808–816, 2002.

[13] J. D. Eshelby. The determination of the elastic field of an ellipsoidal inclusion, and related problems. *The Royal Society*, 241(1226), 1957.

[14] E. Kröner. Bounds for the effective elastic moduli of disordered materials. *Journal of Mechanics and Physics of Solids*, 24:137–155, 1977.

[15] A. V. Hershey. The elasticity of an isotropic aggregate of anisotropic cubic crystals. *Journal of Applied Mechanics*, 21:236, 1954.

[16] R. Hill. Elastic properties of reinforced solids. *Journal of the Mechanics and Physics of Solids*, 11(5):357–372, 1963.

[17] G. Constantinides and F.-J. Ulm. The effect of two types of c-s-h on the elasticity of cement-based materials: Results from nanoindentation and micromechanical modeling. *Cement and Concrete Research*, 34(1):67–80, 2004.

[18] B. Pichler, S. Scheiner, and C. Hellmich. From micron-sized needle-shaped hydrates to meter-sized shotcrete tunnel shells: micromechanical upscaling of stiffness and strength of hydrating shotcrete. *Acta Geotechnica*, 3(4):273–294, 2008.

[19] V. Smilauer and Z. Bittnar. Microstructure-based micromechanical prediction of elastic properties in hydrating cement paste. *Cement and Concrete Research*, 36(9):1708–1718, 2006.

[20] W. R. Drugan and J. R. Willis. A micromechanics-based nonlocal constitutive equation and estimates of representative volume element size for elastic composites. *Journal of the Mechanics and Physics of Solids*, 44(4), 1996.

[21] P. D. Tennis and H. M. Jennings. A model for two types of calcium silicate hydrate in the microstructure of portland cement paste. *Cement and Concrete Research*, 36:855–863, 2000.

[22] T. Mori and K. Tanaka. Average stress in matrix and average elastic energy of materials with misfitting inclusions. *Acta Metallurgica*, 21(5):571–574, 1973.

[23] Y. Benveniste. A new approach to the application of mori-tanakas theory in composite materials. *Mechanics of Materials*, 6(2):147–157, 1987.

[24] A. Saltelli, M. Ratto, T. Andres, F. Campologno, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global Sensitivity Analysis: The Primer*. John Wiley & Sons, West Sussex, England, 2008.

[25] I. M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment*, 1:407–414, 1993.

[26] C. Xu and G. Z. Gertner. Uncertainty and sensitivity analysis for models with correlated parameter. *Reliability Engineering and System Safety*, 93(10):1563–1573, 2008.

[27] M. D. McKay, R. J. Beckham, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis from a computer code. *Technometrics*, 21:239–245, 1979.

[28] P. Acker. Swelling, shrinkage and creep: a mechanical approach to cement hydration. *Materials and Structures / Concrete Science and Engineering*, 37:237–243, 2004.

[29] K. Velez, S. Maximilien, D. Damidot, G. Fantozzi, and F. Sorrentino. Determination by nanoindentation of elastic modulus and hardness of pure constituents of portland cement clinker. *Cement and Concrete Research*, 31:555–561, 2001.

[30] M. Choy, W. Cook, R. Hearmon, J. Jaffe, J. Jerphagon, and S. Kurtz. *Landolt-Bornstein: numerical data and functional relationships in science and technology. Group III: Crystal and solid state physics: Elastic, piezoelectric, pyroelectric, piezoptic, electroptic constants and nonlinear dielectric susceptibilities of crystals*, volume 11. Springer-Verlag, Berlin and Heidelberg, 1979.

# A APPENDIX

## A.1 Model input parameters and their stochastic properties for the Powers model

Table 1: Elastic parameters (** - assumptions).

| Parameter | Mean | SD | PDF | Reference |
|---|---|---|---|---|
| Elastic parameters - $E$ [GPa] | | | | |
| $C_3S$ | 135 | 7.0 | logn | ACKER [28], VELEZ [29] |
| $C_2S$ | 140 | 20.0 | logn | ACKER [28], VELEZ [29] |
| $C_3A$ | 145 | 10.0 | logn | ACKER [28], VELEZ [29] |
| $C_3S$ | 125 | 25.0 | logn | ACKER [28], VELEZ [29] |
| Gypsum | 45.7 | 4.6 | logn | CHOY et al. [30] |
| Portlandite | 38 | 5.0 | logn | CONSTANTINIDES and ULM [17] |
| C-S-H-LD | 21.7 | 2.2 | logn | CONSTANTINIDES and ULM [17] |
| C-S-H-HD | 29.4 | 2.4 | logn | CONSTANTINIDES and ULM[17] |
| Sand | 60 | 15 | logn | ** |
| Aggregate | 65 | 20 | logn | ** |

Table 2: Quantitative phase compositions of the reactants (** - assumptions).

| Parameter | Mean | PDF | Reference |
|---|---|---|---|
| $C_3S$ | 0.622 | u (0.568, 0.676) | VENKOVIC [11] |
| $C_2S$ | 0.152 | u (0.126, 0.178) | VENKOVIC [11] |
| $C_3A$ | 0.106 | u (0.097, 0.115) | VENKOVIC [11] |
| $C_4AF$ | 0.009 | u (0.008, 0.010) | VENKOVIC [11] |
| C-S-H HD | 0.7 | u (0.55, 0.85) | ** |

Table 3: Volume fractions (** - assumptions).

| Parameter | Mean | PDF | Reference |
|---|---|---|---|
| Sand content | 3 u (2, 4) | ** | |
| Aggregate content | 5 u (4, 6) | ** | |

## A.2 Model input parameters and their stochastic properties for the Bernard model

Table 4: Quantitative phase compositions of the reactants.

| Parameter | Mean | PDF | Reference |
|---|---|---|---|
| $C_3S$ | 0.622 | u (0.568, 0.676) | VENKOVIC [11] |
| $C_2S$ | 0.152 | u (0.126, 0.178) | VENKOVIC [11] |
| $C_3A$ | 0.106 | u (0.097, 0.115) | VENKOVIC [11] |
| $C_4AF$ | 0.009 | u (0.008, 0.010) | VENKOVIC [11] |

Table 5: Elastic parameters.

| Parameter | Mean | SD | PDF | Reference |
|---|---|---|---|---|
| Elastic parameters - $E$ [GPa] | | | | |
| $C_3S$ | 135 | 7.0 | logn | ACKER [28], VELEZ [29] |
| $C_2S$ | 140 | 20.0 | logn | ACKER [28], VELEZ [29] |
| $C_3A$ | 145 | 10.0 | logn | ACKER [28], VELEZ [29] |
| $C_3S$ | 125 | 25.0 | logn | ACKER [28], VELEZ [29] |
| Gypsum | 45.7 | 4.6 | logn | CHOY et al. [30] |
| Portlandite | 38 | 5.0 | logn | CONSTANTINIDES and ULM [17] |
| C-S-H-LD | 21.7 | 2.2 | logn | CONSTANTINIDES and ULM [17] |
| C-S-H-HD | 29.4 | 2.4 | logn | CONSTANTINIDES and ULM[17] |
| Sand | 60 | 15 | logn | ** |
| Aggregate | 65 | 20 | logn | ** |
| Elastic parameters - $\nu$ [-] (det* - determinstic) | | | | |
| $C_3S$ | 0.3 | | det* | ACKER [28], VELEZ [29] |
| $C_2S$ | 0.3 | | det* | ACKER [28], VELEZ [29] |
| $C_3A$ | 0.3 | | det* | ACKER [28], VELEZ [29] |
| $C_3S$ | 0.3 | | det* | ACKER [28], VELEZ [29] |
| Gypsum | 0.33 | | det* | CHOY et al. [30] |
| Portlandite | 0.305 | | det* | CONSTANTINIDES and ULM [17] |
| C-S-H-LD | 0.24 | | det* | CONSTANTINIDES and ULM [17] |
| C-S-H-HD | 0.24 | | det* | CONSTANTINIDES and ULM [17] |

Table 6: Activation energies.

| Parameter | Mean | PDF | Reference |
|---|---|---|---|
| Activation energies – $E_{aX}/R$ [K] | | | |
| $C_3S$ | 4500 | u (4110, 4890) | BERNARD [6] |
| $C_2S$ | 2500 | u (2285, 2715) | BERNARD [6] |
| $C_3A$ | 5500 | u (5025, 5975) | BERNARD [6] |
| $C_4AF$ | 4200 | u (3835, 4565) | BERNARD [6] |

Table 7: Kinetic Parameters.

| Parameter | w/c-ratio | Mean | SD | PDF | Reference |
|---|---|---|---|---|---|
| Kinetic parameters: Reaction order – $\kappa_X$ [-] | | | | | |
| $C_3S$ | 0.4 | 1.79 | 0.18 | logn | BERNARD [6] |
| | 0.5 | 1.72 | 0.17 | logn | BERNARD [6] |
| | 0.6 | 1.69 | 0.17 | logn | BERNARD [6] |
| $C_2S$ | 0.4 | 1.03 | 0.10 | logn | BERNARD [6] |
| | 0.5 | 0.96 | 0.10 | logn | BERNARD [6] |
| | 0.6 | 0.90 | 0.10 | logn | BERNARD [6] |
| $C_3A$ | 0.4 | 1.07 | 0.11 | logn | BERNARD [6] |
| | 0.5 | 1.00 | 0.10 | logn | BERNARD [6] |
| | 0.6 | 0.93 | 0.09 | logn | BERNARD [6] |
| $C_4AF$ | 0.4 | 2.37 | 0.24 | logn | BERNARD [6] |
| | 0.5 | 2.30 | 0.23 | logn | BERNARD [6] |
| | 0.6 | 2.23 | 0.22 | logn | BERNARD [6] |
| Kinetic parameters: Diffusion coefficient – $D_X$ [-] | | | | | |
| $C_3S$ | 0.4 | $1.05 \cdot 10^{-10}$ | $0.13 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.5 | $1.64 \cdot 10^{-10}$ | $0.32 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.6 | $6.42 \cdot 10^{-10}$ | $0.77 \cdot 10^{-10}$ | logn | BERNARD [6] |
| $C_2S$ | 0.4 | $6.64 \cdot 10^{-10}$ | $0.80 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.5 | $6.64 \cdot 10^{-10}$ | $0.80 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.6 | $6.64 \cdot 10^{-10}$ | $0.80 \cdot 10^{-10}$ | logn | BERNARD [6] |
| $C_3A$ | 0.4 | $2.64 \cdot 10^{-10}$ | $0.32 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.5 | $2.64 \cdot 10^{-10}$ | $0.32 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.6 | $2.64 \cdot 10^{-10}$ | $0.32 \cdot 10^{-10}$ | logn | BERNARD [6] |
| $C_4AF$ | 0.4 | $1.05 \cdot 10^{-10}$ | $0.13 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.5 | $2.64 \cdot 10^{-10}$ | $0.32 \cdot 10^{-10}$ | logn | BERNARD [6] |
| | 0.6 | $6.42 \cdot 10^{-10}$ | $0.77 \cdot 10^{-10}$ | logn | BERNARD [6] |

Table 8: Volume fractions (** - assumptions).

| Parameter | Mean | PDF | Reference |
|---|---|---|---|
| Sand content | 3 | u (2, 4) | ** |
| Aggregate content | 5 | u (4, 6) | ** |

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SPATIAL CLUSTERING STRATEGIES FOR HIERARCHICAL MULTI-SCALE MODELLING OF METAL PLASTICITY

**Md Khairullah[1], Jerzy Gawad[1], Dirk Roose[1], and Albert Van Bael[2]**

[1]Department of Computer Science, KU Leuven
3001 Leuven, Belgium
e-mail: {md.khairullah, jerzy.gawad, dirk.roose}@cs.kuleuven.be

[2]Department of Materials Engineering, KU Leuven
3001 Leuven, Belgium
e-mail: albert.vanbael@kuleuven.be

**Keywords:** Multi-scale modeling, Polycrystalline materials, Spatial clustering, Dynamic adaptive clustering, Clustering criteria, Approximation error.

**Abstract.** *In multi-scale simulations of material forming processes, macroscopic zones of nearly homogeneous strain response occur. In such zones the evolution of plastic anisotropy at each finite element integration point can be approximated from the properties at a representative point. We show how these zones can be identified by a clustering algorithm and can be utilized to reduce the computational cost of the simulation.*

## 1 INTRODUCTION

The Hierarchical Multi-Scale (HMS) software, developed at KU Leuven, simulates the deformation of polycrystalline metallic alloys. HMS takes into account both the strain-driven evolution of the preferred orientation of crystallites (or texture) and the associated plastic anisotropy, as they significantly influence the mechanical and physical properties of the material [1]. At the macroscopic level, the deformation of the material is described by a Finite Element (FE) model and the anisotropy of the plastic properties is approximated by an analytical plastic potential function. On the other hand, the physics-based crystalline plasticity (CP) code provides the micro-scale stress response. The HMS incorporates two categories of material state: a) the direct material state (e.g. equivalent plastic strain, equivalent yield stress) that is directly needed by the FE code in each time step, and b) the extended material state that has to be updated regularly, but does not have to be accessed in each time step. The parameters of the plastic potential function, together with the texture data (represented by the orientation distribution function [2]) form the extended state variables at the FE integration points. The HMS model assumes that the texture and the coupled plastic anisotropy are initially identical in the whole volume of the material, but may evolve independently in every FE integration point with increasing plastic strain.

In principle, the texture can change at every time increment of the macroscopic FE model, and the parameters of the plastic potential function should be updated at every time step, resulting in a substantial computational cost. The HMS software partly resolves this issue by reconstructing the function not after every time increment, but only if a given deformation-based criterion is satisfied. Hence, the texture state variable remains constant in each time interval between the updating events. In order to decide whether such an update of the texture state variable should take place at a specific integration point, the HMS software tracks the recent history of the deformation tensor $\mathbf{d}(t) \in \mathbb{R}^{3\times3}$ in that point by

$$\mathbf{P}^t = \int_{t_i}^{t} \mathbf{d}(t)dt, \tag{1}$$

which is the plastic strain that has been accumulated (at time $t$) since the previous update (at time $t_i$). The texture related material properties are updated if

$$\|\mathbf{P}^t\| \geq P_{cr}, \tag{2}$$

with $P_{cr}$ is a user-defined threshold. Then, the accumulated plastic strain tensor $\mathbf{P}^t$ is passed to the texture evolution model, which applies appropriate lattice rotations to the crystal orientations to obtain the new orientations [3].

Similarly, reduction of update operations in the spatial domain is also an option. Instead of updating the texture related material properties at each integration point inside a group of points, updating a single representative point is more feasible. In this paper, we present methods to reduce the simulation time of the HMS model by using spatial clustering of the finite element integration points w.r.t. some relevant feature of interest.

The paper is organized as follows. Section 2 presents the principles of the proposed enhancements to the existing hierarchical multi-scale software as well as the description of the test case used for the simulation experiments. Section 3 outlines estimations of the approximation error and the performance gain due to the enhancements. A detailed analysis and discussions of the simulation results are presented in Section 4. The last section contains a summary of the ongoing works, concluding remarks and future research.

## 2 SPATIAL CLUSTERING IN THE HMS SOFTWARE

Clustering is a technique to group data objects, based on information found in the data characterizing the objects and their relationships. The aim is that the objects within a group are similar to one another and different from the objects in other groups. The greater the similarity within a group and the greater the difference between groups, the better the clustering. Spatial clustering organizes objects based on spatial aspects such as distance, connectivity, relative density in space as well as other feature(s) of interest [4]. Due to its generic applicability, clustering is used in a broad range of applications, e.g. in materials modeling, material discovery, general finite element simulations and other fields of engineering [5, 6, 7, 8, 9]. The fundamental assumption in this paper is that similar micro-structural state variables in neighboring integration points subjected to a similar deformation history would evolve along nearly identical trajectories in state space. It can be expected that the derived macroscopic plastic anisotropy would be similar as well. Therefore, we perform the actual update of the plastic anisotropy at a single representative integration point per cluster and propagate the updated properties to the other integration points belonging to the cluster. This significantly reduces the number of updates of the plastic anisotropy and subsequently the overall simulation time. Similar approaches to reduce computational cost of engineering applications based on clustering are reported in [8, 10, 11].

In this work we consistently employ the implementation of the agglomerative clustering algorithm provided by the scikit-learn library for machine learning [12]. This algorithm performs a hierarchical clustering using a bottom-up approach. Initially each object is a cluster on its own, and adjacent clusters are successively merged together. A variance-minimizing approach is used for the merge strategy. It merges the pair of clusters for which the sum of squared differences of the data points within the clusters is minimal. We developed an interface to this clustering algorithm, for which the inputs are a set of features of interest, information about the connectivity among the data points, and the number of clusters to construct. This interface computes the connectivity among the integration points. Note that clustering is carried out based on the data at the integration points and the connectivity among the integration points. This makes the proposed methods completely independent of chosen element/mesh types, which is an additional advantage.

However, the selection of input data, which represent the feature of interest, and the number of cluster is not trivial and are discussed below. We also discuss the cluster representative selection procedure.

**Feature of interest:** As stated in section 1, upon satisfaction of the deformation based update criteria, given in (2), the accumulated plastic strain $\mathbf{P}^t$ is passed to the texture evolution model. Thus, we assume that the accumulated plastic strain, i.e. the recent deformation history, determines the evolution of texture and the associated plastic anisotropy. However, as the updates at the integration points are independent and may occur out of sync, the accumulated plastic strain in a set of integration points can not be compared to measure the similarity in recent history. Rather, the total plastic strain (accumulated since the beginning) that represents the total plastic deformation history at an integration point, is a more logical feature to compare a set of integration points for our purpose. Furthermore, the plastic strain tensor having 9 components (6 independent components) may suffer from the limitations of clustering multidimensional data [13]. Dimensionality reduction (e.g. subspace clustering/ projected clustering) is the typical solution for this. Equivalently, state variable(s) with lower dimension which can correspond to the total plastic strain may be considered as an alternative feature for clustering. HMS incorpo-

rates the traditional approach of elastic predictor and plastic corrector in the elasto-plastic stress integration algorithm [14]. The plastic strain increment $\Delta\boldsymbol{\varepsilon}^p \in \mathbb{R}^5$ is assumed to be

$$\Delta\boldsymbol{\varepsilon}^p = \Delta\varepsilon^p \mathbf{A}, \tag{3}$$

where the scalar $\Delta\varepsilon^p$ represents the magnitude of the plastic strain increment, and $\mathbf{A}$ is the plastic strain rate mode that corresponds to the point on the yield locus in the direction of the trial stress [15]. The increment in equivalent plastic strain $\Delta\bar{\varepsilon} \in \mathbb{R}$ is calculated as

$$\Delta\bar{\varepsilon} = \Delta\varepsilon^p \psi^t(\mathbf{A}), \tag{4}$$

where $\psi^t$ is the plastic potential function at time $t$. If $\boldsymbol{\varepsilon}^{p,t}$ and $\bar{\varepsilon}^t$ respectively are the plastic strain and the equivalent plastic strain at time $t$, at the end of time increment $\Delta t$ they are respectively given by

$$\boldsymbol{\varepsilon}^{p,(t+\Delta t)} = \boldsymbol{\varepsilon}^{p,t} + \Delta\boldsymbol{\varepsilon}^p, \tag{5}$$

and

$$\bar{\varepsilon}^{t+\Delta t} = \bar{\varepsilon}^t + \Delta\bar{\varepsilon}. \tag{6}$$

As the calculation of the increment in plastic strain $\Delta\boldsymbol{\varepsilon}^p$ and in the equivalent plastic strain $\Delta\bar{\varepsilon}$ depends on the set of same variables $\Delta\varepsilon^p$, $\mathbf{A}$, the equivalent plastic strain, which is a scalar state variable at an integration point, can be considered as a scalar representation of the plastic strain, and also a feature of interest for clustering.

**Number of clusters:** Most of the clustering algorithm constructs some prescribed number of clusters instead of finding the optimal number of clusters. This limitation remains a fundamental and largely unsolved problem in cluster analysis. A simple rule of thumb suggests to construct $\sqrt{n/2}$ clusters out of $n$ data points. The Elbow method recommends to choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. However, for several cases the stopping criteria can not be unambiguously evaluated. A number of similar but more concrete approaches, primarily based on cluster evaluation and assessment, have been proposed (e.g. [16, 17]) and reliably incorporated in popular software packages. Computational efficiency of these algorithms is poor as they construct a number of clusters, measure an evaluation metric (e.g. silhouette coefficient), and repeat with an incremented number of clusters. However, if the construction of clusters in the simulation is infrequent, one of these methods can be utilized to decide about the optimal number of clusters. As the cluster evaluation metric itself is computationally costly, in situations where we may frequently need to apply the clustering algorithm, we can utilize some simple heuristics instead. For example, we can measure the distance between the minimum and the maximum of the data points and decide to construct two clusters if the distance is larger than a threshold value. This can be recursively continued inside the generated clusters. However, the clusters generated in this way would lead to sub-optimal clusters, and still we need to choose a good threshold value.

**Cluster representative:** The selection of the cluster representative is quite intuitive. The mean (or, the centroid for multi dimensional data) is commonly considered as the most appropriate representation of the cluster data points, as this value minimizes the distance between the data points and the representative. However, we must select an integration point inside the cluster to represent the cluster for practical purposes (e.g. in our case the cluster representative would contain a number of extended state variables, which are common for all the members). We choose the integration point whose average dissimilarity to all the integration points in the cluster is minimal, known as the medoid, to serve this purpose in our case.

For a brief explanation of the proposed strategy, let us consider the two dimensional miniature model in Figure 1 with only 7 elements. Let each element has only one integration point located at the center. Integration points in each element are labeled with *A*–*G*. The borders of the clusters are marked by dashed lines, and the cluster representatives are encircled. Notice that although the points *C* and *E* show exactly the same field values, they are put in different clusters since point *D* with large difference in the field value is located between them, which violates the connectivity constraint for forming a spatial cluster. In the proposed method only the cluster representatives *B*, *D*, *E*, and *G* update the plastic anisotropy. Representative point *B* propagates its updated properties to *A* and *C*, whereas *G* propagates to *F*. As *D* and *E* are the single member of their own clusters, they only update the plastic anisotropy but need no propagation.

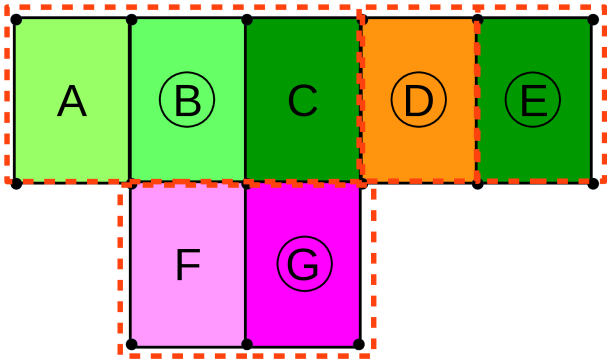



Figure 1: Subdivisions of the FE mesh into clusters with respect to the field value (color scale). The elements are labeled with letters, while the cluster representatives are distinguished by circles.

While the underlying micro-structure, texture and the associated plastic anisotropy are assumed to be the same for the whole cluster, other variables such as the strain and the corresponding stress at the integration points are independent of the clusters and evolve individually.

### 2.1 Selection of test case

In [18] we used a complex geometry specimen to simulate a tensile test. However, this test case can be considered as very simple in four aspects: (i) the homogeneity of plastic strain field across integration points at the beginning of plastic deformation is somehow maintained for the rest of the simulation, (ii) the magnitude of the accumulated plastic strain is sufficient as the clustering criterion because only one of the plastic strain tensor components is dominating, which is equivalent to the strain magnitude, (iii) in principle the specimen was two dimensional, and (iv) the model was comparatively small (2,226 integration points). For these reasons, a more complex deformation is considered in the test case of this paper.

We consider the specimen in Figure 2, which is used in the simulation of a two stage deformation. The FE mesh consists of 12,168 elements of type C3D8R (i.e. 8 node linear brick element with one integration point). The farthest Y-Z surface in the figure is always kept fixed. In the first stage a moment load is applied on the nearest Y-Z surface in the figure along the positive X direction and in the second stage tension is applied on the same surface in the same direction with small displacements along the positive Y and the negative Z directions. Figure 3a shows the equivalent plastic strain field at the integration points at some moment in the first stage during the simulation, where regions of almost homogeneous field values can be easily observed. If one groups the neighboring integration points with similar equivalent plastic strain

in a cluster, a single update is needed per cluster instead of updates at each integration point, based on the assumption of homogeneity of the underlying texture and the plastic anisotropy inside the cluster. The corresponding 9 clusters are represented by distinct colors in Figure 3b.

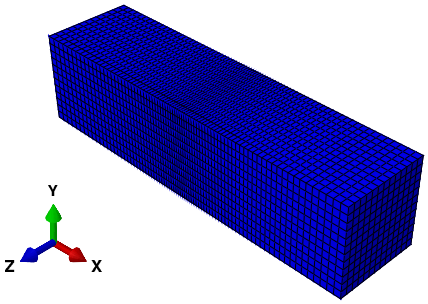
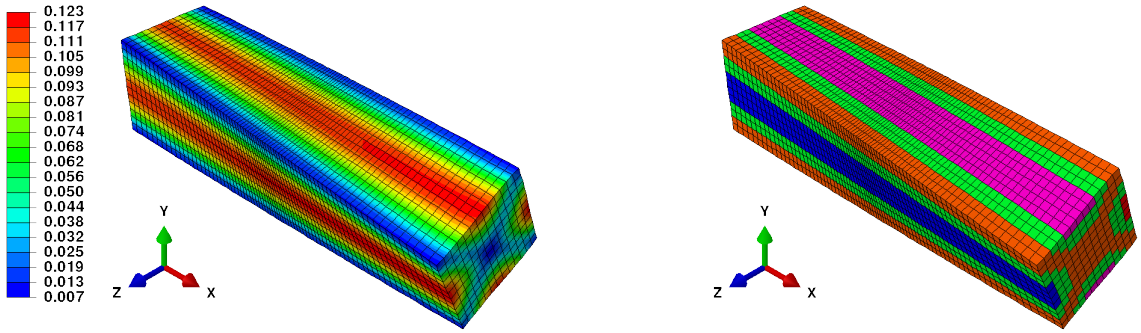



Figure 2: Specimen used in a two step deformation for the simulation experiments. The FE mesh consists of 12,168 C3D8R elements.




(a) (b)

Figure 3: a) The field of equivalent plastic strain in the specimen used in the two-stage deformation simulation; and b) clusters (represented by distinct colors) are constructed according to equivalent plastic strain in the highlighted area.

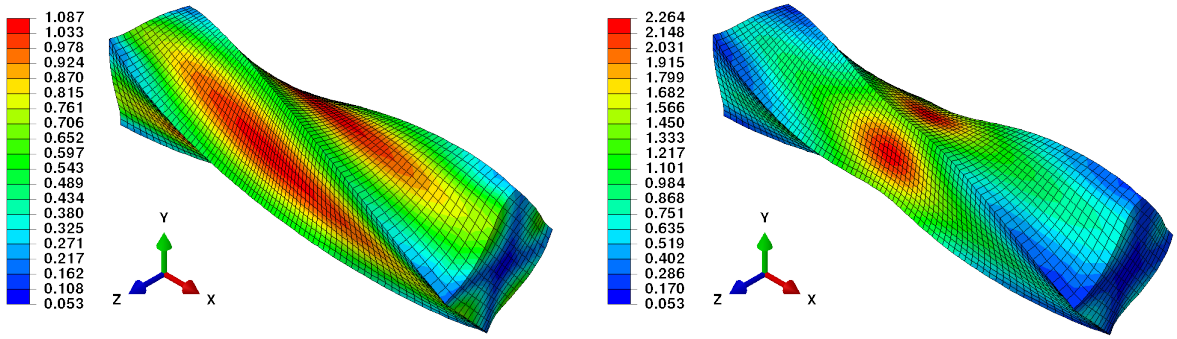
## 2.2 Static clustering

A static clustering strategy employs a fixed configuration of the clusters, which is constructed based on the field values at a particular moment in the simulation and subsequently kept constant. The advantage of this simple approach is that the overhead related to the construction of the clusters is minimal. However, the choice of the particular moment for constructing the clusters is crucial to obtain a good balance between the number of clusters (and thus the computational cost of the simulation) and the accuracy.

We can perform a number of time steps in the actual HMS simulation without clustering, and then construct the clusters based on the field of interest. If the clustering is performed in an early time step, it may be that insufficient deformation history is available, leading to suboptimal clusters. However, postponing the construction of the clusters leads to less savings in the computational cost. Therefore, we suggest to construct the clusters in the time step in which the update criterion is satisfied for the first time.

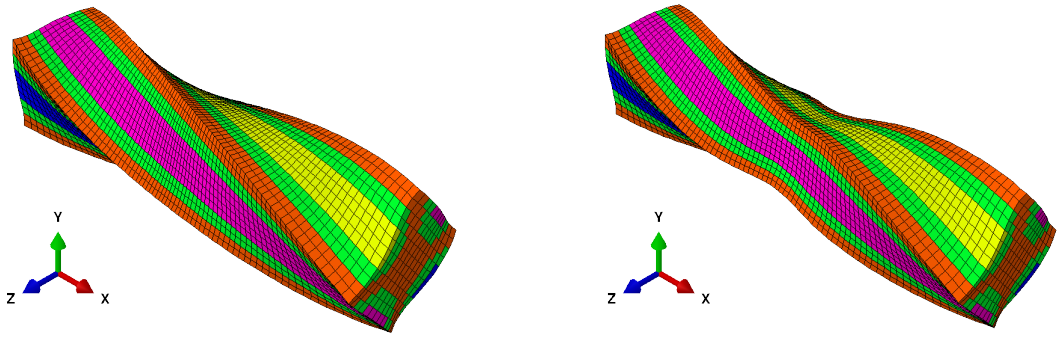
Figure 4a and 4b show the equivalent plastic strain field values at the middle and at the end of the second deformation stage respectively. These clearly indicate that the clusters constructed

by the static clustering approach do not reflect the homogeneous zones in the second stage. This is particularly true if we use a small number of clusters. For example, we can compare the nine clusters in Figure 5a and Figure 5b with the field values in Figure 4a and 4b. The incapability of the static clustering to reflect the variation in time of homogeneity of field values among the integration points is a fundamental drawback of it.




(a) (b)

Figure 4: The field of equivalent plastic strain at a) the middle, and b) at the end of the second deformation stage.




(a) (b)

Figure 5: Statically constructed clusters at a) the middle, and b) at the end of the second deformation stage.

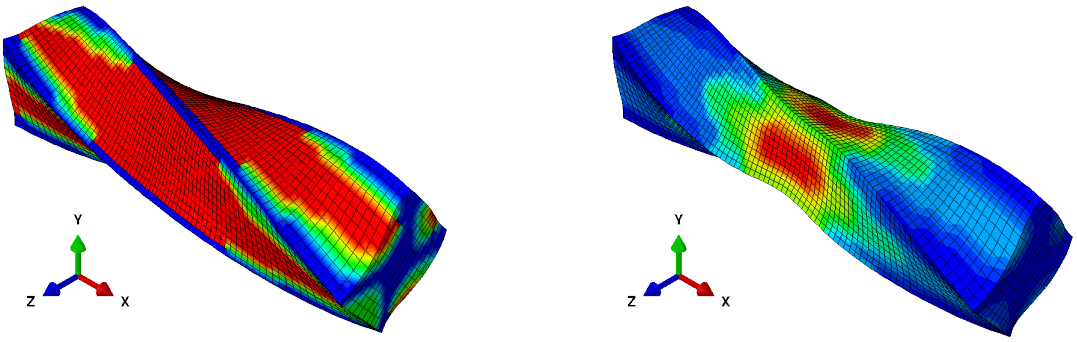
### 2.3 Dynamic adaptive clustering

To overcome the limitations of static clustering, a dynamic adaptive clustering is proposed. In this strategy the integration points are re-assigned to the clusters using criteria based on minimization of the variance w.r.t. the equivalent plastic strain among the cluster members. This dynamic approach is more realistic, since it is able to capture the evolution of strain and, more importantly, it does not rely on a single time step to determine the clusters. Hence, compared with static clustering, we expect improvements in the accuracy and, at the same time, performance gains, in particular if relatively large clusters can be retained for a longer time, leading to fewer updates of the plastic anisotropy.

At the beginning of the simulation, the amount of the plastic strain as well as the equivalent plastic strain is zero at each integration point and we assign all points to a single cluster. A

cluster is split into two clusters if the difference between the maximum and the minimum of the observed field values among the points inside the cluster exceeds a specified threshold value. Of course, this splitting criterion is only tested when the deformation-based updating criteria at an integration point is satisfied. As long as a cluster is not split, the cluster representative prevails the role. At present, for simplicity the adaptation of clustering is based on splitting only without the possibility of merging two clusters.

Clusters constructed with the dynamic adaptive approach are shown in Figure 6a and 6b. Obviously, these clusters better represent the actual homogeneity of field values in Figure 4a and 4b.




(a) (b)

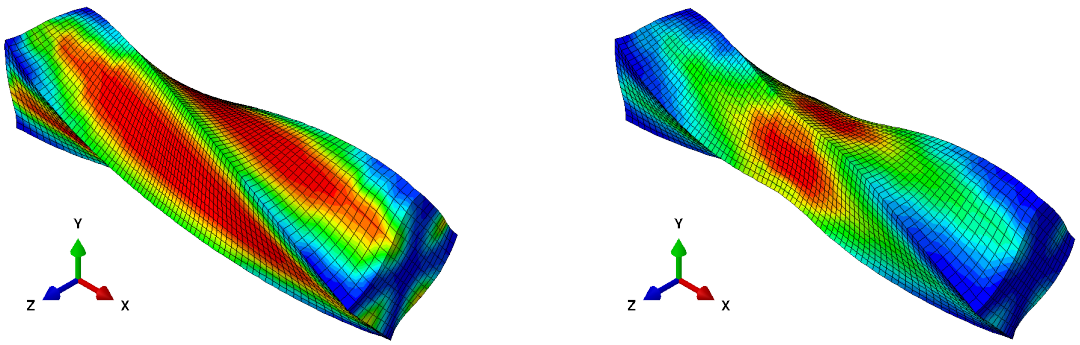
Figure 6: Adaptively constructed clusters at a) the middle, and b) at the end of the second deformation stage. In (a) and (b) the number of clusters are 3 and 62 respectively (only 24 colors are used).

A smaller threshold value for splitting (i.e. a more rigorous splitting condition) leads to an early split of the clusters and ultimately results in better clusters. Figure 7a and Figure 7b represent such a case. Comparison with the field values in Figure 4a and 4b makes it evident. However, such a small threshold value for splitting result in a large number of clusters, and consequently large number of updates and less computational gain.

Note that only 24 colors are used in Figure 6 and Figure 7 to visualize the clusters. In fact, formed clusters are assigned unique labels from 1 to $n_c$, where $n_c$ is the number of clusters (e.g. $n_c = 3$ in Figure 6a and $n_c = 62$ in Figure 6b). These labels are assigned from low to high in an sorted order of the mean data in each cluster, i.e. the cluster with the lowest mean data is assigned label 1 and the cluster with the highest mean data is assigned label $n_c$.

## 3 ERROR ESTIMATION AND COMPUTATIONAL PERFORMANCE

We approximate the anisotropy of plastic properties at each integration point in a cluster by the properties at the cluster representative. This introduces an additional modeling error. More specifically, the plastic anisotropy model is utilized uniformly inside the cluster. The coefficients of the anisotropy model are periodically re-identified to follow the evolution of anisotropy solely at the cluster representative. This approximation also affects the field values (e.g. equivalent plastic strain, equivalent yield stress) that are calculated based on the plastic anisotropy. Hence, we will estimate the approximation error in both the plastic properties and the affected field values. In the rest of the paper the HMS software that is extended with spatial clustering to reduce the computational cost is referred as the *improved HMS* and the original HMS software that does not exploit spatial clustering is referred as the *reference HMS*. Note that




(a) (b)

Figure 7: Adaptively constructed clusters with a smaller threshold value at a) the middle, and b) at the end of the second deformation stage. In (a) and (b) the number of clusters are 2,119 and 6,013 respectively (only 24 colors are used).

the reference HMS is equivalent to the improved HMS using one integration point per cluster. Using this maximal number of clusters results in no approximation error but also no reduction of the computational cost.

### 3.1 Approximation error in terms of plastic properties

The plastic potential function can be utilized to estimate the error in plastic properties regardless of the loading direction. We calculate the Euclidean distance of the plastic potential function values at an integration point, computed with the reference HMS software, and the plastic potential function values at the representative point, computed with the improved HMS software. For an integration point, *m* almost evenly spaced strain rate modes in the 5D strain rate space are chosen [19] and the value of the potential function is calculated for each of them. We define the relative error in plastic properties as the relative difference in the plastic potential function using the formulas

$$e_i^{\psi} = \frac{1}{m}\sum_{j=1}^{m} \left|\frac{\psi(\mathbf{D}_j)_i - \psi'(\mathbf{D}_j)_{cr(i)}}{\psi(\mathbf{D}_j)_i}\right| \times 100\% \tag{7}$$

and

$$\bar{e^{\psi}} = \frac{1}{n}\sum_{i=1}^{n} e_i^{\psi}, \tag{8}$$

where $\psi(\mathbf{D}_j)_i$ and $\psi'(\mathbf{D}_j)_{cr(i)}$ are the values of the plastic potential function at an integration point $i$ in the reference HMS and at the cluster representative $cr(i)$ in the improved HMS respectively, for a chosen strain rate mode $\mathbf{D}_j$. Hence, $e_i^{\psi}$ and $\bar{e^{\psi}}$ represent the relative error respectively in integration point $i$ and in the whole model. In principle we can compute these values at every time increment, but most often we are only interested in these errors at the end of the simulation.

### 3.2 Approximation error in terms of field values

The plastic strain during a time step is affected by the approximation of the plastic anisotropy using the clustering strategy. As the equivalent plastic strain field in the final state of the simulation contains the total plastic deformation history summarizing the total effect of the approximation, is selected for comparison with the reference HMS. In the improved HMS, this scalar

value is computed at each integration point using the plastic anisotropy computed at the cluster representative. Hence, we compute the relative approximation error in the equivalent plastic strain field at point *i* as

$$e_i^{\gamma} = \frac{|\gamma_i - \gamma_i'|}{\gamma_i} \times 100\%, \tag{9}$$

where $\gamma_i$ and $\gamma_i'$ are the equivalent plastic strain values at an integration point $i$ calculated with the reference HMS and the improved HMS respectively. As in the previous case, we also compute the average of the error for the whole model

$$\bar{e^{\gamma}} = \frac{1}{n} \sum_{i=1}^{n} e_i^{\gamma}. \tag{10}$$

Note that (9) and (10) can be utilized to calculate the approximation error for any other relevant scalar variables, for example the equivalent yield stress.

### 3.3 Computational gain

We measure the relative performance improvement in terms of the calculation time by *gain in time* $g_t$, defined as

$$g_t = \frac{t_r}{t_c}, \tag{11}$$

where $t_r$ is the simulation time with the reference HMS software and $t_c$ is the simulation time with the improved HMS.

By using the methods proposed above, we only speed up the part of the software responsible for the evolution of texture and the associated plastic anisotropy. If that part represents a fraction $f$ of the execution time, then the gain in time is limited by

$$g_t^{max} = \frac{1}{1-f}. \tag{12}$$

Hence, if $f \approx 94\%$, as in the simulations performed below, $g_t^{max} \approx 16$. The expected gain in time is

$$g_t^e = \frac{1}{(1-f) + \frac{n_c}{n} f} \leq g_t^{max}, \tag{13}$$

with $n_c$ the number of clusters used in the improved HMS software, cf. Amdahl's law [20]. Note that $g_t^e$ is an approximation of $g_t$, for two reasons. Primarily, the computational cost of all the integration points is assumed to be equal, which is not true in reality. In fact, different integration points in the model undergo a different number of updates based on the position of the point in the model and the applied loads and boundary conditions. For example, the undeformed integration points need no update but the integration points experiencing the largest deformation require the most updates. Additionally, there are some overheads to implement spatial clustering in HMS.

For the dynamic adaptive clustering approach, the gain in time may be additionally affected by the overhead due to construction and maintenance of the clustering information and may not reflect the actual savings in update operations. For this reason, we can measure the computational benefit of the proposed methods by the reduction in update operations. We define *gain in updates* $g_u$ to be

$$g_u = \frac{u_r}{u_c}, \tag{14}$$

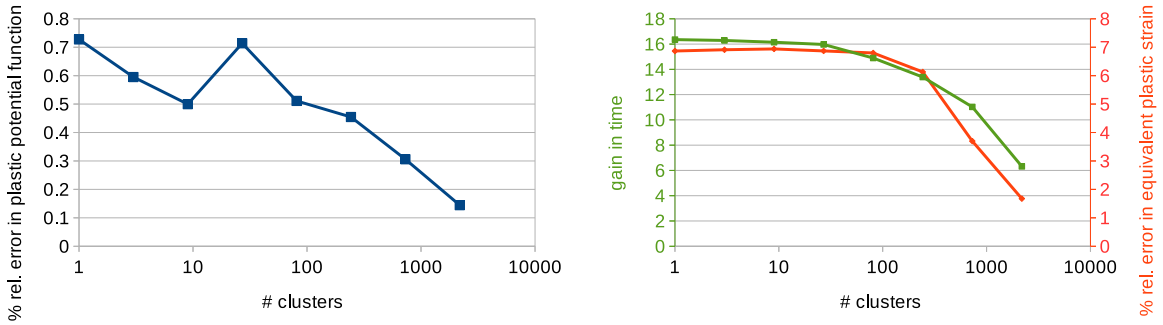
where $u_r$ is the number of updates in the reference HMS and $u_c$ is the number of updates in the HMS, which uses clustering.

## 4 RESULTS AND DISCUSSIONS

The improved HMS software is used to simulate the new test case described in 2.1. We compare the accuracy and the computational gain in the improved HMS and in the reference HMS. In all simulations the texture and plastic anisotropy at an integration point are updated if the norm of the accumulated plastic strain exceeds 0.05, i.e. $P_{cr} = 0.05$ in (2). The calculations were carried out on a twenty-core Intel Xeon compute node (2.8 GHz processor). Multiple cores were used for calculating the parameters of the plastic potential function. The simulation time with the reference HMS is around 47 h, of which 3 h is attributed to the macroscopic FE computations by Abaqus/Explicit 6.14-1 and required approximately 206,000 time increments in total. The remaining 44 h were consumed by 90,502 updates of the texture and the anisotropic plastic properties.

### 4.1 Static clustering

As described in section 2.2, the clusters are formed at the first update of texture and plastic anisotropy to obtain a balance between savings in computational cost and accuracy of the simulation. Figure 8(a) shows the measured approximation error for the plastic anisotropy (plastic potential function) $\bar{e^{\psi}}$, given by (8), for varying number of clusters. We observe a steady decrease in the approximation error with increasing of number of clusters, except for number of clusters around 30. Note that the approximation error is measured at the end of the simulation. So, it is possible that the statically constructed clusters, formed at the first update of texture and plastic anisotropy, are not representative for the spatial variation in plastic properties during the rest of the simulation. Figure 8b presents a comparison between the approximation error in the affected equivalent plastic strain field $\bar{e^{\gamma}}$ given by (10), and the corresponding gain in time $g_t$, given by (11). We see that varying the number of clusters up to 100 has no significant effect on the approximation error as well as on the gain in time. Hence, the error in equivalent plastic strain and the error in plastic potential function evolve differently as a function of the number of clusters. In fact, the equivalent plastic strain has a non-linear relationship with the plastic potential function. Moreover, in (8) we considered almost all directions in the strain rate space, whereas the equivalent plastic strain in (10) is affected only by the direction of corresponding strain rate.




(a) (b)

Figure 8: (a) Approximation error in plastic anisotropy (plastic potential function) $\bar{e^{\psi}}$. (b) Approximation error in equivalent plastic strain $\bar{e^{\gamma}}$ and gain in time $g_t$ for varying number of clusters.

We measured that the evolution of texture and the associated plastic anisotropy consumes

94% of the simulation time (i.e. $f = 0.94$). Figure 9 compares the gain in time $g_t$ obtained by the improved HMS software with the expected gain in time $g_t^e$ for a varying number of clusters, see (13). We observe that the actual gain in time is higher than the expected gain. Possible causes are stated earlier.

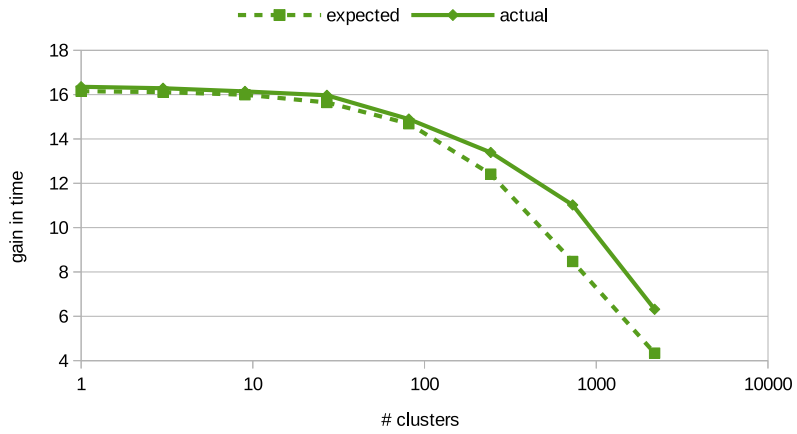



Figure 9: Actual gain in time $g_t$ compared to the expected gain in time $g_t^e$ for varying number of clusters.

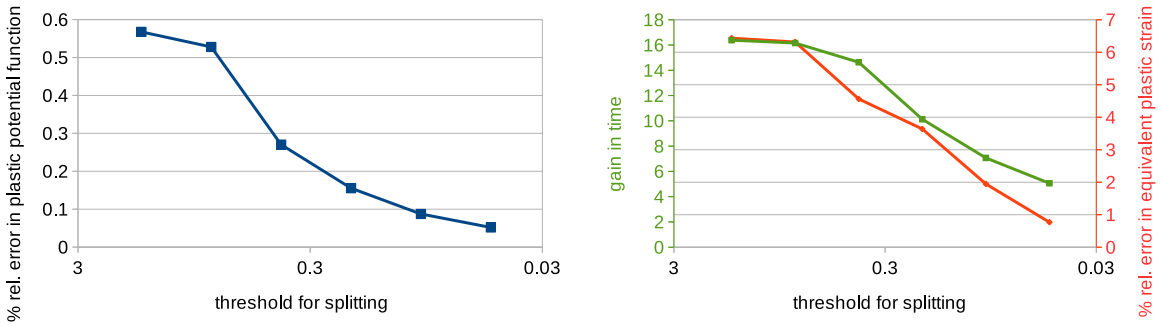
### 4.2 Dynamic adaptive clustering

A set of simulations with the same specimen and same settings were run using adaptive clustering. We mentioned above that a cluster is split if the difference between the maximum and the minimum of the equivalent plastic strain within a cluster exceeds the specified threshold value. If this threshold value increases, a cluster remains undivided for longer time, fewer clusters are constructed and the required number of update operations is reduced. Thus, a larger gain in time is achieved, whereby the approximation error increases. In Figure 10a we observe that the approximation error in plastic potential function $\bar{e^{\psi}}$ decreases rapidly with decreasing threshold values, except for very small threshold values where clustering reached the level of saturation. Figure 10b represents the effect of the adaptive clustering approach on the approximation error in the equivalent plastic strain $\bar{e^{\gamma}}$ and on the gain in time $g_t$ for varying threshold value for splitting. Again, the general tendency of the approximation error $\bar{e^{\gamma}}$ is decreasing with decreasing threshold value (increasing number of effective clusters), but there is not a linear correspondence between the approximation error in plastic potential function $\bar{e^{\psi}}$ and in the equivalent plastic strain $\bar{e^{\gamma}}$.

For both static and dynamic adaptive clustering, we are interested in the evolution of the approximation error in the equivalent plastic strain $\bar{e^{\gamma}}$ and the gain in time $g_t$ when the number of clusters increases. A clear relationship between approximation error and gain in computation time is observed in Figure 8b and Figure 10b. Moreover, the approximation error decreases faster than the gain in time, which is an important advantage.

### 4.3 Comparison between static and dynamic adaptive approach

Figure 11a and Figure 11b compare the clustering approaches w.r.t. obtained gain in time and accuracy. Clearly, the dynamic adaptive approach has a lower approximation error in plastic potential function $\bar{e^{\psi}}$, except when the approximation error is very high or very low. On the




(a) (b)

Figure 10: (a) Approximation error in plastic anisotropy (plastic potential function) $e^{\bar{\psi}}$, and (b) approximation error in equivalent plastic strain $e^{\bar{\gamma}}$ and gain in time $g_t$ for varying threshold value for splitting a cluster.

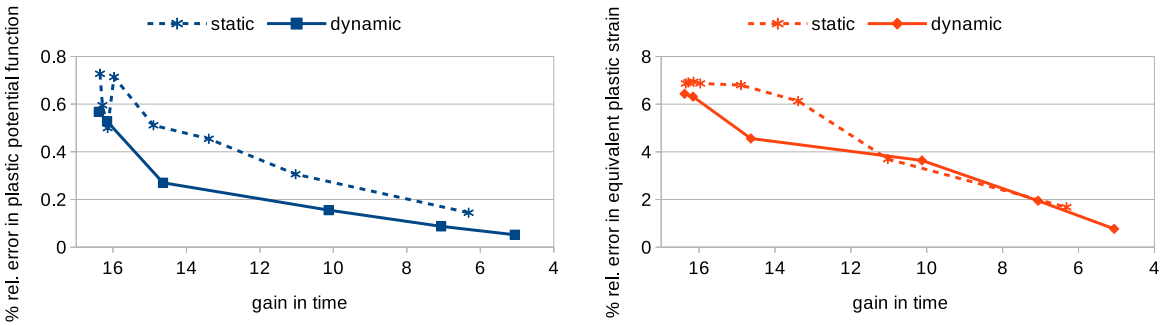
other hand, when we aim at a high gain in time, dynamic adaptive clustering is apparently superior to static clustering w.r.t. approximation error in equivalent plastic strain $e^{\bar{\gamma}}$. However, for low gain in time values both clustering strategies give nearly identical results.

For a large number (e.g. $\geq 729$) of statically constructed clusters, each consisting of a few points, the clusters are not affected much by the evolution of homogeneous field values. On the other hand, in the dynamic adaptive clustering approach with a small threshold value (e.g. $\leq 0.2$) we end up with a large number of clusters and the construction and maintenance of them require a significant amount of computation time. For example, with threshold value of 0.2 we have 1949 clusters at the end of the simulation, which requires at least 1948 split operations. Around 33 m of computation time is spent for this overhead, if a single split takes 1 s (roughly calculated), which is completely avoided in static clustering. This overhead increases exponentially as the number of generated clusters increases exponentially with the threshold value, also seen in Figure 12. However, if we ignore the computation time and pay attention to the savings in updates, the significance of the dynamic adaptive approach can be shown, see Figure 13. Again, the approximation error in plastic potential function is significantly less for the dynamic adaptive approach and the gap reduces for large number of static clusters. The dynamic adaptive approach leads to a lower approximation error also in equivalent plastic strain. Thus, we can infer that the dynamic adaptive approach can better capture the real dynamics and the deformation history.

## 5 CONCLUSIONS AND FUTURE WORK

The results presented above show that spatial clustering can considerably reduce the computational cost of the HMS model while retaining acceptable accuracy. Also, dynamic adaptive clustering is preferable to static clustering as the former has lower approximation error with the same number of updates. Additionally, the approximation error and the computational benefit depend on the number of clusters used. A certain trade-off has to be made: by increasing the number of clusters, the approximation error can be decreased, but at the same time the computational cost increases.

At present, we only consider the equivalent plastic strain value at an integration point as the feature of interest for clustering. As described earlier, each of the strain tensor components influences the evolution of the texture and the associated plastic anisotropy. Thus clustering based on the strain tensor components is expected to generate more accurate clusters and sub-




(a) (b)

Figure 11: Approximation error in (a) plastic potential function, (b) equivalent plastic strain for varying gain in time.

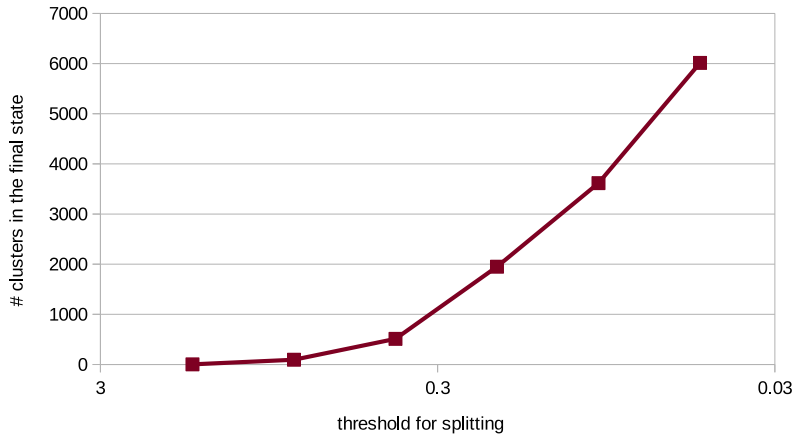
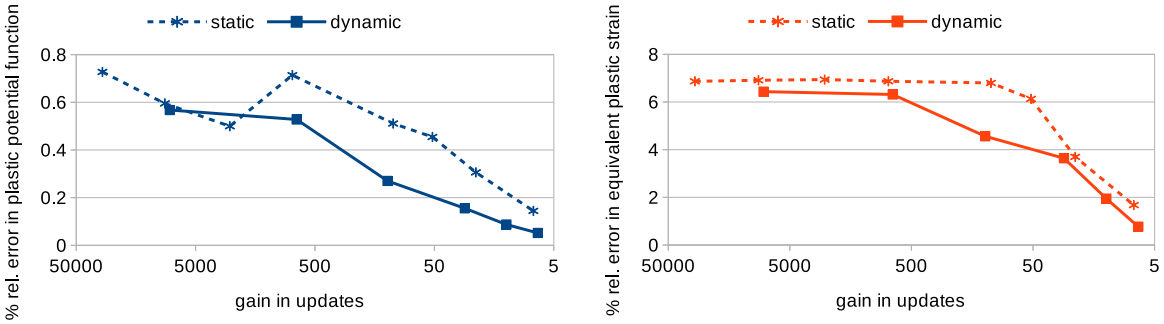



Figure 12: Resulting number of clusters in the dynamic adaptive approach for varying threshold value for splitting a cluster.




(a) (b)

Figure 13: Approximation error in (a) plastic potential function, (b) equivalent plastic strain for varying gain in updates $(g_u)$, given in 14.

sequently a lower approximation error. We intend to implement a strain mode aware clustering and updating.

The overhead of the dynamic adaptive clustering approach is significant in the simulation experiments with the current test case. In the current approach of adaptive clustering, only

splitting of clusters is considered. We can also consider merging two adjacent clusters, which will effectively reduce the number of active clusters and update operations.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from the Knowledge Platform M2Form, funded by IOF KU Leuven, and from the Belgian Federal Science Policy agency, contracts IAP7/19 and IAP7/21. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government department EWI.

## REFERENCES

[1] J. Gawad, A. Van Bael, P. Eyckens, G. Samaey, P. Van Houtte, D. Roose, Hierarchical multi-scale modeling of texture induced plastic anisotropy in sheet forming. *Computational Material Science*, **66**:65–83, 2013.

[2] H.-J. Bunge, Texture analysis in materials science, Butterworth, London, 1982.

[3] P. Van Houtte, S. Li, M. Seefeldt, L. Delannay, Deformation texture prediction: from the Taylor model to the advanced Lamel model, *International Journal of Plasticity*, **21(3)**:589–624, 2005.

[4] J. Han, M. Kamber, A.K.H. Tung, Spatial clustering methods in data mining: a survey. *Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS*, 2001.

[5] C.A. Duartea, T.J. Liszka, W.W. Tworzydlob, Clustered generalized finite element methods for mesh unrefinement, non-matching and invalid meshes. *International Journal for Numerical Methods in Engineering*, 1–28, 2006.

[6] R.L. Bras, T. Damoulas, J.M. Gregoire, A. Sabharwal, C.P. Gomes, R.B.V. Dover, Computational thinking for material discovery: bridging constraint reasoning and learning. *CROCS 2010 - 2nd International Workshop on Constraint Reasoning and Optimization for Computational Sustainability*, 2010.

[7] A.S. Varde, E.A. Rundensteiner, C. Ruiz, D.C. Brown, M. Maniruzzaman, R.D. Sisson Jr., Integrating clustering and classification for estimating process variables in materials science. *Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*, **13**:495–496 (Preface), 2006.

[8] A. Nakano, Fuzzy clustering approach to hierarchical molecular dynamics simulation of multiscale materials phenomena. *Computer Physics Communications*, **105(2–3)**:139–150, 1997.

[9] P.G. Ferreira, C.G. Silva, P.J. Azevedo, R.M.M. Brito, Spatial clustering of molecular dynamics trajectories in protein unfolding simulations. *Computational Intelligence Methods for Bioinformatics and Biostatistics, Lecture Notes in Computer Science*, **5488**:156–166, 2009.

[10] A. Alzraiee, L.A. Garcia, Using cluster analysis of hydraulic conductivity realizations to reduce computational time for Monte Carlo simulations. *Journal of Irrigation and Drainage Engineering*, **138(5)**:424–436, 2012.

[11] M. Vannucci, G.F. Porzio, V. Colla, B. Fornai, Use of clustering and interpolation techniques for the time-efficient simulation of complex models within optimization tasks *5th European Symposium on Computer Modeling and Simulation*, 6–11, 2011.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**:2825–2830, 2011.

[13] H. Kriegel, P. Kröger, A. Zimek, Clustering High-dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering. *ACM Trans. Knowl. Discov. Data*, **3(1)**:1–58, 2009.

[14] J.C. Simo, T.J.R. Hughes, Computational Inelasticity. *Interdisciplinary Applied Mathematics*, **7**, Springer, New York, 1998.

[15] M.L. Wilkins, Calculation of elastic-plastic flow. *Methods in Computational Physics*, **3**:211–263, 1964.

[16] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, **63(2)**:411–423, 2001.

[17] R. Lletí, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, **515(1)**:87–100, 2004.

[18] M. Khairullah, J. Gawad, D. Roose, A. Van Bael, Accelerating the hierarchical multi-scale software by spatial clustering strategies. *Computational Plasticity XIII Fundamentals and Applications*, 932–943, 2015.

[19] P. Van Houtte, S.K. Yerra, A. Van Bael, The Facet method: A hierarchical multilevel modelling scheme for anisotropic convex plastic potentials, *International Journal of Plasticity*, **25(2)**:332–360, 2009.

[20] G.M. Amdahl, Validity of the single-processor approach to achieving large scale computing capabilities. *In AFIPS Conference Proceedings*, **30**:483–485, 1967.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# BLACK-BOX SOLVER FOR ONE DIMENSIONAL MULTISCALE MODELLING USING THE QTT FORMAT

**Ivan V. Oseledets**[1,2], **Maxim V. Rakhuba**[1], **Andrei V. Chertkov**[1]

[1] Skolkovo Institute of Science and Technology
Skolkovo Innovation Center, Building 3, Moscow 143026, Russia
e-mail: {i.oseledets, m.rakhuba}@skoltech.ru, andrei.chertkov@skolkovotech.ru

[2] Institute of Numerical Mathematics, Russian Academy of Sciences
Gubkina St. 8, Moscow 119333, Russia
e-mail: ivan.oseledets@gmail.com

**Keywords:** Stable PDE discretization, TT-format, QTT-format, multiscale modelling, black-box solver.

**Abstract.** *In the present work we propose an efficient black-box solver for one-dimensional multiple scaled diffusion equation. For this problem it has been recently shown [1] that the solution can be represented in a certain low parametric representation, namely the quantized tensor train (QTT) format [2]. The key idea of the QTT format is to make the real space data multidimensional by introducing virtual dimensionalities. The next step is to apply the tensor train (TT) representation [3] to multidimensional data, which leads us to the logarithmic complexity. Hence very fine grids that describe the finest scale can be used.*

*Since the solution of second order multi-scale problems can be represented in the QTT format, simple and efficient solvers can be developed using the existing software for the approximate solution of linear systems in the TT-format. However, if equations are discretized using standard finite element/difference methods, it is not possible to get to very fine meshes, say with $2^{50}$ grid points due to the condition number. On the other hand, the theory guarantees the existence of a good QTT-FEM approximant of the continuous problem. Thus, another discretization should be used to compute it numerically.*

*Our idea is to rewrite the initial formulation in a certain form without derivatives. After that we get an explicit formula, which consists of the inversion of a diagonal matrix and the multiplication by a dense matrix. The latter can be multiplied with logarithmic complexity in the QTT format due to a special structure. The numerical experiment show that this formula gives accurate results and can be used for $2^{50}$ grid points with no problems with conditioning, while total computational time is around several seconds.*

## 1 INTRODUCTION

We consider a model 1D diffusion equation

$$-\frac{\partial}{\partial x}\left(k(x)\frac{\partial}{\partial x}u(x)\right) = f(x), \quad u(0) = u(1) = 0 \tag{1}$$

with coefficient $k(x)$ that has multiple scales. The problem to solve this equation directly is that very fine grid that describes the finest scale has to be introduced. Alternatively one can solve this problem using the approach based on analytical expansions or use specific multiscale finite element methods. Despite these approaches work very efficiently and are well-developed, still they do not have enough generality and rely on the knowledge of analytical behavior of the solution.

In [4] it was shown that the solution of (1) can be represented in a certain low parametric representation, namely the quantized tensor train (QTT) format [2]. The idea behind the QTT approach is as follows. First of all we introduce very fine grid that is able to describe the finest scale of the problem. To work efficiently with such grids we use low-parametric representations, namely tensor decompositions that deal with high-dimensional data. The key idea is to make the real space 1D data multidimensional by introducing virtual dimensionalities, which leads us to logarithmic complexity.

Although the solution can be approximated in the QTT-format [1], it is very diffucult to recover using standard finite difference approaches. Indeed, even if $k \equiv 1$ the simplest discretization scheme reads

$$-\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f(x_i), \quad i = 1, \ldots, n, \quad u_0 = u_{n+1} = 0, \quad h = \frac{1}{n+1}.$$

It is well known that $u_i - u(x_i) = \mathcal{O}(h^2)$, i.e. the smaller the $h$, the better is the approximation. However, in numerical computations we can not take $h$ too small. Let $\epsilon$ be the accuracy of the computations. Then the approximation error of the action of discrete operator can be estimated as

$$\mathcal{O}\left(\frac{\epsilon}{h^2} + h^2\right),$$

which means that the minimal possible grid step is $h \sim \epsilon^{-1/4}$. For the double precision $\epsilon \approx 10^{-16}$ the grid step $h \sim 10^{-4}$ is the minimal possible. In principle, such small grid steps are rarely (not to say never) encountered in standard mathematical modelling, especially when we go from one-dimensional to 2D and 3D problems. However, as was already mentioned we have recently encountered a problem when we need $h$ much beyond the number mentioned above, even $h \sim 2^{-d}$ where $d \geq 20$.

In this paper we propose an explicit formula for finding solution of (1) that resolves the problem with accuracy on very fine grids. We describe how to apply this formula in the QTT format and provide bound estimates for the rank of solution. We also show the relation between the proposed formula and standard second order finite difference scheme. In numerical experiments we illustrate theoretical results and provide comparison with the homogenization approach.

## 2 ROBUST DISCRETIZATION SCHEME

Consider a one-dimensional diffusion equation (1). This problem is equivalent to the minimization of the functional

$$u = \arg\min_{v(0)=v(1)=0} F(v), \quad F(v) = \int_0^1 k\left(\frac{\partial v}{\partial x}\right)^2 dx - 2\int_0^1 vf dx.$$

Now we introduce additional variable

$$v_x = \frac{\partial v}{\partial x} = B(v). \tag{2}$$

From (2) and taking into account the boundary conditions, we can write

$$v(x) = \int_0^x v_x(t)dt.$$

To satisfy the boundary condition at $x = 1$ the function $v_x$ has to satisfy

$$\int_0^1 v_x(t)dt = 0.$$

Finally we have the following optimization problem:

$$u_x = \arg\min \int_0^1 k v_x^2 dx - 2\int_0^1 \mathcal{B}(v_x) f dx, \quad \text{s.t.} \quad \int_0^1 v_x(t)dt = 0, \tag{3}$$

where $u_x = u'$ is the derivative of the solution of equation (1).

Now we replace the integrals in (3) by the rectangular rule on the uniform mesh with grid step $h$ and have the following quadratic optimization problem:

$$F(v_x) = (Dv_x, v_x) - 2(Bv_x, f) = (Dv_x, v_x) - 2(v_x, B^\top f), \quad \text{s.t.} \quad e^\top u_x = 0,$$

where $B$ is the discretization of the operator $\mathcal{B}$, $D$ is a diagonal matrix with

$$d_i = k\left(x_{i-\frac{1}{2}}\right), \quad x_{i-\frac{1}{2}} = \left(i - \frac{1}{2}\right)h, \quad i = 1, \ldots, n$$

and $e$ is the vector of all ones. The unknowns $u_x$ are defined in the midpoints as well. Introducing Lagrange multiplier for the constraint, we have

$$Du_x = B^\top f + \alpha e, \quad e^\top u_x = 0,$$

therefore

$$\alpha = -\frac{e^\top D^{-1} B^\top f}{e^\top D^{-1} e}.$$

Finally the solution is given (not unexpectedly!) by the explicit formula

$$u = Bu_x = BD^{-1}B^\top f - \frac{e^\top D^{-1} B^\top f}{e^\top D^{-1} e} BD^{-1}e. \tag{4}$$

Matrix $B$ plays a crucial role. Let $u$ be defined on the grid points, and the centered second-order finite different scheme reads

$$(u_x)_{i-\frac{1}{2}} = \frac{u_i - u_{i-1}}{h},$$

and the matrix $B$ is given as

$$B_{ij} = \begin{cases} h, & i \geq j, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The formula (4) can be considered as a stable discretization of the original equation. It involves only elementwise operations and Volterra integral operations, and it is easy to see that errors decrease while $h$ goes to zero. Moreover, these operations can be efficiently implemented in the QTT-format, and this is the main motivation why such discretizations are interesting in practice.

## 3 CONNECTION WITH FINITE DIFFERENCE SCHEME

The next theorem shows the relation between standard second order discretization scheme and the formula (4).

**Theorem 1.** *In exact arithmetics solution obtained by the proposed formula* (4) *is equivalent to the solution obtained by the standard second order discretization scheme on uniform grid*

$$\begin{gathered} -\frac{k_{i+1/2}u_{i+1} - (k_{i+1/2} + k_{i-1/2})u_i + k_{i-1/2}u_{i-1}}{h^2} = f(x_i), \quad i = 1, \ldots, n, \\ u_0 = u_{n+1} = 0, \quad h = \frac{1}{n+1}. \end{gathered} \tag{6}$$

*Proof.*

$$\text{Let} \quad B = h \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ \vdots & & \ddots & \\ 1 & \ldots & \ldots & 1 \end{pmatrix}, \quad \text{then} \quad B^{-1} = \frac{1}{h} \begin{pmatrix} 1 & & & \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix},$$

as it is easy to check that $BB^{-1} = I$. Let us denote by $u_x$ vector of approximate derivatives

$$u_x \equiv \left( \frac{u_1 - u_0}{h}, \frac{u_2 - u_1}{h}, \ldots, \frac{u_{n+1} - u_n}{h} \right)^T,$$

and $u = (u_1, \ldots, u_{n+1})$. Due to the fact that $u_0 = 0$ we get

$$u_x = B^{-1}u.$$

Due to (6)

$$B^{-T}DB^{-1}u = \left( f(x_1), \ldots, f(x_n), k_{n+1/2}\frac{u_{n+1} - u_n}{h^2} \right)^T, \tag{7}$$

where

$$D = \text{diag}(k_{1/2}, \ldots, k_{n+1/2}).$$

Using additional information that $u_{n+1} = 0$ we get

$$\left(h^{-2}k_{n+1/2}e_{n+1}e_n^T + B^{-T}DB^{-1}\right)u = f,$$

where $e_i$ is zero vector with only one 1 in the $i$-th position and $f = (f(x_1), \ldots, f(x_n), 0)^T$.

Let us apply Sherman-Woodbury-Morrison formula

$$u = BD^{-1}B^T f - h^{-2}k_{n+1/2}\frac{BD^{-1}B^T e_{n+1}e_n^T BD^{-1}B^T f}{1 + h^{-2}k_{n+1/2}e_n^T BD^{-1}B^T e_{n+1}}. \tag{8}$$

To get (4) let us simplify the latter expression. First of all,

$$B^T e_{n+1} = he, \quad e_n^T B = h(1, \ldots, 1, 0),$$

therefore,

$$BD^{-1}B^T e_{n+1}e_n^T BD^{-1}B^T f = h^{-2}BD^{-1}e\,(1, \ldots, 1, 0)D^{-1}B^T f = h^{-2}BD^{-1}ee^T D^{-1}B^T f$$

Finally the denominator in (8) can be written as

$$1 + h^{-2} k_{n+1/2} e_n^T B D^{-1} B^T e_{n+1} = k_{n+1/2} \left( \frac{1}{k_{n+1/2}} + \frac{1}{k_{1/2}} + \cdots + \frac{1}{k_{n-1/2}} \right) = k_{n+1/2}\, e^T D^{-1} e,$$

As a result, we get formula (4)

$$u = BD^{-1}B^\top f - \frac{e^\top D^{-1} B^\top f}{e^\top D^{-1} e} BD^{-1} e.$$

□

**Remark 2.** *The important consequence of Theorem 1 is that solution obtained by the formula* (4) *converges to the exact solution with the second order.*

**Remark 3.** *From* (7) *it follows that if Dirichlet-Neumann boudary conditions are used, i.e.* $u_0 = 0$ *and* $u_{n+1} = u_n$*, then the formula* (4) *reads* $u = BD^{-1}B^\top f$.

## 4 QTT REPRESENTATION FOR THE ONE DIMENSIONAL CASE

The concept of the QTT looks as follows. Let $n = 2^d$, then the vector has $2^d$ unknowns. We treat this one-dimensional vector as a $d$-dimensional tensor of size $2 \times \ldots \times 2$. This tensor $V$ is then approximated in the *tensor-train* (TT) format. A tensor $V(i_1, \ldots, i_d)$ is said to be in the TT-format, if

$$V(i_1, \ldots, i_d) = G_1(i_1) G_2(i_2) \ldots G_d(i_d),$$

where $G_k(i_k)$ is an $r_{k-1} \times r_k$ matrix for each fixed $i_k$, and $r_0 = r_d = 1$.

The main benefit of the QTT-format is that it leads to logarithmic complexity to represent the vector of unknowns, if the ranks $r_k$ are bounded: we only need to store $\mathcal{O}(dr^2)$ parameters. For elliptic problems, the upper bounds of QTT-ranks were provided in [5] and extended to the highly oscillating case in [4]. The last case is the most practically interesting, since it is exactly the case when astronomically large grids are needed. In order to turn (4) into a computational formula, we need a tensor representation of the matrices and vectors involved. Linear operator acting on tensors from $R^{\otimes_{i=1}^d n_i}$ to $R^{\otimes_{i=1}^d n_i}$, and is naturally represented as a $2d$ tensor

$$A(i_1, \ldots, i_d; j_1, \ldots, j_d).$$

Such linear operator is said to be in the *TT-matrix format*, if

$$A(i_1, \ldots, i_d; j_1, \ldots, j_d) = A_1(i_1, j_1) \ldots A_d(i_d, j_d).$$

For $r_k = 1$ this boils down to the Kronecker product of $2 \times 2$ matrices. The product of two TT-matrices is also a TT-matrix with ranks bounded by the product of the ranks of the terms, thus it is only necessary to put the matrices $B$ and $D^{-1}$ into the QTT-format.

**Lemma 4.** *The matrix* $B$ *defined by* (5) *can be exactly represented in the QTT-format with QTT-ranks equal to* 2*.*

**Lemma 5.** *Let* $d$ *be a vector with* $2^d$ *elements in the QTT-format with QTT-ranks* $r_k$*. Then, the matrix*

$$D = \mathrm{diag}(d),$$

*can be represented in the QTT-format with QTT-ranks* $r_k$*.*

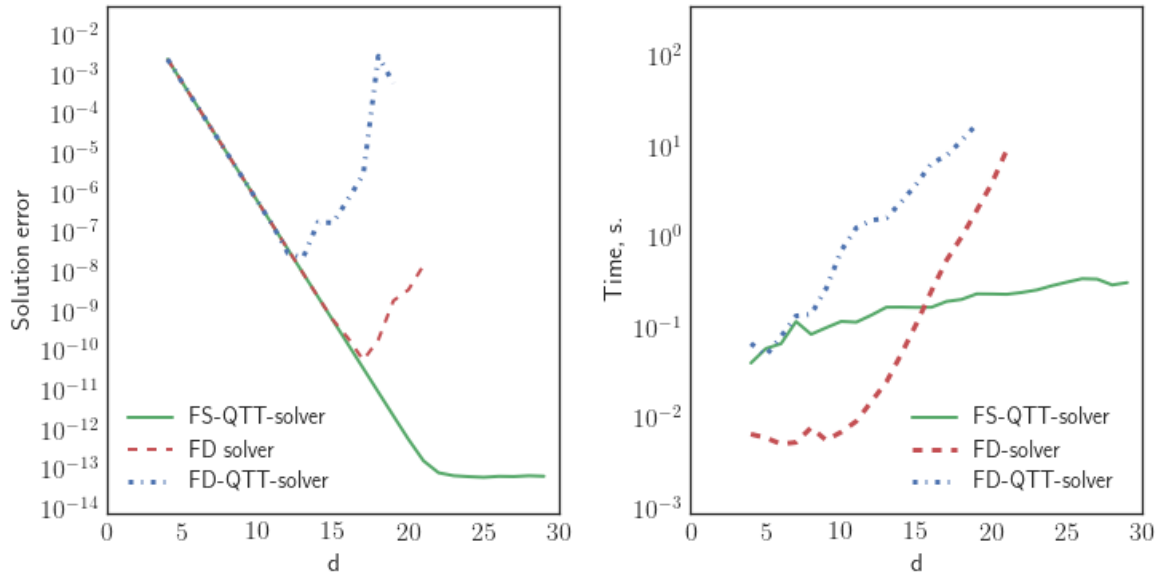



Figure 1: Error of calculated solution $u$ (on the left plot) and total calculation time (on the right plot) w.r.t. the mesh size factor $d$ (total number of grid nodes is $2^d$) for the model PDE with known analytic solution. Results are presented for three different solvers, that are described in the text.

Thus, the most difficult task is to put the vector of values of the function $k^{-1}$ to the QTT-format. For many practically interesting cases, the QTT-ranks are bounded [6, 7, 8]. To get such approximation the cross approximation algorithm is the method of choice [9], which allows to recover the approximation by adaptively sampling $\mathcal{O}(dnr^2)$ points. The right-hand also has to be put into the QTT-format using the same cross approximation procedure.

Provided that both $f$ and $k^{-1}$ are represented in the QTT format, it is easy to find bounds on rank of the solution.

**Lemma 6.** *Let $f$ and $k^{-1}$ has maximal ranks $r_f$ and $r_{k^{-1}}$ correspondingly. In this case maximal rank $r_u$ of the solution satisfies*

$$r_u \leq 2r_{k^{-1}}(2r_f + 1)$$

*Proof.* The proof immediately follows from the fact that the bound on rank of matrix-vector product is product of ranks. In our case

$$\text{rank}(BD^{-1}B^T u) \leq 2 \cdot r_{k^{-1}} \cdot 2 \cdot r_f, \quad \text{rank}(BD^{-1}e) \leq 2 \cdot r_{k^{-1}}.$$

□

## 5 NUMERICAL EXPERIMENTS

In this section we illustrate the theoretical results presented above with numerical experiments. Firstly, we consider a PDE with known analytic solution for validation of the developed solver (denoted hereinafter as finite sum QTT-solver or FS-QTT-solver). After that we consider a more complicated case of multiscale PDE. Special analytic form of multiscale PDE coefficients makes it possible to construct exact homogenized solution and first order correction, hence we can check the accuracy of numerical computations result in the terms of energy.

We compared calculation results obtained by FS-QTT-solver with the results of two solvers based on a finite difference discretization scheme. The first of them (denoted hereinafter as

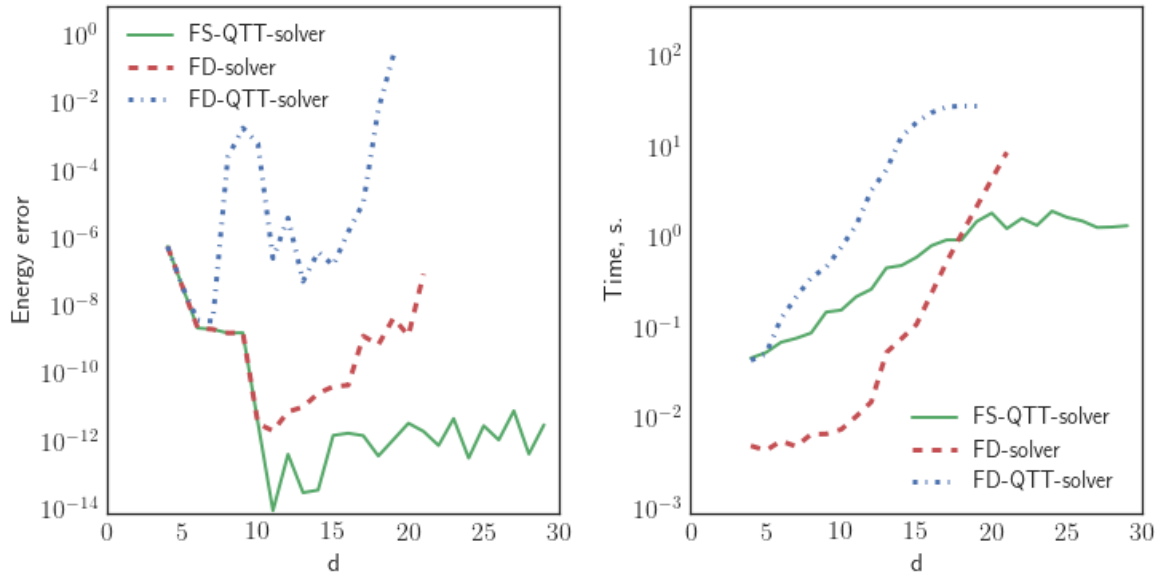



Figure 2: Error of the energy $(D\nabla u, \nabla u)$ (on the left plot) and total calculation time (on the right plot) w.r.t. the mesh size factor $d$ (total number of grid nodes is $2^d$) for the multiscale PDE with scale parameter $\epsilon = 10^{-4}$. Results are presented for three different solvers, that are described in the text.

FD-solver) is constructed in sparse format, and hence can operate only with moderate grids. We also construct a QTT version of the FD-solver (denoted hereinafter as FD-QTT-solver).

### 5.1 Method validation for PDE with known analytic solution

First we consider a PDE with homogeneous Dirichlet boundary conditions:

$$-\left(k(x)u'(x)\right)' = f(x), \quad x \in [0,1], \quad u(0) = u(1) = 0, \tag{9}$$

with $k(x) = 1 + x$ and $f(x) = \pi^2(1+x)\sin(\pi x) - \pi\cos(\pi x)$. This problem has exact analytic solution of the form

$$u(x) = \sin(\pi x). \tag{10}$$

Then for the numerical solution $\hat{u_h}$ on a uniform grid with step $h$ we can calculate an error:

$$E_h^{(1)} = \frac{||\hat{u_h} - u_h||_2}{||u_h||_2},$$

where $u_h$ is the exact solution (10) discretized on the same grid.

The dependence of $E_h^{(1)}$ and the total calculation time on grid size $h$ for FS-QTT, FD and FD-QTT-solver is presented in Figure 1. As follows from the results, all solvers have the same accuracy for small grids ($d < 10$). FD-QTT-solver and FD-solver as anticipated become unstable for finer grids and the second order convergence for larger $d$ is remained only for FS-QTT-solver. At the same time FS-QTT-solver works faster than FD-solver for grids with $d > 15$ as time scales linearly with $d$.

### 5.2 Multiscale problem

Here we consider the multiscale case of equation (9) with $f = -1$ and two-scale coefficient $k^\epsilon$ of the form

$$k(x,y) = k_0(x)k_1(y), \quad k_0(x) = 1 + x, \quad k_1(y) = \frac{2}{3}(1 + cos^2(2\pi y)).$$

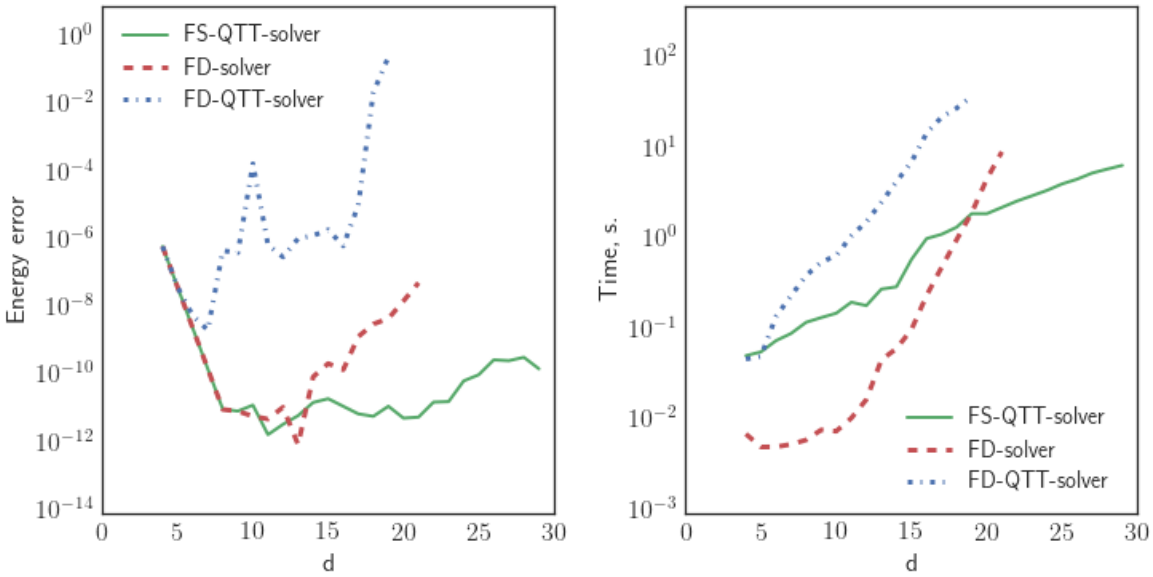



Figure 3: Error of the energy $(D\nabla u, \nabla u)$ (on the left plot) and total calculation time (on the right plot) w.r.t. the mesh size factor $d$ (total number of grid nodes is $2^d$) for the multiscale PDE with scale parameter $\epsilon = 10^{-6}$. Results are presented for three different solvers, that are described in the text.

that is $Y$-periodic function for

$$y = \frac{x}{\epsilon} \in Y = (0, 1).$$

The parameter $\epsilon > 0$ stands for some small scale in the problem.

It is a classical result of homogenization theory [10] that for small $\epsilon$, $u^\epsilon$ and $\nabla u^\epsilon$ can be approximated by

$$u^\epsilon(x) \approx u_0(x) + \epsilon\xi(y)\nabla_x u_0,$$

and

$$\nabla u^\epsilon(x) \approx \nabla_x u_0(x) + \nabla_y \xi(y)\nabla_x u_0,$$

where $y = x/\epsilon$ and $\xi$ is the $Y$-periodic solution of the so-called cell problem:

$$-\left(k^\epsilon(x, y)\xi'(y)\right)'_y = \left(k^\epsilon(x, y)\right)'_y. \tag{11}$$

The considered two-scale limiting equation has the exact homogenized solution

$$u_0(x) = \frac{3}{2\sqrt{2}}\left(x - \frac{\log(1 + x)}{\log 2}\right),$$

and $\xi$ has a form

$$\xi = \left(\frac{1}{2\pi}\tan^{-1}\left(\frac{\tan(2\pi y)}{\sqrt{2}}\right) - y + C\right). \tag{12}$$

and is determined within an additive constant $C$ for fixed $x$.

According to (4) FS-QTT-solver computes both the solution of the equation $\hat{u}$ and it's derivative $\hat{u_x}$, hence we can construct an energy functional: $(D\hat{u_x}, \hat{u_x})$, where $(\cdot, \cdot)$ is a scalar product, and compare its value with the same functional for $\nabla u^\epsilon$ from (11). FD-solver and FD-QTT-solver calculate only solution $\hat{u}$ of the PDE, but for comparison purposes we also need to construct approximation of $\hat{u_x}$. We do it by applying finite difference operator for the obtained

solution $\hat{u}$. Then we can calculate an error:

$$E_h^2 = \frac{|(D\hat{u_x}, \hat{u_x}) - (D\nabla u^\epsilon, \nabla u^\epsilon)|}{(D\nabla u^\epsilon, \nabla u^\epsilon)}.$$

In the case of QTT-based solvers the calculations may be performed for a huge grid sizes and the reference functions (12) and (11) must be constructed on the same grid for comparison. We use a cross-approximation method for this purpose with accuracy two orders of magnitude greater than the one that was used for the numerical solution of the equation and calculate both error norms in the TT-format.

The dependence of $E_h^{(2)}$ and the total calculation time on grid size $h$ for FS-QTT, FD and FD-QTT-solver for scale parameter values $\epsilon = 10^{-4}$ and $\epsilon = 10^{-6}$ are presented in Figure 2 and Figure 3 respectively. As follows from the results, FS-QTT-solver outperforms both FD-QTT-solver and FD-solver in accuracy and in calculation time for fine grids.

## 6 CONCLUSIONS

In this paper we proposed explicit formula that resolves the problem with accuracy on very fine grids. We showed how to use it in the QTT format. We also proved that in exact arithmetics this formula is equivalent to the second-order finite discretization. Numerical experiments illustrated efficiency of the proposed formula.

## ACKNOWLEDGEMENT

The work was supported by the Ministry of Education and Science of Russian Federation, Grant Agreement no. 14.618.21.0004, the unique project identifier RFMEFI61815X0004.

## REFERENCES

[1] V. Kazeev, I. Oseledets, M. Rakhuba, and Ch. Schwab. Qtt-fe approximation for multiscale problems. *Research Report, Seminar for Applied Mathematics*, 2015.

[2] Ivan V Oseledets. Approximation of 2ˆd\times2ˆd matrices using tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2130–2145, 2010.

[3] Ivan V Oseledets. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

[4] Vladimir Kazeev, Ivan Oseledets, Maxim Rakhuba, and Christoph Schwab. Qtt-finite-element approximation for multiscale problems. Technical Report 2016-06, Seminar for Applied Mathematics, ETH Zürich, 2016.

[5] Vladimir Kazeev and Ch Schwab. Quantized tensor-structured finite elements for second-order elliptic pdes in two dimensions. Technical report, SAM research report 2015-24, ETH Zürich, 2015.

[6] I. V. Oseledets. Constructive representation of functions in low-rank tensor formats. *Constr. Approx.*, 37(1):1–18, 2013.

[7] B. N. Khoromskij. $\mathcal{O}(d \log n)$–Quantics approximation of $N$–$d$ tensors in high-dimensional numerical modeling. *Constr. Approx.*, 34(2):257–280, 2011.

[8] L. Grasedyck. Polynomial approximation in hierarchical Tucker format by vector-tensorization. DFG-SPP1324 Preprint 43, Philipps-Univ., Marburg, 2010.

[9] I. V. Oseledets and E. E. Tyrtyshnikov. TT-cross approximation for multidimensional arrays. *Linear Algebra Appl.*, 432(1):70–88, 2010.

[10] G Papanicolau, A Bensoussan, and J-L Lions. *Asymptotic analysis for periodic structures*. Elsevier, 1978.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5-10 June 2016

# SCALE EFFECT IN MICROFLOWS MODELLING WITH THE MICROPOLAR FLUID THEORY

**Anna Kucaba-Pietal**

Rzeszow University of Technology
Powstańców Warszawy 12, 350959 Rzeszów, Poland
anpietal@prz.edu.pl

**Keywords:** Micropolar fluid, Squeeze flow, Hagen-Poiseuille flow.

**Abstract.** *The aim of this paper is to perform a comparative study of the scale effect on the applicability of the theory of micropolar fluids to microflows modelling. Flows which often appear in microfluidics systems: the squeezing flow between flat plates and the Hagen–Poiseuille flow are considered.*
*For each flow the maximal geometrical dimension of the flow field for which the "micropolar ity" of the fluid affects flows characteristics is predicted and its value is calculated for water, blood and electro-rheological suspension. Results indicate that for the Hagen–Poiseuille flow the dimension is ten times greater than for the squeezing flow.*

## 1. INTRODUCTION

Fluid flows in small scales have received increased attention over the last decade due to the fact that modelling and simulating micro- and nanoscale fluid systems is still a challenging task. In such scales the effect of interaction between the fluid molecules and the solid wall molecules is much more significant than in its bulk case described by classical (i.e. Cauchy type) continuum mechanics theories that are not capable of explaining the observed effects such as: microrotation of molecules, density fluctuations or the electrokinetic effect [1]. Recently, the most powerful, and perhaps the most widely used tools for studying such fluid systems are the molecular dynamic and the Monte Carlo methods. However, it is well known that both approaches are computationally expensive.

Other possibility to model flows in small scales is to employ a generalized continuum mechanics approach, in which more material constants allow to capture some microstructural features that are non-detectable by classical theories. The micropolar fluid theory was developed by Eringen in 1966 [2] in the frame of the generalized continuum mechanics. It takes into account the microrotation of molecules (microrotation – spinning motion of molecules that is independent on the rotation of the flow velocity field) and introduces additionally a couple stress tensor. In this theory, field equations are presentable in terms of two independent kinematic vector fields: the velocity and the microrotation vector, and equations involve six material coefficients. Microrotation vector differs from vorticity vector which is defined as rotation of velocity. The stress tensor is not symmetric.

Micropolar fluid flow equations are being reduced to their counterparts in the classical continuum mechanics (Cauchy), when the microrotation is omitted or when the characteristic linear dimension of the flow field is large enough [3]. It indicates that the geometrical size of the flow field plays a crucial role in the applicability of the micropolar fluid theory to microflows modelling. Till now, only the Hagen–Poiseuille flow in circular tube in this aspect was examined in detail [3].

This paper presents the analysis of the scale impact on the effective micropolar fluid theory microflows modelling. The Hagen–Poiseuille flow and the squeezing flow are considered. Flow characteristics are determined. Values of experimentally determined micropolar fluid constants [4] and values predicted on molecular dynamic simulations [5] are used in the calculations. Formulas for the maximal distance between plates for the squeezing flow and the maximal cross-sectional size of the channel for the Hagen–Poiseuille flow, for which the micropolar effects of the fluid affects flow characteristics are established and their values for some real fluids are calculated. Beyond this dimension, it pays off to carry on the calculations on the basis of the classical (Cauchy) hydrodynamics. The comparative study of obtained results for considered flows is performed. The results indicate that the scale effect depends on the type of the flow. The maximal geometrical dimension of the flow field, for which the fluid micropolar effects are negligible, is ten times greater for the squeeze flow than for the Hagen-Poiseuille flow. It indicates that for each type of the fluid flow its "own" limiting dimension value should be estimated for effective modelling by using the micropolar fluid theory.

## 2. MICROPOLAR FLUID FIELD EQUATIONS

The micropolar fluid theory (MFT) is referred as fluids with an asymmetric stress tensor or extended Navier–Stokes theory [6]. The coupling between the hydrodynamic flow degree of freedom which is described by the velocity vector, and the microscopic molecular spin angular velocity degree of freedom which is described by the microrotation vector is visible in field equations. In the most general form, the micropolar field equations which represent

conservation of mass, conservation of impulse and angular momentum for incompressible and viscous fluid are as follows:

$$\mathrm{div}\boldsymbol{V} = 0 \tag{1}$$

$$\rho\frac{\mathrm{d}\boldsymbol{V}}{\mathrm{d}t} = \rho\boldsymbol{f} - \mathrm{grad}\,p - (\mu+\kappa)\mathrm{rotrot}\,\boldsymbol{V} + \kappa\mathrm{rot}\boldsymbol{\omega} \tag{2}$$

$$\rho I\frac{\mathrm{d}\boldsymbol{\omega}}{\mathrm{d}t} = (\alpha+\beta+\gamma)\mathrm{graddiv}\boldsymbol{\omega} - \gamma\mathrm{rotrot}\boldsymbol{\omega} + \kappa\mathrm{rot}\boldsymbol{V} - 2\kappa\boldsymbol{\omega} + \rho\boldsymbol{g} \tag{3}$$

where $\rho$ is density, $\boldsymbol{V} = (V_1, V_2, V_3)$, the velocity field, $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)$ the micro-rotation field, $I$ the gyration parameter, $\boldsymbol{f} = (f_1, f_2, f_3)$, body forces per unit mass, $p$ the hydrostatical pressure, $\mu$ the classical viscosity coefficient, $\kappa$, $\lambda$ the vortex viscosity coefficients and $\alpha$, $\beta$, $\gamma$ are gyroviscosity coefficients satisfying the following inequalities:

$$3\alpha+\beta+\gamma \geq 0,\ 2\mu+\kappa \geq 0,\quad 3\lambda+2\mu+\kappa \geq 0,\quad \gamma \geq \beta,\quad \kappa \geq 0,\quad \gamma \geq 0.$$

The constitutive equations for the stress tensor $\boldsymbol{T} = \{T_{ij}\}$, and the couple stress tensor $\boldsymbol{C} = \{C_{ij}\}$, are expressed by following equations, respectively as:

$$T_{ij} = (-p+\lambda V_{k,k})\delta_{ij} + \mu(V_{i,j}+V_{j,i}) + \kappa(V_{j,i} - \varepsilon_{ijk}\omega_k) \tag{4}$$

$$C_{ij} = \alpha\omega_{k,k}\delta_{ij} + \beta\omega_{i,j} + \gamma\omega_{j,i} \tag{5}$$

$\varepsilon_{ijk}$ – the Levi-Civita tensor, $\delta_{ik}$ – the Kronecker delta.

The micropolar fluid flow equations arrive at the classical Navier–Stokes equations: (i) if transport coefficients: $\alpha,\beta,\gamma$, and $\kappa$ vanish (ii) if microrotation is ignored (iii) when the characteristic linear dimension of the flow field is large enough [3].

## 3. SIMPLE MICROPOLAR FLUID FLOWS

In this section the quantitatively analyze of a scale impact on the effective micropolar fluid microflows modelling is presented. Flows, for which analytical solutions are known, are considered. This way errors arising from numerical solutions are excluded. Two different flows are studied which represent two systems where fluid is confined in narrow channel: i) pressure driven planar Hagen–Poiseuille flow, ii) squeezing flow of narrow film between parallel plates. Such flows often appear in microfluidics systems. The influence of the molecular spin (microrotation) on the hydrodynamic quantities in dependence of characteristic linear flow dimension is analyzed. Exact, analytical solutions of Eqs (1-3) for the flows at hand are provided in the literature for different boundary conditions imposed on velocity and microrotation on the confining surfaces. In this paper, their simplest form is used; vanishing velocity and microrotation on the confined surfaces. Hydrodynamic quantities which characterize each of flows are expressed by analytical formulas. The results are compared with those obtained for classical (i.e. Cauchy type) continuum mechanics theory, in which mass transport is described by the Navier–Stokes equation and which omits microrotation vector.

### 3.1. Flow reduction in a planar Poiseuille flow

A fluid confined between parallel walls located at $z = 0$ with $h$ distance between them is considered. The z direction is the only direction of confinement and the walls are infinite in extent in the x, y-plane. Pressure gradient drives the planar flow. The micropolar equations

(1-3) can be solved analytically with Dirichlet slip boundary conditions where $\boldsymbol{V}(h) = 0$, $\boldsymbol{V}(0) = 0$ and $\boldsymbol{\omega}(\mathrm{h}) = \mathbf{0}$, $\boldsymbol{\omega}(0) = \mathbf{0}$ It has been shown that the relative flow rate reduction due to spinning motion of molecules is given by [4]:

$$Q(h) = Q_m(h)/Q_{\mathrm{N}}(h) \tag{6}$$

where

$$Q_m(h) = Q_{\mathrm{N}}(h)\left\{1 + \frac{2\delta}{2+\delta}\frac{4}{(kh)^2}\left[1 - \frac{\cosh(kh)}{\sinh(kh)}\right]\right\}$$

$$k = \sqrt{\frac{(2\mu+\kappa)\kappa}{(\mu+\kappa)\gamma}}$$

$$\delta = \frac{\kappa(1-\alpha_0)}{\mu_{\mathrm{N}}}$$

$Q_m$ and $Q_n$ are the flow rates when the fluid spinning motion of molecules described by microrotation is included (in the frame of micropolar fluid theory) and when it is ignored (in the classical - i.e. Cauchy type continuum mechanics theory).

### 3.2. Squeezing flow of narrow film between parallel plates

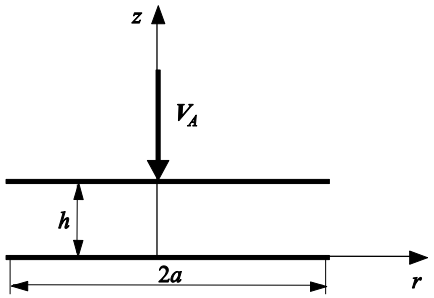



Fig. 1. Geometry of squeezing flow

A two-dimensional flow in fluid layer $0 < z < h$, $r < a$ between two parallel plates circular approaching each other symmetrically is considered. The plate at $z = 0$ is immobile, while the second plate moves with velocity $\boldsymbol{V_A}$.

The micropolar equations (1-3) can be solved analytically in Reynolds approach with Dirichlet slip boundary conditions where $\boldsymbol{V}(h) = \boldsymbol{V_A}$, $\boldsymbol{V(0) = 0}$ and $\boldsymbol{\omega}(h) = \mathbf{0}$, $\boldsymbol{\omega}(0) = \mathbf{0}$ and the relative load capacity increase due to the microrotation is defined as [8]:

$$W(h) = W_m(h)/W_N(h) \tag{7}$$

where:

$$W_m(h) = \frac{\pi\mu V_A a^4}{8h^3\left[\frac{1}{12} + \frac{l^2}{h^2} - \frac{Nl}{2h}\coth\frac{mh}{2}\right]}$$

and

$$W_{\mathrm{N}}(h) = \frac{3\pi\mu_{\mathrm{N}} V_A a^4}{2h^3}$$

Symbols $W_m$ and $W_N$ denote the load capacity when the fluid microrotation is included and when it is ignored i.e. predicted in the frame of the micropolar fluid theory and in the frame of the classical - i.e. Cauchy type continuum mechanics theory respectively.

## 4. RESULTS AND DISCUSION

In this section values of hydrodynamic quantities: (i) $Q(h)$ – relative flow rate reduction for the Hagen–Poiseuille flow (6), (ii) $W(h)$ – relative load capacity increase (7) for squeezing flow are calculated as a function of distance $h$ between walls confined the flow for water, blood and electrorheological suspensions, and plotted in Figures 1 and 2 respectively.

The values of micropolar viscosity coefficients used in calculations were found in literature. For blood, the values $\mu = 2.9 \cdot 10^{-3}$, $\kappa = 2.32 \cdot 10^{-4}$, $\gamma = 10^{-6}$ are listed in [7]. For water the values evaluated by Hansen et al. (2011) using equilibrium Molecular Dynamics are used: $\mu = 0.7 \cdot 10^{-3}$, $\kappa = 0.17 \cdot 10^{-3}$, $\gamma = 2.1 \cdot 10^{-21}$. For electro-rheological suspension, the viscosity coefficient values are given in reference to viscosity $\mu_N$ and reads: $\mu/\mu_N$ = = $1.922 \cdot 10^{-3}$, $\kappa/\mu_{\mathrm{N}} = 5.845$, $\gamma/\mu_{\mathrm{N}} = 9.94 \cdot 10^{-10}$ [4].

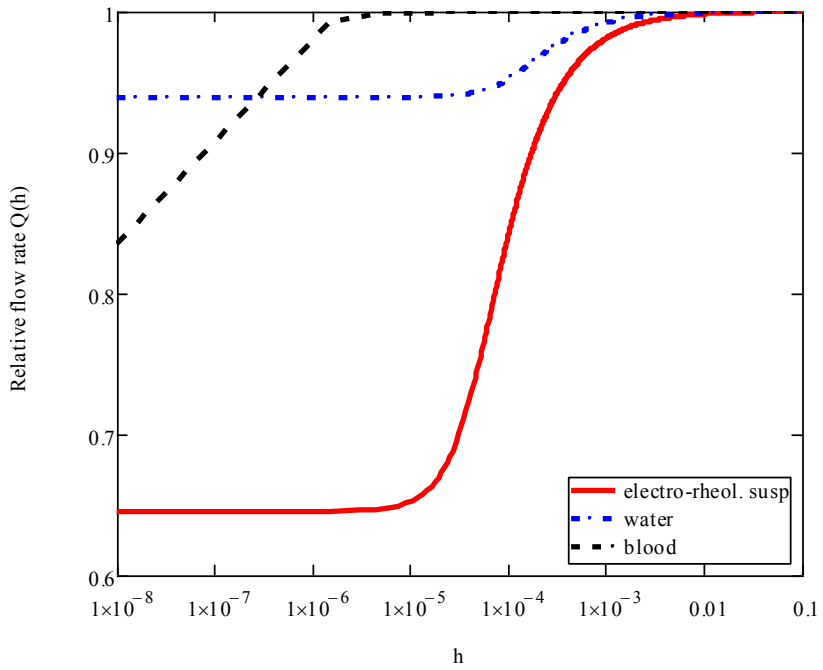



Fig. 2. Relative flow rate $Q(h)$ (6)versus distance $h$ [m] between walls in Hagen–Poiseuille flow

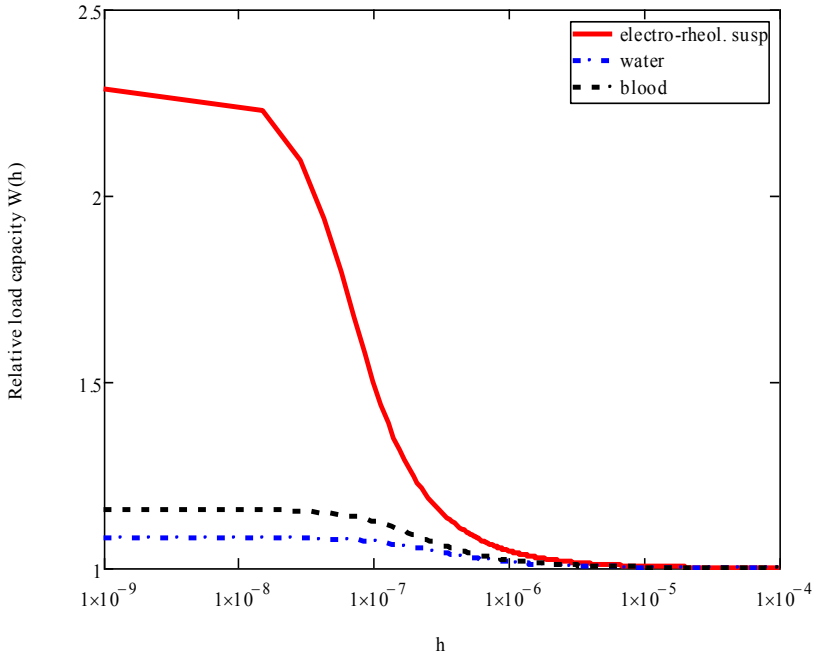



Fig. 3. Relative load capacity *W(h)* (7) versus distance *h* [m] between plates in squeezing flow

It can be seen from data presented in Figure 2 that there exists a flow reduction with microchannels height. Results presented in Figure 3 show that for every fluid, beginning from a certain distance between approaching plates, the load capacity calculated using the micropolar fluid model is bigger than the one calculated with the use of the classical model of the fluid. Comparing relative quantities plots for a given fluid, we observe that values of distances *h* when deviations started are not the same in two different flows. For Hagen-Poisuille the flow rate 1% reduction begins for *h* value greater than the value of *h* for which relative load capacity increment is equal to 1%.

What is more important, it can be observed that every fluid has its "own" value of the limiting distance *h*, starting from which the relative quantities plotted in Figs 2 and 3 begins to change. But comparing the limiting distance *h* for given fluid in two different flows we can observe that the distances differ.

To visualise the observed effect, the relative flow rate defined by formula (6) and the relative load capacity defined by formula (7) are plotted in Figs 4 and 5 in terms of the microstructural parameters *L* and *N, very* often used when compare various micropolar fluid properties. They are defined as follows:

$$N = \sqrt{\frac{\kappa}{2\mu_{\mathrm{N}} + \kappa}}, \quad L = \frac{L_c}{l}, \quad l = \sqrt{\frac{\gamma}{4\mu_{\mathrm{N}}}} \tag{7}$$

where $\mu_N = \mu + \kappa/2$.

Symbol $L_c$ denotes characteristic length for given flow geometry, therefore for considered problems $L_c$ should be replaced by *h* - a distance between walls.

To perform the data comparison, the parameter *N* values are assumed: *N*=0.3, *N*=0.6 and *N*=0.9.

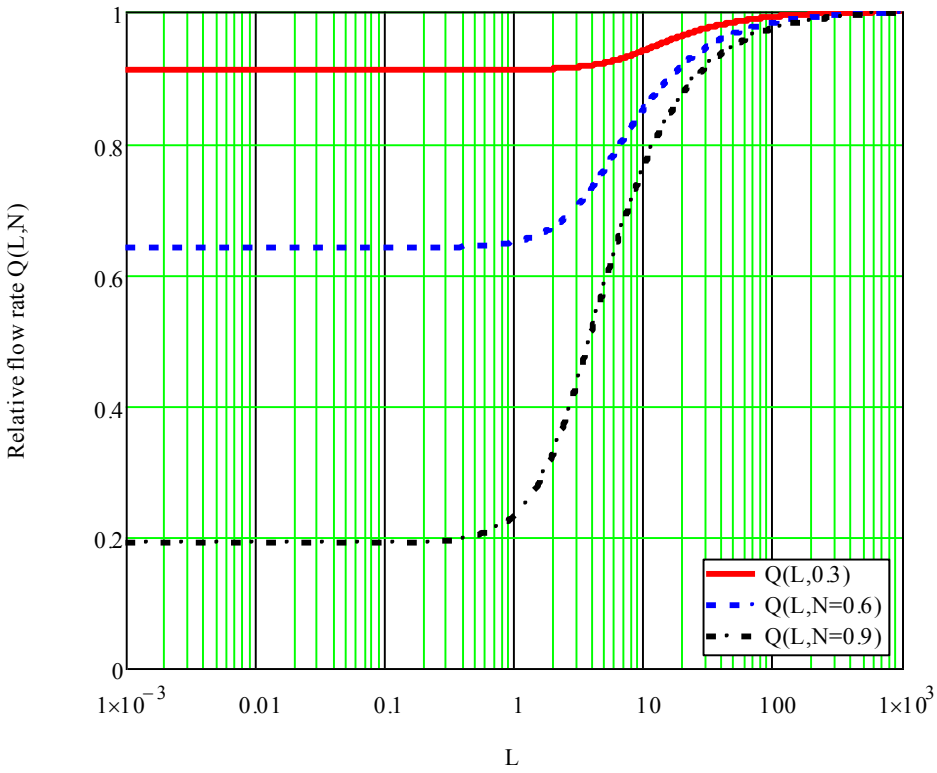



Fig. 4. Relative flow rate Q(L,N) in Hagen–Poiseuille flow versus parameter L

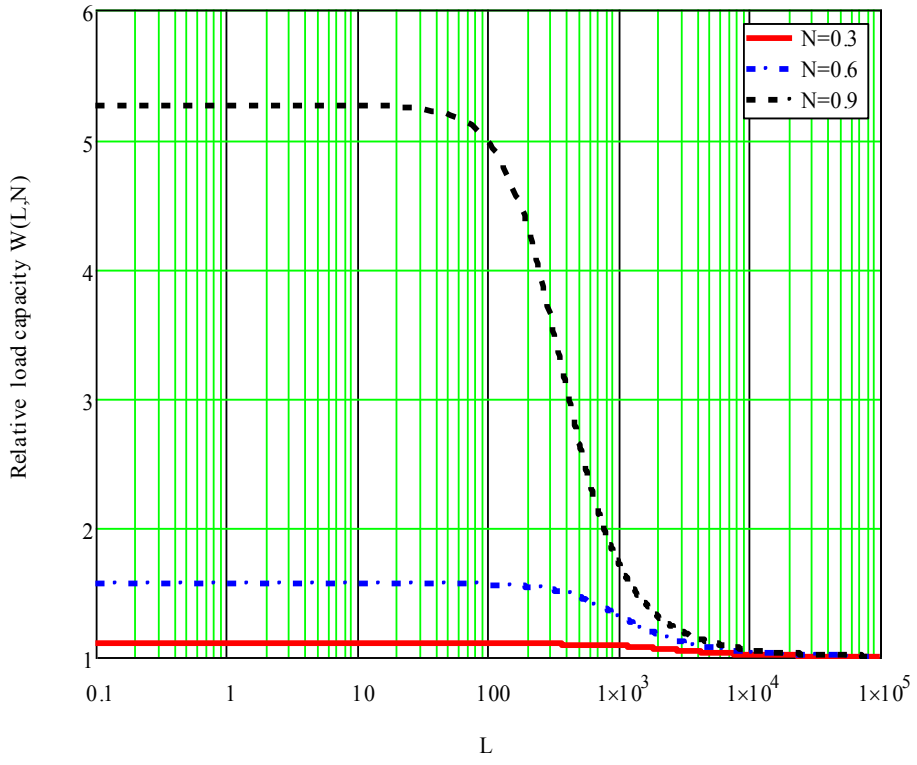



Fig. 5. Relative load capacity W(L,N) in squeezing flow versus parameter L

Comparing plots from Figures 4 and 5 we observe that only for $L$ parameter values satisfying inequality: $L < 10000$ the relative load capacity increases. The flow reduction in the planar Poiseuille flow appears when only $L < 1000$. It means that values of geometrical dimensions of flow field for which the " micropolarity" affect characteristics are different for considered flows. Their values can be easily calculated by use formula (7) and are equal respectively: $h_{max}=10^4\,(\gamma/(4\mu_+2\kappa))^{1/2}$, $h_{max}=10^3\,(\gamma/(4\mu_+2\kappa))^{1/2}$.

## 5. CONCLUSIONS

- Results of the performed analysis prove that sufficient deviations of hydrodynamic quantities from classical results are important on the small scale which depends on rheological properties of the fluid. This means that the coupling between the molecular rotations must be taken into account to give a correct prediction of the flow characteristics on suitable small length scales.
- Length scale for which the fluid micropolar effects are negligible is ten times greater for Hagen–Poiseuille flow then for squeeze flow.
- Comparative study of obtained results for considered flows show that the scale effect depends on the type of the flow.

## REFERENCES

[1] H. Bruus, Theoretical Microfluidics (2008) Oxford, ISBN 978-0-19-923509-4.

[2] A.C. Eringen, (1966), Theory of micropolar fluids, J. Math. Mech. 16, 1, 1-16.

[3] A. Kucaba-Pietal, (2004), Microchannels flow modelling with the micropolar fluid theory, Bulletin of Polish. Academy of Sciences. Techn. Note. 52 (3), s. 209-214, http://bulletin.pan.pl/(52-3)209.pdf.

[4] P. Prokhorenko, N.P. Migoun, M. Stadthaus, (1999), Theoretical Principles of Liquid Penetrant Testing, DVS Verlag, Berlin.

[5] J.S. Hansen, H. Bruus, B.D. Todd, P. Daivis, (2011), Rotational and spin viscosities of water: Application to nanofluidics Phys. Review E 84, 036311.

[6] G. Łukaszewicz, (1999), Micropolar Fluids. Theory and Application, Birkhouser, Berlin.

[7] I. Papautsky, J. Brazzle, T. Ameel, A.B. Frazier, (1999), Laminar fluid behavior in microchannels using micropolar fluid theory, Sensors and Actuators, A, vol. 73, 2, 101-108.

[8] J. Prakash, P. Sinha, (1976), Squeeze film theory for micropolar fluids, Trans. ASME, F98, 1, 139-144.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A COMPARISON OF VARIOUS NODAL DISCONTINUOUS GALERKIN METHODS FOR THE 3D EULER EQUATIONS

**Michael Bergmann, Svetlana Drapkina, Graham Ashcroft, Christian Frey**

German Aerospace Center (DLR),
Institute of Propulsion Technology,
Linder Höhe, 51147 Cologne, Germany
e-mail: {Michael.Bergmann@dlr.de, Svetlana.Drapkina@dlr.de, Graham.Ashcroft@dlr.de,
Christian.Frey@dlr.de}

**Keywords:** Discontinuous Galerkin, Collocation, Spectral Element Methods

**Abstract.** *Recent research has indicated that collocation-type Discontinuous Galerkin Spectral Element Methods (DGSEM) represent a more efficient alternative to the standard modal or nodal DG approaches. In this paper, we compare two collocation-type nodal DGSEM and a standard nodal DG approach in the context of the three-dimensional Euler equations. The nodal DG schemes for hexahedral elements are based on the polynomial interpolation of the unknown solution using tensor product Lagrange basis functions and the use of Gaussian quadrature for integration. In the standard nodal DG approach, we employ uniform interpolation nodes and Legendre-Gauss (LG) quadrature points. The two collocated DGSEM schemes arise from using either LG or Legendre-Gauss-Lobatto (LGL) points as both interpolation and integration nodes. The resulting diagonal mass matrices and the ability to compute the fluxes directly from the solution nodes give rise to highly efficient schemes.*

*The results of the numerical convergence studies highlight, especially at high approximation orders, the performance improvement of the DGSEM schemes compared to the standard DG scheme. Although having advantages in the evaluation of the boundary values over the LG-DGSEM, the lower degree of precision of the LGL quadrature negates this benefit. In addition, without the application of filtering techniques or over-integration, the lower integration accuracy of the LGL-DGSEM leads to numerical instabilities at stagnation points. Hence, the LG-DGSEM is found to be the most efficient scheme as it is more accurate and robust for the considered test cases.*

## 1 INTRODUCTION

Through the permanently growing standards of modern turbomachines with regard to emission and noise pollution, there is a great need for new and more efficient engine technologies. To accurately investigate such technologies, more efficient, flexible and precise numerical tools for the simulation of turbomachinery flows are essential.

In this regard, the discontinuous Galerkin (DG) finite element method has become a very popular method for the computation of unsteady flows as it combines an arbitrary order of accuracy with local data and algorithmic structures. In contrast to the standard finite element method, the polynomial interpolating functions are not restricted by continuity requirements between adjacent elements, cf. [12]. Due to the absence of global continuity, solutions at element interfaces are obtained by techniques similar to those used in upwind finite volume methods, leading to a stable finite element formulation even for advection-dominated problems. As a consequence of their combined robustness, flexibility and accuracy, DG methods are generally more computationally intensive and have higher storage requirements, cf. [1]. The necessity of high-order accurate representation of curved boundaries, as pointed out by [3], enhances both drawbacks.

As an efficient alternative to the standard nodal or modal approaches, spectral collocate forms of the DG space discretization on hexahedral elements have recently been developed, cf. [1, 2, 10, 16]. In these approximations, tensor product Lagrange polynomials are used to interpolate the solutions and Gaussian quadrature for the integrals. Applying the same points as interpolation and quadrature nodes, i.e. collocate them, gives rise to highly efficient schemes. Unfortunately, this choice limits the maximal polynomial degree, which can be exactly integrated. Therefore, numerical instabilities can occur as a consequence of the insufficient quadrature for handling nonlinearities of the governing equations and high-order geometries, cf [2,14].

In this paper, we compare three nodal DG schemes in the context of the three-dimensional Euler equations. In the first scheme, we employ uniform interpolation nodes and Legendre-Gauss (LG) quadrature points. As we don't exploit the collocation, this scheme serves as a reference nodal implementation. The two collocated DG methods arise from using LG or Legendre-Gauss-Lobatto (LGL) points as interpolation and integration nodes, respectively. No over-integration or modal-filtering will be used in any scheme. All DG schemes investigated have been integrated into DLR's in-house CFD code TRACE, which is developed at DLR's Institute of Propulsion Technology, cf. [9].

This paper is organized as follows: In Section 2, the governing equations and the DG methods investigated in this work are presented. In Section 3, the methods are then applied to two inviscid test cases. The last section summarizes the results.

## 2 DISCONTINUOUS GALERKIN APPROXIMATION

### 2.1 Governing equations

Three-dimensional compressible inviscid flow is described by the Euler equations, which can be written in divergence form as

$$\frac{\partial \boldsymbol{q}}{\partial t} + \nabla \cdot \boldsymbol{F}(\boldsymbol{q}) = 0, \tag{1}$$

with suitable initial and boundary conditions. The Cartesian components of the conservative state vector $\boldsymbol{q}$ and flux vectors $\langle \boldsymbol{F}(\boldsymbol{q}), \boldsymbol{n} \rangle = \boldsymbol{F^n}(\boldsymbol{q})$ are

$$\boldsymbol{q} = \begin{pmatrix} \rho \\ \rho \boldsymbol{u} \\ \rho E \end{pmatrix}, \qquad \boldsymbol{F^n}(\boldsymbol{q}) = \begin{pmatrix} \rho \langle \boldsymbol{u}, \boldsymbol{n} \rangle \\ \rho \boldsymbol{u} \langle \boldsymbol{u}, \boldsymbol{n} \rangle + p \boldsymbol{n} \\ \rho \langle \boldsymbol{u}, \boldsymbol{n} \rangle H \end{pmatrix},$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $\boldsymbol{n} = (n_x, n_y, n_z)$ is the outward normal vector to the boundary, $\rho$ is the density, $\boldsymbol{u} = (u, v, w)$ are the Cartesian velocity components, $E$ is the total energy and $H = E + p/\rho$ is the total enthalpy. The pressure is related to the other thermodynamic variables by the equation of state for an ideal gas and can be computed as $p = (\gamma - 1)\rho\,(E - 1/2||\boldsymbol{u}||^2)$, where $\gamma = 1.4$ is the ratio between the specific heats of the fluid.

### 2.2 Discontinuous Galerkin discretization

The DG discretization of the Euler equations is based on the weak formulation, which can be obtained by multiplying Equation (1) by a sufficiently smooth test function $\boldsymbol{v}$ and performing integration by parts,

$$\int_\Omega \left\langle \boldsymbol{v}, \frac{\partial \boldsymbol{q}}{\partial t} \right\rangle d\Omega + \int_{\partial\Omega} \langle \boldsymbol{v}, \boldsymbol{F^n}(\boldsymbol{q}) \rangle \, d\sigma - \int_\Omega \nabla \boldsymbol{v} : \boldsymbol{F}(\boldsymbol{q}) \, d\Omega = 0, \tag{2}$$

where the symbol : denotes the double inner product. The domain $\Omega$ is now subdivided into shape-regular meshes $\mathcal{T}_h = \{K\}$ consisting of non-overlapping elements $K$, where $h$ denotes a piecewise constant mesh function. Hereafter, the solution $\boldsymbol{q}$ and test function $\boldsymbol{v}$ are approximated as piecewise polynomial functions $\boldsymbol{q}_h$ and $\boldsymbol{v}_h$, which are discontinuous across the elements. The space of piecewise polynomials is defined as,

$$V_h = \{v_h \mid v_h \in L^2(\Omega),\ v_h \left(M^K(\boldsymbol{\xi})\right) \in \mathbb{P}^p(K_{\text{ref}}),\ \forall K \in \mathcal{T}_h\}, \tag{3}$$

where the solution vector and the test function belong to $\boldsymbol{q_h}, \boldsymbol{v_h} \in (V_h)^5$. $\mathbb{P}^p(K_{\text{ref}})$ denotes the space of tensor product polynomials up to degree $p$ on a reference element $K_{\text{ref}}$ and $M^K$ is a polynomial continuous invertible mapping for each element $K \in \mathcal{T}_h$ to a reference element $K_{\text{ref}}$ with $\boldsymbol{x} = M^K(\boldsymbol{\xi})$, where $\boldsymbol{x} = (x, y, z)$ are the Cartesian coordinates and $\boldsymbol{\xi} = (\xi, \eta, \zeta)$ the normalized coordinates of the reference element $K_{\text{ref}}$, cf. [3, 6, 9].

By substituting $\boldsymbol{q}$ with $\boldsymbol{q}_h$ and $\boldsymbol{v}$ with $\boldsymbol{v}_h$, the weak formulation (2) for each element $K \in \mathcal{T}_h$ can be rewritten as,

$$\int_K \left\langle \boldsymbol{v}_h, \frac{\partial \boldsymbol{q}_h}{\partial t} \right\rangle d\boldsymbol{x} + \int_{\partial K} \left\langle \boldsymbol{v}_h, \mathcal{H}(\boldsymbol{q}_h^+, \boldsymbol{q}_h^-, \boldsymbol{n}) \right\rangle ds - \int_K \nabla_h \boldsymbol{v}_h : \boldsymbol{F}(\boldsymbol{q}_h) \, d\boldsymbol{x} = 0. \tag{4}$$

Due to discontinuities of the solution $\boldsymbol{q}_h$ at element interfaces, the normal flux vector $\boldsymbol{F^n}(\boldsymbol{q})$ in the boundary integrals is replaced by a normal numerical flux function $\mathcal{H}(\boldsymbol{q}_h^+, \boldsymbol{q}_h^-, \boldsymbol{n})$, which depends on the outward normal vector $\boldsymbol{n}$, the internal interface state $\boldsymbol{q}_h^-$ and the interface state of the neighboring element $\boldsymbol{q}_h^+$. In the current DG code the Roe numerical flux is employed, cf. [6].

### 2.3 Polynomial approximation and numerical integration

The approximated solution $\boldsymbol{q}_h$ and test function $\boldsymbol{v}_h$, belonging to the approximation space $(V_h)^5$, can be expressed as their polynomial expansion,

$$\boldsymbol{q}_h(\boldsymbol{\xi},t) = \sum_{i=1}^{N(p)} \boldsymbol{q}_i(t)\phi_i(\boldsymbol{\xi}), \quad \boldsymbol{v}_h(\boldsymbol{\xi}) = \sum_{j=1}^{N(p)} \boldsymbol{v}_j\phi_j(\boldsymbol{\xi}), \quad \forall \boldsymbol{\xi} \in K_{\text{ref}}. \tag{5}$$

Here, $\phi_i(\boldsymbol{\xi})$ denotes the basis function of the polynomial space $\mathbb{P}^p(K_{\text{ref}})$, $N(p)$ is the number of interpolation points, and $\boldsymbol{q}_i(t)$ and $\boldsymbol{v}_j$ are the expansion coefficients of $\boldsymbol{q}_h$ and $\boldsymbol{v}_h$, respectively. In this work, we consider nodal approximations and, therefore, using Lagrange interpolating polynomials as basis functions. In one space dimension, they are defined on the reference element as follows,

$$l_i(\xi) = \prod_{j=1, j\neq i}^{p+1} \frac{\xi - \xi_j}{\xi_i - \xi_j}, \qquad i = 1, \ldots, (p+1), \tag{6}$$

where $\xi_j$ are the node coordinates. One can easily observe that for Lagrange polynomials

$$l_i(\xi_j) = \delta_{ij}, \tag{7}$$

holds, where $\delta$ is the Kronecker delta. For hexahedral elements, the nodal basis functions are the tensor product of one-dimensional Lagrange polynomials,

$$\phi_i(\boldsymbol{\xi}) = \phi_{mno}(\xi,\eta,\zeta) = l_m(\xi)l_n(\eta)l_o(\zeta). \tag{8}$$

Considering the property of the Lagrange polynomials (7), the unknown expansion coefficients $\boldsymbol{q}_i(t)$ correspond to the solution values at the specific node $\boldsymbol{q}_h(\boldsymbol{\xi}_i, t)$, cf. [2].

Mapping Equation (4) into the reference space, substituting $\boldsymbol{q}_h$ by (5) and testing only against the basis functions $\phi_j(\boldsymbol{\xi})$ for all $j = 1, \ldots, N(p)$, yields,

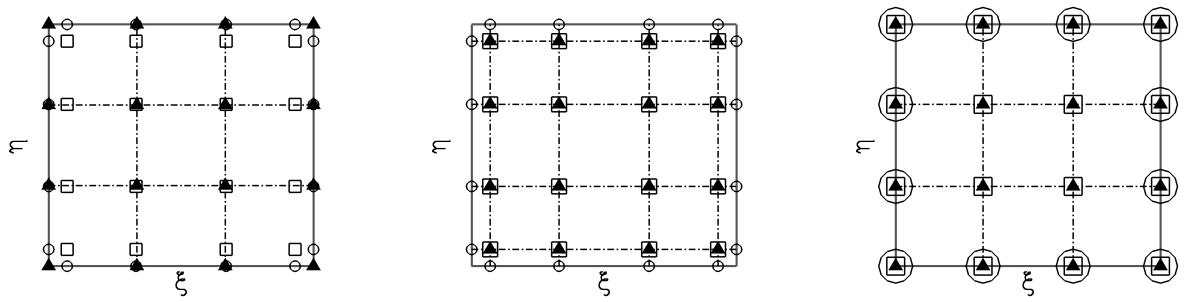
$$\frac{\partial \boldsymbol{q}_i(t)}{\partial t} \int_{K_{\text{ref}}} \phi_j\phi_i |\boldsymbol{J}_K(\boldsymbol{\xi})| d\boldsymbol{\xi} + \int_{\partial K_{\text{ref}}} \phi_j \mathcal{H}(\boldsymbol{q}_h^+, \boldsymbol{q}_h^-, \boldsymbol{n}) S_K(\boldsymbol{\xi})\, d\hat{s} - \int_{K_{\text{ref}}} \left\langle \left(\boldsymbol{J}_K(\boldsymbol{\xi})\right)^{-1} \nabla_\xi \phi_j, \boldsymbol{F}(\boldsymbol{q}_h) \right\rangle |\boldsymbol{J}_K(\boldsymbol{\xi})| d\boldsymbol{\xi} = 0, \tag{9}$$

where $|\boldsymbol{J}_K(\boldsymbol{\xi})|$ and $\left(\boldsymbol{J}_K(\boldsymbol{\xi})\right)^{-1}$ denotes the determinant and inverse of the Jacobian matrix, respectivly and $S_K(\boldsymbol{\xi})$ is the curved face area, cf. [9]. By assembling all element contributions, Equation (9) can be rewritten in the compact form as,

$$\boldsymbol{M}\frac{\partial \boldsymbol{q}}{\partial t} + \boldsymbol{R}(\boldsymbol{q}) = 0. \tag{10}$$

Here, $\boldsymbol{M}$ denotes the mass matrix, $\boldsymbol{q}$ the unknown degrees of freedom (DOF) and $\boldsymbol{R}$ the residual vector, containing the boundary and volume integrals. The integrals are carried out using Gaussian quadrature formulae. For example, the mapped volume integral is approximated as,

$$\int_{K_{\text{ref}}} \left\langle \boldsymbol{J}_K^{-1}(\boldsymbol{\xi}) \nabla_\xi \phi_j(\boldsymbol{\xi}), \boldsymbol{F}(\boldsymbol{q}_h(\boldsymbol{\xi}, t)) \right\rangle |\boldsymbol{J}_K(\boldsymbol{\xi}))| d\boldsymbol{\xi} \approx \sum_{i=1}^{M} \left\langle \boldsymbol{J}_K^{-1}(\boldsymbol{\xi_i}) \nabla_\xi \phi_j(\boldsymbol{\xi_i}), \boldsymbol{F}(\boldsymbol{q}_h(\boldsymbol{\xi_i}, t)) \right\rangle |\boldsymbol{J}_K(\boldsymbol{\xi}_i)| w_i, \tag{11}$$




(a) DGNODAL (b) LG-DGSEM (c) LGL-DGSEM

Figure 1: Interpolation and quadrature nodes on the reference quadrilateral for an approximation of order $p = 3$. The black triangles denote the interpolation nodes, the squares the volume quadrature points and the circles the boundary quadrature points of the specific scheme.

where $\boldsymbol{\xi}_i = (\xi_i, \eta_i, \zeta_i)$ and $w_i$ denotes the quadrature coordinates and weights, respectively. It is now possible to choose various sets of interpolation and quadrature nodes in order to derive different nodal DG formulations. Note, for all schemes presented, the number of quadrature points is set to be equal to the number of interpolation points, e.g. $M = N(p) := (p+1)^3$ for the three-dimensional case.

In the first scheme, we employ equidistant interpolation nodes and LG points as integration nodes. Note, although equidistant interpolation nodes can lead to unstable formulations for very high-order schemes, they are viable for moderate approximation orders, cf. [8, 11]. The degree of accuracy of the Gaussian quadrature with LG quadrature nodes is $2M - 1$, which means that polynomials up to degree $2M - 1$ are integrated exactly. In the following, this formulation is called DGNODAL.

For the purpose of efficiency, we consider nodal collocation-type approximations in the next schemes, i.e. using the same points as interpolation and integration nodes. As a result, many numerical operations can be omitted, which give rise to highly efficient implementation, cf. [2, 15, 16]. Two collocation formulations arise from using LG points or LGL points as both interpolation and quadrature nodes. With LG nodes, the degree of accuracy of the Gaussian quadrature is the same as in the DGNODAL formulation. Whereas, the degree of accuracy of the Gaussian quadrature with LGL nodes is only $2M - 3$. Hence, the inner product of two polynomial functions of order $p$, e.g. appearing in the mass matrix, is not integrated exactly, cf. [13, 16, 18]. The collocation formulations are named in the following as LG-DGSEM and LGL-DGSEM, respectively.

In order to clarify the DG schemes investigated in this work, the volume and boundary quadrature points and interpolation nodes for the two-dimensional case are shown in Figure 1. The DGNODAL scheme represents the reference case in which all numerical operations are performed. Therefore, the solution and the derivatives have to be interpolated onto all the volume and boundary quadrature points. Additionally, the resulting element mass matrix is a dense matrix. Taking the interpolation nodes as integration nodes and considering the property of Lagrange polynomials (7), the mass matrix becomes diagonalized in both DGSEM schemes. Furthermore, the solution is directly available at the quadrature points of the volume integral and does not have to be interpolated. Through tensor product basis functions, the derivatives

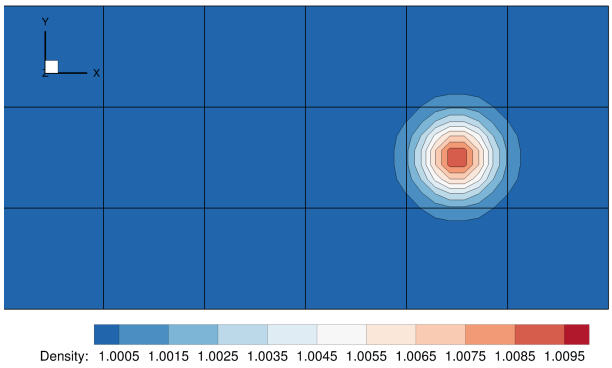



Figure 2: Gaussian pulse. Density contours of $\mathbb{P}^7$ solution on a coarse grid at $z = 0.5$.

can be evaluated at the volume quadrature points using an one-dimensional interpolation. The disadvantage of the LG-DGSEM formulation, compared to the LGL-DGSEM, is however that boundary values have to be interpolated additionally, see Figures 1b and 1c.

### 2.4 Explicit time integration

An explicit three-stage third order accurate Runge-Kutta method is used to integrate Equation (10) in time. The scheme can be written as follows, cf. [6, 9],

$$\begin{aligned} \boldsymbol{q}^{(1)} &= \boldsymbol{q}^n + \Delta t \boldsymbol{R}_M(\boldsymbol{q}^n, t^n), \\ \boldsymbol{q}^{(2)} &= \frac{3}{4}\boldsymbol{q}^n + \frac{1}{4}\Big(\boldsymbol{q}^{(1)} + \Delta t \boldsymbol{R}_M((\boldsymbol{q}^{(1)}, t^n + \Delta t)\Big), \\ \boldsymbol{q}^{n+1} &= \frac{1}{3}\boldsymbol{q}^n + \frac{2}{3}\Big(\boldsymbol{q}^{(2)} + \Delta t \boldsymbol{R}_M((\boldsymbol{q}^{(2)}, t^n + \frac{1}{2}\Delta t)\Big), \end{aligned} \tag{12}$$

with $\boldsymbol{R}_M = \boldsymbol{M}^{-1}\boldsymbol{R}$. Applied to the DG system with elements of order $p$ this method has been shown to be linearly stable for a Courant number less than or equal to $\frac{1}{2p+1}$, cf. [7].

## 3 Numerical Results

### 3.1 Gaussian pulse in density

In order to verify the formal order of accuracy of the DG schemes in practice, we consider a test case with an analytical solution. Therefore, a three-dimensional Gaussian density fluctuation is initialized in a cuboid domain with straight sided edges, cf. [17]. The pressure and velocity fields are uniform. The initial distribution is as follows,

$$\boldsymbol{q}_0 = \begin{pmatrix} \rho_0 + a\ \exp\left[-\ln\left(2\frac{(\boldsymbol{x}-\boldsymbol{x}_0)^2}{\sigma^2}\right)\right] \\ \rho_0 \boldsymbol{u}_0 \\ \frac{p_0}{\gamma - 1} + \frac{\rho_0 ||\boldsymbol{u}_0||^2}{2} \end{pmatrix},$$

where $\boldsymbol{x}_0$ denotes the initial peak coordinates, $\sigma$ the width of the fluctuation and $a$ the amplitude. The analytical solution can obtained by moving the Gaussian pulse with the specified convective velocity $\boldsymbol{u}_0$. In Figure 2, the density contours of the $\mathbb{P}^7$ solution are displayed for a coarse hexahedral mesh. The convergence rates of the schemes are investigated by performing computations on increasingly fine grids. To ensure that the finite temporal accuracy of the time

integration method does not influence the results, a small constant time-step size has be used in all simulations. In Figures 3a, 3b and 3c, the $L^2$-errors of the outlined schemes are plotted against the characteristic mesh size $h$, i.e. the element edge length. The black dash-dot lines illustrate the formal convergence rate $\mathcal{O}(h^{p+1})$ of the DG method. It can be seen clearly that the schemes achieve the expected order of convergence.

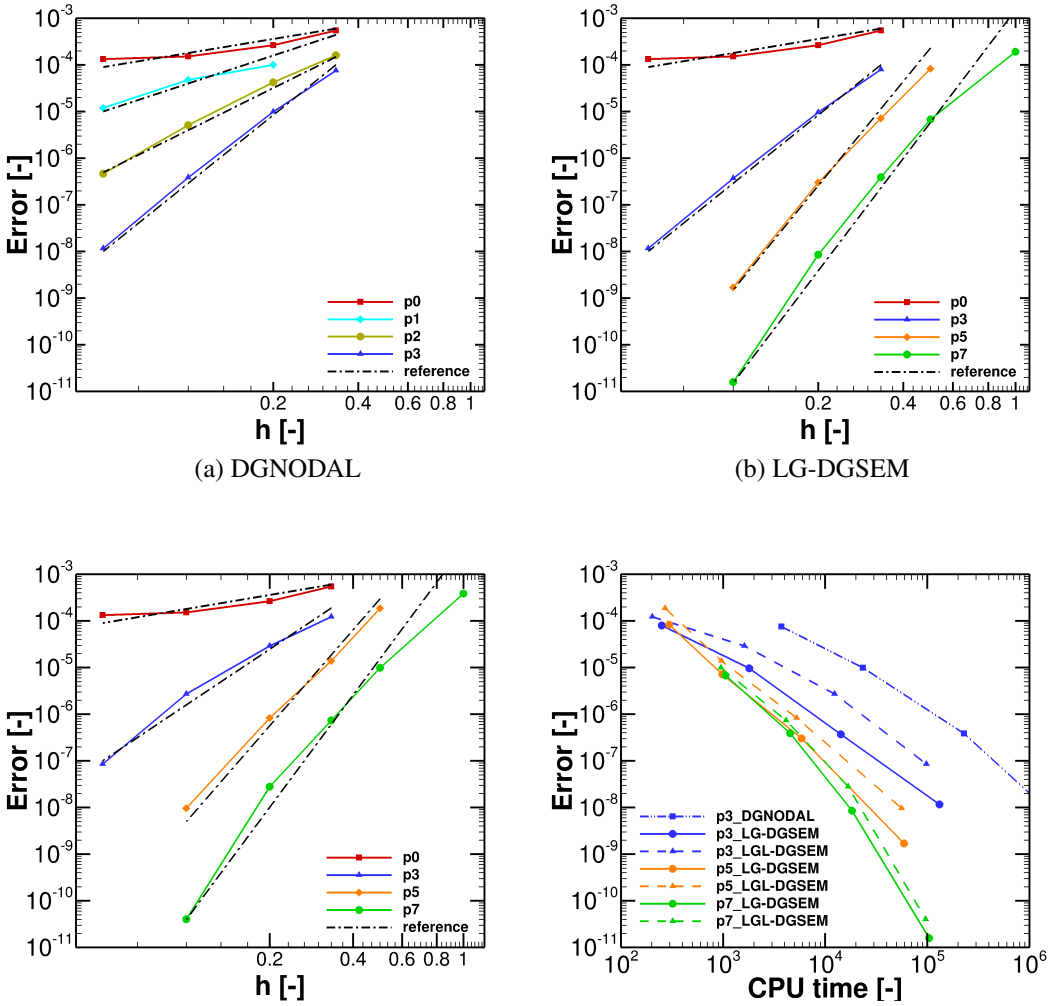
To compare the efficiency, the $L^2$-error is plotted against the run time in Figure 3d. The solid, dashed and dash-dot lines indicate the LG-DGSEM, the LGL-DGSEM and the DGNODAL implementations, respectively. It can generally be observed that the higher-order schemes outperform the lower-order ones beyond a certain accuracy level. Looking at the $\mathbb{P}^3$ results, the theoretical efficiency disadvantage of DGNODAL, compared to both DGSEM schemes, is also clearly evident. Furthermore, the results show that the LG-DGSEM is the most efficient scheme for the inspected approximation orders. Therefore, the higher accuracy of the Gaussian quadrature seems to offset the cost of the additional interpolation required at the element boundaries. However, as the efficiency advantage of the LG-DGSEM decreases with increasing approximation order, this implies that the benefit of the higher integration order decreases relative to the cost of the additional interpolation.

### 3.2 Flow around a NACA0012 airfoil

To investigate the behaviour of the schemes for flows with stagnation points and the usage of high-order boundaries, the inviscid flow around a NACA0012 airfoil is considered. The freestream Mach number is set to $M_\infty = 0.3$ and an incidence angle equal to $\alpha = 1.49°$ is applied. The computations are performed on a coarse mesh containing $200$ hexahedral elements with a single element in the spanwise direction. At the inlet and outlet boundaries Riemann boundary conditions are imposed. The profile geometry is represented by high-order elements of order $q = 3$, which are generated using the process chain introduced in [9]. In Figure 4a, the Mach contours of the $\mathbb{P}^3$ solution of the LG-DGSEM scheme are presented. Same results can be obtained using the DGNODAL implementation. However, computations with LGL-DGSEM turn out be unstable. These instabilities occur near the stagnation point of the leading edge. In [2], *Bassi et. al.* identified the numerical under-integration of the linear terms as the most critical reason for this phenomenon with the decoupling of the DOF at stagnation points being only the trigger of the numerical instabilities at lower orders. By using LGL interpolation and integration nodes with over-integration, i.e. $M = N + 1$ and, therefore, applying non-collocation approximations, both possible reasons are nullified and a stable scheme for the test case can be obtained. Nevertheless, without collocation, the efficiency of the scheme is heavily impaired. Notice that this is just a brief test and has to be further investigated and optimized, cf. for example [4]. In Figure 4b, the pressure coefficient distributions on the profile of the $\mathbb{P}^3$ solutions of DGNODAL, LG-DGSEM and the over-integrated LGL scheme, respectively, are plotted. All schemes show very similar distributions.

In Figure 5a, the pressure coefficient distributions of various polynomial approximation orders, computed with LG-DGSEM, are plotted. The mesh and the geometrical approximation order is the same in all presented simulations, i.e. $q = 3$. It can be observed that the distributions converge with increasing polynomial approximation order.

As a last observation, the pressure coefficient distribution of $\mathbb{P}^5$ LG-DGSEM is compared with solutions obtained with the second-order accurate finite volume (FV) solver in TRACE, cf. [5]. For this comparison two simulations have been performed with the FV solver: one on the coarse DG mesh with $200$ elements and a second on a fine mesh with $9800$ elements. The pressure coefficient distributions are illustrated in Figure 5b. As one can notice, there are




(a) DGNODAL (b) LG-DGSEM

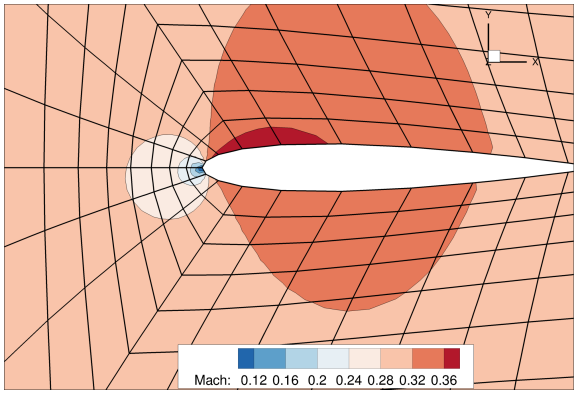
(c) LGL-DGSEM (d) CPU time behaviour

Figure 3: Gaussian pulse. Convergence rates and efficiency rates of the outlined schemes.

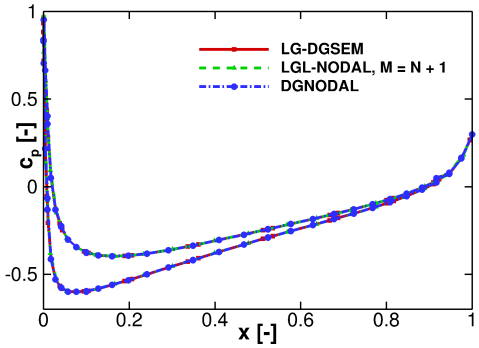
large differences between the FV solution on the coarse mesh and the other results. Moreover, the distribution of the fine FV simulation and $\mathbb{P}^5$ LG-DGSEM are very similar to each other. The slight differences are most likely due to the not fully converged mesh used for the fine FV simulation. However, these results suggest that the proposed DG scheme has the potential of computing highly accurate solutions.

## 4 Conclusions

In this paper, a standard nodal DG approach and two collocation-type nodal DG formulations are compared in the context of the compressible, three-dimensional Euler equations. All schemes considered show the optimal order of accuracy $\mathcal{O}(h^{p+1})$ for an approximation order $\mathbb{P}^p$, whereby both collocation schemes have huge advantages in efficiency compared to the standard nodal DG approach. Matching both collocation methods, LG-DGSEM seems to be more efficient than LGL-DGSEM, although the difference between both schemes decreases with increasing polynomial approximation order as the influence of the lower integration order of LGL quadrature compared to the more efficient boundary value evaluation decreases. Furthermore,

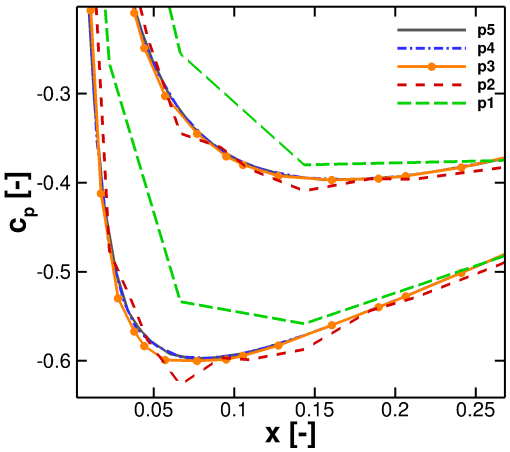



(a) LG-DGSEM

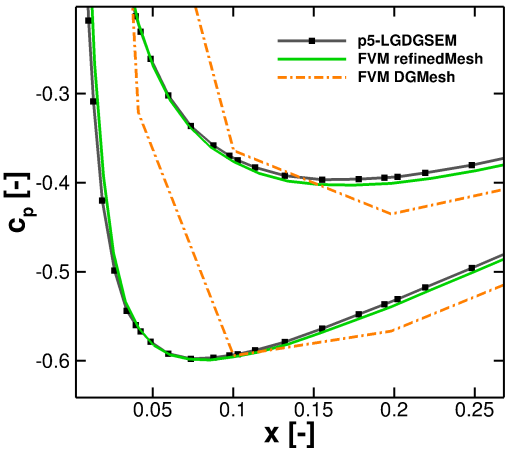



(b) Pressure coefficient distribution on the profile with $q = 3$ of $\mathbb{P}^3$ solutions

Figure 4: NACA0012 airfoil, $M_\infty = 0.3$, $\alpha = 1.49°$, geometric approximation of order $q = 3$. Mach contours of the $\mathbb{P}^3$ solution with LG-DGSEM and pressure coefficient distributions obtained with LG-DGSEM, DGNODAL and the over-integrated, non-collocated nodal LGL formulation.




(a) LG-DGSEM, $p$-refinement, $q = 3$




(b) $\mathbb{P}^5$ LG-DGSEM and 2nd order accurate FV solver

Figure 5: NACA0012 airfoil, $M_\infty = 0.3$, $\alpha = 1.49°$. Pressure coefficient distributions. (a) LG-DGSEM with various polynomial approximation orders; (b) LG-DGSEM with $\mathbb{P}^5$ (solid black line with squares), 2nd order accurate FV scheme computed on the DG mesh, i.e. 200 hexahedral elements (orange dash-dot line) and on a refined mesh containing 9800 elements (green solid line), respectively.

the LGL-DGSEM, without applying over-integration or modal filter, becomes unstable at the stagnation point in the presented steady flow around a NACA0012 airfoil, whereas LG-DGSEM and DGNODAL remain stable. Hence, the LG-DGSEM is found to be the most efficient scheme as it is more accurate and robust for the considered test cases. Finally, comparisons with an extensively validated FV solver indicate the DG method is able to provide accurate results.

## REFERENCES

[1] F. Bassi, N. Franchina, A. Ghidoni, and S. Rebay. Spectral p-multigrid discontinuous Galerkin solution of the Navier-Stokes equations. *International Journal for Numerical Methods in Fluids*, 67(11):1540–1558, 2011.

[2] F. Bassi, N. Franchina, A. Ghidoni, and S. Rebay. A numerical investigation of a spectral-type nodal collocation discontinuous Galerkin approximation of the Euler and Navier-Stokes equations. *International Journal for Numerical Methods in Fluids*, 71(10):1322–1339, 2013.

[3] F. Bassi and S. Rebay. High-order accurate discontinuous solution of the 2D Euler equations. *Journal of computational physics*, 138:251–285, 1997.

[4] A. Beck. *High order discontinuous Galerkin methods for the simulation of multiscale problems*. PhD thesis, Faculty of Aerospace Engineering and Geodesy of the University of Stuttgart, 2015.

[5] K. Becker, K. Heitkamp, and E. Kügeler. Recent progress in a hybrid-grid CFD solver for turbomachinery flows. In *Proceedings Fifth European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010*, 2010.

[6] S. Cherednichenko, C. Frey, and G. Ashcroft. On the application of the discontinuous Galerkin method to turbomachinery flows. In *6th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012)*, pages 2359–2375, September 2012.

[7] B. Cockburn and C. Shu. Runge-Kutta discontinuous Galerkin methods for convection-dominated problems. *Journal of Scientific Computing*, 16(3):173–261, 2001.

[8] M. O. Deville, P. F. Fischer, and E. H. Mund. *High-Order Methods for Incompressible Fluid Flow*. Cambridge University Press, 2002.

[9] S. Drapkina, C. Frey, and G. Ashcroft. On the integration of high-order boundary elements in a 3d discontinuous Galerkin method for turbomachinery flows. In *European Conference on Computational Fluid Dynamics - ECFD VI*, Juli 2014.

[10] F. Giraldo, J. Hesthaven, and T. Warburton. Nodal high-order discontinuous Galerkin methods for the spherical shallow water equations. *Journal of Computational Physics*, 181(2):499 – 525, 2002.

[11] J. S. Hesthaven and T. Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications.* Texts in applied mathematics. Springer, 2008.

[12] K. Hillewaert, N. Chevaugeon, P. Geuzaine, and J.-F. Remacle. Hierarchic multigrid iteration strategy for the discontinuous Galerkin solution of the steady Euler equations. *International Journal for Numerical Methods in Fluids*, 51(9-10):1157–1176, 2006.

[13] G. Karniadakis and S. Sherwin. *Spectral/hp element methods for computational fluid dynamics*. Oxford University Press, 2nd edition, 2005.

[14] R. M. Kirby and G. E. Karniadakis. De-aliasing on non-uniform grids: algorithms and applications. *Journal of Computational Physics*, 191(1):249 – 264, 2003.

[15] D. A. Kopriva. *Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[16] D. A. Kopriva and G. Gassner. On the quadrature and weak form choices in collocation type discontinuous Galerkin spectral element methods. *Journal of Scientific Computing*, 44(2):136–155, 2010.

[17] B. Landmann. *A parallel discontinuous Galerkin code for the Navier-Stokes and Reynolds-averaged Navier-Stokes equations*. PhD thesis, Faculty of Aerospace Engineering and Geodesy of the University of Stuttgart, 2009.

[18] S. A. Teukolsky. Short note on the mass matrix for Gauss-Lobatto grid points. *Journal of Computational Physics*, 283:408–413, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A 3D ISOGEOMETRICAL BOUNDARY ELEMENT ANALYSIS FOR NONLINEAR GRAVITY WAVE PROPAGATION

**Jorge Maestre**[1], **Jordi Pallarés**[1], **Ildefonso Cuesta**[1]

[1]Departament d'Enginyeria Mecànica, Universitat Rovira i Virgili, Av. Països Catalans, 26, 43007 Tarragona, Spain.
{jorge.maestre,jordi.pallares,ildefonso.cuesta}@urv.cat

**Keywords:** BEM, T-spline, NURBS, Gravity wave, Time domain, non-linear waves

**Abstract.** *In this paper we describe a three-dimensional Isogeometric BEM in the time domain based on the T-spline and Non Uniform Rational B-Splines (NURBS) basis to study the free surface behavior. Traditionally, the Lagrange polynomials have been used to discretize the geometry and the BEMs variables. The Lagrange basis are discontinues across the elements although the physical variables are continuous. In dynamics problems with high mesh distortions this approach can produce numerical instabilities that can be quickly propagated. This problem could be overcome using T-spline or NURB basis. The main advantages of this approach are: (1) the control of the continuity and smoothness of the T-spline and NURBS basis, which makes the model numerically stable without the need of artificial smooth techniques;(2) the high order geometrical approximation by non-rational splines;(3) the refinement capabilities without affecting the geometry and BEM's variables; and (4) the direct integration with computer aid geometrical design tools. We use the concept of the Bézier extraction operation which provides an element point of view of the T-Spline and NURBS similar to the traditional finite element. The boundary integral Equation is solved at each time step by the GMRES algorithm and the time marching scheme is performed with a fourth-order Runge-Kutta method to update the model. Additionally, the hydrodynamic force is calculated by an auxiliary boundary equation. Some numerical benchmark examples are analysed to show the accuracy and the stability of the method*

## 1 INTRODUCTION

In the last decades, numerical models have been developed for the prediction of the propagation of gravity waves in the time domain. These models have been applied to simulated tsunami generation and [1] overturning waves [2–4], to design breakwaters [5, 6], to predict the wave pressure impact on structures [7], or to study radiation and diffraction waves produced by a wave-maker [8–11]. Nowadays, because of the increasing interest on the ocean renewable energy, in which the generation systems are installed near-shore or off-shore, new applications of such models are being used to study the fluid-structure interaction considering the effect of the waves [12–16].

Early numerical studies on the non-linear wave propagation appeared in 1976 when Longuet-Higguin and Cokelet [17] presented their bi-dimensional (2D) approach to simulate the transient surface waves. Their approach was based on the potential flow theory and used the Boundary Element Method (BEM) to solve the Laplace's equation in conjunction with a Mixed Eulerian-Lagrange (MEL) technique to update the free surface. Similar models were used to study a wide variety of non-linear water waves problems [18–22], but it was the study of Dommermuth et al. [23] that demonstrated the validity of the potential flow theory applied to the unsteady gravity wave propagation. These authors presented a successful comparison between the numerical results obtained by a 2D-BEM model and experimental results in a water tank. It is worth to mention the studies reported by Grilli et al. [2, 24, 25] who developed an advanced approach using 2D high order BEM. Later, Grilli et al. [3] and Guyenne and Grilli [4] extended the model to three-dimensional (3D) geometries to analyse overturning waves over an arbitrary bottom.

Some full non-linear high-order 3D BEM models have been proposed to simulate the unsteady interaction of the wave with a rigid body. An overview of such works can be found in Tanizawa [26]. Among them, Lee et al. [27] studied the non-linear waves and the hydrodynamic force generated by the movement of a submerged sphere. Bai and Taylor [10] investigated the wave radiation produced by a moving submerged truncated cylinder and the wave diffraction around a vertical cylinder [28]. They used an unstructured triangular mesh constructed with second order Lagrange elements in order to obtain a good approximation of the geometry. Later, these authors extended the model to flared floating structures [29]. Sung and Grilli [30] analysed the waves generated by an advancing surface disturbance. Following the Grilli's works, these authors used a structured mesh formed by isoparametric bi-cubic piecewise overlapping elements that provided a local smoothness of the geometry and of the physical variables. More recently, Hannan et al. [31, 32] studied the interactions between water waves and fully submerged fixed or moving structures. In the same line, but limited to 2D, Dombre et al. [33] extended the early work presented in [34] to study the dynamics of free fully submerged structures.

In addition, other models have been proposed to analyse more specific problems such as the post-breaking phenomenon, violent wave impact against structures, large fluid movement in confined space with steep non-linear waves or viscous and vortex force in the context of floating structure. In such cases, the potential flow theory is not valid and other tools based on Computational Fluid Dynamics (CFD) solvers to resolve the full Navier-Stokes equations or Lattice Boltzmann method (LBM) have been used [35–40]. Most of these problems have been formulated in 2D, and despite the increase of computational power, the full 3D models require a considerable computational cost. Therefore, unless in such specific cases, the potential theory together with BEM provides good results for non-linear gravity waves problems with reduced computational time. Some advantages are: (1) the reduced dimensionality of the domain; (2)

the ability to handle complex 3D geometric models; (3) and the Sommerfeld radiation condition is implicitly satisfied in the Green function.

In the last two decades, the Non-Uniform Rational B-Spline (NURBS) representation has been widely used in the marine industry through of the modern Computational Aid Design (CAD) systems. NURBS exhibits multiple advantages, such as: (1) the quadratic surface can be represented exactly; (2) advanced construction techniques and shape modification tools are available (3) the smoothness of the basis can be managed via knot; and (4) h, p and k-refinement can be applied without modifying the geometry. These characteristics, in combination with other mathematical properties, give to NURBS the ability to handle complex geometries and provide suitable basis for numerical analyses. In this line, the incorporation of NURBS into the BEM to simulate wave-structure interaction has appeared recently in the literature (see for example references [41–46]). These approaches use two different basis functions, one to represent the geometry and the other for physical variables. Hughes et al. [47, 48] and Cottrell et al. [49, 50] introduced the concept of the Isogeometrical Analysis (IGA) in the context of Finite Element Method (FEM) that consists on approximate the full model by the same basis functions used for the geometry. This concept, that has acquired enough maturity in recent years, allows a direct link between the CAD geometry and the engineering analysis tools. Politis et al. [51] extended the IGA to the BEM and applied it to a 2D external potential flow problem. Simpson et al. [52] applied an IGA-BEM method to a 2D elastoestatic problem. More recently, Belibassakis et al. [53] presented a 3D IGA-BEM model based on the Neumann-Kelvin problem to study the ship wave resistance and showed a novel combination of a modern CAD system with hydrodynamic solvers.

Although the NURBS is widely used by the CAD industry, it has some limitations associated with the fact that only surfaces with four edges can be represented. This entails that complicated geometries require many NURBS patches and the smoothness between patches is lost. The T-spline technology has been introduced to overcome these limitations [54]. T-splines are a generalization of the NURBS, in which several NURBS patches can be integrated into a unique T-spline. Moreover, unlike the NURBS, T-splines allow local refinements and unstructured meshes this makes their use attractive for the engineering analyses. In this direction, Scott et al. [55] and Li et al. [56] formulated the analysis-suitable T-spline in which necessary mathematical properties of the basis (as linear independence) are guaranteed to be used in IGA. However, the implementation of T-spline in the IGA requires a more advanced development. Adopting the idea of the Bézier extraction operation, presented by Borden et al. [57] and Scott et al. [58], the T-spline and NURBS can be incorporated efficiently via modifying the shape functions in the classical BEM and FEM. Thomas et al. [59] developed a generalization of such technique. In recent years, the application of the T-spline in the IGA framework to different areas has been developed [60–64]. In the context of IGA-BEM Scott et al. [58] extended the analysis-suitable T-spline to unstructured meshes for elastostatic problems and developed a collocation procedure based on a generalization of the Grenville abscissae. Simpson et al. [65] applied this method to acoustic problems and demonstrated the higher accuracy in comparison to the conventional Lagrange discretization due to the superior geometric approximation. Ginnis et al. [66] exploited IGA-BEM based on the T-spline to analyse in the ship wave resistance problem and Kostas et al. [67] used this method to optimize the hull shape.

Other alternatives to analysis-suitable T-spline are appearing in the literature. These alternatives include hierarchical spline (B-splines, NURBS and T-spline), PHT-splines and LR-splines. In addition to the limited works cited above, the interested reader can be found an overview of the IGA [68].

The literature review performed shows that the IGA-BEM based on the NURBS and analysis-suitable T-spline reveals that is a very recent topic. The objective of this study is the application of this method to solve non-linear unsteady gravity wave propagation and wave-structure interaction problems. We consider the full non-linear potential flow theory in combination with the MEL procedure. In this work we exploit the advantages of the IGA-BEM based on T-spline and NURBS basis in the framework of the Bézier extraction procedure. The high continuity of the basis functions allows to deal with large mesh distortion [69] in the Lagrange formulation as well as it prevents numerical instabilities as the saw-tooth effect that occurs when classical $C^0$ piecewise basis are used. The Bézier extraction concept provides an easy and standard way to introduce the NURBS and T-spline basis in the BEM. In addition, we employ direct integration with CAD software [70] to make and handle 3D geometrical models.

## 2 FORMULATION OF THE PROBLEM

We consider a fluid volume $\Omega(t)$ defined by an arbitrary lateral boundary $\Gamma_W(t)$, composed by one or several surfaces, a bottom surface $\Gamma_S(t)$ and a free surface $\Gamma_F(t)$ located at the top. The fluid can move freely with the time $(t)$ by the action of a submerged moving rigid body $\Gamma_B(t)$ or by imposing a specific initial condition. In Fig. 1 we show a sketch of the domain and the Cartesian coordinate adopted. The plane $z = 0$ corresponds to the free surface at rest.

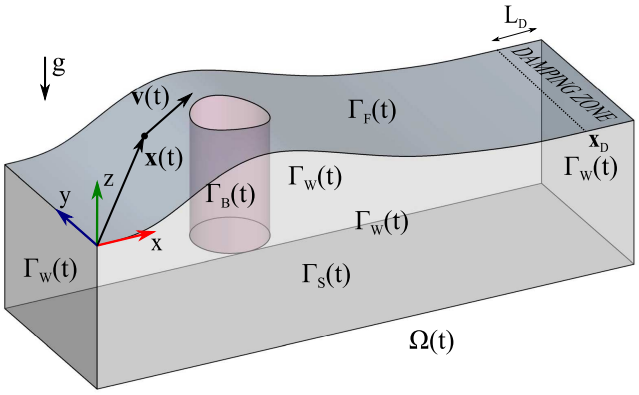



Figure 1: Sketch of the problem

The flow is assumed to be incompressible, inviscid and irrotational fluid. Under these hypotheses, the movement of a fluid particle $\mathbf{x}(t)$ can be described in terms of a velocity potential $(\phi(x, y, z, t))$, such that the velocity of a fluid particle can be expressed as:

$$\mathbf{v} = \boldsymbol{\nabla}\phi \tag{1}$$

The continuity equation in the a fluid domain $\Omega(t)$ can be reduced to a Laplace equation:

$$\nabla^2\phi = 0 \qquad \text{in } \Omega(t) \tag{2}$$

that is subjected to different boundary conditions on the surfaces $\Gamma(t)$ that define the domain.

On the free surface $\Gamma_F(t)$, the kinematic and dynamic conditions in the Lagrangian descrip-

tion are expressed as:

$$\frac{D\mathbf{x}}{Dt} = \boldsymbol{\nabla}\phi \qquad \text{on } \Gamma_F(t) \tag{3}$$

$$\frac{D\phi}{Dt} = -gz + \frac{1}{2}\boldsymbol{\nabla}\phi \cdot \boldsymbol{\nabla}\phi - \frac{p_a}{\rho} \qquad \text{on } \Gamma_F(t) \tag{4}$$

respectively. In Eqs. (3-4) $D/Dt$ denotes the material derivative, $g$ is the gravity acceleration, $p_a$ is the atmospheric pressure and $\rho$ is the density of the fluid.

The instantaneous kinematic conditions on the lateral boundaries $\Gamma_W(t)$, the bottom $\Gamma_S(t)$ and wetted body surfaces $\Gamma_B(t)$ are:

$$\frac{\partial\phi}{\partial n} = 0 \qquad \text{on } \Gamma_W(t) \text{ and } \Gamma_S(t) \tag{5}$$

$$\frac{\partial\phi}{\partial n} = \mathbf{v}_B \cdot \mathbf{n} \qquad \text{on } \Gamma_B(t) \tag{6}$$

with $\mathbf{n}$ the normal pointing out of fluid and $\mathbf{v}_B$ the velocity vector of the rigid body. In the case of a free floating body, the velocity should be computed using the second Newton's law. Here we consider the rigid body motion given by $\mathbf{v}_B = \dot{\mathbf{x}}_G + \dot{\boldsymbol{\theta}}_G \times \mathbf{r}_B$ , where $\dot{\mathbf{x}}_G$ and $\dot{\boldsymbol{\theta}}_G$ are the velocity and rotation vectors referred to the center of mass, and $\mathbf{r}_B = \mathbf{x}_B - \mathbf{x}_G$ is the location of a body particle respect to the center of mass.

As it can be noted in the formulation expressed above, the lateral boundaries are considered solid and impermeable. This means that waves are reflected and kept into the domain as it would occur in a tank. In some cases treated in this study, we are interested on simulating an infinite domain. For these cases, an artificial damping method is used to absorb the wave energy near the boundaries [10, 71]. This is done by modifying the kinematic and dynamic boundary conditions over a finite zone of the free surface, known as damping zone, located near the boundaries as follows:

$$\frac{D\mathbf{x}}{Dt} = \boldsymbol{\nabla}\phi - \mu(\mathbf{x})(\mathbf{x} - \mathbf{x}_0) \qquad \text{on } \Gamma_F(t) \tag{7}$$

$$\frac{D\phi}{Dt} = -gz + \frac{1}{2}\boldsymbol{\nabla}\phi \cdot \boldsymbol{\nabla}\phi - \frac{p_a}{\rho} - \mu(\mathbf{x})(\phi - \phi_0) \qquad \text{on } \Gamma_F(t) \tag{8}$$

where $\mathbf{x}_0$ and $\phi_0$ are reference values at the rest state, and $\mu(\mathbf{x})$ is the damping factor, which is gradually increasing along the damping zone as:

$$\mu(\mathbf{x}) = \begin{cases} \alpha \left(\frac{l_D(\mathbf{x})}{L_D}\right)^2 & \text{in damping zone} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

In Eq. 9 $\alpha$ is the maximum damping factor, $L_D$ is the length of the zone and $l_D(\mathbf{x})$ is distance between $\mathbf{x}_D$ and the specific location considered (see Fig. 1).

## 3 FORMULATION OF THE METHOD

### 3.1 Geometry

Considering a 3D BEM model, the geometry of the problem is defined uniquely by the boundary surfaces $\Gamma$ of the fluid domain. In this paper we use NURBS and cubic T-spline technology which are widely used in CAD applications. Particularly, we focus on 3D surfaces

where the physical space is given by $\Gamma \in \Re^{d_s}$ with $d_s = 3$ and the parameter space can be reduced to $\widehat{\Gamma} \in \Re^{d_p}$ with $d_p = 2$.

A NURBS surface is described in terms of a control mesh and a valid pair of knot vectors. A knot vector is denoted by $\Xi^i = \{\xi_1^i, \xi_2^i, ..., \xi_{1+n_i+p_i}^i\}$, where $\xi_k^i \in \Re$ is the kth knot value, and $n_i$ is the number of B-spline basis functions of degree $p_i$ in the $i$ direction. Hereinafter, we denote the multi-index $\mathbf{p} = \{p_1, p_2\}$, $\mathbf{n} = \{n_1, n_2\}$ and the parametric coordinate $\boldsymbol{\xi} = \{\xi^1, \xi^2\}$. The control mesh is a set of vertices that forms a structured set of quadrilateral elements. Each vertex has associated a control point $\mathbf{P}_A \in \Re^{d_s}$ and a weight $w_A \in \Re$, where $A = 1, 2, ..., n_{cp}$ is the global index that denotes all vertices $n_{cp} = \prod_{i=1}^{d_p} n_i$.

If we define the parametric space by $\widehat{\Gamma} = [\xi_1^1, \xi_{1+n_1+p_1}^1] \times [\xi_1^2, \xi_{1+n_2+p_2}^2]$, the NURBS surface that maps the parametric space to the physical space $\mathbf{x} : \widehat{\Gamma} \rightarrow \Gamma$ can be expressed as:

$$\mathbf{x}(\boldsymbol{\xi}) = \frac{\sum_{A=1}^{n_{cp}} \mathbf{P}_A w_A N_A(\boldsymbol{\xi})}{\sum_{A=1}^{n_{cp}} w_A N_A(\boldsymbol{\xi})} = \sum_{A=1}^{n_{cp}} \mathbf{P}_A R_A(\boldsymbol{\xi}) \qquad \boldsymbol{\xi} \in \widehat{\Gamma} \tag{10}$$

where $N_A(\boldsymbol{\xi})$ is the bivariate B-spline basis functions associated with the global vertex $A$ and $R_A(\boldsymbol{\xi})$ denotes the rational basis functions made by the weights and the B-spline basis functions.

Note that the NURBS is limited to surfaces defined by four edges. To represent more complex geometries it is necessary to divide the models in different rectangular patches. This limitation can be overcome naturally using the T-spline framework T-splines which allows the existence of hanging and extraordinary points within the control mesh. For a cubic T-spline, the mapping is expressed in a similar way as in Eq. (10). However, unlike NURBS, the surface is locally parametrized by defining of a local pair of knot interval vectors $\Delta\Xi_A^i = \{\Delta\xi_{A,1}^i, \Delta\xi_{A,2}^i, ..., \Delta\xi_{A,1+p_i}^i\}$ with $\Delta\xi_i = \xi_{i+1} - \xi_i$, associated to each vertex $A$ where the T-spline basis functions is restricted to be non-null. This provides the capacity of making unstructured meshes and local refinements. For more details about the T-spline or NURBS theory, the reader is refereed to [47, 55, 72–75].

### 3.2 A finite element structure

The idea is to discretize the physical domain with a series of independent elements $\mathbf{x}^e(\tilde{\boldsymbol{\xi}})$ defined over a parent domain $\tilde{\boldsymbol{\xi}} \in \widetilde{\Gamma}$ [57]. Considering the compact support property of the basis functions, there exists a number of them, $n_e$, which are non-null over each element $e$. Then, we can construct a matrix $A = IEN(e, a)$, that maps the local basis functions number $a$ in the element $e$ to the global function $A$. The local geometry, restricted to each element domain, is defined as:

$$\mathbf{x}^e(\tilde{\boldsymbol{\xi}}) = \frac{\sum_{a=1}^{n_e} \mathbf{P}_a^e w_a^e N_a^e(\tilde{\boldsymbol{\xi}})}{\sum_{a=1}^{n_e} w_a^e N_a^e(\tilde{\boldsymbol{\xi}})} \qquad \tilde{\boldsymbol{\xi}} \in \widetilde{\Gamma} \tag{11}$$

and in compact form it can be rewritten as:

$$\mathbf{x}^e(\tilde{\boldsymbol{\xi}}) = \frac{(\mathbf{P}^e)^T \mathbf{W}^e \mathbf{N}^e(\tilde{\boldsymbol{\xi}})}{\mathbf{w}^e \mathbf{N}^e(\tilde{\boldsymbol{\xi}})} \qquad \tilde{\boldsymbol{\xi}} \in \widetilde{\Gamma} \tag{12}$$

where $\mathbf{P}^e = \{\mathbf{P}_a^e\}_{a=1}^{n_e}$ is a set of the control points, $\mathbf{w}^e = \{W_a^e\}_{a=1}^{n_e}$ is a vector of weights, $\mathbf{W}^e = diag(\mathbf{w}^e)$ is a diagonal matrix and $\mathbf{N}^e(\tilde{\boldsymbol{\xi}}) = \{N_a^e(\tilde{\boldsymbol{\xi}})\}_{a=1}^{n_e}$ is a vector of the basis functions relative to the element $e$.

It should be noted that the number of T-spline basis functions contained in each element can be different, while in the case of NURBS are exactly equal to $\prod_{i=1}^{d_p}(p_i + 1)$. Furthermore, the

T-spline and NURBS basis functions affect to different elements and their definition depends on the topology. An elegant and efficient way to standardize the T-spline and NURBS basis functions to the same set of compact basis functions restricted to a unique element is using the Bézier extraction operator. This operator determines the Berstein basis combination ($B_{\mathbf{p}}(\tilde{\boldsymbol{\xi}})$ of $\mathbf{p}$ degree) in which the T-spline and NURBS basis functions are decomposed.

Let us define a vector of the Berstein basis functions $\mathbf{B}(\tilde{\boldsymbol{\xi}}) = \{B_{a,\mathbf{p}}(\tilde{\boldsymbol{\xi}})\}_{a=1}^{n_B}$ with $n_B = \prod_{i=1}^{d_p}(p_i+1)$. The geometry, can be expressed as:

$$\mathbf{x}^e(\tilde{\boldsymbol{\xi}}) = \frac{(\mathbf{P}^e)^T\mathbf{W}^e\mathbf{C}^e\mathbf{B}^e(\tilde{\boldsymbol{\xi}})}{\mathbf{w}^e\mathbf{W}^e\mathbf{C}^e\mathbf{B}^e(\tilde{\boldsymbol{\xi}})} = (\mathbf{P}^e)^T\mathbf{R}^e(\tilde{\boldsymbol{\xi}}) \qquad \tilde{\boldsymbol{\xi}} \in \widetilde{\Gamma} \tag{13}$$

being $\mathbf{C}^e \in \Re^{n_e \times n_B}$ the matrix extractor operator corresponding to the element $e$ that only depends on the topology. This method provides a Bézier-element point of view such that the geometry is decomposed in a set of Bézier elements without changing the properties of the geometry. In addition, this method provides a equivalent treatment for T-spline and NURBS. A detailed explanation of the Bézier extraction can be found in [57–59].

### 3.3 Isogeometric BEM

Given the domain $\Omega(t)$ delimited for a piecewise smooth boundary $\Gamma(t)$, the external Boundary Integral Equation (BIE) applied to the potential problem can be written as:

$$C(\mathbf{x}_0)\phi(\mathbf{x}_0,t) = \int_{\Gamma(t)} \left[ G(\mathbf{x},\mathbf{x}_0)\frac{\partial\phi(\mathbf{x},t)}{\partial n} \right] d\Gamma - \fint_{\Gamma(t)} \left[ \phi(\mathbf{x},t)\frac{\partial G(\mathbf{x},\mathbf{x}_0)}{\partial n} \right] d\Gamma \tag{14}$$

The symbol $\fint$ denotes that the integral is evaluated in the Cauchy Principal Value sense, $C(\mathbf{x}_0)$ is the solid angle defined at the source point $\mathbf{x}_0 \in \Gamma$, that can be calculated directly with the expression proposed by Mantic [76] or indirectly as described in Section 4.1, and $G(\mathbf{x},\mathbf{x}_0)$ is the fundamental solution of the Laplace's problem. In the examples treated here, we consider three possible orthogonal symmetric planes parallel to the Cartesian planes, such that the geometric complexity and computational costs can be reduced. Using the method of the images, the Green function can be expressed as:

$$G(\mathbf{x},\mathbf{x}_0) = \frac{-1}{4\pi}\sum_{i=1}^{8}\left[\frac{1}{\|\mathbf{x}-\mathbf{x}\prime_i\|}\right] \tag{15}$$

where:

$$\begin{cases} \mathbf{x}\prime_1 = \mathbf{x}_0 \\ \mathbf{x}\prime_2 = \{x_0, y_0, 2z_{sym} - z_0\} \\ \mathbf{x}\prime_3 = \{x_0, 2y_{sym} - y_0, z_0\} \\ \mathbf{x}\prime_4 = \{x_0, 2y_{sym} - y_0, 2z_{sym} - z_0\} \\ \mathbf{x}\prime_5 = \{2x_{sym} - x_0, y_0, z_0\} \\ \mathbf{x}\prime_6 = \{2x_{sym} - x_0, y_0, 2z_{sym} - z_0\} \\ \mathbf{x}\prime_7 = \{2x_{sym} - x_0, 2y_{sym} - y_0, z_0\} \\ \mathbf{x}\prime_8 = \{2x_{sym} - x_0, 2y_{sym} - y_0, 2z_{sym} - z_0\} \end{cases} \tag{16}$$

with $x = x_{sym}$, $y = y_{sym}$ and $z = z_{sym}$ being the location of the symmetric planes.

Considering a Isogeometric Analysis (IGA), the variables introduced in the BIE (i.e. the potential and the flux) are defined in a finite dimensional space given by the same basis functions

than the geometry. This provides the BIE variables and the geometry the intrinsic properties of the T-spline (or NURBS) as demonstrated in the previous section. The BIE variables can be written as:

$$\phi^e(\tilde{\boldsymbol{\xi}},t) = \mathbf{R}^e(\tilde{\boldsymbol{\xi}})^T\boldsymbol{\phi}^e(t) \tag{17}$$

$$\frac{\partial\phi^e}{\partial n}(\tilde{\boldsymbol{\xi}},t) = q^e(\tilde{\boldsymbol{\xi}},t) = \mathbf{R}^e(\tilde{\boldsymbol{\xi}})^T\boldsymbol{q}^e(t) \tag{18}$$

where $\boldsymbol{\phi}^e(t) = \{\phi_a^e(t)\}_{a=1}^{n_e}$ and $\boldsymbol{q}^e(t) = \{q_a^e(t)\}_{a=1}^{n_e}$ are the temporal potential and temporal flux vector corresponding to the local basis functions of each element, respectively.

Introducing Eqs. (17) and (13) into Eq. (14), and applying the result on a particular source point $\mathbf{x}_i^{e_i} = \mathbf{R}_i^{e_i}(\tilde{\boldsymbol{\xi}}_i^{e_i})^T\mathbf{P}^{e_i}$ contained in the element $e_i$, we obtain the discrete IGA BIE:

$$\begin{aligned} C(\mathbf{x}_i^{e_i})(\mathbf{R}_i^{e_i})^T\boldsymbol{\phi}^{e_i} = \sum_{e=1}^{N}\Bigg[\int_{\Gamma^e(t)}\left[G(\mathbf{x}^e,\mathbf{x}_i^{e_i})(\mathbf{R}^e)^T\boldsymbol{q}^e\right]\left|\frac{\partial\mathbf{x}^e}{\partial\tilde{\boldsymbol{\xi}}}\right|d\Gamma(\tilde{\boldsymbol{\xi}}) \\ -\!\!\!\!\!\!\int_{\Gamma^e(t)}\left[(\mathbf{R}^e)^T\boldsymbol{\phi}^e\frac{\partial G(\mathbf{x}^e,\mathbf{x}_i^{e_i})}{\partial n}\right]\left|\frac{\partial\mathbf{x}^e}{\partial\tilde{\boldsymbol{\xi}}}\right|d\Gamma(\tilde{\boldsymbol{\xi}})\Bigg] \end{aligned} \tag{19}$$

In Eq. (19) $\left|\frac{\partial\mathbf{x}^e}{\partial\tilde{\boldsymbol{\xi}}}\right|$ is the Jacobian determinant of the geometrical mapping, $N$ is the number of elements. For more details about the calculation of the Jacobian and the tangential derivatives, the reader is referred to [57, 77]. Hereinafter, we drop the upper index $e_i$ and for convenience we keep the subindex $i$ to refer to the source point.

In this study we employ a collocational method, such that the BIE must be satisfied for a set of $i = 1, 2, ..., n_{cp}$ specific source points known as collocation points. Particularly, we use a generalization of the Greville abscissae as presented in Scott et al. [58], since it provides accurate results in the context of the IGA BEM. However, we relax the continuity condition for the position of control points such as these are allowed to lie in sharp edges and corners. As demonstrated in [65], this relaxation does not reduce the accuracy of the results and it facilitates the special treatment of the velocity in these locations, as shown in Section 4.3.

## 4 NUMERICAL IMPLEMENTATION

### 4.1 Numerical integration

It is well known that the first and second kernel of the BIE (see Eq. 14) contain a weak and a strong singularity, respectively, when the point of evaluation approaches to the collocation point. The integrals are defined, but they have to be suitably evaluated to avoid numerical problems. This topic has been studied intensively and an overview of the available techniques can be found in [78–87].

In this study we regularize the second kernel using the rigid body method [78], that provides an indirect way to calculate the jump term as:

$$C(\mathbf{x}_0) = -\!\!\!\!\!\!\int_{\Gamma(t)}\frac{\partial G(\mathbf{x},\mathbf{x}_0)}{\partial n}d\Gamma \tag{20}$$

Introducing Eq. 20 in Eq. 14 gives:

$$\int_{\Gamma(t)}\left[(\phi(\mathbf{x},t)-\phi(\mathbf{x}_0,t))\frac{\partial G(\mathbf{x},\mathbf{x}_0)}{\partial n}\right]d\Gamma = \int_{\Gamma(t)}\left[G(\mathbf{x},\mathbf{x}_0)\frac{\partial\phi(\mathbf{x},t)}{\partial n}\right]d\Gamma \tag{21}$$

The weak singularity is solved using a local change of variables over the element containing the collocation point, as proposed by Lachat and Watson [79]. Once regularized, these kernels can be numerically evaluated by applying the quadrature rule to each element in the parent domain. We use the standard Gauss-Legendre quadrature. Recently, new specific quadrature rules for NURBS have been proposed to take into account the continuity of the basis functions [80]. In addition, if the collocation point is not located on the element but very close, the integrals become nearly singular and their calculation can exhibit numerical instability. For this case, we use an adaptive regular element subdivision on the parent domain.

### 4.2 System of equations

The discretization of the BIE in all collocation points leads to a set of $n_{cp}$ equations with $2n_{cp}$ variables corresponding to the potential and the flux at the control points. This set of equations can be written compactly as,

$$(\mathbf{C}\boldsymbol{\mathcal{T}} + \boldsymbol{\mathcal{H}})\,\boldsymbol{\phi} = \boldsymbol{\mathcal{G}}\mathbf{q} \tag{22}$$

in which, using the global notation, $\boldsymbol{\phi} = \{\phi\}_{A=1}^{n_{cp}}$ is the potential vector, $\mathbf{q} = \{q\}_{A=1}^{n_{cp}}$ is the flux vector, and the matrices $\mathbf{C}$, $\boldsymbol{\mathcal{T}}$, $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{G}} \in \Re^{n_{cp}\times n_{cp}}$ are given by:

$$\boldsymbol{\mathcal{H}}_{iA} = \int_{\Gamma(t)} \frac{\partial G(\mathbf{x}, \mathbf{x}_i)}{\partial n} R_A d\Gamma \tag{23}$$

$$\boldsymbol{\mathcal{G}}_{iA} = \int_{\Gamma(t)} G(\mathbf{x}, \mathbf{x}_i) R_A d\Gamma \tag{24}$$

$$\boldsymbol{\mathcal{T}}_{iA} = R_A(\mathbf{x}_i) \tag{25}$$

$$\mathbf{C}_{iA} = -\delta_{iA} \sum_{k=1}^{n_{cp}} \boldsymbol{\mathcal{H}}_{ik} \tag{26}$$

As it has been shown in Section 2, the problem under consideration consists in a mixed boundary problem in which the potential, once integrated in time, is known on the free surface and its flux on the rest of the boundaries. The discretization of the boundary conditions on the set of collocation points yields,

$$\mathbf{R}(\mathbf{x}_i)^T \boldsymbol{\phi} = \bar{\phi}(\mathbf{x}_i) \qquad \forall \mathbf{x}_i \in \Gamma_F(t) \tag{27}$$

$$\mathbf{R}(\mathbf{x}_i)^T \mathbf{q} = \bar{q}(\mathbf{x}_i) \qquad \forall \mathbf{x}_i \in \Gamma_W(t) \cup \Gamma_S(t) \cup \Gamma_B(t) \tag{28}$$

We reorder the system of equations (22), (27) and (28) to obtain a linear system that can be written as $\mathbf{Ka} = \mathbf{b}$, where $\mathbf{a} \in \Re^{2n_{cp}}$ is a vector containing the control variables, $\mathbf{b} \in \Re^{2n_{cp}}$ is a vector with the boundary conditions at the collocation points and $\mathbf{K} \in \Re^{2n_{cp}\times 2n_{cp}}$ is in general a dense non-symmetric matrix. The minimum residual method (GMRES) have been found to be efficient and suitable for solving the system of equations [10, 88]. In the present study the tolerance has been set to $10^{-12}$. Moreover, to correctly model the physical variables at the sharp edge and corners, we use the semi-discontinuous basis technique, as presented in [58], such that the flux is allowed to be discontinuous while the potential is continuous.

### 4.3 Velocity field

Before the update of the free surface, we need to calculate the velocity field. Following the idea of IGA, we use the same set of T-spline (or NURBS) basis functions to approximate the

velocity which can be expressed as,

$$\mathbf{v}(\tilde{\boldsymbol{\xi}}, t) = \mathbf{R}(\boldsymbol{\xi})^T \mathbf{v}(t) \tag{29}$$

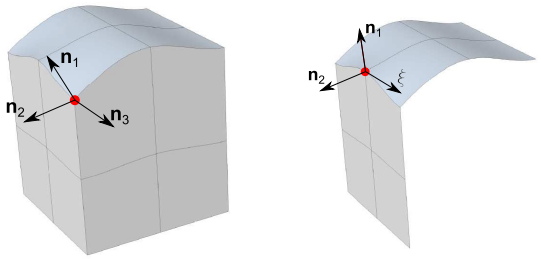
where $\mathbf{v}(t) = \{\mathbf{v}_A(t)\}_{A=1}^{n_{cp}}$ is the global velocity vector.

Using Eq. (1), the velocity of the fluid particle at any point of the domain can be determined as a function of the potential, of the flux and of the Jacobian of the mapping:

$$\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} \frac{\partial x}{\partial \xi^1} & \frac{\partial y}{\partial \xi^1} & \frac{\partial z}{\partial \xi^1} \\ \frac{\partial x}{\partial \xi^2} & \frac{\partial y}{\partial \xi^2} & \frac{\partial z}{\partial \xi^2} \\ n_x & n_y & n_z \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \phi}{\partial \xi^1} \\ \frac{\partial \phi}{\partial \xi^2} \\ \frac{\partial \phi}{\partial n} \end{bmatrix} \tag{30}$$

In the case of a damping zone on the free surface, the terms on the right hand side have to be corrected as shown in Eq. (7).

The sharp edges and corners where several surfaces meet need a special attention. Because the derivatives of the potential and the flux are discontinuous at the intersections, the velocity associated to each surface can be numerically different and also it can be incompatible with the movement of each boundary. An usual procedure to overcome this problem is, first compute an average velocity and second make a normal projection of the intersections on the new boundary position during the time marching [10]. In this work we employ an alternative procedure that consists on calculating an unique velocity compatible with the normal velocity of each surface. We consider two possible cases as shown in Fig. 2. The first one (Fig. 2a) correspond to a corner at the intersection of three surfaces. In this case, the velocity can be determined using the follow equation:




(a) (b)

Figure 2: (a) Corner at the intersection of three surfaces and (b) edge defined by the intersection two surfaces.

$$\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} n_x^1 & n_y^1 & n_z^1 \\ n_x^2 & n_y^2 & n_z^2 \\ n_x^3 & n_y^3 & n_z^3 \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \phi}{\partial n^1} \\ \frac{\partial \phi}{\partial n^2} \\ \frac{\partial \phi}{\partial n^3} \end{bmatrix} \tag{31}$$

where $n^i$ are the normals corresponding to each surface.

The second case (Fig. 2b) is found in the edge defined by the intersection of two surfaces. The velocity can be computed as,

$$\begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = \begin{bmatrix} n_x^1 & n_y^1 & n_z^1 \\ n_x^2 & n_y^2 & n_z^2 \\ \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} & \frac{\partial z}{\partial \xi} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \phi}{\partial n^1} \\ \frac{\partial \phi}{\partial n^2} \\ \frac{\partial \phi}{\partial \xi} \end{bmatrix} \tag{32}$$

In Eq. (32) $\xi$ represents the common axis along the direction of the edge. Note that, since $\phi$ is continuous, its derivative along the edge is unique.

Now, if we particularize Eq. (30) for the collocation points contained inside the surfaces, Eq. (32) for the collocation points on the edges and Eq. (31) for the corners, we obtain a system of equations (Eq. 33) that allows the determination of the global velocity vector.

$$\mathbf{v}(t) = \boldsymbol{\mathcal{T}}^{-1}\boldsymbol{\mathcal{V}} \tag{33}$$

In Eq. (33) $\boldsymbol{\mathcal{V}} = \{\mathbf{v}_i(\boldsymbol{\xi}_i, t)\}_{i=1}^{n_{cp}}$ is the velocity vector at the collocations points and $\boldsymbol{\mathcal{T}}$ is an invertible and well conditioned matrix. Note that this matrix, that we denote as the interpolation matrix, allows the mapping of the information of the control points to the collocation points. In addition, it can be applied to any variable because we use an IGA.

Therefore, using this procedure, the velocity field is forced to be unique and compatible with the normal movement of the body or the walls. Moreover, the intersections can be managed directly through collocation points due to the interpolatory nature of the basis functions at the sharp edges or corners. However, a slight separation of the intersections from the body or the walls can appear during the time marching. The points on these intersections are relocated back by the normal projection technique as proposed by Bai and Taylor [10].

### 4.4 Hydrodynamic pressure

The pressure force exerted by the fluid on the surface of a body can be computed as,

$$F_B = \int_{\Gamma_B(t)} p\mathbf{n} d\Gamma \tag{34}$$

The pressure field at any point can be calculated using the Bernoulli equation:

$$p = -\rho\left(\frac{\partial\phi}{\partial t} + gz + \frac{1}{2}\boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\phi\right) \tag{35}$$

The last term on the right hand side of the Eq. 35 can be computed directly once the BIE (14) is computed. The term $\frac{\partial\phi}{\partial t}$, which is not known at this stage, can be approximated by using temporal integration schemes. However, this can produce numerical errors. Here, we use an auxiliary BIE to find a more accurate approach. As proposed by Dombre et al. [3, 33], an analogous BIE for $\psi = \partial\phi/\partial t$ can be considered. This auxiliary BIE does not increase significantly the computational cost because the discretization is the same as in the main BIE. Only, the boundary conditions have to be changed. The equations for the new variable $\psi$ can be written as,

$$\nabla^2\psi = 0 \qquad \text{in } \Omega(t) \tag{36}$$

$$\psi = -gz + \frac{1}{2}\boldsymbol{\nabla}\phi\cdot\boldsymbol{\nabla}\phi - \frac{p_a}{\rho} \qquad \text{on } \Gamma_F(t) \tag{37}$$

$$\frac{\partial\psi}{\partial n} = 0 \qquad \text{on } \Gamma_W(t) \text{ and } \Gamma_S(t) \tag{38}$$

$$\frac{\partial\psi}{\partial n} = \frac{d(\mathbf{v}_B\cdot\mathbf{n})}{dt} - (\mathbf{v}_B\cdot\boldsymbol{\nabla}\boldsymbol{\nabla}\phi)\cdot\mathbf{n} - \boldsymbol{\nabla}\phi\frac{d\mathbf{n}}{dt} \qquad \text{on } \Gamma_B(t) \tag{39}$$

where $\frac{d}{dt}$ denotes Lagrangian time derivative following the body motion.

### 4.5 Time marching

As shown in the formulation of the problem, the position of the fluid particles and the physical variables, Eq. (3) and Eq. (4), depend on the time. Therefore, to obtain the temporal evolution of the free surface, a temporal integration scheme is needed. In the context of strong non-linear gravity waves, several schemes has been proposed [3,29,71,90]. In this study, we use the the standard 4th order Runge Kutta method (RK4), which has been shown to be numerically stable and accurate.

An important aspect of temporal integration is the adequate selection of the time step to provide an accurate and stable temporal resolution. This depends on the integration scheme, physical discretization and the particular problem under consideration. For periodic wave problems the time step is usually chosen as a fraction of the wave period such that the physical phenomenon is represented accurately [71]. For a solitary wave, Grilli et al. [3] proposed an adaptive time steeping method as a function of the spatial discretization and the Courant number, $\mathcal{C}$ (Eq. 40). They demonstrated that there exists an optimal Courant number such that the error is minimum. We use this method, adapted to the IGA-BEM, and investigate about the optimal value of the Courant number.

$$\Delta t = \mathcal{C} Le_{min} / \sqrt{gh} \tag{40}$$

In Eq 40 $Le_{min}$ is the minimum characteristic element size defined as the square root of its surface and $h$ is the depth of the bottom surface.

The steps for the time marching procedure are:

**1**. The initial geometrical model is defined and an initial boundary condition for the potential ($\bar{\phi}_0$) and the flux ($\bar{q}_0$) are specified on the free surface and on the rest of the surfaces, respectively.

**2**. A time step ($\Delta t_n$) is selected. This can be set dynamically at each step depending on the type of problem.

**3**. The time integration is carried out to update the position of the fluid particles and potential ( $\phi_{n+1}$, $\mathbf{x}_{n+1}$) from the actual time ($t_n$) to the next time step ($t_{n+1} = t_n + \Delta t_n$). Due to the non-linear behaviour, several sub-steps are required.

- **3**.1. The BIE equation (22) is calculated, and the boundary conditions (Eq. (27) and Eq. (28)) are applied at the each sub-step ($k$). A linear system of equations is obtained and solved by using the GMRES.
- **3**.2. The velocity is computed using Eq. (30), Eq. (31) and Eq. (32).
- **3**.3. The potential is updated to the next sub-step ($\phi_{n+1}^{k+1}$) by applying the integration scheme to the kinematic relation for the free surface (Eq. (3) or Eq. (8)).
- **3**.4. The positions of the fluid particles are updated to the next sub-step ($\mathbf{x}_{n+1}^{k+1}$) using the known velocity field. If fluid particles are separated from the walls or body surface, these are relocated back to the boundaries using a normal projection.
- **3**.5. The flux is computed in the next sub-step ($q_{n+1}^{k+1}$) using the kinematic equations for the walls, bottom surface, and body surface.
- **3**.6. For the RK4 scheme, this process is repeated four times ($k = 1$ to $k = 4$).

**4**. The fluid forces on the body are computed (Eq. (35) and Eq. (34)).

**5**. The volume and total energy is calculated in order to monitor their temporal evolution as,

$$V(t) = \int_{\Gamma(t)} z(\boldsymbol{\lambda}_z \mathbf{n}) d\Gamma \tag{41}$$

$$e(t) = \frac{\rho g}{2} \int_{\Gamma(t)} z^2(\boldsymbol{\lambda}_z \mathbf{n}) + \phi q d\Gamma \tag{42}$$

where $\boldsymbol{\lambda}_z$ is the unit vertical vector.

**6**. Steps 2 to 5 are carried out iteratively until a maximum time is reached.

It should be noted that we are employing a collocational method and all variables are evaluated at the collocations points. The values of the variables on the control points can be obtained using the interpolation matrix.

## 5 NUMERICAL RESULTS

In this section we show some numerical examples to validate the present IGA-BEM formulation. Each example is focused on a particular computational aspect of the method. The first example is a train of small amplitude waves and we analyse the spatial convergence of the IGA-BEM solution. In the next example we simulate a solitary wave to evaluate the temporal stability and the accuracy of the method. In the third example we check the ability of the method to handle wave-structure problems by simulating the waves generated in a tank.

All geometric models were construct using the T-spline plugin of Rhino [70]. Symmetric boundary conditions were imposed at the symmetric planes of the problems. These planes have the zero flux condition as the wall and are not discretized because symmetric conditions are implicitly incorporated in the Green function. Moreover, unless otherwise is specified, the time was made dimensionless using the time scale $\sqrt{h/g}$ with $h$ being the depth of water tank. The dimensionless variables are denoted with $^*$.

### 5.1 Periodic wave: convergence of the spatial discretization

The objective of this example is to analyse the convergence of the IGA-BEM method presented in this study. The test, which has analytical solution, consists on the propagation of steady regular gravity waves over a horizontal bottom surface. For the case of small wave amplitude with respect to the depth and the wavelength $\lambda$, the dynamic boundary conditions on the free surface (Eq. (4)) can be linearised and the exact solution to potential and the free surface elevation ($\eta$) can be written as:

$$\bar{\phi}(x, z, t) = \frac{A\omega}{k} \frac{\cosh(k(z+h))}{\cosh(kh)} \sin(kx - \omega t) \tag{43}$$

$$\bar{\eta} = A\cos(kx - \omega t) \tag{44}$$

where $A$ is the amplitude, $h$ is the depth, $\omega$ is the wave frequency, and $k = 2\pi/\lambda$ is the wave number.

This is a 2D periodic problem where the gravity waves propagate along the x-direction and they are invariant along y-direction. Since the presented method is formulated in 3D, we study a representative domain ($\Omega$) defined by its length $x = [0, 2\lambda]$, depth $z = [0, -\lambda]$ and width

$y = [0, 2\lambda]$. The parameters are set by the relation between the amplitude and the wave number $kA = 0.2$, and the dispersion equation $\omega^2 = gk\tanh(kh)$. Moreover, a specific time $\bar{t} = \pi/(2\omega)$ is chosen such the flux across the lateral boundaries along the longitudinal direction are zero.

Considering three symmetric planes at $x = 2\lambda$, $y = 2\lambda$ and $z = -\lambda$, the geometry is discretized with three NURBS surfaces. Note that these surfaces can be converted into a unique T-spline patch. We set a number of Bézier elements that varies in a range from 8 to 24 regular elements with an increment of 4 elements per wavelength along the longitudinal direction, while along the other two directions we utilize 24 Bézier elements. All weights are set to unity. Moreover, three different basis degree $p = 2, 3, 4$ are selected.

It should be noted that a sinusoidal function cannot be represented exactly by a rational spline, and a geometrical error is expected. To minimize this error we use a $L^2$-projection method. Fig. 3a shows the $L^2$-error between the elevation function and the free surface representation. It can be seen that the geometrical discretization shows an optimal convergence (i.e. convergence rate $\geq p + 1$). In the same line, given the potential function expressed in Eq. (43), we use the $L^2$-projection technique to apply the boundary conditions on the free surface. Its convergence is also optimal as shown in Fig. 3b.

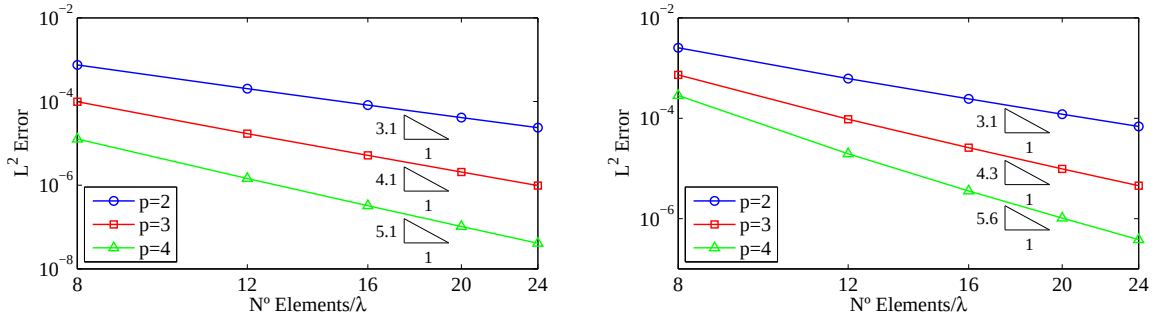



Figure 3: a) $L^2$-Error between the elevation function and the NURBS surface that represent the free surface in a central domain $[\lambda \times \lambda]$. b) $L^2$-Error between the potential function and the boundary conditions applied on the free surface in the same domain.

In Fig. 4 we show the point-wise error between the analytic and BEM solution ($\|w - w^h\|$ with $w^h$ being the flux or potential) for three levels of refinement (8, 16 and 24 elements per wavelength) and the standard cubic basis degree. The error on the Neumann boundaries is considerable lower (one order of magnitude smaller) than on the Dirichlet boundary (free surface). Moreover, small fluctuations are detected near the intersections of the free surface due to lateral boundary effects that become negligible for a distance greater than $\lambda/2$.

Given the previous findings, we limit the converge study of the IGA-BEM solution to a central region $[\lambda \times \lambda]$ of the free surface. The result is given in Fig. 5 where an optimal convergence ratio is shown for all cases.

### 5.2 Solitary wave: temporal accuracy

Our next benchmark example consists on a solitary wave propagating over constant depth. This example has been studied by Tanaka [91], who obtained a 2D full non-linear numerical solution, and then used by other authors to validate their models [3,25,92]. In this subsection we

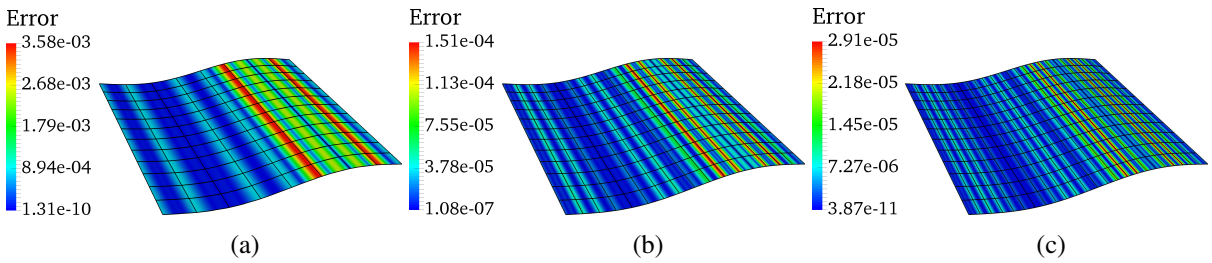



Figure 4: Point-wise error between the analytic and numeric (IGA-BEM) solution in the center region $[\lambda \times \lambda]$ of the free surface using cubic basis functions for three levels of refinement: a) 8 elements/$\lambda$, b) 16 elements/$\lambda$ and c) 24 elements/$\lambda$.

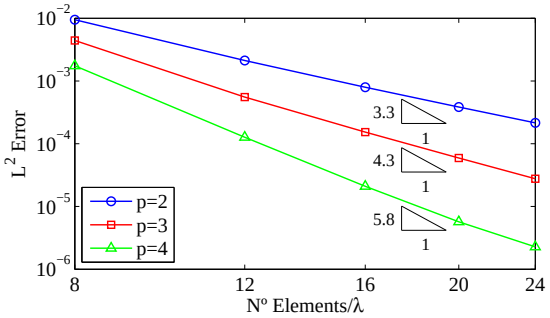



Figure 5: $L^2$-Error between the analytical and numerical (IGA-BEM) solution for the central domain $[\lambda \times \lambda]$ on the free surface.

analyse the temporal stability and accuracy of the present method by computing the numerical errors along the time with respect to the steady solution presented by Dutykha and Clamond [92].

We consider a wave of amplitude $A = 0.6h$, where $h$ is the depth of the free surface at the rest. The domain is given by a length $x = [0, 20h]$, width $y = [0, 2h]$ and depth $z = [0, -h]$. The wave is propagated along the longitudinal direction (x-direction) and at the initial stage the crest is located at $x = 7.5h$.

As in the previous example, we set three symmetric planes at at $x = 20h$, $y = 2h$ and $z = -h$ and the model is represented by three NURBS surfaces with weight unity. The free surface elevation and the potential are established with the $L^2$-projection method applied over the solution proposed by Dutykha and Clamond [92]. We consider two longitudinal discretizations with element lengths of $0.25$ and $0.5$. The element length along y-direction is $0.5$ and $0.25$ along z-direction. We use quadratic and cubic NURBS basis functions. Moreover, simulations with an equivalent model using the standard quadratic Lagrangian basis functions has been performed for comparison purposes.

We carried out some simulations for different Courant numbers (calculated based on the length of the elements of the free surface along x-direction, given the invariance along y-direction) in the range of $0.3 \leq \mathcal{C} \leq 1.0$. The instantaneous shape, energy and volume of the wave have been monitored during a period of $\Delta t^* = 4$. We found that the wave is kept stable during the simulation for all cases and the elements are concentrated around the crest, as it can be seen in Fig.6. In Fig. 7 we show the average of the relative energy error ($\epsilon_e = (e(t) - e(0))/e(0)$), volume error ($\epsilon_v = (V(t) - V(0))/V(0)$) and $L^2$-error of the shape normalized by $h\sqrt{A_F}$ with $A_F$ being the area of the free surface. It can be seen that the errors

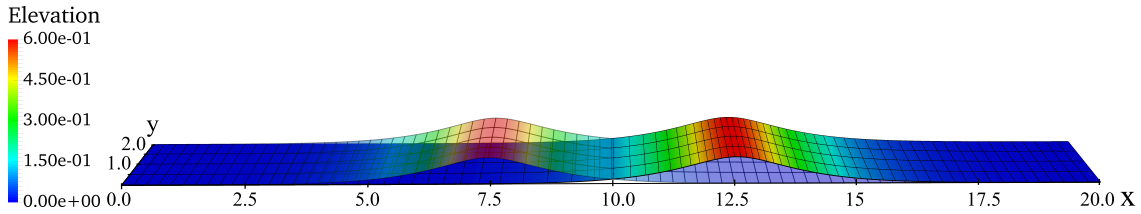



Figure 6: Time evolution of the wave shape. The translucent surface correpond to he initial condition and the solid surface to $t^* = 4$. Simulation with the fine mesh using cubic NURBS basis functions.

are mostly constant for the complete range of Courant numbers considered. This demonstrates the robustness and stability of the scheme of integration. These errors decrease as the physical discretization is finer and the basis degree increase.

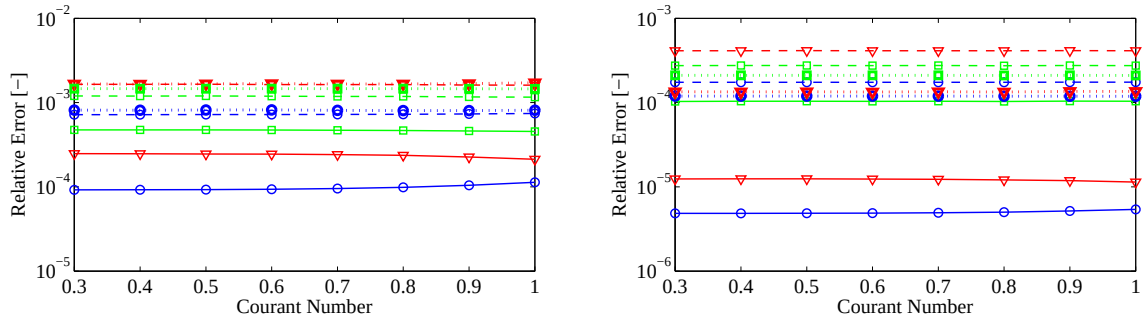
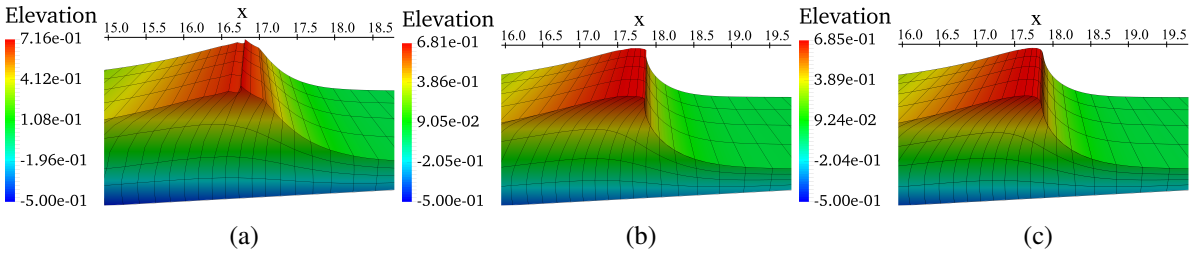



Figure 7: Relative energy (triangle), volumen (circle) and shape (square) errors for a cubic NURBS (continuouos line), quadratic NURBS (dash line) and cuadratic Lagrangian (dotted line) model, for a) the case of a coarse mesh and b) the fine mesh .

We also analysed the numerical stability of the same solitary wave propagating over a 1:15 slope that starts at the initial position of the crest ($x = 7.5h$). This configuration, that produces the overturn of the wave, was studied in detail by Grilli et al. [2] using a 2D BEM mode. Later these results have been used used to validate a 3D BEM formulation [3]. Here we consider the numerical model described above for the fine mesh and a Courant number of $0.5$. Additionally, we discretize the bottom surface with the same number of uniform elements than the free surface and we extend the length of domain up to $x = 22h$. The simulation is carried out until the wave reaches the breaking point (at $t^*_{BP} = 7.78$). At larger times the lateral mesh becomes very distorted and the numerical errors increase very fast. In fact, for the quadratic Lagrange model, numerical instabilities appear before that point (at $t^* = 7.11$), as it can be observed in Fig. 8a, and the simulation stops. This problem is associated with the lack of smoothness across the elements that produces a sawtooth effect near the crest. However, the models using NURBS are stable for $t^* <= 7.78$ and the free surface remains smooth (see Fig. 8a and b). For the cubic NURBS model the wave reaches a maximum height of $6.85 \cdot 10^{-1}h$ which is very close to the reported in [2] ($6.89 \cdot 10^{-1}h$), while for the quadratic NURBS model is lightly lower ($6.81 \cdot 10^{-1}h$) due to the lower degree of accuracy. In Fig. 9 we compare the central section of the cubic NURBS model and the results presented by Grilli et al. [2] at the breaking time. The relative errors for volume and energy are $\epsilon_v = 1.21 \cdot 10^{-4}$ and $\epsilon_e = 7.62 \cdot 10^{-5}$, respectively, and the shape is in very good agreement with the data reported in [2].




(a) (b) (c)

Figure 8: Wave shape for a) the Lagrangian model at $t^* = 7.11$, when numerical inestabilities appear, and for b) the quadratic and c) cubic NURBS models at the breaking time (at $t^* = 7.78$.)

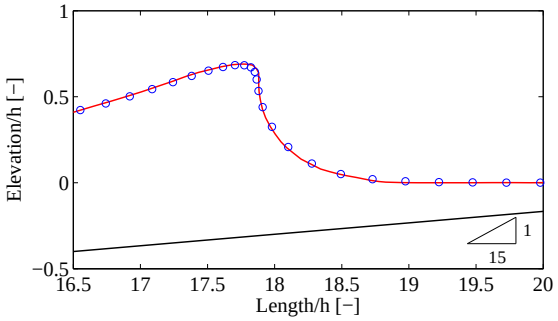



Figure 9: Wave shape of the central section (at $y = 1h$) for the present cubic NURBS model (circles) and for the results presented by Grilli et al. [2] using a 2D BEM model (continuous line) at the breaking time ($t^* = 7.78$).

### 5.3 Wave in a cylindrical tank: Fluid-structure interaction

The objective of this section is to demonstrate the ability of the present formulation to reproduce fluid-structure phenomenon considering the effect of the waves propagating on the free surface of a tank.

We consider a cylindrical water tank with a solid cylindrical bar located at the center. The radius of the tank is $r_T = 10h$ and the bar is $r_B = 1h$ with $h$ being the depth of the tank. Initially a wave is established in the middle point between the the bar and the external wall of the tank. The elevation of the wave is given by a Gaussian pulse $z = Aexp[-(x - (r_B + r_T)/2)^2 - y^2]$ with $A = 0.1h$ being the amplitude of the wave. This problem has been solved by Chern et al. [93] using a Pseudo-Spectral Matrix Element Method and then used by Bai and Taylor [10] to validated their BEM formulation. More details about the problem can be found in [10, 93].

The geometry is discretized using three regular cubic NURBS surfaces that represent the free surface, the bar and the external wall. Two symmetry planes are considered at $y = 0$ and $z = -h$. We set $26 \times 25$ Bézier elements (radial $\times$ circumferential) for the free surface, and $6 \times 25$ (vertical $\times$ circumferential) for the body and external wall. There are a total of 1316 control points.

A comparison of the surface elevation adjacent to the body ($[x = r_B, y = 0]$) and horizontal force along the time between the present results and those reported by Chern et al. [93] is shown in Fig. 10. The simulation is run until $t^* = 30$ with a time step $\Delta t^* = 0.1$. As it can be seen, the present results are in good agreement with those in [93] where we have used less DoF than the required by [10] to obtain a close approximation (this is, 1316 vs 4208 DoF of the mesh b in Fig. 3a).

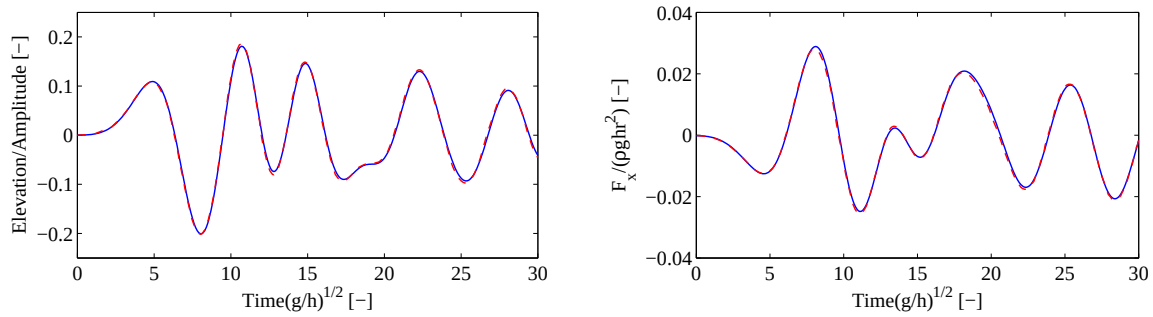



Figure 10: Time evolution of a) surface elevation adjacent to the body ($[x = r_B, y = 0]$) and b) horizontal force ($F_x/(\rho g h r^2)$). Present results: continuous line and results in [93]: dash line.

## 6 CONCLUSIONS

In this paper a 3D IGA-BEM formulation using T-spline and NURBS basis to study fully non-linear gravity wave propagation in the time domain has been presented. The use of splines basis provides a high geometrical approximation and also it can be directly integrated with computer aid geometrical design tools. The properties of these basis within the BEM have been analysed using numerical examples. First, an optimal spatial convergence of the method have been demonstrated simulating a periodic wave. Following, the temporal stability and accuracy have been analysed simulating the propagation of a solitary gravity wave. The numerical errors are kept practically constant to a low order over a wide range of time steps demonstrating the robustness of the time marching scheme. Moreover, the smoothness of the basis provides stability to the method without the need of artificial smooth techniques, and it allows relatively large distortion of elements. In addition, the ability of the method to the simulate the fluid-structure interaction with fixed bodies has been shown simulating the waves field in a tank.

## ACKNOWLEDGMENTS

This study has been supported by the Spanish Ministerio de Economía y Competitividad under project CTQ2013-46799-C2-1-P.

## REFERENCES

[1] S. T. Grilli, S. Vogelmann, P. Watts, Development of a 3d numerical wave tank for modeling tsunami generation by underwater landslides, Engineering Analysis with Boundary Elements 26 (4) (2002) 301–313.

[2] S. T. Grilli, I. A. Svendsen, R. Subramanya, Breaking criterion and characteristics for solitary waves on slopes, Journal of Waterway, Port, Coastal and Ocean Engineering 123 (3) (1997) 102–112.

[3] S. T. Grilli, P. Guyenne, F. Dias, A fully non-linear model for three-dimensional overturning waves over an arbitrary bottom, International Journal for Numerical Methods in Fluids 35 (7) (2001) 829–867.

[4] P. Guyenne, S. T. Grilli, Numerical study of three-dimensional overturning waves in shallow water, Journal of Fluid Mechanics 547 (2006) 361–388.

[5] N. Mizutani, A. M. Mostafa, K. Iwata, Nonlinear regular wave, submerged breakwater and seabed dynamic interaction, Coastal Engineering 33 (2–3) (1998) 177–202.

[6] H. K. Johnson, T. V. Karambas, I. Avgeris, B. Zanuttigh, D. Gonzalez-Marco, I. Caceres, Modelling of waves and currents around submerged breakwaters, Coastal Engineering 52 (10–11) (2005) 949–969.

[7] S. Sun, G. Wu, Fully nonlinear simulation for fluid/structure impact: A review, Journal of Marine Science and Application 13 (3) (2014) 237–244.

[8] M. Au, C. Brebbia, Diffraction of water waves for vertical cylinders using boundary elements, Applied Mathematical Modelling 7 (2) (1983) 106–114.

[9] W. Koo, M. Kim, Numerical simulation of nonlinear wave and force generated by a wedge-shape wave maker, Ocean Engineering 33 (8-9) (2006) 983–1006.

[10] W. Bai, R. Eatock Taylor, Higher-order boundary element simulation of fully nonlinear wave radiation by oscillating vertical cylinders, Applied Ocean Research 28 (4) (2006) 247–265.

[11] G. Ducrozet, F. Bonnefoy, D. L. Touzé, P. Ferrant, A modified High-Order Spectral method for wavemaker modeling in a numerical wave tank, European Journal of Mechanics - B/Fluids 34 (2012) 19 – 34.

[12] P. Ferrant, D. L. Touzé, K. Pelletier, Non-linear time-domain models for irregular wave diffraction about offshore structures, International Journal for Numerical Methods in Fluids 43 (2003) 1257–1277.

[13] S. Jeon, Y. Cho, M. Seo, J. Cho, W. Jeong, Dynamic response of floating substructure of spar-type offshore wind turbine with catenary mooring cables, Ocean Engineering 72 (2013) 356–364.

[14] T. Shivaji Ganesan, D. Sen, Time-domain simulation of large-amplitude wave–structure interactions by a 3d numerical tank approach, Journal of Ocean Engineering and Marine Energy 1 (3) (2015) 299–324.

[15] R. A. Watai, F. Ruggeri, C. M. P. Sampaio, A. N. Simos, Development of a time domain boundary element method for numerical analysis of floating bodies' responses in waves, Journal of the Brazilian Society of Mechanical Sciences and Engineering 37 (5) (2015) 1569–1589.

[16] W. Koo, M. Kim, D. Lee, S. Hong., Nonlinear time-domain simulation of pneumatic floating breakwater, International Journal of Offshore and Polar Engineering 16 (1) (2006) 25–32.

[17] M. S. Longuet-Higgins, E. D. Cokelet, The deformation of steep surface waves on water. i. a numerical method of computation, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 350 (1660) (1976) 1–26.

[18] T. Nakayama, K. Washizu, The boundary element method applied to the analysis of two-dimensional nonlinear sloshing problems, International Journal for Numerical Methods in Engineering 17 (11) (1981) 1631–1646.

[19] T. Nakayama, Boundary element analysis of nonlinear water wave problems, International Journal for Numerical Methods in Engineering 19 (7) (1983) 953–970.

[20] D. Skyner, A comparison of numerical predictions and experimental measurements of the internal kinematics of a deep-water plunging wave, Journal of Fluid Mechanics 315 (1996) 51–64.

[21] S. Maiti, D. Sen, Computation of solitary waves during propagation and runup on a slope, Ocean Engineering 26 (11) (1999) 1063–1083.

[22] M. Abou-Dina, M. Helal, Boundary integral method applied to the transient, nonlinear wave propagation in a fluid with initial free surface elevation, Applied Mathematical Modelling 24 (8–9) (2000) 535–549.

[23] D. G. Dommermuth, D. K. P. Yue, W. M. Lin, R. J. Rapp, E. S. Chan, W. K. Melville, Deep-water plunging breakers: a comparison between potential theory and experiments, Journal of Fluid Mechanics 189 (1988) 423–442.

[24] S. T. Grilli, J. Skourup, I. Svendsen, An efficient boundary element method for nonlinear water waves, Engineering Analysis with Boundary Elements 6 (2) (1989) 97–107.

[25] S. T. Grilli, I. A. Svendsen, Corner problems and global accuracy in the boundary element solution of nonlinear wave flows, Engineering Analysis with Boundary Elements 7 (4) (1990) 178–195.

[26] K. Tanizawa, The state of the art on numerical wave tank, Proceedings of the 4th Osaka Colloquium on seakeeping performance of ships (2000) 95–114.

[27] C. C. Lee, Y. H. Liu, C. H. Kim, Simulation of nonlinear waves and forces due to transient and steady motion of submerged sphere, International Journal of Offshore and Polar Engineering 416 (1994) 174–182.

[28] W. Bai, R. E. Taylor, Numerical simulation of fully nonlinear regular and focused wave diffraction around a vertical cylinder using domain decomposition, Applied Ocean Research 29 (1–2) (2007) 55–71.

[29] W. Bai, R. E. Taylor, Fully nonlinear simulation of wave interaction with fixed and floating flared structures, Ocean Engineering 36 (3–4) (2009) 223–236.

[30] H. G. Sung, S. T. Grilli, BEM Computations of 3-D Fully Nonlinear Free-Surface Flows Caused by Advancing Surface Disturbances, International Journal of Offshore and Polar Engineerin 18 (2008) 292–301.

[31] M. Hannan, W. Bai, K. Ang, Modeling of fully nonlinear wave radiation by submerged moving structures using the higher order boundary element method, Journal of Marine Science and Application 13 (1) (2014) 1–10.

[32] W. Bai, M. Hannan, K. Ang, Numerical simulation of fully nonlinear wave interaction with submerged structures: Fixed or subjected to constrained motion, Journal of Fluids and Structures 49 (2014) 534–553.

[33] E. Dombre, M. Benoit, D. Violeau, C. Peyrard, S. T. Grilli, Simulation of floating structure dynamics in waves by implicit coupling of a fully non-linear potential flow model and a rigid body motion approach, Journal of Ocean Engineering and Marine Energy 1 (1) (2014) 55–76.

[34] E. Guerber, M. Benoit, S. T. Grilli, C. Buvat, A fully nonlinear implicit model for wave interactions with submerged structures in forced or free motion, Engineering Analysis with Boundary Elements 36 (2012) 1151–1163.

[35] G. Chen, C. Kharif, S. Zaleski, J. Li, Two-dimensional Navier–Stokes simulation of breaking waves, Physics of Fluids 11 (1) (1999) 121–133.

[36] S. Guignard, R. Marcer, V. Rey, C. Kharif, P. Fraunié, Solitary wave breaking on sloping beaches: 2-D two phase flow numerical simulation by SL-VOF method, European Journal of Mechanics - B/Fluids 20 (1) (2001) 57–74.

[37] I. Hadžić, J. Hennig, M. Perić, Y. Xing-Kaeding, Computation of flow-induced motion of floating bodies, Applied Mathematical Modelling 29 (12) (2005) 1196–1210.

[38] M. A. Rahman, N. Mizutani, K. Kawasaki, Numerical modeling of dynamic responses and mooring forces of submerged floating breakwater, Coastal Engineering 53 (10) (2006) 799–815.

[39] Y. Li, M. Lin, Regular and irregular wave impacts on floating body, Ocean Engineering 42 (2012) 93–101.

[40] A. Ghasemi, A. Pathak, M. Raessi, Computational simulation of the interactions between moving rigid bodies and incompressible two-fluid flows, Computers & Fluids 94 (2014) 1–13.

[41] M. Okan, S. Umpleby, Free surface flow around arbitrary two-dimensional bodies by b-splines, International Shipbuilding Progress 32 (372) (1985) 182–187.

[42] D. Kring, P. Sclavounos, Numerical stability analysis for time-domain ship motion simulations, Journal of Ship Research 39 (4) (1995) 313–320.

[43] T. Peltzer, B. Rosenthal, W. Reppun, D. Kring, W. Milewski, B. Connell, Multi-body seakeeping design optimization, Proc. Eighth Int. Conf. Hydrodyn. (ICHD2008).

[44] B. Kim, Y. Shin, A NURBS Panel Method for Three-dimensional Radiation and Diffraction Problems, Journal of Ship Research 47 (2) (2003) 177–186.

[45] G.-D. Kim, C.-S. Lee, J. Kerwin, A B-spline based higher order panel method for analysis of steady flow around marine propellers, Ocean Engineering 34 (14-15) (2007) 2045–2060.

[46] Z. Gao, Z. Zou, A NURBS-based high-order panel method for three-dimensional radiation and diffraction problems with forward speed, Ocean Engineering 35 (11-12) (2008) 1271–1282.

[47] T. Hughes, J. Cottrell, Y. Bazilevs, Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement, Computer Methods in Applied Mechanics and Engineering 194 (39–41) (2005) 4135–4195.

[48] T. Hughes, Isogeometric analysis: Progress and challenges, Conference on Mathematical Methods for Curves & Surfaces (MMCS'08), 2008.

[49] J. Cottrell, T. Hughes, A. Reali, Studies of refinement and continuity in isogeometric structural analysis, Computer Methods in Applied Mechanics and Engineering 196 (41-44) (2007) 4160–4183.

[50] J. Cottrell, T. Hughes, Y. Bazilevs, Isogeometric Analysis: Toward Integration of CAD and FEA, John Wiley & Sons, 2009.

[51] C. Politis, A. Ginnis, P. Kaklis, K. Belibassakis, C. Feurer, An isogeometric BEM for exterior potential-flow problems in the plane, 2009, pp. 349–354.

[52] R. Simpson, S. Bordas, J. Trevelyan, T. Rabczuk, A two-dimensional isogeometric boundary element method for elastostatic analysis, Computer Methods in Applied Mechanics and Engineering 209-212 (2012) 87–100.

[53] K. Belibassakis, T. Gerostathis, K. Kostas, C. Politis, P. Kaklis, A. Ginnis, C. Feurer, A BEM-isogeometric method for the ship wave-resistance problem, Ocean Engineering 60 (2013) 53–67.

[54] T. Sederberg, J. Zheng, A. Bakenov, A. Nasri, T-splines and T-NURCCs, Vol. 22, 2003, pp. 477–484.

[55] M. Scott, X. Li, T. Sederberg, T. Hughes, Local refinement of analysis-suitable T-splines, Computer Methods in Applied Mechanics and Engineering 213-216 (2012) 206–222.

[56] X. Li, J. Zheng, T. W. Sederberg, T. J. Hughes, M. A. Scott, On linear independence of T-spline blending functions, Computer Aided Geometric Design 29 (1) (2012) 63–76.

[57] M. Borden, M. Scott, J. Evans, T. Hughes, Isogeometric finite element data structures based on Bézier extraction of NURBS, International Journal for Numerical Methods in Engineering 87 (1-5) (2011) 15–47.

[58] M. Scott, R. Simpson, J. Evans, S. Lipton, S. Bordas, T. Hughes, T. Sederberg, Isogeometric boundary element analysis using unstructured T-splines, Computer Methods in Applied Mechanics and Engineering 254 (0) (2013) 197–221.

[59] D. Thomas, M. Scott, J. Evans, K. Tew, E. Evans, Bézier projection: A unified approach for local projection and quadrature-free refinement and coarsening of NURBS and T-splines with particular application to isogeometric design and analysis, Computer Methods in Applied Mechanics and Engineering 284 (2015) 55–105.

[60] D. Benson, Y. Bazilevs, M.-C. Hsu, T. Hughes, A large deformation, rotation-free, isogeometric shell, Computer Methods in Applied Mechanics and Engineering 200 (13-16) (2011) 1367–1378.

[61] C. Verhoosel, M. Scott, T. Hughes, R. de Borst, An isogeometric analysis approach to gradient damage models, International Journal for Numerical Methods in Engineering 86 (1) (2011) 115–134.

[62] C. Verhoosel, M. Scott, R. De Borst, T. Hughes, An isogeometric approach to cohesive zone modeling, International Journal for Numerical Methods in Engineering 87 (1-5) (2011) 336–360.

[63] D. Schillinger, L. Dedè, M. Scott, J. Evans, M. Borden, E. Rank, T. Hughes, An isogeometric design-through-analysis methodology based on adaptive hierarchical refinement of NURBS, immersed boundary methods, and T-spline CAD surfaces, Computer Methods in Applied Mechanics and Engineering 249-252 (2012) 116–150.

[64] Z. Kacprzyk, K. Ostapska-Luczkowska, Isogeometric analysis as a new fem formulation-simple problems of steady state thermal analysis, Procedia Engineering 91 (2014) 87–92.

[65] R. Simpson, M. Scott, M. Taus, D. Thomas, H. Lian, Acoustic isogeometric boundary element analysis, Computer Methods in Applied Mechanics and Engineering 269 (2014) 265–290.

[66] A. Ginnis, K. Kostas, C. Politis, P. Kaklis, K. Belibassakis, T. Gerostathis, M. Scott, T. Hughes, Isogeometric boundary-element analysis for the wave-resistance problem using T-splines, Computer Methods in Applied Mechanics and Engineering 279 (0) (2014) 425–439.

[67] K. Kostas, A. Ginnis, C. Politis, P. Kaklis, Ship-hull shape optimization with a T-spline based BEM-isogeometric solver, Computer Methods in Applied Mechanics and Engineering 284 (2015) 611–622.

[68] V. Nguyen, C. Anitescu, S. Bordas, T. Rabczuk, Isogeometric analysis: An overview and computer implementation aspects, Mathematics and Computers in Simulation 117 (2015) 89–116.

[69] S. Lipton, J. Evans, Y. Bazilevs, T. Elguedj, T. Hughes, Robustness of isogeometric structural discretizations under severe mesh distortion, Computer Methods in Applied Mechanics and Engineering 199 (5–8) (2010) 357–373.

[70] T-splines, inc, http://www.tsplines.com.

[71] H. G. Sung, H. S. Choi, Implicit formulation with the boundary element method for nonlinear radiation of water waves, Engineering Analysis with Boundary Elements 34 (5) (2010) 511–529.

[72] L. Piegls, W. Tiller, The NURBS Book, 2nd Edition, Springer, 1996.

[73] Y. Bazilevs, V. Calo, J. Cottrell, J. Evans, T. Hughes, S. Lipton, M. Scott, T. Sederberg, Isogeometric analysis using T-splines, Computer Methods in Applied Mechanics and Engineering 199 (5-8) (2010) 229–263.

[74] E. Evans, M. Scott, X. Li, D. Thomas, Hierarchical T-splines: Analysis-suitability, Bézier extraction, and application as an adaptive basis for isogeometric analysis, Computer Methods in Applied Mechanics and Engineering 284 (2014) 1–20.

[75] X. Li, Some properties for analysis-suitable T-splines, Journal of Computational Mathematics 33 (4) (2015) 428–442.

[76] V. Mantic, A new formula for the C-matrix in the Somigliana identity, Journal of Elasticity 33 (3) (1993) 191–201.

[77] M. Scott, M. Borden, C. Verhoosel, T. Sederberg, T. Hughes, Isogeometric finite element data structures based on Bézier extraction of T-splines, International Journal for Numerical Methods in Engineering 88 (2) (2011) 126–156.

[78] C. A. Brebbia, J. Domínguez, Boundary Elements: An Introductory Course, second edition Edition, WIT Press, Computational Mechanics, 1992.

[79] J. Lachat, J. Watson, Effective numerical treatment of boundary integral equations: A formulation for three-dimensional elastostatics, International Journal for Numerical Methods in Engineering 10 (5) (1976) 991–1005.

[80] C. Adam, T. Hughes, S. Bouabdallah, M. Zarroug, H. Maitournam, Selective and reduced numerical integrations for NURBS-based isogeometric analysis, Computer Methods in Applied Mechanics and Engineering 284 (2015) 732–761.

[81] G. Xie, J. Zhang, Y. Dong, C. Huang, G. Li, An improved exponential transformation for nearly singular boundary element integrals in elasticity problems, International Journal of Solids and Structures 51 (6) (2014) 1322–1329.

[82] G. Xie, F. Zhou, J. Zhang, X. Zheng, C. Huang, New variable transformations for evaluating nearly singular integrals in 3d boundary element method, Engineering Analysis with Boundary Elements 37 (9) (2013) 1169–1178.

[83] B. I. Yun, A generalized non-linear transformation for evaluating singular integrals, International Journal for Numerical Methods in Engineering 65 (12) (2006) 1947–1969.

[84] J. Telles, R. Oliveira, Third degree polynomial transformation for boundary element integrals: Further improvements, Engineering Analysis with Boundary Elements 13 (2) (1994) 135–141.

[85] X.-W. Gao, The radial integration method for evaluation of domain integrals with boundary-only discretization, Engineering Analysis with Boundary Elements 26 (10) (2002) 905–916.

[86] X.-W. Gao, An effective method for numerical evaluation of general 2d and 3d high order singular boundary integrals, Computer Methods in Applied Mechanics and Engineering 199 (45–48) (2010) 2856–2864.

[87] J. C. F. Telles, A self-adaptive co-ordinate transformation for efficient numerical evaluation of general boundary element integrals, International Journal for Numerical Methods in Engineering 24 (5) (1987) 959–973.

[88] A. Brancati, M. Aliabadi, I. Benedetti, Hierarchical adaptive cross approximation gmres technique for solution of acoustic problems using the boundary element method, Computer Modeling in Engineering and Sciences 43 (2) (2009) 149–172.

[89] G. Wu, R. Taylor, The coupled finite element and boundary element analysis of nonlinear interactions between waves and bodies, Ocean Engineering 30 (3) (2003) 387–400.

[90] S. Boo, C. Kim, Fully nonlinear diffraction due to a vertical circular cylinder in a 3-D HOBEM numerical wave tank, Proceedings of the International Offshore and Polar Engineering Conference 3 (1996) 23–30.

[91] M. Tanaka, The stability of solitary waves, Physics of Fluids 29 (3) (1986) 650–655.

[92] D. Dutykh, D. Clamond, Efficient computation of steady solitary gravity waves, Wave Motion 51 (1) (2014) 86–99.

[93] M. Chern, A. Borthwick, R. Taylor, Simulation of non-linear free surface motions in a cylindrical domain using a chebyshev-fourier spectral collocation method, International Journal for Numerical Methods in Fluids 36 (4) (2001) 465–496.

[94] P. Ferrant, A coupled time and frequency approach for nonlinear wave radiation, In Eighteenth Symposium on Naval Hydrodynamics (1991) 67–83.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# HYBRID OPENMP/MPI PARALLELIZATION OF A HIGH–ORDER DISCONTINUOUS GALERKIN CFD/CAA SOLVER

**F. Bassi[1], A. Colombo[1], A. Crivellini[2] and M. Franciolini[2]**

[1]Università degli Studi di Bergamo, Dipartimento di Ingegneria e Scienze Applicate
Viale Marconi 5, 24044, Dalmine (BG), Italy
e-mail: {francesco.bassi,alessandro.colombo}@unibg.it

[2] Università Politecnica delle Marche, Dipartimento di Ingegneria Industriale e Scienze Matematiche
Via Brecce Bianche 12, 60131, Ancona, Italy
e-mail: {a.crivellini@,m.franciolini@pm.}univpm.it

**Keywords:** Discontinuous Galerkin, parallelization, CFD, CAA, OpenMP, HPC.

**Abstract.** *This paper describes the implementation of an hybrid OpenMP/MPI parallelization strategy in a Discontinuous Galerkin solver used for DNS and LES or CAA computations, to fruitfully exploit the modern massively parallel HPC facilities. It is usually believed that the sheared memory view of OpenMP can easily increase the parallel efficiency of codes dealing with multi–core clusters. The idea consists of running calculations on those machines restricting as much as possible the use of the MPI library to the communications between nodes and exploiting the shared memory paradigm within a node. However, in practice, the achievement of a real parallel performance gain is not straightforward. Moreover, as far as DG solvers are concerned, almost nothing is reported in the current literature about the hybrid MPI/OpenMP implementation. In this work a colouring algorithm has been employed for OpenMP. The resulting hybrid strategy performs quite satisfactory, since generally it is more efficient of the pure MPI implementation. However, the performances are heavily dependent on hardware platforms, as well as on computational details such as the polynomial order of space discretization or the number of computational elements. Several scalability tests have been performed, resulting in the conclusion that the best performance can be achieved only with a proper choice of the number of MPI partition and OpenMP threads to be used within a single node. The reliability of the method was here assessed by solving the Taylor Green vortex problem at Reynolds numbers equal to 800 and 1600 and the Linear Euler acustic scattering from a rigid sphere.*

## 1 Introduction

In recent years Discontinuous Galerkin (DG) methods have emerged as one of the most promising high–order discretization techniques for CFD. DG methods have been successfully applied to the simulation of turbulent flows by solving the Reynolds averaged Navier–Stokes (RANS) equations [1, 2, 3], and more recently, have also been found very well suited for the Direct Numerical Simulation (DNS) as well as Implicit Large Eddy Simulation (ILES) of turbulent flows [4, 5] due to their favourable dispersion and dissipation properties. One of the advantages of DG methods is the use of a compact stencil, which is independent of the degree of polynomial approximation and is thus well suited for massively parallel implementations.

The most common strategy for exploiting HPC facilities relies on the usage of the Message Passing Interface (MPI) library, which is suitable for both the shared and distributed memory platforms. As an alternative, on shared memory systems only, the Open Multi-Processing (OpenMP) paradigm can be used. While the former strategy is widely used in the CFD/CAA community, almost nothing is reported in literature about the latter with reference to DG methods. To the authors' knowledge only [6] partially deals with this subject. However, even if OpenMP is confined to shared memory systems, some advantages of this parallelization strategy over the MPI approach are clear. The first one is the small programming effort required for the parallelization, at least partial, of an executable. In fact, it is possible to use few work–sharing directives in the most time consuming section of the source code. With this approach a relevant reduction of the execution time can be obtained almost immediately using the multi–core CPUs. At a later stage the developer can eventually improve the parallel performance considering additional pieces of source code, resulting in an incremental development. The second theoretical advantage refers to the expected performance gain achievable on multi– and many–core platforms.

Especially for modern HPC facilities, the use of the MPI library within a computing node can be a not optimal choice when compared to a sheared memory-oriented parallelism that can drastically reduce the cost of the inter-core communication. For this reason, a hybrid OpenMP/MPI implementation is considered to improve the parallel efficiency of codes running on clusters of multi–core nodes. However, several papers highlight that for a large number of cores the achievement of high OpenMP parallel efficiencies is not trivial and it is not clear if all the applications and hardware platforms can effectively benefit from the hybrid approach [7]. There are several reasons for those evidences: 1) the code should be fully parallelized even in the OpenMP context; 2) the use of OpenMP implies overheads due to the initialization of parallel and/or work–sharing regions; 3) the overheads due to the explicit and implicit synchronizations of some of the work-sharing constructs; 4) the penalization due to false sharing of cache lines between processors; 5) the reduction of memory bandwidth of cache coherent Non Uniform Memory Access (ccNUMA) hardware used for the more recent and common multi–core chips.

Some of these arguments shed some light on the great influence of the hardware characteristics on the parallel effectiveness. At this regards it is particularly relevant the number of NUMA regions within a node and the resulting latency encountered when transferring data from the random access memory (RAM) of one of these regions to the cache memory of the core owned by a different region. It is worth mentioning that, to date, OpenMP does not supply any instruction to force the placement of one data in a particular NUMA region. The use of a first touch policy within the source code or, alternatively, the use of some operating system utilities/commands, such as `numactl`, are viable shortcuts for reaching this objective. In fact, they are able only partially to control the affinity of a data to a socket. Even if they can improve the OpenMP

performance, the effectiveness of these approaches may change with the platform hardware and its operating system. OpenMP efficiency is also strongly affected by the used compiler. The parallelization of the OpenMP directives is demanded to the compiler, thus depending on how efficiently it implements the standard. This is not the only compiler characteristic impacting the parallel efficiency. Just switching on the OpenMP option the effectiveness of the compiler optimizations may vary significantly mainly within a parallel loop. Additionally, the choiche of a proper parallelization strategy based on OpenMP is the key ingredient of an efficient solver. While with an MPI parallelization the use of a domain decomposition algorithm is almost unavoidable, the shared memory parallel programming context offers more opportunities. Here we explore a strategy commonly referred as colouring approach, which retains all the OpenMP advantages and it can be very simply introduced in a already existing MPI solver. Moreover, it is also well suited to deal with hardware accelerators, an opportunity considered in the current OpenMP 4.0 standard.

Despite DG discretization is particularly well suited for a pure MPI implementation, due to the compactness of the discretization which produces a limited amount of information exchanged between grid partitions, the proposed hybrid OpenMP/MPI implementation perform well as shown in the numerical tests reported in Section 6. However, the best performance can be achieved only with a proper choice of the number of partitions to be inserted within a single node. The advantages are mainly evident for large numbers of computational cores within a computing node, when the pure MPI implementation suffers for high communications overheads. Our results show at most a 38% gain in terms of parallel efficiency of the hybrid OpenMP/MPI implementation with respect to the pure MPI strategy.

## 2 DG approximation of the Navier–Stokes equations

The complete set of Navier–Stokes equations can be written as:

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = 0, \tag{1}$$

$$\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_j}(\rho u_j u_i + p\delta_{ji}) = \frac{\partial \tau_{ji}}{\partial x_j}, \tag{2}$$

$$\frac{\partial}{\partial t}(\rho e_0) + \frac{\partial}{\partial x_j}(\rho u_j h_0) = \frac{\partial}{\partial x_j}\left(u_i \tau_{ij} - q_j\right), \tag{3}$$

where the total energy $e_0$ and total enthalpy $h_0$, pressure $p$, the total stress tensor $\tau_{ij}$ and the heat flux vector $q_j$ are given by

$$\begin{aligned}
e_0 &= e + u_k u_k/2, \\
h_0 &= e_0 + p/\rho, \\
p &= (\gamma - 1)\rho\left(e_0 - u_k u_k/2\right), \\
\tau_{ij} &= 2\mu\left(S_{ij} - \frac{1}{3}\frac{\partial u_k}{\partial x_k}\delta_{ij}\right), \\
q_j &= -\frac{\mu}{\mathrm{Pr}}\frac{\partial h}{\partial x_j}.
\end{aligned}$$

Here $e$ is the internal energy, $h$ is the enthalpy, $\gamma = c_p/c_v$ is the ratio of gas specific heats, $\mathrm{Pr}$ is the molecular Prandtl number and

$$S_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$$

is the mean strain-rate tensor.

The governing equations (1)–(3) can be written in the following compact form

$$\frac{\partial \mathbf{u}}{\partial t} + \boldsymbol{\nabla} \cdot \mathbf{F}_c(\mathbf{u}) + \boldsymbol{\nabla} \cdot \mathbf{F}_v(\mathbf{u}, \boldsymbol{\nabla}\mathbf{u}) = \mathbf{0}, \tag{4}$$

where $\mathbf{u}, \in \mathbb{R}^M$ denote the vectors of the $M$ conservative variables and source terms, $\mathbf{F}_c, \mathbf{F}_v \in \mathbb{R}^M \otimes \mathbb{R}^N$ denote the inviscid and viscous flux functions, respectively, and $N$ is the space dimension. In order to obtain the weak formulation of the equations, one can multiply equation (4) by an arbitrary test function $\phi$ and integrate over the domain $\Omega$. Integrating by parts the divergence terms, the weak form of equation (4) reads:

$$\int_\Omega \phi \frac{\partial \mathbf{u}}{\partial t}\,\mathrm{d}\mathbf{x} - \int_\Omega \boldsymbol{\nabla}\phi \cdot \mathbf{F}(\mathbf{u}, \boldsymbol{\nabla}\mathbf{u})\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \phi \mathbf{F}(\mathbf{u}, \boldsymbol{\nabla}\mathbf{u}) \cdot \mathbf{n}\,\mathrm{d}\sigma = \mathbf{0}, \tag{5}$$

where $\mathbf{F}$ is the sum of the inviscid and viscous fluxes.

In order to build a DG discretization of equation (5), we consider an approximation $\Omega_h$ of $\Omega$ and a triangulation $\mathcal{T}_h = \{K\}$ of $\Omega_h$, *i.e.*, a set of $ne$ non-overlapping elements $K$ not necessarily simplices. We denote with $\mathcal{F}_h^i$ the set of internal element faces, with $\mathcal{F}_h^b$ the set of boundary element faces and with $\mathcal{F}_h = \mathcal{F}_h^i \cup \mathcal{F}_h^b$ their union. We moreover set

$$\Gamma_h^i = \bigcup_{F \in \mathcal{F}_h^i} F, \qquad \Gamma_h^b = \bigcup_{F \in \mathcal{F}_h^b} F, \qquad \Gamma_h = \Gamma_h^i \cup \Gamma_h^b. \tag{6}$$

The solution is approximated on $\mathcal{T}_h$ as a piecewise polynomial function possibly discontinuous on element interfaces, *i.e.*, we assume the following space setting for each component $u_{h_i} = u_{h_1}, \ldots, u_{h_M}$ of the numerical solution $\mathbf{u}_h$:

$$u_{h_i} \in \Phi_h \stackrel{\text{def}}{=} \left\{\phi_h \in L^2(\Omega) : \phi_h|_K \in \mathbb{P}_k(K)\ \forall K \in \mathcal{T}_h\right\} \tag{7}$$

for some polynomial degree $k \geq 0$, $\mathbb{P}_k(K)$ being the space of polynomials of global degree at most $k$ on the element $K$. To overcome the ill-conditioning of elemental mass matrices for higher-order polynomials on high aspect ratio and curved elements we have choose a hierarchical and orthogonal set of shape functions defined in physical space. This set is obtained by using a modified Gram-Schmidt procedure, considering as a starting point a set of monomial functions of the same degree $k$.

Following Brezzi et al. [8], it is also convenient to introduce the jump and average trace operators, which on a generic internal face $F \in \mathcal{F}_h^i$ are defined as:

$$[\![q]\!] \stackrel{\text{def}}{=} q^+\mathbf{n}^+ + q^-\mathbf{n}^-, \qquad \{\cdot\} \stackrel{\text{def}}{=} \frac{(\cdot)^+ + (\cdot)^-}{2}, \tag{8}$$

where $q$ denotes a generic scalar quantity and the average operator applies to scalars and vector quantities. By definition, $[\![q]\!]$ is a vector quantity. These definitions can be suitably extended to faces intersecting $\partial\Omega$ accounting for the weak imposition of boundary conditions.

The discrete counterpart of equation (5) for a generic element $K \in \mathcal{T}_h$ then reads:

$$\int_K \phi_h \frac{\partial \mathbf{u}_h}{\partial t} \, \mathrm{d}\mathbf{x} - \int_K \boldsymbol{\nabla}_h \phi_h \cdot \mathbf{F}(\mathbf{u}_h, \boldsymbol{\nabla}_h \mathbf{u}_h) \, \mathrm{d}\mathbf{x} + \int_{\partial K} \phi_h \mathbf{F}(\mathbf{u}_h|_K, \boldsymbol{\nabla}_h \mathbf{u}_h|_K) \cdot \mathbf{n} \, \mathrm{d}\sigma = \mathbf{0}. \quad (9)$$

Due to the discontinuous functional approximation, the flux function in the third term of equation (9) is not uniquely defined. Moreover, a consistent and stable discretization of viscous fluxes must account for the effect of interface discontinuities on the gradient $\boldsymbol{\nabla}_h \mathbf{u}_h$. Uniqueness of interface fluxes is achieved by introducing a suitably defined numerical flux which couples the solution in adjacent elements and ensures local conservation.

The inviscid $\widehat{\mathbf{F}}_c$ and viscous $\widehat{\mathbf{F}}_v$ parts of the numerical flux are treated independently. For the former, exact Riemann flux, approximate Roe flux–difference splitting or the approximate Van Leer flux-difference splitting as modified by Hänel [9] were used, while for the viscous flux discretization we adopted the BR2 scheme presented in[10, 11] and theoretically analyzed in [8, 12] (where it is referred to as BRMPS). The numerical viscous flux is given by:

$$\widehat{\mathbf{F}}_v \left( \mathbf{u}_h^{\pm}, (\boldsymbol{\nabla}_h \mathbf{u}_h + \eta_F \mathbf{r}_F([\![\mathbf{u}_h]\!]))^{\pm} \right) \stackrel{\text{def}}{=} \{ \mathbf{F}_v \left( \mathbf{u}_h, \boldsymbol{\nabla}_h \mathbf{u}_h + \eta_F \mathbf{r}_F([\![\mathbf{u}_h]\!]) \right) \}, \quad (10)$$

where, according to [8, 12], the penalty factor $\eta_F$ must be greater than the number of faces of the elements. A very interesting feature of the outlined viscous flux discretization scheme is that it couples only the unknowns already coupled by the inviscid flux discretization scheme, irrespective of the degree of polynomial approximation of the solution.

Summing equation (9) over the elements and accounting for the above considerations, we obtain the DG formulation of problem (5), which then requires to find $u_{h_1}, \ldots, u_{h_M} \in \Phi_h$ such that:

$$\begin{aligned} \int_{\Omega_h} \phi_h \frac{\partial \mathbf{u}_h}{\partial t} \, \mathrm{d}\mathbf{x} - \int_{\Omega_h} \boldsymbol{\nabla}_h \phi_h \cdot \left( \mathbf{F}_c \left( \mathbf{u}_h \right) + \mathbf{F}_v \left( \mathbf{u}_h, \boldsymbol{\nabla}_h \mathbf{u}_h + \mathbf{r} \left( [\![\mathbf{u}_h]\!] \right) \right) \right) \, \mathrm{d}\mathbf{x} \\ + \int_{\Gamma_h} [\![\phi_h]\!] \cdot \left( \widehat{\mathbf{F}}_c \left( \mathbf{u}_h^{\pm} \right) + \widehat{\mathbf{F}}_v \left( \mathbf{u}_h^{\pm}, (\boldsymbol{\nabla}_h \mathbf{u}_h + \eta_F \mathbf{r}_F([\![\mathbf{u}_h]\!]))^{\pm} \right) \right) \, \mathrm{d}\sigma = \mathbf{0}, \end{aligned} \quad (11)$$

for all $\phi_h \in \Phi_h$. The lifting operator $\mathbf{r}_F$, which is assumed to act on the jumps of $\mathbf{u}_h$ componentwise, is defined as the solution of the following problem:

$$\int_{\Omega_h} \boldsymbol{\phi}_h \cdot \mathbf{r}_F \left( \mathbf{v} \right) \, \mathrm{d}\mathbf{x} = - \int_F \{ \boldsymbol{\phi}_h \} \cdot \mathbf{v} \, \mathrm{d}\sigma, \forall \boldsymbol{\phi}_h \in \left[ \Phi_h \right]^N, \, \mathbf{v} \in \left[ L^1 \left( e \right) \right]^N, \quad (12)$$

and the function $\mathbf{r}$ is related to $\mathbf{r}_F$ by the equation:

$$\mathbf{r}(\mathbf{v}) \stackrel{\text{def}}{=} \sum_{F \in \mathcal{F}_h} \mathbf{r}_F(\mathbf{v}). \quad (13)$$

At the boundary of the domain, the numerical flux function appearing in equation (11) must be consistent with the boundary conditions of the problem. In practice, this is accomplished by properly defining a boundary state which accounts for the boundary data and, together with the internal state, allows to compute the numerical fluxes and the lifting operator on the portion $\Gamma_h^b$ of the boundary $\Gamma_h$, see [10, 13].

## 3 Time integration

Numerical integration of Eq. (11) by means of suitable Gauss quadrature rules leads to a system of nonlinear ODEs that can be written as

$$\mathbf{M}\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} + \mathbf{R}\left(\mathbf{U}\right) = \mathbf{0}, \tag{14}$$

where $\mathbf{U}$ is the vector of the degrees of freedom, $\mathbf{R}\left(\mathbf{U}\right)$ is the residuals vector, and $\mathbf{M}$ is the block diagonal mass matrix, which is, in our implementation, equal to the identity matrix due to the choice of orthonormalized shape functions. It is worth noting that the degrees of freedoms (DOFs) are related to the solution via the relation

$$u_{h,i} = U_{i,j}\phi_j$$

for all the elements $K \in \mathcal{T}_h$, where $i = 1, ..., m$, being $m$ the number of conservative variables, and $j = 1, ..., N_{dof}^K$, being $N_{dof}^K$ the number of degrees of freedom on the element $K$. In this paper we integrate in time the system (14) by using the explicit three stage, order three, strong stability–preserving Runge–Kutta scheme of Gottlieb et al. [14]. The time integration of Eq. (14) by means of this low-storage scheme can be written as

$$\begin{aligned}
\mathbf{U}^{(0)} &= \mathbf{U}^n, \\
\mathbf{U}^{(i)} &= \sum_{k=0}^{i-1} \left(\alpha_{i,k}\mathbf{U}^{(k)} + \Delta t \beta_{i,k}\mathbf{R}(\mathbf{U}^{(k)})\right), \qquad i = 1, ..., s \\
\mathbf{U}^{n+1} &= \mathbf{U}^{(s)}.
\end{aligned} \tag{15}$$

where $s$ is the number of stages, $\alpha_{i,k}, \beta_{i,k}$ are the coefficients of the scheme [14]. For the sake of convenience, the dependence of the residuals vector by the independent variables $\mathbf{U}$ will be omitted in the rest of the paper, moreover $\mathbf{R}$ will be splitted into the volumes, $\mathbf{R}_K$ with $K \in \mathcal{T}_h$, and the surfaces, $\mathbf{R}_F$ with $F \in \mathcal{F}_h$, contributions. The entries of $\mathbf{R}_K$ and $\mathbf{R}_F$ can be found by comparison with Eq. (11).

## 4 Parallel OpenMP implementation

In order to implement the hybrid parallelization strategy, based upon the MPI and OpenMP paradigms, it is first convenient to study an efficient and practical use of the OpenMP within our existing algorithm. This can be done by considering the main characteristics of a modal DG framework. In fact, it is typically convenient to arrange by elements the main arrays $\mathbf{U}$ and $\mathbf{R}$ and to perform the evaluation of the latter vector according to two main loops. In the first one, performed over the $ne$ elements, all the volume contributions, $\mathbf{R}_K$, are evaluated. In the second, performed over the faces $F \in \mathcal{F}_h$, the numerical fluxes contributions $\mathbf{R}_F$ are handled. Again, the latter operation can be splitted in two parts: the first covers all the internal faces, $\mathbf{R}_{F_i}$ with $F \in \mathcal{F}_h^i$, the second takes in account all the boundaries faces, $\mathbf{R}_{F_b}$ with $F \in \mathcal{F}_h^b$.

In the contest of OpenMP, the evaluation of the $\mathbf{R}_K$ contribution to the residual vector can be safely parallelized, for example using the Fortran `!$OMP do` work sharing directive, without data race condition problems. The use of this straightforward strategy is not possible for the loop over the faces, since for each face the residuals of both the $K^-$ and $K^+$ elements are simultaneously updated. Performing this loop in parallel require special attention, since two threads may overwrite, at once, the same memory location of the $\mathbf{R}$ vector.

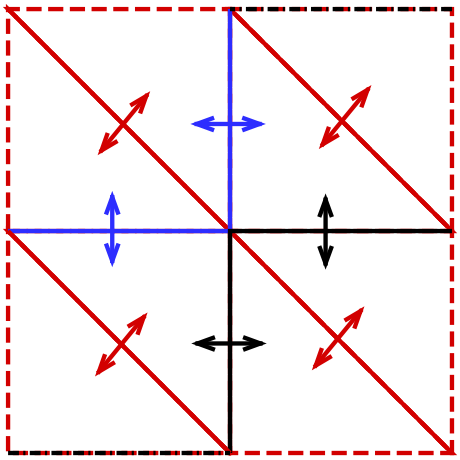


Figure 1: Representation of the colouring algorithm

To solve this issue we adopted a colouring strategy. The colouring procedure is based on a naive and simple to code algorithm similar to that used in [15, 16]. Here the main difference is that the colouring procedure is applied to the faces and not to the elements. The key idea is that one element cannot have more than one face of the same colour. The algorithm is described as follows: when the face $F$ has to be coloured, one must check the already assigned colours of the neighbour faces. The scope is to assign a different colour to $F$. If there is not a available colour in the current palette the number of colours is increased by one and this new colour is used for $F$. Note that the neighbour faces correspond to all the faces owned by the two elements connected by the face $F$. Once that all the faces are coloured the contribution of numerical fluxes to the residual vector can be performed in parallel for all the faces of the same colour, ensuring that the data race condition is avoided. In Figure 1 the $\mathbf{R}_{F_i}$ components due to the four red internal faces can be summed concurrently (in the figure this operation is represented by an arrow) to the $\mathbf{R}$ global vector, the same consideration holds true for the two black and for the two blue faces. It is clear that this algorithm is not optimized to obtain a group of colours characterized by an equal number of faces. However, it is fast and it performs well as will be shown in the remaining of the paper. Although it is evident that the colours distribution change with the initial ordering of the faces, as well as with grid reordering, we verified that the overall parallel performance is not significantly affected. Moreover, as pointed out in [16], it is possible to add a a posteriori phase to the algorithm to balance the number of faces for each colour.

Regarding the boundary conditions, the same algorithm is used. Indeed, in this circumstance, almost all the boundary faces have the same colour and only if one element owns more than one boundary face a further colour is needed. Even in this case all the boundary of the same colour can be safely evaluated in parallel, see Figure 1 in which the boundaries are depicted with dashed lines. Finally, it is worth mentioning that the selected algorithm is well suited to deal with accelerators, such as GPUs [16] or the Xeon Phi coprocessors in its offload mode, and that it can be easily implemented in the context of an MPI solver resulting in a hybrid strategy.

Hybrid MPI/OpenMP implementation can also take advantage of possibly non–blocking communications which are overlapped with computations. More in detail, the $\mathbf{R}_{F_i}$ evaluation and the exchange of messages required for updating the ghost elements DOFs are simultaneously performed. Note that for the diffusive volume terms our DG implementation uses the

BR2 scheme [1], which relies on the lifting operator $\mathbf{r}\left([\![\mathbf{u}_h]\!]\right)$, that depends on the jumps of the variables at the faces. It is more computationally convenient to evaluate these jumps in the loop over the faces and to compute only later the $\mathbf{R}_K$ residual contributions. Due to this implementation it is clearly not possible to overlap the last task with the MPI communications in the viscous case.

## 5 The Taylor Green Vortex problem

The Taylor Green Vortex problem is a canonical test case used for studying the mechanism upon which large scales vortices interact with each other producing turbulence decay and thus allowing the study of the energy dissipation processes. Since the geometry, consisting on a periodic cube with sides of length $2\pi L$, is considerably simple, it has been used extensively for testing the effects of numerical methods, grid resolution and turbulence modelling strategies like DNS and LES on the simulation of turbulent flows.

The problem consists of a cubic volume of fluid that contains a smooth initial distribution of vorticity, given by the following initial field

$$u = V_0 \sin\left(\frac{x}{L}\right) \cos\left(\frac{y}{L}\right) \cos\left(\frac{z}{L}\right), \tag{16}$$

$$v = -V_0 \cos\left(\frac{x}{L}\right) \sin\left(\frac{y}{L}\right) \cos\left(\frac{z}{L}\right), \tag{17}$$

$$w = 0, \tag{18}$$

$$\rho = \rho_0, \tag{19}$$

$$p = p_0 + \frac{1}{16}\left(\cos\left(\frac{x}{L}\right) + \cos\left(\frac{y}{L}\right)\right)\left(\cos\left(\frac{z}{L}\right) + 2\right). \tag{20}$$

The initial vortices start interacting together, resulting in a transition to turbulence and in the following turbulence decay for sufficiently high Reynolds numbers. As there is no forcing term on the equations, the flow field will slowly dissipate all the initial kinetic energy and it will come to rest.

The flow conditions are those used in [17]. The Mach number $M$ is equal to 0.1, resulting in a nearly–incompressible simulation, and the Prandtl number is 0.72. Two different Reynolds numbers have been employed, $Re = \rho_0 V_0 L/\mu = 800$ and $1600$. Both the Reynolds numbers are high enough to make the solution unstable and to produce a nearly isotropic turbulent decay. In order to monitor the development of the turbulent flow over time, two quantities have been employed. The first one is the time derivative of the kinetic energy defined as

$$k \stackrel{\text{def}}{=} \frac{1}{16\pi^3 \rho_0} \int_\Omega \rho u_k u_k \, \mathrm{d}\mathbf{x}. \tag{21}$$

The second one is the theoretical dissipation rate, evaluated as

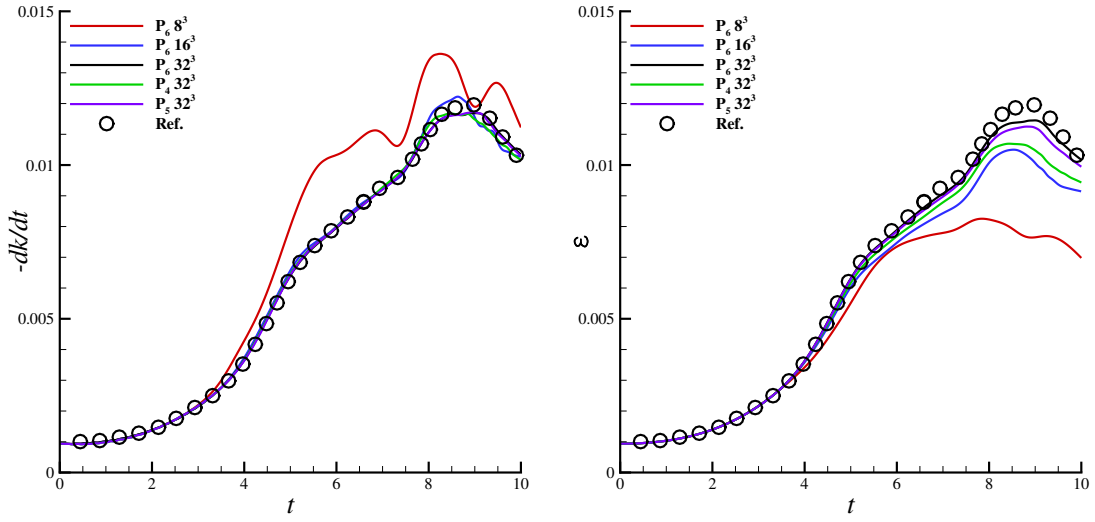
$$\varepsilon \stackrel{\text{def}}{=} \frac{2\mu}{8\pi^3 \rho_0} \int_\Omega \frac{\partial u_i}{\partial x_j} \frac{\partial u_i}{\partial x_j} \, \mathrm{d}\mathbf{x}. \tag{22}$$

Theoretically, in the incompressible limit, the following relation holds

$$-\frac{\mathrm{d}k}{\mathrm{d}t} = \varepsilon.$$

The numerical difference between these two quantities is thus somehow connected to the quality of the simulation and its numerical dissipation.

The solutions were advanced with a time step size corresponding to a maximum CFL number of $0.1$, for both the $\mathbb{P}_4$ and the $\mathbb{P}_5$ approximations, or to $0.08$, for $\mathbb{P}_6$. These values results in negligible time discretization errors. Figure 2 shows the kinetic energy time derivative and the theoretical dissipation rate for the case at $Re = 800$ for different grid sizes as well as for different order of polynomial approximation. The figure highlights the increment in resolution obtained increasing both the order of accuracy of the solution and the grid dimension. It is clear from Figure 2(a) that the evolution in time $t$, non–dimensionalized by $V_0$ and $L$, of the kinetic energy time derivative is not captured by the coarsest grid ($8^3$ elements). However, the solutions are in good agreement with the reference DNS data (Brachet et al. [18], digitalized from Ref. [17]) starting from the $16^3$ $\mathbb{P}_6$ case to the more resolved cases. Only a small difference can be noted between the $\mathbb{P}_5$ and the $\mathbb{P}_6$ solutions obtained on the finest grid. The largest value of this difference is close to $t = 8.5$, when the dissipation reaches its maximum value. Figure 2(b) depicts the theoretical dissipation rate evaluated via Eq.(22). In this case, slightly larger differences between $\varepsilon$ and reference data can be observed. This result suggests that for the most of




(a) Kinetic energy $k$ time derivative

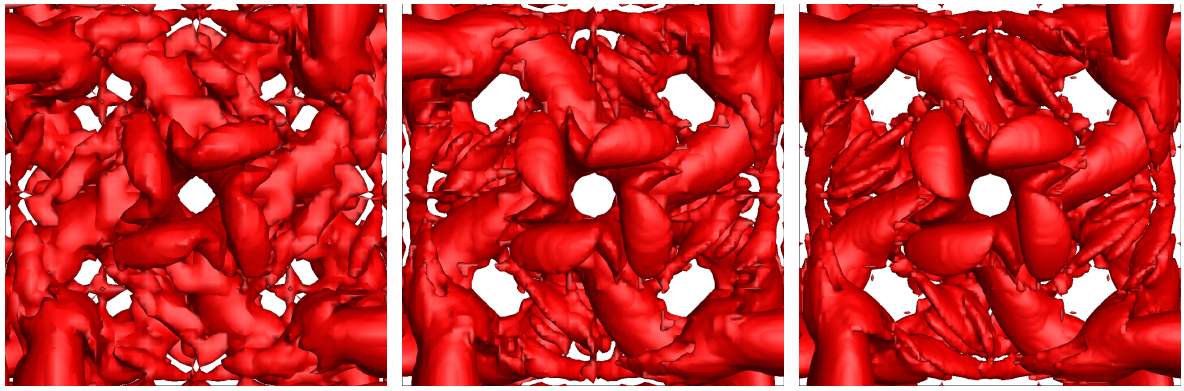
(b) Theoretical dissipation rate, $\varepsilon$

Figure 2: Taylor Green vortex at $Re = 800$. Results for different numerical resolutions.

the cases our computations under–resolve the turbulent scales. It is worth noting that the $16^3$ $\mathbb{P}_6$ and $32^3$ $\mathbb{P}_4$ solutions produce nearly the same numerical dissipation over time. However, they are obtained with a quite large difference of DOFs, which are approximately equal to $70^3$ for the former and to $105^3$ for the latter case. This behaviour shows the benefits of using an increased order of polynomial approximation instead of refining the mesh. Figure 3 shows the $\lambda_2 = -1.5$ iso–surface plot as reported in Ref. [17]. The resulting flow field resembles the reference one, and as the calculation is refined the turbulent structures become more clean and defined, especially for the $32^3$ $\mathbb{P}_6$ case.

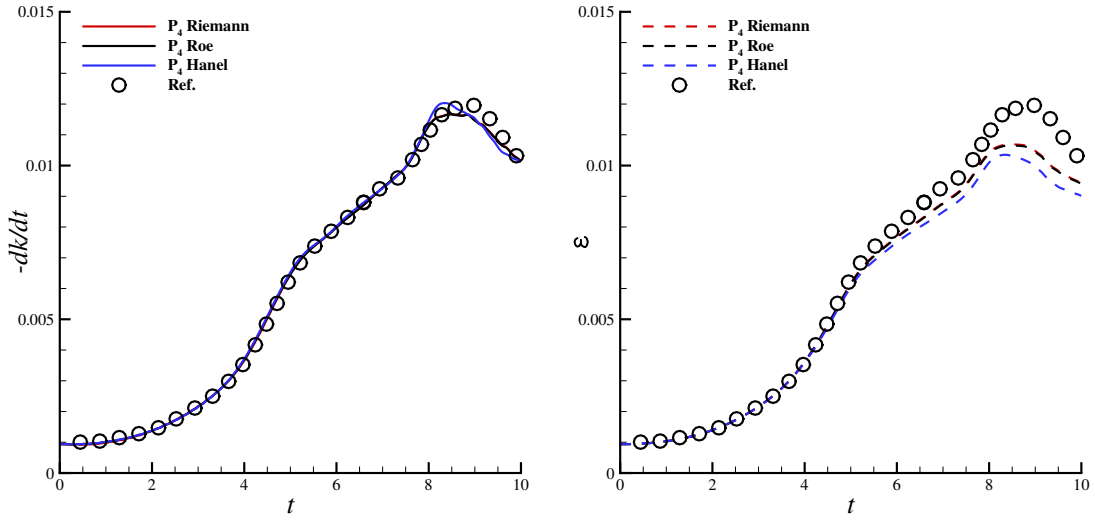
The effects of using different convective fluxes can be observed in Figure 4. The Roe approximate solver produces a negligible difference with respect to the exact Riemann solver, that exploits the numerical iterative procedure of Gottlieb et al. [19]. The Hänel flux splitting is revealed to be less accurate, as the theoretical dissipation rate, see Figure 4(b), is lower if compared to the reference data and to the results obtained with the other fluxes.

Figure 5 shows the results for $Re = 1600$, with the same grids and polynomial orders of



(a) $16^3$ $\mathbb{P}_6$ (b) $32^3$ $\mathbb{P}_4$ (c) $32^3$ $\mathbb{P}_6$

Figure 3: Taylor–Green vortex at $Re = 800$. $\lambda_2 = -1.5$ iso–surface at $t = 8.5$.



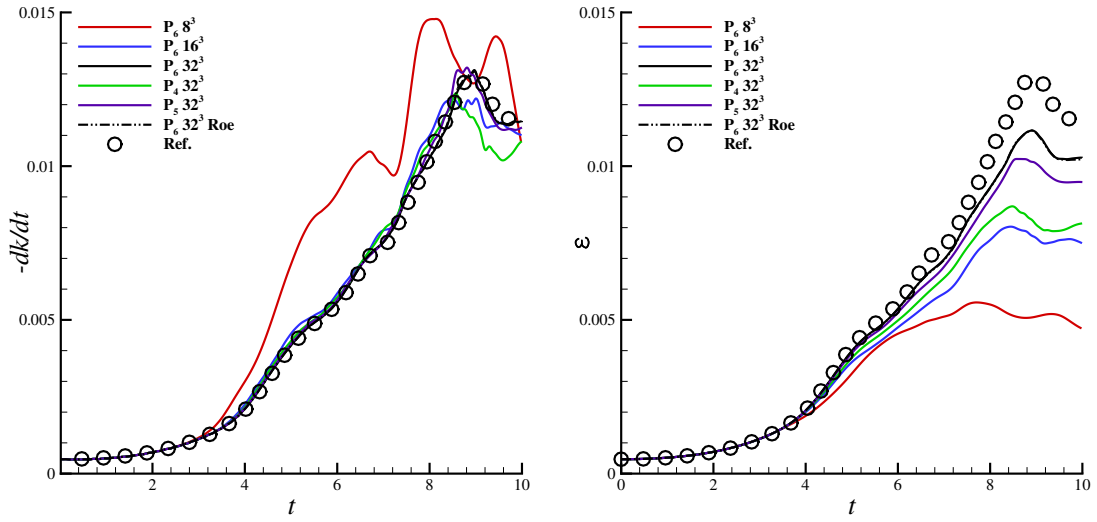

(a) Kinetic energy $k$ time derivative (b) Theoretical dissipation rate, $\varepsilon$

Figure 4: Taylor Green vortex at $Re = 800$. Results for different numerical fluxes at $32^3$ $\mathbb{P}_4$.

the previous test case. Since the smallest scales of turbulent decreases in size as the Reynolds number increases, the turbulence is clearly under–resolved. Figure 5(a) shows that the kinetic energy dissipation rate, for the test case $32^3$ $\mathbb{P}_6$, is in good agreement with the reference DNS $512^3$ pseudo–spectral solution [20]. Concerning the theoretical dissipation rate, see Figure 5(b), the difference from the reference data grows, showing that the numerical setup for this Reynolds number filters part of the smallest turbulent scales. Moreover, the use of the Roe numerical flux, employed for the finest test case, produces negligible difference in dissipation from the exact Riemann solver, but reduces of about 20% the computational time, see Table 1.

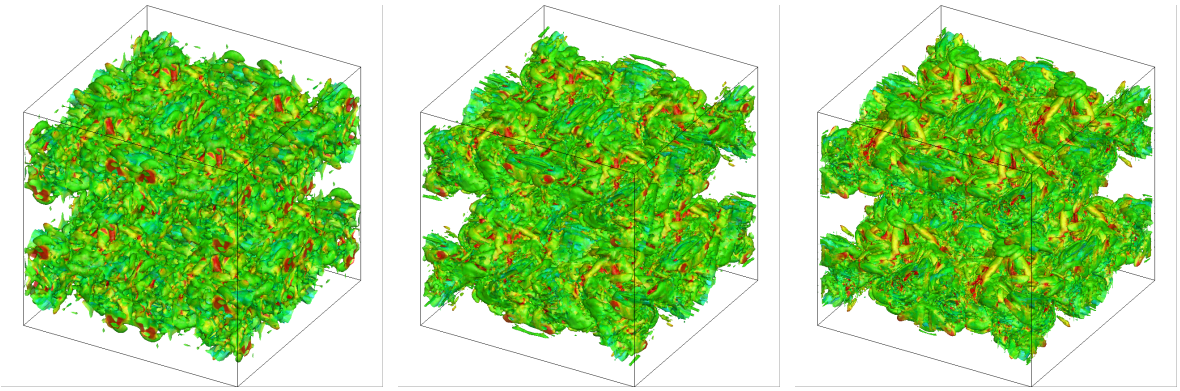
Figure 6 shows the $\lambda_2 = -1.5$ iso–surface coloured by the vorticity magnitude. It appears clear that as the resolution is increased the turbulent structures become smaller and more defined.

Figure 7 reports the sum $-\,\mathrm{d}k/\,\mathrm{d}t - \varepsilon$, which ideally should be zero. The differences from zero are related to the lack of numerical resolution. For both the solutions presented, as the polynomial order as well as the grid resolution increases, the imbalance tends to diminish. A



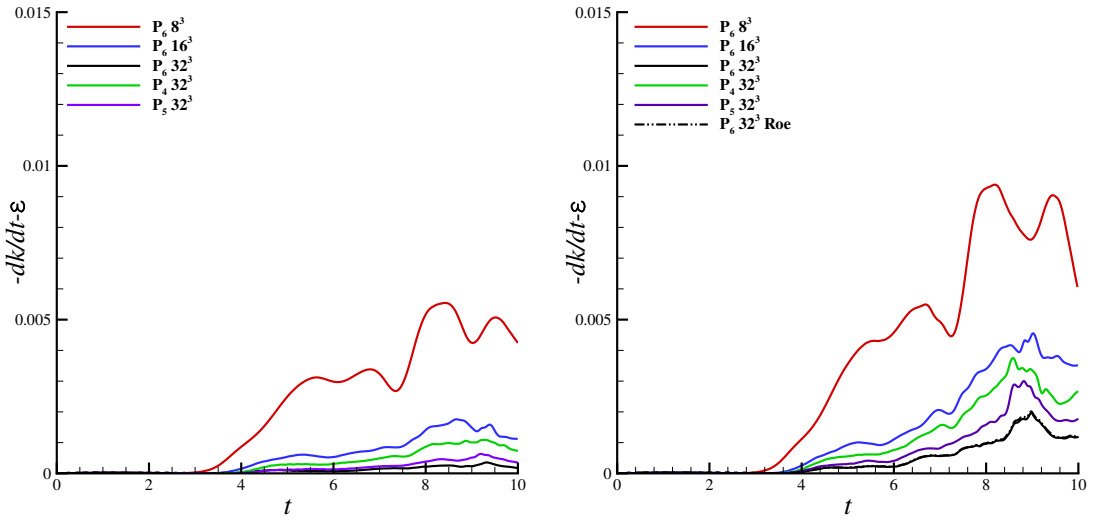

(a) Kinetic energy $k$ time derivative (b) Theoretical dissipation rate, $\varepsilon$

Figure 5: Taylor Green vortex at $Re = 1600$. Results for different numerical resolutions. Exact Riemann solver employed unless otherwise stated.



(a) $16^3$ $\mathbb{P}_6$ (b) $32^3$ $\mathbb{P}_4$ (c) $32^3$ $\mathbb{P}_6$

Figure 6: Taylor–Green vortex at $Re = 1600$. $\lambda_2 = -1.5$ iso–surface at $t = 10$. Iso–surface coloured by non–dimensional vorticity magnitude.




(a) $Re = 800$ (b) $Re = 1600$

Figure 7: Taylor Green vortex. $-\varepsilon - \mathrm{d}k/\mathrm{d}t$ imbalance as function of non–dimensional time.

peak value around $t = 8.5$ is produced, due to the fact that the dissipation reaches a maximum and thus the turbulent scales produced in the flow field are the smallest. This clearly stresses the resolution properties of the schemes. As shown in Figure 7(a), the $32^3$ $\mathbb{P}_6$ computation is able to solve almost completely the turbulent structures at $Re = 800$. However, the same numerical discretization filters the smallest scales as the Reynolds is increased, see Figure 7(b). Additionally, the Roe flux–difference splitting confirms its benefits over the exact Riemann solver, not affecting the numerical resolution but increasing the efficiency of the computation.

| Case | DOFs | $\Delta\varepsilon_1$ | $\Delta\varepsilon_2$ | Work Units |
|---|---|---|---|---|
| $8^3$ $\mathbb{P}_6$ | $35^3$ | $3.2 \cdot 10^{-1}$ | $7.3 \cdot 10^{-1}$ | $3.0 \cdot 10^3$ |
| $16^3$ $\mathbb{P}_6$ | $70^3$ | $8.8 \cdot 10^{-2}$ | $3.5 \cdot 10^{-1}$ | $5.3 \cdot 10^4$ |
| $32^3$ $\mathbb{P}_4$ | $105^3$ | $2.1 \cdot 10^{-1}$ | $2.9 \cdot 10^{-1}$ | $2.1 \cdot 10^5$ |
| $32^3$ $\mathbb{P}_5$ | $122^3$ | $6.1 \cdot 10^{-2}$ | $2.3 \cdot 10^-1$ | $5.6 \cdot 10^5$ |
| $32^3$ $\mathbb{P}_6$ | $140^3$ | $2.5 \cdot 10^{-2}$ | $1.5 \cdot 10^{-1}$ | $9.3 \cdot 10^5$ |
| $32^3$ $\mathbb{P}_6$, Roe | $140^3$ | $2.5 \cdot 10^{-2}$ | $1.5 \cdot 10^{-1}$ | $7.7 \cdot 10^5$ |

Table 1: Taylor Green vortex at $Re = 1600$. Error norms and work units for the hybrid MPI/OpenMP solver

The computational efforts, expressed in work units as defined in the 1st International Workshop on High-Order CFD Methods [21], for the $Re = 1600$ case are reported in Table 1 together with the following error norms

$$\Delta\varepsilon_1 = \frac{||dk/dt - (dk/dt)_{ref}||_{L\infty}}{\max\left(-(dk/dt)_{ref}\right)},$$
$$\Delta\varepsilon_2 = \frac{||-\epsilon - dk/dt||_{L\infty}}{\max(-(dk/dt)_{ref})},$$

where the subscript $ref$ indicates the Van Ress et al. [20] solution taken as reference in that Workshop also. It is worth noting that, since we use a polynomial approximation of global

degree at most of $k$ for a $\mathbb{P}_k$ solution, our finest simulation contains only about $140^3$ DOFs for each of the physical variable. Clearly the theoretical order of convergence is not reached, but it should be considered that the reported results are somehow under–resolved. Anyway, if compared with the results reported in [21], our data appear satisfactory both in terms of solution quality and computational efficiency.

## 6 Parallel performance of the Hybrid OpenMP/MPI

To obtain reliable indications about the parallel behaviour of the proposed hybrid implementation we report benchmarks results obtained on different multi–core machines, including clusters. The characteristics of each of these computational platforms (hardware and software) have a relevant influence on the parallel performance, thus a brief description of these architecture will precede each numerical experiment. If not otherwise stated, the TGV problem has been considered here to asses the performances of the the different parallel programming paradigms. The timings were measured advancing the solution for 10 Runge–Kutta steps, which ensures the independence of the results from eventual delay due to possible background processes of the operative system. We chose three different grid sizes, $8^3$, $16^3$ and $32^3$ as well as three different polynomial orders, $\mathbb{P}_1$, $\mathbb{P}_3$ and $\mathbb{P}_5$.

To get indications reflecting the real usage of the solver, they were used in a completely standard configuration, for instance even the $\|\mathbf{R}\|_{L_2}$ and $\|\mathbf{R}\|_{L_\infty}$ values were written to a file at the beginning of each time step. This task is performed exclusively to allow the user control of the computation status and it is not strictly required by the DG method. To maximize the performance we used in all the cases the OpenMP capability to bind threads to cores, while, as regards the threads affinity, generally we observed that the way in which the threads are assigned to the cores is not particularly relevant. The only exception is when a core pair of the AMD 6276 Opteron CPU is employed, as it will be explained in the next sections.

For the sake of compactness in this paper we will use the following notation to describe a parallel setting of a particular simulation: an hybrid MPI/OpenMP computation will be denoted by a triple of positive integers $(n, m, t)$, where $n$ is the number of the employed nodes, $m$ represents the MPI processes running in each of this node and $t$ reflects the OpenMP threads activated for each of the MPI process. Clearly the product of these three numbers indicates the total number of threads used, and usually this number is at most equal to the number of the available cores. For a pure MPI run we will use only the couple of numbers $(n, m)$, since obviously $t$ is always equal to 1. Finally, note that in this work all the computational grids were partitioned by means of the Metis [22] library.

### 6.1 Results on the 1st multi–core platform

The first system consists of two nodes interconnected by a point to point Infiniband network. Each node is equipped with two AMD Opteron 6276 processors, each with sixteen cores working at a fixed 2.3 GHz frequency, since the application power management (APM) of the operating system, Linux Debian 3.2.32-1, was disabled. Moreover even the AMD Turbo Core technology was also turned off since it can dynamically increase up to 3.2 GHz the operating frequency of a core. It is obvious that using these technologies it is not possible to produce reliable and repeatable indications about the OpenMP scalability.

The sixteen cores AMD 6276 CPU is a multi–chip module built by two eight–cores CPU die placed on a single die package (interconnected by a HyperTransport link). Each CPU in the package has its own memory controller, which is shared by its 8 cores. This configuration leads

to 4 NUMA regions, see [23]. Another relevant aspect of this hardware platform is that the cores are organized in core pair, sharing some components such as the floating point units and the L2 cache. Due to this hardware configuration the maximum overall performance of one socket should be not equal to sixteen times the single core performance. Indeed, we have verified that the usage of both the cores of the pair, running two identical instances of the serial DG solver, a significant performance reduction can be observed. In fact, the execution time becomes larger, approximately equal to $1.25$ times the standard wall clock time. Moreover, this performance degradation is not visible if cores of different pairs are employed in the computation. As a direct consequence, the theoretical maximum speed–up achievable using all the $32$ cores is limited to about $25.6$, corresponding to an apparent parallel efficiency of $80\%$.

The code is compiled, and optimized following the AMD indications found in [24], with the Open64 compiler suite. We verified that this choice, on our specific hardware, increases extensively the performances of our solver. The MVAPICH-2.2.1 library, which is an open source library suited for Infiniband networks and derived from the MPICH-3 library, was used for MPI communications. This library uses shared memory channels for the inter–node communications, thus the message passing between cores sharing the memory is faster then message passing between nodes.

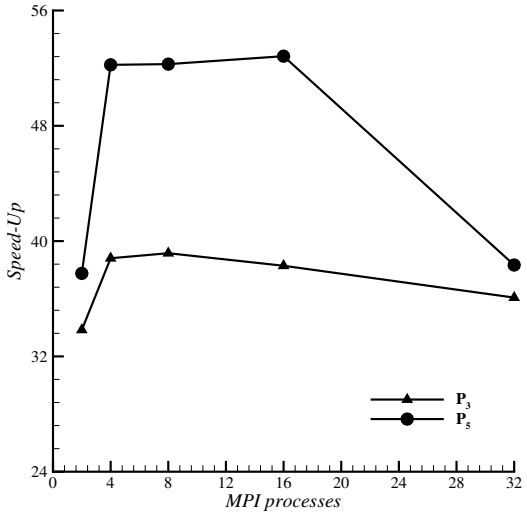



Figure 8: Hybrid MPI/OpenMP solver, Speed–up using 32 AMD Opteron 6276 cores on the $32^3$ grid.

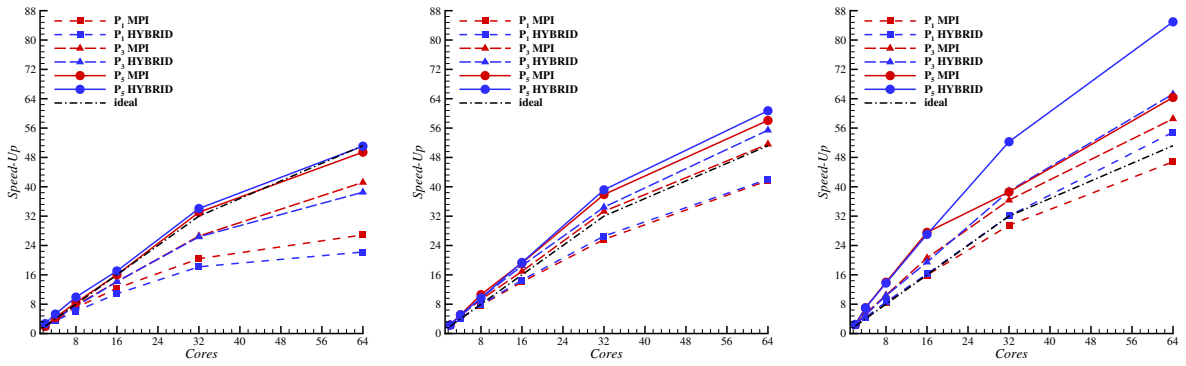
According to Guo et al. [25] we firstly investigated how the performance is affected by the subdivision of MPI partitions and OpenMP threads. In fact, those authors observed a significant performance reduction using $16$ or $32$ threads on a node of the HECToR Cray XE6 system, which consists of exactly two AMD Opteron 6276 CPUs. In their work this behaviour was ascribed to the latencies and the bandwidth reduction related to the presence of the 4 NUMA regions. To obtain a better parallel efficiency the authors used 4 MPI process per node and 8 OpenMP threads per each MPI process. The same idea has been applied to determine which configuration is able to exploit better performance of the hybrid code. For the $32^3$ grid a comparison between different strategies have been performed and reported in Figure 8. It should be noted that the tests have been done using only the half of the cores available on the two nodes to avoid the use of the core pairs and the intrinsic computational penalty. It can be seen clearly that the Speed–Up grows when the total MPI partitions are between 4 and 16 reaching a plateau.

This observation is confirmed both for the $\mathbb{P}_3$ case and for the $\mathbb{P}_5$ as well. From those evidences it can be deduced that the optimal use of the hybrid algorithm coincides with having a number of MPI partitions comparable to the number of NUMA regions. The strong scalability curves of Figure 9 has been obtained following this approach. Table 2 reports the resulting optimal subdivision between the OpenMP threads and MPI processes, using the compact notation introduced above.

| Cores | MPI | Hybrid MPI/OpenMP |
|---|---|---|
| 1 | $(1,1)$ | $(1,1,1)$ |
| 2 | $(1,2)$ | $(1,2,1)$ |
| 4 | $(1,4)$ | $(1,4,1)$ |
| 8 | $(1,8)$ | $(1,4,2)$ |
| 16 | $(1,16)$ | $(1,4,4)$ |
| 32 | $(2,16)$ | $(2,4,4)$ |
| 64 | $(2,32)$ | $(2,4,8)$ |

Table 2: Subdivision between MPI processes and OpenMP threads on the AMD Opteron 6276 system

In Figure 9 the Speed–Up trends for the three different grids, increasing the polynomial order from $\mathbb{P}_1$ to $\mathbb{P}_5$, are reported. Figure 9(a), which is related to the coarsest ($8^3$) grid, clearly highlights that the parallel efficiency raises with the polynomial order and that it reaches the ideal scaling at $\mathbb{P}_5$. Note that the pure MPI implementation is more efficient at the lower order approximation, but at $\mathbb{P}_5$ the hybrid version is slightly more convenient.




(a) $8^3$ grid (b) $16^3$ grid (c) $32^3$ grid

Figure 9: Performance of the hybrid MPI/OpenMP and the pure MPI solvers on four AMD Opteron 6276 CPUs for different grid sizes and $\mathbb{P}_1$, $\mathbb{P}_3$ and $\mathbb{P}_5$.

For the $16^3$ grid, see 9(b), the pure MPI code version performs considerably well, especially at $\mathbb{P}_5$ when the Speed–Up curve is considerably above the ideal scaling law. This behaviour is partially due to the poor performance of the serial version of the code in dealing with a large memory allocation which is typical of the high–order approximation. The hybrid curve is in this circumstance always above the one obtained using only the MPI parallelization strategy. Those observations and trends are confirmed increasing the grid size, see Figure 9(c). The results show that at highest computational loads the Speed–Up can be highly super–linear. It is worth noting

that with the $32^3$ grid, and $5^{\text{th}}$ degree of polynomial approximation, the amount of memory required exceeds 29Gb, and probably the TLB misses (Translation Lookaside Buffer) reduces the application performance. In this case the *HugePage* support in the Linux kernel should be enabled to limit this overhead. Moreover, the memory requirement exceeds the available RAM of each NUMA region (the total amount of RAM for each node is 64Gb) and the memory bandwidth and latency could be the bottlenecks. Just by partitioning the execution in two MPI processes, since every process deals with the half of the total memory, the parallel efficiency is highly super–linear, producing an efficiency of $120\%$. With 4 MPI processes the efficiency still raises up to $170\%$, and this gain is affecting all the other numerical tests. However only the hybrid implementation is able to retain this performance advantage (over the serial application) when $32$ and $64$ cores are used. Using all the cores this behaviour results in a performance improvement of about the $38\%$.

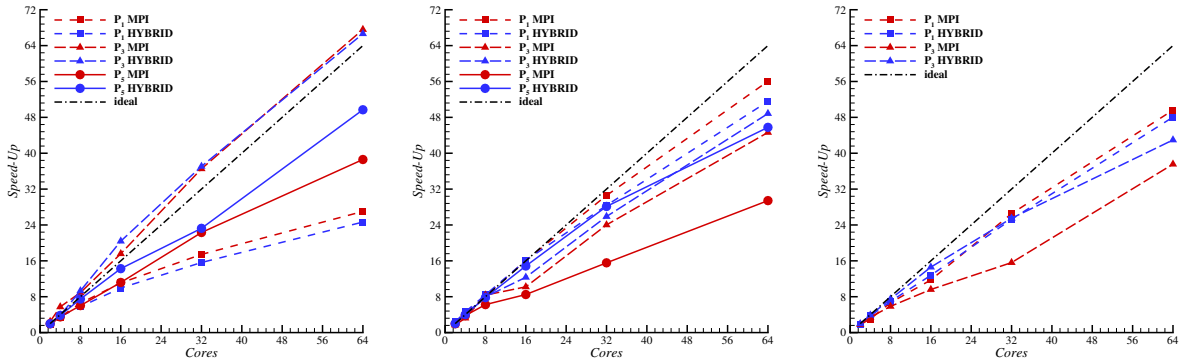
### 6.2 Results on the 2nd multi–core platform

The second cluster is built by 8 nodes interconnected by a gibabit ethernet that takes advantage of a single switch. This network serves also the Network File System (NFS). Due to these hardware characteristics the system is expected to scale poorly, especially when all the available 64 cores are involved in a parallel job. Each node consists of two NUMA regions equipped by a quad–cores AMD Opteron 2378 CPUs, running at $2.4$ GHz. The operating system is the same employed in the first machine, while in this circumstance we used the GNU compilers 4.7.2 and the MPICH–3.1.4 library.

The strong scalability curves are reported in Figure 10. Three different grid sizes, $8^3$, $16^3$ and $32^3$ have been employed for the tests. For the first two grids, $\mathbb{P}_1$, $\mathbb{P}_3$ and $\mathbb{P}_5$ polynomial approximations have been used, while for the finest one, the scalability has been performed only using $\mathbb{P}_1$ and $\mathbb{P}_3$ polynomial approximation, due to the limitation in memory for running the scalar version of the $\mathbb{P}_5$ case. Despite this platform produced quite counterintuitive results (the Speed–Up does not always increase raising the polynomial order and/or the grid size) the hybrid OpenMP/MPI algorithm produces a remarkable improvement of the performance over the pure MPI approach, especially as the order of polynomial approximation grows. The advantage is above 15% as far as cases of considerable computational load are considered. Also with this machine super–linear scalability effects, due to cache efficiency, can be observed in some cases, mainly at $8^3$ $\mathbb{P}_3$. The experimented subdivision between MPI partitions and OpenMP threads is reported in Table 3.

| Cores | MPI | Hybrid MPI/OpenMP |
|---|---|---|
| 1 | $(1,1)$ | $(1,1,1)$ |
| 2 | $(1,2)$ | $(1,2,1)$ |
| 4 | $(1,4)$ | $(1,2,2)$ |
| 8 | $(1,8)$ | $(1,2,4)$ |
| 16 | $(2,8)$ | $(2,2,4)$ |
| 32 | $(4,8)$ | $(4,2,4)$ |
| 64 | $(8,8)$ | $(8,2,4)$ |

Table 3: Subdivision between MPI processes and OpenMP threads on the AMD Opteron 2378 system




(a) $8^3$ grid (b) $16^3$ grid (c) $32^3$ grid

Figure 10: Performance of the hybrid MPI/OpenMP and the pure MPI solvers on sixteen AMD Opteron 2378 CPUs for different grid sizes and $\mathbb{P}_1$, $\mathbb{P}_3$ and $\mathbb{P}_5$.

### 6.3 Results on the BlueGene/Q platform

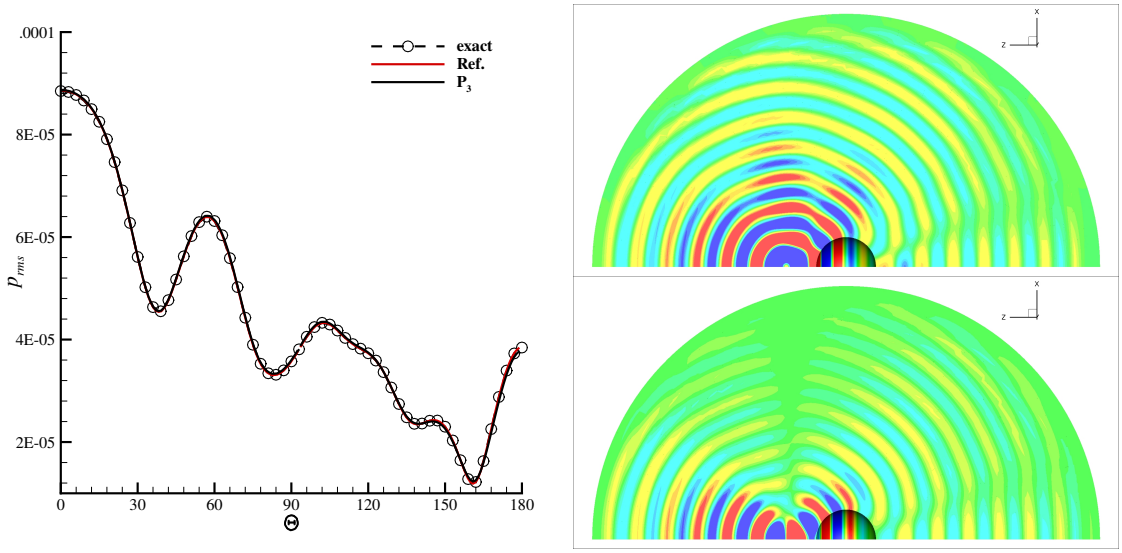
Although we still have to verify the performances of the hybrid Navier–Stokes DG solver on large HPC facilities, we already experimented the CAA (Computational Aeroacoustics) version of the current code on the IBM–BlueGene/Q platform. This machine, named FERMI, is available at CINECA, and it is characterized by a massively parallel architecture consisting of 10.240 nodes using sixteen cores of the IBM PowerPC A2 1.6 GHz chip. These cores are able to execute concurrently up to 4 hardware threads and better performance is expected exploiting more than one thread for each core. Each node is characterized by a symmetric access to all the memory, resulting in a Uniform Memory Access (UMA) machine.

The Problem 4 of Category 1 of Second Computational Aeroacoustics (CAA) Workshop on Benchmark Problems [26] was considered as test case. The problem of the acoustic scatter from a rigid sphere has been computed by solving the Linearized Euler Equations (LEE). The acoustic perturbations are generated by a spatially distributed acoustic source and the mean flow is at rest. The computational domain has been discretized using an unstructured grid consisting of 89628 hexahedral elements and the solution was computed using a $\mathbb{P}_3$ polynomial approximation.

Figure 11(a) shows that the computed acoustic directivity pattern compares very favourably with the analytical solution derived by Morris, [27], while Figure 11(b) presents a couple of pictures of the $p$ and $u_3$ contours that help clarifying the solution structure.

As regards the parallel performance, Figure 12(a) reports the measured Speed-Up. Using up to 16 threads, equal to the number of the cores, for each node we obtain good parallel performances adopting both the approaches. The hybrid MPI/OpenMP performs slightly better, for example its parallel efficiency at 4096 cores, corresponding to 22 elements per core, is 83.5% against the 77.2% pure MPI value. However, with 64 threads for node, 4 hardware threads for each core, the parallel performance of the hybrid code strongly outperforms that of the pure MPI one. This gain increases, at a fixed grid size, raising the number of the employed computational nodes. When using 16384 threads each process deals with an average value of 5.5 grid cells.

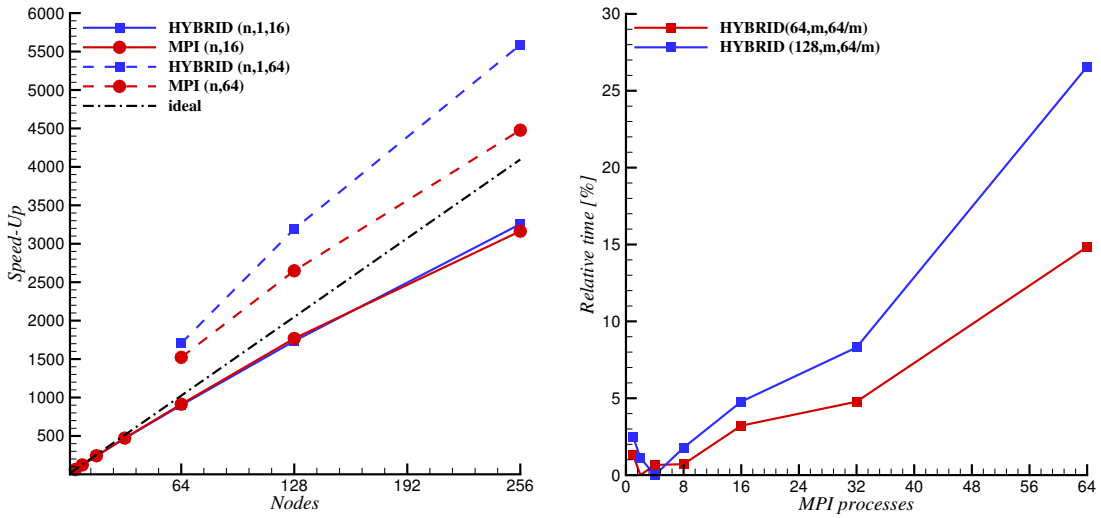
Figure 12(b) shows that the relative wall clock time, with respect to the best configuration, using all the 64 hardware threads and varying the combination between the MPI processes and the OpenMP threads. It comes clear that with this UMA platform the better performance is obtained by using a small number of MPI process for each node.




(a) Acoustic directivity pattern

(b) top $p$, bottom $u_3$

Figure 11: Acoustic scattering from a sphere. $\mathbb{P}_3$ solution, Ref. solution is from [28], exact solution is due to Morris [27]




(a) Strong scalability

(b) $(n, m, 64/m)$ relative time

Figure 12: Acoustic scattering from a sphere. Performance of the hybrid MPI/OpenMP and pure MPI implementations on the BlueGene/Q platform.

## 7 Conclusions

In this work we have considered two possibly OpenMP and hybrid OpenMP/MPI implementations of a high–order DG solver for CFD and CAA applications. OpenMP-based parallelism for shared memory architectures is usually expected to give large gains when compared to pure MPI implementations on massively parallel computers. Although the coding effort required by an OpenMP parallelization is quite small, obtaining an efficient solver and clear usage guidelines (optimal load repartition among OpenMP and MPI processes) is not trivial and many issues has to be considered. Here, a satisfactory parallel efficiency has been obtained using a simple yet effective colouring algorithm, tested on several problems.

Numerical experiments indicate that our OpenMP/MPI implementation improves the parallel efficiency of the solver on clusters made of largely multi-core nodes when compared to pure MPI, although special care is required with NUMA architectures to prevent the reduction of the memory bandwidth of CPUs. This suggest that an optimal usage of the computational resources can be obtained by placing a single MPI process in each NUMA region.

## 8 Acknowledgements

We acknowledge “Centro per le Tecnologie Didattiche e la Comunicazione” University of Bergamo for the resources provided by CINECA, within the “Convenzione di Ateneo Università degli Studi di Bergamo”, for the availability of high performance computing and support.

## REFERENCES

[1] F. Bassi, A. Crivellini, S. Rebay, and M. Savini. Discontinuous Galerkin solution of the Reynolds averaged Navier-Stokes and k - $\omega$ turbulence model equations. *Computers & Fluids*, (34):507–540, 2005.

[2] A. Crivellini, V. D’Alessandro, and F. Bassi. A Spalart-Allmaras turbulence model implementation in a discontinuous Galerkin solver for incompressible flows. *Journal of Computational Physics*, 241:388 – 415, 2013.

[3] A. Crivellini, V. D’Alessandro, and F. Bassi. High-order discontinuous Galerkin solutions of three-dimensional incompressible RANS equations. *Computers & Fluids*, 81:122 – 133, 2013.

[4] A. Crivellini, V. D’Alessandro, and F. Bassi. Assessment of a high-order discontinuous Galerkin method for incompressible three–dimensional Navier–Stokes equations: Benchmark results for the flow past a sphere up to Re=500. *Computers & Fluids*, 86:442 – 458, 2013.

[5] F. Bassi, L. Botti, A. Colombo, A. Crivellini, A. Ghidoni, and F. Massa. On the development of an implicit high-order Discontinuous Galerkin method for DNS and implicit LES of turbulent flows. *European Journal of Mechanics, B/Fluids*, 2015. article in press.

[6] C. C. de Wiart and K. Hillewaert. Development and validation of a massively parallel high-order solver for DNS and LES of industrial flows. In *IDIHOM: Industrialization of High-Order Methods-A Top-Down Approach*, pages 251–292. Springer, 2015.

[7] Martin J. Chorley and David W. Walker. Performance analysis of a hybrid MPI/OpenMP application on multi-core clusters. *Journal of Computational Science*, 1(3):168 – 174, 2010.

[8] F. Brezzi, G. Manzini, D. Marini, P. Pietra, and A. Russo. Discontinuous Galerkin approximations for elliptic problems. *Numer. Methods Partial Differential Equations*, 16:365–378, 2000.

[9] D. Hänel, R. Schwane, and G. Seider. On the accuracy of upwind schemes for the solution of the Navier–Stokes equations. AIAA Paper 87-1105 CP, AIAA, July 1987. Proceedings of the AIAA 8th Computational Fluid Dynamics Conference.

[10] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–108, Antwerpen, Belgium, March 5–7 1997. Technologisch Instituut.

[11] F. Bassi and S. Rebay. A high order discontinuous Galerkin method for compressible turbulent flows. In C.-W. Shu B. Cockburn, G.E. Karniadakis, editor, *Discontinuous Galerkin Methods. Theory, Computation and Applications*, volume 11 of *Lecture Notes in Computational Science and Engeneering*. Springer-Verlag, 2000. *First Internation Symposium on Discontinuous Galerkin Methods*, May 24–26, 1999, Newport, RI, USA.

[12] D. N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[13] F. Bassi and S. Rebay. High-order accurate discontinuous finite element solution of the 2D Euler equations. *J. Comput. Phys.*, 138:251–285, 1997.

[14] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001.

[15] Yohei Sato, Takanori Hino, and Kunihide Ohashi. Parallelization of an unstructured Navier–Stokes solver using a multi-color ordering method for OpenMP. *Computers & Fluids*, 88:496 – 509, 2013.

[16] Dimitri Komatitsch, David Michéa, and Gordon Erlebacher. Porting a high-order finite-element earthquake modeling application to NVIDIA graphics cards using CUDA. *Journal of Parallel and Distributed Computing*, 69(5):451 – 460, 2009.

[17] Gregor J Gassner and Andrea D Beck. On the accuracy of high-order discretizations for underresolved turbulence simulations. *Theoretical and Computational Fluid Dynamics*, 27(3-4):221–237, 2013.

[18] Marc E Brachet, Daniel I Meiron, Steven A Orszag, BG Nickel, Rudolf H Morf, and Uriel Frisch. Small-scale structure of the Taylor–Green vortex. *Journal of Fluid Mechanics*, 130:411–452, 1983.

[19] JJ Gottlieb and CPT Groth. Assessment of riemann solvers for unsteady one-dimensional inviscid flows of perfect gases. *Journal of Computational Physics*, 78(2):437–458, 1988.

[20] Wim M Van Rees, Anthony Leonard, DI Pullin, and Petros Koumoutsakos. A comparison of vortex and pseudo-spectral methods for the simulation of periodic vortical flows at high reynolds numbers. *Journal of Computational Physics*, 230(8):2794–2805, 2011.

[21] Z.J. Wang, Krzysztof Fidkowski, Rémi Abgrall, Francesco Bassi, Doru Caraeni, Andrew Cary, Herman Deconinck, Ralf Hartmann, Koen Hillewaert, H.T. Huynh, Norbert Kroll, Georg May, Per-Olof Persson, Bram van Leer, and Miguel Visbal. High-order CFD methods: current status and perspective. *International Journal for Numerical Methods in Fluids*, 72(8):811–845, 2013.

[22] G. Karypis and V. Kumar. METIS, a software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. Technical Report Version 4.0, University of Minnesota, Department of Computer Science/Army HPC Research Center, 1998.

[23] Advanced Micro Devices, Inc. AMD Opteron 6200 series processors, Linux tuning guide, 2012. Downloadable from `http://developer.amd.com/resources/documentation-articles/developer-guides-manuals/`.

[24] Advanced Micro Devices, Inc. AMD Opteron 4200/6200 series processors compiler options quick reference guide, 2012. Downloadable from `http://developer.amd.com/resources/documentation-articles/developer-guides-manuals/`.

[25] Xiaohu Guo, Michael Lange, Gerard Gorman, Lawrence Mitchell, and Michèle Weiland. Developing a scalable hybrid MPI/OpenMP unstructured finite element model. *Computers & Fluids*, 110:227 – 234, 2015. ParCFD 2013.

[26] C.K.W. Tam and J.C. Hardin. *Second Computational Aeroacoustics (CAA): Workshop on Benchmark Problems*. NASA conference publication. NASA, 1997.

[27] P.J. Morris. Scattering of Sound by a Sphere: Category 1: Problems 3 and 4. In C.K.W Tam and J.C. Hardin, editors, *Second Computational Aeroacoustics (CAA) Workshop on Benchmark Problems*, 1997. NASA CP 3352.

[28] A. Crivellini and F. Bassi. A three–dimensional parallel discontinuous Galerkin solver for acoustic propagation studies. *International Journal of Aeroacoustics*, 2(2):157–173, 2003.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE CHOICE OF SHOCK CAPTURING SCHEMES FOR THE SOLUTION OF THE LWR TRAFFIC FLOW EQUATION USING A HIGH ORDER MODAL DISCONTINUOUS GALERKIN DISCRETIZATION

**Alberto Costa Nogueira Junior**[1], **João Lucas de Sousa Almeida**[1], **Cláudio Alessandro de Carvalho Silva**[2]

[1]IBM Research Brazil
Rua Tutóia, 1157 - São Paulo/SP - 04007-900
e-mail: {albercn, jlsa}@br.ibm.com

[2] Thorus Scisoft Tecnologia
Rua Francisco Glicério, 1561 - Valinhos/SP - 13271-200
e-mail: cacs@thorus-scisoft.com.br

**Keywords:** Traffic Flow, LWR Equation, High-Order, Modal, Discontinuous Galerkin, Shock Capturing.

**Abstract.** *In this paper we studied different techniques to stabilize the numerical solution of the LWR traffic flow equation when using a modal version of the high order discontinuous Galerkin finite element method (DGFEM). Based on a set of three standard examples of traffic flow initial conditions, we compared four shock capturing strategies to control Gibbs phenomenon around discontinuous solutions of the car density conserved variable: (1) generalized slope limiter ($\Pi^N$), (2) hierarchical slope limiter, (3) sub-cell shock capturing with elementwise constant artificial viscosity, and (4) sub-cell shock capturing with local $C^0$ artificial viscosity. Although such stabilizing techniques were well documented for aerospace applications, our experiments revealed important features which are worth mentioning. Firstly, regardless of the polynomial order of the approximation, both limiters required quite refined meshes to achieve sharp shock resolution. Secondly, shock capturing based on elementwise constant viscosity produced small traveling spikes over the solution profiles destroying the locality of the method and resulting in excessive smearing of the discontinuities. Finally, $C^0$ artificial viscosity addition produced sharp shock resolutions with very coarse meshes and high order polynomial approximations. Overall, this technique demonstrated the better cost-benefit in terms of computational effort (number of degrees of freedom) to achieve a certain level of shock resolution.*

## 1 INTRODUCTION

The LWR nonlinear equation due to Lighthill, Witham and Richards [1, 2, 3] is one of the most popular PDEs used to describe traffic flow problems. The model dynamics is written as a scalar hyperbolic conservation law in one dimension [4]. The LWR mathematical formulation assumes a monotonically decreasing relation between speed and density of cars leading to a strictly concave form of the convective flux associated to the hyperbolic PDE [5].

In this paper we discretized the LWR equation using the modal version of the high order discontinuous Galerkin method (DGM) [6]. The Lax-Friedrichs' numerical flux was chosen to approximate the solution of Riemman problems at element boundaries within the mesh. Details of the DGM numerical formulation implemented in this work can be found in [7]. As the unsteady LWR scalar equation naturally develops shocks along time integration and we wanted to use high order approximations, we were specially interested in shock capturing strategies to stabilize solutions and get rid of undesirable wiggles.

The main goal of the present study is to implement and compare different shock capturing strategies to stabilize the numerical solution of the LWR equation and establish the most suited way to deal with Gibbs phenomenon in terms of shock resolution and computational efficiency. Four well known stabilizing techniques were considered to accomplish this task: (1) generalized slope limiter ($\Pi^N$) [8]; (2) hierarchical slope limiter [9]; (3) sub-cell shock capturing with elementwise constant artificial viscosity [11] and (4) sub-cell shock capturing with local $C^0$ artificial viscosity [12]. We used the LDG numerical flux [10, 8] to discretize the $2^{nd}$ order diffusive operator associated to the viscous formulation of the LWR equation when considering the stabilization techniques based on artificial viscosity. A five step $4^{th}$ order strong stability preserving Runge-Kutta explicit integrator [8] was used to evolve the semi-discrete formulation in time. Numerical examples for typical traffic flow situations were chosen to validate and compare the algorithmic performance. The following examples, with three different initial conditions, were evaluated: (a) red light at road exit; (b) green light at half way of road length and (c) exponential car density at road entrance.

This paper is organized as follows: section 2 introduces the LWR equation and its corresponding boundary value problem, section 3 briefly describes the DGFEM discretization used to numerically approximate the solution of the LWR boundary value problem, section 4 discusses the four different shock capturing strategies considered to stabilize the numerical solutions, section 5 discusses the results obtained when comparing all shock capturing mechanisms through a set of three typical examples of traffic flow initial conditions, finally, section 6 summarizes the main results and point out the strongest and weakest features of each shock capturing scheme used in our numerical experiments.

## 2 THE LWR TRAFFIC FLOW MODEL

To model traffic flow behavior along single roads the present study considers the Lighthill-Whitham-Richards (LWR) equation which can be written as follows,

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial f(u(x,t))}{\partial x} = 0,\ x \in [L,R] = \Omega, \tag{1}$$

with appropriate initial and boundary conditions given by

$$u(x,0) = u_0(x),\ u(L,t) = u_L(t),\ u(R,t) = u_R(t), \tag{2}$$

where $u(x,t)$ is the traffic density, $x \in \mathbb{R}$ and $t \in \mathbb{R}_+$ are the space and time coordinates and $\Omega$ is the problem domain, respectively.

The main conserved quantity in the LWR model is the traffic density which is defined as the average number of cars per unit length of road. A fundamental assumption of this model is the car speed as a monotonically decreasing function of the car density. Such hypothesis, postulated in [5], implies the following quadratic flux function $f : \mathbb{R} \to \mathbb{R}$ in equation 1,

$$f(u) = V_{max}\left(u - \frac{u^2}{u_{max}}\right) \tag{3}$$

where the constants $V_{max}$ and $u_{max}$ stand for maximum car speed and maximum car density, respectively.

Fig. 1 depicts the monotonically decreasing car speed as a function of car density as well as the resulting quadratic flux described in equation 3.

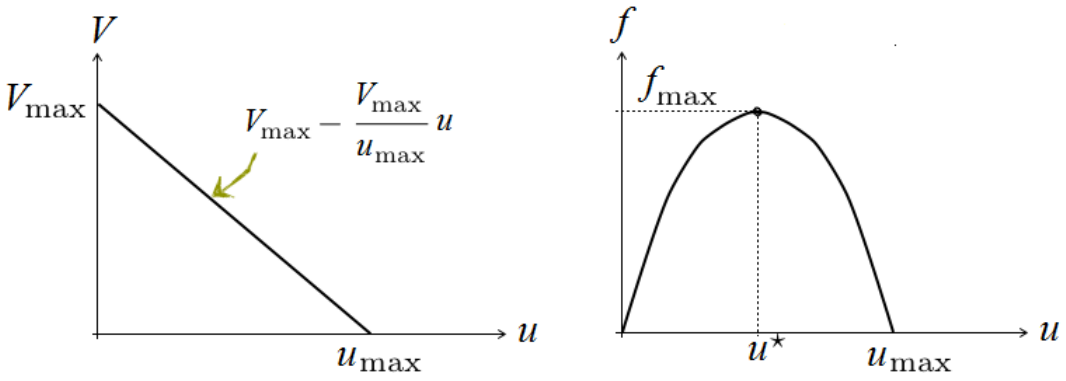



Figure 1: Monotonically decreasing car speed and quadratic traffic flux.

## 3 DGFEM DISCRETIZATION

Following a standard DGFEM approach, we rewrite the boundary value problem defined by equations 1 and 2 in a weak form valid for each grid element separately, and enforce a global approximate solution choosing a numerical flux at each element boundary to establish communication among elements.

We consider an approximation of $\Omega$ by $K$ non-overlapping elements, $x \in [x_l^k, x_r^k] = \Omega^k$ and for each grid element $\Omega^k$, we start forming the local residual as follows

$$x \in \Omega^k : \mathcal{R}_h(x,t) = \frac{\partial u_h^k}{\partial t} + \frac{\partial f_h^k}{\partial x}, \tag{4}$$

where $u_h^k$ and $f_h^k$ are numerical approximations to the traffic density, $u(x,t)$, and the physical flux, $f(u)$, respectively. We also assume that both functions can be expressed as polynomials of order N

$$x \in \Omega^k : \; u_h^k(x,t) = \sum_{j=1}^{N+1} \hat{u}_j^k(t)\psi_j^k(x), \tag{5}$$

$$f_h^k(u_h(x,t)) = \sum_{j=1}^{N+1} \hat{f}_j^k(t)\psi_j^k(x). \tag{6}$$

where $\psi_j^k(x)$ are polynomial basis functions belonging to a finite dimensional vector space $V_h^k = span\left\{\psi_j^k(\Omega^k)\right\}_{j=1}^{N+1}$. In this work, we considered the Legendre polynomials, $P_n(\xi)$, $\xi \in [-1,1]$, as our set of modal basis functions defined over the reference element $\hat{\Omega}$, i.e., $\psi_j^k(x(\xi)) = P_n(\xi)$.

Returning to equation 4, we require the residual to be orthogonal to every basis function $\psi_j^k(x)$ belonging to the vector space $V_h^k$,

$$\int_{\Omega^k} \mathcal{R}_h(x,t)\psi_i^k(x)\,dx = 0, \;\; 1 \leq i \leq N+1, \tag{7}$$

After integrating by parts the expression 7, we arrive at the local weak form of the LWR equation

$$\int_{\Omega^k} \left( \frac{\partial u_h^k}{\partial t}\psi_i^k(x) - f_h^k(u_h^k)\frac{d\psi_i^k(x)}{dx} \right) dx = -\left[\psi_i^k(x) f^*\right]_{x_l^k}^{x_r^k}, \; 1 \leq i \leq N+1. \tag{8}$$

At this point it is necessary to define the numerical flux $f^*$. We chose the Lax-Friedrichs flux because of its simplicity and effectiveness.

After some tedious algebra manipulation, we can write the matrix form of our local semi-discrete weak statement [7]

$$\begin{aligned} \frac{h^k}{2}\mathcal{M}\frac{d}{dt}\hat{\boldsymbol{u}}_h^k - \mathcal{S}^T\hat{\boldsymbol{f}}_h^k = \; & \left[ -\frac{1}{2}\mathcal{F}_l^{k+1}\left(\hat{\boldsymbol{f}}_h^{k+1} - C\hat{\boldsymbol{u}}_h^{k+1}\right) - \frac{1}{2}\mathcal{F}_r^k\left(\hat{\boldsymbol{f}}_h^k + C\hat{\boldsymbol{u}}_h^k\right) \right. \\ & \left. + \frac{1}{2}\mathcal{F}_r^{k-1}\left(\hat{\boldsymbol{f}}_h^{k-1} + C\hat{\boldsymbol{u}}_h^{k-1}\right) + \frac{1}{2}\mathcal{F}_l^k\left(\hat{\boldsymbol{f}}_h^k - C\hat{\boldsymbol{u}}_h^k\right) \right], \end{aligned} \tag{9}$$

where $\mathcal{M}$ and $\mathcal{S}$ are the local mass and stiffness matrices, respectively. The term $h^k/2$ is the jacobian associated to the affine transformation between the reference element $\hat{\Omega} = [-1,1]$ and each grid element $k$-th. In the above formula we also identify the *Lift* matrices, $\mathcal{F}$, associated to the boundaries of each element. These matrices are defined as the dyadic products of elemental basis functions evaluated at the boundaries of the standard element $\hat{\Omega}$ [7]. The vector variables $\hat{\boldsymbol{u}}_h^k$ and $\hat{\boldsymbol{f}}_h^k$ account for the degrees of freedom of the model and their meaning were already described by expressions 5 and 6. It is worth mentioning that we locally recover $\hat{f}_i^k$ coefficients, in terms of the approximate solution $u_h^k$, through a $L^2$ projection as detailed in [7]. Finally, the variable $C$ is defined as the upper bound of the characteristic wave speed of the approximate numerical solution.

When dealing with artificial viscosity based techniques for shock capturing, we must write a viscous version of the original LWR equation. Thus, we add a second order diffusive term at the right hand side of equation 1 and define an auxiliary variable $q$ to keep the DG approach of the numerical approximation which contains only first order derivatives. The resulting system of first order equations is now given as,

$$\begin{aligned} & \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(f(u) - \sqrt{\varepsilon}q) = 0 \\ & q = \sqrt{\varepsilon}\frac{\partial u}{\partial x}\,, \end{aligned} \tag{10}$$

where $\varepsilon(x)$ is the artificial viscosity function.

As before, we assume that the auxiliary variable $q$ which approximates the derivative of the conserved variable $u(x,t)$ can be written as a polynomial expansion of order $N$ as follows,

$$q_h^k(x,t) = \sum_{j=1}^{N+1} \hat{q}_j^k(t)\psi_j^k(x). \tag{11}$$

Following the same discretization steps applied to the original equation we arrive at the local weak statement of the modified traffic flow problem,

$$\int_{\Omega^k} \frac{\partial u_h^k}{dt}\psi_i^k(x)\,dx - \int_{\Omega^k} f_h^k(u_h^k)\frac{d\psi_i^k(x)}{dx}\,dx + \int_{\Omega^k} \sqrt{\varepsilon}\, q_h^k(u_h^k)\frac{d\psi_i^k(x)}{dx}\,dx = -\left[\psi_i^k(x)f^*\right]_{x_l^k}^{x_r^k}$$
$$+ \left[\psi_i^k(x)(\sqrt{\varepsilon}q_h^k)^*\right]_{x_l^k}^{x_r^k},\ 1 \le i \le N+1, \tag{12}$$

$$\int_{\Omega^k} q_h^k(u_h^k)\psi_i^k(x)\,dx + \int_{\Omega^k} u_h^k\frac{d}{dx}(\sqrt{\varepsilon}\psi_i^k(x))\,dx = \left[\psi_i^k(x)(\sqrt{\varepsilon}u_h^k)^*\right]_{x_l^k}^{x_r^k},\ 1 \le i \le N+1. \tag{13}$$

The numerical fluxes $(\sqrt{\varepsilon}q)^*$ and $(\sqrt{\varepsilon}u)^*$ above, corresponding to the second order dissipative operator, were approximated by the Local Discontinuous Galerkin (LDG) flux which shows optimal convergence rates [10].

After choosing Lax-Friedrichs and LDG as our numerical fluxes and upon substitution of the polynomial expansions for $u_h^k$, $f_h^k$, $q_h^k$ in equations 12 and 13, we can write the final matrix form of the semi-discrete DGFEM approximation of the viscous version of LWR traffic flow equation,

$$\frac{h^k}{2}\mathcal{M}\frac{d}{dt}\hat{\boldsymbol{u}}_h^k - \mathcal{S}^T\hat{\boldsymbol{f}}_h^k + (\mathcal{S}^{\sqrt{\varepsilon}})^T\hat{\boldsymbol{q}}_h^k = \Big[-\frac{1}{2}\mathcal{F}_l^{k+1}\left(\hat{\boldsymbol{f}}_h^{k+1} - C\hat{\boldsymbol{u}}_h^{k+1}\right) - \frac{1}{2}\mathcal{F}_r^k\left(\hat{\boldsymbol{f}}_h^k + C\hat{\boldsymbol{u}}_h^k\right)$$
$$+ \frac{1}{2}\mathcal{F}_r^{k-1}\left(\hat{\boldsymbol{f}}_h^{k-1} + C\hat{\boldsymbol{u}}_h^{k-1}\right) + \frac{1}{2}\mathcal{F}_l^k\left(\hat{\boldsymbol{f}}_h^k - C\hat{\boldsymbol{u}}_h^k\right)$$
$$+ \sqrt{\varepsilon(x_r^k)}\,\mathcal{F}_r^k\hat{\boldsymbol{q}}_h^k - \sqrt{\varepsilon(x_r^{k-1})}\,\mathcal{F}_r^{k-1}\hat{\boldsymbol{q}}_h^{k-1}\Big], \tag{14}$$

$$\frac{h^k}{2}\mathcal{M}\hat{\boldsymbol{q}}_h^k + (\tilde{\mathcal{S}}^{\sqrt{\varepsilon}})^T\hat{\boldsymbol{u}}_h^k = +\frac{1}{2}\left(\sqrt{\varepsilon(x_l^{k+1})} + \sqrt{\varepsilon(x_r^k)}\right)\mathcal{F}_l^{k+1}\hat{\boldsymbol{u}}_h^{k+1}$$
$$-\frac{1}{2}\left(\sqrt{\varepsilon(x_r^{k-1})} + \sqrt{\varepsilon(x_l^k)}\right)\mathcal{F}_l^k\hat{\boldsymbol{u}}_h^k, \tag{15}$$

where $\hat{\boldsymbol{q}}_h^k$ is the vector of degrees of freedom associated to the auxiliary variable $q$ and $\mathcal{S}^{\sqrt{\varepsilon}}$ and $\tilde{\mathcal{S}}^{\sqrt{\varepsilon}}$ are modified stiffness matrices which take into account the artificial viscosity function $\varepsilon(x)$ [7]. It is worth noting that as $\hat{\boldsymbol{f}}_h^k$, the vector $\hat{\boldsymbol{q}}_h^k$ is also locally computed.

## 4 SHOCK CAPTURING SCHEMES

As a typical nonlinear PDE based on a conservation law, the LWR equation naturally develops shocks along time integration. When high order discretization techniques such as DGFEM are used to approximate solutions for this equation, stabilizing schemes are necessary to control undesired oscillations in the vicinity of discontinuities due to the Gibbs phenomenon.

We chose two standard approaches to stabilize the numerical solutions computed by the high order DGFEM solver: limiters and sub-cell shock capturing based on artificial viscosity adding. In the first case we implemented a generalized slope limiter ($\Pi^N$) as described in [8] and

a hierarchical slope limiter proposed in [9]. In the second approach we implemented the sub-cell shock capturing scheme as originally proposed in [11] which uses elementwise constant artificial viscosity. We also implemented a modified version of this strategy which considers the artificial viscosity as a $C^0$ function [12]. The details of the mathematical formulation of the sub-cell shock capturing scheme applied to the LWR traffic flow equation can be found in [7].

## 5 NUMERICAL EXPERIMENTS

The LWR equation was approximated considering three different initial conditions: (1) Red light at the exit of the road; (2) Green light at half way of the road length and (3) Exponential density configuration at the entrance of the road. All examples were set with appropriate Dirichlet or Neumann boundary conditions. For each of these initial conditions the numerical solutions were marched in time for $2.0$ minutes.

The five step $4^{th}$ order strong stability preserving Runge-Kutta explicit integrator suggested in [8] was used to evolve the semi-discrete formulation (equations 9 and 14) in time.

We organized our experiments in four subsections: the first subsection presents the overall aspect of the numerical solution of the LWR equation for the three initial conditions considered. For each initial condition, we depict the behavior of the four shock capturing schemes we have chosen. In the second subsection, we compare the performance of both limiters, namely, hierarchical slope limiter (HL) and generalized slope limiter (GSL) in isolation when increasing the number of grid elements ($K$) and polynomial order ($N$). Likewise, in the third subsection, we compare the performance of both stabilizing schemes based on artificial viscosity, namely, sub-cell shock capturing with elementwise constant artificial viscosity (Constant-SCSC) and sub-cell shock capturing with local $C^0$ artificial viscosity ($C^0$-SCSC) in isolation in terms of the number of grid elements and polynomial order. Finally, we compare the two schemes showing the better performance (limiter vs. artificial viscosity) in terms of the number of degrees of freedom (NDOF) to achieve an acceptable resolution of the jump discontinuity.

### 5.1 Traffic initial condition examples

#### 5.1.1 Red light

For the red light example a traffic light indicating stop is positioned at the end of a road of $4\ Km$ long. Cars accumulate quickly near the road exit, where we have the maximum allowed density of cars between 3 and $4\ Km$. There is an incoming traffic of $50\%$ the maximum allowed density ($u = 0.5u_{max}$). This is a standard benchmark problem for the LWR equation because it has an exact solution given by a moving shock wave traveling backward from the exit to the entrance of the road. The remaining parameters of the problem were defined as follows: car speed limit $V_{max} = 1.0\ Km/min$, and maximum car density per unit length $u_{max} = 10.0\ cars/Km$. Non homogeneous Dirichlet boundary conditions were set at both limits of the road with the following values: $u(x = 0) = 5.0\ cars/Km$, at left, and, $u(x = 4) = 10.0\ cars/Km$, at right.

The red light problem solutions were stabilized with the four shock capturing schemes described in section 4. For limiters, we set the number of grid elements and the polynomial order of approximation as $K = 500$ and $N = 2$. In the case of artificial viscosity based schemes, we set $K = 20$ and $N = 10$. Figure 2 depicts the 3D graphics of the numerical solution of this problem for the four shock capturing schemes.

We can clearly notice that the Constant-SCSC scheme produced many spikes over the numerical solution in different time instants. The other schemes showed very similar solution

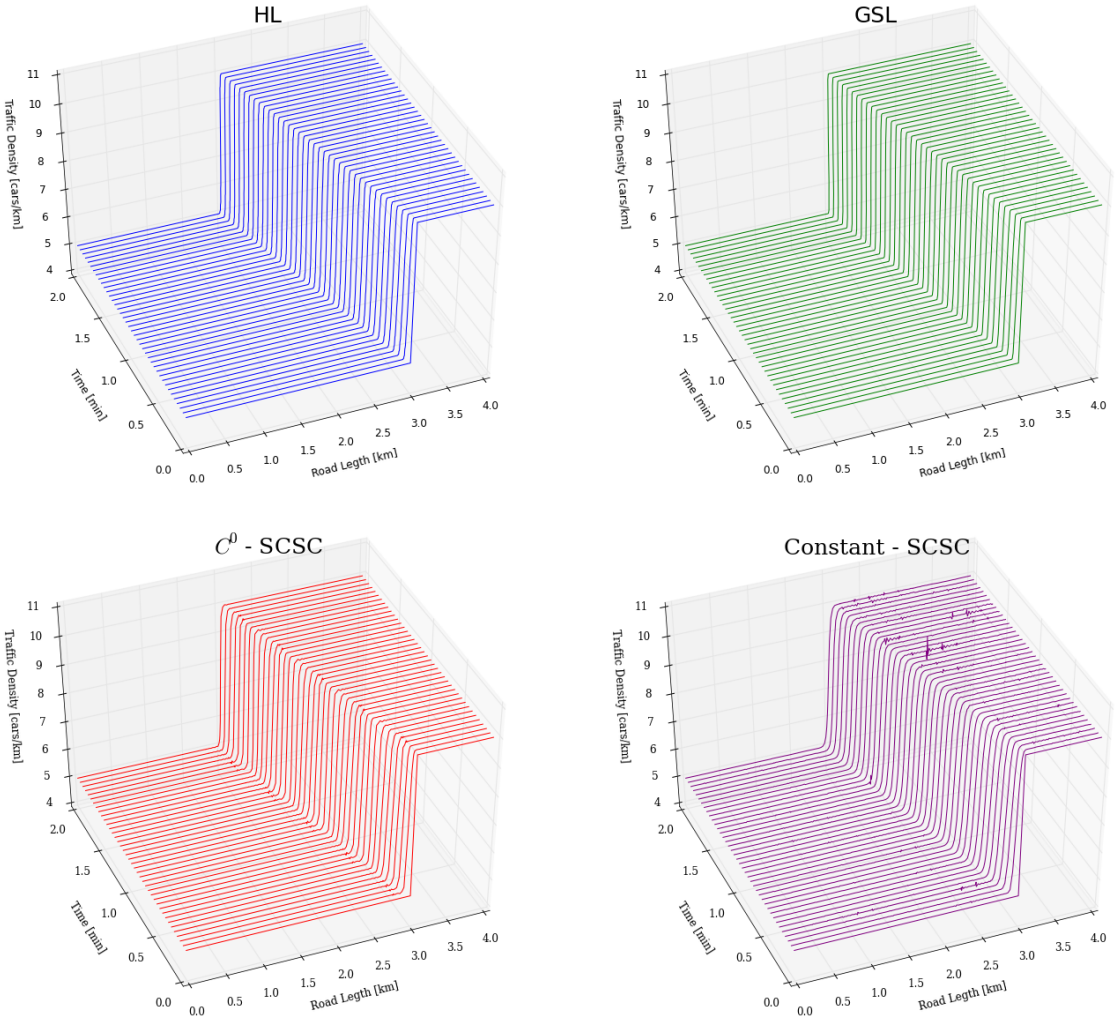



Figure 2: Red Light problem solutions using the four shock capturing schemes.

profiles.

### 5.1.2 Green light

For the green light example we suppose there exists a traffic light at the half way of a road of $4\ Km$ long where cars are bumper-to-bumper. The traffic density decreases linearly to zero towards the entrance of the road and ahead of the traffic light the road is clear. When the traffic light turns green cars start to move and the car density profile changes its original triangular shape. The car density profile starts with a jump discontinuity at the wave front and develops a shock at the rear of the propagating wave. For this example, the following setup parameters were used: car speed limit $V_{max} = 1.0\ Km/min$, and maximum car density per unit length $u_{max} = 10.0\ cars/Km$. Homogeneous Dirichlet boundary conditions were set at both limits of the road.

As in the previous example, numerical solutions were stabilized with the four shock capturing schemes and the following configurations were set for limiters: $K = 500$ and $N = 2$, and for artificial viscosity based schemes: $K = 20$ and $N = 10$. Figure 3 depicts the 3D graphics of the numerical solution of the green light problem for the four shock capturing schemes.

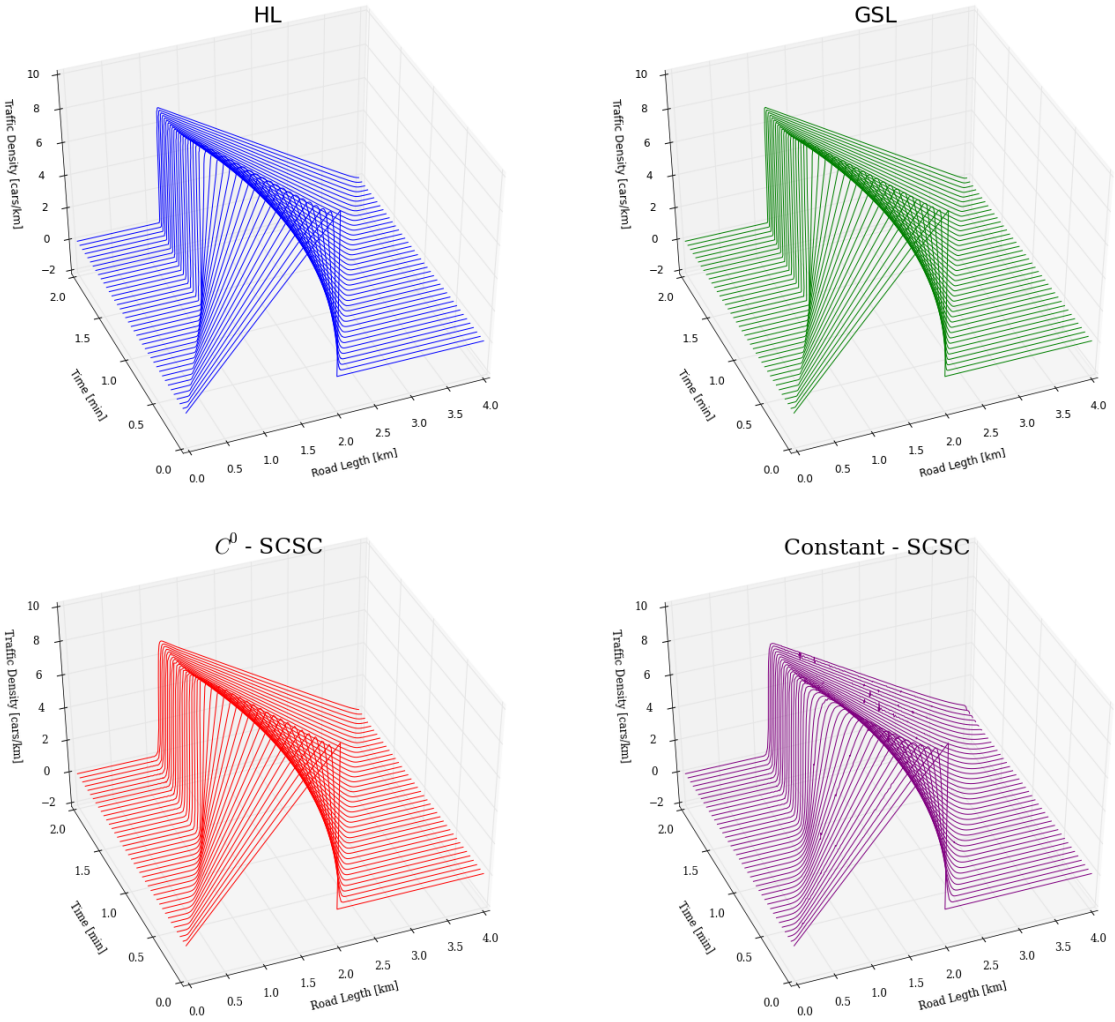



Figure 3: Green light problem solutions using the four shock capturing schemes.

The same observations made for the numerical solutions of the red light example are valid for the green light problem.

### 5.1.3 Exponential

The exponential problem is a rather theoretical case study for the LWR equation which shows a shock formation from a smooth initial profile. In this case, there is no traffic light along a road of $2\pi$ $Km$ long. The remaining setup parameters were chosen as follows: speed limit $V_{max} = 1.0\ Km/min$, and maximum traffic density $u_{max} = 1.0\ car/Km$. At left, it was set a non homogeneous time dependent Dirichlet boundary condition given as $u(t) = 1.0 + \sin(t)$ and, at right, a homogeneous Neumann boundary condition was prescribed.

Again, we can see in the figure 4 the 3D graphics representing the numerical solution of the exponential problem for the four shock capturing schemes. In this case, again, the following setups were fixed for limiters: $K = 500$ and $N = 2$, and for artificial viscosity based schemes: $K = 20$ and $N = 10$.

For this example, all the numerical solutions revealed very similar shapes and no spikes could be noticed over the solution stabilized by the Constant-SCSC scheme, at least visually.

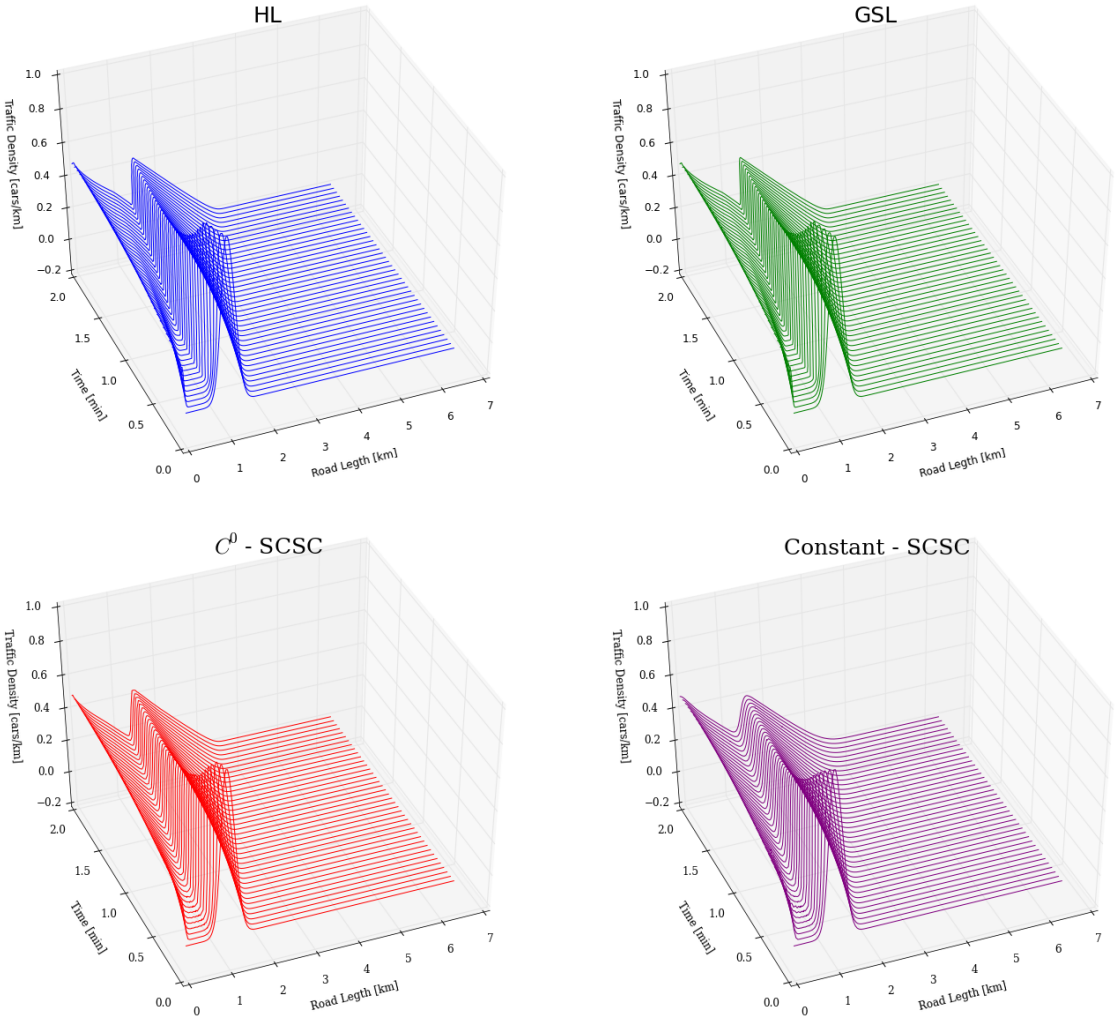



Figure 4: Exponential problem solutions using the four shock capturing schemes.

## 5.2 Limiters

This subsection aims at checking the ability of each slope limiter, HL and GSL, to improve shock resolution when increasing the number of grid elements, $K$, and the polynomial order, $N$. For each traffic problem studied, these quantities were fixed as $K = 100, 170, 240, 300$, and $N = 2, 6, 10$. Snapshots of the shock formation were taken at $t = 1.2, 1.0, 0.7\ min$ for the red light, green light and exponential examples, respectively.

### 5.2.1 Hierarchical slope limiter

Figures 5, 6 and 7 depicts the performance of the hierarchical slope limiter as a function of the total number of grid elements and the polynomial order of approximation.

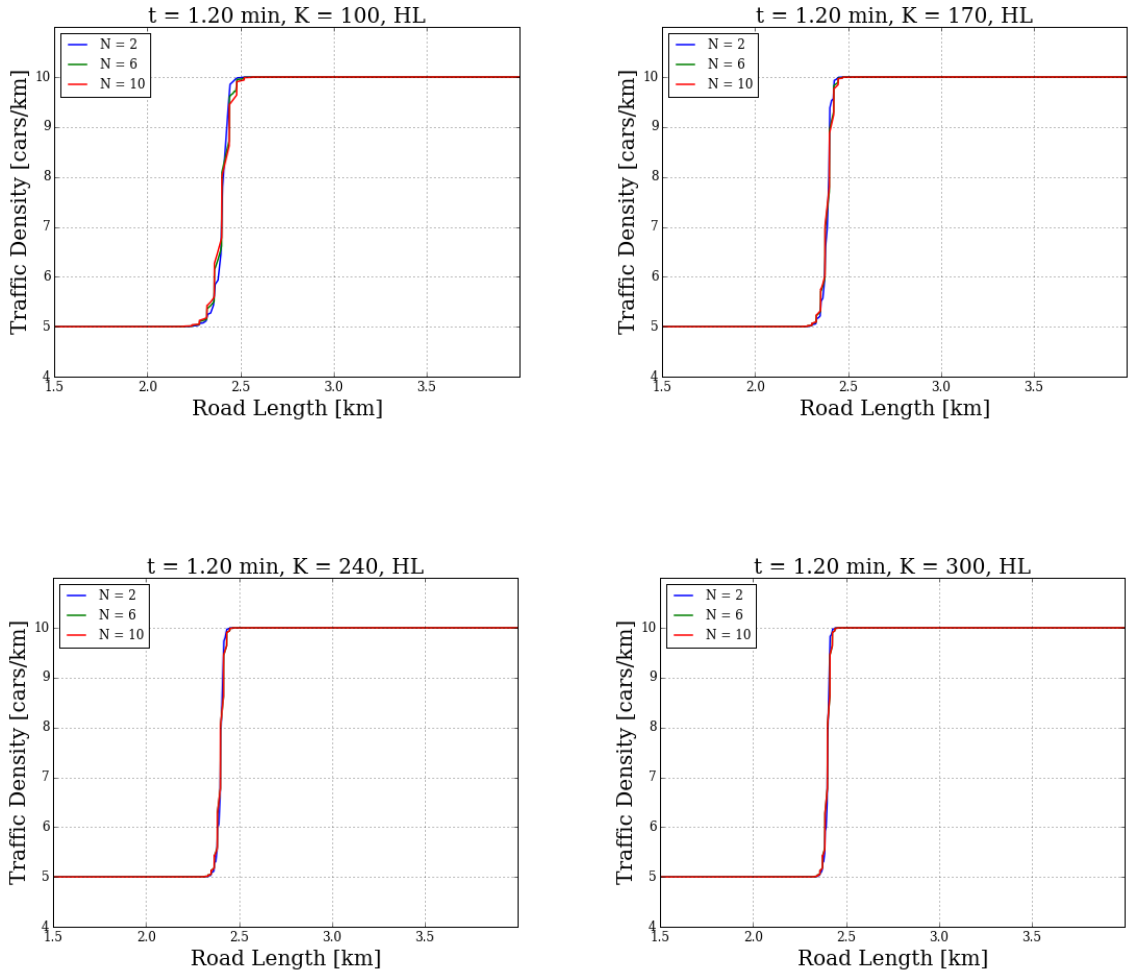
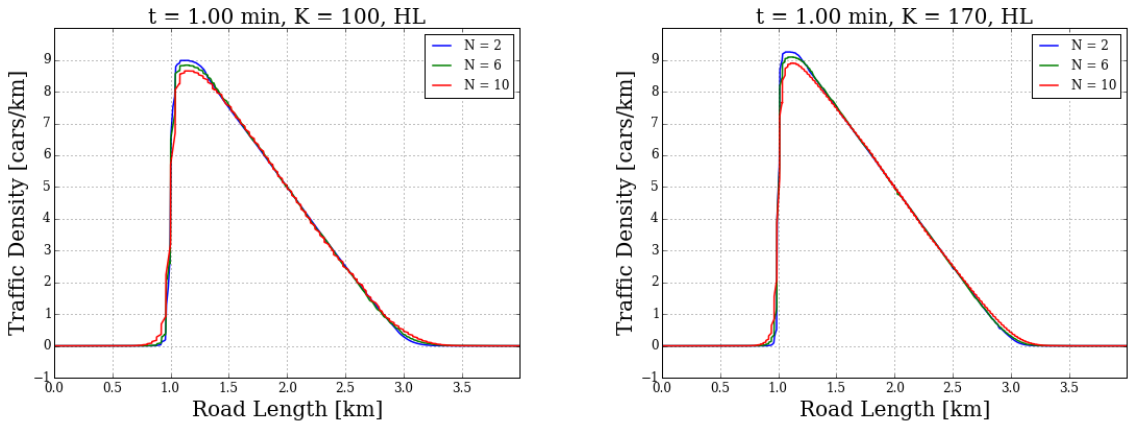



Figure 5: Red Light initial condition.



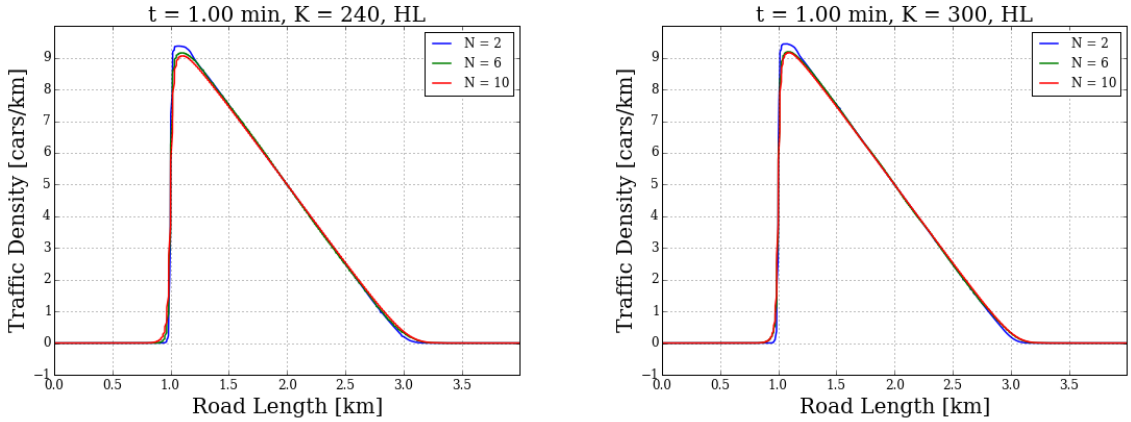



Figure 6: Green Light initial condition.

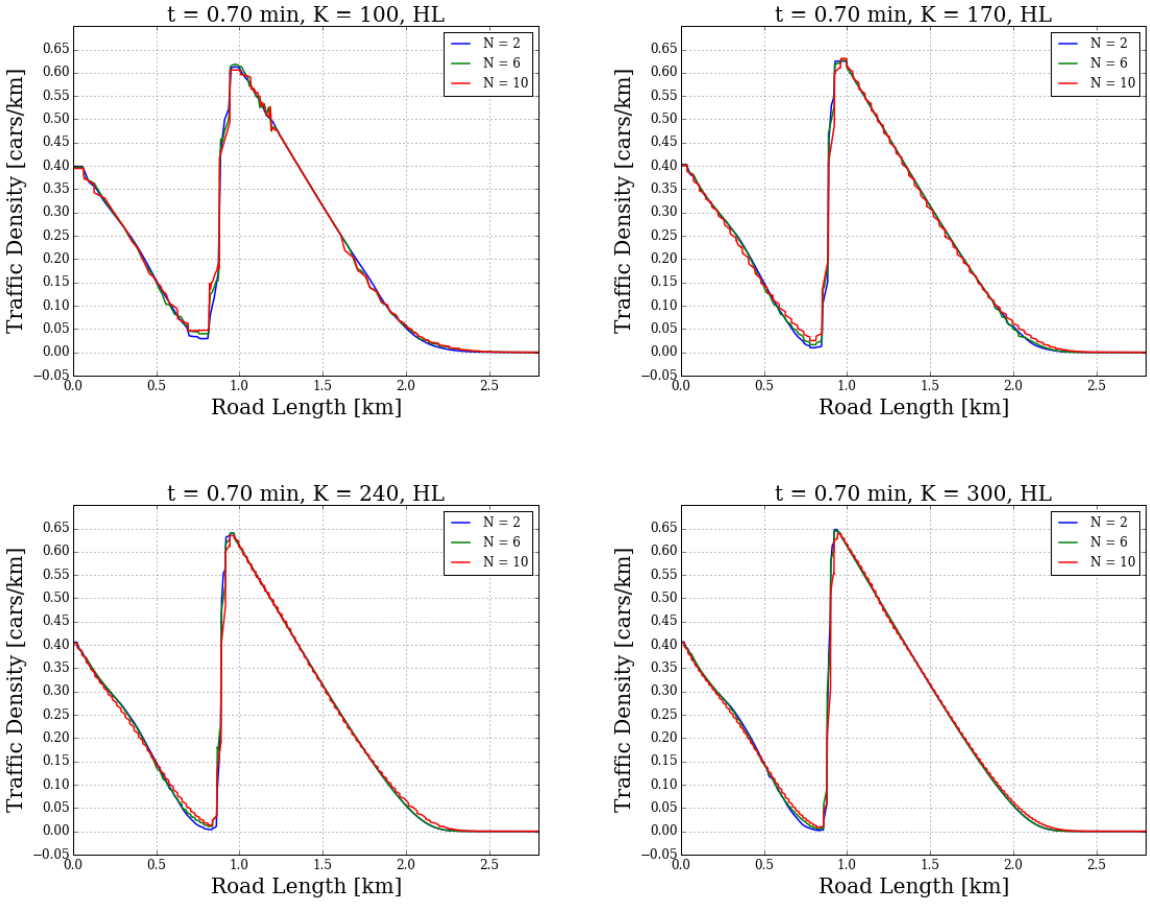



Figure 7: Exponential initial condition.

We can observe from these graphics that the increase of grid elements rather than the polynomial order is the best way to improve the shock resolution. Moreover, we can note that the greater the polynomial order the worse the overall quality of the solutions for a fixed number of elements.

### 5.2.2 Generalized slope limiter

Figures 8, 9 and 10 depicts the performance of the generalized slope limiter ($\Pi^N$) as a function of the total number of grid elements and the polynomial order of approximation.

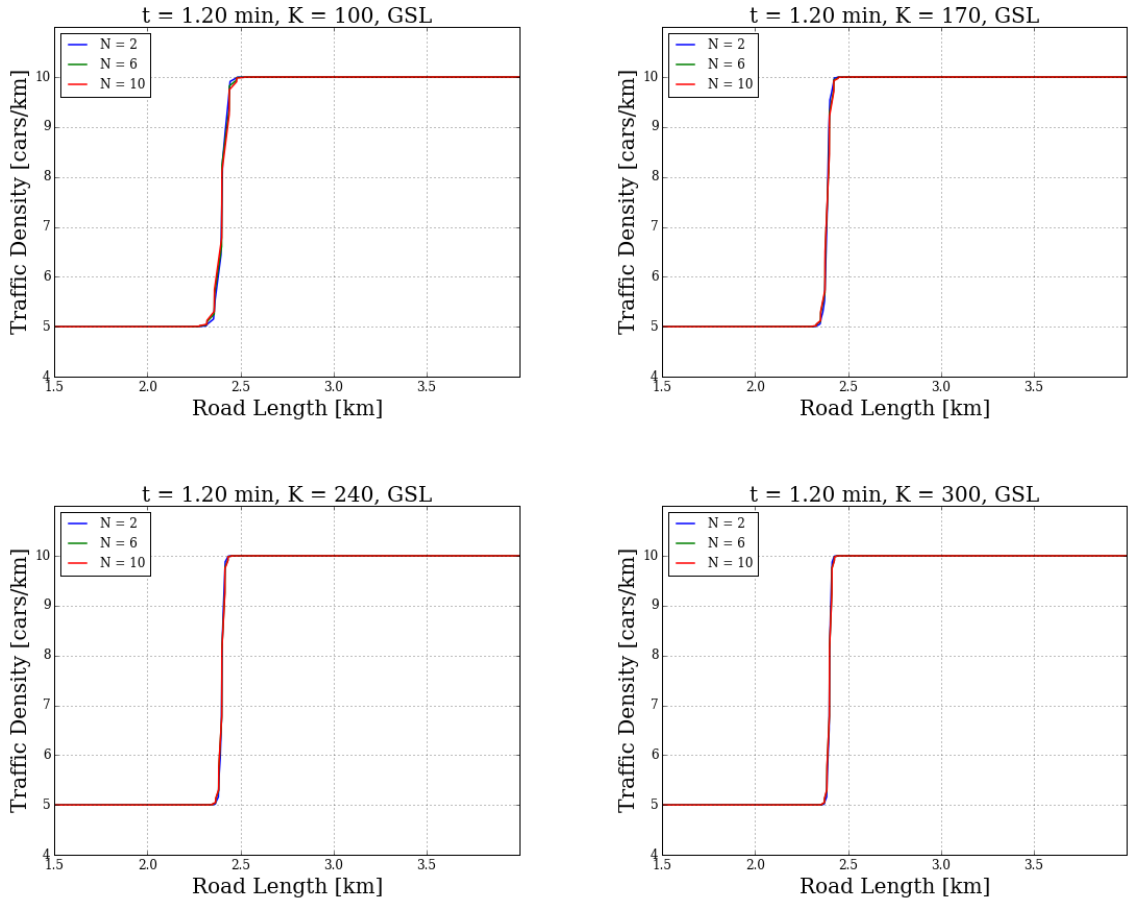
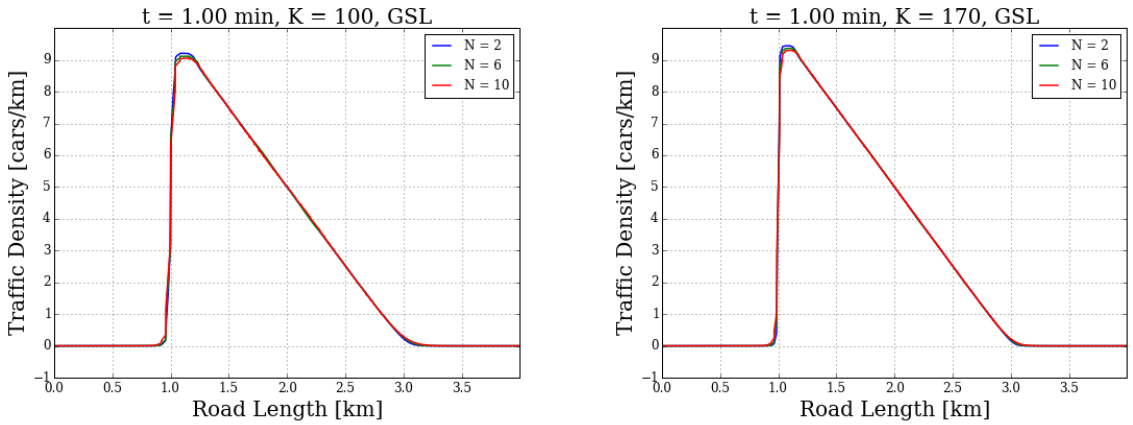



Figure 8: Red Light initial condition



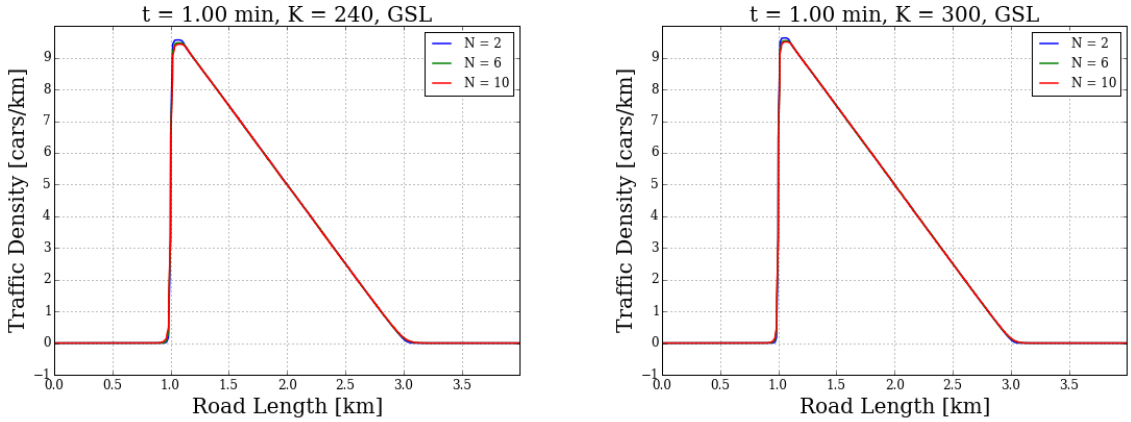



Figure 9: Green Light initial condition

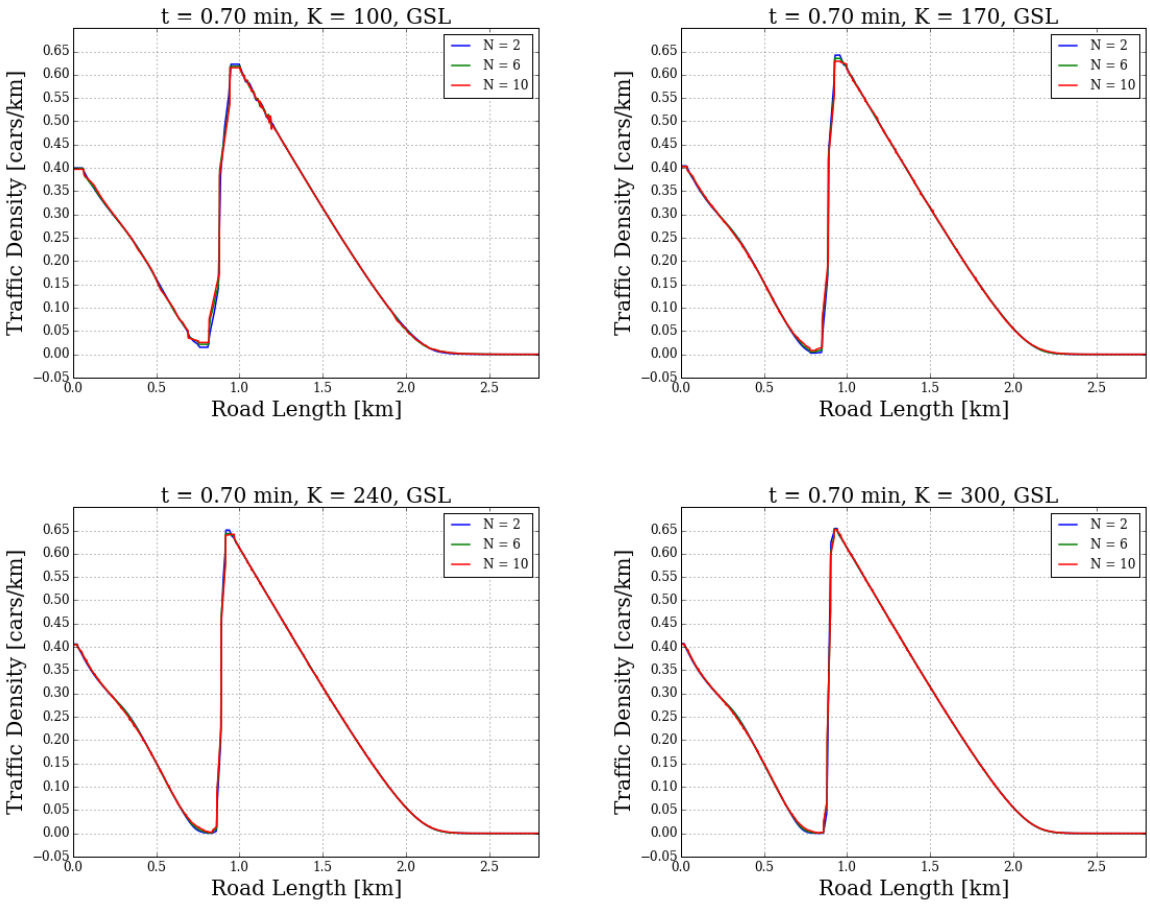



Figure 10: Exponential initial condition.

We can notice that the GSL showed better resolution capabilities when using a large number of elements as in the case of the HL despite being less sensitive to polynomial order increment. Our goal with this first set of experiments was to prove that limiters are not greatly affected by the increment of polynomial order to display sharp shock resolution requiring quite refined grids to achieve reasonable sharpness.

### 5.3 Sub-cell shock capturing schemes

As in the previous study using limiters, this subsection intends to verify the ability of the artificial viscosity based schemes, Constant-SCSC and $C^0$-SCSC, to improve shock resolution as a function of the number of grid elements, $K$, and the polynomial order, $N$. We fixed $K = 40$ for the Constant-SCSC scheme and $K = 20$ for the $C^0$-SCSC scheme as the former scheme required more degrees of freedom to show minimally acceptable results. For both schemes, the polynomial order were set as $N = 2, 6, 10$. Again, snapshots of the shock formation were taken at $t = 1.2, 1.0, 0.7\ min$ for the red light, green light and exponential examples, respectively. It is worth mention that sub-cell shock capturing strategies required much smaller time steps than limiters to ensure stable solutions. Overall, limiters used four times larger time steps than sub-cell shock capturing schemes.

#### 5.3.1 Constant-SCSC

Figures 11, 12 and 13 depicts the overall performance of the constant-SCSC scheme as a function of the total number of grid elements and the polynomial order of approximation.

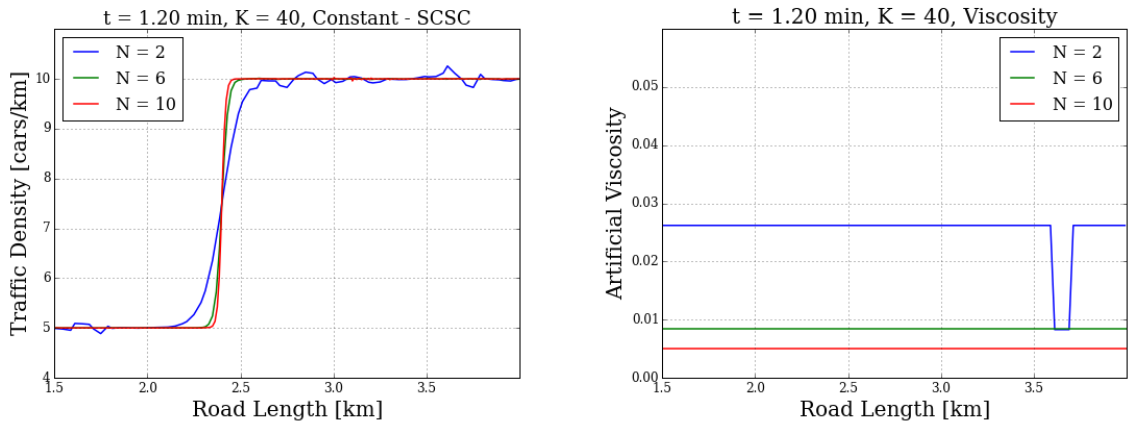



Figure 11: Red Light initial condition

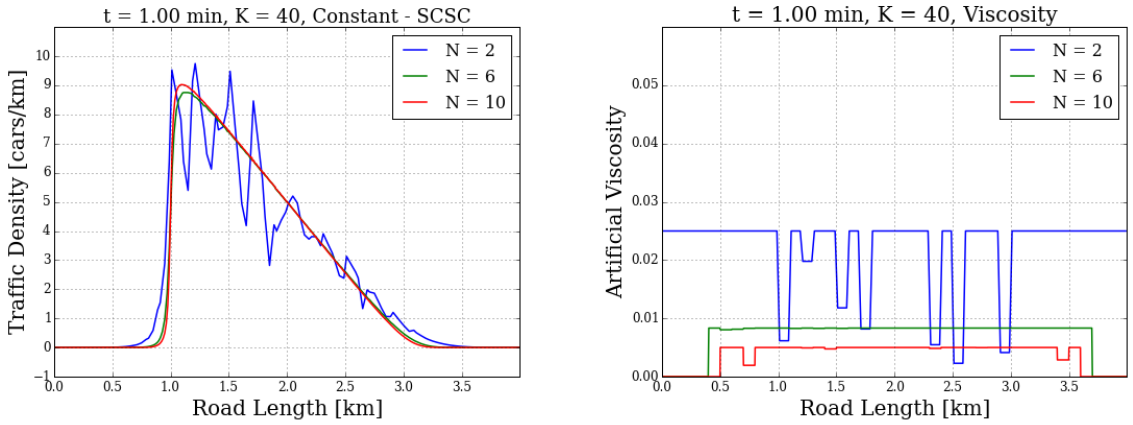



Figure 12: Green Light initial condition

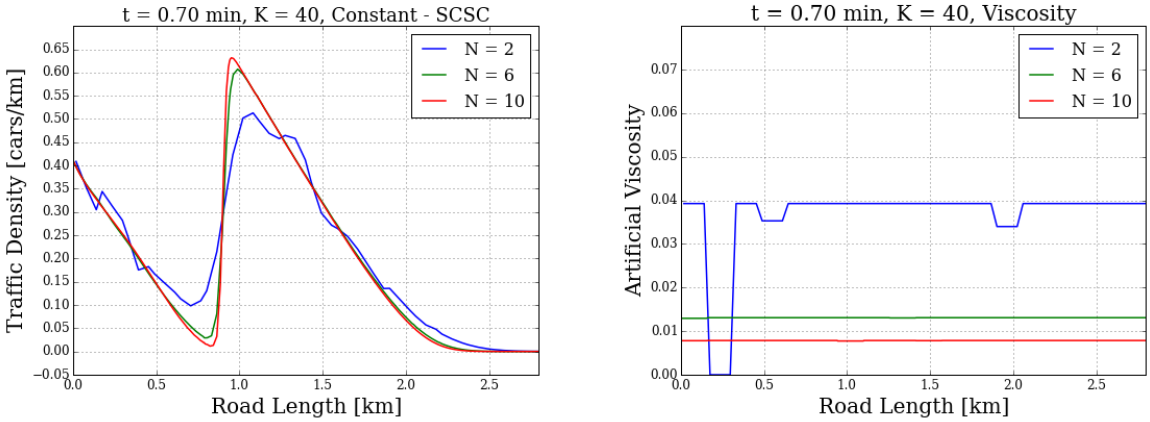



Figure 13: Exponential initial condition

From the above mentioned figures, it is evident that the sub-cell shock capturing scheme based on a globally discontinuous viscosity function $\varepsilon(x)$ produced very poor quality results for low polynomial orders. Even with high order polynomial approximations such scheme developed small traveling spikes over the solution profiles destroying the locality of the method and resulting in excessive smearing of the jump discontinuities.

### 5.3.2 $C^0$−SCSC

Figures 14, 15 and 16 depicts the overall performance of the $C^0$-SCSC scheme as a function of the total number of grid elements and the polynomial order of approximation.

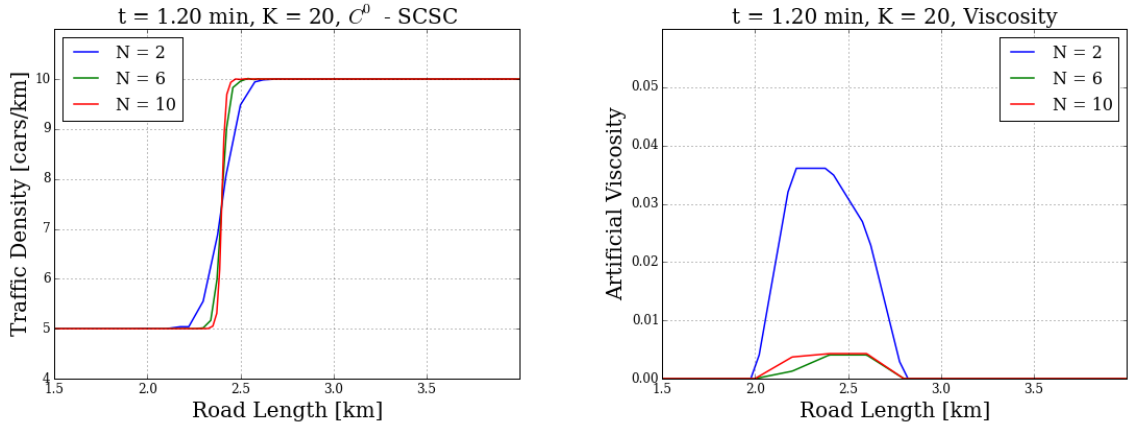



Figure 14: Red Light initial condition

Contrary to what has been revealed by the Constant-SCSC approach, the $C^0$-SCSC scheme demonstrated very reliable even for very coarse grids. Sharpness in shock resolution was clearly improved with polynomial order increasing and the locality of the added artificial viscosity was definitely proved. This second set of experiments aimed at showing that artificial viscosity mechanisms perform much better with very high order approximations and larger elements, and high resolution of the shock representation is achieved with the increment of the polynomial order.

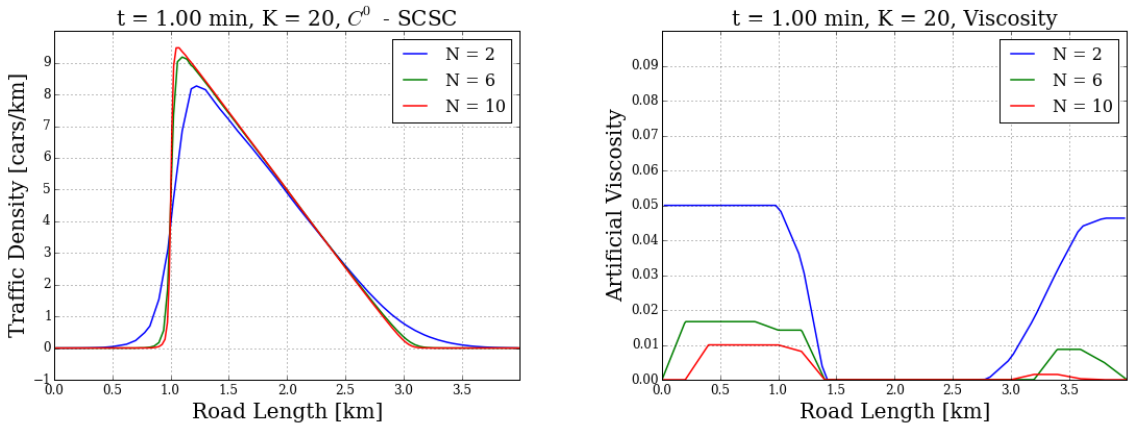



Figure 15: Green Light initial condition

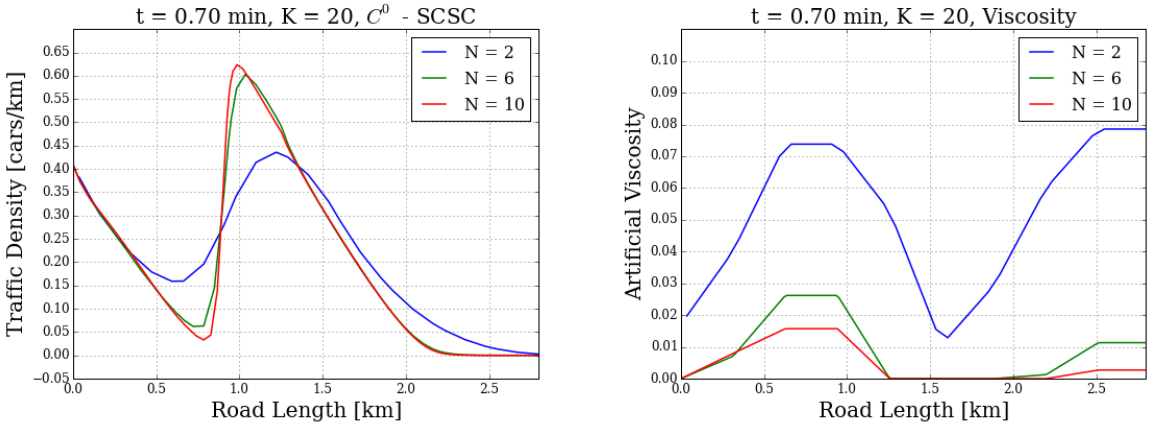


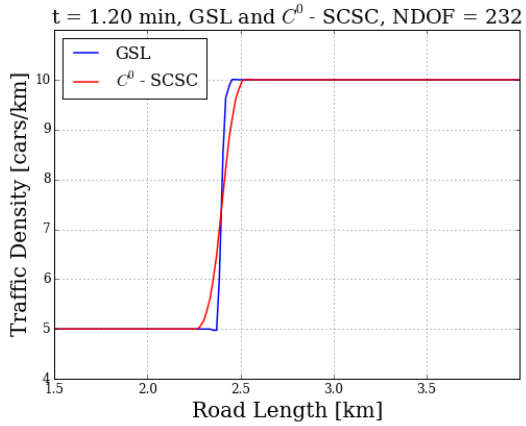

Figure 16: Exponential initial condition

## 5.4 Limiter vs. Sub-cell shock capturing

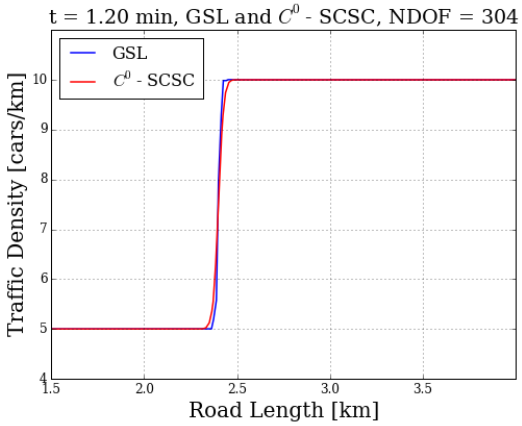
This subsection was designed to evaluate which shock capturing scheme shows the best cost-benefit to sharply resolve the shock formation using an unbiased criterion to measure computational performance. To achieve such goal, we compare GSL to $C^0$-SCSC in terms of the total degrees of freedom required by each scheme to reach a reasonable resolution of jump discontinuities.

Figures 17, 18 and 19 compare the overall performance of the GSL and the $C^0$-SCSC schemes as a function of the total number of degrees of freedom. Snapshots of the shock formation were taken at $t = 1.2, 1.0, 0.7\ min$ for the red light, green light and exponential examples, respectively. The number of elements and the polynomial orders are indicated in the caption of each figure for both schemes. For GSL we kept $N = 1$ and increased the number of elements $K$ as we have previously demonstrated the weak impact of polynomial order in the shock resolution. In the case of $C^0$-SCSC, we increased $N$ as much as we reduced $K$ once the most remarkable feature of this scheme is to achieve sharp shock resolution with very coarse grids and high order polynomial approximations.

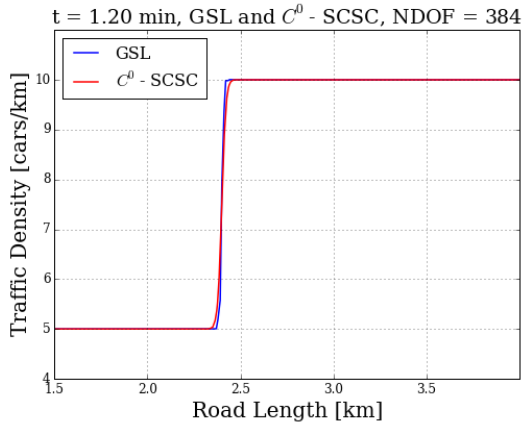
Results from both techniques showed very similar for the red light example. With just a few degrees of freedom to approximate the solution, $C^0$-SCSC was very dissipative but GSL



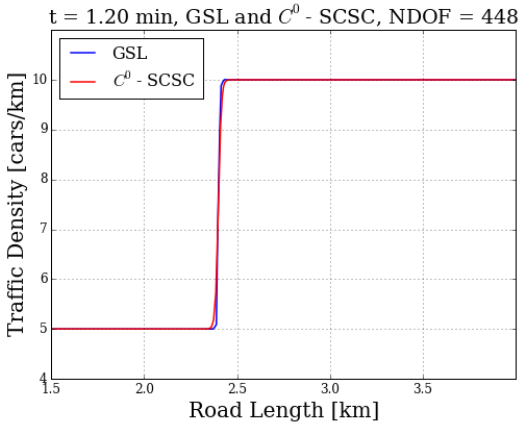

(a) GSL: K = 116, N = 1. $C^0$- SCSC: K = 116, N = 1




(b) GSL: K = 152, N = 1. $C^0$- SCSC: K = 76, N = 3



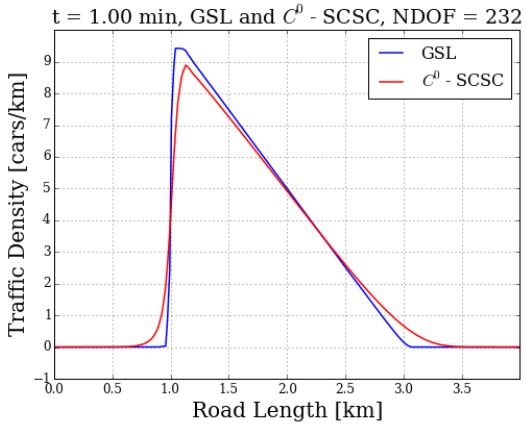

(c) GSL: K = 192, N = 1. $C^0$- SCSC: K = 64, N = 5



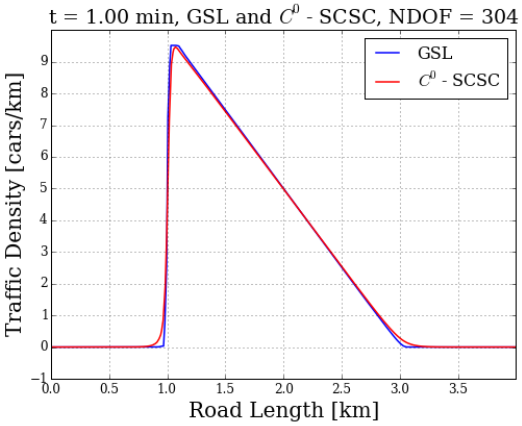

(d) GSL: K = 224, N = 1. $C^0$- SCSC: K = 56, N = 7

Figure 17: Red light initial condition

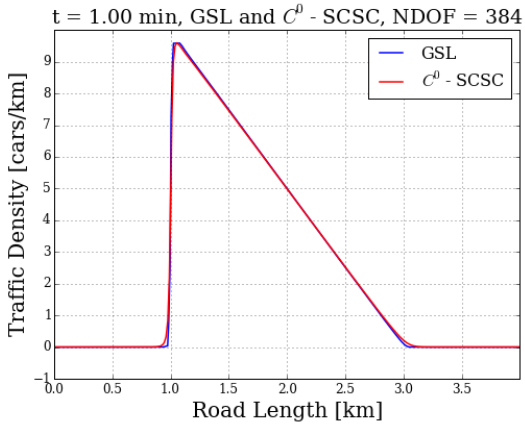
showed a small oscillation at the base of the discontinuity. For the maximum number of degrees of freedom, GLS seemed to slightly trim the sharp corners of the shock.



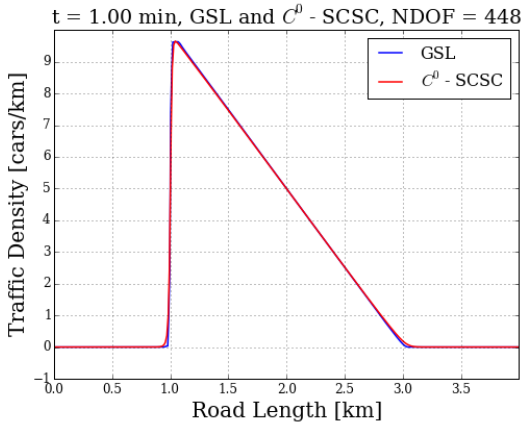

(a) GSL: K = 116, N = 1. $C^0$- SCSC: K = 116, N = 1




(b) GSL: K = 152, N = 1. $C^0$- SCSC: K = 76, N = 3



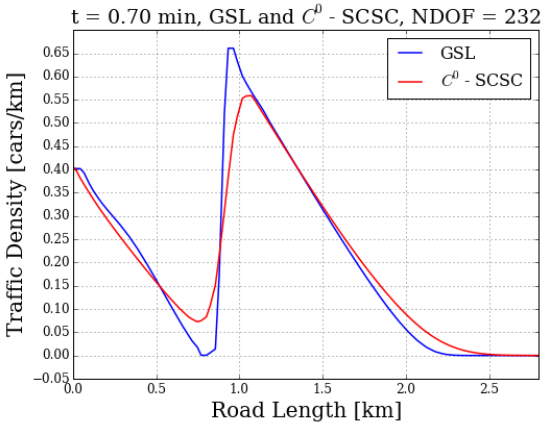

(c) GSL: K = 192, N = 1. $C^0$- SCSC: K = 64, N = 5



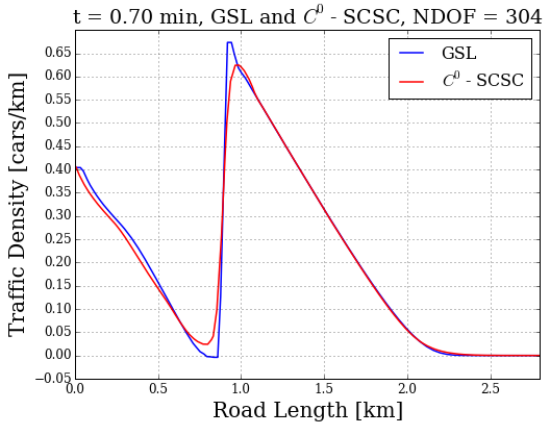

(d) GSL: K = 224, N = 1. $C^0$- SCSC: K = 56, N = 7

Figure 18: Green light initial condition

For the green light example, GSL trimmed the top corner of the shock in every case. As expected, $C^0$-SCSC solutions were slightly more dissipative but still able to sharply characterize the shock resolution.

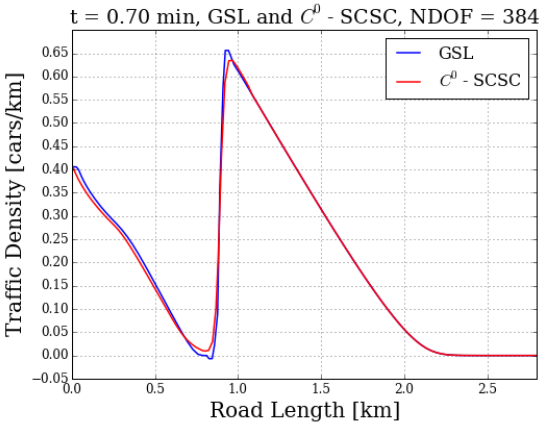



(a) GSL: K = 116, N = 1. $C^0$- SCSC: K = 116, N = 1

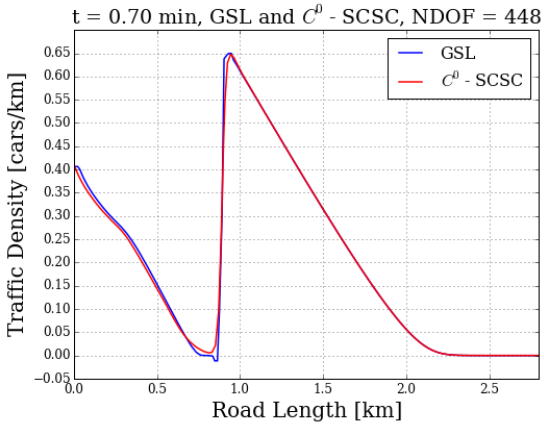



(b) GSL: K = 152, N = 1. $C^0$- SCSC: K = 76, N = 3

For the exponential example, $C^0$-SCSC scheme proved superior as GSL was not able to control instabilities at the base of the jump discontinuity and also trimmed the top corner of the shock in every situation. As already observed, the $C^0$-SCSC scheme worked quite well for very high order approximations.




(c) GSL: K = 192, N = 1. $C^0$- SCSC: K = 64, N = 5




(d) GSL: K = 224, N = 1. $C^0$- SCSC: K = 56, N = 7

Figure 19: Exponential initial condition.

## 6 CONCLUSIONS

This paper discussed the implementation and verification of four different strategies to stabilize the numerical solution of the LWR equation discretized with a modal version of the high order discontinuous Galerkin method. We selected two efficient limiters (HL and GSL) [8, 9] as well as two artificial viscosity based schemes (Constant-SCSC and $C^0$-SCSC) to stabilize our numerical solutions. We proposed three examples of initial conditions to test our implementations: (a) red light at road exit; (b) green light at half way of road length and (c) exponential car density at road entrance.

We were able to prove that limiters are not greatly affected by the increment of polynomial order to display sharp shock resolution requiring quite refined grids to achieve reasonable results. The HL revealed more sensitive to polynomial order refinement as it produced poorer results when compared to the GSL for high order approximations.

Both sub-cell shock capturing schemes demonstrated good robustness to sharply represent shock structures with very coarse grids. High resolution was achieved for increasing polynomial order. However, it is worth remarking that sub-cell shock capturing with elementwise constant artificial viscosity produced traveling spikes over the solution profiles during time integration causing the shock sensor activation most of the spatial domain and therefore destroying the acclaimed locality of the method. It also resulted in excessive smearing of the shock structure.

Sub-cell shock capturing with $C^0$ artificial viscosity [12] proved very reliable to stabilize solutions in the presence of jump discontinuities although it required smaller time steps for explicit time integration. Such stabilizing scheme displayed the most noticeable results exhibiting sharp shock resolution features in very coarse meshes with very little dissipation. Overall, such scheme revealed the best cost-benefit in terms of computational efficiency based on the number of degrees of freedom for a certain level of shock resolution.

## ACKNOWLEDGEMENTS

The authors would like to thank the IBM Research Brazil for the technical and financial support which enabled the development of this research initiative.

## REFERENCES


[1] M.J. Lighthill, G.B. Whitham, On kinematic waves II. A theory of traffic flow on long crowded roads, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, **229(1178)**, 317–345, 1955.

[2] P.I. Richards, Shock waves on the highway, *Operations Research*, **4(1)**, 42–51, 1956.

[3] W. Ceulemans, M.A. Wahab, K. De Proft, G. Wets, Modelling Traffic Flow with Constant Speed using the Galerkin Finite Element Method. *Proceedings of the World Congress on Engineering*, Vols. **I** and **II**, 993–999, 2009.

[4] C. Suncica, B. Piccoli, J.M. Qiu, T. Ren, Runge-Kutta Discontinuous Galerkin Method for Traffic Flow Model on Networks. *ArXiv e-prints*, 1403.3750, math.NA, 2014.

[5] B.D. Greenshields, A study of highway capacity. *Highway Research Board Proceedings*, **14**, 448–477, 1935.

[6] G.E. Karniadakis, S.J. Sherwin, *Spectral/hp Element Methods for Computational Fluid Dynamics,* $2^{nd}$ *Edition*. Oxford University Press, 2005.

[7] A.C. Nogueira Jr., J. Bernardo, S. Moore. Solution of the LWR Nonlinear Traffic Flow Model using a Discontinuous Galerkin FEM Approach. *Proceedings of the CILAMCE XXXVI - 36th Ibero-Latin American Congress on Computational Methods in Engineering*, Rio de Janeiro, Brazil, Nov. 22-25, 2015.

[8] J.S. Hesthaven, T. Warburton, *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer, 2008.

[9] L. Krivodonova, Limiters for high-order discontinuous Galerkin methods. *Journal of Computational Physics*, **226**, 879–896, 2007.

[10] B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, **35**, 2440–2463, 1998.

[11] P-O. Persson, J. Peraire, Sub-Cell Shock Capturing for Discontinuous Galerkin Methods. *Proceedings of the* $44^{th}$ *AIAA Aerospace Sciences Meeting and Exhibit*, **112**, 2006.

[12] A. Klöckner, T. Warburton, J.S. Hesthaven. Viscous Shock Capturing in a Time-Explicit Discontinuous Galerkin Method. *Mathematical Modelling of Natural Phenomena*, **6(3)**, 57–83, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPARISON OF FINITE VOLUME HIGH-ORDER SCHEMES FOR THE TWO-DIMENSIONAL EULER EQUATIONS

**Jens Wellner**

German Aerospace Center (DLR),
Institute of Propulsion Technology,
Linder Höhe, 51147 Cologne, Germany
e-mail: Jens.Wellner@dlr.de

**Keywords:** High-order, Finite volume, Euler equations, MUSCL, WENO, compact scheme

**Abstract.** *The present paper describes the use of various high-order schemes in a finite volume formulation for use on structured grids. In particular, WENO and compact upwind schemes as well as central and upwind schemes are investigated. Their efficiency and accuracy are compared to the well-established second- and third-order MUSCL family of schemes. The different schemes are analyzed and tested numerically using canonical flow problems in the context of the two-dimensional Euler equations. Results are shown for the two-dimensional advection of a density pulse to quantify the dissipation and dispersion properties of the schemes. To investigate the non-oscillatory nature of the schemes and the resolution of discontinuities the Riemann problem is analyzed. The simulations are performed on uniform as well as non-uniform meshes.*

# 1 INTRODUCTION

Over the past decades there has much research into the design and application of high-order accurate numerical methods. In the context of computational fluid dynamics (CFD) high-order methods are especially desirable to simulate flows with complicated structures. In compressible flows, the existence of shocks, interfaces and other discontinuities present additional challenges to high-order schemes as, alongside providing high-order accuracy in smooth regions of the solution, they must remain non-oscillatory and stable in regions of high nonlinearity.

This paper examines the behavior of various high-order schemes in a finite volume formulation for the use of structured grids and compares their efficiency and accuracy to the well-established second- and third-order MUSCL family of schemes. In particular, weighted essentially non-oscillatory (WENO) and compact schemes show great potential for further improvements due to the fact that these schemes are active fields of research since many years. But also central and upwind schemes are good candidates for very efficient high-order schemes by the reason of their relatively simple nature.

The different schemes are analyzed and tested on both uniform and non-uniform grids in the context of the two-dimensional Euler equations. The accuracy, convergence and resolution are studied for the advection of a Gaussian entropy pulse and the non-oscillatory behavior across the discontinuities is investigated with the classical shock tube test problem by Sod [1].

# 2 THE FINITE VOLUME SCHEMES

## 2.1 Governing Equations

The flow of two-dimensional, compressible, inviscid fluid can be described in conservation form by the Euler equations:

$$\frac{\partial Q}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} = 0 \tag{1}$$

where $Q = (\rho, \rho u, \rho v, \rho e)^T$ and

$$F = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u\left(\rho e + p\right) \end{pmatrix}, \quad G = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v\left(\rho e + p\right) \end{pmatrix}.$$

Here $p$, $\rho$, $u$, $v$ and $e$ are the pressure, density, velocity components and the total energy per unit mass, respectively. The pressure is given by the equation of state for perfect gas:

$$p = \left(\gamma - 1\right)\rho\left(e - \frac{1}{2}\left(u^2 + v^2\right)\right)$$

where $\gamma$ is the ratio of specific heats.

Considering a structured mesh, the finite volume formulation is derived by integrating Eq. 1 on an interval $\left[x_{i-1/2,j}, x_{i+1/2,j}\right] \times \left[y_{i,j-1/2}, y_{i,j+1/2}\right]$

$$\frac{\partial}{\partial t}\bar{Q}_{i,j} = -\frac{1}{\Delta_i x \Delta_j y}\left(F_{i+1/2,j} - F_{i-1/2,j} + G_{i,j+1/2} - G_{i,j-1/2}\right) \tag{2}$$

where

$$\bar{Q}_{i,j} = \frac{1}{\Delta_i x \Delta_j y}\int_{x_{i-1/2,j}}^{x_{i+1/2,j}}\int_{y_{i,j-1/2}}^{y_{i,j+1/2}} Q dx dy$$

is the cell average of $\bar{Q}$ in the $(i,j)$-th cell. The Gaussian quadrature used to integrate the flux on each cell interface is second-order accurate.

Temporal integration in Eq. 2 can be accomplished by treating this as a system of ordinary differential equations, via a method-of-lines approach. In particular, the explicit third-order Runge-Kutta method of Gottlieb and Shu [2] will be used. This method is high-order accurate and total-variation diminishing (TVD) in the sense that the temporal operator does not increase the solution's total variation in time.

$$\begin{aligned}\bar{Q}^{(1)} &= \bar{Q}^{(n)} + \Delta t \mathbf{R}(\bar{Q}^{(n)}) \\ \bar{Q}^{(2)} &= \frac{3}{4}\bar{Q}^{(n)} + \frac{1}{4}\bar{Q}^{(1)} + \frac{1}{4}\Delta t \mathbf{R}(\bar{Q}^{(1)}) \\ \bar{Q}^{(n+1)} &= \frac{1}{3}\bar{Q}^{(n)} + \frac{2}{3}\bar{Q}^{(2)} + \frac{2}{3}\Delta t \mathbf{R}(\bar{Q}^{(2)})\end{aligned}$$

Where $\mathbf{R}$ is the right hand side of Eq. 2.

The fluxes in $\mathbf{R}$ are approximated by solving the local Riemann problems at the cell interfaces with a solver due to Roe [3] for all asymmetric discretization schemes. E.g. for

$$F_{i+1/2,j} = \frac{1}{2}\left(F^L_{i+1/2,j} + F^R_{i+1/2,j}\right) - |A|\left(U^L_{i+1/2,j} + U^R_{i+1/2,j}\right)$$

where $A$ is the Roe matrix with $U = (\rho, u, v, p)^T$. In the case of symmetric schemes, for which $U^L$ and $U^R$ are identical the fluxes are directly evaluated from the flow state at the cell interface position.

### 2.2 Discretization Schemes

To obtain the face fluxes the left and right states of $U$ are computed. There are several choices of the variable to be reconstructed, such as the primitive, conservative or characteristic variables. In the context of this paper for all investigated schemes the primitive variables are used even if it is recommended to use the characteristic variables in context of WENO schemes for shock simulation (see e.g. Borges et al. [4]). The reconstruction with all investigated schemes are accomplished in a dimension-by-dimension fashion. For the sake of brevity only the reconstruction of the left state is denoted here.

The first class of discretization schemes are the well-established second- and third-order MUSCL (Monotonic Upstream-Centered Scheme for Conservation Laws) family.

$$U^L_{i+1/2} = \bar{U}_i + \frac{1}{4}\left[(1+\kappa)\, r_i \Phi\left(1/r_i\right) + (1-\kappa)\,\Phi\left(r_i\right)\right]\left(\bar{U}_i - \bar{U}_{i-1}\right) \tag{3}$$

The function $\Phi\left(r_i\right)$ is a limiter function to limit the slopes $r_i$. The parameter k determines the spatial accuracy of the interpolation. In this work the minmod, Van Leer, superbee, monotonized central, Van Albada and Van Albada 2 limiter functions are considered for uniform meshes. Zeng [5] proposed also non-uniform versions of the limiter functions for minmod, Van Leer, superbee, monotonized central and Van Albada. Leng et al. [6] developed an optimized MUSCL scheme to minimize the dissipation and dispersion error in combination with the monotonized central limiter. This family of schemes is usually second or third-order, since it is difficult to develop limiter functions for higher-order MUSCL schemes.

The second class of discretization schemes are the WENO (weighted essentially non-oscillatory) schemes.

$$U^L_{i+1/2} = \sum_{k=0}^{r-1} \omega_k q^r_k \tag{4}$$

$q_k^r$ are $r$th-order accurate polynomial interpolants evaluated at $x_{i+1/2}$ computed by

$$q_k^r = \sum_{l=0}^{r-1} a_{kl}^r \bar{U}_{i-r+k+l+1}$$

where $a_{kl}^r$ are stencil coefficients. The weights are defined by

$$\omega_k = \frac{\alpha_k}{\sum_{k=1}^{r-1} \alpha_k} \tag{5}$$

and

$$\alpha_k = \frac{C_k^r}{(\epsilon + \beta_k)^2}.$$

where

$$\beta_k = \sum_{m=1}^{r-1} \left( \sum_{l=1}^{r-1} d_{kml}^r \bar{U}_{i-r+k+l+1} \right)^2$$

and $\epsilon$ is small number designed to prevent division by zero. The smoothness indicators $\beta_k$ become large when discontinuities are present within the stencil $k$ and remain relatively small otherwise.

In the case of odd-order WENO schemes considered here an one point upwind biased interpolation polynomials are used for $q_k^r$. For the classical fifth-order WENO scheme after Jiang and Shu [7] the above definition results in a convex combination of three third-order candidate stencils. This version is denoted as WENOJS in the rest of the paper. In critical points where multiple derivatives vanish the weight $\omega_k$ are far away from the optimal weights $C_k^r$ and the order of convergence decreases dramatically. In the literature several approaches have been proposed to recover the formal order of accuracy in the smooth regions of the solution. In this work, in addition to the above formulation in Eq. 5, the modified weighting procedures WENOM after Hendricks et al. [8], WENOZ after Borges et al. [4], WENOYC after Yamaleev and Carpenter [9], WENOZS after Shen and Zha [10] and WENON after Xiaoshuai and Yuxin [11] are considered. Within the context of WENO schemes the sensitivity parameter $\epsilon$ is also a subject of active research and can cause a drop of the rate of convergence in some cases. In this work both constant values ranging between $10^{-60}$ and $10^{-2}$ are used as well as methods suggested by Peer et al. [12], Yamaleev and Carpenter [9] and Don and Borges [13].

Martin et al. [14] developed a central symmetric WENO scheme. Since in classical WENO schemes the center of the stencil collection lies slightly upwind of the reconstruction point the WENO scheme proposed by Martin et al. uses an additional stencil to form a symmetric central scheme which is capable of providing zero dissipation.

Additional class of schemes investigated are compact schemes. The general form of the interpolation formula for the cell faces is

$$\alpha U_{i-1/2}^L + U_{i+1/2}^L + \beta U_{i+3/2}^L = a\bar{U}_{i-2} + b\bar{U}_{i-1} + c\bar{U}_i + d\bar{U}_{i+1}. \tag{6}$$

Here only compact upwind interpolations with narrow stencils are considered so as to use efficient matrix inversion algorithms of a tridiagonal matrix, such as the Thomas algorithm. The used coefficients of Eq. 6 can be found in Tab. 1. For a more detailed description of the schemes see Broeckhoven et al. [15]. Following the notation of Broeckhoven et al. two different types of upwind-biased interpolations are investigated. The implicit (left-hand side) part purely central and the explicit (right-hand side) part upwind-biased (IC$x$EU$yz$). Here $x$ is the number of

| Scheme | Order | $\alpha$ | $\beta$ | $a$ | $b$ | $c$ | $d$ |
|---|---|---|---|---|---|---|---|
| IC3EU22 | 3 | $-\frac{1}{5}$ | $-\frac{1}{5}$ | 0 | $-\frac{3}{10}$ | $\frac{9}{10}$ | 0 |
| IC3EU33 | 4 | $-\frac{1}{8}$ | $-\frac{1}{8}$ | $\frac{1}{8}$ | $-\frac{5}{8}$ | $\frac{5}{4}$ | 0 |
| IC3EU43 | 5 | $-\frac{3}{14}$ | $-\frac{3}{14}$ | $\frac{1}{84}$ | $-\frac{5}{84}$ | $\frac{67}{84}$ | $\frac{19}{28}$ |
| IU22EU21 | 3 | $\frac{1}{2}$ | 0 | 0 | 0 | $\frac{5}{4}$ | $\frac{1}{4}$ |
| IU32EU32 | 5 | $\frac{1}{2}$ | $\frac{1}{6}$ | 0 | $\frac{1}{18}$ | $\frac{19}{18}$ | $\frac{5}{9}$ |

Table 1: Overview of different coefficients of compact schemes

points in the implicit and $y$ the number of points in the explicit part. $z$ indicates the position of $\bar{U}_i$. And both implicit and explicit part upwind-biased (IU$xy$EU$zw$). Here $x$ is the number of points in the implicit and $z$ the number of points in the explicit part. $y$ indicates the position of $U_{i+1/2}$ and $w$ indicates the position of $\bar{U}_i$. For periodic problems these schemes are solved using the Thomas Algorithm for periodic tridiagonal systems. Otherwise the standard Thomas Algorithm is used by applying different boundary closures to IU32EU32 (compactBC), which can be found in Pirozzoli [16] and Ghosh and Baeder [17]. Also the impact to the accuracy and stability of the application of an explicit central filter is investigated to reduce non-physical oscillations.

The last class of schemes are simple explicit upwinding (Eq. 7), symmetric (Eq. 8) and asymmetric (Eq. 9) central schemes without any stabilization techniques such as limiters or filtering. These schemes are based upon Lagrange interpolation polynomials which are also used to construct the WENO schemes. Therefore, they can be considered as WENO schemes with fixed optimal weights $C_k^r$.

$$U^L_{i+1/2} = a\bar{U}_{i-3} + b\bar{U}_{i-2} + c\bar{U}_{i-1} + d\bar{U}_i + e\bar{U}_{i+1} + f\bar{U}_{i+2} + g\bar{U}_{i+3} \tag{7}$$

$$U^{L,R}_{i+1/2} = a\bar{U}_{i-3} + b\bar{U}_{i-2} + c\bar{U}_{i-1} + d\bar{U}_i + d\bar{U}_{i+1} + c\bar{U}_{i+2} + b\bar{U}_{i+3} + a\bar{U}_{i+4} \tag{8}$$

$$U^L_{i+1/2} = a\bar{U}_{i-3} + b\bar{U}_{i-2} + c\bar{U}_{i-1} + d\bar{U}_i + e\bar{U}_{i+1} + f\bar{U}_{i+2} + g\bar{U}_{i+3} + h\bar{U}_{i+4} \tag{9}$$

Tab. 2 provides an overview of the reviewed methods, their order of accuracy and used acronym for the rest of this paper. The theoretical order of convergence is denoted with a number after the scheme acronym if appropriate.

## 3 NUMERICAL RESULTS

The application of the described schemes to the two-dimensional Euler equations is discussed in this section. The accuracy and convergence properties are studied on a smooth problem that involves the advection of a Gaussian entropy pulse. The non-oscillatory nature of the schemes and the resolution of discontinuities are studied in this section using the Riemann problem.

### 3.1 Entropy Advection

The advection of a smooth density pulse over a periodic domain is considered in this section. The pressure and velocity fields are constant, thus reducing the Euler equations to a linear advection equation for density, with the advection speed as the freestream velocity. The exact

| Scheme | Acronym | Implemented Order | Equation |
|---|---|---|---|
| Monotonic Upstream-Centered Scheme for Conservation Laws | MUSCL | 1, 2, 3 | 3 |
| Optimized MUSCL | oMUSCL | 2 | - |
| MUSCL for non-uniform grids | nuMUSCL | 2 | - |
| Weighted Essentially Non-oscillatory Schemes | WENO | 3, 5, 7 | 4 |
| Central symmetric WENO | CWENOSym | 4, 6, 8 | - |
| Compact upwind scheme | compact | 3, 4, 5 | 6 |
| one point upwind biased scheme | upwind | 3, 5, 7 | 7 |
| central symmetric schemes | centralSym | 2, 4, 6, 8 | 8 |
| centralSym for non-uniform grids | nuCentralSym | 2, 4, 6, 8 | 8 |
| central asymmetric schemes | centralASym | 2, 4, 6, 8 | 9 |
| optimized centralASym | oCentralSym | 5, 7 | 9 |

Table 2: Overview of review discretization schemes

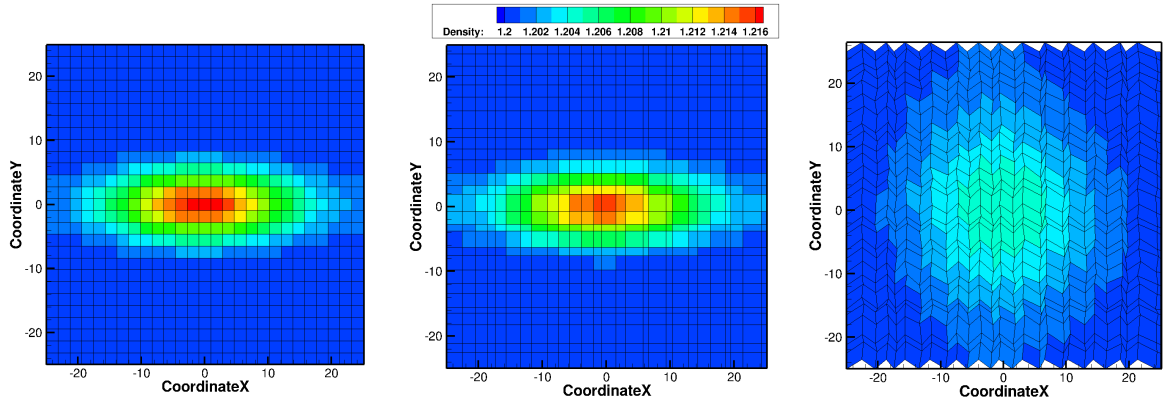
solution is given by:

$$\begin{aligned}
\rho\left(x,y,t\right) &= \rho_\infty + Ae^{-\left((x-u_\infty t)^2+y^2\right)} \\
u\left(x,y,t\right) &= u_\infty \\
v\left(x,y,t\right) &= v_\infty \\
p\left(x,y,t\right) &= p_\infty
\end{aligned}$$

In the present example, the freestream conditions are taken as $\rho_\infty = 1.2\,\mathrm{kg/m^3}$, $u_\infty = 180\,\mathrm{m/s}$, $v_\infty = 0\,\mathrm{m/s}$, $p_\infty = 94040\,\mathrm{Pa}$ and the amplitude of the entropy wave is taken as $A = 0.05\,\mathrm{kg/m^3}$. The domain is taken as $[-25\,\mathrm{m}, 25\,\mathrm{m}] \times [-25\,\mathrm{m}, 25\,\mathrm{m}]$ and periodic boundaries are assumed on both sides. The solution is obtained after one cycle ($t = L/u_\infty = 0.28\,\mathrm{s}$) and the $L_1$ error norms are compared. An initial grid with $21$ points is taken and progressively refined. The CFL number corresponding to the grid with $21$ points is $0.08$ and is decreased with each grid refinement, to ensure that errors due to time discretization converge at the same order as those due to space discretization.

The simulations were carried out on three different grids, see Fig. 1. This first consists of uniform distributed cells. The second is a non-uniform grid where every $i$ constant line (and every $j$ constant line) has been shifted in the $x$-direction (or respectively the $y$-direction) by a random distance. The last investigated grid is additionally distorted by a random shift of every vertex and every second $i$ constant line shifted in $y$-direction by a constant distance.

Fig 2 shows the convergence rate of the various scheme classes on the uniform grid. The $L_1$ errors from the simulations with different formulation in one class of schemes are averaged. This way it is possible to compare not only single schemes with each other but rather to get an overview of the different classes. With every grid refinement the error decreases. But since every scheme shows the theoretical order of convergence for this simple test case a clustering of schemes can be observed.




(a) uniform grid (b) non-uniform grid (c) distorted grid

Figure 1: Solution of the advection of a density pulse with MUSCL1

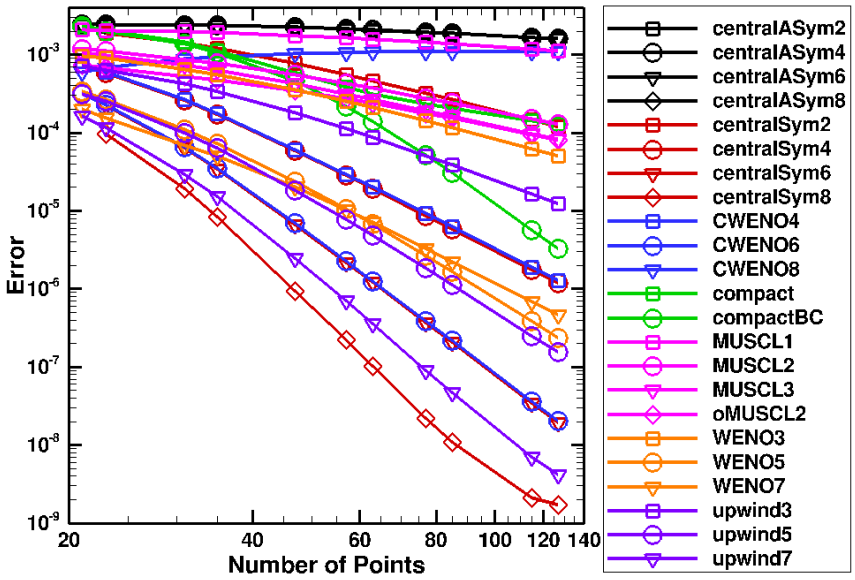
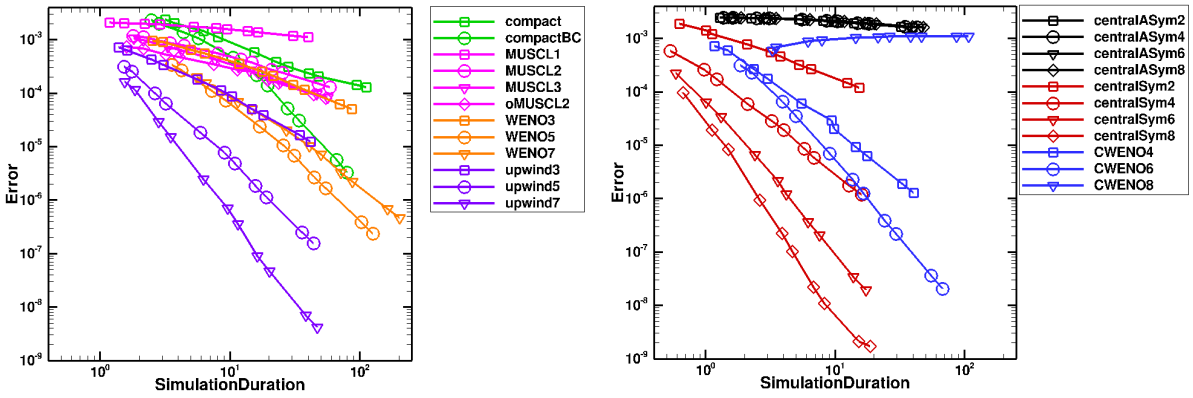



Figure 2: Convergence rate of various scheme classes for the advection of a density pulse on uniform grids

In order to choose an efficient scheme, it is helpful to consider the simulation duration instead of the number of cells for a given accuracy. In Fig. 3 error versus the wall-clock time required to obtain the solution on the given grid. Each symbol represents one grid tested with all schemes. Here it is easy to see that the simple schemes use significant less computational effort to reach a certain level of error. E.g. the advantage of the upwind5 compared to the WENO5 scheme is evident in Fig. 3(a) but not in Fig. 2. Since the upwind and central schemes are constructed without any shock-capturing abilities, they offer a very efficient possibility to simulate wave propagation phenomenon in smooth flow regimes. Also the high-order versions are preferable due to lower dissipation. In the category of shock-capturing schemes, in the direct comparison of MUSCL3 and WENO3, it appears that the computational cost to determine the non-linear weights in the WENO schemes can not be compensated by a lower error. Especially the oMUSCL2 scheme shows advantages on coarse grids. Due to the wave number optimization they are able to provide lower dissipation on these grids. The compact schemes are not able to show the ability to provide a high spectral resolution. Here spurious oscillations appear even with the use of upwind biased compact schemes which contaminate the solution. To address this problem an explicit central filter (see [18]) was used but it was found to reduce the accuracy. Still compact schemes including an inversion of a tridiagonal matrix seems to be faster than WENO schemes. For the class of central schemes shown in Fig. 3(b) there is a similar pattern. The centralSym schemes are very efficient since there is no need to solve a Riemann problem at the cell interfaces. So the computational effort is smaller than for any upwind biased scheme. However, their use is limited to only a few very simple flow conditions. CWENO schemes have also no need for a Riemann solver but the computation of the expensive non-linear weights is needed. The high error in the solutions with the centralASym is due to the fact than the advection speed is overestimated.




(a) upwind biased schemes (b) central schemes

Figure 3: Efficiency of various scheme classes for the advection of a density pulse on uniform grids

On non-uniform grids all schemes investigated produce significantly higher errors and even the rate convergence drops to mainly second-order. Furthermore, it is found that the upwind schemes designed for uniform grids produce the lowest error. Only the non-uniform formulation of the MUSCL2 schemes can give comparable results. Most notably the compact schemes are very sensitive to non-uniform spacing and have only first-order accuracy. However, most central schemes were not able to provide acceptable solutions due to the high stretching ratios in the grid. For the third grid type, the distorted grid, the situation is even worse. No scheme class is able to achieve a first-order convergence rate. Actually the nuMUSCL2 have a convergence

rate of $0.48$. This scheme is designed to take a stretching ratio into account, but still relies on cell orthogonal to the flow direction.

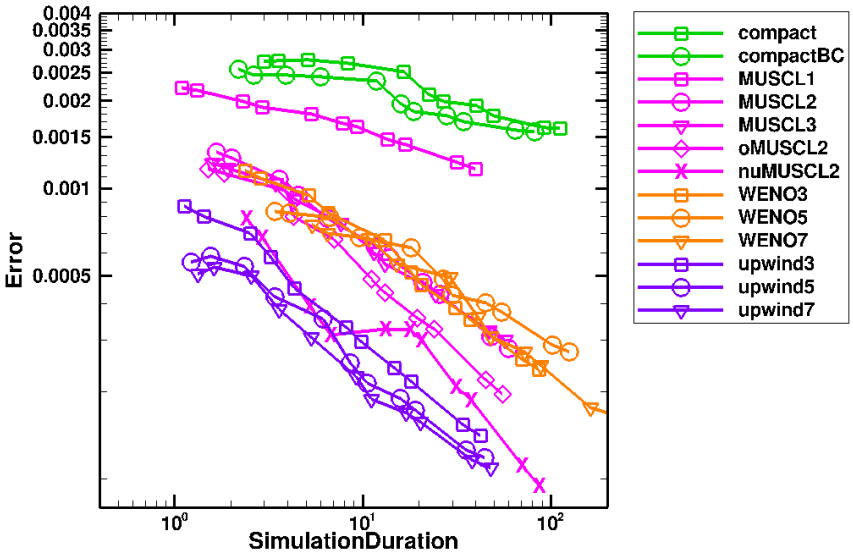



Figure 4: Efficiency of upwind biased schemes for the advection of a density pulse on non-uniform grids

### 3.2 Sod Shock Tube Problem

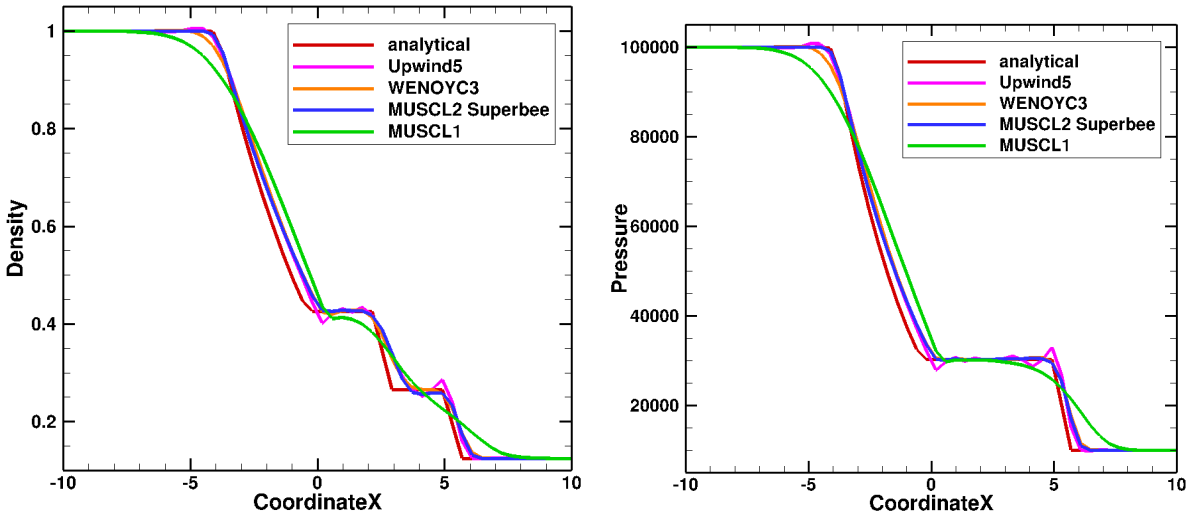
The non-oscillatory nature of the schemes and the resolution of discontinuities are assessed using a Riemann problem. The problem consists of an initial discontinuity that decomposes into a rarefaction wave, a contact discontinuity and a shock wave, corresponding to each of the characteristic fields of the Euler equations. The Sod shock tube problem [1] is an inviscid benchmark problem. The exact solution to this problem is obtained using a Riemann solver. The initial conditions are:

$$\rho_L = 1\,\mathrm{kg/m^3},\ u_L = 0\,\mathrm{m/s},\ v_L = 0\,\mathrm{m/s},\ p_L = 100000\,\mathrm{m/s}$$
$$\rho_R = 0.125\,\mathrm{kg/m^3},\ u_R = 0\,\mathrm{m/s},\ v_R = 0\,\mathrm{m/s},\ p_R = 10000\,\mathrm{m/s}$$

The domain is taken as $[-10\,\mathrm{m}, 10\,\mathrm{m}] \times [-10\,\mathrm{m}, 10\,\mathrm{m}]$ and the initial discontinuity is placed at $x = 0\mathrm{m}$. The solution is obtained at $t = 0.01\,\mathrm{s}$ on a grid with $20$ points, a CFL number of $0.1$ and is decreased with each grid refinement.

The analytical solution for the density and pressure at $y = 0\mathrm{m}$ is shown in Fig. 5. In addition, the simulation results obtained on an uniform grid with the first-order MUSCL scheme, the most efficient schemes of the MUSCL and WENO classes and upwind scheme are presented. Due to the relatively high dissipation of the MUSCL1 scheme the discontinuity is smeared while the Upwind5 scheme shows strong oscillations. The MUSCL2 scheme with the superbee limiter and WENOYC3 are able to sharply resolve the contact discontinuity and the shock wave.

An overview of the efficiency averaged for the different classes can found in Fig. 6. The results for compact schemes are not shown since it was not possible to obtain stable solutions for this test case. The other central schemes as expected suffer from severe oscillations and are not able to produce comparable smooth solutions. The WENO and upwind (a.k.a. optimal WENO) classes show a similar efficiency behavior. The calculation of the non-linear weights in the WENO schemes produces a large computational overhead that is not compensated by the




(a) Density (b) Pressure

Figure 5: Solution of Sod shock tube problem with exemplary schemes

higher accuracy. However, this observation does no longer hold for stronger shocks since the stencil of the simple upwind schemes uses cells containing the shock and becomes unstable. Through the construction of WENO schemes the use of these cells is reduced. The candidate stencils which contain the discontinuities are assigned a nearly zero weight in the formation of the final interpolation polynomial. The well-established MUSCL scheme class is able to create fast a solution with only minimal oscillations at a low level of errors.

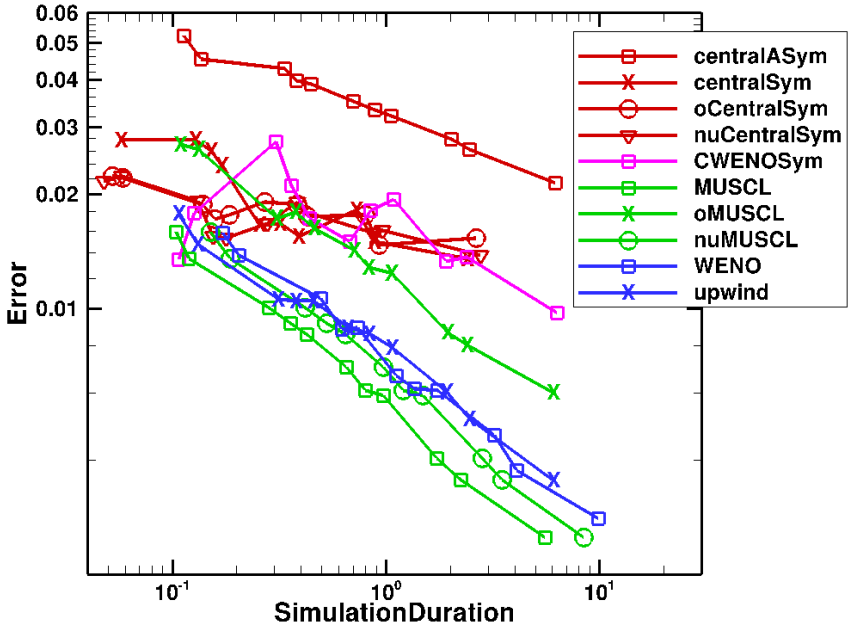



Figure 6: Efficiency for the Sod shock tube problem

To have a more detailed look only the best versions of each upwind biased class is shown in Fig. 7. For the MUSCL class the second-order schemes seem to represent the best choice. Within the context of the WENO schemes only the seventh-order method is not able the compensate the computational overhead with a lower error. Note, even the most efficient versions

of these shock-capturing schemes need significantly more computational time than the upwind schemes. Since this test case simulates a moderate shock problem the oscillations are not critical for the stability of the simulation. But in the presence of stronger shocks the MUSCL2 and WENO5 versions show better results. The non-uniform version of the MUSCL family is also

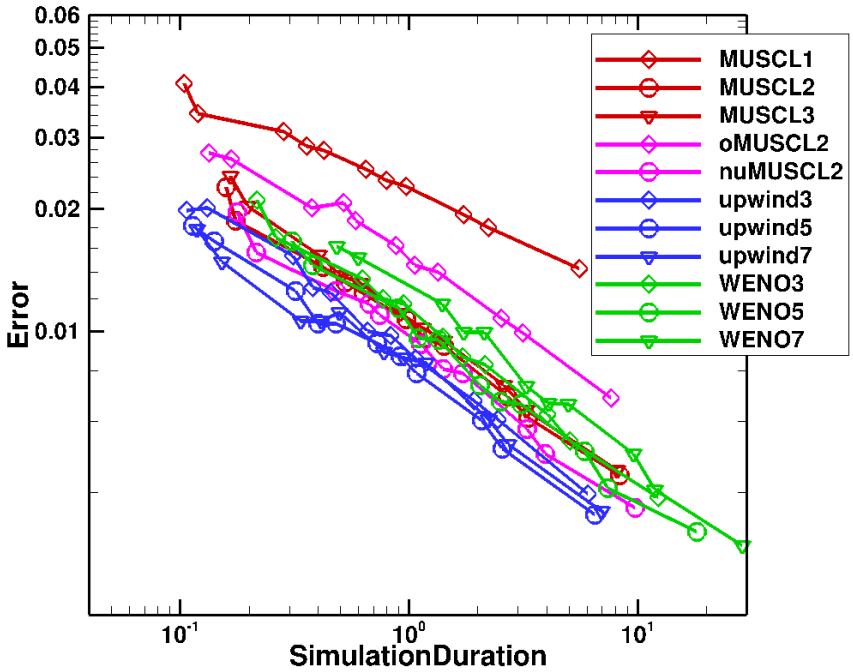



Figure 7: Efficiency of best versions of upwind biased scheme classes

able to reach the same accuracy, but is slightly more expensive. When applied to non-uniform grids, however these methods are found to be more efficient than their uniform counterparts as non-uniform parameter provides greater accuracy.

Of particular interest for the accuracy of WENO schemes especially near critical points, is the formulation of the non-linear weights. In Fig. 8 the analysis of all used accuracy order and sensitivity parameter formulations illustrated. Here the average of all WENO schemes, alongside the method with the best and worst efficiency is shown. The WENOM formulation has a significant larger computational cost. The other versions lie close together while the WENOYC seems to be the best choice within the WENO family.

In Tab. 3 the convergence order of the analyzed sets of schemes is documented. No class is able to provide a first-order convergence for the $L_1$-error even on uniform grids. The WENO schemes have a slightly better convergence rate than the MUSCL schemes. However, in terms of efficiency they are not able to outperform the MUSCL schemes due to the expensive calculation of the non-linear weights. In general, in the case of stronger shocks or non-uniform or distorted grids the total error increases and the order of convergence drops. Although the effect is only marginal this holds also for non-uniform schemes.

## 4 CONCLUSIONS

In this paper, the well-established MUSCL family of schemes are compared to modern shock-capturing WENO schemes, compact schemes of high-order accuracy and simple upwind and central schemes. Numerical experiments were performed by means of academic test cases within the framework of the two-dimensional Euler equations on uniform and non-uniform

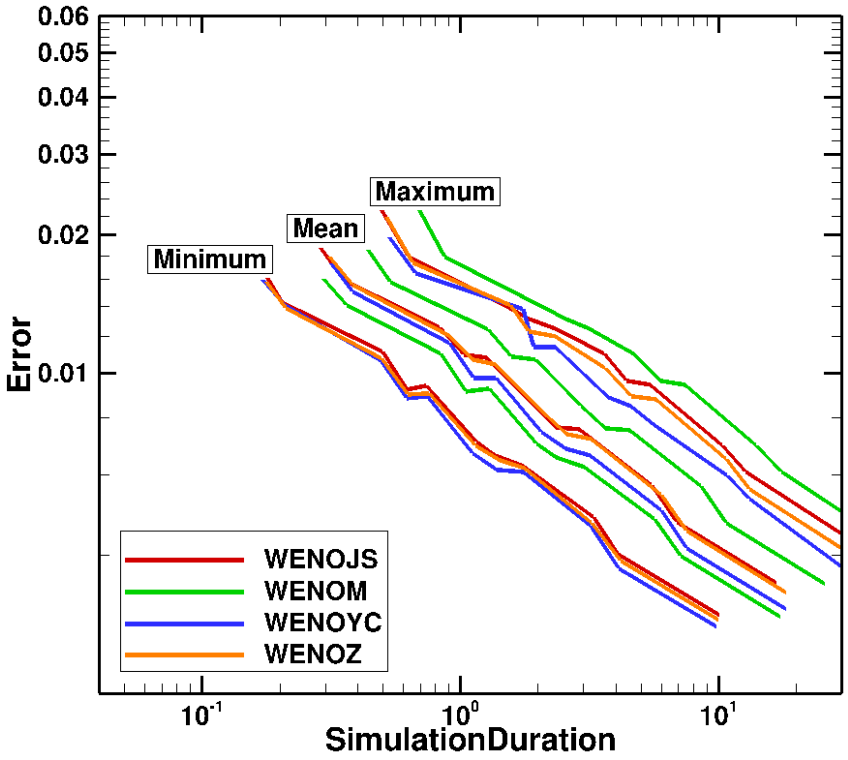



Figure 8: Efficiency distribution of WENO weight formulations

| Scheme class | | Convergence Order | |
|---|---|---|---|
| | | uniform grid | non-uniform grid |
| MUSCL | | 0.83 | 0.70 |
| | MUSCL1 | 0.54 | 0.52 |
| | MUSCL2 | 0.86 | 0.73 |
| | MUSCL3 | 0.89 | 0.72 |
| | oMUSCL2 | 0.69 | 0.41 |
| | nuMUSCL2 | 0.89 | 0.81 |
| WENO | | 0.87 | 0.78 |
| | WENOJS | 0.87 | 0.75 |
| | WENOM | 0.86 | 0.66 |
| | WENOYC | 0.90 | 0.83 |
| | WENOZ | 0.87 | 0.77 |
| upwind | | 0.80 | 0.77 |
| centralASym | | 0.47 | 0.40 |
| centralSym | | 0.45 | 0.34 |

Table 3: Order of convergence for Sod shock tube problem on uniform grids for different scheme classes

grids. It is found, for a given computational effort, that the high-order schemes are better able to accurately propagate wave disturbances. In particular, the simple upwind and central schemes show high efficiency in smooth flow regimes on uniform grids. On non-uniform grids, however, the order of accuracy drops for all investigated schemes and the absolute error level increases. Here the extra computational cost for non-uniform coefficients is found to be justified. In the case of discontinuities, the shock-capturing schemes are found to be superior over the simple schemes, since these schemes are not able to obtain an oscillation free stable solution. In terms of efficiency the MUSCL family provides good results. The WENO schemes show comparable results, but are not able to outperform the MUSCL schemes due to the high computational cost of the non-linear weights.

## ACKNOWLEDGMENTS

The author would like to thank the Numerical Methods department at the Institute of Propulsion Technology at DLR and in particular Graham Ashcroft for his advice and many helpful discussions on this research.

## REFERENCES

[1] Gary A. Sod. A survey of finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.*, 27:1 – 31, 1978.

[2] Sigal Gottlieb and Chi-Wang Shu. Total variation diminishing Runge-Kutta schemes. *Mathematics of computation*, 67(211):73 – 85, January 1998.

[3] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357–372, 1981.

[4] Rafael Borges, Monique Carmona, Bruno Costa, and Wai Sun Don. An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws. *Journal of Computational Physics*, 227(6):3191 – 3211, 2008.

[5] Xianyi Zeng. A general approach to enhance slope limiters on non-uniform grids. *SIAM J. Sci. Comput.*, January 2013.

[6] Yan Leng, XinLiang Li, DeXun Fu, and YanWen Ma. Optimization of the MUSCL scheme by dispersion and dissipation. *Science China Physics, Mechanics and Astronomy*, 55(5):844 – 853, 2012.

[7] Guang-Shan Jiang and Chi-Wang Shu. Efficient implementation of weighted ENO schemes. *J. Comput. Phys.*, 126(1):202 – 228, June 1996.

[8] Andrew K. Henrick, Tariq D. Aslam, and Joseph M. Powers. Mapped weighted essentially non-oscillatory schemes: Achieving optimal order near critical points. *J. Comput. Phys.*, 207(207):542 – 567, March 2005.

[9] Nail K. Yamaleev and Mark H. Carpenter. Third-order energy stable WENO scheme. *J. Comput. Phys.*, 228(8):3025 – 3047, 2009.

[10] Yiqing Shen and Gecheng Zha. Improvement of weighted essentially non-oscillatory schemes near discontinuities. *Computers & Fluids*, 96:1 – 9, February 2014.

[11] Wu Xiaoshuai and Zhao Yuxin. A high-resolution hybrid scheme for hyperbolic conservation laws. *Int. J. Numer. Meth. Fl.*, 78(3):162 – 187, 2015.

[12] A.A.I. Peer, M.Z. Dauhoo, and M. Bhuruth. A method for improving the performance of the WENO5 scheme near discontinuities. *Applied Mathematics Letters*, 22(11):1730 – 1733, 2009.

[13] Wai-Sun Don and Rafael Borges. Accuracy of the weighted essentially non-oscillatory conservative finite difference schemes. *J. Comput. Phys.*, 250:347 – 372, October 2013.

[14] M. P. Martin, E. M. Taylor, M. Wu, and V. G. Weirs. A bandwidth-optimized WENO scheme for the effective direct numerical simulation of compressible turbulence. *J. Comput. Phys.*, 220(1):270 – 289, December 2006.

[15] Tim Broeckhoven, Sergey Smirnov, Jan Ramboer, and Chris Lacor. Finite volume formulation of compact upwind and central schemes with artificial selective damping. *J. Sci. Comput.*, 21(3):341 – 367, December 2004.

[16] Sergio Pirozzoli. Conservative hybrid compact-WENO schemes for shock-turbulence interaction. *J. Comput. Phys.*, 178(1):81 – 117, May 2002.

[17] Debojyoti Ghosh and James D. Baeder. Compact Reconstruction Schemes with Weighted ENO Limiting for Hyperbolic Conservation Laws. *SIAM J. Sci. Comput.*, 34(3):A1678 – A1706, January 2012.

[18] Mei Zhuang and Christoph Richter. Computational aeroacoustics and its applications, 2007. Notes of the lecture series Computational Aeroacoustics at the Technical University Berlin.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NURBS-BASED ISOGEOMETRIC ANALYSIS FOR BALLISTIC EVALUATION OF TITANIUM PLATES

**Mattia Montanari**[1], **Petros Siegkas**[1], **Antonio Pellegrino**[1], **Nik Petrinic**[1]

[1]Department of Engineering Science, University of Oxford
Parks Road, OX1 3PJ, Oxford, UK
e-mail: mattia.montanari@eng.ox.ac.uk

**Keywords:** Isogeometric analysis, NURBS, Impact testing, Titanium alloy, High strain rate.

**Abstract.** *The variation diminishing property of non-uniform rational B-splines (NURBS) makes NURBS-based isogeometric analysis attractive for applications involving travelling stress waves generated, for example, by impulsive loads. Over their finite elements counterparts, the advantage of NURBS basis functions is the higher continuity at every point of single-patch domains. This research investigates the applicability of isogeometric analysis for the ballistic performance of titanium alloys. Numerical results, in the framework of explicit time-integration, present the effect for $k$-refined trivariate NURBS. A comparison with the classical finite element analysis is also presented. Although the global continuity of NURBS results into a superior approximation of the travelling waves, it does not allow physical discontinuities in the displacement field. This prevents, for example, a projectile to penetrate a target plate. It is argued that the realistic physical response can be captured by adopting material models incorporating strain-rate-sensitive failure criteria.*

## 1 INTRODUCTION

Ballistic tests provide greater insight into the dynamic response of mechanical components. Usually, ballistic tests aim to measure the ballistic limit of a target: the threshold velocity for which penetration occurs. Nowadays, these tests are not restricted to military applications; bird-strike and safety equipment tests are only few of the many case studies involving ballistic experiments.

The complexity and the high costs of experimental test are often a limitation to the accurate measurement of the ballistic limit. Numerical simulations provide tangible benefits and are, in fact, an integral part of the design process for mechanical components undergoing ballistic loads. The general tool adopted in many engineering communities is the finite element analysis (FEA). Several works using this method have been published in the fields of aerospace and mechanical engineering to model ballistic tests on steel [1], woven fabric [2] and titanium plates [3]. We now want to expand these studies to the a higher-order approximation method by means of isogeometric analysis.

Isogeometric analysis (IGA) [4] is a numerical paradigm that has several advantages compared to the classical FEA. Firstly, by employing non-uniform rational B-splines (NURBS) as basis functions, IGA avoids the geometrical inaccuracies introduced by the FEA meshes. Moreover, spline provide a superior approximation to stress travelling waves, making IGA appealing for dynamic simulations.

During perforation, FEA models are usually coupled with element erosion techniques that have major drawbacks [5, 6]. By taking away mass from the system, these methods inevitably compromise the inertia and thus the dynamic response of the system itself. If the features of spalled plug or residual crater are not of primary interest, and if the ultimate goal of the investigation is to predict whether penetration would occur or not, element erosion can be avoided altogether. More accurate numerical methods allowing for cracks growth have been proposed in the literature. These approaches face enormous computational challenges as they have to take into account for geometrical instabilities and discontinuities [7]. The extended finite element method (X-FEM) [8], reproducing kernel particle method (RKPM) [9] and phase field models [10] are just few of the well-established approaches available in the literature. Nevertheless, the advantages of such methods come at the price of a higher computational cost. For this study a pragmatic and closer to in-service engineering applications is pursued.

To predict the ballistic limit we propose to combine IGA with a material model with failure criterion. For this work the constitutive model developed by Børvik et al [11] was selected. Originally formulated starting from the Johnson-Cook [12] material model and the continuum damage mechanics model by Lamaitre [13], this is one of the most diffused constitutive models for the simulation of ballistic tests. At last, we provide a comparison of numerical results obtained with IGA and FEA.

## 2 NURBS-BASED ISOGEOMETRIC ANALYSIS

Higher-order NURBS shape function have a great advantage over their finite element counterparts, that is, splines do not oscillate nor overshoot near discontinuities. Instead, higher-order polynomials yield to the so-called Gibbs phenomenon [14]. Another advantage is that NURBS are found in the every computer aided design (CAD) software packages for engineers. NURBS-based IGA represents CAD models exactly and thus reduces the gap between CAD and numerical analysis.

A NURBS curve of order $p$ is defined over a one-dimensional parametric domain as:

$$\mathbf{C}(u) = \frac{\sum_{i=0}^{n} N_{i,p}(u) w_i \mathbf{P}_i}{\sum_{i=0}^{n} N_{i,p}(u) w_i} \qquad 0 \leq u \leq 1 \tag{1}$$

The control points $\mathbf{P}_i$ map the curve from parametric to Euclidean three-dimensional space $\mathbb{R}^3$. To each control point is associated a weight $w_i$ which, together with the B-spline basis function $N_{i,p}(u)$, defines the rational basis function $R_{i,p}(u)$ such that:

$$\mathbf{C}(u) = \sum_{i=0}^{n} R_{i,p}(u) \mathbf{P}_i \tag{2}$$

The most common formulation of B-spline basis function for computer application is the recursive Cox-de Boor [15] formula:

$$\begin{aligned} N_{i,0}(u) &= \begin{cases} 1 & \text{if} \quad u_i \leq u < u_{i+1} \\ 0 & \text{otherwise} \end{cases} \\ N_{i,p}(u) &= \frac{u - u_i}{u_{i+p} - u_i} N_{i,p-1}(u) + \frac{u_{i+p+1} - u}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(u) \end{aligned} \tag{3}$$

The functions in Eqs. (1) and (2) are defined on a knot vector $\Xi$, which is a nondecreasing and nonuniform sequence of real numbers between 0 and 1. Each entry is a knot, and each knot has multiplicity at the most $p + 1$. Only the extremes of $\Xi$ have multiplicity necessarily equal to $p + 1$. For example, a general knot vector may be written as

$$\Xi = \{\underbrace{0, \ldots, 0}_{\text{p+1}}, \xi_1, \ldots, \xi_m, \underbrace{1, \ldots, 1}_{\text{p+1}}\} \tag{4}$$

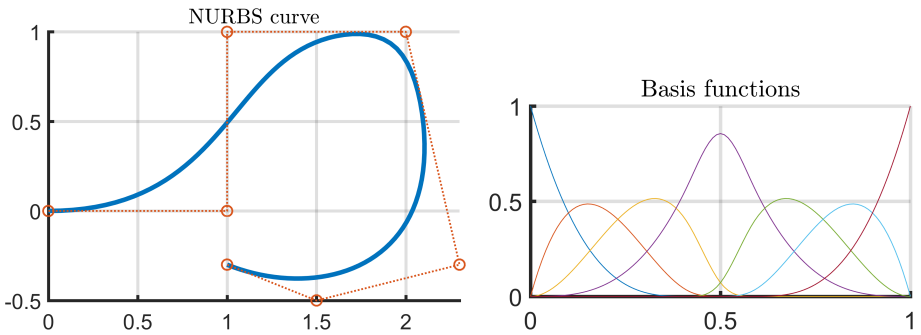
For the tests presented in this work, the set of knots $\{\xi_j\}$ is obtained by evaluating the equation below at $m = n - (p + 1)$ equally distributed points between $-a$ and $a$.

$$\xi_j = \frac{\tan t_j}{2a} + 0.5 \quad t \in [-a, a] \quad j = 1, \ldots, m \tag{5}$$

For the present study $a = 7\pi/3$. Figure 1 shows an example of NURBS curve with knot vector obtained from Eq. (5) for $m = 3$ and $w_i = 1$ for all $i = 1, \ldots, n$.

The curve $\mathbf{C}(u)$ inherits a series of properties from its basis functions. The functions $R_{i,p}(u)$ form a partition of unity ($\sum_{i=0}^{n} R_{i,p}(u) = 1 \, \forall \, u$) and have compact support ($R_{i,p} = 0$ if $u \notin [u_i, u_{i+p+1})$ ). Similar properties exist for the Lagrange polynomials used in FEA. The difference is that the rational function $R_{i,p}(u)$ are nonnegative ($R_{i,p}(u) \geq 0 \, \forall \, i, n \, \text{and} \, 0 \leq u \leq 1$), do not necessarily interpolate the Kronecker delta functions [14] and benefit from the variation diminishing property [16].

The most striking consequence of the latter property is the absence of Gibbs phenomenon for higher-order spline interpolants. This is a spurious numerical oscillation that occurs when a sharp discontinuity (e.g. the step function) is approximated by a high-order polynomial. The inefficiency of Lagrange polynomials to describe jumps and discontinuities is inevitably seen in the classic FEA. The variation diminishing property of NURBS curves establishes that a plane cannot intersect more times the curve $\mathbf{C}(u)$ than control polygon. Intuitively, this property ensures that a NURBS curve does not oscillate more than its control grid.




(a) (b)

Figure 1: (a) Cubic NURBS curve $\mathbf{C}(u)$ defined over the knot vector $\Xi = [0, 0, 0, 0, 0.3819, 0.5, 0.6181, 1, 1, 1, 1]$ illustrated with its control grid. (b) Representation of the basis functions $N_{i,p}(u)$ of $\mathbf{C}(u)$.

From the tensor product of three NURBS curves, a trivariate solid element may be obtained. This is a vector-evaluated function of three parameters. Each parametric variable may have an order $p$ and an $h$-refinement different from the other two. As a result, tensor product NURBS solids can be illustrated as a solid cube in parametric domain. By mapping the parametric cube one can, for example, obtain the uniaxial test specimen in Figure 2. For this study the parametric cube was mapped into a simple plate.

It should be noted that the variation diminishing property does not exists for NURBS surfaces and solids. For these, however, the property applies somewhat in a weaker sense. Reference is made to [18] for NURBS and to [19] for NURBS-based IGA.

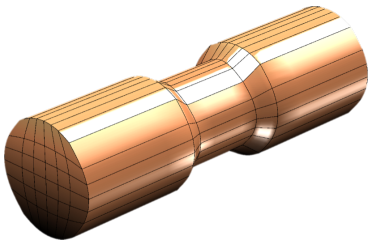


Figure 2: Uniaxial test specimen model described by tensor product solid NURBS.

## 3 NUMERICAL MODEL

### 3.1 Description

The numerical model of the ballistic test includes a target, a rounded nose projectile and four semi-spherical supports. Figure 3 illustrates the model at time $t = 0$. For this model the effect of friction is considered negligible and a rigid body assumption is made for the projectile and the supports. These assumptions will have a minor effect on the predicted physical response, but do not cancel the outcome of the comparison between IGA and FEA.

The projectile impacts the target at a velocity of 400 m/s , which is well above the ballistic limit of the plate. The target is a 7 mm thick, 127 mm wide titanium plate mounted onto four fixed supports. The mass of the projectile is equal to 67.3 g. The contact between the plate, supports and projectile uses a standard penalty master-slave formulation.

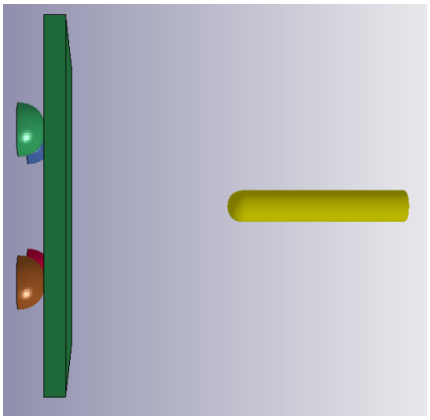


Figure 3: Perspective view of the ballistic test numerical model

### 3.2 Spatial discretization

For the discretization of the titanium plate, IGA uses a single solid NURBS patch, while FEA employs solid brick elements. The starting point to produce the NURBS patch is a quadratic trivariate model (Figure 4a). If a higher-order approximation is required, the order of this reference patch is elevated accordingly. Afterwards, the NURBS is *h*-refined by knot insertion. The knots inserted in the two parametric direction creating the face of the plate are defined by Eq. (5). The three models depicted in Figures 4b-4d (only the top-right quarter of each model is shown), are obtained for values of $m = 10, 15, 23$ respectively.

The FEA model employs Lagrange first-order polynomials to describe the titanium plate. For the numerical solution under-integrated hexahedral finite element are used. After a mesh sensitivity study, it was observed that a mesh with approximately 20000 elements would have generated convergent results in terms of projectile velocity after perforation. However, a finer mesh with about the double of number of elements was used to mitigate the hourglass effect. Detailed information about the spatial discretizations are presented in Table 1.

### 3.3 Material model

In this work we use the material model proposed by Børvik et at [11]. This constitutive model couples viscoplasticity with ductile damage and was designed for modelling impacts that exhibit penetration. Some application of this model that can be found in literature are [17, 20].

The model describes the equivalent stress using the following equation

$$\sigma = \left(A + B\bar{\varepsilon}_p^n\right)\left(1 - \frac{T - T_r}{T_m - T_r}\right)\left(1 + \frac{\dot{\bar{\varepsilon}}_p}{\dot{\varepsilon}_0}\right)^C \tag{6}$$

where $A$, $B$, $C$, $n$, $m$ are material constants. The accumulated plastic strain is $\bar{\varepsilon}_p$, the equivalent strain rate and the reference strain rate are $\dot{\bar{\varepsilon}}_p$ , $\dot{\varepsilon}_0$, respectively. The second term in the equation above considers the temperature effect; $T$ is the absolute temperature, while $T_r$ and $T_m$ are room and melting temperatures, respectively.

This model couples damage with the constitutive equation by means of following the fracture strain law:

$$\varepsilon_f = \left[D_1 + D_2 \exp\left(\frac{D_3\sigma_m}{\overline{\sigma}}\right)\right]\left[1 + \frac{\dot{\bar{\varepsilon}}_p}{\dot{\varepsilon}_0}\right]^{D_4}\left[1 + D_5\frac{T - T_r}{T_m - T_r}\right] \tag{7}$$

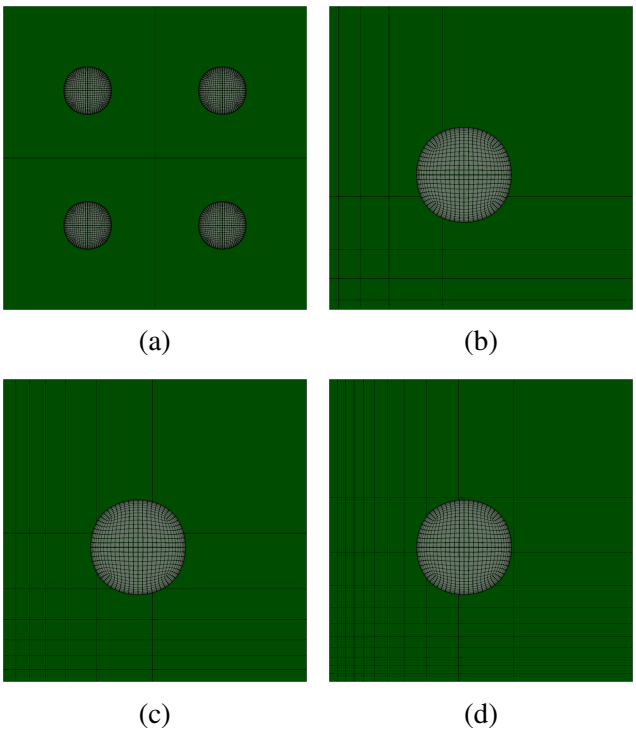


Figure 4: Spatial discretization for the IGA model. (a) Backside view with supports and NURBS patch. (b)-(d) Top-right quarters of the model after various *h*-refinements.

This is an empirical formulation for which the material constants $D_1, \ldots, D_5$ need to be calibrated experimentally. The evolution damage rule is relatively simple and given as

$$\dot{D} = \begin{cases} 0 & \text{for} \quad \overline{\varepsilon}_p \leq \varepsilon_d \\ D_c \frac{\dot{\overline{\varepsilon}}_p}{\varepsilon_f - \varepsilon_d} & \text{for} \quad \overline{\varepsilon}_p > \varepsilon_d \end{cases} \tag{8}$$

where $\varepsilon_d$ is the value of the damage threshold. The material fails as the damage parameter $D$ reaches the critical value $D_c$.

The advantages of this constitutive model are the simple calibration and its computational efficiency. An experimental campaign was conducted in our laboratory to obtain the material parameters which will be disclosed in a future publication.

### 3.4 Results

Numerical tests are carried out using the commercial software LS-DYNA. This allows to use solid NURBS element as well as the material model described in the previous section. The velocity of the projectile and the stress state of the target are the main outcomes of the numerical tests presented in this work.

Herein, we present the numerical results for an initial velocity of the projectile $v = 400$ m/s. All tests, for both IGA and FEA, predicted penetration. The results are summarised in Table 1.

Figure 5 shows the velocity of the projectile for both IGA and FEA. The comparison do not show good agreement. FEA simulation result in a much lower velocity after penetration. In this sense IGA produces somewhat conservative results. At first sight this might be attribute to volumetric locking effect. It is however interesting to observe that NURBS basis functions of order $p = 3$ (test 052) approximate the FEA solution better than those of $p = 2$ order.

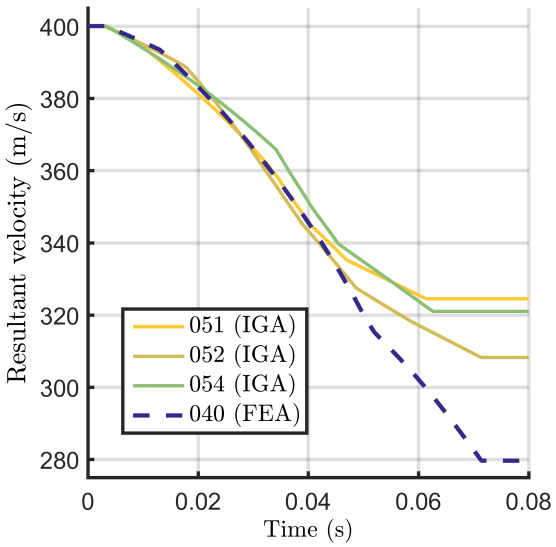
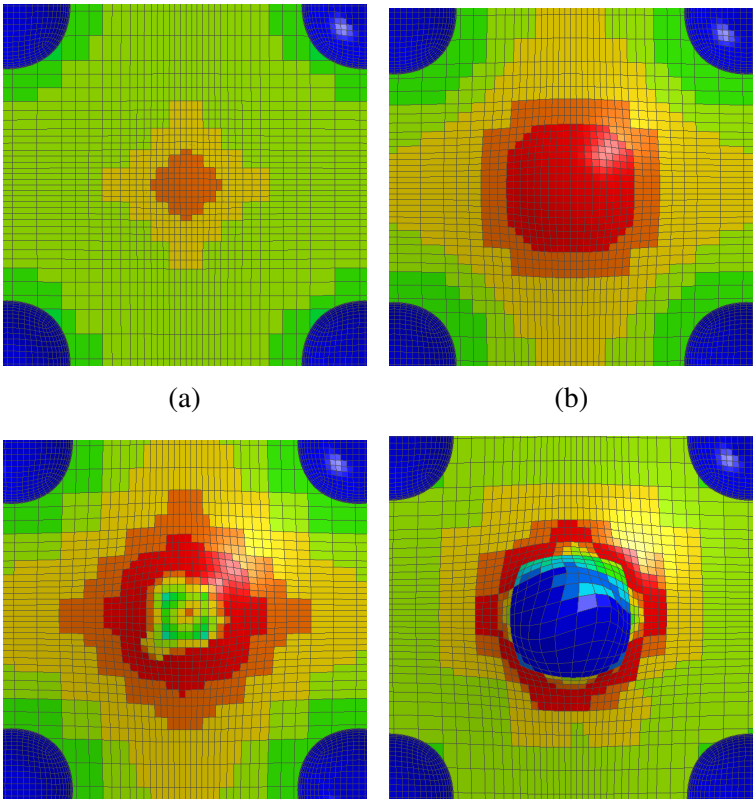



Figure 5: Comparison between second-order IGA (tests 051 and 054), third-order IGA (test 052) and FEA (test 040).

Table 1: Results for projectile initial velocity $v = 400$ m/s.

| Test ID | Method | $p$ | $m$ | *Nodes* | $v_{final}$ (m/s) | CPU$_{time}$ (h) |
|---|---|---|---|---|---|---|
| 051 | IGA | 2 | 10 | 726 | 324.5 | 23 |
| 052 | IGA | 3 | 10 | 1008 | 308.2 | 40 |
| 054 | IGA | 2 | 15 | 2048 | 321.0 | 61 |
| 040 | FEA | 1 | - | 40368 | 279.7 | 64 |

The stress state of the target plate for the IGA test 052 is illustrated in Figure 6. The backside view of four solution frames represent: supports, deformed meshes (with magnified displacement) and von Mises stress contour plots. It can be seen that, as the simulation evolves, the stress builds up (Figures 6a and 6b ). Once failure emerges, the stress in the damaged area drops (Figure 6c ) until complete penetration of the plate (Figure 6d ).



(a) (b)

(c) (d)

Figure 6: Von Mises stress contour plots from the backside view of the target at different time frames: (a) $t = 0.0009$, (b) $t = 0.0386$, (c) $t = 0.0460$ and (d) $t = 0.0800$. The red colour corresponds to stress of 800 MPa and above, while the blue colour indicates zero stress.

## 4 CONCLUSIONS

This study presents preliminary results of a ballistic test model that couples IGA with continuum damage mechanics. The numerical results are compared with classic FEA. For both methods mesh sensitivity is preformed and the constitutive model successfully predicts failure of the titanium target plate.

When penetration occurs the dynamic responses predicted by IGA and FEA diverge. Given the different nature and properties of the basis functions employed by the two methods, this difference is justifiable. An experimental validation, together with a systematic study on volume locking effects, will be addressed in future works.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of this research by the UK Technology Strategy Board and Rolls-Royce plc through the Strategic Investment in LOw-carbon Engine Technology (SILOET 2) project.

## REFERENCES

[1] Iqbal, M.A. , Senthil, K. , Bhargava, P. ,Gupta, N.K. (2015). The characterization and ballistic evaluation of mild steel, *International Journal of Impact Engineering*, 78:98-113.

[2] Ercan Sevkat, (2012). Experimental and numerical approaches for estimating ballistic limit velocities of woven composite beams, *International Journal of Impact Engineering*, 45:16-27.

[3] Hubert W. Meyer Jr., Kleponis, D. S. (2001). Modeling the high strain rate behavior of titanium undergoing ballistic impact and penetration, *International Journal of Impact Engineering*, Volume 26:509-521.

[4] Hughes, T. J. R. , Cottrell, J. A. , and Bazilevs,Y. (2005). Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement *Computer Methods in Applied Mechanics and Engineering*, 194:4135-4195.

[5] Chen E.P. (1992). Numerical simulation of shear induced plugging in HY100 steel plates. *International Journal of Damage Mechanics*, 1:132-143.

[6] Sewell, D. A., Ong, A. C. J. , Hallquist, J. (1990). Penetration calculations using an erosion algorithm in Dyna. *Proceedings of 12th International Symposium on Ballistics*, San Antonio, USA.

[7] Belytschko, T., Liu, W. K., Moran, B. , Elkhodary, K., (2013). *Nonlinear Finite Elements for Continua and StructuresJohn*, Wiley & Sons.

[8] Mos, N., Dolbow, J., Belytschko, T. (1999). A finite element method for crack growth without remeshing. *International Journal for Numerical Methods in Engineering*, 46:131-150.

[9] Barbieri, E., Petrinic, N., (2013). Three-dimensional crack propagation with distance-based discontinuous kernels in meshfree methods *Computational Mechanics*, 53:325-342.

[10] Borden, M. J., Hughes,T. J. R., Landis, C. M., Verhoosel, C. V., (2014). A higher-order phase-field model for brittle fracture: Formulation and analysis within the isogeometric analysis framework, *Computer Methods in Applied Mechanics and Engineering*, 273:100-118.

[11] Børvik T., Hopperstad, O. S., Berstad, T., Langseth, M., (2001). A computational model of viscoplasticity and ductile damage for impact and penetration, *European Journal of Mechanics - A/Solids*, 20:685-712.

[12] Johnson, G.R., Cook, W.H., (1983). A constitutive model and data for metals subjected to large strains, high strain rates and high temperatures.*Proceedings of Seventh International Symposium on Ballistics*.

[13] Lemaitre, J., Chaboche, J. L. (1990). *Mechanics of Solid Materials*. Cambridge University Press.

[14] Trefethen, Lloyd N. (2013). *Approximation theory and approximation practice*, SIAM.

[15] De Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag.

[16] Schoenberg, I.J., Greville, T.N.E., (1967). On spline functions. *Inequalities*, 1:255-291.

[17] X.G. Li, X.-L. Gao, S. Kleiven, (2016). Behind helmet blunt trauma induced by ballistic impact: A computational model, *International Journal of Impact Engineering*, 91:56-67

[18] Piegl L., Tiller, W., (1997). *The NURBS Book*. Springer.

[19] Cottrell, J. A., Hughes, T. J. R. , and Bazilevs, Y., (2009). *Isogeometric Analysis: Toward Integration of CAD and FEA*. John Wiley & Sons.

[20] Børvik, T.,Hopperstad, O. S., Dey, S., Pizzinato, E. V., Langseth, M., Albertini, C., (2005) Strength and ductility of Weldox 460 E steel at high strain rates, elevated temperatures and various stress triaxialities. *Engineering Fracture Mechanics*, 72:1071-1087.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# HYBRID RIEMANN SOLVERS FOR LARGE SYSTEMS OF CONSERVATION LAWS

**Birte Schmidtmann**[1], **Mariia Astrakhantceva**[2], **and Manuel Torrilhon**[1]

[1] MathCCES, RWTH Aachen University
Schinkelstr. 2, 52062 Aachen, Germany
e-mail: {schmidtmann, mt}@mathcces.rwth-aachen.de

[2] Lucerne School of Business
Zentralstrasse 9, Postfach 2940, 6002 Luzern, Switzerland
e-mail: mariia.astrakhantceva@hslu.ch

**Keywords:** Finite Volume Method, incomplete Riemann solvers, conservation laws, hyperbolic systems, ideal Magnetohydrodynamics.

**Abstract.** *In this paper we present a new family of approximate Riemann solvers for the numerical approximation of solutions of hyperbolic conservation laws. They are approximate, also referred to as incomplete, in the sense that the solvers avoid computing the characteristic decomposition of the flux Jacobian. Instead, they require only an estimate of the globally fastest wave speeds in both directions. Thus, this family of solvers is particularly efficient for large systems of conservation laws, i.e. with many different propagation speeds, and when no explicit expression for the eigensystem is available. Even though only fastest wave speeds are needed as input values, the new family of Riemann solvers reproduces all waves with less dissipation than HLL, which has the same prerequisites, requiring only one additional flux evaluation.*

## 1 INTRODUCTION

In the finite volume method, integrating conservation laws over a control volume leads to a formulation which requires the evaluation of the flux function at cell interfaces. Because the exact information is not available, a sequence of local Riemann problems needs to be solved [4]. The initial states for these problems are typically given by the left and right adjacent cell values, $(u_i^n, u_{i+1}^n)$ for each $i$ [8]. These local Riemann problems have to be solved many times for finding the numerical solution, therefore, the Riemann solver is a building block of the finite volume method. Note that the numerical flux function at interfaces also appears in the Discontinuous Galerkin (DG) formulation. Therefore, Riemann solvers are also needed in DG.

Over the last decades, many different Riemann solvers have been developed, see e.g. [8]. The challenge is that the solver needs to be computationally efficient and easy to implement. At the same time, it needs to yield accurate results which do not create artificial oscillations. The latter is ensured by requiring the solver to be monotone.

Riemann solvers can be classified into complete and incomplete schemes, depending on whether all present characteristic fields are considered in the model or not. According to this classification, the upwind scheme and Roe's scheme [6], respectively, are complete schemes. They yield good, monotone results, however, an evaluation of the eigensystem of the flux Jacobian is needed. This characteristic decomposition might be expensive to compute, especially for large systems, and in some cases, an analytic expression is not available at all. However, if possible, using Roe's scheme typically yields the best resolution of the Riemann wave fan, since all waves are well-resolved. Therefore, this scheme can be considered as the optimum, also taking into account that less dissipation than upwinding leads to non-monotone solutions. In order to keep its high resolution and at the same time reducing the computational cost, there have been many attempts to approximate the upwind scheme without solving the eigenvalue problem, see e.g. [9, 2] and references therein.

In this article, we are interested in approximate Riemann solvers. Their advantage is the easy implementation and the requirement of few characteristic information. However, incomplete Riemann solvers contain more dissipation than the upwind scheme and thus, yield lower resolution, especially of slow waves. Nevertheless, in many test cases, these Riemann solvers may be sufficient to obtain good results, especially if the system contains only fast waves.

## 2 FINITE VOLUME METHOD

We consider a hyperbolic system of conservation laws in one space dimension of the form

$$\partial_t U(x,t) + \partial_x f(U(x,t)) = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}^+ \tag{1}$$

$$U(x,0) = U_0(x) \tag{2}$$

where the unknown variable vector is given by $U : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^N$, with some initial conditions $U_0(x)$ and a flux function $f : \mathbb{R}^N \to \mathbb{R}^N$, such that the Jacobian $A(U) = Df(U)$ only has real eigenvalues. In finite volume methods, the domain is subdivided into cells $C_i = [x_{i-1/2}, x_{i+1/2}]$ and Equation (1) is integrated over a cell $C_i$. The cell average at time $t^n$, given by

$$\frac{1}{\Delta x} \int_{C_i} U(x, t^n) dx, \tag{3}$$

is approximated by $\bar{U}_i^n$. The update of the approximation at time $t^{n+1} = t^n + \Delta t$ reads [5]

$$\bar{U}_i^{n+1} = \bar{U}_i^n - \frac{\Delta t}{\Delta x} \left( \hat{f}_{i+1/2} - \hat{f}_{i-1/2} \right) \tag{4}$$

and contains the numerical flux function $\hat{f}_{i+1/2} = \hat{f}(U^-_{i+1/2}, U^+_{i+1/2})$. This function takes as input values the left and right limiting values of the variable vector $U$ at the cell interface $i + 1/2$.

The choice of the numerical flux function - or Riemann solver - of the method, determines the mathematical properties of the scheme, such as accuracy and monotonicity. Thus, the choice of the solver is crucial for the resulting scheme.

In this work, we consider the general formulation of the numerical flux, given by

$$\hat{f}(U_L, U_R) = \frac{1}{2}\left(f(U_L) + f(U_R)\right) + \frac{1}{2} D(U_L, U_R)\left(U_L - U_R\right) \tag{5}$$

with a dissipation matrix $D$ which depends on the left and right states $U_L$ and $U_R$, respectively. $D$ also depends on characteristics of the flux Jacobian $A(U)$.

Let us denote by $\lambda_{\text{min}}(U)$ and $\lambda_{\text{max}}(U)$ the minimal and maximal eigenvalue of $A(U)$. The spectral radius at $U$, i.e. the maximum absolute characteristic speed, is given by $\bar{\lambda} = \max\{|\lambda_{\text{min}}|, |\lambda_{\text{max}}|\}$.

## 3 REVIEW OF RIEMANN SOLVERS

In this section we recall some well-known Riemann solvers which are necessary for the development of the new family of hybrid Riemann Solvers. First, we note that the dissipation matrix $D$ completely dictates the numerical flux function (5) and hence the numerical scheme. Therefore, we break down the following discussion to comparing the dissipation matrices of the schemes. Furthermore, for the sake of comparability, we introduce the notation of the dimensionless scalar dissipation function $d(\nu)$:

The flux Jacobian $A(U)$ can be diagonalized as $A(U) = T(U)\,\Lambda(U)\,T(U)^{-1}$, where $\Lambda(U)$ is the eigenvalue matrix $\Lambda(U) = \text{diag}(\lambda_1(U), \ldots, \lambda_N(U))$, with $\lambda_1 < \lambda_2 < \ldots < \lambda_N$, and $T(U)$ the corresponding eigenvector matrix. Since the dissipation matrix $D$ is a function of the flux Jacobian $A$, it can be shown that

$$\frac{\Delta t}{\Delta x} D(A) = T^{-1}\,\frac{\Delta t}{\Delta x}\,\text{diag}(\tilde{d}(\lambda_1), \ldots, \tilde{d}(\lambda_N))\,T = T^{-1}\,\text{diag}(d(\nu_1), \ldots, d(\nu_N))\,T, \tag{6}$$

holds. Here, $\nu_i = \lambda_i \Delta t/\Delta x$ and $d(\nu)$ is the dimensionless scalar dissipation function. The eigenvector matrix $T$ is the same, independent of $D$.

We will compare the dissipation functions of different schemes in a $\nu - d(\nu)$-plot at the end of this section.

For a linear system, where we can write $f(U) = A\,U$ with some constant matrix $A \in \mathbb{R}^{N\times N}$, the dissipation matrix of the upwind scheme reads

$$D_{\text{up}} = |A| \quad \leftrightarrow \quad d_{\text{up}}(\nu) = |\nu|. \tag{7}$$

For non-linear systems with general flux function, the scheme has been extended by Roe [6]. The dissipation matrix of Roe's scheme is given by

$$D_{\text{Roe}} = |\tilde{A}| \quad \leftrightarrow \quad d_{\text{Roe}}(\nu) = |\nu|, \tag{8}$$

with a so-called Roe Matrix $\tilde{A}$, satisfying $\tilde{A}(U_L, U_R)(U_L - U_R) = F(U_L) - F(U_R)$. The upwind Godunov solver and its non-linear extension, the Roe solver are complete Riemann solvers. Now follows a list of incomplete solvers with decreasing dissipation.

The dissipation function of the monotone Lax-Friedrichs scheme is

$$D_{\text{LF}} = \frac{\Delta x}{\Delta t} I \quad \leftrightarrow \quad d_{\text{LF}}(\nu) = 1. \tag{9}$$

A solver which decreases the dissipation is the Rusanov scheme, also referred to as local Lax-Friedrichs scheme. It takes into account the globally fastest eigenvalue of the system:

$$D_{\text{LLF}} = \max\{\bar{\lambda}(U_L), \bar{\lambda}(U_R)\} I \quad \leftrightarrow \quad d_{\text{LLF}}(\nu) = \max(|\nu_{\min}|, |\nu_{\max}|). \tag{10}$$

Harten, Lax and van Leer [3] further decreased the dissipation, especially of slow waves, by considering the fastest and slowest waves of the system:

$$D_{\text{HLL}} = \frac{|\lambda_L| - |\lambda_R|}{\lambda_L - \lambda_R} \tilde{A} - \frac{|\lambda_L|\,\lambda_R - |\lambda_R|\,\lambda_L}{\lambda_L - \lambda_R} I \tag{11a}$$

$$\leftrightarrow d_{\text{HLL}} = \frac{|\nu_L| - |\nu_R|}{\nu_L - \nu_R} \nu - \frac{|\nu_L|\,\nu_R - |\nu_R|\,\nu_L}{\nu_L - \nu_R}. \tag{11b}$$

Here, $\lambda_L = \lambda_{\min}(U_L)$, $\lambda_R = \lambda_{\max}(U_R)$, and $\nu_{L/R} = \lambda_{L/R}\Delta t/\Delta x$.

A scheme which further reduces the dissipation is the Lax-Wendroff scheme,

$$D_{\text{LW}} = \frac{\Delta t}{\Delta x} \tilde{A}^2 \quad \leftrightarrow \quad d_{\text{LW}}(\nu) = \nu^2. \tag{12}$$

Discontinuities are approximated with steep gradients using the Lax-Wendroff scheme, however, the method is known to cause oscillations at discontinuities because it is non-monotone [4], see also Fig. 1.

## 4 A FAMILY OF NEW HYBRID RIEMANN SOLVERS

We want to construct Riemann solvers which require as few information as HLL but are less dissipative. This is advantageous for the resolution, especially for slow waves, i.e. $\lambda \approx 0$. Demanding the same input values as HLL, we require the knowledge or an estimate of the globally slowest and fastest characteristic waves of the system. Only one additional flux evaluation shall be required.

### 4.1 $P_2$ - conditions

We now pick up three requirements which have been proposed by Degond et al. [1]. The resulting monotone Riemann solver, named $P_2$, is based on a quadratic dissipation function, fully determined by

$$d_{P_2}(\nu_{\min}) = d_{\text{up}}(\nu_{\min}) = |\nu_{\min}|, \tag{13a}$$

$$d_{P_2}(\nu_{\max}) = d_{\text{up}}(\nu_{\max}) = |\nu_{\max}|, \text{ and} \tag{13b}$$

$$d'_{P_2}(\bar{\nu}) = d'_{\text{up}}(\bar{\nu}) = \text{sign}(\bar{\nu}), \quad \bar{\nu} = \begin{cases} \nu_{\max}, & \text{if } |\nu_{\max}| \geq |\nu_{\min}| \\ \nu_{\min}, & \text{if } |\nu_{\min}| > |\nu_{\max}|. \end{cases} \tag{13c}$$

This function automatically fulfills $d_{P_2}(\nu) \geq |\nu|$ for $\nu \in [\nu_{\min}, \nu_{\max}]$ which means that it is monotone in this region. This can also be seen in Fig. 1, which additionally shows that $P_2$ is less dissipative than HLL, especially for $\nu \approx 0$. The dissipation function $d_{P_2}(\nu)$ can be written in the simple form

$$d_{P_2}(\nu) = d_{\text{HLL}}(\nu) + \alpha(\nu - \nu_{\min})(\nu - \nu_{\max}), \tag{14}$$

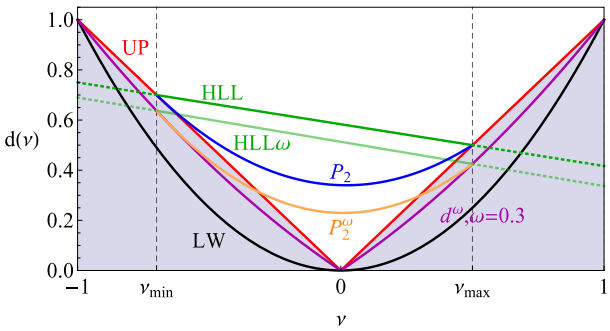



Figure 1: Scalar, non-dimensional dissipation functions of different Riemann solvers.

with

$$\alpha = \frac{\nu_{\max} - \nu_{\min} - \big||\nu_{\max}| - |\nu_{\min}|\big|}{(\nu_{\max} - \nu_{\min})^2}. \tag{15}$$

### 4.2 $P_2^\omega$ - Beyond Monotonicity

The main idea of $P_2^\omega$ is to construct a quadratic dissipation function, in the form of $P_2$. This new function shall be closer to the absolute value function, i.e. the upwind scheme, for all waves $\lambda_i$ (and thus for all $\nu_i$) of the hyperbolic system. Since we do not want to increase neither the number of input information, nor the number of flux evaluations, we lower the dissipation function by a certain amount. This amount is described by a parameter $\omega \in [0, 1]$, which determines the monotonicity behavior of the solver. For $\omega = 0$ we recover the monotone $P_2$ solver, and for $\omega = 1$, the non-monotone Lax-Wendroff solver. All intermediate members of the $P_2^\omega$ family are slightly non-monotone for a certain range of waves. However, we show in this section that under some mild assumptions, the results do not show spurious oscillations.

#### 4.2.1 Monotonicity Study

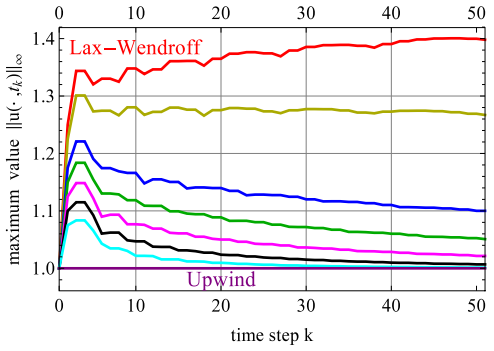
Before we introduce this family of Riemann solvers, we state and validate some observations made by Torrilhon [9]. Firstly, it was perceived that the MUSTA fluxes introduced by Toro [7] slightly drop below the upwind flux, which means that they do not fully lie in the monotonicity preserving region. Thus, as expected, the numerical solutions obtained with MUSTA fluxes show some non-monotone behavior. However, this behavior is far from the oscillations created by the Lax-Wendroff scheme. Additionally, the observed oscillations of MUSTA solutions decay in time and disappear after a certain number of time steps, cf. [9, Fig. 5, p. A2084]. These results are essentially independent of the grid size.

These interesting results were observed for a dissipation function which slightly drops below the absolute value function. Let us introduce the dissipation function $d_\omega(\nu)$, which is the weighted average of the dissipation functions of the monotone upwind scheme and the non-monotone Lax-Wendroff scheme,

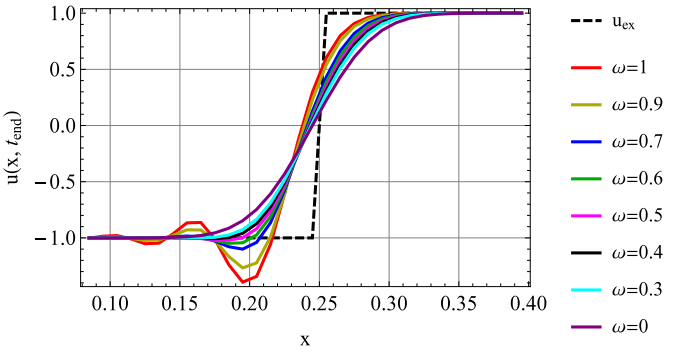
$$d_\omega(\nu, \omega) = \omega\, d_{\text{LW}}(\nu) + (1 - \omega)\, d_{\text{up}}(\nu) \quad \omega \in [0, 1]. \tag{16}$$

For $\omega = 0$ we recover the monotone upwind scheme $d_{\omega=0}(\nu) = d_{UP}(\nu)$ and for $\omega = 1, d_{\omega=1}(\nu) = d_{LW}(\nu)$ holds true. Fig. 1 shows $d_\omega(\nu, \omega)$ for $\omega = 0.3$.

The aim of this section is to study the monotonicity behavior of $d_\omega(\nu)$ and produce similar effects as studied in [9]. Let us therefore investigate the solutions of the numerical flux function




(a) Maximum value of the solution $u$ as a function of the number of time steps.




(b) Zoom of the solution for different $\omega$.

Figure 2: Test case with initial condition $\text{sign}(x)$ on $x \in [-1, 1]$ with $n = 200$ grid cells, $CFL = 0.5$, and end-time $T_{\text{end}} = 0.25$, which corresponds to 50 time steps.

Eq. (5) with Eq. (16) for different values of $\omega$. The test problem is the scalar transport equation $u_t + u_x = 0$ with initial condition $u_0(x) = \text{sign}(x)$ on the interval $[-1, 1]$. The jump evolves with time on a grid with $n = 200$ grid cells until $T_{\text{end}} = 0.25$. The Courant number is set to $CFL = 0.5$, which shows the maximal deviation of Lax-Wendroff from Upwind.

The numerical results for all tested values of $\omega$ are shown in Fig. 2b. It can be easily seen, that $\omega = 0$ and $\omega = 1$ correspond to the upwind and the Lax-Wendroff schemes. That is, for $\omega = 1$ we can observe the well-known oscillations. As $\omega$ decreases, the oscillations also decrease and for $\omega \leq 0.4$ no oscillations can be seen anymore. This can also be seen in Fig. 2a, which shows the maximum value of the solutions $u$ depending on the number of time steps. Here, 50 time steps correspond to $T_{\text{end}} = 0.25$. This figure shows that the oscillations, which appear in the Lax-Wendroff scheme do not decrease over time. However, as soon as the monotone upwind scheme has enough weight, the oscillations start decreasing over time. This phenomenon can be explained by looking at the modified equations [4] of upwind, Lax-Wendroff and the their weighted average. In the modified equation of the latter, the diffusive upwind term remains the dominant term. This avoids the creation of oscillations caused by the Lax-Wendroff dispersion term, when this term has enough weight. More details on the modified equations will be presented in future work.

$d_\omega(\nu, \omega)$ could also be used as a limiter function, increasing $\omega$ in smooth parts of the solution and setting it to 0 (i.e. recovering the upwind solver) when discontinuities are encountered.

### 4.2.2 HLL$\omega$

Based on the findings above, let us define a Riemann solver which is a modification of HLL with less dissipation. We shall call this solver HLL$\omega$. In the same way HLL is constructed [3, 8], the dissipation function of this solver is of the form $d_{\text{HLL}\omega}(\nu) = b_0 + b_1\nu$. Here, the coefficients depend on $\omega$, i.e. $b_0 = b_0(\omega), b_1 = b_1(\omega)$. Additionally, instead of intersecting with the absolute value function at $\nu_{\text{min}}$ and $\nu_{\text{max}}$, HLL$\omega$ fulfills the following constraints:

$$d_{\text{HLL}\omega}(\nu_{\text{min}}) = d_\omega(\nu_{\text{min}}), \quad d_{\text{HLL}\omega}(\nu_{\text{max}}) = d_\omega(\nu_{\text{max}}). \tag{17}$$

The dissipation function is shown in Fig. 1, where it is well-visible, that HLL$\omega$ is less dissipative than HLL and is non-monotone for some wave speeds.

### 4.2.3 $P_2^\omega$

Now we can come back to the aim of this section, the construction of a new family of approximate Riemann solvers - $P_2^\omega$. The dissipation functions of these solvers are similar to $d_{P_2}$, only closer to the absolute value function for all emerging wave speeds of the system. The dissipation functions are given by

$$d_{P_2^\omega}(\nu) = d_{\text{HLL}\omega}(\nu) + \beta(\nu - \nu_{\text{min}})(\nu - \nu_{\text{max}}), \tag{18}$$

with

$$\beta = \omega + (1-\omega)\frac{\nu_{\text{max}} - \nu_{\text{min}} - \big||\nu_{\text{max}}| - |\nu_{\text{min}}|\big|}{(\nu_{\text{max}} - \nu_{\text{min}})^2}. \tag{19}$$

Note that $\beta = \omega + (1-\omega)\alpha$, with the $P_2$ coefficient $\alpha$ (15). Thus, it is easy to verify that for $\omega = 0$, the monotone $P_2$ solver is recovered.

It can be seen in Fig. 1 that $P_2^\omega$ is less dissipative than HLL and $P_2$. However, it does not fully lie in the monotonicity preserving region, thus, one would expect some non-monotone behavior. Nevertheless, we observe that oscillations appearing close to discontinuities disappear after a certain number of time steps. Thus, the final result obtained with $P_2^\omega$ is non-oscillatory.

The choice of $\omega$ remains problem-dependent. However, $\omega \leq 0.5$ turned out to be a good choice and will be used in the following.

## 5 NUMERICAL RESULTS - APPLICATION TO IDEAL MHD

We will apply the new family of Riemann solvers to the equations of ideal magnetohydrodynamics (MHD) and compare it to other solvers. Ideal MHD describes the flow of plasma, assuming infinite electrical resistivity. The equations of ideal MHD in one-dimensional processes read

$$\partial_t \begin{pmatrix} \rho \\ \rho v_x \\ \rho \mathbf{v_t} \\ \mathbf{B_t} \\ E \end{pmatrix} + \partial_x \begin{pmatrix} \rho v_x \\ \rho v_x^2 + p + \frac{1}{2}\mathbf{B_t^2} \\ \rho v_x \mathbf{v_t} - B_x \mathbf{B_t} \\ v_x \mathbf{B_t} - B_x \mathbf{v_t} \\ (E + p + \frac{1}{2}\mathbf{B_t^2}) v_x - B_x \mathbf{B_t} \cdot \mathbf{v_t} \end{pmatrix} = 0 \tag{20}$$

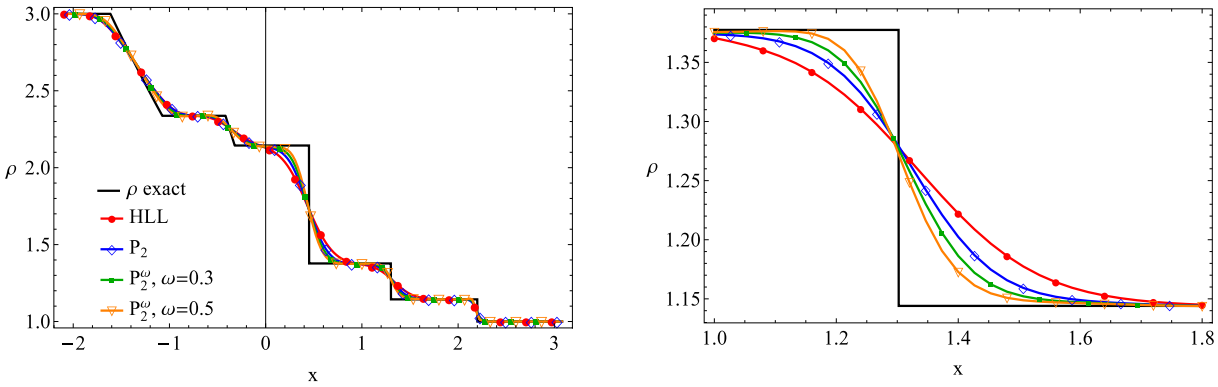
with density $\rho$, normal and tangential velocities $v_x$, and $\mathbf{v_t} = (v_y, v_z)$, respectively. The normal magnetic field $B_x$ is constant in the one-dimensional case, the tangential magnetic field is $\mathbf{B_t} = (B_y, B_z)$ and the energy $E$ is given by

$$E = \frac{1}{\gamma - 1} p + \frac{1}{2}\rho(v_x^2 + \mathbf{v_t^2}) + \frac{1}{2}\mathbf{B_t^2} \tag{21}$$

in terms of the pressure $p$. The adiabatic constant $\gamma$ is set to $5/3$. Since the normal magnetic field $B_x$ is constant, system (20) contains seven equations for the seven unknowns, $\mathbf{U} = (\rho, v_x, \mathbf{v_t}, p, \mathbf{B_t})$, exhibiting seven characteristic velocities, and therefore can be considered as a large system of conservation laws.

Let us consider the Riemann problem given by

$$\mathbf{U}_0(x) = \begin{cases} \mathbf{U}_L = (3, 0, (0,0), 3, (1,1)) & \text{if } x < 0 \\ \mathbf{U}_R = (1, 0, (0,0), 1, (\cos(1.5), \sin(1.5))) & \text{if } x \geq 0, \end{cases} \tag{22}$$




(a) Density profile for different solvers.

(b) Zoom of the slow shock around $x = 1.3$.

Figure 3: Solution of the ideal MHD equations (20) with initial conditions (22) in the domain $x \in [-4, 4]$ with $n = 300$ grid cells, $CFL = 0.9$, $\Delta t = 0.01$, and end time $T_{\text{end}} = 1.0$.

and $B_x = 1.5$. The computational domain is $[-4, 4]$, and the solutions depicted in Fig. 3 have been obtained with $N = 300$ grid cells, CFL number $0.9$, $\Delta t = 0.01$, and end time $T_{\text{end}} = 1.0$. Fig. 3a shows the solution of the density profile obtained with HLL, $P_2$ and $P_2^\omega$, $\omega = 0.3, 0.5$. The exact solution has been obtained by [10]. Fig. 3b shows a zoom of the slow right moving shock at $x = 1.3$, where the resolution of all solvers can be nicely compared. It can be stated that $P_2$ increases the resolution compared to HLL, causing a steeper gradient. $P_2^\omega$ further increases the steepness of the gradient for increasing $\omega$, due to decreasing dissipation. This effect is present in an even stronger form at the contact discontinuity, which corresponds to a slower wave. At the fast shock, the differences between the solutions of the four solvers are less significant, because this discontinuity relates to a larger wave speed $\lambda$. This observation corresponds well to Fig. 1, where we showed that the differences of the discussed dissipation functions are larger for slower waves.

## 6 CONCLUSIONS

This paper presented a family of approximate hybrid Riemann solvers, $P_2^\omega$, for non-linear hyperbolic systems of conservation laws. The solvers do not require the characteristic decomposition of the flux Jacobian, only an estimate of the maximal propagation speeds in both directions is needed. The family of solvers contains a parameter $\omega$ which orders the solvers from fully-monotone to fully non-monotone. The intermediate solvers contain monotone as well as non-monotone parts. We showed that these intermediate family members, even though containing non-monotone parts for certain wave speeds, do not lead to oscillatory solutions.

Extremely slow waves and stationary waves will still be approximated with higher dissipation than the upwind scheme, however, the computational cost of the new solvers is lower. Compared to solvers with similar prerequisites, the new Riemann solvers are able to rigorously decrease the dissipation of the scheme.

The numerical examples underline the excellent performance of the new family of solvers with respect to other solvers.

## REFERENCES

[1] P. Degond, P.F. Peyrard, G. Russo, P. Villedieu, Polynomial upwind schemes for hyperbolic systems. *Comptes Rendus de l'Académie des Sciences-Série I-Mathématiques*, **328(6)**, 479–483, 1999.

[2] M.J. Castro, J.M. Gallardo, A. Marquina, A class of incomplete Riemann solvers based on uniform rational approximations to the absolute value function. *J. Sci. Comput.*, **60(2)**, 363–389, 2014.

[3] A. Harten, P.D. Lax, B. van Leer, On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, **25(1)**, 35–61, 1983.

[4] R.J. LeVeque, *Numerical methods for conservation laws*. Basel: Birkhäuser, 1992.

[5] R.J. LeVeque, *Finite volume methods for hyperbolic problems*. Cambridge University press, 2002.

[6] P.L. Roe, Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, **43(2)**, 357–372, 1981.

[7] E.F. Toro, MUSTA: A multi-stage numerical flux. *Appl. Numer. Math.*, **56(10)**, 1464–1479, 2006.

[8] E.F. Toro, *Riemann solvers and numerical methods for fluid dynamics: a practical introduction*. Springer, 2013.

[9] M. Torrilhon, Krylov–Riemann Solver for Large Hyperbolic Systems of Conservation Laws. *SIAM J. Sci.Comput.*, **34(4)**, A2072–A2091, 2012.

[10] M. Torrilhon, https://web.mathcces.rwth-aachen.de/mhdsolver/

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE USE OF ENRICHED BASE FUNCTIONS IN THE THREE-DIMENSIONAL SCALED BOUNDARY FINITE ELEMENT METHOD

**Sascha Hell**[1] **and Wilfried Becker**[1]

[1]Institute of Structural Mechanics, Technical University of Darmstadt
Franziska-Braun-Straße 7, 64287 Darmstadt, Germany
e-mail: {hell,becker}@fsm.tu-darmstadt.de

**Keywords:** Scaled Boundary Finite Element Method, enriched base functions, three-dimensional linear elasticity, quadratic eigenvalue problem

**Abstract.** *The Scaled Boundary Finite Element Method (SBFEM) is a semi-analytical method which combines the advantages of the Boundary Element Method (BEM) and the Finite Element Method (FEM). Like in the BEM, only the boundary needs to be discretized. On the other hand, the SBFEM is based on the virtual work principle and does not need any fundamental solutions. If the scalability condition is fulfilled, a separation of variables representation can be employed leading to a quadratic eigenvalue problem and a linear equation system which can be solved by standard methods. The SBFEM has proven its high efficiency and accuracy in the presence of stress singularities, especially in 2D fracture mechanics when the singularity is entirely located within the considered domain.*

*However, in 3D elasticity problems, there can also be singularities on the discretized boundary itself. Then, the SBFEM suffers from drawbacks also well known from the standard FEM, i.e. moderate accuracy and bad convergence. To overcome these deficiencies in such 3D cases, we propose the enrichment of the standard separation of variables representation with analytical fields which are known to exactly fulfill the local boundary conditions:*

$$\boldsymbol{u} = \underbrace{\mathbf{N}(\eta_1, \eta_2)\, \boldsymbol{u}(\xi)}_{\text{standard}} + \underbrace{\mathbf{F}(r, \varphi)\, \boldsymbol{a}(\xi)}_{\text{enrichment}}. \tag{1}$$

*The examples of a single plane crack and two perpendicularly meeting cracks in an isotropic continuum are considered. It is demonstrated that the method's original excellent accuracy and convergence are regained, at a minimum cost of additional degrees of freedom (DOF). The normalized errors in the solution of the quadratic eigenvalue problem already become negligibly small for very coarse boundary meshes. The obtained convergence orders are often optimal and sometimes even superconvergence is observed.*

## 1 INTRODUCTION

This work is concerned with the efficient numerical solution of three-dimensional boundary value problems of linear elasticity which involve stress singularities. Stress singularities in linear elasticity theory involve infinite stresses and strains at certain points or lines which typically represent discontinuities of geometry (e.g. cracks and notches), material (e.g. multi-material junctions) or loading. Although such singularities cannot actually occur in a real material as they are locally reduced by non-linear deformations, in brittle materials, their near-field still is a good approximation to the real stress and displacement fields (e.g. already in 1927 their occurrence was considered by Knein [19] because he found his pressure test blocks to unexpectedly often fail at the corners instead of their center). Usually, as e.g. by Williams [38], the near-field solution at a singular point for the displacements and stresses respectively is represented by a power law function series of the kind

$$\underline{u}(r,\varphi,\vartheta) = \sum_{j=1}^{\infty} c_{uj} r^{\lambda_j} \underline{\phi}_{uj}(\varphi,\vartheta) \qquad , \qquad \underline{\underline{\sigma}}(r,\varphi,\vartheta) = \sum_{j=1}^{\infty} c_{\sigma j} r^{\lambda_j - 1} \underline{\underline{\phi}}_{\sigma j}(\varphi,\vartheta) \tag{2}$$

given in spherical coordinates $r, \varphi, \vartheta$. Here, $\lambda_j$ are complex exponents, $\underline{\phi}_{uj}$ are vector and $\underline{\underline{\Phi}}_{\sigma j}$ are tensor functions of the angle coordinates $\varphi$ and $\vartheta$. Line singularities can be similarly represented in e.g. cylindrical coordinates. The free constants $c_{uj}$ respectively $c_{\sigma j}$ are determined by conformance to the boundary conditions.

It is considerably more difficult to find solutions for three-dimensional boundary value problems containing singularities than for two-dimensional ones. This resulted in the development and application of appropriate numerical (e.g. FEM [44]) but also semi-analytical methods (e.g. [6, 28, 4]). The probably most employed semi-analytical method for such elasticity problems is the FEM eigenanalysis, which has been used by a number of researchers for several 2D (e.g. [40, 42, 15, 30, 34, 43]) and 3D applications (e.g. [5, 32, 24, 13, 31, 9, 20, 26, 18]). The FEM eigenanalysis is based on the approximation of the displacements by a separation of variables representation with FE-shape functions in the circumferential direction, i.e. only the boundary needs to be discretized. This representation is substituted in the Galerkin method generally leading to a Cauchy-Euler differential equation system of second order. This, in turn, can be transformed into a quadratic eigenvalue problem. The solution of the quadratic eigenvalue problem yields approximations for the quantities $\lambda_j$ and $\underline{\phi}_{uj}$ in eq. (2) but not for the free constants $c_{uj}$, which would be necessary to solve the complete boundary value problem. Generally, the solution of the complete quadratic eigenvalue problem is computationally expensive which practically was the main reason that the method was hardly used for the solution of complete boundary value problems.

The Scaled Boundary Finite Element Method (SBFEM) [34, 8, 39] is a variant of the FEM eigenanalysis technique. To the authors' knowledge, it was the first used to solve complete boundary value problems, but only for 2D situations. Its striking advantages in the treatment of 2D boundary value problems with stress singularities were consequently exploited (e.g. [33]) up to the development of codes which automatically predict crack growth ([41, 29, 7]). For 3D situations, it has lately been used e.g. by the current authors in [16].

A major challenge of the SBFEM in 3D situations containing singularities is that its accuracy and convergence properties suffer when line singularities reach the boundary mesh, meaning a loss of the method's initial advantages in the treatment of boundary value problems with singularities. Indeed, this is a phenomenon which also is well-known in the standard FEM (e.g. [35, 44]) and the usual remedy is a graded mesh with a strong refinement towards the

singularity. This approach has also been used for the FEM eigenanalysis (e.g. [2]) but still led to linearized eigenvalue problems so large that a complete solution with the QR-algorithm was avoided.

The QR-algorithm is computationally very expensive as its computational effort grows cubically with the number of degrees of freedom (DOF). But in return, it gives the complete set of eigenvalues and eigenvectors needed for the solution of the complete boundary value problem. Other solution methods for the quadratic eigenvalue problem, which have been implemented (e.g. [9]), are a lot more efficient but only partly solve the eigenvalue problem, i.e. they only give a small set of eigenpairs or only give the eigenvalues but not the eigenvectors. Thus, they can compensate the reduced convergence and accuracy with a strong refinement of the boundary mesh around the singularity but cannot be used to solve the complete boundary value problem. Consequently, an approach making it possible to still solve the complete eigenvalue problem while increasing the accuracy of the eigenpairs at a minimum cost of additional DOF is sought-after.

To the authors' knowledge, Lee and Im [23] have been the only ones engaged in this venture so far. They used simple quarter-point element formulations [17] to resolve a classical crack singularity on the boundary mesh. As this approach is limited to the classical crack singularity order, we propose a more general approach based on an enrichment of the separation of variables approach for the displacement field. In the past 20 years, such enrichment approaches have been intensely studied in the framework of the XFEM introduced by Belytschko and coworkers [10, 12]. From this stock of knowledge, we selected the DOF-efficient plateau method [22] in conjunction with blending element formulations [11, 37] and Fleming's decomposition of the crack tip near-field [10] for our implementation.

In section 2, the prerequisites of the newly developed method are presented. In section 3, its advantages regarding accuracy and convergence are demonstrated for the generic examples of a simple crack [1, 25] and two meeting cracks in a homogeneous isotropic cube [16]. Section 4 contains the major conclusions of our analyses.

## 2 THE SCALED BOUNDARY FINITE ELEMENT METHOD WITH ENRICHED DISPLACEMENT REPRESENTATIONS

Our three-dimensional SBFEM formulation (cf. [39], [16]) with an enriched displacement field is based on the virtual work principle

$$\delta W_i = \int_V \underline{\underline{\sigma}} : \delta\underline{\underline{\varepsilon}}\,\mathrm{d}V = \int_V \underline{f}^* \cdot \delta\underline{u}\,\mathrm{d}V + \int_{S_t} \underline{t}^* \cdot \delta\underline{u}\,\mathrm{d}A = \delta W_a \tag{3}$$

in which $\underline{\underline{\sigma}}, \delta\underline{\underline{\varepsilon}}, \underline{f}^*, \delta\underline{u}, \underline{t}^*$ are the stress tensor, the virtual strain tensor, body forces, virtual displacements and prescribed boundary tractions respectively. $\delta W_i$ is the internal and $\delta W_a$ the external virtual work. $V$ is the volume of the considered domain and $S_t$ the part of the domain's boundary with prescribed traction boundary conditions. We write stresses and strains in vector notation $\boldsymbol{\sigma}$ and $\boldsymbol{\varepsilon}$ and use a differential operator $\mathbf{L}$

$$\mathbf{L}^{\mathrm{T}} = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 & 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ 0 & \frac{\partial}{\partial y} & 0 & \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \\ 0 & 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \end{bmatrix} \tag{4}$$

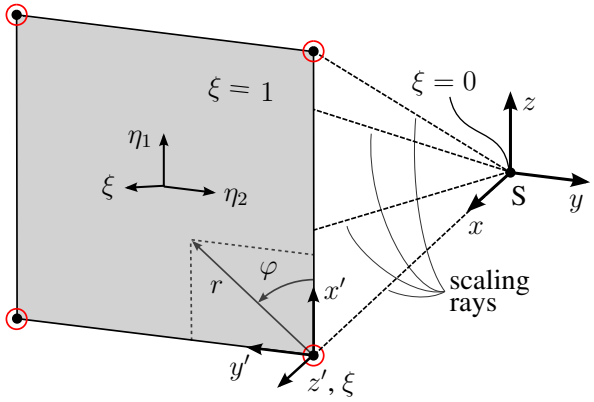



Figure 1: Plane, 4-node scaled boundary finite element forming a square-based pyramid (scaling center $S$ with $\xi = 0$ at origin of a global cartesian coordinate system $x, y, z$). An isoparametric scaled boundary coordinate system $\xi, \eta_1, \eta_2$ and a conical coordinate system $\xi, r, \varphi$ are present, the latter such that $r = 0$ at the crack front (singularity line) and $\varphi = \pm\pi$ at the crack faces. Enriched nodes are circled in red.

such that the equilibrium equations and strain-displacement relations can be written in matrix form

$$\sigma_{ij,j} + f_j = 0 \qquad \rightarrow \qquad \mathbf{L}^{\mathrm{T}}\boldsymbol{\sigma} + \boldsymbol{f} = \mathbf{0} \tag{5}$$

$$\varepsilon_{ij} = (u_{i,j} + u_{j,i})/2 \qquad \rightarrow \qquad \boldsymbol{\varepsilon} = \mathbf{L}\,\boldsymbol{u} \tag{6}$$

with $i, j = 1, 2, 3$. The latter relation is equally valid for the virtual displacements and strains. Furthermore, linear material behavior in the form of Hooke's law with the elasticity tensor in matrix notation $\mathbf{C}$ is assumed:

$$\boldsymbol{\sigma} = \mathbf{C}\,\boldsymbol{\varepsilon}. \tag{7}$$

The kinematic relations (6) and Hooke's law (7) are used in the virtual work principle (3) such that it can be given in terms of the displacements $\boldsymbol{u}$ and virtual displacements $\boldsymbol{\delta u}$:

$$\int_V \boldsymbol{\delta u}^{\mathrm{T}} \mathbf{L}^T \mathbf{C}\,\mathbf{L}\,\boldsymbol{u}\,\mathrm{d}V = \int_V \boldsymbol{\delta u}^{\mathrm{T}} \boldsymbol{f}\,\mathrm{d}V + \int_{S_t} \boldsymbol{\delta u}^{\mathrm{T}} \boldsymbol{t}^*\,\mathrm{d}A. \tag{8}$$

A fundamental prerequisite in the application of the SBFEM is that the considered elasticity problem is geometrically scalable, i.e. it must be possible to connect any point on a simply connected boundary to a chosen scaling center by a straight line (scaling ray), without this line meeting the boundary at any other point. An exception only is possible if the scaling center itself is located on the boundary and a scaling ray coincides with the boundary. In fact, this exceptional case is beneficially exploited in the analysis of cracks, V-notches and multimaterial corners. However, if the scalability condition is fulfilled, the displacement field can be expressed by a separation of variables representation. In the standard SBFEM for bounded domains (only one simply connected, outer boundary as it is exclusively considered within this work), the boundary is discretized creating pyramidal scaled boundary finite elements (SBFE). For each SBFE, an isoparametric formulation with a scaled boundary coordinate system is introduced (Fig. 1). The scaling coordinate $\xi$ runs from the scaling center ($\xi = 0$) to the discretized boundary ($\xi = 1$). The boundary coordinates $\eta_1, \eta_2$ specify the position on the discretized boundary ($\eta_1, \eta_2 \in [-1, 1]$) each being zero at the center of the SBFE. This coordinate system is not necessarily orthonormal. Then, the approximated displacement field $\tilde{\boldsymbol{u}}_{st}^{(e)}$ of a standard SBFE $e$ is represented as

$$\tilde{\boldsymbol{u}}_{st}^{(e)}(\xi, \eta_1, \eta_2) = \mathbf{N}(\eta_1, \eta_2)\,\boldsymbol{u}_n^{(e)}(\xi) \tag{9}$$

with 2D-element-shape-functions $\mathbf{N}(\eta_1, \eta_2)$ and nodal displacement functions $\boldsymbol{u}_n^{(e)}(\xi)$ on the scaling rays of the boundary nodes ($\xi \in [0, 1]$). Unless stated otherwise, the following derivations all apply to a single SBFE $e$. So from now on, the labeling $(\cdot)^{(e)}$ is omitted for brevity.

Our proposition is to enrich this separation of variables representation (9) by an additional field to better fulfill local boundary conditions at discontinuities, e.g. at crack faces:

$$\tilde{\boldsymbol{u}}(\xi, \eta_1, \eta_2) = \tilde{\boldsymbol{u}}_{st}(\xi, \eta_1, \eta_2) + \tilde{\boldsymbol{u}}_{enr}(\xi, \eta_1, \eta_2). \tag{10}$$

It is often advantageous to define a new conical coordinate system $\xi, r, \varphi$ for this additional field. Its origin is placed at the scaling center as well and $\xi$ again is the scaling coordinate. The scaled distance from the considered singularity line is denoted by $r$, while $\varphi$ represents the angular coordinate. This coordinate system again is not necessarily orthonormal. Then, the separation of variables representation for the enriched field can be written in conical coordinates

$$\tilde{\boldsymbol{u}}_{enr}(\xi, r, \varphi) = \mathbf{F}(r, \varphi)\, \boldsymbol{a}(\xi). \tag{11}$$

with the enrichment functions $\mathbf{F}(r, \varphi)$ and their free coefficients $\boldsymbol{a}(\xi)$ which can be functions of the scaling coordinate. The coordinates $r, \varphi$ can also be given in terms of the boundary coordinates $\eta_1, \eta_2$ if and only if the $r, \varphi$-plane and the $\eta_1, \eta_2$-plane are identical.

To be able to use the resulting enriched separation of variables representation in the virtual work principle (8), the differential operator $\mathbf{L}$ needs to be partitioned into the partial differential operators of the introduced coordinate systems.

$$\mathbf{L} = \mathbf{L}_x \frac{\partial}{\partial x} + \mathbf{L}_y \frac{\partial}{\partial y} + \mathbf{L}_z \frac{\partial}{\partial z} = \mathbf{L}_\xi \frac{\partial}{\partial \xi} + \mathbf{L}_{\eta_1} \frac{1}{\xi} \frac{\partial}{\partial \eta_1} + \mathbf{L}_{\eta_2} \frac{1}{\xi} \frac{\partial}{\partial \eta_2} = \mathbf{L}_\xi \frac{\partial}{\partial \xi} + \mathbf{L}_r \frac{1}{\xi} \frac{\partial}{\partial r} + \mathbf{L}_\varphi \frac{1}{\xi r} \frac{\partial}{\partial \varphi} \tag{12}$$

The partial differential operators $\mathbf{L}_x, \mathbf{L}_y, \mathbf{L}_z$ of the cartesian coordinate system and the corresponding Jacobi matrices $\mathbf{J}_\eta, \mathbf{J}_{r\varphi}$ are used to determine the partial differential operators $\mathbf{L}_\xi$, $\mathbf{L}_{\eta_1}$, $\mathbf{L}_{\eta_2}$, $\mathbf{L}_r$, $\mathbf{L}_\varphi$ of the conical and the scaled coordinate system. $\mathbf{L}_\xi$ happens to be identical for the conical and the scaled boundary coordinate system if the boundary is plane.

In the general case, the scaling center is located at certain coordinates $\boldsymbol{x_0}$ in a global cartesian coordinate system $x, y, z$. Then, the cartesian coordinates can be expressed in terms of the scaled boundary coordinates

$$\boldsymbol{x} = \boldsymbol{x_0} + \xi\, \boldsymbol{x_\eta}(\eta_1, \eta_2) \tag{13}$$

and in terms of the conical coordinates

$$\boldsymbol{x} = \boldsymbol{x_0} + \xi\, \boldsymbol{x_{r\varphi}}(r, \varphi) \tag{14}$$

where both $\boldsymbol{x_\eta}(\eta_1, \eta_2)$ and $\boldsymbol{x_{r\varphi}}(r, \varphi)$ denote the coordinates on the boundary ($\xi = 1$). These two equations describe the coordinate transformations needed for the determination of the Jacobi matrices. In this notation, the Jacobi matrices $\mathbf{J}_\eta(\eta_1, \eta_2)$ and $\mathbf{J}_{r\varphi}(r, \varphi)$ are functions of only the boundary coordinates so that the volumetric differential can be written as

$$\mathrm{d}V = \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z = \|\mathbf{J}_\eta(\eta_1, \eta_2)\|\, \xi^2\, \mathrm{d}\xi\, \mathrm{d}\eta_1\, \mathrm{d}\eta_2 = \|\mathbf{J}_{r\varphi}(r, \varphi)\|\, \xi^2 r\, \mathrm{d}\xi\, \mathrm{d}r\, \mathrm{d}\varphi. \tag{15}$$

Before the enriched displacement field approximation (10) is substituted in the virtual work principle (8), some terms of the differential operator and the enriched displacement field approximation should be combined. We distinguish between the standard SBFEM part of the displacement field approximation (9) and the enriching part (11). The shape functions $\mathbf{N}(\eta_1, \eta_2)$

in eq. (9) are combined with the partial differential operators of the scaled boundary coordinate system, resulting in the matrices $\mathbf{B}_\xi^{st}(\eta_1,\eta_2), \mathbf{B}_\eta^{st}(\eta_1,\eta_2)$.

$$\left[\mathbf{L}_\xi \frac{\partial}{\partial \xi} + \mathbf{L}_{\eta_1} \frac{1}{\xi} \frac{\partial}{\partial \eta_1} + \mathbf{L}_{\eta_2} \frac{1}{\xi} \frac{\partial}{\partial \eta_2}\right] \mathbf{N}(\eta_1,\eta_2) = \mathbf{B}_\xi^{st}(\eta_1,\eta_2) \frac{\partial}{\partial \xi} + \mathbf{B}_\eta^{st}(\eta_1,\eta_2) \frac{1}{\xi} \tag{16}$$

On the other hand, the enrichment functions $\mathbf{F}(r,\varphi)$ in eq. (11) are combined with the partial differential operators of the conical coordinate system, resulting in the matrices $\mathbf{B}_\xi^{enr}(r,\varphi)$ and $\mathbf{B}_{r\varphi}^{enr}(r,\varphi)$.

$$\left[\mathbf{L}_\xi \frac{\partial}{\partial \xi} + \mathbf{L}_r \frac{1}{\xi} \frac{\partial}{\partial r} + \mathbf{L}_\varphi \frac{1}{\xi r} \frac{\partial}{\partial \varphi}\right] \mathbf{F}(r,\varphi) = \mathbf{B}_\xi^{enr}(r,\varphi) \frac{\partial}{\partial \xi} + \mathbf{B}_{r\varphi}^{enr}(r,\varphi) \frac{1}{\xi} \tag{17}$$

Restructuring of the matrices leads to the following formulation of the product of the differential operator and the enriched displacement field approximation $\tilde{\boldsymbol{u}}$ (strain field approximation $\tilde{\boldsymbol{\varepsilon}}$, cf. eq. (6):

$$(\tilde{\boldsymbol{\varepsilon}} =)\ \mathbf{L}\,\tilde{\boldsymbol{u}}(\xi,\eta_1,\eta_2) = \left( \underbrace{\begin{bmatrix} \mathbf{B}_\xi^{st} \\ \mathbf{B}_\xi^{enr} \end{bmatrix}}_{\mathbf{B}_\xi(\eta_1,\eta_2)} \frac{\partial}{\partial \xi} + \underbrace{\begin{bmatrix} \mathbf{B}_\eta^{st} \\ \mathbf{B}_{r\varphi}^{enr} \end{bmatrix}}_{\mathbf{B}_\eta(\eta_1,\eta_2)} \frac{1}{\xi} \right) \underbrace{\begin{bmatrix} \boldsymbol{u}_n(\xi) \\ \boldsymbol{a}(\xi) \end{bmatrix}}_{\boldsymbol{u}_{nc}(\xi)} \tag{18}$$

with both coordinates $r$ and $\varphi$ being expressed in terms of $\eta_1$ and $\eta_2$. The vector $\boldsymbol{u}_{nc}$ contains the complete set of degrees of freedom on the boundary. Analogously, its virtual counterpart will be denoted $\boldsymbol{\delta u}_{nc}$. As it is assumed that this approximation also holds for the virtual displacements, the internal virtual work can be written as

$$\delta W_i = \int_V \left[\boldsymbol{\delta u}_{nc,\xi}^{\mathrm{T}}\, \mathbf{B}_\xi^{\mathrm{T}} + \boldsymbol{\delta u}_{nc}^{\mathrm{T}}\, \mathbf{B}_\eta^T \frac{1}{\xi}\right] \mathbf{C} \left[\mathbf{B}_\xi\, \boldsymbol{u}_{nc,\xi} + \mathbf{B}_\eta\, \boldsymbol{u}_{nc} \frac{1}{\xi}\right] \mathrm{d}V \tag{19}$$

This expression is identical to the standard SBFEM derivation except for the divergent formulation of the matrices $\mathbf{B}_\xi$ and $\mathbf{B}_\eta$. Expansion of the product and integration by parts over $\xi$ results in an expression which can be integrated separately. Integration over the boundary coordinates $\eta_1, \eta_2$ yields the matrices

$$\begin{aligned}
\mathbf{E}_0 &= \int_{S_\xi} \mathbf{B}_\xi^{\mathrm{T}}(\eta_1,\eta_2)\, \mathbf{C}\, \mathbf{B}_\xi(\eta_1,\eta_2)\, \|\mathbf{J}_\eta(\eta_1,\eta_2)\|\, \mathrm{d}\eta_1\, \mathrm{d}\eta_2, \\
\mathbf{E}_1 &= \int_{S_\xi} \mathbf{B}_\eta^{\mathrm{T}}(\eta_1,\eta_2)\, \mathbf{C}\, \mathbf{B}_\xi(\eta_1,\eta_2)\, \|\mathbf{J}_\eta(\eta_1,\eta_2)\|\, \mathrm{d}\eta_1\, \mathrm{d}\eta_2, \\
\mathbf{E}_2 &= \int_{S_\xi} \mathbf{B}_\eta^{\mathrm{T}}(\eta_1,\eta_2)\, \mathbf{C}\, \mathbf{B}_\eta(\eta_1,\eta_2)\, \|\mathbf{J}_\eta(\eta_1,\eta_2)\|\, \mathrm{d}\eta_1\, \mathrm{d}\eta_2.
\end{aligned} \tag{20}$$

$S_\xi$ denotes the discretized boundary with constant scaling coordinate $\xi$. A numerical integration procedure is implemented but attention has to be paid that a sufficiently high integration order is used to account for the enrichment functions. In fact, we distinguish between integration of the purely standard SBFEM parts of the matrices, for which we can use appropriate low order Gauss integration schemes, and all other parts which also depend on the enrichment functions.

Nevertheless, we omitted special integration schemes otherwise used in literature (e.g. [22, 27, 37]) because we already achieved adequate accuracy with a simple high order Gauss integration. However, more sophisticated integration schemes surely would further improve the numerical efficiency.

If the considered boundary value problem involves more than one SBFE, which generally is the case, the element matrices $\mathbf{E}_0, \mathbf{E}_1, \mathbf{E}_2$ are similarly assemled like the element stiffness matrices in the standard Finite Element Method (FEM [44]). Then, they are substituted in the internal virtual work yielding one integral term and one boundary term which is to be evaluated at the boundaries $\xi = 0$ and $\xi = 1$.

$$\begin{aligned} - \quad & \int_0^1 \boldsymbol{\delta u}_{nc}^{\mathrm{T}}(\xi) \Big[ \xi^2 \mathbf{E}_0 \boldsymbol{u}_{nc}(\xi)_{,\xi\xi} + \xi \Big[ 2\mathbf{E}_0 - \mathbf{E}_1 + \mathbf{E}_1^{\mathrm{T}} \Big] \boldsymbol{u}_{nc}(\xi)_{,\xi} + \Big[ \mathbf{E}_1^T - \mathbf{E}_2 \Big] \boldsymbol{u}_{nc}(\xi) \Big] \, \mathrm{d}\xi \\ + \quad & \boldsymbol{\delta u}_{nc}^{\mathrm{T}}(\xi) \Big[ \xi^2 \mathbf{E}_0 \boldsymbol{u}_{nc}(\xi)_{,\xi} + \xi \mathbf{E}_1^{\mathrm{T}} \boldsymbol{u}_{nc}(\xi) \Big]_{\xi=0}^{1} \quad = \quad \delta W_i \end{aligned} \tag{21}$$

If it is assumed that virtual external work terms are nonzero only at the discretized boundary $S_\xi$ (i.e. stress free boundary conditions at all other boundaries), these terms are independent of the scaling coordinate so that they only interact with the boundary term of the internal virtual work. Then, the virtual work principle is fulfilled if the integrand of the integral term in the internal virtual work is zero on the one hand, which yields a homogeneous Cauchy-Euler differential equation system of second order. On the other hand, all remaining terms in the virtual work principle form a linear equation system for enforcing the boundary conditions at the discretized boundary $S_\xi$.

Cauchy-Euler differential equation systems can be solved using standard methods leading to a quadratic eigenvalue problem with eigenvalues $\lambda$ and eigenvectors $\boldsymbol{\phi}$.

$$\Big( \lambda^2 \mathbf{E}_0 + \lambda \Big[ \mathbf{E}_0 - \mathbf{E}_1 + \mathbf{E}_1^T \Big] + \Big[ \mathbf{E}_1^T - \mathbf{E}_2 \Big] \Big) \, \boldsymbol{\phi} = \mathbf{0} \tag{22}$$

We solve the quadratic eigenvalue problem by, first, transforming it into a linear one at the cost of doubling the degrees of freedom, and second, applying the QR-algorithm implemented in the *eig()* function in the numerical mathematics software MATLAB.

If the solution does not contain any Jordan blocks, i.e. the geometric multiplicity of all the eigenvalues is equal to their algebraic one, the solution can be written as

$$\boldsymbol{u}_{nc}(\xi) = \boldsymbol{\Phi} \, \xi^{\boldsymbol{\lambda}} \, \boldsymbol{c} \tag{23}$$

with $\xi^{\boldsymbol{\lambda}}$ being a diagonal matrix with entries $\xi^{\lambda_i}$ and $\boldsymbol{c}$ being a vector of free constants which are to be determined from the linear equation system for the boundary conditions at $S_\xi$. The matrix $\boldsymbol{\Phi}$ contains all eigenvectors $\boldsymbol{\phi}_i$, which in turn comprise the nodal boundary displacements and the enrichment functions' coefficients at the discretized boundary $S_\xi$.

Interpreting the eigenvectors $\boldsymbol{\phi}_i$ as deformation modes and the eigenvalues $\lambda_i$ as their associated decay rates, this representation just matches the power series representation (2) for 3D boundary value problems containing singularities.

## 3 NUMERICAL RESULTS AND DISCUSSION

Next, we present some results that we obtained using this newly developed method. First, the situation of a simple crack in an isotropic homogeneous material is considered as a benchmark example before we proceed to the more complex structural situation of two meeting cracks as it has been studied in [16].

In both cases, a cubic domain containing the discontinuities is modeled as depicted in fig. 2. Only the boundary is discretized and plane 4- or 8-node SBFEs with linear or quadratic shape functions respectively are used. The scaling center is located at the cube's center and the cracks

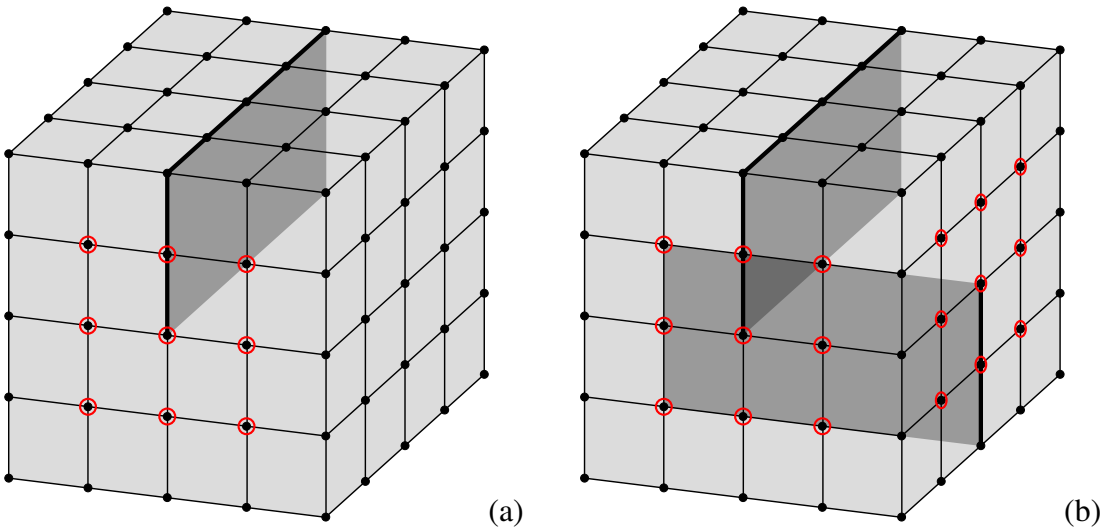



Figure 2: SBFE boundary meshes of cubes consisting of plane, 4-node SBFEs with (a) one crack (shaded) and (b) two meeting cracks. Enriched nodes are circled (red). Cracks are modeled using double nodes. The scaling center $S$ is placed at the cube's center.

are modeled by inserting double nodes where a crack breaks the boundary. Consequently, the crack just reaches the scaling center and the crack front coincides with scaling rays. This also means that, contrary to the common practice in the popular XFEM which involves similar crack enrichments, the cracks are located between the elements instead of within them and there is no need for Heavyside enrichments.

The only enrichment functions needed are based on the classical crack tip near-field (e.g. [1, 14]). We adopt the decomposition of Fleming et al. [10]:

$$\mathbf{F}(r,\varphi) = \sqrt{r} \cdot [\sin(\varphi/2), \cos(\varphi/2), \sin(\varphi/2)\sin(\varphi), \cos(\varphi/2)\sin(\varphi)]. \tag{24}$$

Only nodes close to a crack front breaking through the boundary are enriched. For saving degrees of freedom (DOF) in the quadratic eigenvalue problem, which is essential for the numerical efficiency of its solution, the plateau method, e.g. used by [22, 11, 37], is adopted. In this method, all enriched nodes in the domain of influence of one crack tip are constrained to the same enrichment function coefficients, leading to only 4 additional DOF per dimension and per crack tip. In our formulation, it does not have to be accounted for crack tips but for locations where the crack front meets the boundary, leading to only $4 \cdot 3 \cdot 2 = 24$ additional DOF per crack.

As only certain domains and not all nodes are enriched, so-called blending elements are also needed. We follow the implementation of [11, 37] and denote the set of all nodes of one element $\mathcal{N}$ and the set only containing the element's enriched nodes $\mathcal{P}$. Furthermore, the values of the enrichment functions at the nodes are subtracted from the enrichment so that the functions $\boldsymbol{u}_n(\xi)$ remain the actual displacement functions at the scaling rays. Then, the complete separation of variables representation for the displacements in index notation can be written as

$$\begin{aligned}\tilde{u}_i(\xi,\eta_1,\eta_2) = &\sum_{k\in\mathcal{N}} N_k(\eta_1,\eta_2)\,\tilde{u}_{ik}(\xi) \quad + \\ &\sum_{l\in\mathcal{P}} N_l(\eta_1,\eta_2)\left[\sum_{m=1}^{n_F}\left(F_m(r,\varphi) - \sum_{k\in\mathcal{N}} N_k(\eta_1,\eta_2)\,F_m(r_k,\varphi_k)\right) a_{im}(\xi)\right]\end{aligned} \tag{25}$$

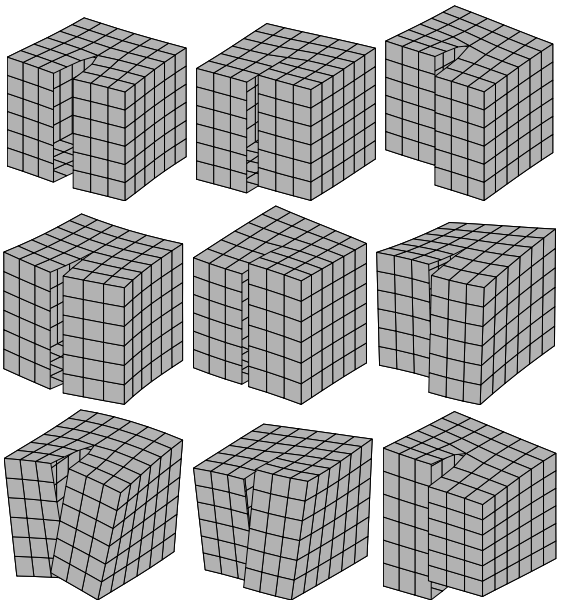


Figure 3: Deformation modes of a crack in an isotropic continuum for eigenvalues $\lambda = 0.5$ (1st line) and $\lambda = 1.5$. 1st line: classical crack modes I, II and III [1]. 2nd and 3rd line: modes IV, V, VI and VII,VIII,IX.

with $i = 1, 2, 3$ and $n_F$ being the number of enrichment functions (here: $n_F = 4$). The coefficients of $\boldsymbol{u}_n(\xi)$ are simply denoted $\tilde{u}_{ik}(\xi)$. The quantities $r_k$ and $\varphi_k$ are the coordinates of the respective element nodes.

In fact, all elements on a cube face exhibiting a boundary breaking crack front contain enriched nodes and, consequently, require high order integration. Of course, the effect of the high integration order on the computational effort for the numerical integration is significant. For example, the computation time on a standard desktop PC (without exploiting any parallelization techniques), increases from 1.3 seconds (all elements with low order integration) to 9.3 seconds (high order integration at two cube faces) for a cube with 6 elements along each edge. But due to the increasing dominance of the computational effort necessary for solving the quadratic eigenvalue problem, the effect of the high integration order on the overall computational effort fades out with increasing mesh density.

The effect of the enrichment on the computational effort for solving the quadratic eigenvalue problem quickly becomes negligible as the number of elements $n_e$ along each edge of a cube is increased because the enrichment for one crack only adds 24 DOFs while a standard SBFEM approach for a cube with one crack and $n_e = 6$ already results in $[(6n_e^2+2)+(2n_e-1)]\cdot 3 = 687$ DOF ($n_e = 10$: 1863 DOFs; $n_e = 20$: 7323 DOFs). On a standard desktop PC, the solution of the linearized quadratic eigenvalue problem using MATLAB's *eig()* function (QR-algorithm implementation) takes about 2.5 seconds for $n_e = 6$ but already about 42 seconds for $n_e = 10$ and 1655 seconds for $n_e = 20$. This, again, strongly emphasizes the need for a DOF-efficient formulation and implementation.

### 3.1 Crack in isotropic homogeneous continuum (benchmark example)

The benchmark example of an isotropic homogeneous continuum containing a single plane crack is considered (fig. 2a)) [1, 14]. In the three-dimensional solution, there are the three classical singular deformation modes with $\lambda = 0.5$ and six deformation modes with $\lambda = 1.5$. Please note, that three of the six deformation modes with decay rate $\lambda = 1.5$ cannot be found by a two-dimensional SBFEM analysis as they involve a displacement variation in the direction

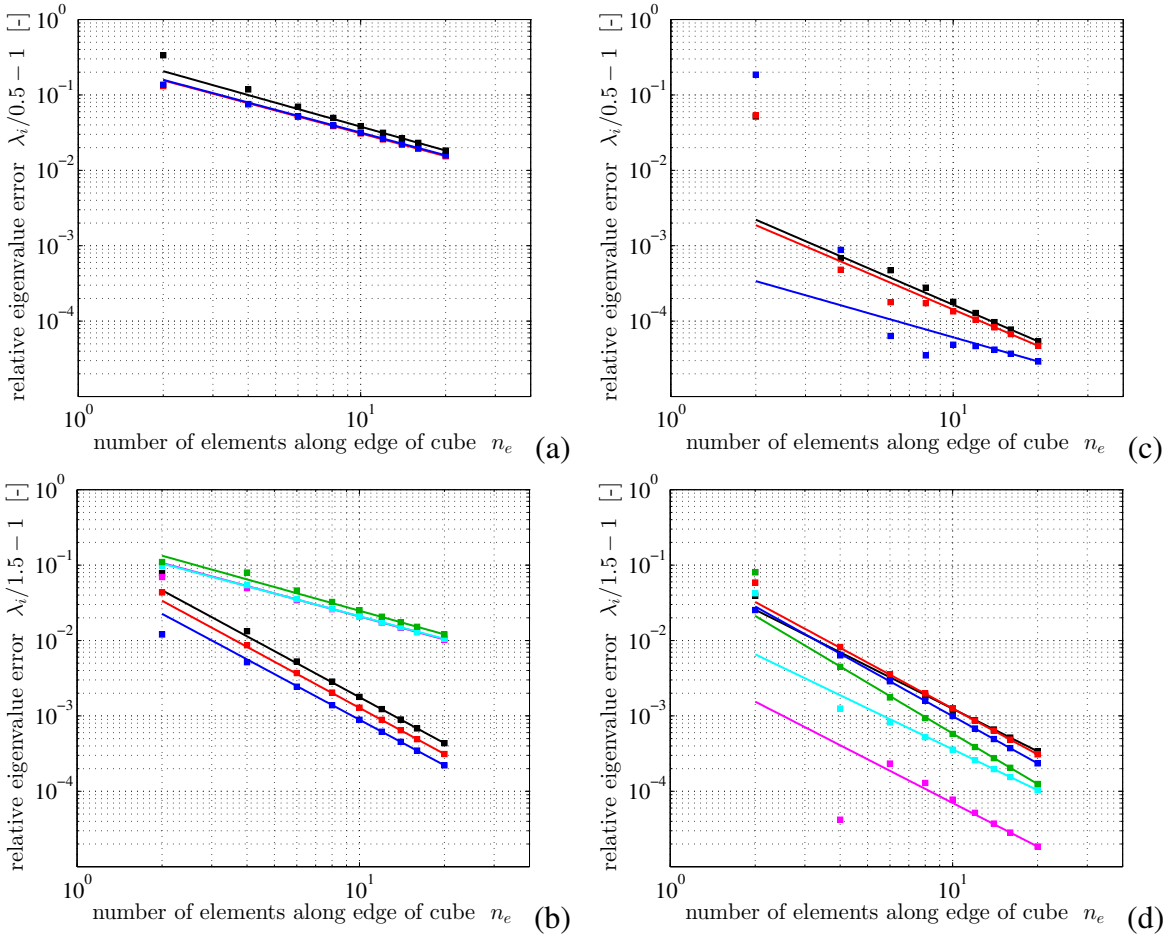



Figure 4: Convergence of some eigenvalues $\lambda_i$ using the standard (a,b) and the enriched SBFEM (c,d), both with uniform meshing and linear 4-node SBFE (C4) (cf. fig. 2a). Plots (a,b) contain the classical crack modes I (red), II (blue) and III (black). Plots (c,d) contain the deformation modes IV (red), V (blue), VI (black), VII (magenta), VIII (cyan) and IX (green).

of the crack front (cf. fig. 3, 3rd line, deformation modes VII, VIII, IX).

In double logarithmic plots, fig. 4 shows the relative error of the calculated eigenvalues $\lambda_i$ from the quadratic eigenvalue problem solution (22) to the analytical values of $0.5$ and $1.5$ for a uniform meshing with $n_e = 2, 4, 6, 8, 10, 12, 14, 16$ and $20$ elements with linear shape functions (C4). As the computational effort quickly increases with an increasing number of DOF in the quadratic eigenvalue problem, computations with higher mesh densities were avoided. Yet at a glance, it can be seen that the enrichment of the displacement field leads to a tremendous reduction of the approximation error of the considered eigenvalues. Indeed, the relative error is reduced by several orders of magnitude, even for very low discretization levels: while the standard formulation with $n_e = 20$ elements along each edge of the cube and a computation time of about 30 minutes is not even capable of pushing the error in the singular modes below $1\%$ (fig. 4a), the enriched formulation already attains an error $< 0.1\%$ for a discretization with only $n_e = 4$ and an overall computation time of 5 seconds (fig. 4c). It is to be noted that $n_e = 4$ actually is the lowest discretization which fully benefits from the enrichment because the discretization with $n_e = 2$ only contains blending and standard elements but no fully enriched ones.

The convergence orders mostly are considerably improved as well. However, at very low rel-

| deformation mode | I | II | III | IV | V | VI | VII | VIII | IX |
|---|---|---|---|---|---|---|---|---|---|
| $m_\lambda$ SBFEM (C4) | 1.0 | 1.0 | 1.0 | 2.0 | 2.0 | 2.0 | 1.0 | 1.0 | 1.0 |
| $m_\lambda$ enrSBFEM (C4) | 1.6 | 1.6 | 1.1 | 2.0 | 2.1 | 1.9 | 2.0 | 1.8 | 2.2 |

Table 1: Comparison of the approximated convergence orders $m_\lambda$ of the standard (SBFEM) and the enriched formulation (enrSBFEM) with linear shape functions (C4) for considered deformation modes.

ative errors an improvement seems difficult, especially because the convergence is not always monotonic and only comparably low mesh densities were considered. Especially in the case of crack mode II (fig. 4c blue line and dots), we expect a further increase of the approximated convergence order with decreasing element size. Table 1 gives the approximated convergence orders $m_\lambda$ of the eigenvalues $\lambda_i$. They were approximated from the last few values using a fitting procedure. The convergence orders achieved by the standard SBFEM formulation do not reach the optimal convergence order of $m_\lambda^{\text{opt}} = 2$ (e.g. [2]) for most of the considered deformation modes. For deformation modes I to III and VII to IX, we get the reduced convergence order of $m_\lambda \approx 1$ which indeed is the expected value when a singularity of order 0.5 is present on the mesh. Surprisingly, for the deformation modes IV, V and VI, which contrary to the deformation modes VII, VIII, IX also occur in a 2D SBFEM analysis, optimal convergence is achieved anyway but we do not have an explanation for this observation so far. However, the enriched formulation achieves about optimal convergence for all deformation modes IV to IX. For the deformation modes I to III, the convergence order is smaller but the relative error of the eigenvalues is of an almost negligible order of magnitude anyway.

Fig. 5 shows the relative error of the calculated eigenvalues $\lambda_i$ from the quadratic eigenvalue problem solution (22) for a uniform meshing with $2, 4, 6, 8, 10$ and $12$ elements $n_e$ along the edge of the considered cube with quadratic shape functions (C8). A uniform mesh with $n_e = 12$ already accounts for $[(18n_e^2 + 2) + (4n_e - 1)] \cdot 3 = 7923$ DOF and a computation time of almost one hour for solving the complete eigenvalue problem. Thus, no finer meshes were considered, here. For SBFE with quadratic shape functions, an optimal convergence order of $m_\lambda^{\text{opt}} = 4$ is expected but not achieved for any eigenvalue or formulation considered. In the cases for which optimal convergence was achieved with linear shape functions (C4), now, the convergence order is bounded by the decay rate of the next two-dimensional deformation mode on the meshed plane which cannot be resolved by a polynomial. This is the decay rate and eigenvalue $\lambda = 1.5$ resulting in a maximum convergence order of $m_\lambda = 3$ as long as this deformation mode is not included in the enrichment. Nevertheless, wherever this increased convergence rate is achieved, there also is a considerable gain in accuracy. Otherwise, the error reduction may seem rather small when discretizations with the same number of nodes along each edge of the cube are compared. But there is an additional gain in efficiency as the employed serendipity elements (C8) have only 8 nodes instead of 9 and, thus, less DOF than a comparable mesh with 4-node elements. Finally, it should be mentioned that, naturally, the approximation of eigenvalues $\lambda = 2$ is distinctly superior compared to SBFE with linear shape functions (C4). Consequently, 8-node SBFE (C8) seem preferable whenever they are applicable.

Up to this point, only the error of the eigenvalues $\lambda_i$ has been studied. We would like to complete this section with a consideration of the quality of the eigenvectors $\boldsymbol{\phi}_i$, i.e. the actual deformation modes (angular functions in eq. (2)). Here, we concentrate on the classical crack modes, which pertain to the eigenvalue $\lambda = 0.5$. The first check involves the coefficients $a_{im}(\xi)$ of the enrichment functions $F_m(r, \varphi)$ (eq. 24). To be able to exactly reproduce the analytical displacement fields [25, 14] given below, they must fulfill certain relations which we are going

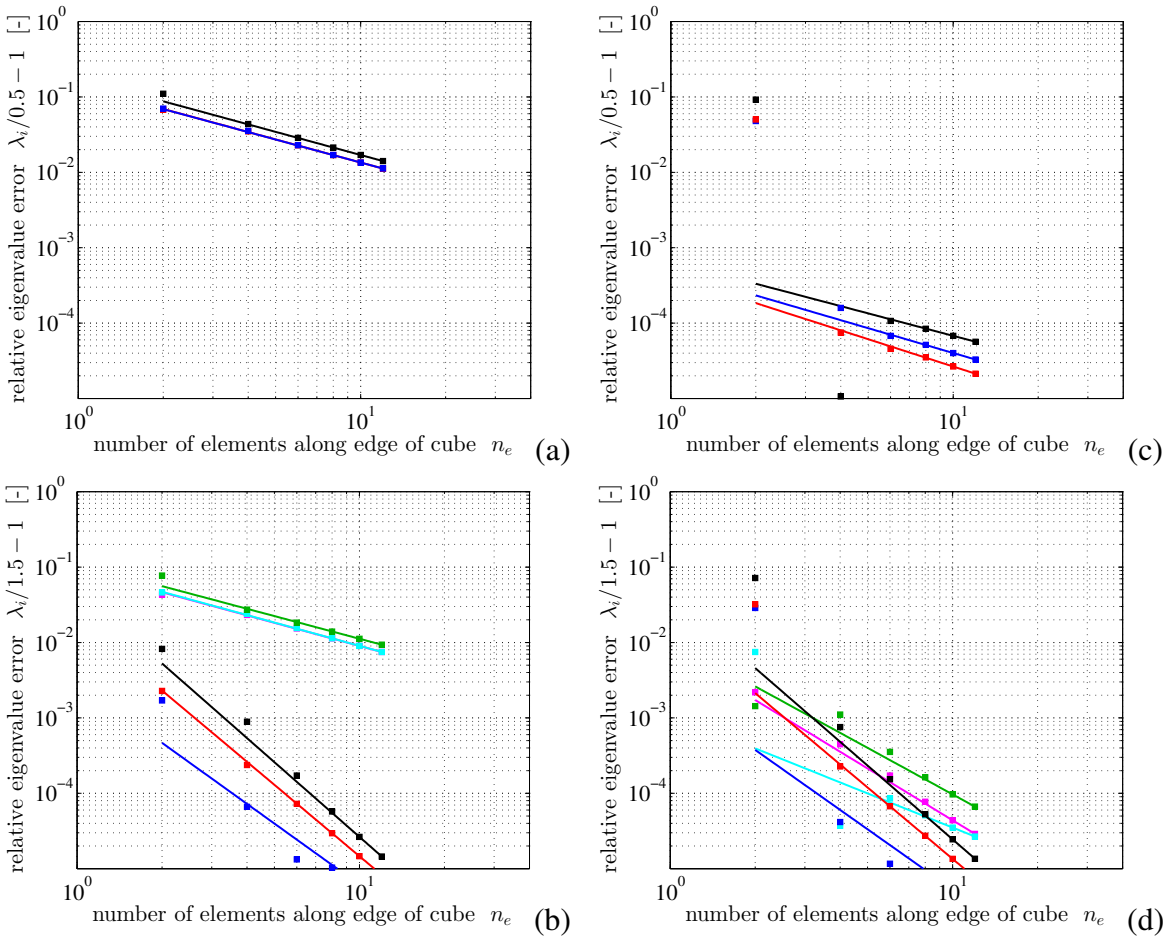



Figure 5: Convergence of some eigenvalues $\lambda_i$ using the standard (a,b) and the enriched SBFEM (c,d), both with uniform meshing and quadratic 8-node serendipity elements (C8). Plots (a,b) contain the classical crack modes I (red), II (blue) and III (black). Plots (c,d) contain the deformation modes IV (red), V (blue), VI (black), VII (magenta), VIII (cyan) and IX (green).

to examine in the following (cf. Table 2 column 2).

$$\underline{u}_I(r,\varphi) = \frac{K_I}{2G}\sqrt{\frac{r}{2\pi}}(\kappa - \cos(\varphi)) \begin{bmatrix} \cos(\varphi/2) \\ \sin(\varphi/2) \\ 0 \end{bmatrix} \qquad \text{mode I} \tag{26}$$

$$\underline{u}_{II}(r,\varphi) = \frac{K_{II}}{2G}\sqrt{\frac{r}{2\pi}} \begin{bmatrix} \sin(\varphi/2)[2+\kappa+\cos(\varphi)] \\ \cos(\varphi/2)[2-\kappa-\cos(\varphi)] \\ 0 \end{bmatrix} \qquad \text{mode II} \tag{27}$$

$$\underline{u}_{III}(r,\varphi) = \frac{2K_{III}}{G}\sqrt{\frac{r}{2\pi}} \begin{bmatrix} 0 \\ 0 \\ \sin(\varphi/2) \end{bmatrix} \qquad \text{mode III} \tag{28}$$

The shear modulus is denoted by $G$ and Poisson's ratio by $\nu$. Further, it is $\kappa = 3 - 4\nu$ for plane strain assumptions, which, different from plane stress assumptions, make use of the full elasticity tensor and are completely consistent with the governing equations of 3D linear elasticity.

For a simpler comparison, we use the following enrichment functions for our reference cal-

| | required relations | analytical | enrSBFEM (8x8C4) | relative deviation |
|---|---|---|---|---|
| crack mode I | $a_{12}/a_{14} = -\kappa$ | -1.8 | -1.800213 | +0.012% |
| | $a_{21}/a_{23} = -\kappa$ | -1.8 | -1.799897 | -0.006% |
| crack mode II | $a_{11}/a_{13} = \kappa + 2$ | +3.8 | +3.802018 | +0.053% |
| | $a_{22}/a_{24} = \kappa - 2$ | -0.2 | -0.202246 | +1.123% |

Table 2: Comparison of analytical and numerical coefficient relations for crack modes I and II.

culation with $n_e = 8$ and linear shape functions (C4):

$$\mathbf{F}(r, \varphi) = \sqrt{r} \cdot [\sin(\varphi/2), \cos(\varphi/2), \sin(\varphi/2)\cos(\varphi), \cos(\varphi/2)\cos(\varphi)]. \tag{29}$$

They are more convenient because they better conform to the analytical field (26-28) but are still equivalent to Fleming's approach [22]. Table 2 shows the analytical coefficient relations and the results obained numerically for a material with a Poisson's ratio $\nu = 0.3$. It can be seen that the relative deviation of the numerically obtained values from the analytical ones for crack modes I and II is rather small. A further requirement for the exact representation of the analytical fields of crack modes I and II is that all other coefficients should be zero. Indeed, most of them are numerically zero and only single coefficients reach values of maximum 0.1% (mode I) and 0.03% (mode II) of the largest coefficient pertaining to the respective crack mode. Crack mode III only contains one analytical coefficient $a_{31}$ which is none-zero, so the only requirement resulting from this crack mode is that all other coefficients are zero, which also is fulfilled (maximum relative deviation 0.006%).

After the comparison of the enrichment contribution of the displacement approximation, which naturally is only valid within the enriched pyramidal parts of the cube, now, the behavior along the crack front is considered. The left plot in Fig. 6 shows the variation of the magnitude of the displacements of each crack mode along a path running parallel to the crack front at a distance of 1/6 (dashed lines) and 1/3 (solid lines) of the considered cube's edge length. The figure for the example of $n_e = 6$ elements along a cube's edge with linear shape functions (C4) is symmetric as it is to be expected. It can be seen that the variation of the magnitude of displacements is marginal. The right plot showing the relative displacement magnitude with respect to the boundary values confirms this observation as the relative deviation remains $< 0.7\%$. The contribution of the enrichments to the displacement magnitude is neglected. This results in a piecewise linear form whose first (solid lines) respective second (dashed lines) section corresponds to a blending element where the largest part of the deviation initiates.

The actual magnitude of the displacements is of minor interest as the deformation modes' scaling indeed is arbitrary and only affects the stress intensity factors in a complete boundary value problem which we here do not intend to solve. Under the obvious assumption that the analytical fields are best matched within the enriched elements, it is most important, how the displacements inside the cube deviate from those at the enriched cube faces. From the obtained results for the rather coarse mesh, we conclude this deviation to be of negligible magnitude.

### 3.2 Two meeting cracks in homogeneous isotropic continuum

In this section, the more complex structural situation of two plane, perpendicularly meeting cracks (fig. 2b) studied in [16] is revisited and the convergence behavior of the first six eigenvalues resulting in stress singularities is investigated. The associated deformation modes are depicted and labeled in figure 7.

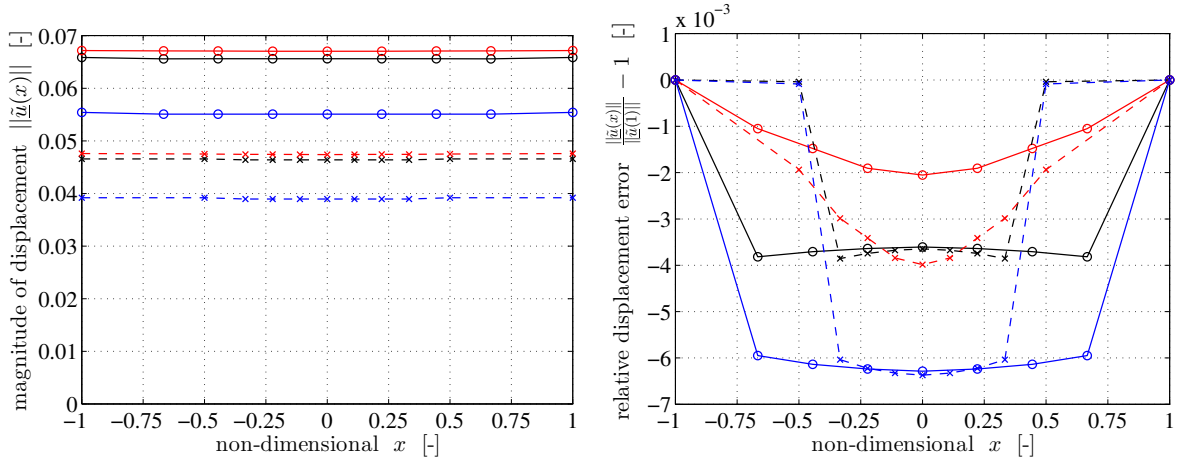



Figure 6: Plot of magnitude of displacements (left) and relative deviation of the displacement magnitude with respect to the boundary displacements (right) over a non-dimensional coordinate parallel to the crack front for crack modes I (red), II (blue), III (black). They are given on one crack face at 1/3 (dashed lines) and 2/3 (solid lines) of the distance of the crack front to the outer boundary for the example of $n_e = 6$ (C4). The displacement contribution of the enrichments is neglected.

As no analytical and, accordingly, no exact reference solution is known for this situation, the converged values are identified by a Richardson extrapolation, i.e. a function of the type $\lambda = C\, n_e^{m_\lambda} \lambda^{ext}$ is fitted to the numerical data. The quantities $n_e$ (number of elements along edge of cube) and $m_\lambda$ (convergence order) have already been introduced in the previous section. The eigenvalue approximated by the numerical procedure is denoted as $\lambda$. $\lambda^{ext}$ is the extrapolated "exact" eigenvalue and $C$ is a free constant.

| deformation mode | *co1* | *co2* | *cs1* | *cs2* | *ct1* | *ct2* |
|---|---|---|---|---|---|---|
| $\lambda^{ext}$ SBFEM | 0.38260 | 0.67315 | 0.50749 | 0.50749 | 0.33982 | 0.79260 |
| $\lambda^{ext}$ enrSBFEM | 0.38241 | 0.67299 | 0.50775 | 0.50775 | 0.33846 | 0.79206 |
| $m_\lambda$ SBFEM (C4) | 1.0 | 1.2 | 1.0 | 1.0 | 1.2 | 1.0 |
| $m_\lambda$ enrSBFEM (C4) | 1.8 | 1.8 | 1.5 | 1.5 | 2.9 | 2.0 |

Table 3: Comparison of the identified convergence orders $m_\lambda$ of the standard (SBFEM) and the enriched formulation (enrSBFEM) with linear shape functions (C4) for the considered deformation modes.

Table 3 shows the results from this extrapolation procedure for the standard SBFEM formulation and the enriched one. It can be seen that the extrapolated eigenvalues $\lambda^{ext}$ in essence coincide for both methods, but also that the associated orders of convergence differ considerably. While the convergence order of the standard formulation hardly exceeds the value of $m_\lambda = 1$, this value is easily surpassed by the enriched formulation. The eigenvalue associated to deformation mode *ct1* converges considerably faster than the others. In fact, even superconvergence seems to be achieved for this deformation mode but, up to date, we lack an explanation for this phenomenon.

From fig. 8, it can furthermore be concluded that the relative error in the eigenvalues is reduced by approximately two orders of magnitude when the enriched formulation is employed.

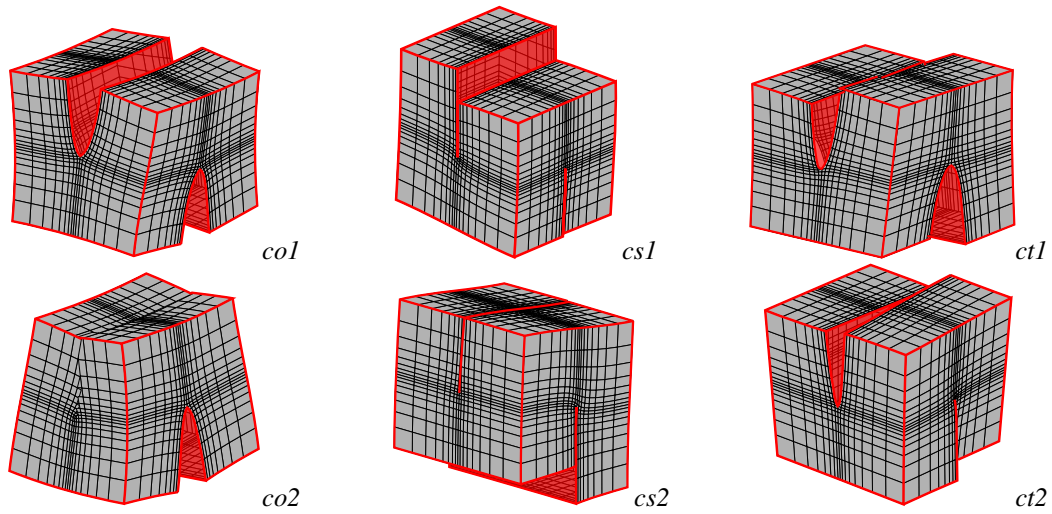



Figure 7: Deformation modes of two plane, perpendicularly meeting cracks in a homogeneous isotropic continuum for a cubical boundary mesh: 2 crack opening modes (*co1/2*), 2 crack shearing modes (*cs1/2*), 2 crack twisting modes (*ct1/2*).

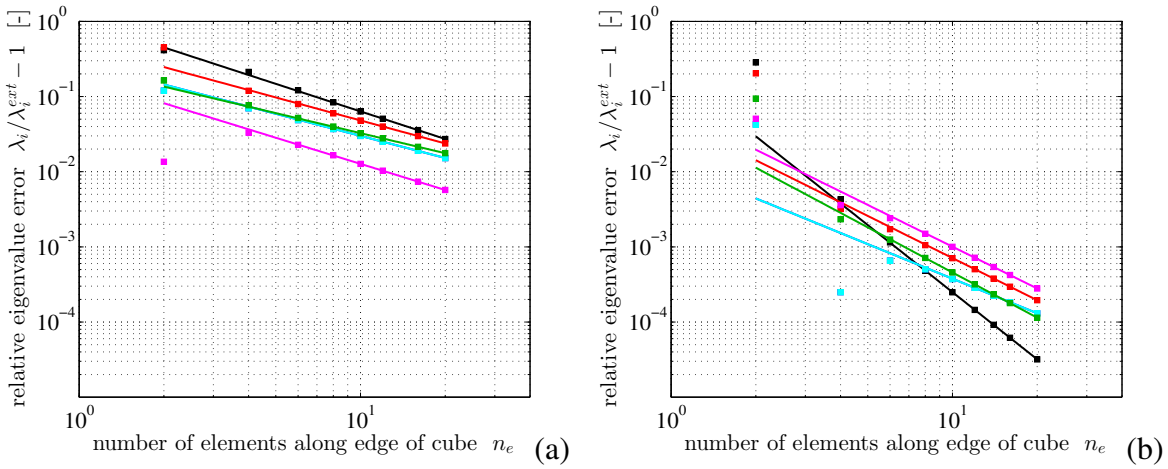



Figure 8: Convergence of first six eigenvalues $\lambda_i$ which result in stress singularities using the standard (a) and the enriched SBFEM (b), both with uniform meshing and linear 4-node SBFE (C4). The corresponding deformation modes are *co1* (red line), *co2* (magenta), *cs1* (blue), *cs2* (cyan), *ct1* (black) and *ct2* (green).

## 4 SUMMARY AND CONCLUSIONS

It was our goal to improve the efficiency of the solution of three-dimensional linear elastic boundary value problems containing stress singularities. In the semi-analytical FEM eigenanalysis techniques like the SBFEM, a separation of variables representation can be employed in the virtual work principle if the scalability requirement is fulfilled. This leads to a differential equation system (DES) of second order which can be transformed into a quadratic eigenvalue problem. The obtained solution of the DES is of power series form. Additionally, a linear equation system (LES) for the conformance of the DES solution to the boundary conditions is obtained. Only the boundary needs to be discretized, typically using linear shape functions similar to those in the standard FEM. This results in a considerable reduction of the degrees of freedom so that the low computational efficiency of the eigenvalue solvers often only has a negligible effect on the overall efficiency of the method. Consequently, the SBFEM has already proven its suitability in the two-dimensional case.

However, in the three-dimensional case, we have demonstrated its deficiencies when a stress singularity is present on the discretized boundary. Then, the relative error increases by orders of magnitude and the convergence order is reduced by a factor of 2. As a similar phenomenon is also known from standard FEM techniques, the employment of their remedies seems natural.

We selected the enrichment approach to be the most suitable one and implemented a separation of variables representation enriched with the classical crack's near-field. For the example of a single plane crack in an isotropic continuum, we demonstrated that, at the cost of only 24 additional degrees of freedom, excellent accuracy of both the eigenvalues and eigenvectors can already be achieved for a very coarse boundary mesh. Moreover, due to the implemented plateau method, this number of additional degrees of freedom is independent of other meshing parameters. It only depends on the number of enrichment functions and the number of singular points on the boundary mesh. The obtained convergence order also is substantially improved, often optimal and sometimes even superconvergence is observed.

For the more complex example of two plane cracks meeting perpendicularly in a homogeneous isotropic continuum, the results could not be compared to an analytic reference solution. So, the converged eigenvalues were estimated using a Richardson extrapolation and the relative error was computed with respect to these values. Indeed, the standard SBFEM and the enriched formulation yielded almost the same extrapolated eigenvalues, but with a very different relative accuracy and convergence order.

We conclude that the proposed approach, which can easily be coupled with standard 3D finite elements, has turned out to be uniquely DOF-efficient. In our opinion, it thus shows the potential of providing easy access to the general solution of complete 3D linear elastic boundary value problems containing stress singularities (instead of only determining singularity order and possibly also mode). Although, its excellent accuracy and convergence properties have so far only been shown for the classical crack singularity, we hope to extend its applicability to more general cases (notch singularities, multi-material junctions, etc.) in future works.

## REFERENCES

[1] T. L. Anderson and T. L. Anderson. *Fracture mechanics: fundamentals and applications*. CRC press, 2005.

[2] T. Apel, A.-M. Sändig, and S. I. Solov'ev. Computation of 3D vertex singularities for linear elasticity: Error estimates for a finite element method on graded meshes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 36:1043–1070, 11 2002.

[3] S. N. Atluri, A. S. Kobayashi, and M. Nakagaki. An assumed displacement hybrid finite element model for linear fracture mechanics. *International Journal of Fracture*, 11(2):257–271, 1975.

[4] R. S. Barsoum. Theoretical basis of the finite element iterative method for the eigenvalue problem in stationary cracks. *International Journal for Numerical Methods in Engineering*, 26(3):531–539, 1988.

[5] Z. P. Bažant and L. F. Estenssoro. Surface singularity and crack propagation. *International Journal of Solids and Structures*, 15(5):405–426, 1979.

[6] J. P. Benthem. State of stress at the vertex of a quarter-infinite crack in a half-space. *International Journal of Solids and Structures*, 13(5):479–492, 1977.

[7] S. Dai, C. Augarde, C. Du, and D. Chen. A fully automatic polygon scaled boundary finite element method for modelling crack propagation. *Engineering Fracture Mechanics*, 133:163–178, 2015.

[8] A. J. Deeks and J. P. Wolf. A virtual work derivation of the scaled boundary finite-element method for elastostatics. *Computational Mechanics*, 28(6):489–504, 2002.

[9] A. Dimitrov, H. Andrä, and E. Schnack. Efficient computation of order and mode of corner singularities in 3D-elasticity. *International Journal for Numerical Methods in Engineering*, 52(8):805–827, 2001.

[10] M. Fleming, Y. A. Chu, B. Moran, T. Belytschko, Y. Y. Lu, and L. Gu. Enriched element-free Galerkin methods for crack tip fields. *International Journal for Numerical Methods in Engineering*, 40(8):1483–1504, 1997.

[11] T.-P. Fries. A corrected XFEM approximation without problems in blending elements. *International Journal for Numerical Methods in Engineering*, 75(5):503–532, 2008.

[12] T.-P. Fries and T. Belytschko. The extended/generalized finite element method: An overview of the method and its applications. *International Journal for Numerical Methods in Engineering*, 84(3):253–304, 2010.

[13] F. Ghahremani. A numerical variational method for extracting 3D singularities. *International Journal of Solids and Structures*, 27(11):1371–1386, 1991.

[14] D. Gross and T. Seelig. *Bruchmechanik: Mit einer Einführung in die Mikromechanik*. Springer Berlin Heidelberg, 2011.

[15] L. Gu and T. Belytschko. A numerical study of stress singularities in a two-material wedge. *International Journal of Solids and Structures*, 31(6):865–889, 1994.

[16] S. Hell and W. Becker. The scaled boundary finite element method for the analysis of 3D crack interaction. *Journal of Computational Science*, 9:76–81, 2015. Computational Science at the Gates of Nature.

[17] R. D. Henshell and K. G. Shaw. Crack tip finite elements are unnecessary. *International Journal for Numerical Methods in Engineering*, 9(3):495–507, 1975.

[18] N. Khaji and M. I. Khodakarami. A semi-analytical method with a system of decoupled ordinary differential equations for three-dimensional elastostatic problems. *International Journal of Solids and Structures*, 49(18):2528–2546, 2012.

[19] M. Knein. *Zur Theorie des Druckversuchs*, volume 7 of *Abhandlungen aus dem Aerodynamischen Institut an der Technischen Hochschule Aachen*, pages 43–62. Springer Berlin Heidelberg, 1927.

[20] H. Koguchi and T. Muramoto. The order of stress singularity near the vertex in three-dimensional joints. *International Journal of Solids and Structures*, 37(35):4737–4762, 2000.

[21] M. Kuna. *Numerische Beanspruchungsanalyse von Rissen: Finite Elemente in der Bruchmechanik*. Vieweg+Teubner, Wiesbaden, 2008.

[22] P. Laborde, J. Pommier, Y. Renard, and M. Salaün. High order extended finite element method for cracked domains. *International Journal for Numerical Methods in Engineering*, vol. 64(n? 3):pp. 354–381, September 2005.

[23] Y. Lee and S. Im. On the computation of the near-tip stress intensities for three-dimensional wedges via two-state M-integral. *Journal of the Mechanics and Physics of Solids*, 51(5):825–850, 2003.

[24] D. Leguillon and E. Sanchez-Palencia. *Computation of singular solutions in elliptic problems and elasticity*. John Wiley & Sons, Inc., 1987.

[25] J. Lemaitre and J.-L. Chaboche. *Mechanics of solid materials*. Cambridge University Press, 1994.

[26] C. Mittelstedt and W. Becker. Efficient computation of order and mode of three-dimensional stress singularities in linear elasticity by the boundary finite element method. *International Journal of Solids and Structures*, 43(10):2868–2903, 2006.

[27] S. Mohammadi. *XFEM fracture analyses of composites*. John Wiley & Sons, Ltd, 2012.

[28] N. V. Movchan and J. R. Willis. Surface-breaking crack in an elastic half-space. *Journal of Engineering Mathematics*, 37(1-3):143–154, 2000.

[29] E. T. Ooi, H. Man, S. Natarajan, and C. Song. Adaptation of quadtree meshes in the scaled boundary finite element method for crack propagation modelling. *Engineering Fracture Mechanics*, 144:101–117, 2015.

[30] S. S. Pageau and S. B. Biggers. A finite element approach to three-dimensional singular stress states in anisotropic multi-material wedges and junctions. *International Journal of Solids and Structures*, 33(1):33–47, 1996.

[31] C. R. Picu and V. Gupta. Three-dimensional stress singularities at the tip of a grain triple junction line intersecting the free surface. *Journal of the Mechanics and Physics of Solids*, 45(9):1495–1520, 1997.

[32] N. Somaratna and T. C. T. Ting. Three-dimensional stress singularities in anisotropic materials and composites. *International Journal of Engineering Science*, 24(7):1115–1134, 1986.

[33] C. Song. Analysis of singular stress fields at multi-material corners under thermal loading. *International Journal for Numerical Methods in Engineering*, 65(5):620–652, 2006.

[34] C. Song and J. P. Wolf. The scaled boundary finite-element method—alias consistent infinitesimal finite-element cell method—for elastodynamics. *Computer Methods in Applied Mechanics and Engineering*, 147(3–4):329–355, 1997.

[35] P. Tong and T. H. H. Pian. On the convergence of the finite element method for problems with singularity. *International Journal of Solids and Structures*, 9(3):313–321, 1973.

[36] P. Tong, T. H. H. Pian, and S. J. Lasry. A hybrid-element approach to crack problems in plane elasticity. *International Journal for Numerical Methods in Engineering*, 7(3):297–308, 1973.

[37] G. Ventura, R. Gracie, and T. Belytschko. Fast integration and weight function blending in the extended finite element method. *International Journal for Numerical Methods in Engineering*, 77(1):1–29, 2009.

[38] M. L. Williams. Stress Singularities Resulting From Various Boundary Conditions. *Journal of Applied Mechanics*, 19(4):526–528, 1952.

[39] J. P. Wolf. *The scaled boundary finite element method*. John Wiley & Sons, 2003.

[40] Y. Yamada and I. Nishiguchi. Concept of general parametric elements and applications. *Computers & Structures*, 19(1–2):277–284, 1984. Special Memorial Issue.

[41] Z. Yang. Fully automatic modelling of mixed-mode crack propagation using scaled boundary finite element method. *Engineering Fracture Mechanics*, 73(12):1711–1731, 2006.

[42] J. R. Yeh and I. G. Tadjbakhsh. Stress Singularity in Composite Laminates by Finite Element Method. *Journal of Composite Materials*, 20(4):347–364, 1986.

[43] N. Zhang and P. F. Joseph. A nonlinear finite elment eigenanalysis of singular stress fields in bimaterial wedges for plane strain. *International Journal of Fracture*, 94(4):299–319, 1998.

[44] O. C. Zienkiewicz and R. L. Taylor. *The finite element method for solid and structural mechanics*. Butterworth-Heinemann, 2005.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# METHODOLOGY OF COMBINED APPLICATIONS OF DIRECTIONAL DERIVATIVES AND EXTENDED FINITE ELEMENT METHOD (X-FEM) FOR SOLVING VIBRATION EIGENVALUE PROBLEMS

**D. Serbichenko[1], P. Cosson[1] and P. Rozycki[1]**

[1] Institut de recherche en Génie Civil et Mécanique (GeM) - UMR CNRS 6183
Ecole Centrale de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3 (France)
e-mail: daria.serbichenko@ec-nantes.fr, {pascal.cosson,patrick.rozycki}@ec-nantes.fr

**Keywords:** Directional Derivatives; X-FEM; Vibrations; Eigenvalue problem.

***Abstract.*** *This paper presents a new methodology for solving the eigenvalue problem for time dependent structures. The time dependent structures of interest are structures with a moving discontinuity such as crack or structures with moving free (external/internal) surfaces. For the last case, they can result from a removal of material during a machining process or from a deterioration of the structure's geometry. The methodology that we developed, is based on a combination of the eXtended Finite Element Method (X-FEM) and the Directional Derivatives method. X-FEM enables to overcome the drawbacks of conformity and remeshing: indeed, using standard FEM, a moving discontinuity in time within a structure requires not only that the mesh must conform to the discontinuity geometry but also to fully remesh the structure as much as necessary to follow the discontinuity in time. In order to alleviate this last point, the directional derivatives are a powerful tool because they allow to estimate the evolution of quantities from on reference domain to another one. In our case, they will allow to estimate the solutions of the eigenvalue problem. We suggest on the first sections to remind the main keys of both methods and we present then the combined methods in order to solve an eigenvalue problem. The application will be done on a one-dimensional eigenvalue problem and the numerical results will be presented to demonstrate the accuracy and the advantages of selected approaches. We conclude on the future prospects of the current work that mainly consist of to develop the methodology at the second order in order to increase the accuracy and to find a criteria in order to automatize the combined methods.*

## 1 INTRODUCTION

The eigenvalue problems are of a great importance and play a crucial role in many engineering fields. In the aeronautical field for instance, the identification of natural frequencies and natural shapes during the design of a plane is compulsory before its commercialization to ensure the security of passengers. A modal analysis is usually done for the whole plane as well as for some structural parts such as engines or blades. In the field of machining, we can also cite the problem of quality insurance for the dimensions of machined pieces: the reliability of the processing machines must be ensured in order to avoid some critical interactions of their possible vibrations modes on the constraints of the dimensions of the piece. At last, in some other cases, the design of optimized structures to prevent failure due to instabilities and vibrations brings the problem of the determination of optimal physical parameters so that the load carrying capacity or the fundamental natural frequency is maximized.

Nowadays, there is a need for solving eigenvalue problems for time dependent structures. Changes of a structure can occur due to the propagation of cracks. They can also appear due to the fact that the shapes of a structure are modified because of the removal of material during its machining process or because of the deterioration of its geometry. They can also happen during design process when the structure has to be optimized with respect to frequency criteria. To investigate such issues, standard ways involve remeshing the structure in order to follow its changes and to conform to its current geometry. They also involve solving the eigenvalue problem for each configuration of the structure which has been meshed. Besides, we can also notice a possible loss of accuracy as the data is mapped from the old mesh to the new one. In other words, the FEM applications can lead to lot of limitations and complexities for engineers and can really be time consuming.

To offer an alternative solution to the above issues, we propose in this paper a new methodology combining two approaches. The first one is the Extended Finite Element Method (X-FEM) [1-2] which is based on the Partition of Unity [3]. The aim of this method is to alleviate the mesh constraints that come from discontinuities, using the level sets [4-7]. That is the main reason that X-FEM is effectively applied to solve many problems in material modeling, crack propagation problems [8] and structures with holes (or free surface) [9,10]. For our problem, the X-FEM method solves the mesh dependency of the structure due to the changing of boundaries or/and due to the discontinuities.

Since the simulation in order to follow the discontinuity in time and its consequence on the natural frequencies imposes to continuously compute the eigenvalue problem, it seems that this method is not quite optimal in terms of CPU time cost. In order to avoid computing the eigenvalue problem on each configuration where information is needed, we focus on a second approach: the Directional Derivatives [11-13]. This mathematical tool allows us to get the evolution of a quantity from one domain to another one. Applied to our problem, it becomes possible to estimate natural frequencies and natural shapes for a given configuration without solving the eigenvalue problem for this configuration.

In the next sections of this paper, we generalize joint application of the above techniques in one unique approach; we first introduce the state of eigenvalue problem, expose some key points of the X-FEM method and the basic definitions for the Directional Derivatives of first order in the general case. We then suggest applying the joint techniques (X-FEM and Directional Derivatives) for solving the eigenvalue problem in the one-dimensional case. In a last section, we discuss about the obtained results for the 1D case and we present the works and the main key-points for the application of the methodology in 2D case.

## 2 STATE OF THE PROBLEM

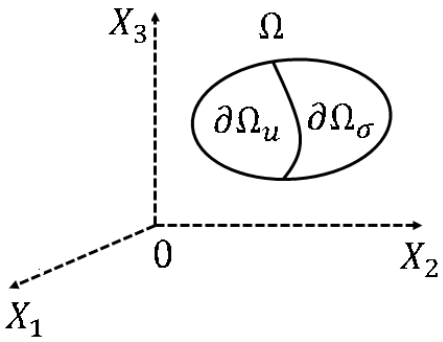



Figure 1: The eigenvalue problem associated with free oscillations.

The free-vibration natural frequencies and mode shapes of a linear structural system can be computed by solving the following eigenvalue problem. With the assumption of small perturbations in the framework of the linear elasticity theory, the free oscillations of a structure (Figure 1) are governed by the following equations system:

- **local equations**

$$div\ \bar{\bar{\sigma}}(u) = \rho\ddot{u} \tag{1}$$

- **behavior law**

$$\bar{\bar{\sigma}}(u) = \bar{\bar{\bar{\bar{E}}}} : \bar{\bar{\epsilon}}(u)\ \ with\ \ \bar{\bar{\epsilon}}(u) = \frac{\nabla u + \nabla^T u}{2} \tag{2}$$

- **boundary conditions**

$$u(X,t) = 0\ \forall X \in \partial\Omega_u \tag{3}$$

$$\bar{\bar{\sigma}}(u)\vec{n} = 0\ \forall X \in \partial\Omega_\sigma \tag{4}$$

Where:

- $u(X,t)$ is the displacement field;
- $\boldsymbol{\epsilon}(u)$ is the deformation field induced by $u(X,t)$ in the framework of the assumption of small perturbations;
- $\bar{\bar{\sigma}}(u)$ is the stress field induced by $u(X,t)$;
- $\partial\Omega_u$ is the part of boundary where the displacement field is known; on $\partial\Omega_u$, boundary conditions are essential or geometrical ones;
- $\partial\Omega_\sigma$ is the part of the boundary where the stress field is known; on $\partial\Omega_\sigma$, boundary conditions are natural ones;
- $\vec{n}$ is the external normal on the boundary $\partial\Omega_\sigma$;
- $\rho$ is the density;
- $\bar{\bar{\bar{\bar{E}}}}$ is elasticity tensor.

Usually, the solution of (1) is set in a following way:

$$u(X,t) = U(X) \cdot q(t)$$

where $U(X)$ is a displacement field (function of the variable $X$) and $q(t)$ is a scalar function of time. We obtain:

$$div\ \bar{\bar{\sigma}}(U)q(t) = \rho U\ddot{q} \tag{5}$$

For longitudinal vibrations of an uniform bar, the equation (5) becomes $EU_{,x^2}q(t) = \rho U\ddot{q}(t)$. Due to stability of solution in time, it is necessary that $\ddot{q} = -\omega^2 q$ with $\omega$ constant. That leads for longitudinal vibrations: $EU_{,x^2} = -\rho\omega^2 U$.

For the general case, we can write:

$$div\,\bar{\bar{\sigma}}(U) = -\rho\lambda U \;\; with \;\; \lambda \geq 0 \tag{6}$$

If the set of kinematically admissible displacement fields is defined by:

$$\mathcal{U}_{ad}(\Omega) = \left\{\widehat{W} \in H^1(\Omega) \,/\, \widehat{W}(X) = 0 \;\forall X \in \partial\Omega_u\right\}$$

The solutions of the eigenvalues problem $(\lambda, U^\lambda)$ satisfies the three following relationships:

$$\left(\lambda, U^\lambda\right) \in \mathbb{R}^+ \times \mathcal{U}_{ad}(\Omega) \quad ; \quad div\,\bar{\bar{\sigma}}\left(U^\lambda\right) = -\rho\lambda U^\lambda \quad ; \quad \bar{\bar{\sigma}}(U^\lambda(X))\vec{n} = 0 \;\forall X \in \partial\Omega_\sigma$$

and therefore $\forall\,\widehat{W} \in U_{ad}(\Omega)$, we obtain:

$$\int_\Omega \bar{\bar{\sigma}}\left(U^\lambda\right) : \bar{\bar{\epsilon}}\left(\widehat{W}\right) d\Omega = \lambda \int_\Omega \rho U^\lambda \,\widehat{W} d\Omega \tag{7}$$

Owing to the fact that the stress tensor is symmetric, we can write the relation (7) as:

$$\int_\Omega \bar{\bar{\sigma}}\left(U^\lambda\right) : \nabla\widehat{W} d\Omega = \lambda \int_\Omega \rho U^\lambda \,\widehat{W} d\Omega$$

The above relation corresponds to the standard variational formulation of the eigenvalue problem that occurs in Dynamics. We obtain that the solutions $(\lambda, U^\lambda)$ satisfy:

$$\begin{gathered}\left(\lambda, U^\lambda\right) \in \mathbb{R}^+ \times \mathcal{U}_{ad}(\Omega) \\ \int_\Omega \bar{\bar{\sigma}}\left(U^\lambda\right) : \nabla\widehat{W} d\Omega = \lambda \int_\Omega \rho U^\lambda \,\widehat{W} d\Omega \;\; \forall\,\widehat{W} \in \mathcal{U}_{ad}(\Omega)\end{gathered} \tag{8}$$

### 2.1 The X-FEM method

To solve the eigenvalue problem for time dependent structures, we use the X-FEM method which introduces a local enrichment of the approximation space in order to treat any kind of discontinuities within a structure (material discontinuity, cracks, free surfaces). These enrichment functions can include an analytical solution or an *a priori* knowledge of the experimental solution of test results. They are a combination of nodal functions that are associated with a mesh and the product of shape functions, which describe the discontinuity. This approach allows to independently model a discontinuity from the mesh. In the case of discontinuity's propagation, the method also avoids remeshing at each step. Besides, the enrichment functions are only added locally, i.e. when the domain requires to be enriched. As a result, the algebraic system of equations consists of two types of unknowns: standard degrees of freedom and enriched degrees of freedom. All the above features provide some important advantages to the X-FEM method compared to the standard FEM method for modeling arbitrary discontinuities.

The feature of X-FEM is to add special functions to some nodes of the global mesh domain. The purpose of these additional functions is to enrich the approximation of the existing displacement fields. The description of discontinuities in the framework of the X-FEM is

usually realized by the level-set method: it allows to locate the cutting elements by the level set and to determine the nodes of interest for which the additional degrees of freedom will be added. According to the method of partition of unity [3], the X-FEM displacement approximation $u^{XFEM}(x)$ is assumed to be as the following general form [6]:

$$u_{XFEM} = u_{FEM} + u_{enr} = \underbrace{\sum_{i \in I} N_i(x) u_i}_{FEM} + \underbrace{\sum_{j \in J} N_j(x) F(x) a_j}_{enrichment} \tag{9}$$

Where:

- $N_i$ are shape functions of node $i$;
- $u_i$ are unknown of finite element part at node $i$;
- $j \in J$ is nodal subset of the enrichment;
- $N_j$ are partition of unity functions of node $i$;
- $F(x)$ is enrichment function;
- $a_j$ are the additional degrees of freedom.

## 2.2 Definition of the changing of configuration in two-dimensional case

As it has been said, the changing of a structure can concern, for instance, the moving of its boundaries or the growth of a crack. So, for a given structure, each position of its boundaries, or each position of a crack, allows to define one configuration of this structure. For each configuration, we can associate one eigenvalue problem. Consequently, eigenvalues and eigenshapes will depend of the configuration of structure and will change as the configuration of the structure will evolves. The main benefit of X-FEM method lies in the fact that it becomes possible to follow the discontinuities within a structure without remeshing: the mesh dependency vanishes. But using merely X-FEM, the eigenvalue problem must still be solved on every configuration where an estimation of natural frequencies is needed.

To overcome the above shortcoming that can be time consuming, we propose to use the directional derivatives. They are a universal tool to avoid multiple calculations. It can be used to estimate the evolution of a quantity with respect to the change of configuration. In our case, the directional derivatives will provide a way to estimate the derivative of eigensolutions with respect to a change of configuration of the structure. Therefore, the eigenvalue problem will only be solved for some configurations until the accuracy has decreased drastically.

In order to take into account the time dependency of the structure, we introduce a function $q(X)$ which governs the changing of the shape from a reference configuration (noted $\Omega_0$) to a current one (noted $\Omega(\tau)$). $\tau$ is a scalar parameter which allows to follow the evolution of the structure and $X$ represents the position of a point on the reference configuration. If we note $x_\tau$ its position on the current configuration, we have (10):

$$x_\tau = X + \vec{q}(X) \cdot \tau \tag{10}$$

Concerning the function $\vec{q}(X)$ which governs the changing of configuration, the following remarks can be done:

- **Remark 1:**
  *The transformation between both configurations must be a bijective one.*
- **Remark 2:**
  *The transformation doesn't change the part of boundary where displacement conditions are imposed.*

- **Remark 3:**
  *It should also be noted that the changing of configuration characterizes a changing of geometry in the initial considered domain, but the material remains the same. The density and the elasticity tensor for both configurations are the same.*

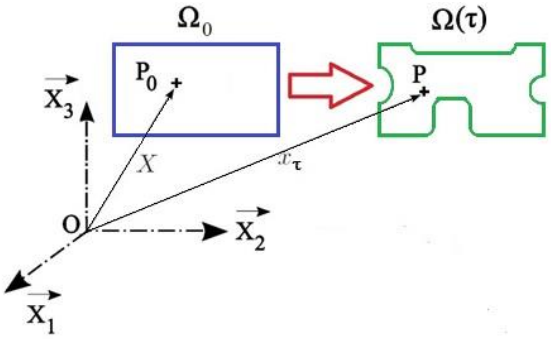



Figure 2: The transformation from $\boldsymbol{\Omega_0}$ to $\boldsymbol{\Omega(\tau)}$.

The transformation's gradient from the reference configuration to the current one is defined by (11):

$$\bar{\bar{F}} = \nabla \vec{x}_\tau = \mathbb{I} + \tau \nabla \vec{q}\left(\vec{X}\right) \tag{11}$$

The determinant of the gradient transformation can be obtained in the following way:

$$det\bar{\bar{F}} = \begin{vmatrix} 1+\tau q_{1,1} & \tau q_{1,2} \\ \tau q_{2,1} & 1+\tau q_{2,2} \end{vmatrix} = 1 + \tau div\vec{q} + \frac{\tau^2}{2}((div\vec{q})^2 - \nabla^T\vec{q} : \nabla\vec{q}) \tag{12}$$

A displacement field $v$ on the current configuration can be regarded as a displacement field on the reference configuration. We have:

$$v(x_\tau) = v(X + \vec{q}(X) \cdot \tau) = V(\tau, X) \tag{13}$$

It follows that:

$$grad\ v(x_\tau) = \nabla V(\tau, X) \cdot \bar{\bar{F}}^{-1} \tag{14}$$

From relation (11), it can be readily deduced that for small values of parameter $\tau$, we have:

$$\bar{\bar{F}}^{-1} = \mathbb{I} - \tau \nabla \vec{q}\left(\vec{X}\right) + \tau^2 \nabla \vec{q}\left(\vec{X}\right) \nabla \vec{q}\left(\vec{X}\right) + \tau^2 \eta(\tau)\ \ with\ \lim_{\tau \to 0} \|\eta(\tau)\| = 0 \tag{15}$$

### 2.3 Expressions of the directional derivatives for eigenvalues and eigenshapes

The solution of eigenvalue problem, illustrated on Figure 2 for initial and current configurations, are noted (16):

$$\begin{gathered} \left(\lambda^i, U^i(X)\right) \in \mathbb{R}^+ \times U_{ad}(\Omega_0) \\ \left(\lambda^i(\tau), u_\tau^i(x_\tau)\right) \in \mathbb{R}^+ \times U_{ad}\left(\Omega(\tau)\right)\ \ with\ \ u_\tau^i(x_\tau) = u^i(\tau, X) \end{gathered} \tag{16}$$

Where $U_{ad}(\Omega_0)$ and $U_{ad}\left(\Omega(\tau)\right)$ are the sets of admissible functions for both reference and current configurations. With these definitions, the directional derivatives of eigenvalue $D\lambda^i[q]$ and eigenshape $Du^i[q]$ for the first order are given by:

$$D\lambda^i[\vec{q}] = \lim_{\tau \to 0} \frac{\lambda^i(\tau) - \lambda^i}{\tau} \tag{17}$$

$$Du^i[\vec{q}] = \lim_{\tau \to 0} \frac{u^i(\tau, X) - U^i(X)}{\tau} \tag{18}$$

Starting from expressions (17) and (18), it becomes possible to obtain an estimation of eigenvalues and eigenshapes for the current configuration as soon as the directional derivatives are known. We have:

$$\lambda^i(\tau) = \lambda^i + \tau D\lambda^i[\vec{q}] \tag{19}$$

$$u^i(\tau, X) = U^i(X) + \tau Du^i[\vec{q}] \tag{20}$$

The both above expressions represent the connection between previous domain and current one and express the eigenvalues at the current time from the reference time as well as the eigenshape $u^i(\tau, X)$, which depends on the parameter $\tau$ and the space. More precisely, as it can be noted, the eigenvalues and the eigenshapes on the current configuration are written in terms of the solution of the eigenvalue problem on the reference configuration and its directional derivatives, both quantities being computing on the reference configuration. Therefore, knowing the solution of the eigenvalue problem on the reference configuration and its directional derivatives, the solution of the eigenvalue problem on the current configuration can be approximated, according to (19) and (20), by varying the value of parameter $\tau$.

### 2.4 Variational formulation of eigenvalue problem in two-dimensional case

To obtain the expression of the directional derivatives of eigenvalues and eigenshapes, one can start from the variational formulation of the eigenvalue problem that has to be solved in Dynamics. On the initial configuration, this variational formulation can be set as follows:

$$\int_{\Omega_0} \bar{\bar{\sigma}}\left(U^i(X)\right) : \nabla \widehat{W} d\Omega_0 = \lambda^i \int_{\Omega_0} \rho U^i(X)\, \widehat{W} d\Omega_0$$
$$\forall \widehat{W} \in U^*_{ad}(\Omega_0) \; and \; U^*_{ad}(\Omega_0) = \left\{\widehat{W} \in H^1(\Omega_0) / \widehat{W}(X) = 0 \; \forall X \in \partial\Omega_u(0)\right\} \tag{21}$$

The variational formulation of the eigenvalue problem on the current configuration takes the form:

$$\int_{\Omega(\tau)} \bar{\bar{\sigma}}\left(u^i_\tau(x_\tau)\right) : \nabla \widehat{w} d\Omega = \lambda^i(\tau) \int_{\Omega(\tau)} \rho u^i_\tau(x_\tau)\, \widehat{w} d\Omega$$
$$\forall \widehat{w} \in U^*_{ad}\big(\Omega(\tau)\big) \; and \; U^*_{ad}(\Omega(\tau)) = \left\{\widehat{w} \in H^1\big(\Omega(\tau)\big) / \widehat{w}(x_\tau) = 0 \; \forall x_\tau \in \partial\Omega_u(\tau)\right\} \tag{22}$$

As the function doesn't change the part of the boundary where displacements are imposed, we have:

$$\forall \widehat{w} \in U^*_{ad}\big(\Omega(\tau)\big) \Leftrightarrow \forall \widehat{W} \in U^*_{ad}(\Omega_0)$$

Starting from the relation (22), using the relation (10) between $x_\tau$ and $X$ and taking into consideration the relation (14), one can note that solution $\left(\lambda^i(\tau), u(\tau, X)\right)$ of the eigenvalue problem on the current configuration must satisfy:

$$\int_{\Omega_0} \bar{\bar{\sigma}}\left(u^i(\tau,X)\right) : (\nabla\widehat{W}\bar{\bar{F}}^{-1}) det\bar{\bar{F}} d\Omega_0 = \lambda^i(\tau) \int_{\Omega_0} \rho u^i(\tau,X)\widehat{W} det\bar{\bar{F}} d\Omega_0 \;\; \forall \widehat{W} \in U^*_{ad}(\Omega_0)$$
$$\bar{\bar{\sigma}}\left(u^i(\tau,X)\right) = \bar{\bar{\bar{\bar{E}}}} : \left(\nabla u^i(\tau,X)\bar{\bar{F}}^{-1}\right)^s \tag{23}$$

It has to be noticed that all integrals of the relation (23) are established on the reference configuration. Therefore, the solution of eigenvalue problem ( $\lambda^i(\tau), u^i(\tau,X)$ ) for configuration $\Omega(\tau)$ is an element of the set $\mathbb{R}^+ \times U^*_{ad}(\Omega_0)$ satisfying for any element $\widehat{W}$ of $U^*_{ad}(\Omega_0)$:

$$\underbrace{\int_{\Omega_0} \left[\bar{\bar{\bar{\bar{E}}}} : \left(\nabla u^i(\tau,X)\bar{\bar{F}}^{-1}\right)^s\right] : (\nabla\widehat{W}\bar{\bar{F}}^{-1}) det\bar{\bar{F}} d\Omega_0}_{A(\tau)} = \underbrace{\lambda^i(\tau) \int_{\Omega_0} \rho u^i(\tau,X)\, \widehat{W} det\bar{\bar{F}} d\Omega_0}_{B(\tau)} \tag{24}$$

In the neighborhood of $\tau = 0$, we can write $A(\tau)$ and $B(\tau)$ as an expansion in series of the variable $\tau$. By taking into account only the zero and first orders, we obtain:

$$A(\tau) = A(0) + \left.\frac{dA}{d\tau}\right|_{\tau=0} \cdot \tau + O(\tau^2) \tag{25}$$

$$B(\tau) = B(0) + \left.\frac{dB}{d\tau}\right|_{\tau=0} \cdot \tau + O(\tau^2) \tag{26}$$

The identification of terms of order zero, namely $A(0) = B(0)$, leads to the definition of the eigenvalue problem on the reference configuration. The equality of terms of first order leads to:

$$\left.\frac{dA}{d\tau}\right|_{\tau=0} = \left.\frac{dB}{d\tau}\right|_{\tau=0}$$

The derivative of $A(\tau)$, with respect to the parameter $\tau$, is a sum of four terms, each one depending of $\tau$:

$$\begin{aligned}\frac{dA(\tau)}{d\tau} = &\underbrace{\int_{\Omega_0} \left[\bar{\bar{\bar{\bar{E}}}} : \left(\nabla \frac{\partial u^i(\tau,X)}{\partial\tau}\bar{\bar{F}}^{-1}\right)^s\right] : \left(\nabla\widehat{W}\bar{\bar{F}}^{-1}\right) det\bar{\bar{F}} d\Omega_0}_{a_1(\tau)} + \\ &\underbrace{\int_{\Omega_0} \left[\bar{\bar{\bar{\bar{E}}}} : \left(\nabla u^i(\tau,X)\frac{\partial\bar{\bar{F}}^{-1}}{\partial\tau}\right)^s\right] : (\nabla\widehat{W}\bar{\bar{F}}^{-1}) det\bar{\bar{F}} d\Omega_0}_{a_2(\tau)} + \\ &\underbrace{\int_{\Omega_0} \left[\bar{\bar{\bar{\bar{E}}}} : \left(\nabla u^i(\tau,X)\bar{\bar{F}}^{-1}\right)^s\right] : \left(\nabla\widehat{W}\frac{\partial\bar{\bar{F}}^{-1}}{\partial\tau}\right) det\bar{\bar{F}} d\Omega_0}_{a_3(\tau)} +\end{aligned} \tag{27}$$

$$\underbrace{\int_{\Omega_0}\left[\bar{\bar{\bar{E}}}:\left(\nabla u^i(\tau,X)\bar{\bar{F}}^{-1}\right)^s\right]:\left(\nabla\widehat{W}\bar{\bar{F}}^{-1}\right)\frac{\partial det\bar{\bar{F}}}{\partial\tau}d\Omega_0}_{a_4(\tau)}$$

The derivative of $B(\tau)$, with respect to the parameter $\tau$, is a sum of three terms, each one also depending of variable $\tau$:

$$\frac{dB(\tau)}{d\tau}=\underbrace{\frac{\partial\lambda^i(\tau)}{\partial\tau}\int_{\Omega_0}\rho u^i(\tau,X)\,\widehat{W}det\bar{\bar{F}}d\Omega_0}_{b_1(\tau)}+$$

$$\underbrace{\lambda^i(\tau)\int_{\Omega_0}\rho\,\frac{\partial u^i(\tau,X)}{\partial\tau}\widehat{W}det\bar{\bar{F}}d\Omega_0}_{b_2(\tau)}+\underbrace{\lambda^i(\tau)\int_{\Omega_0}\rho u^i(\tau,X)\,\widehat{W}\frac{\partial det\bar{\bar{F}}}{\partial\tau}d\Omega_0}_{b_3(\tau)} \tag{28}$$

Starting from expression (27), using relations (11), (12), (14), (15) and (20), we finally obtain, by taking the limits when $\tau$ vanishes:

$$\left.\frac{dA}{d\tau}\right|_{\tau=0}=\int_{\Omega_0}\left[\bar{\bar{\bar{E}}}:\left(\nabla Du^i[\vec{q}]\right)^s\right]:\left(\nabla\widehat{W}\right)d\Omega_0+$$
$$\int_{\Omega_0}\left[-\bar{\bar{\bar{E}}}:\left(\nabla U^i\nabla\vec{q}\right)^s-(\bar{\bar{\bar{E}}}:\left(\nabla U^i\right)^s)\nabla^T\vec{q}+(\bar{\bar{\bar{E}}}:\left(\nabla U^i\right)^s)div\vec{q}\right]:\left(\nabla\widehat{W}\right)d\Omega_0 \tag{29}$$

In a similar way, starting from expression (28), using relations (12), (19) and (20), it can been shown, by taking the limits when $\tau$ vanishes, that:

$$\left.\frac{dB}{d\tau}\right|_{\tau=0}=D\lambda^i[\vec{q}]\int_{\Omega_0}\rho U^i\,\widehat{W}d\Omega_0+\lambda^i\int_{\Omega_0}\rho Du^i[\vec{q}]\,\widehat{W}d\Omega_0$$
$$+\lambda^i\int_{\Omega_0}\rho U^i\,\widehat{W}div\vec{q}d\Omega_0 \tag{30}$$

Consequently, the expressions (29) and (30) of $\left.\frac{dA}{d\tau}\right|_{\tau=0}$ and $\left.\frac{dB}{d\tau}\right|_{\tau=0}$ lead to the following relation between the directional derivatives of eigenvalue and eigenshape. We have:

$$\int_{\Omega_0}\left[\bar{\bar{\bar{E}}}:\left(\nabla Du^i[\vec{q}]\right)^s\right]:\left(\nabla\widehat{W}\right)d\Omega_0-D\lambda^i[\vec{q}]\int_{\Omega_0}\rho U^i\,\widehat{W}d\Omega_0$$
$$-\lambda^i\int_{\Omega_0}\rho\,Du^i[\vec{q}]\;\widehat{W}d\Omega_0$$
$$=\int_{\Omega_0}\left[\bar{\bar{\bar{E}}}:\left(\nabla U^i\nabla\vec{q}\right)^s+\left(\bar{\bar{\bar{E}}}:\left(\nabla U^i\right)^s\right)\nabla^T\vec{q}\right]:\left(\nabla\widehat{W}\right)d\Omega_0$$
$$+\int_{\Omega_0}\left[-\left(\bar{\bar{\bar{E}}}:\left(\nabla U^i\right)^s\right):\left(\nabla\widehat{W}\right)+\lambda^i\rho U^i\widehat{W}\right]div\vec{q}\;d\Omega_0 \tag{31}$$

Clearly, the relation (8) defines an eigenvalue problem for a self-adjoint system. The self-adjointness comes from the symmetry of integrals and the nullity of boundary conditions. It can be ascertained through integrations by parts. It follows from this self-adjointness that the

eigenfunctions $U^i(X)$ constitute a complete orthogonal set of infinite dimension [14] which we use as basis for $U_{ad}(\Omega_0)$. Consequently, using the expansion theorem for self-adjoint systems, the directional derivative of eigenshape on the reference configuration can be set as:

$$Du^i[\vec{q}] = \sum a_{ki} U^k(X) \tag{32}$$

In relation (32), it should be pointed out that the series $\sum a_{ki} U^k(X)$ converge in energy to $Du^i[\vec{q}]$. The expressions of the directional derivatives for eigenvalues and eigenshapes can be deduced from relation (31) using the expansion (32) of $Du^i[\vec{q}]$ and the orthogonality relations between eigenshapes on the reference configuration.

## 3 RESULTS

This section deals with the numerical calculation of the eigenvalue problem for a clamped-free bar which is illustrated on Figure 3. The length of the bar which contains two different sections is named $L$. It is discretized using 200 finite elements. In this one-dimensional case, the discontinuity corresponds to the change of the cross section of the bar: in this particular case, the level set is reduced to one unique point. It follows that the enrichment function has to be built for only one element (the one containing the cross section discontinuity).

The used material is steel for which Young's modulus $E$ is equal to $2.06\ 10^{11}\ Pa$ and the density is defined by $\rho = 7860\ kg/m^3$. The length of the bar $L$ is $2\ m$. The initial position of the discontinuity (change of cross section) is noted $X_0$. The sections $S_1$ and $S_2$ are equal respectively to $12.566\ 10^{-4}\ m^2$ and $8.0425\ 10^{-4}\ m^2$. Our purpose is to obtain an estimation of natural frequencies and natural shapes when the position of the discontinuity moves, that is to say when the size of each part is varied, the whole length of the bar staying unchanged.

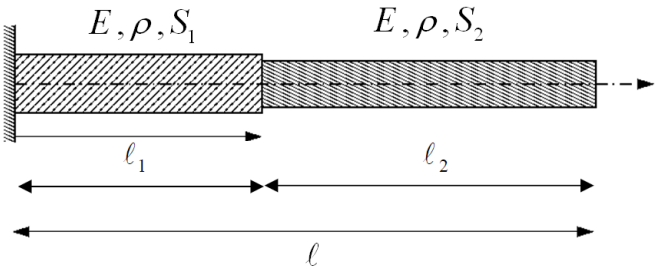



Figure 3: Bar with a discontinuity of the cross section.

For our problem, the bar is a free-clamped one. As the length of the whole remains constant, we have to consider a function $q(X)$ satisfying $q(0) = q(L) = 0$.

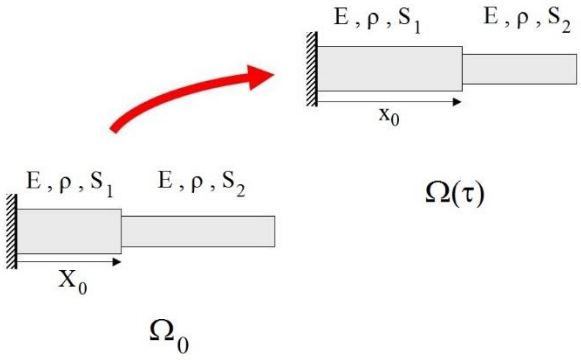



Figure 4: The transformation for one-dimensional case.

The calculation of the directional derivatives of the eigenvalues and the eigenshapes was carried out for one-dimensional case. Starting from the relations (19) and (22) we have:

$$\int_0^L ES\frac{\partial}{\partial X}\left(Du^i[q]\right)\frac{\partial \widehat{W}}{\partial X}dX - \lambda^i\int_0^L \rho S Du^i[q]\widehat{W}(X)dX - D\lambda^i[q]\int_0^L \rho S u^i(X)\widehat{W}(X)dX = \\ = \int_0^L \left[ES\frac{\partial U^i(X)}{\partial X}\frac{\partial \widehat{W}}{\partial X} + \lambda^i \rho S U^i(X)\widehat{W}(X)\right]q'(X)dX$$

As the above expression is verified, $\forall \widehat{W} \in \mathrm{U}_{\mathrm{ad}}^{*}(\Omega_0)$, we obtain, after discretization:

$$\left([K] - \lambda^i(\tau)[\mathrm{M}]\right)\{Du^i[q]\}_{EF} - D\lambda^i[q][\mathrm{M}]\{U^i\}_{EF} \\ = \left([\mathrm{K}Dq] + \lambda^i(\tau)[\mathrm{M}Dq]\right)\left\{U^i\right\}_{EF} \tag{33}$$

Where:

- $[\mathrm{K}] = \sum_k \int_{element\ k} ES\left[\frac{\partial N}{\partial \xi}\right]^T\left[\frac{\partial N}{\partial \xi}\right]dX$
- $[\mathrm{M}] = \sum_k \int_{element\ k} \rho S[N(\xi)]^T[N(\xi)]dX$
- $[\mathrm{K}Dq] = \sum_k \int_{element\ k} ES\left[\frac{\partial N}{\partial \xi}\right]^T\left[\frac{\partial N}{\partial \xi}\right]q'(X)dX$
- $[\mathrm{M}Dq] = \sum_k \int_{element\ k} \rho S[N(\xi)]^T[N(\xi)]q'(X)dX$

Due to the orthogonality of eigenshapes and after pre-multiplying left and right sides of expression (35) by $\left\{\mathrm{U}^{\mathrm{i}}\right\}_{\mathrm{EF}}^{\mathrm{T}}$, we obtain:

$$D\lambda^i[q] = -\frac{\left\{U^i\right\}_{EF}^T\left([\mathrm{K}Dq] + \lambda^i(\tau)[\mathrm{M}Dq]\right)\left\{U^i\right\}_{EF}}{\{U^i\}_{EF}^T[\mathrm{M}]\{U^i\}_{EF}} \tag{34}$$

Owing to relation (32), we assume a directional derivative $\left\{\mathrm{Du}^{\mathrm{i}}[\mathrm{q}]\right\}_{\mathrm{EF}}$ in the form (35).

$$\left\{Du^i[q]\right\}_{EF} = \sum_{k\neq i}\alpha_k^i\{U^k(X)\}_{EF} \tag{35}$$

Starting from the relation (33), introducing the expression (35) of the directional derivative $\left\{\mathrm{Du}^{\mathrm{i}}[\mathrm{q}]\right\}_{\mathrm{EF}}$, pre-multiplying by $\left\{\mathrm{U}^{\mathrm{k}}\right\}_{\mathrm{EF}}^{\mathrm{T}}$ with $\mathrm{k} \neq \mathrm{i}$ and using the orthogonality relation between eigenshapes on the reference configuration leads to:

$$\alpha_k^i = \frac{\{U^k\}_{EF}^T\left([\mathrm{K}Dq] + \lambda^i(\tau)[\mathrm{M}Dq]\right)\left\{U^i\right\}_{EF}}{\{U^k\}_{EF}^T[\mathrm{M}]\{U^k\}_{EF}\cdot(\lambda^k - \lambda^i)} \tag{36}$$

Thus, the approximate expression of $\left\{\mathrm{u}^{\mathrm{i}}(\tau, \mathrm{X})\right\}_{\mathrm{EF}}$ can be written as:

$$\left\{u^i(\tau, X)\right\}_{EF} = \left\{U^i\right\}_{EF} + \tau\cdot\sum_{k\neq i}\alpha_k^i\{U^k\}_{EF} \tag{37}$$

For a bar of two different sections, the exact solution of the eigenvalue problem is given by:

$$u_1^i(X) = \sin W_i X \ \ for\ 0 \le X \le l_1$$

$$u_2^i(X) = A_2^i \cos W_i X + \ B_2^i \sin W_i X \ \ for\ 0 \le X \le l_2$$

where $A_2^i = \sin W_i l_1$ and $B_2^i = \frac{S_1}{S_2}\cos W_i l_1$. $W_i$ satisfy:

$$\cos W_i(l_1 + l_2) + \frac{S_1 - S_2}{S_1 + S_2}\cos W_i(l_1 - l_2) = 0$$

It leads the opportunity to compare exact solution of this problem and the solution obtained with the proposed method.

For studying the dependence of the eigenvalues and eigenshapes with respect to the position of the discontinuity, we have to take a function $q(X)$ which allows to move the position of the discontinuity, the total length of the bar is remained to be constant. Consequently, the function $q(X)$ has to satisfy:

- $q(0) = 0$ for the fixed extremity;
- $q(L) = 0$ in order to impose the total length of the bar unchanged;
- $q(X_0) = 1$ if we have noted $X_0$ the initial position of the discontinuity.

We here present the results that we have obtained by choosing a polynomial piecewise function for $q(X)$. Two initial positions of the discontinuity have been studied:

$$X_0 = 0.9854m \quad \text{and} \quad X_0 = 1.1854m$$

For these two cases, the functions $q(X)$ that we have considered are presented on the Figure 5.

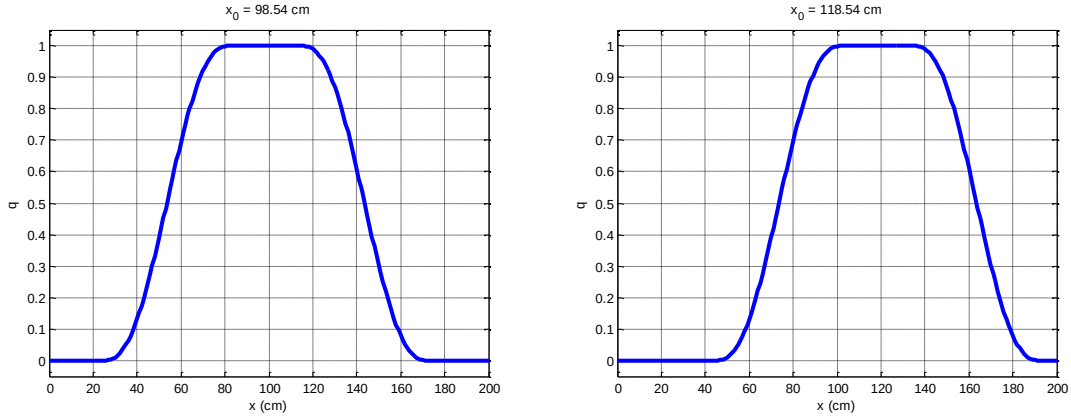



Figure 5: Functions **q**(**X**) for both initial position cases.

The estimated values of natural frequencies have been obtained owing to relation (19). As our formulation only uses directional derivatives of first order, the approximation of the eigenvalue is linear. Directional derivatives have been computed according to relation (34). The estimated values have been compared to exact ones. To compare both solutions, the frequency deviation can be used. It corresponds to the relative error between both methods. For the $i^{th}$ natural frequency, this frequency deviation is defined by

$$\left| \frac{f_{i,estimated} - f_{i,exact}}{f_{i,exact}} \right|$$

The graphs of Figure 6, Figure 7 and Figure 8 show the evolution of the first three frequencies with respect to the position of the discontinuity. For each natural frequency, the exact values and estimated ones have been represented for two positions of the initial discontinuity.

$$X_0 = 0.9854m \quad \text{and} \quad X_0 = 1.1854m$$

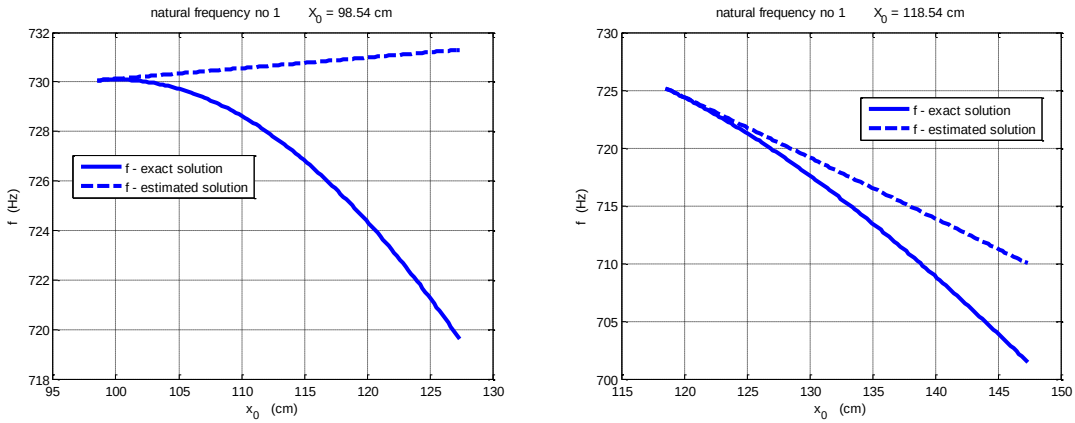



Figure 6: Dependence of the first natural frequency on position of discontinuity (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

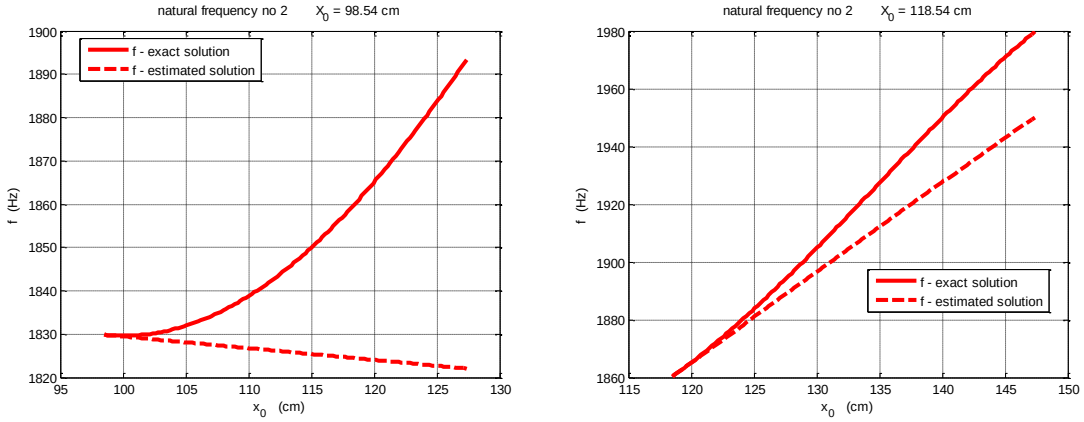



Figure 7: Dependence of the second natural frequency on position of discontinuity (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

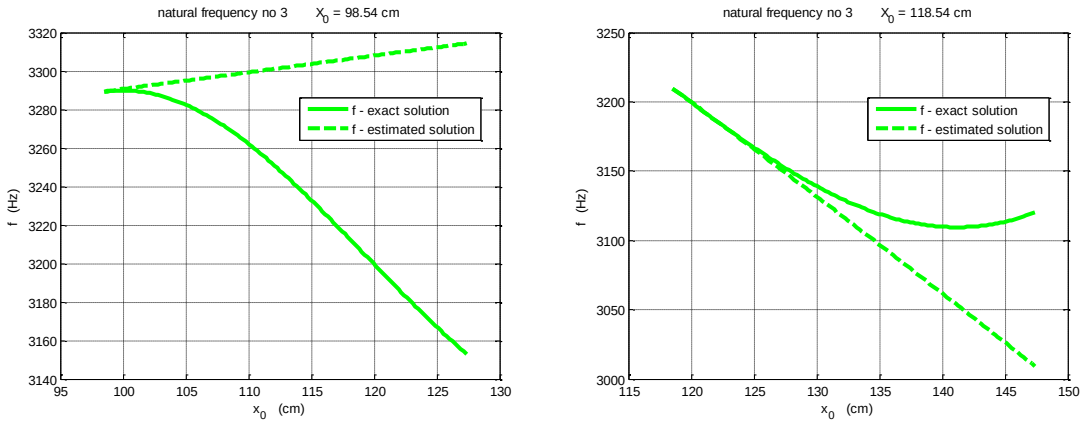



Figure 8: Dependence of the third natural frequency on position of discontinuity (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

As seen in these graphs, the curves of exact and estimated solutions are very close. But the exact solution has a parabolic shape. The approximate solution, obtained with relation (19), is linear. On these curves, it can be observed that when directional derivatives of eigenvalue are close to zero (for instance when $X_0 = 0.9854m$), the gap between exact natural frequencies

and estimated ones increases rapidly. It provides from the fact that according to relation (19), our approximation is linear. Therefore, when directional derivatives are close to zero, the accuracy is weak. In order to overcome this drawback, it would be useful to calculate directional derivatives of second order which would allow to define a parabolic approximation of eigenvalues and eigenshapes. In the second case, when $X_0 = 1.1854m$, directional derivatives of eigenvalues are not close to zero and the gap between both solutions decreases significantly. In this second case, the linear approximation provides more accurate results.

Graphs of the Figure 9 show the frequency deviation for the three first natural frequencies when directional derivatives are calculated in $X_0 = 1.1854m$. The relative error is less than 3% for the three natural frequencies.

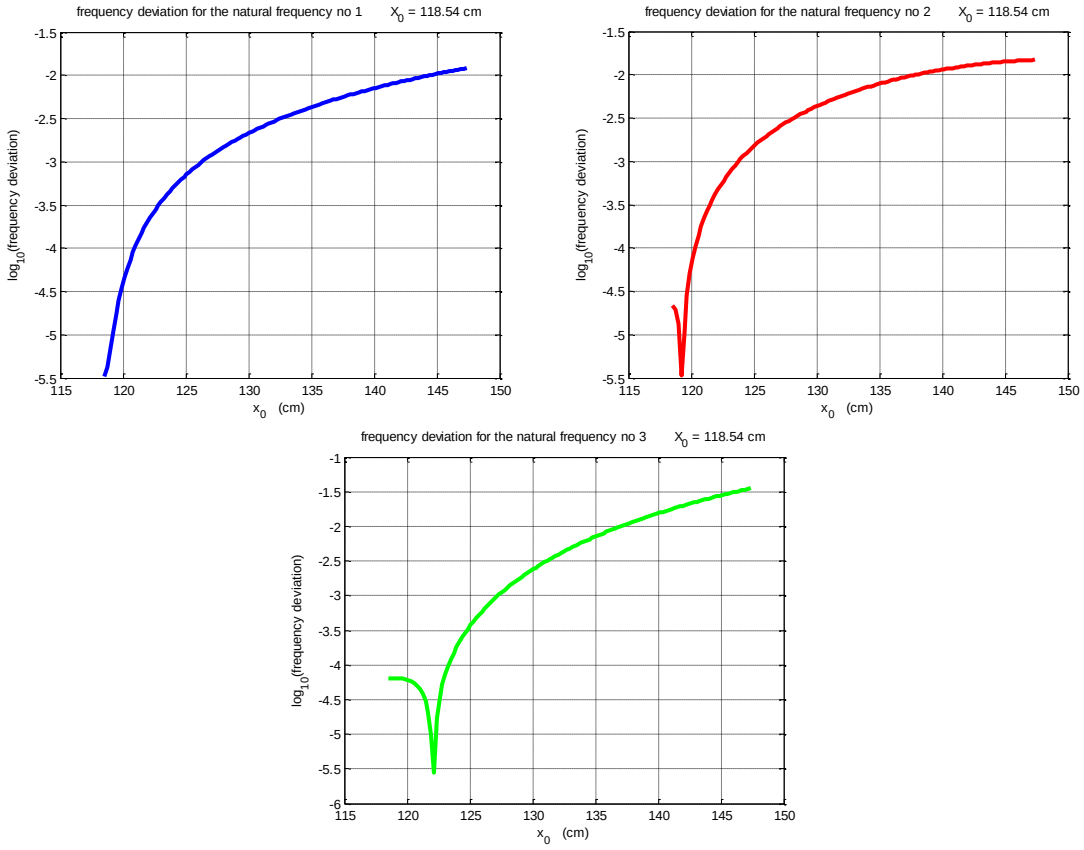



Figure 9: Frequency deviation for the three first natural frequencies when directional derivatives are calculated for $\boldsymbol{X_0 = 1.1854m}$.

The evolution of the three first eigenshapes is given Figure 10 for several values of parameter $\tau$. Exact eigenshapes and estimated ones are represented. The estimated eigenshapes have been obtained according to relation (20). Just like for the eigenvalues, our approximation for eigenshapes is linear. The directional derivatives $Du^i[q]$ are calculated with 20 modes.

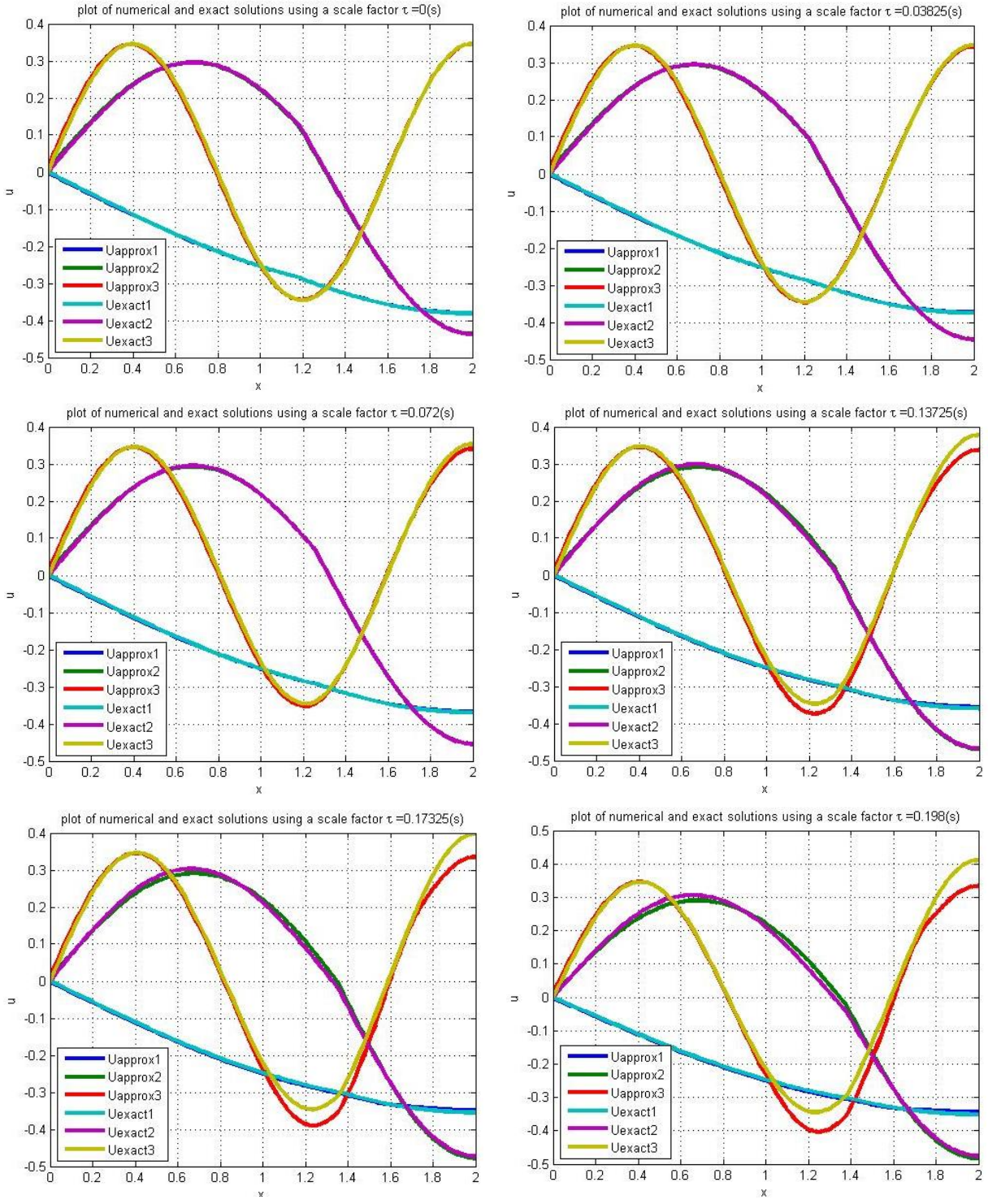



Figure 10: Comparison of exact and approximate solutions.

For the eigenshapes, to make an accurate comparison between estimated values and theoretical ones, we use the Modal Assistance Criterium. When we compare the estimated and theoretical eigenshapes of the same order, the MAC is close to one and decreases as τ grows up. On top of that, the MAC between estimated and theoretical eigenshapes of different order is nearly to zero. From these two results, we can deduce that

- the correlation between exact eigenshapes and estimated ones is good;
- orthogonality properties are verified between estimated eigenshapes.

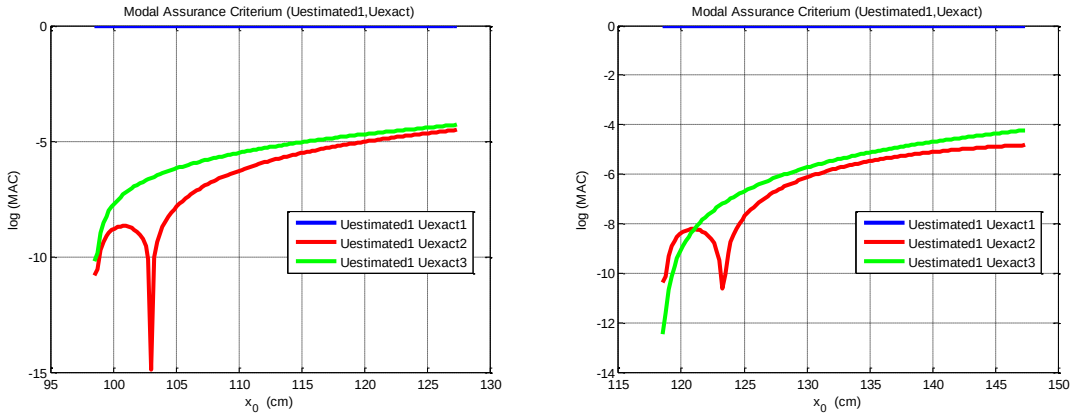



Figure 11: Modal Assistance Criterium for the first natural shape (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

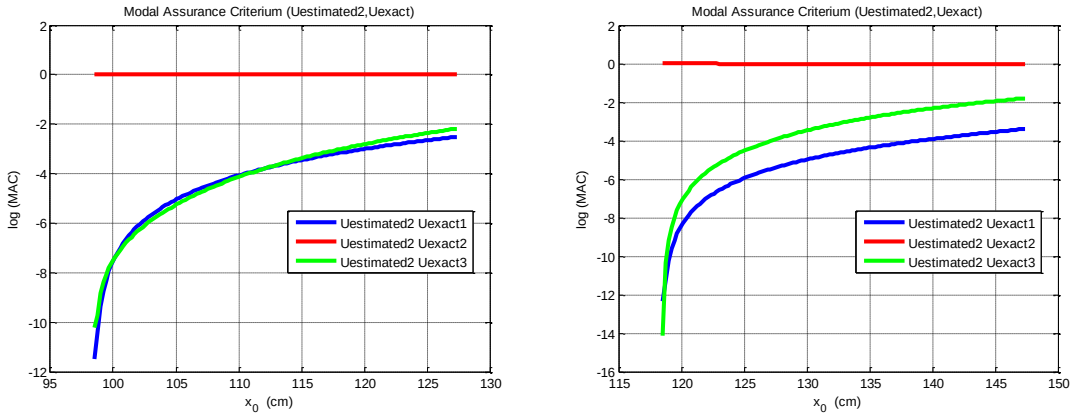



Figure 12: Modal Assistance Criterium for the second natural shape (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

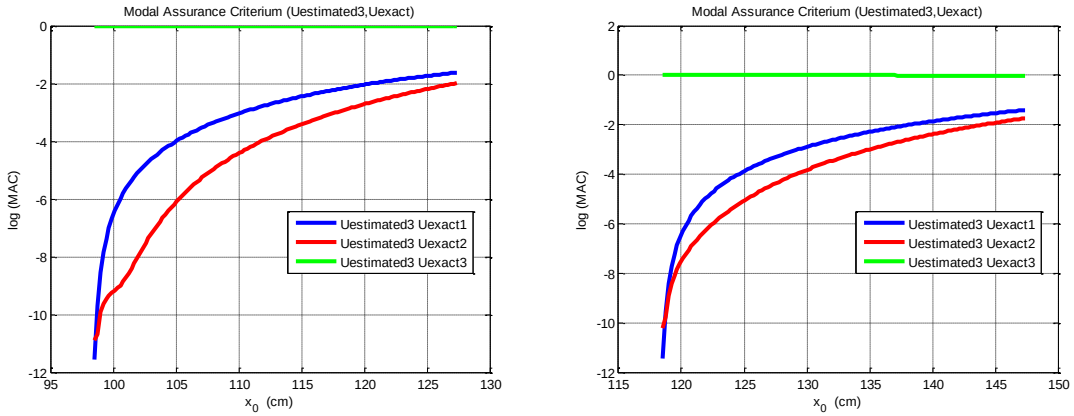



Figure 13: Modal Assistance Criterium for the third natural shape (directional derivatives are calculated for $\boldsymbol{X_0 = 0.9854m}$ in the first case and $\boldsymbol{X_0 = 1.1854m}$ in the second case).

## 4 CONCLUSION

A methodology for solving the eigenvalue problem for a time dependent structures has been proposed. This methodology combines both the directional derivatives and X-FEM methods. The convenience of the use of X-FEM is the possibility to avoid remeshing. A discontinuity can be depicted with a Level Set function. This Level Set function is used to define a local enrichment for taking into account this discontinuity. Consequently, due to enrichment, the mesh is not mandatory to be conform to the geometry. The application of directional derivative allows to avoid many calculations at every time step and gives an approximation of the solutions of the eigenvalue problem. This study focuses on the fact that directional derivatives is a convenient tool for following the evolution of natural frequencies and shapes.

The main idea of our approach consisted in obtaining the expression of derivatives for eigenvalues and eigenshapes in a given direction which is described by a function $q(X)$. This function determines the transformation from a reference configuration to a current one. Starting from the variational formulation of the eigenvalue problem on the current configuration, derivatives of eigenvalues and eigenshapes can be expressed in terms of the solutions of the eigenvalue problem for the reference domain by the mean of this function $q(X)$. Once we have computed the directional derivatives of eigenvalues and eigenshapes, it becomes possible to evaluate the evolution of these values when the domain is modified. To obtain solutions at the required time step (that is to say on the current configuration), it is sufficient to know information from the initial one. The main advantage of this procedure lies in the fact that numerical calculations are done without time consuming.

The numerical results for a one-dimensional structure varying in time and the accuracy of the new technique were presented. Geometric transformation of the eigenvalue problem for one-dimensional element with discontinuity was defined. To compute the natural shapes for a current configuration, the directional derivatives of natural shapes have been performed on the reference configuration, these directional derivatives have been set in terms of natural shapes of the reference configuration.

Using the general methodology which combines of the offered methods gives possibility to illustrate the application of this technique on a two-dimensional eigenvalue problem for a plate with shapes which are moving. One of the aspects for future work is to look upon the expressions of directional derivatives of eigenvalue problem for second order to improve the existing linear approximation. Another issue which has to be solved is to find a criteria to determine when accuracy has decreased significantly which means that directional derivatives have to be computed. The numerical implementation will be tested on on real simple and more complex cases.

## REFERENCES

[1] T. Belytschko and T. Black, Elastic crack growth in finite elements with minimal re-meshing. *International journal for numerical methods in engineering*, **45**, 601-620, 1999.

[2] N.Moës, John Dolbow, and T. Belytschko, A finite element method for crack growth without remeshing. *International journal for numerical methods in engineering*, **46**, 131-150, 1999.

[3] J.M. Melenk and I. Babuska, The partition of unity finite element method: basic theory and applications. *Computer Methods in Applied Mechanics and Engineering*, **139**, 289-314, 1996.

[4] S. Osher and J. Sethian, Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, **79**, 712-749, 1988.

[5] M. Duflot, A study of representation of crack with level sets. *International journal for numerical methods in engineering*, **70**, 1261-1302, 2007.

[6] S. Bordas and B. Moran, Enriched finite elements and level sets for damage tolerance assessment of complex structures. *Engineering Fracture Mechanics*, **73**, 1176–1201, 2006.

[7] M. Stolarska, D. Chopp, Modeling thermal fatigue cracking in integrated circuits by level sets and the extended finite element method. *International Journal of Engineering Science*, **41**, 2381–2410, 2003.

[8] D.L. Chopp, N. Sukumar, Fatigue Crack Propagation of Multiple Coplanar Cracks with the Coupled Extended Finite Element/Fast Marching Method. *International Journal of Engineer-ing Science*, **41** (8), 845–869, 2003.

[9] N. Sukumar, D. Chopp, N. Moës and T. Belytschko, Modelling holes and inclusions by level sets in the extended finite-element method. Computer Methods in Applied Mechanics and Engineering 190, 6183-6200, 2001.

[10] N. Moës, M. Cloirec, P. Cartraud, J.F. Remacle, A computational approach to handle complex microstructure geometries. *Computer Methods in Applied Mechanics and Engineering*, **192**, 3163-3177, 2003.

[11] E. Taroco, Shape sensitivity analysis in linear elastic fracture mechanics. *Computer Methods in Applied Mechanics and Engineering*, 697-712, 2000.

[12] K. Choi, N. Kim, *Design Sensitivity Analysis of Linear Structural Systems*. In K. Choi. N. Kim, Structural Sensitivity Analysis and Optimization 1: Linear Systems and Applications. Berlin: Springer Science & Business Media, 2006.

[13] J.P. Zolésio, *The material derivative (or speed) method for shape optimization. Optimization of Distributed Parameter Structures*, Sijthoff and Noordhoff, Alphen aan den Rihn, Netherlands, 1089-1151, 1981.

[14] L. Meirovitch, *Principles and Techniques of Vibrations*, Prentice-Hall International Editions, 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# XFEM ANALYSIS – INCLUDING BUCKLING – OF COMPOSITE SHELLS CONTAINING DELAMINATION

**Wilhelm J.H. Rust**[1], **Saleh Yazdani**[2], **and Peter Wriggers**[2]

[1] Hochschule Hannover – University of Applied Sciences and Arts, Ricklinger Stadtweg 120, 30459 Hannover, Germany
wilhelm.rust@hs-hannover.de

[2] Institute of Continuum Mechanics, Leibniz Universität Hannover, Appelstraße 11, 30167 Hannover, Germany
e-mails: yazdani@ikm.uni-hannover.de; wriggers@ikm.uni-hannover.de

**Keywords:** XFEM, Delamination, Shells, Buckling.

**Abstract.** *Based on a first-order shear-deformation theory a four-noded shell element is formulated for composite laminates with and without a delamination. Interlaminar stresses are calculated to evaluate a delamination criterion. The presentation will show how and how accurate. Once the danger of delamination is detected the displacement field is enriched in the sense of the eXtended FEM to account for a discontinuity at an arbitrary position. The remaining strength after starting delamination is simulated by a mixed-mode cohesive zone model based on energy release. Contact when reclosing a discontinuity is included. Large rotations are covered by Green's strain.*

*The developed formulation is tested for shell problems by comparing its results with available benchmark tests for in- and out-of-plane load cases. The feasibility and practicality of the presented model and its advantage over the approach using two shell elements at the predefined plane of the discontinuity, connected by a third one, a cohesive zone element, is demonstrated. Then, linear and non-linear buckling analyses for composite laminates containing a delamination are performed. It is shown how imperfections of the delamination type, i.e. concerning the discontinuity, used for non-linear analysis are based on the modes from linear buckling.*

*The full process from starting with one layer of elements over detecting a delamination to simulating its growth and its interaction with buckling is outlined.*

## 1 INTRODUCTION

When modeling delamination this can be done with two shell or solid elements over each other where all the interface area is bonded but the delaminated region. In such a model delamination propagation can be accounted for by a fracture criterion or a stress criterion in bonded contact, residual strength by a cohesive zone element. However, this requires double elements for the continuous region, too. Furthermore, if the delamination can appear between arbitrary layers not known before starting the analysis *n* thin solid elements and *n*-1 cohesive zones above each other are needed leading to enormous model sizes.

In the present formulation, the discontinuity can be activated at any region of interest without extra simulation effort. To do so, shape functions are enriched by defining extra degrees of freedom (DOF), holding the property of the partition of unity [1]. Thus, the behaviour of the discontinuous subdomains is described using these DOFs. A flat-shell formulation based on a First-order Shear Deformable Theory (FSDT) is developed. The proposed model is implemented in the finite element method using four-noded elements. Due to the robustness and simplicity of lower-order theories in the non-linear simulations [2], in the present formulation the XFEM topology and the geometrically non-linear terms can be effectively combined. The shell formulation is verified for a benchmark test problem of shells in non-linear regime. Then, several numerical tests are carried out for the linear and non-linear delamination buckling analyses of composite laminates. More tests an details are published in [8]. In the next section, the formulation is briefly described.

## 2 FORMULATION

All the formulations are developed with respect to the local Cartesian coordinate system which is located on each flat element. Later, a transformation from local to the global Cartesian coordinate system is required to include the coupling between membrane and bending components of shell structures [3, 4].

### 2.1 Discontinuous flat-shell element

The first-order displacement field in the local Cartesian coordinate system is enriched by XFEM as follows:

$$\begin{aligned} U(x,y,z) &= u^0(x,y) + H(z_d)a_u^0(x,y) + z\left(\theta_y^0(x,y) + H(z_d)a_{\theta_y}^0(x,y)\right) \\ V(x,y,z) &= v^0(x,y) + H(z_d)a_v^0(x,y) - z\left(\theta_x^0(x,y) + H(z_d)a_{\theta_x}^0(x,y)\right) \\ W(x,y,z) &= w^0(x,y) + H(z_d)a_w^0(x,y) \end{aligned} \tag{1}$$

where $u^0$, $v^0$, $w^0$, $\theta_x^0$, and $\theta_y^0$ are the normal DOFs while $a_u^0$, $a_v^0$, $a_w^0$, $a_{\theta_x}^0$, and $a_{\theta_y}^0$ are associated to the enriched DOFs. The superscript 0 indicates the middle plane kinematic variables. $H(z_d)$ is the Heaviside function, let it be $H^*$, shifted by $z_d$ which gets values zero and one below and above the discontinuous surface, respectively:

$$H(z_d) := H^*(z - z_d) = \begin{cases} 1 & for\ z > z_d \\ 0 & for\ z \le z_d \end{cases}$$

This function is being used to activate the extra DOFs when they are required. Therefore, two subdomains ($\Omega^-$ and $\Omega^+$) are identified by $z_d$, being the location of delamination in the thickness direction. The introduced discontinuous element and subdomains are shown in Figure 1.

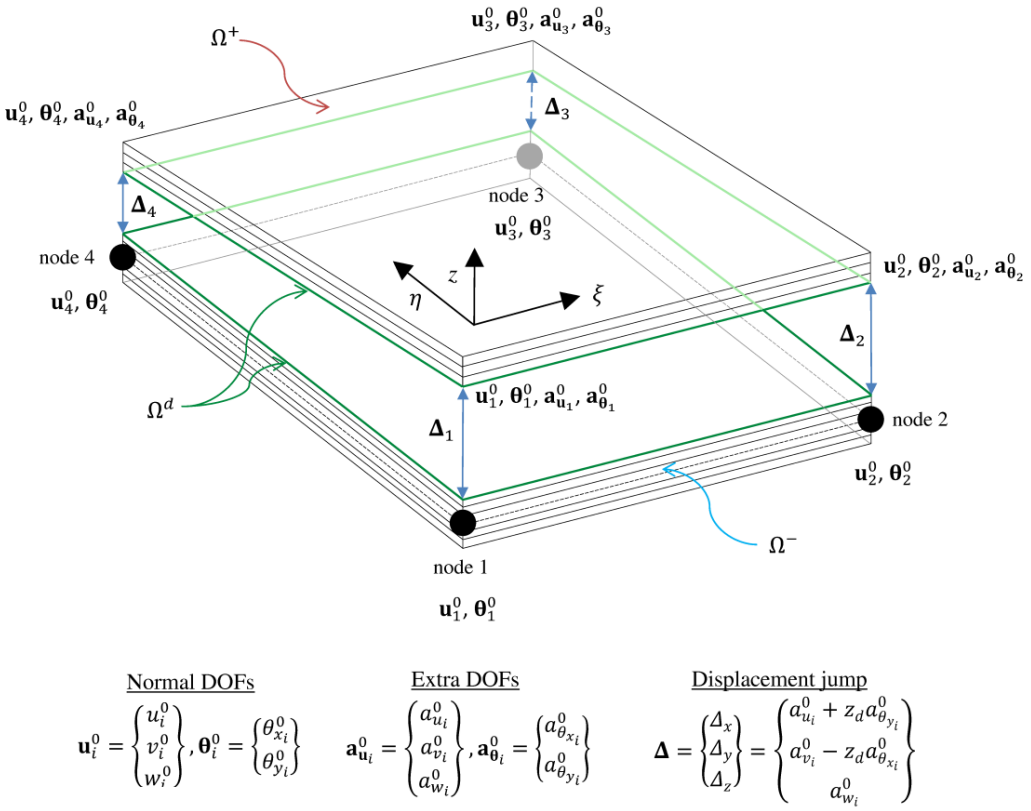



Figure 1: A discontinuous XFEM element.

By inserting the proposed displacement field into the Green-strain tensor and considering the quadratic terms associate to the in-plane components, the strain field is calculated as

$$
\begin{aligned}
\varepsilon_x &= u_{,x}^0 + z\theta_{y,x}^0 + H(z_d)\left(a_{u,x}^0 + z a_{\theta_y,x}^0\right) + \frac{1}{2}\left(u_{,x}^0 + H(z_d)a_{u,x}^0\right)^2 + \frac{1}{2}\left(v_{,x}^0 + H(z_d)a_{v,x}^0\right)^2 + \frac{1}{2}\left(w_{,x}^0 + H(z_d)a_{w,x}^0\right)^2 \\
\varepsilon_y &= v_{,y}^0 - z\theta_{x,y}^0 + H(z_d)\left(a_{v,y}^0 - z a_{\theta_x,y}^0\right) + \frac{1}{2}\left(u_{,y}^0 + H(z_d)a_{u,y}^0\right)^2 + \frac{1}{2}\left(v_{,y}^0 + H(z_d)a_{v,y}^0\right)^2 + \frac{1}{2}\left(w_{,y}^0 + H(z_d)a_{w,y}^0\right)^2 \\
\gamma_{xy} &= \left(u_{,y}^0 + H(z_d)a_{u,y}^0\right) + \left(v_{,x}^0 + H(z_d)a_{v,x}^0\right) + z\left(\theta_{y,y}^0 + H(z_d)a_{\theta_y,y}^0\right) - z\left(\theta_{x,x}^0 + H(z_d)a_{\theta_x,x}^0\right) + \\
&\left(u_{,x}^0 + H(z_d)a_{u,x}^0\right)\left(u_{,y}^0 + H(z_d)a_{u,y}^0\right) + \left(v_{,x}^0 + H(z_d)a_{v,x}^0\right)\left(v_{,y}^0 + H(z_d)a_{v,y}^0\right) + \left(w_{,x}^0 + H(z_d)a_{w,x}^0\right)\left(w_{,y}^0 + H(z_d)a_{w,y}^0\right) \\
\gamma_{xz} &= w_{,x}^0 + \theta_y^0 + H(z_d)\left(a_{w,x}^0 + a_{\theta_y}^0\right) \\
\gamma_{yz} &= w_{,y}^0 - \theta_x^0 + H(z_d)\left(a_{w,y}^0 - a_{\theta_x}^0\right)
\end{aligned}
\tag{2}
$$

The higher-order terms in Eq. (2) are utilized to capture the large deflection response of the laminates. By substituting the acquired strain field into the constitutive equation of orthotropic materials, the stress field is acquired. Due to the lack of parabolic distribution of transverse shear stresses in this theory, the so-called shear correction factor is calculated. In addition, since the formulation is based on a lower-order approach which shows locking for very thin laminates, the shear strain field has been replaced by an assumed one.

The principle of virtual work is applied to obtain the governing equations. To achieve a quadratic convergence in the context of the Newton-Raphson algorithm, the obtained formulations should be linearized with respect to all normal and enriched unknowns [2, 7]. A transformation from the local Cartesian coordinate system to the global one has been performed to

form the flat elements on the curved surface. This transformation is based on the standard procedure using four rotational matrices (in XFEM formulation eight rotational matrices) explained by Zienkiewicz and Taylor in Ref. [4]. Henceforth, owing to the transformation into the curve surface the coupling between membrane and bending components are formed as well as two extra variables, being the drilling DOFs. One of the drilling variables belongs to the finite element part of the problem whereas the other one corresponds to the XFEM.

### 2.2 Drilling degrees of freedom

To avoid the singularity which is faced in the solution process of co-planar elements due to missing stiffness related to drilling dofs pointing out of plane, the penalty method is implemented[4]. In this case, the rotational stiffness coefficient is related to the membrane stiffness components in such a manner that the overall equilibrium equation is not disturbed. Therefore, the drilling potential energy, containing a penalty parameter, is defined as

$$\Pi_{drilling} = \frac{1}{2}\int_{\Omega} P\left(\omega^0 - \theta_z^0\right)^2 d\Omega \tag{3}$$

where $P$ is the penalty parameter, $\theta_z^0$ is the drilling degree of freedom, and $\omega$ is the in-plane rotation of the shell which is related to the membrane DOFs as follows

$$\omega^0 = \frac{1}{2}\left(v_{,x}^0 - u_{,y}^0\right) \tag{4}$$

The amplitude of the penalty parameter can be chosen as $P=10^{-4}E_1h$. The stiffness which is driven based on the presented potential energy can be included for all elements whether they are co-planar or not. It is of importance to conquer the singularity by the aforementioned methods in the local Cartesian coordinate system before the transformation into the global coordinate system.

### 2.3 Contact formulation

In the linear buckling analysis, it is possible to obtain a few mode shapes concerning the penetration of discontinuous surfaces. Though, these mode shapes cannot be physically interpreted. In addition, performing the non-linear analysis may lead to the overlaps between contact subdomains in the post-buckling regime. Therefore, a simple contact formulation based on the penalty method has been taken into account. The relative displacements at the interface region can be simply calculated through the enhanced DOFs in where the contact force values are retrieved as

$$F_c = P\delta_z \tag{5}$$

where $F_c$ is the contact force, $P$ is the penalty stiffness value, and $\delta_z$ is the gap between the subdomains. Once the overlapping of the subdomains lead to negative values for the relative displacement $\delta_z$ during the non-linear buckling analysis, the above formulation is activated and the contact forces hold the equilibrium at the contact area. It is noted that the penalty stiffness value should be chosen sufficiently high to ensure a perfect bonding at the interface region.

All the developed formulations are set up as a user element routine in ANSYS 14.5 commercial software. In the next section, some numerical case studies are carried out based on the developed formulation.

## 3 NUMERICAL TESTS

In this section numerical tests are carried out to verify the accuracy of the XFEM model in predicting the structural response of composite laminates containing the delamination. In addition, convergence studies are performed to check the precision of the formulation. The material properties of studied laminates are given in Table 1.

| Type | $E_{11}$ | $E_{22}$ | $\nu_{12}$ | $G_{12}$ | $G_{13}$ | $G_{23}$ |
|---|---|---|---|---|---|---|
| Shell [5] | 3300 Pa | 1100 Pa | 0.3 | 660 Pa | 660 Pa | 660 Pa |
| Plate [6] | 142 GPa | 10.8 GPa | 0.3 | 5.49 GPa | 5.49 GPa | 3.72 GPa |

Table 1: Material properties of studied laminates.

### 3.1 Nonlinear static analysis

The three layers cross-ply [0°/90°/0°] and [90°/0°/90°] cylindrical panels are analysed. The dimensions of the shell are: $L$=508 m, $R$=2540 m, and $\phi$=0.1 rad. The shell is supposed to be hinged at the edges along the axial direction while it is free on the other edges. Two different thicknesses $h_1$=6.35 m and $h_2$=12.7 m are modelled and they are subjected to an external point load at the centre. Due to the symmetry, only a quarter of the shell is simulated. A schematic view of the shell is shown in Figure 2.

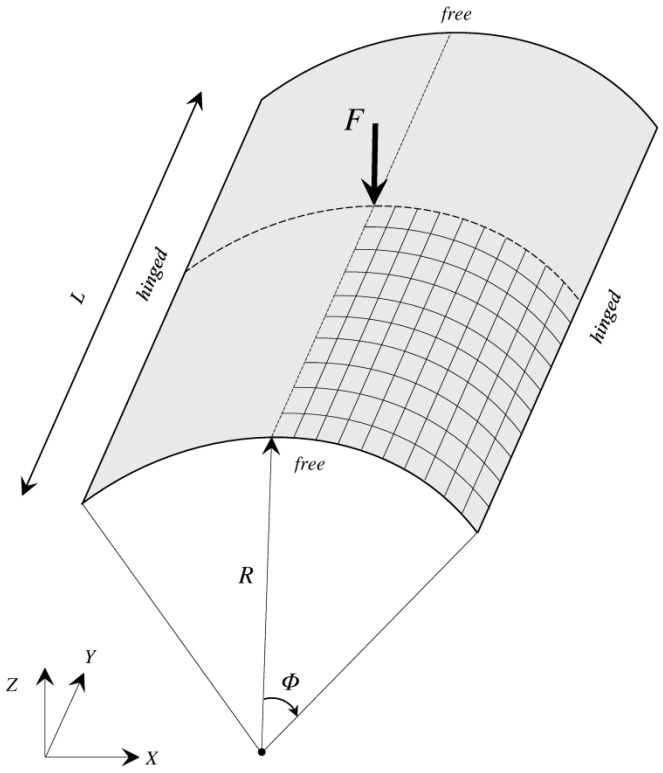



Figure 2: The schematic view of the hinged cylindrical shell.

In order to precisely track the response of the laminate under the applied load, arc-length solution available in ANSYS software is activated. The load versus transverse displacement response of the laminates at the point where the load is applied is depicted in Figure 3.

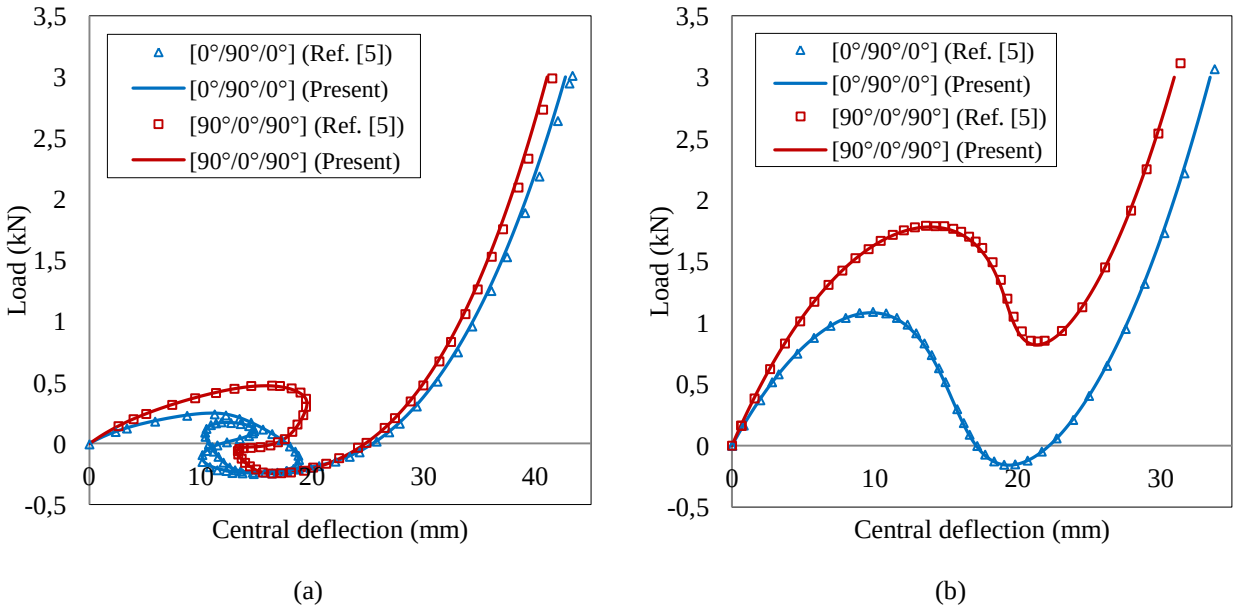



Figure 3: Load-deflection response of the cylindrical shell: (a) $h$=6.35 m; (b) $h$=12.7 m.

As it is shown in Figure 3, the results agree very well with the ones reported in Ref [5].

### 3.2 Delamination buckling analysis

Next, the buckling analysis of multi-layered laminated plate $[45°/-45°/0°/90°]_s$ is investigated. The dimensions of the plate are $a$=0.5 m, $b$=0.025 m, and $h$=0.001 m where a delamination of length $t$ is inserted exactly in the middle of the thickness, between the fourth and fifth layers. The plate is considered as fully clamped on one of the small edges whereas subjected to axial compressive load on the other small edge. The schematic view of the studied laminate is shown in Figure 4.

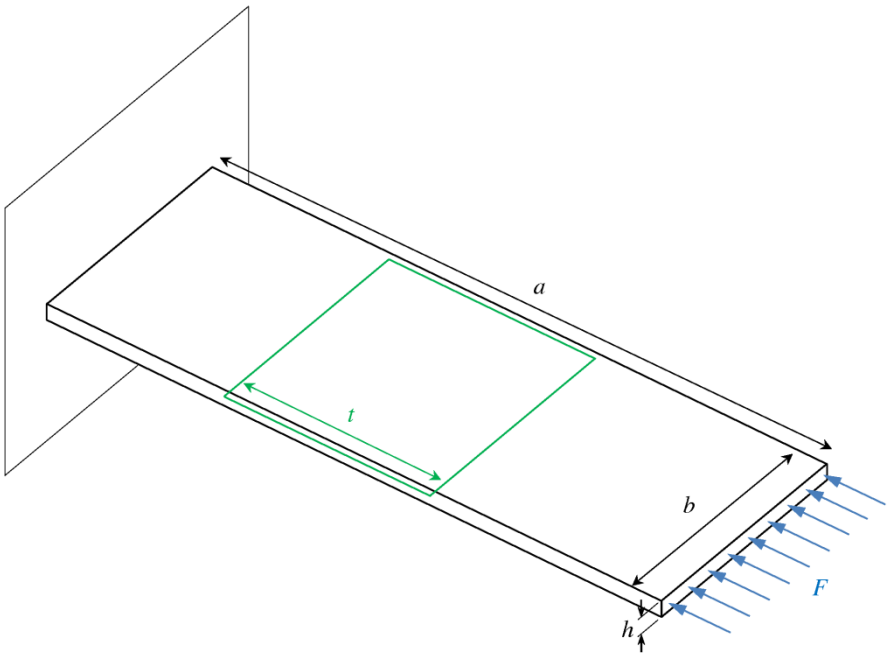



Figure 4: Schematic view of the delaminated composite plate.

In the simulation process of the continuous part, only the normal degrees of freedom are considered. However, extra degrees of freedom are activated in where the discontinuous region exists. To survey the precision of the XFEM model developed, a convergence study on the linear buckling analysis of composite laminate with the delamination length $t$=40 mm is carried out.

For the linear buckling analysis, the tangent stiffness matrix can be split up into linear and non-linear parts. Therefore, the tangent operator $\mathbf{K}_T$ can be written as

$$\mathbf{K}_T = \mathbf{K}_L + \mathbf{K}_U + \mathbf{K}_\sigma \tag{6}$$

where $\mathbf{K}_L$ is the linear stiffness matrix, $\mathbf{K}_U$ is the non-linear part of the stiffness matrix which is related to the initial deformation, and $\mathbf{K}_\sigma$ is the other non-linear part which is the so-called geometric stiffness matrix or initial stress matrix, in case of a Green-strain formulation not directly depending on the displacements, only via the stress. Hence, the linear eigenvalue analysis can be formulated as follows [7]

$$\left(\mathbf{K}_L + \lambda_i \left(\mathbf{K}_\sigma\right)\right)\phi_i = \mathbf{0} \tag{7}$$

where $\lambda_i$ is the critical load factor whereas $\Phi_i$ is the corresponding mode shape. The first nine predicted buckling loads are reported in Table 2.

| Method | Element size | Buckling mode | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Present | 0.005 m | 18.122 | 93.933 | 271.33 | 418.66 | 455.50 | 764.19 | 791.48 | 917.11 | 994.67 |
| | 0.0025 m | 17.984 | 91.723 | 255.49 | 375.17 | 422.26 | 710.01 | 710.88 | 825.61 | 892.06 |
| | 0.00125 m | 17.923 | 91.057 | 251.12 | 364.37 | 413.41 | 685.82 | 694.39 | 801.73 | 856.76 |
| ABAQUS (S4) [6] | Element number | 17.9435 | 93.7975 | 269.975 | 418.400 | 452.825 | - | - | - | - |
| XSHELL [6] | 200 | 18.148 | 94.0725 | 272.200 | 420.725 | 456.975 | - | - | - | - |

Table 2: Convergence study on linear buckling loads of delaminated plate ($t$=40 mm).

A fair convergence property is achieved for the model containing extra DOFs. The obtained results are closer to the ones of ABAQUS (S4) standard element in where the delaminated region was simulated through defining double nodes. The corresponding first and fourth mode shapes of the finest mesh scheme are depicted in Figure 5. All post-processing concerning the delaminated state is performed in the standard postprocessor of ANSYS by creating virtual elements and assigning results obtained from extra dofs to them.

Next, the length of the delamination is altered to 45 mm and 50 mm and the first nine buckling loads are compared in Table 3. The element size for the present model is supposed to be 0.00125 m.

| $t$ (mm) | Method | Buckling mode | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 45 | Present | 17.331 | 80.522 | 214.04 | 282.01 | 392.91 | 562.48 | 635.95 | 714.14 | 725.66 |
| | XSHELL [6] | 17.937 | 90.702 | 256.900 | 390.730 | 446.800 | - | - | - | - |
| 50 | Present | 16.634 | 72.425 | 182.85 | 224.15 | 372.36 | 458.72 | 595.60 | 640.83 | 649.24 |
| | XSHELL [6] | 16.850 | 74.378 | 196.450 | 242.790 | 410.500 | - | - | - | - |

Table 3: Buckling loads of composite plates with delamination ($t_1$=45 mm and $t_2$=50 mm).

As it was expected, by increasing the delaminated area the predicted buckling loads are reduced. The calculated buckling loads, especially the fundamental ones, are in a good agreement with Ref [6].

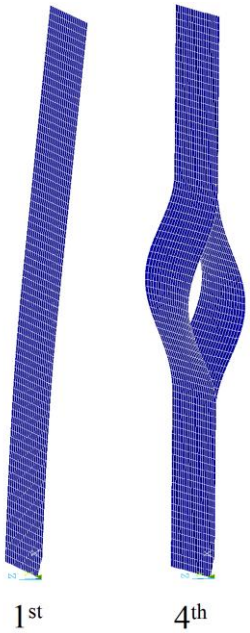



Figure 5: Linear buckling mode shapes of delaminated composite plate ($t$=40 mm).

In the final example, the non-linear buckling analysis of the previous plate is carried out. In order to enable tracking the post-buckling response of the laminate under compressive load condition, very small imperfection amplitudes has been utilized. In the present XFEM formulation we are restricted to only one middle plane, even for discontinuous regions. Therefore, the application of geometric imperfection for the aforementioned problems is not provided. Void layers of variable thickness as well as non-zero initial jumps for the $z$-coordinates of the upper region had been under discussion to get stress-free imperfections but finally a force-type perturbation was chosen. This type of imperfection is applied in literatures to the discontinuous layers when the plate with delamination under compression is investigated. This method can be effectively implemented in the XFEM topology through the availability of the extra DOFs. Herein, a new algorithm which describes a method to apply forces leading to a deformation proportional to the buckling mode has been applied. The flowchart view of this algorithm is shown in Figure 6. The proposed method is applied to ANSYS version 14.5 commercial software.

The algorithm is started with performing a linear buckling analysis. By depicting the linear buckling mode shapes, one can identify whether the critical load corresponds to the local delamination mode shape or not. If so, a factor of dislocations is utilized in the form of displacement constraints to retrieve the nodal reaction forces in a linear static analysis. Next, the force values of the extra DOFs are applied to the discontinuous section of model and the non-linear analysis is initiated in its first step. By doing so, the deformations proportional to that buckling mode are calculated. Thereafter, the same analysis is being continued after applying compressive loads. It is noted that when the linear buckling analysis of laminates with discontinuity is carried out, the critical buckling mode shape is not necessarily represents the local delamination mode shape. Therefore, one further step is required to include the geometrical

imperfection from the critical buckling mode shape. By following the above algorithm, the possibility of opening multi bifurcation paths is provided.

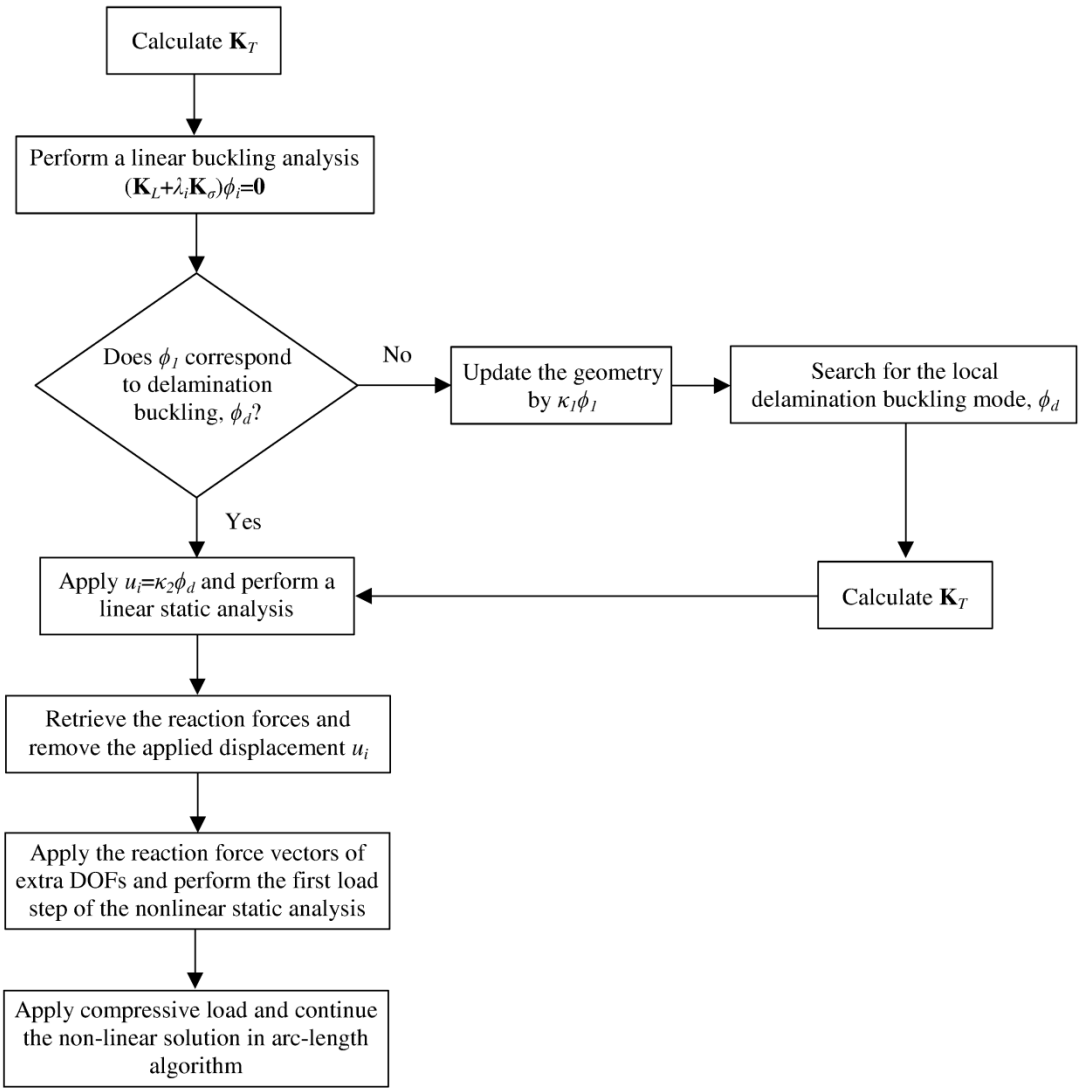



Figure 6: Flowchart of the non-linear delamination buckling analysis.

In the previous linear buckling study, the critical buckling mode was associated to the bending of the structure whereas the fourth one was corresponded to the outward separation of layers. Therefore, in order to carefully track the response of the laminate, a combination effect of the mode I and IV is considered. The amplitudes of imperfections are assumed as $\kappa_1=h$ and $\kappa_2=0.025h$. It should be mentioned that excluding the bending imperfection will lead to negative pivot values in the vicinity of the critical buckling load. The element size is supposed to be 1.25 mm. The load versus the transverse displacement of the edge in where the compressive load was applied is compared with the undamaged one in Figure 7.

A meaningful reduction of the maximum carried load is observed for the laminated plates containing the greater delaminated surface. Moreover, higher load magnitudes are supported by the undamaged plate than the delaminated ones. Taking the results of the linear buckling analysis as the reference, the maximum supported load for the undamaged laminate is close to the critical one. However, less load amplitudes was supported by the delaminated ones.

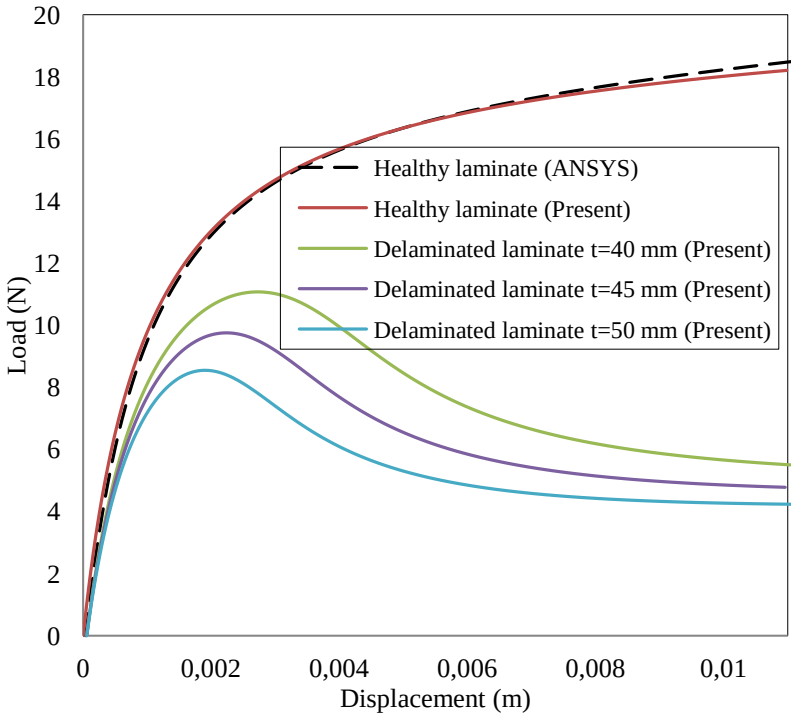



Figure 7: Axial load versus transverse displacement of composite laminated plate under compressive load.

The deformed shape of the composite laminates with delamination $t$=40 mm at the last load step is shown in Figure 8.

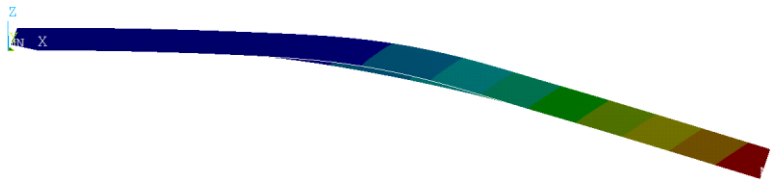


Figure 8: Deformed shape of the composite laminated plate with delamination under compressive load.

## 4 DELAMINATION INITIATION AND PROPAGATION

The danger of delamination is checked by a stress criterion based on interlaminar shear and normal stress obtained from equilibrium. However, the determination of interlaminar shear stress requires the derivative of in-plane normal stress which is constant in a low order element. Therefore, nodal averaging of in-plane stress is carried out before forming the derivative.

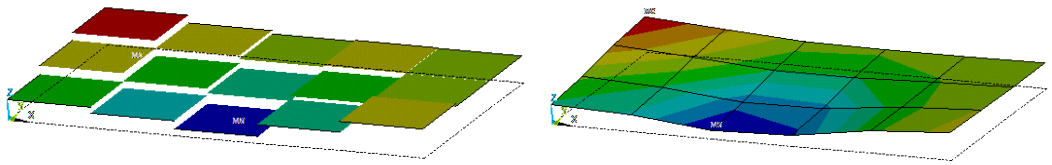


Figure 9: Elemental and averaged stress.

Furthermore, interlaminar normal stress is in equilibrium with the change of transverse shear stress, requiring another derivative. This can deteriorate the accuracy. In the example shown in Figure 10 this mainly holds for the normal stress.

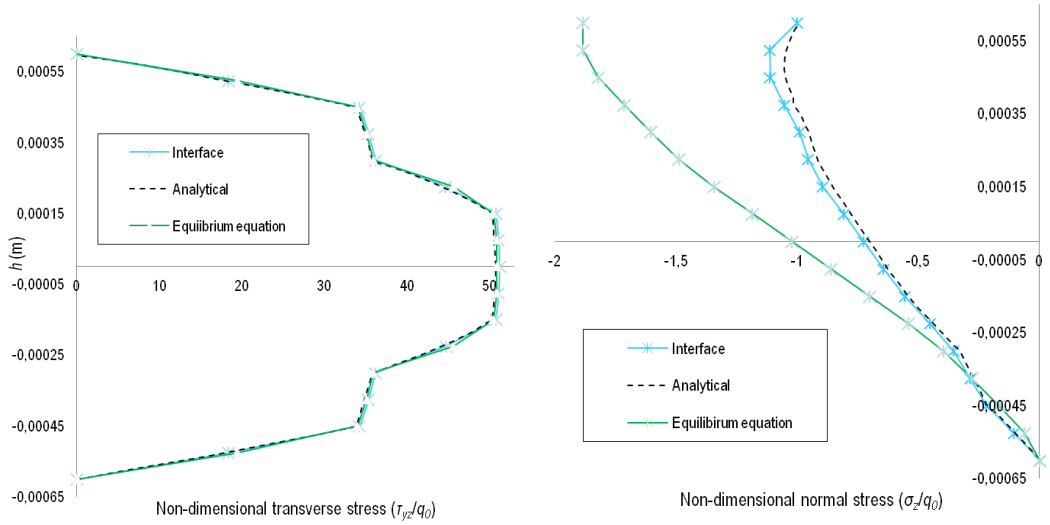



Figure 10: Shear- and out-of-plane normal-stress distribution, comparison of interfacial stress and stress from equilibrium with analytical solution from [9].

The simulation of delamination propagation with this XFEM shell element including a mixed-mode cohesive-zone model and the influence of the integration scheme has been presented by the authors e.g. in [8]. The relative displacements in the interface between lower an upper part are directly obtained from the extra dofs $a_i$ multiplied by the element shape functions. Instead of using a stress-strain relation the interfacial stress depends on the relative motion. Hence, this stress distribution is continuous over the element edges and one order higher than that from strain anywhere else in the finite element. All interfacial stress components are available from the cohesive model. Therefore, for delamination initiation and propagation the algorithm is as follows:

- start with continous model – one element over the thickness for all layers
- at certain load levels
  - calculate interlaminar stress from equilibrium between all layers
  - evaluate the delamination criterion
  - if the criterion indicate danger of delamination at a certain location set the element and its neighborhood to discontinuous and introduce the interfacial cohesive model
- continue with loading

## 5 CONCLUSIONS

The delamination buckling analysis of composite laminates is investigated. Thus, a new XFEM model based on a lower order laminate theory is developed. In the present approach, the formulation is enhanced by adding extra DOFs; and subsequently, the delaminated surface can be described by only one four-noded element. Therefore, no extra element are needed to simulate the subdomains. Moreover, the definition of contact element and imperfections for the post-buckling analysis are facilitated through the available enriched DOFs. The performance of the model is successfully verified with the available results in literature.

## REFERENCES


[1] N. Moës, J. Dolbow, T. Belytschko, A finite element method for crack growth without remeshing. *International Journal of Numerical Methods in Engineering*, **46.1**, 131-150, 1999.

[2] P. Wriggers, *Nonlinear finite element methods*. Springer Science & Business Media, 2008.

[3] E. Oñate, *Structural Analysis with the Finite Element Method. Linear Statics: Volume 2: Beams, Plates and Shells*. Springer, 2013.

[4] O.C. Zienkiewicz, R.L. Taylor, *The finite element method, Vol. 2: Solid Mechanics, 5th Edition*. McGraw Hill, 2000.

[5] K.Y. Sze, X.H. Liu, S.H. Lo, Popular benchmark problems for geometric nonlinear analysis of shells. *Finite elements in analysis and design*, **40.11**, 1551-1569, 2004.

[6] T. Nagashima, H. Suemasu, X-FEM analyses of a thin-walled composite shell structure with a delamination. *Computers & structures*, **88.9**, 549-557, 2010.

[7] W. Rust, *Non-linear Finite Element Analysis in Structural Mechanics*. Springer, 2015.

[8] S. Yazdani, W.J.H. Rust, P. Wriggers, *An XFEM approach for modelling delamination in composite laminates*, Composite Structures (2016), pp. 353-364

[9] T.B. Hartman, M.W. Hyer, S.W. Case, Stress Recovery in Composite Laminates. In: *52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, Denver, Colorado; 4–7 April 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THREE-DIMENSIONAL CRACK PROPAGATION ANALYSIS USING MESHLESS POINT COLLOCATION METHOD

**Eiji Tanaka**[1]

[1]Institute of Technology, Shimizu Corporation
3-4-17 Etchujima, Koto-ku, Tokyo 135-8530 Japan
e-mail: tanaka_e@shimz.co.jp

**Keywords:** Point collocation method, Meshless methods, Constrained moving least squares approximation, crack propagation, cohesive crack

**Abstract.** *This paper presents a meshless point collocation method for three-dimensional crack propagation. The meshless point collocation method is based on direct discretization of strong-form governing equations to achieve a truly meshless scheme that does not require mesh structures or a numerical integration procedure. These characteristics of the point collocation method enable the direction of an arbitrary crack propagation to be defined purely according to the fracture criterion of the mechanical model because node-wise topology can be easily transformed by the simple addition of nodes and polygonal surfaces. The main advantages of the presented method are that there is no need to consider the consistency between the mesh connectivity and continuity of fracture zone locations and that the crack propagation procedure is relatively simplified with no additional implementation features from numerical integration used in weak formulations. The constrained moving least squares approximation (C-MLS) is employed for meshless interpolation so that essential boundary conditions can easily be applied. A cohesive crack model is directly introduced in combination with the crack growth procedure to express the fracture of quasi-brittle materials such as concrete. The accuracy and robustness of the presented method were demonstrated through its application to several numerical examples.*

## 1 INTRODUCTION

The meshless method is an attractive numerical technique that can solve various natural phenomena and engineering problems that conventional mesh methods are not applicable to. To date, a number of computational methods have appeared, such as smoothed particle hydrodynamics (SPH) [1, 2], the diffuse element method (DEM) [3], the element-free Galerkin method (EFGM) [4], the reproducing kernel particle method (RKPM) [5], and the partition of unity finite element method (PUFEM) [6]. Although these meshless methods have been applied to various engineering problems and become popular in computational mechanics, each still exhibits several disadvantages. Some are not completely meshless and use background meshes or other mesh structures over the problem domain because of the numerical integration in weak-form governing equations. Most also have difficulties with treating essential boundary conditions because of the lack of interpolation properties. For example, the moving least squares approximation (MLS) [7] lacks the Kronecker delta property. Further research is required on meshless methods to overcome these issues.

The point collocation approach—of which examples include the finite point method (FPM) [8], meshfree point collocation method (MPCM) [9], and particle difference method (PDM) [10]—is known to achieve a completely meshless method based on node-wise strong formulation. There are two characteristics to transforming the point collocation approach into a truly meshless method: using strong-form governing equations to avoid any complex implementation features caused by the numerical integration scheme of the weak formulation and using interpolation schemes that do not require meshes to construct shape functions, such as the reproducing kernel approximation and MLS approximation. Compared to other meshless methods, the point collocation approach is a relatively simple scheme because most meshless methods need some additional mesh structures or numerical integration procedure. Point collocation methods have been applied to fluid flow problems [8], elasticity of solid mechanics [11], solid elastoplasticity [12, 13], and crack growth problems [14] in two dimensions.

This paper presents a meshless point collocation method for the analysis of three-dimensional fracture. The constrained moving least squares (C-MLS) approximation, which is a meshless interpolation technique proposed by Noguchi et al. [15, 16], is introduced to easily treat essential boundary conditions. The topological change due to crack propagation is simply represented by the addition and subtraction of nodes and surfaces. A cohesive crack model is employed to deal with the fracture of quasi-brittle materials such as concrete. The cohesive crack model suggested by Barenblatt [17] is widely exploited to address the fracture of engineering materials to which classical linear fracture mechanics cannot be applied. The cohesive crack model assumes that a virtually narrow band called the cohesive zone exists ahead of a crack tip and represents the fracture process zone. The resistance to fracture is considered to be a gradual process in which separation between incipient crack surfaces is given by the cohesive traction of the cohesive zone following a cohesive constitutive law. The cohesive constitutive law represents the failure characteristics of the materials and is governed by cohesion at the atomistic scale, crack-bridging ligaments, interlocking of grains, and other factors. Crack growth occurs when the cohesive traction applied to the outer end of the cohesive zone disappears. Several numerical examples were considered, and the results were compared with those in the literature.

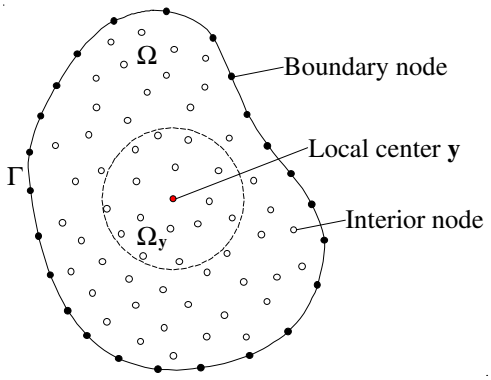



Figure 1: Support domain $\Omega_y$ for C-MLS approximation of the local center $\mathbf{y}$

## 2 ANALYSIS METHODS

### 2.1 Constrained moving least squares approximation

This section introduces the C-MLS approximation procedure with some extensions and modifications based on the original scheme [15, 16]. Consider a domain $\Omega$ surrounded by a boundary $\Gamma$ in $\mathbb{R}^n$, as shown in Fig. 1. Suppose that a continuous function $u(\mathbf{x}) : \bar{\Omega} \to \mathbb{R}$ is to be approximated and that the nodal values $u_I$ at $\mathbf{x}_I \in \bar{\Omega}$, $I = 1,2,...,N$ are given, where $N$ is the number of nodes included in the support domain $\bar{\Omega}$. For the sake of convenience, multi-index notation has been adopted. Let $\mathbf{x} = (x_1,...,x_n)$ be an $n$-dimensional vector and $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_n)$ be an $n$-tuple of non-negative integers. Then, the $\alpha$th power of $\mathbf{x}$ is interpreted as $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n}$ and $|\boldsymbol{\alpha}| = \sum_{i=1}^{n} \alpha_i$. In addition, $D_{\mathbf{x}}^{\boldsymbol{\alpha}} := \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \dots \partial_{x_n}^{\alpha_n}$ indicates the $\alpha$th partial derivative operator with respect to $\mathbf{x}$. The $\alpha$th derivative of the differentiable function $f(\mathbf{x})$ is given by

$$D_{\mathbf{x}}^{\boldsymbol{\alpha}} f(\mathbf{x}) = \frac{\partial^{|\boldsymbol{\alpha}|} f(\mathbf{x})}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \dots \partial_{x_n}^{\alpha_n}}. \tag{1}$$

The C-MLS approximation employs an approximate function based on a Taylor series instead of the normal polynomial approximation used in general MLS. The local approximation of $u(\mathbf{x})$ derived from the $m$th order Taylor polynomial is written as

$$u^h(\mathbf{x},\mathbf{y}) = u_{\mathbf{y}} + \mathbf{p}_m^T(\mathbf{x},\mathbf{y})\mathbf{a}(\mathbf{y}) \tag{2}$$

where $\mathbf{p}_m^T$ is constructed from all of the elements in the set $\mathscr{P}_m = \left\{ (\mathbf{x}-\mathbf{y})^{\boldsymbol{\alpha}}/\boldsymbol{\alpha}! \;\middle|\; \boldsymbol{\alpha} \in \mathscr{A}_m \right\}$, $\mathbf{a}(\mathbf{y})$ is the corresponding unknown coefficient vector made from the elements in the set $\mathscr{D}_m = \left\{ D_{\mathbf{x}}^{\boldsymbol{\alpha}} u(\mathbf{y}) \;\middle|\; \boldsymbol{\alpha} \in \mathscr{A}_m \right\}$, and $\mathscr{A}_m = \left\{ \boldsymbol{\alpha} \;\middle|\; 1 \le |\boldsymbol{\alpha}| \le m \right\}$. In addition, the number of all elements in the set $\mathscr{A}_m$ is $L = (n+m)!/(n!m!) - 1$. The above equation implies that a constraint condition is imposed on the approximate function used in MLS approximation, and all of the derivative approximations for $u(\mathbf{y})$ up to the $m$th order are obtained by $\mathbf{a}(\mathbf{y})$. In the sense of a discrete $L_2$ norm, the residual $J$ is defined by

$$J = \sum_{I=1}^{N} w\left( \frac{\mathbf{x}_I - \mathbf{y}}{\rho_{\mathbf{y}}} \right) \left[ \mathbf{p}_m^T(\mathbf{x}_I,\mathbf{y})\mathbf{a}(\mathbf{y}) + u_{\mathbf{y}} - u_I \right]^2 \tag{3}$$

where $w$ is the support of the weight function and $\rho_{\mathbf{y}}$ (i.e., the dilation parameter) determines the size of the support domain $\bar{\Omega}$. The best local approximation is derived by minimizing the

functional $J(\mathbf{y})$ in Eq. (3) under the stationary condition. This yields the following normal equation:

$$\mathbf{a}(\mathbf{y}) = \mathbf{M}^{-1}(\mathbf{y})\mathbf{B}(\mathbf{y})\mathbf{u} \tag{4}$$

$$\mathbf{M}(\mathbf{y}) = \sum_{I=1}^{N} w\left(\frac{\mathbf{x}_I - \mathbf{y}}{\rho_{\mathbf{y}}}\right)\mathbf{p}_m(\mathbf{x}_I,\mathbf{y})\mathbf{p}_m^T(\mathbf{x}_I,\mathbf{y}) \tag{5}$$

$$\mathbf{B}(\mathbf{y}) = \left(w\left(\frac{\mathbf{x}_1 - \mathbf{y}}{\rho_{\mathbf{y}}}\right)\mathbf{p}_m(\mathbf{x}_1,\mathbf{y}),\ldots,w\left(\frac{\mathbf{x}_N - \mathbf{y}}{\rho_{\mathbf{y}}}\right)\mathbf{p}_m(\mathbf{x}_N,\mathbf{y})\right) \tag{6}$$

where $\mathbf{u} = (u_1 - u_{\mathbf{y}}, u_2 - u_{\mathbf{y}}, \ldots, u_N - u_{\mathbf{y}})^T$. The local approximation of $u(\mathbf{x})$ is interpreted as

$$u^h(\mathbf{x},\mathbf{y}) = u_{\mathbf{y}} + \mathbf{p}_m^T(\mathbf{x},\mathbf{y})\mathbf{M}^{-1}(\mathbf{y})\mathbf{B}(\mathbf{y})\mathbf{u}. \tag{7}$$

Thus, the C-MLS approximation clearly satisfies $u^h(\mathbf{y},\mathbf{y}) = u_{\mathbf{y}}$ because a constraint condition is imposed, which means that no special implementation is required to treat the essential boundary conditions. These are the distinctive characteristics of C-MLS compared to MLS [7] and other approximation schemes.

Consider the derivative approximation of $u(\mathbf{x})$ at the local center $\mathbf{y}$. Then, the derivative coefficients of the Taylor polynomial can be related to differentiation with respect to $\mathbf{x}$ by substituting $\mathbf{x} = \mathbf{y}$ as $D_{\mathbf{x}}^{\boldsymbol{\alpha}} u^h(\mathbf{x},\mathbf{y})\big|_{\mathbf{x}=\mathbf{y}} \approx D_{\mathbf{x}}^{\boldsymbol{\alpha}} u(\mathbf{y}) = D_{\mathbf{x}}^{\boldsymbol{\alpha}} u(\mathbf{x})$. Equation (4) is transformed as shown below:

$$\begin{pmatrix} D_{\mathbf{x}}^{\boldsymbol{\alpha}_1} u(\mathbf{x}) \\ \vdots \\ D_{\mathbf{x}}^{\boldsymbol{\alpha}_L} u(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \Phi_1^{[\boldsymbol{\alpha}_1]}(\mathbf{x}) & \ldots & \Phi_N^{[\boldsymbol{\alpha}_1]}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \Phi_1^{[\boldsymbol{\alpha}_L]}(\mathbf{x}) & \ldots & \Phi_N^{[\boldsymbol{\alpha}_L]}(\mathbf{x}) \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} \tag{8}$$

where $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_L$ are the elements in the set $\mathscr{A}_m$ sorted in a lexicographic order and $\Phi_I^{[\boldsymbol{\alpha}_K]}(\mathbf{x})$ denotes the $\boldsymbol{\alpha}_K$th derivative of the nodal shape function. If $K$ is omitted, $\Phi_I^{[\boldsymbol{\alpha}_K]}(\mathbf{x})$ is expressed as

$$\Phi_I^{[\boldsymbol{\alpha}]}(\mathbf{x}) = \mathbf{e}_{\boldsymbol{\alpha}}^T \mathbf{M}^{-1}(\mathbf{x})\mathbf{c}_I(\mathbf{x}) \tag{9}$$

$$\mathbf{c}_I(\mathbf{x}) = \begin{cases} w\left(\dfrac{\mathbf{x}_I - \mathbf{x}}{\rho_{\mathbf{x}}}\right)\mathbf{p}_m(\mathbf{x}_I,\mathbf{x}) & \text{for } \mathbf{x}_I \neq \mathbf{x} \\ -\displaystyle\sum_{\mathbf{x}_I \neq \mathbf{x}} w\left(\dfrac{\mathbf{x}_I - \mathbf{x}}{\rho_{\mathbf{x}}}\right)\mathbf{p}_m(\mathbf{x}_I,\mathbf{x}) & \text{for } \mathbf{x}_I = \mathbf{x} \end{cases} \tag{10}$$

where $\mathbf{e}_{\boldsymbol{\alpha}}^T = (0,\ldots,1,\ldots,0)$ is a unit vector of the same order as the polynomial basis, whose non-zero component is related to the $\boldsymbol{\alpha}$th derivative. Then the derivative approximation $D_{\mathbf{x}}^{\boldsymbol{\alpha}} u(\mathbf{x})$ is given by

$$D_{\mathbf{x}}^{\boldsymbol{\alpha}} u(\mathbf{x}) = \sum_{I=1}^{N} \Phi_I^{[\boldsymbol{\alpha}]}(\mathbf{x}) u_I. \tag{11}$$

Therefore, C-MLS does not need direct differentiation of the shape function, which is a costly process in most meshless methods. This means that the computation speed for the derivatives can be considerably improved, especially when computing high-order derivative approximations.

The weight function used in all numerical calculations is non-smooth and is given as follows:

$$w(\mathbf{r}) = \begin{cases} \left(1 - \|\mathbf{r}\|^{\frac{1}{2}}\right)^2 & \text{for } \|\mathbf{r}\| \leq 1 \\ 0 & \text{for } \|\mathbf{r}\| > 1 \end{cases} \tag{12}$$

where $\mathbf{r} = (\mathbf{x}_I - \mathbf{x})/\rho_{\mathbf{x}}$. Although several numerical methods based on MLS approximation require a differentiable weight function to approximate the derivatives, C-MLS does not need the weight function to be differentiable because the derivative of the approximation includes no derivatives of the weight function. Yoon [10] stated that non-smooth functions are more suitable for strongly formulated meshless methods than smooth ones.

The dilation parameter $\rho_{\mathbf{x}}$ indicates the support radius of the weight function and how many nodes are included in the local approximation. Although some early meshless methods used a constant radius probe regardless of the node density [4, 5], several surveys [18] have noted that this may affect the approximation and suggested the variable dilation parameter to adjust the support range depending on the position of the local center and node density. In the present work, $\rho_{\mathbf{x}}$ was chosen to maximize the support radius under the condition $N \leq 7.89L$ where $m = n = 3$.

### 2.2 Point collocation scheme

Next, the framework of the point collocation method based on the strong formulation is explained. Let $\Omega$ be a bounded domain in $\mathbb{R}^n$ with the boundary $\Gamma$, as shown in Fig. 1. Then, the nonlinear problem is governed by the following set of partial differential equations:

$$\mathsf{A}_\Omega \mathbf{u} = \mathbf{f}_\Omega \quad \text{in } \Omega \tag{13}$$

$$\mathsf{A}_\Gamma \mathbf{u} = \mathbf{f}_\Gamma \quad \text{on } \Gamma \tag{14}$$

where $\mathsf{A}_\Omega$ denotes the partial differential operator of the given problem and $\mathsf{A}_\Gamma$ denotes the boundary value operator for the boundary condition. The above equations can be interpreted as

$$\mathbf{r}_\Omega(\mathbf{u}) = \mathsf{A}_\Omega \mathbf{u} - \mathbf{f}_\Omega = \mathbf{0} \quad \text{in } \Omega \tag{15}$$

$$\mathbf{r}_\Gamma(\mathbf{u}) = \mathsf{A}_\Gamma \mathbf{u} - \mathbf{f}_\Gamma = \mathbf{0} \quad \text{on } \Gamma \tag{16}$$

where $\mathbf{r}_\Omega(\mathbf{u})$ and $\mathbf{r}_\Gamma(\mathbf{u})$ are the residual functions in the domain and on the boundary, respectively. To solve the nonlinear problems, Newton's method is employed. Then, the residuals are linearized as follows:

$$\mathrm{D}\mathbf{r}_\Omega(\mathbf{u})[\delta\mathbf{u}] = \mathsf{K}_\Omega \delta\mathbf{u} \quad \text{in } \Omega \tag{17}$$

$$\mathrm{D}\mathbf{r}_\Gamma(\mathbf{u})[\delta\mathbf{u}] = \mathsf{K}_\Gamma \delta\mathbf{u} \quad \text{on } \Gamma \tag{18}$$

where $\mathsf{K}_\Omega$ and $\mathsf{K}_\Gamma$ represent the directional derivative operators in the domain and on the boundary, respectively. The residuals and unknown function $\mathbf{u}$ can be computed by iterating the following scheme when updating the $(k)$th iteration step to the $(k+1)$th iteration step:

$$\mathbf{r}_\Omega^{(k)} = \mathsf{A}_\Omega \mathbf{u}^{(k)} - \mathbf{f}_\Omega \tag{19}$$

$$\mathbf{r}_\Gamma^{(k)} = \mathsf{A}_\Gamma \mathbf{u}^{(k)} - \mathbf{f}_\Gamma \tag{20}$$

$$\mathsf{K}_\Omega \delta\mathbf{u} = -\mathbf{r}_\Omega^{(k)} \tag{21}$$

$$\mathsf{K}_\Gamma \delta\mathbf{u} = -\mathbf{r}_\Gamma^{(k)} \tag{22}$$

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \delta\mathbf{u} \tag{23}$$

Both Eqs. (21) and (22) are linear, so $\delta\mathbf{u}$ can be obtained by using a numerical solution of algebraic equations. The operators $\mathbf{r}_\Omega(\mathbf{u})$ and $\mathbf{r}_\Gamma(\mathbf{u})$ are also constructed in each step up to the unknown function $\mathbf{u}^{(k)}$. Most meshless methods based on the particle collocation approach discretize Eqs. (13) and (14) directly for the numerical solution. In contrast, the presented method introduces Eqs. (21) and (22) to extend the point collocation approach so that nonlinear problems can be solved by using the linearization and the iteration scheme.

### 2.3 Governing equations

In the classical theory of elasticity, the analytical solution for crack problems is obtained to solve the equilibrium equation by imposing boundary conditions on the crack faces. Similarly, the point collocation method can express cracks by imposing the traction boundary condition on the crack faces. Consider a solid mechanics problem with a small deformation in the domain $\Omega$ with the crack $\Gamma_c$ $(\Gamma_{c+}+\Gamma_{c-})$, i.e., bounded by $\Gamma$ $(\Gamma_t \cup \Gamma_u \cup \Gamma_c)$ as shown in Fig. 2. At each point $\mathbf{x} \in \Omega$, the problem is governed by the following set of differential equations:

$$\mathrm{div}\boldsymbol{\sigma} + \mathbf{b} = \mathbf{0} \quad \text{in } \Omega \tag{24}$$
$$\boldsymbol{\sigma}\mathbf{n} = \mathbf{t} \quad \text{on } \Gamma_t \tag{25}$$
$$\mathbf{u} = \bar{\mathbf{u}} \quad \text{on } \Gamma_u \tag{26}$$
$$\boldsymbol{\sigma}\mathbf{n}_{c+} = \mathbf{t}_{c+} \quad \text{on } \Gamma_{c+} \tag{27}$$
$$\boldsymbol{\sigma}\mathbf{n}_{c-} = \mathbf{t}_{c-} \quad \text{on } \Gamma_{c-} \tag{28}$$

where $\boldsymbol{\sigma}$ is the Cauchy stress tensor; $\mathbf{b}$ is the body force; $\mathbf{n}$, $\mathbf{n}_{c+}$, and $\mathbf{n}_{c-}$ are the outward surface normal vectors; $\mathbf{t}$ and $\mathbf{t}_{c+} = -\mathbf{t}_{c-}$ are the tractions; and $\bar{\mathbf{u}}$ is the prescribed displacement. This means that Eq. (14) directly applies to not only $\Gamma_t$ but also $\Gamma_c$. In other words, the boundary condition of cracks is simply embodied by the addition of nodes on crack faces and by imposing a node-wise strong formulation. Then, the residual functions are given as

$$\mathbf{r}_\Omega = \mathrm{div}\boldsymbol{\sigma} + \mathbf{b} \quad \text{in } \Omega \tag{29}$$
$$\mathbf{r}_{\Gamma_t} = \boldsymbol{\sigma}\mathbf{n} - \mathbf{t} \quad \text{on } \Gamma_t \tag{30}$$
$$\mathbf{r}_{\Gamma_u} = \mathbf{u} - \bar{\mathbf{u}} \quad \text{on } \Gamma_u \tag{31}$$
$$\mathbf{r}_{\Gamma_{c+}} = \boldsymbol{\sigma}\mathbf{n}_{c+} - \mathbf{t}_{c+} \quad \text{on } \Gamma_{c+} \tag{32}$$
$$\mathbf{r}_{\Gamma_{c-}} = \boldsymbol{\sigma}\mathbf{n}_{c-} - \mathbf{t}_{c-} \quad \text{on } \Gamma_{c-} \tag{33}$$

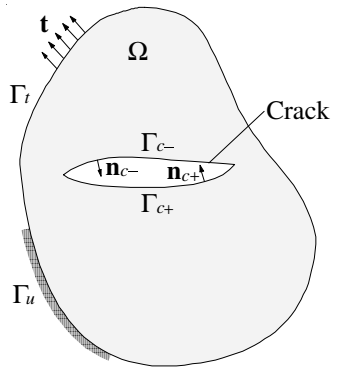



Figure 2: Schematic drawing of a cracked body

and the linearization is derived as

$$\left[\mathrm{D}\mathbf{r}_{\Omega}[\delta\mathbf{u}]\right]_i = \frac{\partial \mathsf{D}_{ijkl}}{\partial x_j}\frac{\partial^s \delta u_k}{\partial^s x_l} + \mathsf{D}_{ijkl}\frac{\partial}{\partial x_j}\frac{\partial^s \delta u_k}{\partial^s x_l} \quad \text{in } \Omega \tag{34}$$

$$\mathrm{D}\mathbf{r}_{\Gamma_t}[\delta\mathbf{u}] = (\mathsf{D} : \nabla^s \delta\mathbf{u})\mathbf{n} \quad \text{on } \Gamma_t \tag{35}$$

$$\mathrm{D}\mathbf{r}_{\Gamma_u}[\delta\mathbf{u}] = \delta\mathbf{u} \quad \text{on } \Gamma_u \tag{36}$$

$$\mathrm{D}\mathbf{r}_{\Gamma_{c+}}[\delta\mathbf{u}] = (\mathsf{D} : \nabla^s \delta\mathbf{u})\mathbf{n}_{\mathbf{c}+} - \mathrm{D}\mathbf{t}_{c+}[\delta\mathbf{u}] \quad \text{on } \Gamma_{c+} \tag{37}$$

$$\mathrm{D}\mathbf{r}_{\Gamma_{c-}}[\delta\mathbf{u}] = (\mathsf{D} : \nabla^s \delta\mathbf{u})\mathbf{n}_{\mathbf{c}-} - \mathrm{D}\mathbf{t}_{c-}[\delta\mathbf{u}] \quad \text{on } \Gamma_{c-} \tag{38}$$

where the summation convention is applied in Eq. (34) and $\mathsf{D}$ is called a tangent modulus tensor. The operator $\nabla^s$ is defined as

$$\nabla^s \delta\mathbf{u} = \frac{\partial^s \delta u_i}{\partial^s x_j} = \frac{1}{2}\left(\frac{\partial \delta u_i}{\partial x_j} + \frac{\partial \delta u_j}{\partial x_i}\right) \tag{39}$$

with the summation convention. The tangent modulus for the isotropic elastic material is written by

$$\mathsf{D} = 2G\mathsf{I}_d + K\mathbf{I}\otimes\mathbf{I} \tag{40}$$

where $K$ is the bulk modulus, $G$ is the transverse elasticity modulus, $\mathsf{I}_d$ is the deviatoric projection tensor, and $\mathbf{I}$ is the identity tensor. By discretizing Eqs. (29)–(38) through the C-MLS approximation, linear algebraic equations related to Eqs. (21) and (22) for the iteration procedure are obtained from

$$\mathbf{K}\delta\mathbf{u} = -\mathbf{r} \tag{41}$$

where $\mathbf{K}$ is a coefficient matrix (i.e., Jacobian matrix) and $\mathbf{r}$ is the residual vector. When the $(k)$th step is updated to the $(k+1)$th step in the iteration for Newton's method, the above formulation is decomposed as

$$\begin{bmatrix} \mathbf{K}_{\Omega} \\ \mathbf{K}_{\Gamma_t} \\ \mathbf{K}_{\Gamma_u} \\ \mathbf{K}_{\Gamma_{c+}} \\ \mathbf{K}_{\Gamma_{c-}} \end{bmatrix} \begin{pmatrix} \delta\mathbf{u}_{\Omega} \\ \delta\mathbf{u}_{\Gamma_t} \\ \delta\mathbf{u}_{\Gamma_u} \\ \delta\mathbf{u}_{\Gamma_{c+}} \\ \delta\mathbf{u}_{\Gamma_{c-}} \end{pmatrix} = - \begin{pmatrix} \mathbf{r}_{\Omega}^{(k)} \\ \mathbf{r}_{\Gamma_t}^{(k)} \\ \mathbf{r}_{\Gamma_u}^{(k)} \\ \mathbf{r}_{\Gamma_{c+}}^{(k)} \\ \mathbf{r}_{\Gamma_{c-}}^{(k)} \end{pmatrix} \tag{42}$$

where $\mathbf{r}_{\Omega}^{(k)}$, $\mathbf{r}_{\Gamma_t}^{(k)}$, $\mathbf{r}_{\Gamma_u}^{(k)}$, $\mathbf{r}_{\Gamma_{c+}}^{(k)}$, $\mathbf{r}_{\Gamma_{c-}}^{(k)}$, $\mathbf{K}_{\Omega}$, $\mathbf{K}_{\Gamma_t}$, $\mathbf{K}_{\Gamma_u}$, $\mathbf{K}_{\Gamma_{c+}}$, and $\mathbf{K}_{\Gamma_{c-}}$ are sub-matrices obtained by the discretization of Eqs. (29)–(38), respectively. In particular, $\mathbf{K}_{\Gamma_u}$ is simply expressed as follows owing to the C-MLS approximation:

$$\mathbf{K}_{\Gamma_u} = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \end{bmatrix} \tag{43}$$

where the elements of $\mathbf{I}$ are located corresponding to $\delta\mathbf{u}_{\Gamma_u}$. The coefficient matrix is sparse and banded but is generally non-symmetric because it is based on the strong formulation. Equation (41) should be linearly independent so as not to cause singularity of the coefficient matrix.

The steps of the method for updating the time step $t_n$ to $t_{n+1}$ are written below:

*Phase 1.* Construct the coefficient matrix $\mathbf{K}$ of discretized Eqs. (34)–(38).

*Phase 2.* Obtain $\delta\mathbf{u}$ by using the numerical solution for Eq. (41) and update the displacement $\mathbf{u}_{n+1}^{(k)}$ to $\mathbf{u}_{n+1}^{(k+1)}$ for the $(k+1)$th iteration step given by Eq. (23).

*Phase 3.* Update the stress $\boldsymbol{\sigma}_{n+1}^{(k+1)}$ and traction on the crack faces $\mathbf{t}_{c+} = -\mathbf{t}_{c-}$.

*Phase 4.* Compute the residual vector $\mathbf{r}_{n+1}^{(k+1)}$ by using discretized Eqs. (29)–(33).

*Phase 5.* If the convergence criteria are satisfied, end the iteration and take the next time step; otherwise, continue the iteration by returning to Phase 1.

The shape function approximated by C-MLS is constructed only once before the above procedure because of the hypothesis of a small deformation. Newton's method is widely used to solve nonlinear problems, so this procedure has a framework based on Newton's method that is basically common to conventional numerical methods such as the finite element method (FEM). However, the coefficient matrix and linear algebraic equations derived from the strong-form governing equations distinctly differ from those employed in FEM when the weak form is assumed. Hence, the presented method, which needs no meshes or numerical integration, can achieve a simpler scheme and reduce the computational costs for iterations.

### 2.4 Crack propagation modeling

In order to deal with the displacement discontinuity due to cracks, the so-called visibility criterion [19] used in other meshless methods is employed with the C-MLS approximation. In contrast to the conventional FEM, the meshless interpolation procedure used in meshless methods such as MLS can adjust the size and shape of its domain of influence for the crack discontinuity. The visibility criterion simply cuts through the domain of influence and ignores the nodes placed at the opposite side of the crack, as shown in Fig. 3. As a result, the C-MLS approximation modified by the visibility criterion allows the displacement field to contain jump discontinuities across the crack face.

Most other meshless methods employ a special function in the approximation to deal with the singularity of the stress field near the crack tip [19], but there have been few reports suggesting a special function that can be applied to complex three-dimensional singularity. In this work, the solution at the concave corner including the crack tip is defined by superposition of multiple nodal solutions. As an example, the methodology for the two-dimensional case is depicted in Fig. 4 (a). Consider the concave corner node and two lines modeling surfaces $\Gamma_1$ and $\Gamma_2$ with outward normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively. Assume that the natural boundary condition of the concave corner node is approximated by the following average:

$$\boldsymbol{\sigma}_1\mathbf{n}_1 + \boldsymbol{\sigma}_2\mathbf{n}_2 = \mathbf{t}_1 + \mathbf{t}_2 \tag{44}$$

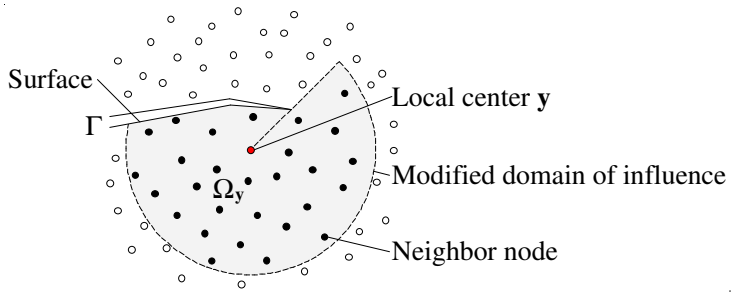
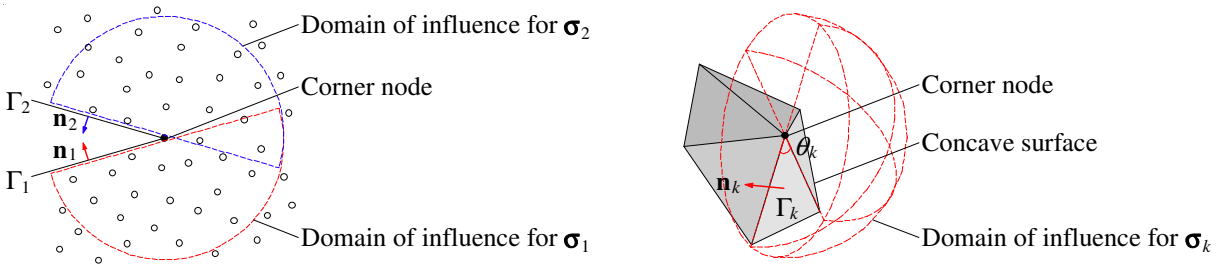
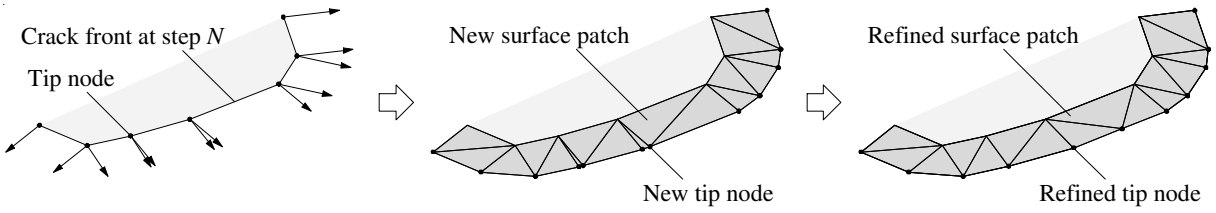



Figure 3: Modification of the domain of influence by the visibility criterion




(a) two dimensions (b) three dimensions

Figure 4: Corner node located at concave boundary surfaces




(a) Locate new tip nodes (b) Extend crack front at step $N$+1 (c) Refine crack front at step $N$+1

Figure 5: Crack propagation modeled in surface patches and tip nodes

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are the surface tractions. The stress $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ are computed by using shape functions approximated by C-MLS with the visibility criterion corresponding to each surface. These are derived for each domain of influence, as shown in Fig. 4 (a). In three dimensions, the crack faces is represented by a polyhedron with triangle patches, as illustrated in Fig. 4 (b), and approximate the natural boundary condition of the concave corner node as the following extension of the two-dimensional case:

$$\sum_{k=1}^{n} \theta_k \boldsymbol{\sigma}_k \mathbf{n}_k = \sum_{k=1}^{n} \theta_k \mathbf{t}_k \tag{45}$$

where $n$ is the number of surface patches around the corner node and $\theta_k$, $\boldsymbol{\sigma}_k$, $\mathbf{n}_k$, and $\mathbf{t}_k$ are the angle between two sides, the stress, and the traction corresponding to each surface patch. When using individual nodal values, the stress at the corner node is represented by $\boldsymbol{\sigma} = (\sum_{k=1}^{n} \theta_k \boldsymbol{\sigma}_k \mathbf{n}_k)/(\sum_{k=1}^{n} \theta_k)$. The above procedure is similar to the treatment for the natural boundary condition at the corner in the finite difference method (FDM) in terms of averaging the directional derivatives for the surface patches to obtain the approximation. The accuracy of this procedure for the singularity of a crack tip was tested in an example described below.

Next, the procedure of crack propagation is explained. The point collocation method is a truly meshless method based on a strong formulation. Thus, geometric changes to the analysis model, such as the addition and subtraction of nodes and surfaces, are applied more easily than in conventional mesh methods. As shown in Fig. 5, the crack front is composed of tip lines and tip nodes on a collection of surface patches representing crack faces. Crack propagation is defined as the motion of the crack front. The advance of the crack front is expressed by the locations of the newly added tip nodes, and the direction of the crack propagation is assumed according to the maximum principal stress. For each tip node, the direction of crack propagation is defined as perpendicular to the direction of the maximum principal stress and the direction along the tip line. In other words, two direction corresponding to the tip lines

are defined, and two new tip nodes are generated for each tip node. The length of the crack growth $d_c$ for each increment is predefined as a sufficiently small value to simulate the actual crack shape. Once the positions of the new tip nodes are determined, the crack faces need to be updated. This is easily done in two dimensions by the addition of tip lines. In three dimensions, a series of surface patches is constructed. In addition, the surface patches are refined if the crack front edge is too long or too short. Specifically, a crack edge that is longer or shorter than the user-specified threshold is subdivided or deleted, as shown in Fig. 5. Most conventional mesh methods have difficulty with applying the above geometric changes due to crack propagation because the change in a system based on the weak formulation causes additional features to be implemented. For instance, the change in element formulation due to the addition and subtraction of nodes, consistency between the mesh connectivity and continuity of the fracture zone location, and others need to be considered. In contrast, the presented method maintains simplicity and effectiveness due to the strong formulation. Therefore, this is a significant advantage of the presented method for dealing with crack propagation.

### 2.5 Fracture criterion and cohesive crack model

This section describes the physical model of the crack propagation. The presented method employs a Rankine-type fracture criterion where the crack is assumed to be initiated when the maximum principal stress exceeds the uniaxial tensile strength. The crack face is introduced perpendicular to the direction of the maximum principal stress, as noted previously. For some geomaterials such as concrete, this provides sufficiently accurate results if no significant lateral compression is present [20].

When the fracture criterion is met, discrete crack faces are introduced by adding the surface patches and nodes. Then, a cohesive crack model is assumed, i.e., cohesive traction is applied along newly generated crack faces. In other words, the crack faces of the model represent the interface between the cohesive zone and the surrounding material. The cohesive traction depends on a cohesive law that relates the traction to the jump in displacement across the discontinuity of the crack faces. The relative displacement across the crack faces is represented by

$$\boldsymbol{w}_c = \mathbf{x}_{c-} - \mathbf{x}_{c+} \tag{46}$$

where $\mathbf{x}_{c+}$ and $\mathbf{x}_{c-}$ are the positions of the nodes on the pair of surface patches. In this work, only the normal component of the cohesive traction was considered. The cohesive traction prescribed on the surface patch $\Gamma_{c+}$ is written as follows:

$$\mathbf{t}_{c+} = t_c(w_c)\mathbf{n}_{c+} \tag{47}$$

where $\mathbf{t}_{c+} = -\mathbf{t}_{c-}$ is the cohesive traction on $\Gamma_{c+}$, $t_c(w_c)$ is the cohesive traction force determined by the cohesive law, $w_c$ is the normal component of $\boldsymbol{w}_c$, and $\mathbf{n}_{c+}$ is the surface normal. The cohesive traction is linearized as follows:

$$\mathrm{D}\mathbf{t}_{c+}[\delta\mathbf{u}] = \frac{\partial t_c}{\partial w_c}\left(\mathbf{n}_{c+} \otimes \mathbf{n}_{c+}\right)\delta\boldsymbol{w}_c \tag{48}$$

where $\partial t_c/\partial w_c$ is the modulus computed by the cohesive law. In this study, an exponential cohesive law was employed. As shown in Fig. 6, the energy dissipated by the development of the crack opening matches the fracture energy. Thus, $t_c$ is related to $w_c$ by the following softening relationship:

$$t_c(w_c) = f_t \exp\left(-\frac{f_t}{G_f} w_c\right) \tag{49}$$

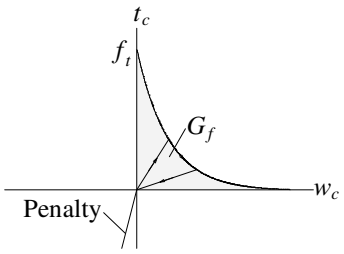



Figure 6: Schematic drawing of the exponential cohesive law

where $G_f$ is the specific fracture energy of the material and $f_t$ indicates the traction force at failure. In addition, note that a penalty force is introduced to consider the crack surface overlap. Most conventional methods apply the cohesive crack model to the material constitutive law to easily treat the discontinuity. In contrast, the presented method directly expresses the framework of the cohesive crack model because the crack surfaces and cohesive traction are simply introduced in combination with the crack growth procedure. Particularly in this regard, the presented method is attractive for the analysis of brittle and quasi-brittle fracture.

## 3 NUMERICAL EXAMPLES

### 3.1 Near-tip crack field problem

A patch test for a singular field was performed. A closed-form solution for a crack can be derived by using the known near-tip field in a domain about the crack tip. A cuboid region was considered as shown in Fig. 7, and the known near-tip displacement fields from the solution for a mode I crack were prescribed along the boundaries parallel to the tip line. Symmetric boundary conditions were imposed on the surfaces perpendicular to the tip line to assume the plane strain condition, and traction-free boundary conditions were applied to the crack faces. This example is an intrinsically two-dimensional problem, and the displacement for the near-tip field is given by

$$u_x = \frac{K_I}{2G}\sqrt{\frac{r}{2\pi}}\cos\left(\frac{\theta}{2}\right)\left[\kappa - 1 + 2\sin^2\left(\frac{\theta}{2}\right)\right] \tag{50}$$

$$u_z = \frac{K_I}{2G}\sqrt{\frac{r}{2\pi}}\sin\left(\frac{\theta}{2}\right)\left[\kappa + 1 - 2\cos^2\left(\frac{\theta}{2}\right)\right] \tag{51}$$

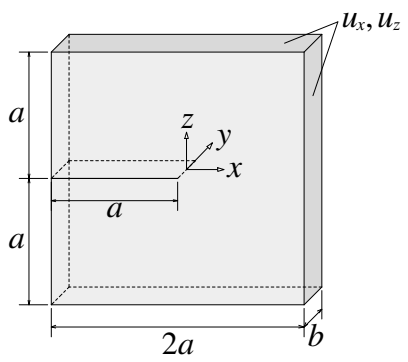
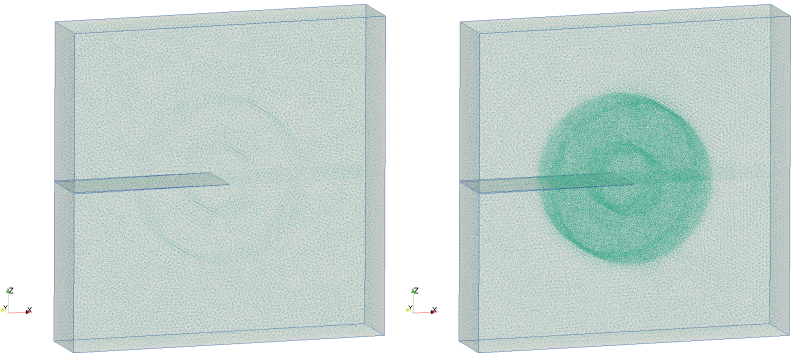



Figure 7: Cuboid region for a near-tip crack problem



(a) 95,335 nodes (b) 206,195 nodes

Figure 8: Node distribution for a near-tip crack problem

Table 1: Recalculated stress intensity factors for a near-tip crack problem

| Case | Number of nodes | Node density | Recalculated SIFs | Relative error (%) |
|---|---|---|---|---|
| 1 | 15,160 | Homogeneous | 0.99059 | 0.94 |
| 2 | 95,335 | Homogeneous | 0.99789 | 0.21 |
| 3 | 300,514 | Homogeneous | 1.00256 | 0.26 |
| 4 | 206,195 | Heterogeneous | 1.00005 | 0.01 |

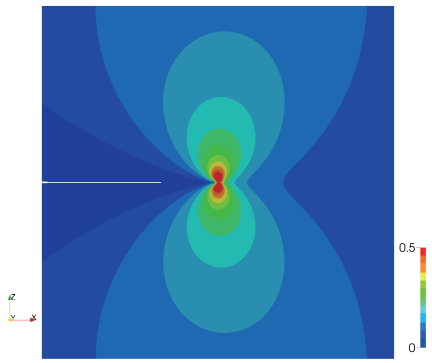



Figure 9: Von Mises stress in the section at $y = 0$ for a near-tip crack problem

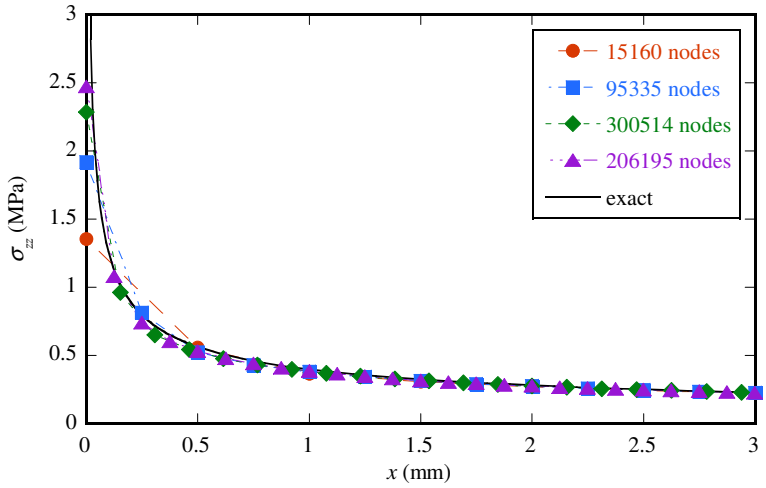



Figure 10: Distribution of the normal stress $\sigma_{zz}$ for the near-tip field along the line $y = z = 0$ for a near-tip crack problem

where $r$ is the distance from the crack tip line in the cylindrical coordinate system, $\theta$ is the angle measured from the line ahead of the crack, $K_I$ is the stress intensity factor, and $\kappa$ is the Kolosov constant defined as $\kappa = 3 - 4\nu$ for the plane strain [21]. The parameters of $a = 10$ mm, $b = 4$ mm, $K_I = 1$ MPa $\cdot$ mm$^{1/2}$, Young's modulus $E = 200$ GPa, and Poisson's ratio $\nu = 0.3$ were used. For the convergence test, four cases for the number of nodes given in Table 1 were used. As shown in Fig. 8, irregular patterns of the node distribution were adopted for each case. The node density of the case with 206,195 nodes was locally refined near the crack tip, and the others were nearly homogeneous.

Table 1 provides the stress intensity factors (SIFs) for the numerical results recalculated by the M-integral method [22], and Fig. 9 illustrates the von Mises stress distribution in a section of $y = 0$ for the case with 206,195 nodes. The recalculated SIFs showed good agreement with the prescribed value in every case. In addition, Fig. 10 plots the numerical results for the distribution of normal stress $\sigma_{zz}$ ahead of the crack tip along the line of $y = x = 0$ with the exact solution. The results indicated that the modeling approach described above worked successfully, even though the near-tip field was not enriched to capture the stress singularity at the crack tip strictly. Furthermore, it is confirmed that the C-MLS approximation provides the accurate numerical results as well as other meshless approximations.

### 3.2 Double-edge notched specimen under tension

The second example was the simple fracture analysis of a uniaxial tension test on a double-edge notched (DEN) specimen. The geometry and prescribed boundary conditions are illustrated in Fig. 11. The specimen was weakened by two notches to initialize fracture. Because

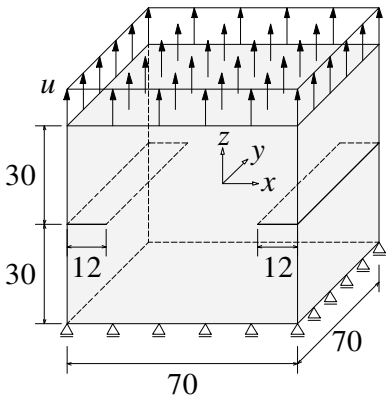



Figure 11: Geometry and applied boundary conditions for a DEN test

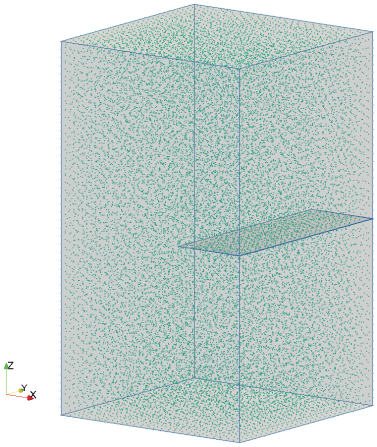
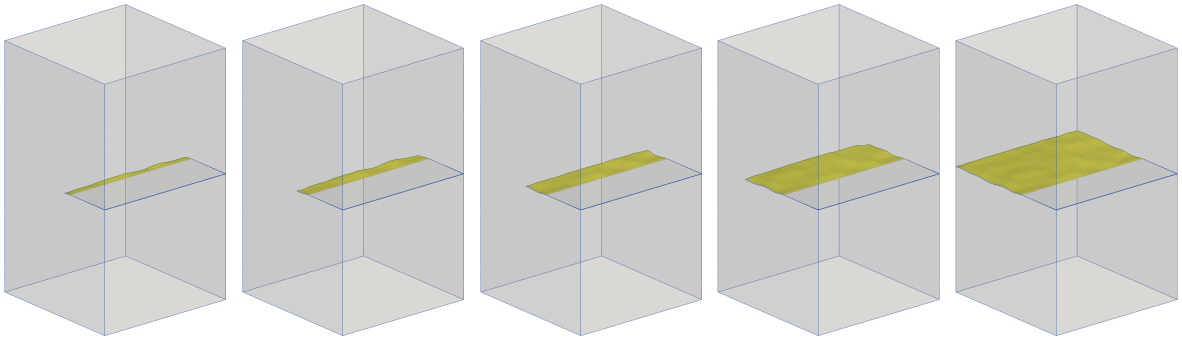


Figure 12: Node distribution for a DEN test

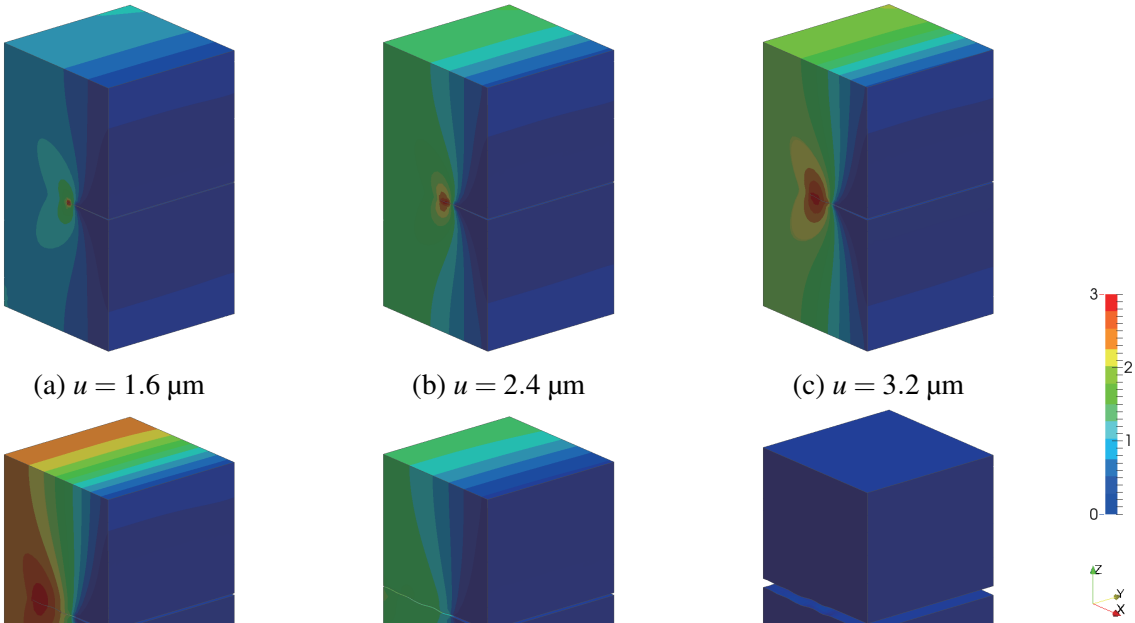


(a) $u = 1.6$ µm (b) $u = 2.4$ µm (c) $u = 3.2$ µm (d) $u = 4.0$ µm (e) $u = 4.8$ µm

Figure 13: Crack propagation for a DEN test

of the symmetry, one quarter of the specimen was discretized with proper symmetric boundary conditions, as displayed in Fig. 12. This also implies that the symmetry of the crack propagation was enforced. The numerical simulation was carried out for the irregular node distribution of 23,988 nodes and the crack increment length of $d_c = 1.04$ mm. The material parameters were taken to be equivalent to those of Comi et al. [23] as follows: $E = 36$ GPa, $\nu = 0.1$, $f_t = 3.0$ MPa and $G_f = 0.011$ N/mm. A displacement control procedure was employed based on the uniform displacement in the $z$-direction $u$ prescribed on the top of the specimen.

Fig. 13 depicts the resulting crack evolution and Fig. 14 provides the maximum principal stress distribution where the displacements are scaled by a factor of 100. Two planar cracks extended from each notch until they joined within the specimen, and the specimen was completely cut by the crack at the end. Note that the fracture process can be nonsymmetric if the symmetric property is not considered because a possible scenario is that only one of the two cracks extends while the other crack unloads owing to the material imperfection. Fig. 15 compares the load–displacement response with the numerical results obtained by using the extended finite element method (XFEM) with a continuum damage model [23]. Both response curves had a rapid drop after the peak load, although the falls in the load after the rapid drop were different. The numerical results of the presented method seem to qualitatively agree with the experimental results [24].




(a) $u = 1.6$ µm (b) $u = 2.4$ µm (c) $u = 3.2$ µm

(d) $u = 4.0$ µm (e) $u = 4.8$ µm (f) $u = 20.0$ µm

Figure 14: Maximum principal stress for a DEN test

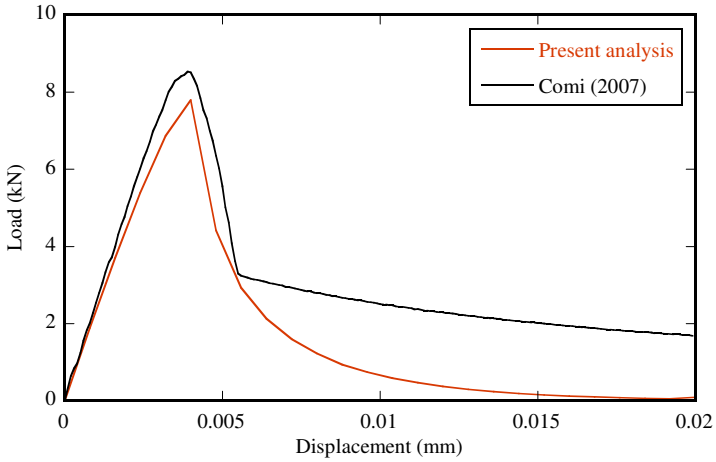



Figure 15: Reaction load versus prescribed displacement for a DEN test

### 3.3 Anchor pull-out test

The next example was the pull-out test of a steel anchor embedded in a cylindrical concrete block. The dimensions and material parameters were set to $E = 30$ GPa, $\nu = 0.2$, $f_t = 3.0$ MPa, and $G_f = 0.106$ N/mm, which are similar to those used by Rots [20] and Gasser and Holzapfel [25]. Fig. 16 displays a quarter of the geometric data and applied boundary conditions. A vertical load $F$ was imposed on the center of the steel disk, and the embedded disc was pulled against a counter pressure placed concentric with the disc on the surface until the

surrounding concrete fractured. As shown in Fig. 17, a quarter of the whole structure was modeled, and symmetric boundary conditions were applied. Note that the three-dimensional model was used in this work to investigate the presented method in three dimensions, although this problem is intrinsically axisymmetric and can be modeled more simply. The irregular node distribution of 23,232 nodes and crack increment length of $d_c = 18$ mm were used. The steel anchor was not explicitly modeled; instead, an incremental displacement $u$ in the $z$-direction was prescribed at the nodes of the upper contact surface between the steel disc and concrete by assuming that the steel disc was rigid. The numerical analysis is conducted under displacement control; a constant incremental displacement of 0.025 mm was imposed in each load step until

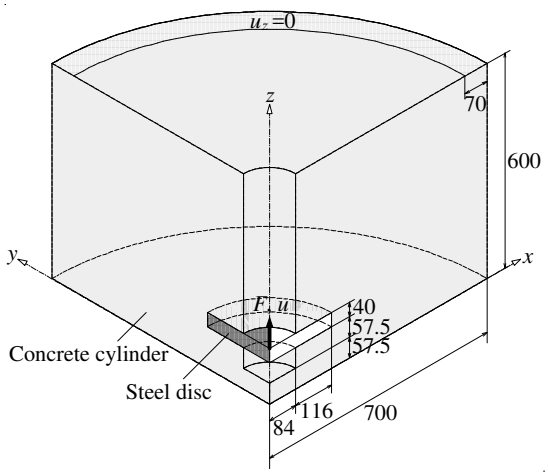



Figure 16: Geometry and applied boundary conditions for an anchor pull-out test

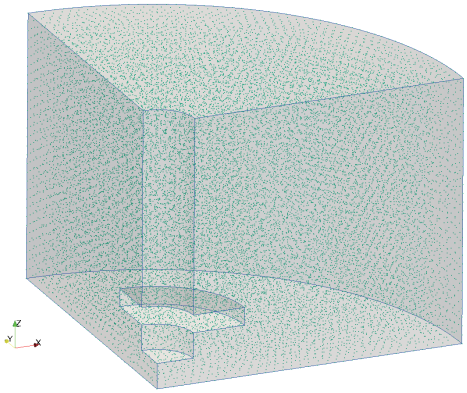
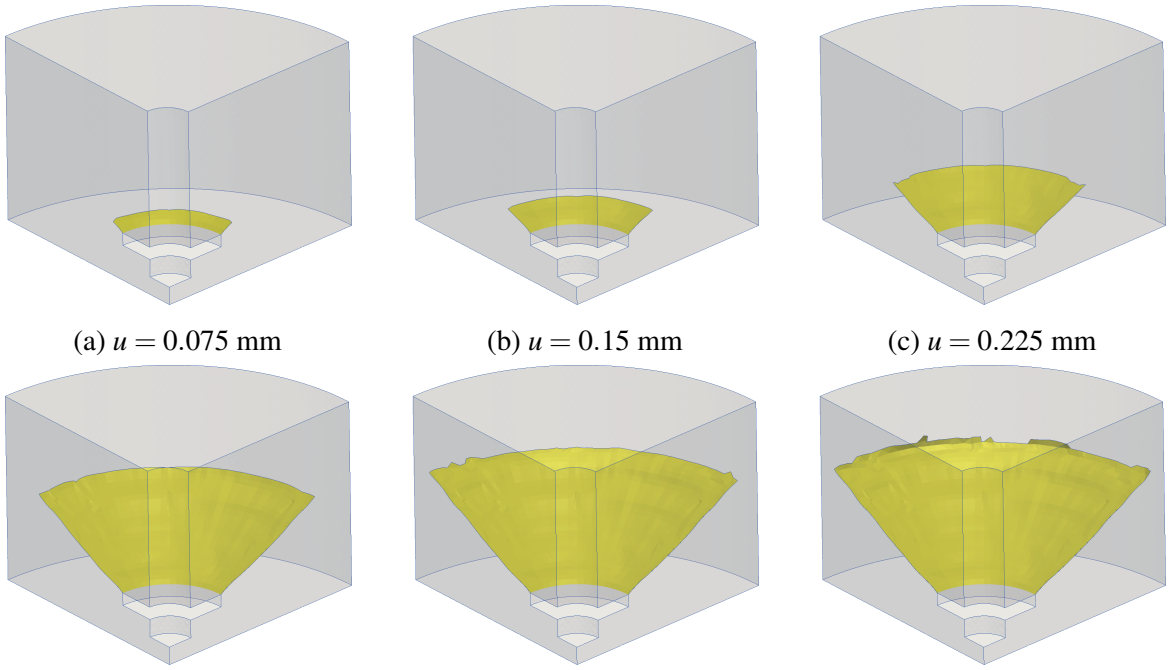


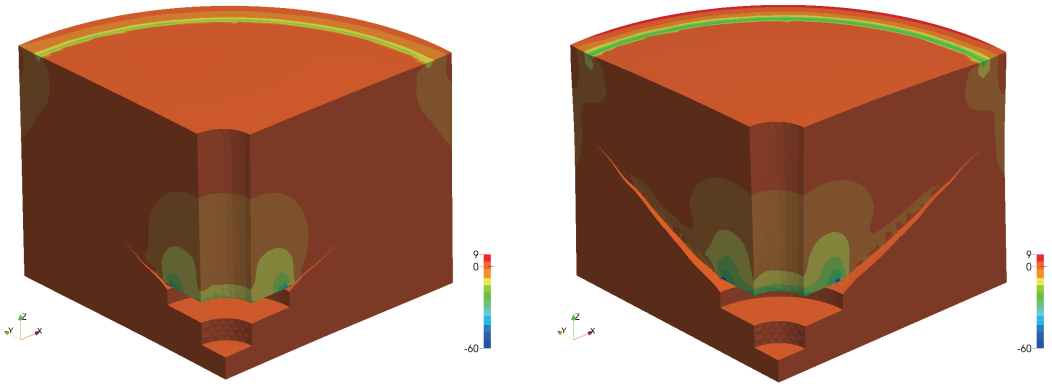
Figure 17: Node distribution for an anchor pull-out test



(a) $u = 0.075$ mm (b) $u = 0.15$ mm (c) $u = 0.225$ mm

(d) $u = 0.275$ mm (e) $u = 0.35$ mm (f) $u = 0.4$ mm

Figure 18: Crack propagation for an anchor pull-out test



(a) $u = 0.225$ mm (b) $u = 0.4$ mm

Figure 19: Normal stress in the loading direction $\sigma_{zz}$ for an anchor pull-out test

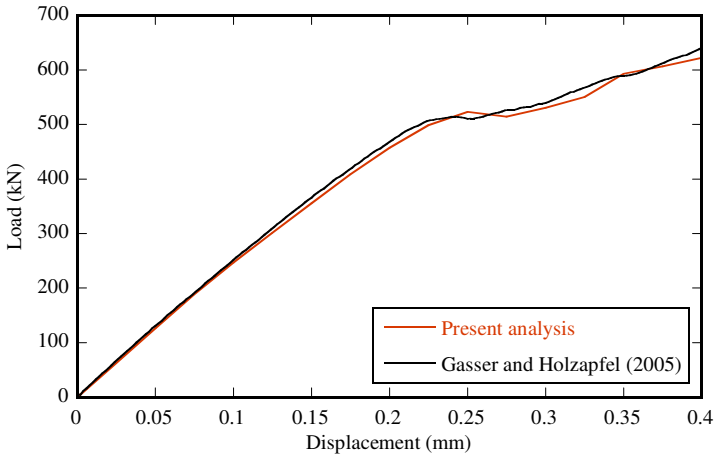



Figure 20: Reaction load versus prescribed displacement for an anchor pull-out test

the displacement reached the final value of 0.4 mm.

Fig. 18 depicts the propagating crack surface of the concrete cylinder, where the different stages are associated with different levels of prescribed displacements. The crack initiated from the outer edge of the upper contact surface of the steel disc into the concrete, and the final crack surface exhibited a conical shape similar to those observed in the experimental results [20] and numerical results of the literature [25]. Fig. 19 shows the normal stress distribution plotted on the deformed representation corresponding to the results of the load steps at the applied displacements $u = 0.225$ mm and $u = 0.4$ mm, respectively. Finally, Fig. 20 shows the evolution of the corresponding load versus the prescribed displacement, where the resulting curve is compared with the numerical result from Gasser and Holzapfel [25] using PUFEM. The load–displacement curve was almost linear until a load of about 500 kN was reached. Afterwards, there was a short drop in the reaction load with an acceleration of the succeeding crack propagation. The accelerated crack growth ended when the crack front reached below the ring zone to which the displacement boundary condition $u_z = 0$ was applied. Finally, the load increased again when the displacement went beyond about 0.28 mm. Both the aforementioned phenomena and load–displacement response showed good agreement with the numerical results in the literature [25].

## 4 CONCLUSIONS

This paper presents a meshless point collocation method for three-dimensional crack propagation. The presented method yields a truly meshless fracture analysis by using C-MLS approximation, node-wise discretization based on the strong formulation, and crack propagation represented by the addition of nodes and surfaces. The C-MLS approximation provides derivative approximations without direct differentiation and does not require any special implementation to treat essential boundary conditions. The geometric change due to crack growth is simply represented by the addition and subtraction of nodes and surfaces, and the cohesive crack model is employed in association with the crack propagation procedure to directly embody the fracture of quasi-brittle materials in a sufficient manner. Numerical examples were performed to demonstrate the accuracy and effectiveness of the presented method. This method has the potential to solve other problems involving geometric and boundary nonlinearity, such as large deformations.

## REFERENCES

[1] L. B. Lucy, A numerical approach to the testing of the fission hypothesis, Astronom. J. 82 (1977) 1013–1024.

[2] R. A. Gingold, J. J. Monaghan, Smoothed particle hydrodynamics: theory and application to non-spherical stars, Mon. Not. R. Astron. Soc. 181 (3) (1977) 375–389.

[3] B. Nayroles, G. Touzot, P. Villon, Generalizing the finite element method: diffuse approximation and diffuse elements, Comput. Mech. 10 (5) (1992) 307–318.

[4] T. Belytschko, Y. Y. Lu, L. Gu, Element-free Galerkin methods, Int. J. Numer. Meth. Eng. 37 (2) (1994) 229–256.

[5] W. K. Liu, S. Jun, S. Li, J. Adee, T. Belytschko, Reproducing kernel particle methods for structural dynamics, Int. J. Numer. Meth. Eng. 38 (10) (1995) 1655–1679.

[6] J. M. Melenk, I. Babuška, The partition of unity finite element method: basic theory and applications, Comput. Meth. Appl. Mech. Eng. 139 (1) (1996) 289–314.

[7] P. Lancaster, K. Salkauskas, Surfaces generated by moving least squares methods, Math. Comp. 37 (155) (1981) 141–158.

[8] E. Onate, S. Idelsohn, O. Zienkiewicz, R. Taylor, A finite point method in computational mechanics. Applications to convective transport and fluid flow, Int. J. Numer. Meth. Eng. 39 (22) (1996) 3839–3866.

[9] N. Aluru, A point collocation method based on reproducing kernel approximations, Int. J. Numer. Meth. Eng. 47 (6) (2000) 1083–1121.

[10] Y.-C. Yoon, J.-H. Song, Extended particle difference method for weak and strong discontinuity problems: part I. Derivation of the extended particle derivative approximation for the representation of weak and strong discontinuities, Comput. Mech. 53 (6) (2014) 1087–1103.

[11] E. Oñate, F. Perazzo, J. Miquel, A finite point method for elasticity problems, Comput. Struct. 79 (22) (2001) 2151–2163.

[12] L. P. Pozo, F. Perazzo, A. Angulo, A meshless FPM model for solving nonlinear material problems with proportional loading based on deformation theory, Adv. Eng. Softw. 40 (11) (2009) 1148–1154.

[13] W. Tu, Y. Gu, P. Wen, Effective shear modulus approach for two dimensional solids and plate bending problems by meshless point collocation method, Eng. Anal. Boundary Elem. 36 (5) (2012) 675–684.

[14] S.-H. Lee, Y.-C. Yoon, Meshfree point collocation method for elasticity and crack problems, Int. J. Numer. Meth. Eng. 61 (1) (2004) 22–48.

[15] H. Noguchi, T. Noguchi, Application of new gridless method using least square method with constraint condition to structural problems, in: Proceedings of the Conference on Computational Engineering and Science, Vol. 2, 1997, pp. 391–394.

[16] H. Minaki, H. Noguchi, Large deformation elastic-plastic analysis by SMAC-SPH using constrained moving least squares method, in: The Computational Mechanics Conference, Vol. 2005, 2005, pp. 763–764.

[17] G. Barenblatt, The formation of equilibrium cracks during brittle fracture. General ideas and hypotheses. Axially-symmetric cracks, J. Appl. Math. Mech. 23 (3) (1959) 622–636.

[18] D. W. Kim, H.-K. Kim, Point collocation method based on the FMLSRK approximation for electromagnetic field analysis, IEEE Trans. Magn. 40 (2) (2004) 1029–1032.

[19] M. Fleming, Y. Chu, B. Moran, T. Belytschko, Y. Lu, L. Gu, Enriched element-free Galerkin methods for crack tip fields, Int. J. Numer. Meth. Eng. 40 (8) (1997) 1483–1504.

[20] J. G. Rots, Computational modeling of concrete fracture, Ph.D. thesis, Delft University of Technology (1988).

[21] M. Williams, On the stress distribution at the base of a stationary crack, J. Appl. Mech. 24 (1957) 109–114.

[22] J. Yau, S. Wang, An analysis of interface cracks between dissimilar isotropic materials using conservation integrals in elasticity, Eng. Fract. Mech. 20 (3) (1984) 423–432.

[23] C. Comi, S. Mariani, U. Perego, An extended fe strategy for transition from continuum damage to mode I cohesive crack propagation, Int. J. Numer. Anal. Methods Geomech. 31 (2) (2007) 213–238.

[24] D. A. Hordijk, Local approach to fatigue of concrete, Ph.D. thesis, TU Delft, Delft University of Technology (1991).

[25] T. C. Gasser, G. A. Holzapfel, Modeling 3D crack propagation in unreinforced concrete using PUFEM, Comput. Meth. Appl. Mech. Eng. 194 (25) (2005) 2859–2896.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A cell-based smoothed point interpolation method for axisymmetric problems

**A. Tootoonchi[*] and A. Khoshghalb[*]**

* School of Civil and Environmental Engineering, UNSW Australia, Sydney, NSW, 2052, Australia

**KEY WORDS:** axisymmetric, meshfree methods, smoothing gradient operation, saturated porous media, point interpolation method.

**Abstract.** *A cell-based smoothed point interpolation method (CSPIM) for the numerical modelling of saturated porous media in axisymmetric conditions is proposed. Spatial discretisation is performed using a simple triangulation. Both displacement and fluid pressure gradients are smoothed by incorporating the smoothing gradient operation technique. Thus, the integration over the supporting domains (i.e. elements) is transformed to that over the boundary of the elements. The field variables, the displacement and excess pore water pressure, are interpolated using point interpolation shape functions (polynomial and radial), which possess the Kronecker function property and facilitate imposing the essential boundary conditions. The global property matrices of the discretised system of equations are derived using the generalised smoothed Galerkin method and the three-point time discretisation scheme for discretisation of the governing equations in space and time, respectively. A novel approach is introduced for the computation of the property matrices, which avoids the singularity problem that will otherwise arise when cell-based smoothed interpolation method is used in axisymmetric conditions. The salient feature of the proposed method is that it incurs no additional computational costs, and it does not compromise on accuracy of the method. The validity of the proposed method is investigated by simulation of Cryer's benchmark consolidation problem. The numerical results are in excellent agreement with the analytical solution.*

## Introduction

A wide class of efficient smoothed point interpolation methods (SPIMs) have been recently developed [1, 2] by incorporating strain smoothing operation [2] to point interpolation methods (PIMs). In SPIMs, instead of using a compatible strain field, a smoothed strain field is constructed through a smoothing operation performed over smoothing domains. The smoothed strain field overcomes the problems associated with continuity of the approximation field over the problem domain through elimination of the need for the derivatives of the shape functions. Different smoothing domains can be adopted for the smoothing operation leading to different SPIMs [3-6]. The simplest SPIM is perhaps the cell-based SPIM (CSPIM) in which the cells of a triangular background mesh are used as the smoothing domains. To date, CSPIM has been successfully applied to plane strain problems in geotechnical engineering with results showing its superiority to similar methods (e.g. finite element methods) in terms of both efficiency and convergence rate [7].

Axisymmetric problems are of considerable importance in geotechnical engineering because of their many applications including simulation of uni-axial and tri-axial loading conditions, pile installation, and tunnels. Nonetheless, application of SPIMs to axisymmetric problems has received little attention in the literature. This may be attributed to the difficulties associated with application of SPIMs in axisymmetric conditions due to the existence of Gauss points on the boundary of the smoothing domains on the axis of the symmetry. In particular, CSPIM cannot be directly extended to axisymmetric conditions due to singularity problem arises in the analysis.

In this paper, a novel approach is presented to extend the formulation of the CSPIM to axisymmetric conditions. The approach is general in nature and can also be applied to other SPIMs to extend their application to axisymmetric problems. The technique presented does not include any approximation and therefore, does not adversely affect the accuracy of the numerical procedure. The validity of the proposed method is verified by analysis of Cryer's benchmark problem [8].

## Governing equations

The equations governing flow-deformation of saturated porous media, first presented by Biot [9], are derived based on the momentum balance of the solid and fluid phases and mass conservation of the fluid phase, as

$$\boldsymbol{\partial}^T\left[\boldsymbol{D}\dot{\boldsymbol{\varepsilon}} + \eta\dot{\tilde{p}}_f\boldsymbol{\delta}\right] + \dot{\boldsymbol{F}} = \boldsymbol{0} \tag{1}$$

$$\boldsymbol{\nabla}^T\left(\frac{\boldsymbol{k}_f}{\mu_f}\left(\boldsymbol{\nabla}\tilde{p}_f + \rho_f\boldsymbol{g}\right)\right) = a_f\dot{\tilde{p}}_f - \boldsymbol{\nabla}^T\dot{\tilde{\boldsymbol{u}}} \tag{2}$$

with

$$a_f = n\left(c_f - c_s\right) + \eta c_s \quad , \quad \eta = 1 - \frac{c_s}{c} \tag{3}$$

where $\boldsymbol{D}$ is the drained stiffness matrix of the soil, $\boldsymbol{\varepsilon}$ is the soil skeleton strain, $\boldsymbol{\delta}$ is the Kronecker delta, $\boldsymbol{F}$ is the vector of body force per unit volume, $a_f$ is the apparent compressibility of the fluid, $c_f$ is the compressibility of the fluid, $c_s$ is the compressibility of solid grains, $c$ is the drained compressibility of the solid skeleton and $\eta$ is the effective stress parameter. $\boldsymbol{k}_f$ is the intrinsic permeability matrix of the soil, $\mu_f$ is the dynamic viscosity of the fluid and $\rho_f$ is the intrinsic mass densities of the fluid. $\tilde{\boldsymbol{u}}$ stands for the displacement vector of the soil skeleton and $\tilde{p}_f$ is the pore fluid pressure (the tilde denotes a continuous field as opposed to the discretised field introduced later). $\boldsymbol{g}$ is the vector of gravitational acceleration and $n$ is the soil porosity. $\boldsymbol{\partial}$ and $\boldsymbol{\nabla}$ are the differentiation matrix and the gradient operator, respectively.

**Construction of the shape functions**

Radial Point Interpolation Method (RPIM) [10] is adopted in this study for the construction of the nodal shape functions. To derive RPIM shape functions, the field approximation function, $u(\boldsymbol{x})$, in the space coordinates $\boldsymbol{x}$=$[r, z]$ is approximated at any point in the problem domain with the following series representation:

$$u(\boldsymbol{x}) = \sum_{i=1}^{p} a_i R_i(\boldsymbol{x}) + \sum_{j=1}^{l} b_j p_j(\boldsymbol{x}) = \boldsymbol{R}^T(\boldsymbol{x})\boldsymbol{a} + \boldsymbol{p}^T(\boldsymbol{x})\boldsymbol{b} \tag{4}$$

in which $R_i(\boldsymbol{x})$ and $p_j(\boldsymbol{x})$ are the radial basis functions (RBFs) and monomial terms, respectively; $a_i$ is the coefficient for radial basis $R_i(\boldsymbol{x})$, and $b_j$ is the coefficient for the polynomial basis $p_j(\boldsymbol{x})$. $p$ is the number of nodes in the compact support domain of the point of interest, and $j$ is the number of monomial terms. Detailed discerption of derivation of shape functions can be found in previous works [10-14].

**Cell-based smoothed strains**

In CSPIM, the cells of the background mesh serve as the smoothing domains, with the smoothed strains computed for each cell through the smoothing operation technique. The smoothed strain for cell $k$ can be calculated from [2],

$$\hat{\boldsymbol{\varepsilon}}_k = \hat{\boldsymbol{B}}_1 \boldsymbol{u} = \frac{1}{\Omega^k} \int_{\Omega^k} \boldsymbol{\nabla} u(\boldsymbol{x}) d\Omega^k = \frac{1}{\Omega^k} \int_{\Gamma^k} \boldsymbol{L}_n u(\boldsymbol{x}) d\Gamma^k \tag{5}$$

where $\hat{\boldsymbol{B}}_1$ is the smoothed strain-displacement matrix for each smoothing domain, which is obtained by the summation of the smoothed strain-displacement matrices computed at the Gauss points on the edges of the smoothing domain, as follows:

$$\hat{\boldsymbol{\varepsilon}}_k = \sum_{i=1}^{q} \hat{\boldsymbol{B}}_{1i} \boldsymbol{u}_i \qquad\qquad i \in S_s \tag{6}$$

where $S_s$ is the set of $q$ support nodes that includes all the nodes involved in the interpolation of the quadrature points located on all segments of the boundary $\Gamma^k$ for the smoothing domain $\Omega^k$. $\boldsymbol{L}_n$ in equation (6) is unit outward operator for each segment of the boundary of cell $k$. Further details about the computation of the smoothed strain-displacement matrix can be found in [7].

### Node selection schemes

In CSPIM, a number of node selection schemes, known as T-schemes, have been proposed for the selection of the supporting nodes at each point of interest. Comprehensive details of the T-Schemes can be found in Liu and Zhang [2]. In the current study, a simple node selection scheme, referred to as T4, is adopted. This scheme selects four nodes of the two cells that share the edge hosting the Gauss point of interest. A linear approximation is applied when the Gauss point of interest is located on the boundary of the domain.

### Spatial and temporal discretisations

Using the Generalised Smoothed Galerkin (GS Galerkin) approach [2], the spatial discretisation of equations (1) and (2) are obtained as follows [7]:

$$\widehat{\boldsymbol{E}}\boldsymbol{U} + \varrho\widehat{\boldsymbol{C}}\boldsymbol{P}_f = \boldsymbol{R} \tag{7}$$

$$\varrho\widehat{\boldsymbol{C}}^T\dot{\boldsymbol{U}} - \widehat{\boldsymbol{H}}\boldsymbol{P}_f - a_f\boldsymbol{M}\dot{\boldsymbol{P}}_f = \boldsymbol{Q} \tag{8}$$

The smoothed-displacement matrix for an axisymmetric setting is as follows

$$\widehat{\boldsymbol{B}}_1 = \sum_{i\in S_s}\widehat{\boldsymbol{B}}_{1i} = \sum_{i\in S_s}\begin{bmatrix}\hat{b}_{ir} & 0\\ 0 & \hat{b}_{iz}\\ \hat{b}_{iz} & \hat{b}_{ir}\\ \frac{\phi_i}{r} & 0\end{bmatrix} \tag{9}$$

where

$$\hat{b}_{il} = \frac{1}{2\Omega^k}\sum_{\mathrm{m}=1}^{N_{seg}}\left[L_m^k\sum_{\mathrm{n}=1}^{N_{gau}} w_n\,\phi_i(\boldsymbol{x}_{mn})\;n_l(\boldsymbol{x}_{mn})\right]\quad (l=r,z) \tag{10}$$

where $N_{seg}$ is the number of line segments, $L_m^k$ is the length of the $m$th segment of $\Gamma^k$, $n_l$ is the component of the unit outward normal and $N_{gau}$ is the number of Gauss points used in each segment of $\Gamma^k$, which is taken two in this study. $\boldsymbol{x}_{mn}$ is the $n$th Gauss point of the $m$th segment of $\Gamma^k$, and $w_n$ is the Gauss integration weight. $\phi_i(\boldsymbol{x}_{mn})$ is the shape function value for node $i \in S_s$ at the point of interest $\boldsymbol{x}_{mn}$. If the current quadrature point $i \notin S_n$, then $\phi_i(\boldsymbol{x}_{mn}) = 0$.

As can be seen, the term $\phi_i/r$ in equation (9) leads to a singularity problem when the Gauss point of interest is on the axis of symmetry. To overcome this problem, the smoothed strain-displacement matrix is decomposed into two matrices:

$$\widehat{\boldsymbol{B}}_1 = \widehat{\boldsymbol{B}}_{1s} + \boldsymbol{B}_{1\Theta} \tag{11}$$

where

$$\widehat{\boldsymbol{B}}_{1s} = \sum_{i\in S_s} \widehat{\boldsymbol{B}}_{1si} = \sum_{i\in S_s} \begin{bmatrix} \hat{b}_{ir} & 0 \\ 0 & \hat{b}_{iz} \\ \hat{b}_{iz} & \hat{b}_{ir} \\ 0 & 0 \end{bmatrix} \tag{12}$$

$$\boldsymbol{B}_{1\theta} = \sum_{i\in S_s} \boldsymbol{B}_{1\theta i} = \sum_{i\in S_s} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{\phi_i}{r} & 0 \end{bmatrix} \tag{13}$$

The decomposition separates the terms causing singularity from the rest of the strain-displacement matrix. Given that no smoothing is required for $\phi_i/r$ terms, the integrations involving this term can be carried out over the smoothing domains rather than the boundary of the smoothing domains, resulting in the removal of the singularity problem. The procedure involves no additional computational cost as in any case, the shape function values are calculated at the Gauss points inside the smoothing domains for the calculation of $\boldsymbol{M}$ and $\widehat{\boldsymbol{C}}$, even in plane strain problems [7]. For the stiffness matrix,

$$\widehat{\boldsymbol{E}} = \sum_{k=1}^{N_{SD}} 2\pi r\left(\widehat{\boldsymbol{B}}_{1s}^T \boldsymbol{D} \widehat{\boldsymbol{B}}_{1s} + \widehat{\boldsymbol{B}}_{1s}^T \boldsymbol{D} \widehat{\boldsymbol{B}}_{1\theta} + \widehat{\boldsymbol{B}}_{1\theta}^T \boldsymbol{D} \widehat{\boldsymbol{B}}_{1s} + \widehat{\boldsymbol{B}}_{1\theta}^T \boldsymbol{D} \widehat{\boldsymbol{B}}_{1\theta}\right)\Omega^k \tag{14}$$

The first term in equation (14) is calculated similar to the computation of the stiffness matrix in CSPIM of plane strain problems [9]. The second and third terms are calculated adopting an approach similar to that proposed in [1] for the calculation of the $\widehat{\boldsymbol{C}}$ matrix. The last term is calculated over the smoothing domains using the standard Gauss integration approach.

The coupling matrix is calculated as follows

$$\widehat{\boldsymbol{C}} = \sum_{k=1}^{N_{\mathrm{SD}}} \widehat{\boldsymbol{C}}^{SD} = \sum_{k=1}^{N_{\mathrm{SD}}} 2\pi \left(\int_{\Omega^k} (\widehat{\boldsymbol{B}}_{2s} + \boldsymbol{B}_{2\Theta})^T \boldsymbol{\Phi}^p r \, d\Omega^k\right) \tag{15}$$

where $\boldsymbol{\Phi}^p$ is the matrix of shape functions. The first part of the integration is calculated similar to the $\widehat{\boldsymbol{C}}$ matrix in plain strain formulation [7]. The second term is obtained over the smoothing domain using the standard Gauss integration approach.

The Permeability matrix ($\widehat{\boldsymbol{H}}$) and the compressibility matrix of the fluid phase ($\boldsymbol{M}$) can be computed with small modifications to the formulations presented in [7], as follows

$$\widehat{\boldsymbol{H}} = \sum_{k=1}^{N_{\mathrm{SD}}} \widehat{\boldsymbol{H}}^{SD} = \sum_{k=1}^{N_{\mathrm{SD}}} \frac{2\pi}{\mu_f} \widehat{\boldsymbol{B}}_3^T \boldsymbol{k}_f \widehat{\boldsymbol{B}}_3 \int_{\Omega^k} r \, d\Omega^k \tag{16}$$

$$\boldsymbol{M} = \sum_{k=1}^{N_{\mathrm{SD}}} \boldsymbol{M}^{SD} = \sum_{k=1}^{N_{\mathrm{SD}}} 2\pi \int_{\Omega^k} \boldsymbol{\Phi}^{pT} \boldsymbol{\Phi}^p r d\Omega^k \tag{17}$$

Equations (7) and (8) are then discretised in time using a three point time discretisation scheme [15] resulting in the fully discretised systems of equations, as follows

$$\boldsymbol{K}\boldsymbol{W} = \boldsymbol{Y} \tag{18}$$

$$\boldsymbol{K} = \begin{bmatrix} A\widehat{\boldsymbol{E}}_{2N\times 2N} & A\varrho\widehat{\boldsymbol{C}}_{2N\times N} \\ A\varrho\widehat{\boldsymbol{C}}^{\boldsymbol{T}}{}_{N\times 2N} & -\left(\Delta t\widehat{\boldsymbol{H}} + Aa_f\boldsymbol{M}\right)_{N\times N} \end{bmatrix}_{3N\times 3N} \tag{19}$$

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{U}^{t+\alpha\Delta t}{}_{2N\times 1} \\ \boldsymbol{P}_f^{t+\alpha\Delta t}{}_{N\times 1} \end{bmatrix}_{3N\times 1} \tag{20}$$

$$\boldsymbol{Y} = \begin{bmatrix} A(\boldsymbol{R}^{t+\alpha\Delta t})_{2N\times 1} \\ (\Delta t\boldsymbol{Q}^{t+\alpha\Delta t} + B\varrho\widehat{\boldsymbol{C}}^T\boldsymbol{U}^t - C\varrho\widehat{\boldsymbol{C}}^T\boldsymbol{U}^{t-\Delta t} - Ba_f\boldsymbol{M}\boldsymbol{P}_f^t + Ca_f\boldsymbol{M}\boldsymbol{P}_f^{t-\Delta t})_{N\times 1} \end{bmatrix}_{3N\times 1} \tag{21}$$

## Verification

The benchmark consolidation problem of Cryer is adopted here for verification purposes. In this problem, a uniformly distributed surface load of $p_0 = 1\ kPa$ is concentrically applied to a sphere of radius $a = 1.0\ m$. Only a quarter of the sphere was modelled, because of the symmetry. The outer boundary of the sphere was assumed permeable and the permeability of the sphere was assumed $k = 1\ m/s$. The geometry and boundary conditions of the problem, along with the triangular background mesh used for the spatial discretisation are shown in Figure 1. Three different cases were studied using three different sets of elastic properties as presented in Table 1. Case 3 here approximates an incompressible material ($\nu$ = 0.5). In all cases, the first dimensionless time step was taken as 0.0005 with a time step growth factor of 1.2.

Table 1. Soil properties used in the analysis of Cryer's problem

| *Case Number* | *ν* | *E* |
|---|---|---|
| 1 | 0.0 | 1.0 |
| 2 | 0.333 | 0.666 |
| 3 | 0.496 | 0.0299 |

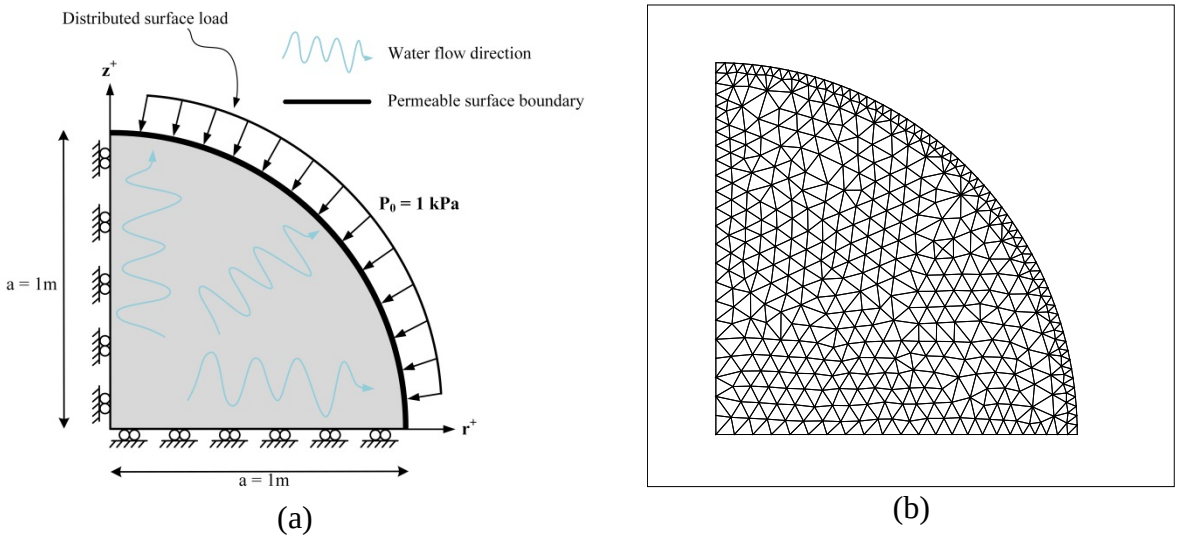



Figure 1. consolidation problem of Cryer; a) Geometry and boundary conditions; b) Spatial discretisation used in this study

As an example, the numerical results for the excess pore water pressure contours from the analysis of case 1 at dimensionless time 0.01979 is shown in Figure 2. The symmetry of the geometry and the applied load result in the symmetry of the excess pore water pressure contour, as expected.

The analytical solution for the excess pore water pressure generated at the centre of the sphere versus dimensionless radial displacement of the surface of the sphere, and also the variation of the dimensionless surface settlement with time were derived by Cryer [8]. For the verification of the proposed formulation, the results of the numerical analysis are compared with the analytical solution as presented in Figure 3. As can be seen, there is an excellent agreement between the results of the proposed method and the analytical solution for all three cases studied. In all cases, the pore pressure increases over the entire sphere to the value of the applied load upon the application of the load. The pore pressure then initially increases at the centre of the sphere in cases 1 and 2, before decreasing due to consolidation process. This effect, known as the Mandel-Cryer effect [8, 16], is perfectly captured in the numerical analyses. For case 3, when the soil is incompressible, no increase of pore water pressure is observed which also agrees with the analytical solution. The Mandle-Cryer effect cannot be captured using the original Terzaghi's formulation for consolidation and can only be simulated when the governing equations for the solid and water phases are properly coupled. More discussion on this matter can be found in [17].

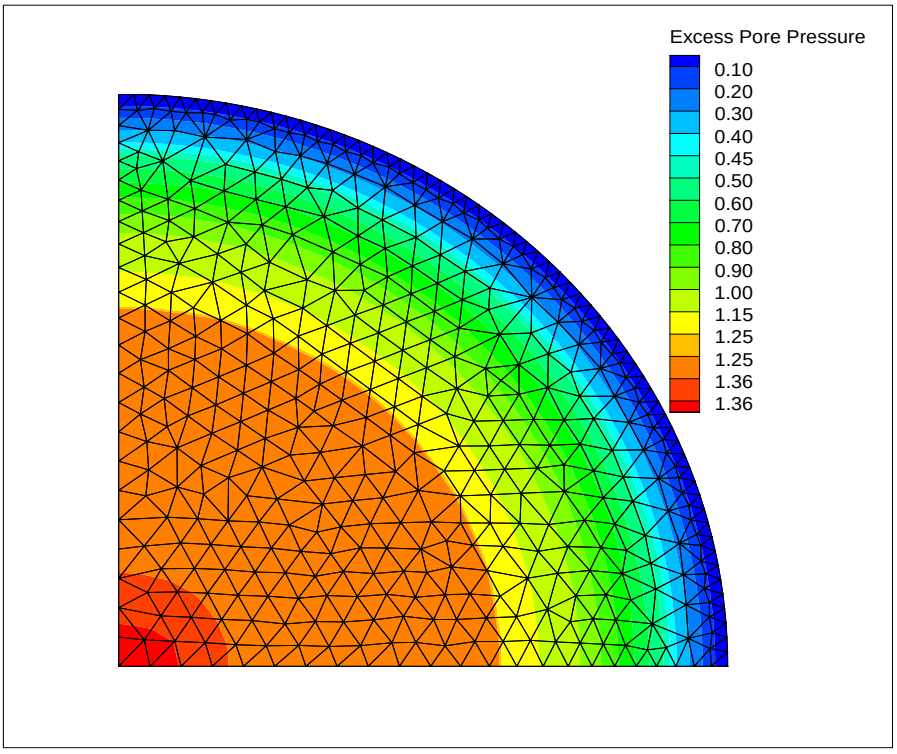



Figure 2. Excess pore pressure contour for Cryer's problem at dimensionless time of 0.01979.

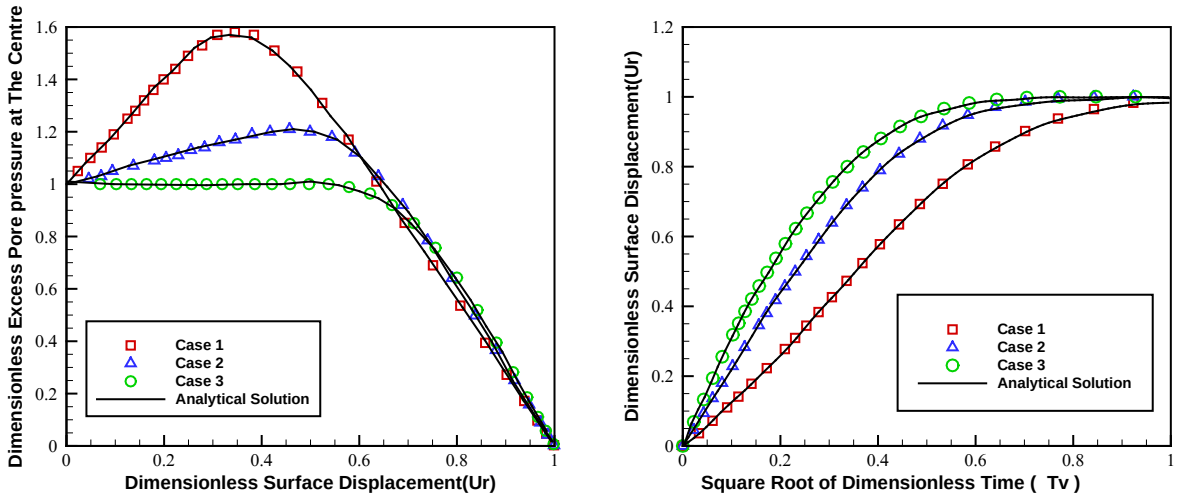



Figure 3. Comparison between numerical results and analytical solution; (a) Dimensionless excess pore pressure at the centre of the sphere versus dimensionless surface displacement; (b) dimensionless surface displacement versus $\sqrt{T_v}$

## Conclusion

A cell-based smoothed point interpolation method (CSPIM) for the analyses of axisymmetric poro-elastic problems has been proposed. This method is based on a novel approach for systematic decomposition of the system matrices to remove the inherent singularity problem in SPIMs when an axisymmetric problem is considered. Solid and fluid phase property matrices and coupling matrices of a fully discretised system of equations have been derived in an axisymmetric setting based on the proposed approach. The method has been verified by numerical analysis of Cryer's benchmark problem of consolidation of a saturated sphere. It has been shown that the proposed method can accurately reproduce the analytical solution and in particular, the Mandel-Cryer effect is precisely captured when a compressible medium is considered.

## References

1. Liu, G., G. Zhang, K. Dai, Y. Wang, Z. Zhong, G. Li, and X. Han, *A linearly conforming point interpolation method (LC-PIM) for 2D solid mechanics problems.* International Journal of Computational Methods, 2005. **2**(04): p. 645-665.
2. Liu, G.-R. and G.-Y. Zhang, *Smoothed point interpolation methods: G space theory and weakened weakforms*. 2013.
3. Cui, X., G. Liu, G. Li, X. Zhao, T. Nguyen-Thoi, and G. Sun, *A smoothed finite element method (SFEM) for linear and geometrically nonlinear analysis of plates and shells.* Comput Model Eng Sci, 2008. **28**(2): p. 109-125.
4. Liu, G. and G. Zhang, *Edge-based smoothed point interpolation methods.* International Journal of Computational Methods, 2008. **5**(04): p. 621-646.
5. Liu, G. and G. Zhang, *A normed G space and weakened weak (W2) formulation of a cell-based smoothed point interpolation method.* International Journal of Computational Methods, 2009. **6**(01): p. 147-179.

6. Liu, G. and G. Zhang, *Upper bound solution to elasticity problems: A unique property of the linearly conforming point interpolation method (LC-PIM).* International Journal for Numerical Methods in Engineering, 2008. **74**(7): p. 1128-1161.
7. Tootoonchi, A., A. Khoshghalb, G. Liu, and N. Khalili, *A cell-based smoothed point interpolation method for flow-deformation analysis of saturated porous media.* Computers and Geotechnics, 2016. **75**: p. 159-173.
8. Cryer, C., *A comparison of the three-dimensional consolidation theories of Biot and Terzaghi.* The Quarterly Journal of Mechanics and Applied Mathematics, 1963. **16**(4): p. 401-412.
9. Biot, M.A., *General theory of three-dimensional consolidation.* Journal of applied physics, 1941. **12**(2): p. 155-164.
10. Wang, J. and G. Liu, *A point interpolation meshless method based on radial basis functions.* International Journal for Numerical Methods in Engineering, 2002. **54**(11): p. 1623-1648.
11. Liu, G.-R., *Meshfree methods: moving beyond the finite element method*. 2010: CRC press.
12. Liu, G.-R. and Y. Gu, *A point interpolation method for two-dimensional solids.* International Journal for Numerical Methods in Engineering, 2001. **50**(4): p. 937-951.
13. Wang, J. and G. Liu, *On the optimal shape parameters of radial basis functions used for 2-D meshless methods.* Computer methods in applied mechanics and engineering, 2002. **191**(23): p. 2611-2630.
14. Khoshghalb, A. and N. Khalili, *A stable meshfree method for fully coupled flow-deformation analysis of saturated porous media.* Computers and Geotechnics, 2010. **37**(6): p. 789-795.
15. Khoshghalb, A., N. Khalili, and A. Selvadurai, *A three-point time discretization technique for parabolic partial differential equations.* International Journal for Numerical and Analytical Methods in Geomechanics, 2011. **35**(3): p. 406-418.
16. Mandel, J., *Consolidation des sols (étude mathématique)*.* Geotechnique, 1953. **3**(7): p. 287-299.
17. Gibson, R., R. Schiffman, and S. Pu, *Plane strain and axially symmetric consolidation of a clay layer on a smooth impervious base.* The Quarterly Journal of Mechanics and Applied Mathematics, 1970. **23**(4): p. 505-520.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EFFICIENT SHAPE FUNCTION COSTRUCTION OF EFG MESHLESS METHOD

**P. Metsis**[1], **and M. Papadrakakis**[1]

[1] *Institute of Structural Analysis and Antiseismic Research*
*National Technical University of Athens*
*Zografou Campus, Athens 15780*
*(pmetsis@gmail.com, mpapadra@central.ntua.gr)*

**Keywords:** Meshless methods; Element free Galerkin (EFG); shape function construction; Adaptivity

**Abstract.** *Meshless methods have shown increased accuracy and better convergence rates compared to other well-known simulation methods in a variety of computational mechanics problems. In adaptivity analysis, besides the effortless node addition at each step, their computational cost is relatively higher, since the construction of updated shape functions and shape functions derivatives that are used to construct the stiffness matrix is burdensome.*

*EFG and other similar methods use the moving least squares procedure, where the approximation of the displacement field is expressed as a polynomial with non-constant coefficients. In order to determine these coefficients, a weighted residual is solved for every point of interest, usually the Gauss points. Each time the domain of influence changes or a new node is added or subtracted from the domain, the shape functions and their derivatives have to be re-calculated for the influenced areas, requiring a significant amount of computational effort.*

*In this work, the shape function construction procedure is analyzed in depth and several new approaches are proposed ranging from an explicit analytical form for typical arrangement of nodes -which corresponds to a significant percentage of typical problem domains- to complex hierarchical formulations in the context of an h-type refinement scheme. The addition of new nodes and subsequently the re-calculation of the influenced moment matrices, that are necessary for obtaining the shape functions and their derivatives and subsequently for the construction of the stiffness matrix, are properly addressed. Unconventional types of weight and shape functions are also proposed and the results are critically assessed.*

## 1 INTRODUCTION

Meshless methods (MMs) have been used for solving problems that are difficult or impossible to be dealt with by mesh related procedures of established simulation methods like the finite element method (FEM). Problems of large deformations, discontinuities, contact and moving boundaries or problems demanding very high accuracy, are some of the problems MMs are mostly used. In MMs there is no need to construct a mesh while the solution and its derivatives are continuous throughout the domain of interest leading to more accurate stress distributions as well as to less sensitive to distortions due to large deformations. In areas where increased accuracy is needed, refinement by adding nodes (h-adaptivity) may implemented effortlessly. Moreover, established methods like FEM may be partially enriched with features of some MMs, or be coupled with, enhancing their effectiveness.

One of the first and most prominent meshless methods is the element free Galerkin (EFG) method introduced by Belytschko et al.. EFG requires only nodal data and no element connectivity is needed to construct the shape functions. However, a global background cell structure is necessary for the numerical integration. Moreover, since the number of interactions between nodes and/or integration points is heavily increased, due to large domains of influence, the resulting matrices are more densely populated and the computational cost for the formulation and solution of the problem is much higher than in the conventional FEM.

The calculation of the shape functions and shape functions derivatives that are used to construct the stiffness matrix add significantly to the abovementioned computational cost. EFG uses the moving least squares procedure, where the approximation of the displacement field is expressed as a polynomial with non-constant coefficients. In order to determine these coefficients, a weighted residual is solved for every point of interest, usually the Gauss points. Each time the domain of influence changes or a new node is added or subtracted from the domain, the shape functions and their derivatives have to be re-calculated for the influenced areas. Thus, in the context of an adaptive scheme this procedure can be burdensome.

## 2 BASIC INGREDIENTS OF THE MESHLESS EFG METHOD

The approximation of a scalar function $u$ in the meshless EFG method can be written as

$$u(\mathrm{x},t)=\sum_{i\in S}\Phi_i(\mathrm{x})u_i(t) \tag{1}$$

where $\Phi_i$ are the shape functions, $u_i$ are the values of the scalar function $u$ at node $i$ located at position $\mathrm{x}_i$, and $S$ is the set of nodes $i$ for which $\Phi_i(\mathrm{x})$ obtains non zero values. The shape functions in eq.(1) are only approximants and not interpolants, since generally $u_i \neq u(\mathrm{x}_i)$.

The shape functions $\Phi_i$ are obtained from the weight coefficients $w_i$, which are functions of a distance parameter $r=\|\mathrm{x}_i-\mathrm{x}\|/d_i$ where $d_i$ defines the domain of influence (DoI) of each node $i$. The size primarily, but also the shape of the domain of influence, is crucial to solution accuracy, stability and computational cost, as it co-defines the degree of continuity between the nodes and the bandwidth of the resulting system matrices.

If $\mathrm{p}(\mathrm{x})$ is a complete polynomial of length $m$ and $\mathrm{a}(\bar{\mathrm{x}})$ contains non-constant coefficients that depend on x:

$$\mathrm{a}(\bar{\mathrm{x}}) = [a_0(\mathrm{x}) \quad a_1(\mathrm{x}) \quad a_2(\mathrm{x}) \quad \dots \quad a_m(\mathrm{x})]^T \tag{2}$$

then the approximation $u^h$ is expressed as a polynomial of length $m$ with non-constant coefficients. The local approximation around a point $\bar{\mathrm{x}}$, evaluated at a point x is given by

$$u_L^h(\mathrm{x}, \bar{\mathrm{x}}) = \mathrm{p}^T(\mathrm{x})\mathrm{a}(\bar{\mathrm{x}}) \tag{3}$$

In two dimensional problems, the linear basis $\mathrm{p}(\mathrm{x})$ is given by

$$\mathrm{p}^T(\mathrm{x}) = [1 \quad x \quad y], \qquad m = 3 \tag{4}$$

and the quadratic basis by

$$\mathrm{p}^T(\mathrm{x}) = [1 \quad x \quad y \quad x^2 \quad y^2 \quad xy], \qquad m = 6 \tag{5}$$

The minimization of a functional $J(\mathrm{x})$ defined by a weighted average over all nodes $i \in 1, \dots, n$:

$$J(\mathrm{x}) = \sum_{i=1}^{n} w(\mathrm{x} - \mathrm{x}_i)\left[u_L^h(\mathrm{x}_i, \mathrm{x}) - u_i\right]^2 = \sum_{i=1}^{n} w(\mathrm{x} - \mathrm{x}_i)\,[\mathrm{p}^T(\mathrm{x}_i)\mathrm{a}(\mathrm{x}) - u_i]^2 \tag{6}$$

determines the parameters $a_j(\mathrm{x})$ at any point x. In eq.(6) the parameters $u_i$ are specified by the difference between the local approximation $u_L^h(\mathrm{x}, \bar{\mathrm{x}})$ and the nodal value $u_i$, while the weight function satisfies the condition $w(x - x_i) \neq 0$.

An extremum of $J(\mathrm{x})$ with respect to the coefficients $a_j(\mathrm{x})$ can be obtained by setting the derivative of $J$ with respect to $\mathrm{a}(\mathrm{x})$ equal to zero. This condition gives the following relation

$$\mathrm{A}(\mathrm{x})\mathrm{a}(\mathrm{x}) = \mathrm{W}(\mathrm{x})\mathrm{u} \tag{7}$$

where the moment matrix $\mathrm{A}(\mathrm{x})$ is defined as

$$\mathrm{A}(\mathrm{x}) = \sum_{i=1}^{n} w(\mathrm{x} - \mathrm{x}_i)\mathrm{p}(\mathrm{x}_i)\,\mathrm{p}^T(\mathrm{x}_i) \tag{8}$$

and

$$\mathrm{W}(\mathrm{x}) = [w(\mathrm{x} - \mathrm{x}_1)\mathrm{p}(\mathrm{x}_1) \quad w(\mathrm{x} - \mathrm{x}_2)\mathrm{p}(\mathrm{x}_2) \quad \dots \quad w(\mathrm{x} - \mathrm{x}_n)\mathrm{p}(\mathrm{x}_n)] \tag{9}$$

The approximants $u^h$ can be defined by solving for $\mathrm{a}(\mathrm{x})$ in eq.(7) and substituting into eq.(3):

$$u^h(\mathrm{x}) = \mathrm{p}^T(\mathrm{x})[[\mathrm{A}(\mathrm{x})]^{-1}\mathrm{W}(\mathrm{x})\mathrm{u}] \tag{10}$$

which together with eq.(1) leads to the derivation of the shape function $\Phi_i$ associated with node $i$ at point x:

$$\Phi_i(\mathrm{x}) = \mathrm{p}^T(\mathrm{x})[\mathrm{A}(\mathrm{x})]^{-1}\mathrm{W}(\mathrm{x}_i) \tag{11}$$

A solution of a local problem $\mathrm{A}(\mathrm{x})\mathrm{z} = \mathrm{p}(\mathrm{x})$ of size $m \times m$ is performed whenever the shape functions are to be evaluated. This constitutes a drawback of moving least squares (MLS)-based MMs since the computational cost can be substantial, while it is possible for the moment matrix $\mathrm{A}(\mathrm{x})$ to be ill conditioned. In adaptive procedures, where nodes are gradually added and the shape functions of enriched areas are recalculated at each adaptive step, the required computational effort can be considerably increased.

The Galerkin weak form of the above formulation gives the discrete algebraic equation

$$\mathrm{Ku} = \mathrm{f} \tag{12}$$

with stiffness matrix components $\mathrm{K}_{ij}$

$$\mathrm{K}_{ij} = \int_{\Omega} \mathrm{B}_i^T \mathrm{E}\mathrm{B}_j d\Omega \tag{13}$$

and force vector $\mathrm{f}_i$

$$\mathrm{f}_i = \int_{\Gamma_\mathrm{t}} \Phi_i \bar{\mathrm{t}} d\Gamma + \int_{\Omega} \Phi_i \mathrm{b} d\Omega \tag{14}$$

In 2D problems, matrix B is given by

$$\mathrm{B}_i = \begin{bmatrix} \Phi_{i,x} & 0 \\ 0 & \Phi_{i,y} \\ \Phi_{i,y} & \Phi_{i,x} \end{bmatrix} \tag{15}$$

and subsequently in 3D problems by

$$\mathrm{B}_i = \begin{bmatrix} \Phi_{j,x} & 0 & 0 \\ 0 & \Phi_{j,y} & 0 \\ 0 & 0 & \Phi_{j,z} \\ \Phi_{j,y} & \Phi_{j,x} & 0 \\ 0 & \Phi_{j,z} & \Phi_{j,y} \\ \Phi_{j,z} & 0 & \Phi_{j,x} \end{bmatrix} \tag{16}$$

Due to the lack of the Kronecker delta property of shape functions, the essential boundary conditions cannot be imposed the same way as in FEM. Several techniques are available such as Lagrange multipliers, penalty method and EFG - FEM coupling.

For the integration of eq.(13), virtual background cells are considered by dividing the problem domain into integration cells over which a Gaussian quadrature is performed:

$$\int_{\Omega} \mathrm{f}(\mathrm{x}) d\Omega = \sum_{J} \mathrm{f}(\xi_J) \omega_{\Xi} det J^{\xi}(\xi) \tag{17}$$

where $\xi$ are the local coordinates and $detJ^{\xi}(\xi)$ is the determinant of the Jacobian.

## 3 FORMULATION OF THE SHAPE FUNCTIONS BY PARTS

The shape functions of eq.(11) are calculated after computing the moment matrix A(x) from eq.(8). Suppose that $A(x)^{initial}$ is the moment matrix of the initial node distribution of a particular domain of influence (DoI) and some new nodes are added to that DoI. The enriched moment matrix $A(x)^{enriched}$ is given by:

$$A(x)^{enriched} = \sum_{k=1}^{n_enriched} w(\mathrm{x}-\mathrm{x}_k)\mathrm{p}(\mathrm{x}_k)\,\mathrm{p}^T(\mathrm{x}_k) \tag{18}$$

where the weight function $w(\mathrm{x}-\mathrm{x}_k)$ and the polynomial $\mathrm{p}(\mathrm{x}_k)$ depend only on the position of the node under consideration. Thus for cubic spline weight function:

$$w(\mathrm{x}-\mathrm{x}_i) = \begin{cases} \frac{2}{3} - 4(x-x_i)^2 + (x-x_i)^3, for\ (x-x_i) \le \frac{1}{2} \\ \frac{4}{3} - 4(x-x_i) + 4(x-x_i)^2 - \frac{4}{3}(x-x_i)^3, for\ \frac{1}{2} < (x-x_i) \le 1 \\ 0, for\ (x-x_i) > 1 \end{cases} \tag{19}$$

$A(x)^{enriched}$ consists of two parts that can be computed independently to each other.

$$A(x)^{enriched} = \sum_{i=1}^{n_initial} w(\mathrm{x}-\mathrm{x}_i)\mathrm{p}(\mathrm{x}_i)\,\mathrm{p}^T(\mathrm{x}_i) \quad + \sum_{j=1}^{n_additional} w(\mathrm{x}-\mathrm{x}_j)\mathrm{p}(\mathrm{x}_j)\,\mathrm{p}^T(\mathrm{x}_j) \tag{20}$$

$$A(x)^{enriched} = A(x)^{\text{initial}} + A(x)^{additional}$$

where $n_enriched = n_initial + n_additional$.

The enriched shape function $\Phi_i$ associated with node $i$ at point x is calculated from eq.(11) as follows:

$$\begin{aligned} \Phi_i(\mathrm{x})^{enriched} &= \mathrm{p}^T(\mathrm{x})\left[A(x)^{enriched}\right]^{-1}\mathrm{W}(\mathrm{x}_i) \\ &= \mathrm{p}^T(\mathrm{x})\left[A(x)^{\text{initial}} + A(x)^{additional}\right]^{-1}\mathrm{W}(\mathrm{x}_i) \end{aligned} \tag{21}$$

where the issue of the inverse of the sum of two matrices arises. Following the Binomial inverse theorem:

If $a$ and $a+b$ are invertible, then

$$(a+b)^{-1} = a^{-1} + X \tag{22}$$

with

$$X = -(I + a^{-1}b)^{-1}\,a^{-1}\,b\,a^{-1}$$

the inverse matrix of eq.(21) can be easily calculated since $(A(x)^{\text{initial}})^{-1}$ is already computed from the initial step. Thus,

$$\begin{aligned}\Phi_i(\mathrm{x})^{enriched} &= \mathrm{p}^T(\mathrm{x})\left[A(x)^{\text{initial}} + A(x)^{additional}\right]^{-1}\mathrm{W}(\mathrm{x}_i) \\ &= \mathrm{p}^T(\mathrm{x})\left[A(x)^{\text{initial}}\right]^{-1}\mathrm{W}(\mathrm{x}_i) \;+\; \mathrm{p}^T(\mathrm{x})\,[\mathrm{X}]\,\mathrm{W}(\mathrm{x}_i)\end{aligned} \tag{23}$$

where

$$X = -\left(I + \{A(x)^{\text{initial}}\}^{-1}A(x)^{additional}\right)^{-1}\{A(x)^{\text{initial}}\}^{-1}A(x)^{additional}\{A(x)^{\text{initial}}\}^{-1} \tag{24}$$

Then, since

$$\Phi_i(\mathrm{x})^{initial} = p^T(x)\left[A(x)^{initial}\right]^{-1}W(x_i) \tag{25}$$

and

$$\Phi_i(\mathrm{x})^{additional} = \mathrm{p}^T(\mathrm{x})\,[\mathbf{X}]\,\mathrm{W}(\mathrm{x}_i) \tag{26}$$

the enriched shape function is given by:

$$\Phi_i(\mathrm{x})^{enriched} = \Phi_i(\mathrm{x})^{initial} + \Phi_i(\mathrm{x})^{additional} \tag{27}$$

The above derivation expresses the enriched shape functions as the sum of the initial shape functions and the contribution of the newly added nodes.

The proposed formulation is demonstrated in the following one-dimensional problem. Consider an interval $0 \le x \le 4$ divided into four unequal parts by five nodes, as shown in Fig. 1.

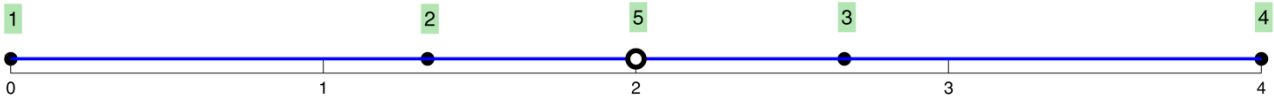



**Fig. 1: Node distribution 1D example. ● initial nodes, ○ added node**

The typical expression of the shape functions for each node is given by eq.(11) which is evaluated at each Gauss point. The linear basic function $\mathbf{p}^T(x) = \begin{bmatrix}1 & x\end{bmatrix}$ is used. For a certain Gauss point, eq.(11) can be written as:

$$\underbrace{\boldsymbol{\Phi}(\mathrm{x}_{Gp})}_{1\mathrm{xm}} = \underbrace{p^T(x_{Gp})}_{1\mathrm{x2}}\underbrace{\left[A_{1:5}(\mathrm{x}_{Gp})\right]^{-1}}_{2\mathrm{x2}}\underbrace{W(\mathrm{x}_{Gp})}_{2\mathrm{xm}} \tag{28}$$

where

$$\mathrm{p}^T(\mathrm{x}_{Gp}) = [1 \quad \mathrm{x}_{Gp}] \tag{29}$$

Eq.(8) for the initial nodes m and the additional node n is written:

$$\boldsymbol{A}_{1:(m+n)}(\mathrm{x}_{Gp}) = \sum_{i=1}^{(m+n)} w(\mathrm{x}_{Gp} - \mathrm{x}_i)\mathrm{p}(\mathrm{x}_i)\,\mathrm{p}^T(\mathrm{x}_i) = \sum_{i=1}^{(m+n)} w(\mathrm{x}_{Gp} - \mathrm{x}_i)\begin{bmatrix}1\\ \mathrm{x}_i\end{bmatrix}[1 \quad \mathrm{x}_i] \tag{30}$$

while eq.(23) becomes:

$$\begin{aligned}\Phi_{1:(m+n)}(\mathrm{x}_{Gp})^{enriched} &= \Phi_{1:m}(\mathrm{x}_{Gp}) + \Phi_n(\mathrm{x}_{Gp}) = \\ &= p^T(\mathrm{x}_{Gp})\left[A_{1:(m+n)}(\mathrm{x}_{Gp})\right]^{-1} W_{1:m}(\mathrm{x}) + \mathrm{p}^T(\mathrm{x}_{Gp})\left[A_{1:(m+n)}(\mathrm{x}_{Gp})\right]^{-1} W_n(\mathrm{x})\end{aligned} \tag{31}$$

For the example of Fig. 1, the first term includes the value of the shape functions for the first four nodes (m=4), at a certain Gauss point, and the second term the value of the shape function for the fifth node (n=1).

The weighted moment matrix **A** can also be separated into two parts as follows:

$$\begin{aligned}A_{1:(m+n)}(\mathrm{x}_{Gp}) &= A_{1:m}(\mathrm{x}_{Gp}) + A_n(\mathrm{x}_{Gp}) = \\ &= \sum_{i=1}^{m} w(x_{Gp} - x_i)p(x_i)\,p^T(x_i) + w(x_{Gp} - x_n)p(x_n)p^T(x_n)\end{aligned} \tag{32}$$

Using the *Binomial inverse theorem,* we obtain

$$\left[A_{1:(m+n)}(\mathrm{x}_{Gp})\right]^{-1} = \left[A_{1:m}(\mathrm{x}_{Gp}) + A_n(\mathrm{x}_{Gp})\right]^{-1} = \left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} + X \tag{33}$$

where $$X = -\left\{I + \left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} A_n(\mathrm{x}_{Gp})\right\}^{-1}\left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} A_n(\mathrm{x}_{Gp})\left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} \tag{34}$$

Since $\left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1}$ is already calculated in the previous (initial) step, the above equation is reduced computationally to one inversion and multiplication of matrices. Finally by combining the eq.(31) and eq.(33) the values for the shape functions of the five nodes, at a certain Gauss point can be written in parts as follows:

$$\Phi_{1:m}(\mathrm{x}_{Gp}) = \underbrace{p^T(\mathrm{x}_{Gp})\left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} W_{1:m}(\mathrm{x})}_{Initial\ \Phi_{1:m}} + \underbrace{p^T(x_{Gp})[X]W_{1:(m+n)}(x)}_{influence\ of\ nodes\ n} \tag{35}$$

and

$$\Phi_n(\mathrm{x}_{Gp}) = p^T(\mathrm{x}_{Gp})\left[A_{1:m}(\mathrm{x}_{Gp})\right]^{-1} W_n(\mathrm{x}) + p^T(x_{Gp})[X]W_n(x) \tag{36}$$

The first term of eq.(35) describes the values for the shape functions of the initial four nodes which divide the domain into three sections. The second part of the eq.(35) modifies the initial part, when the fifth node is added into the problem domain in order to add the contribution of the added node and retain the partition of unity.

Fig. 2 and Fig. 3 show the shape functions of the first four nodes before and after the enrichment with the fifth node, for a cubic spline weight function. Fig. 4 illustrates the shape function of the fifth node. The factor $d_m$ of the support domain is chosen to be 2.2.

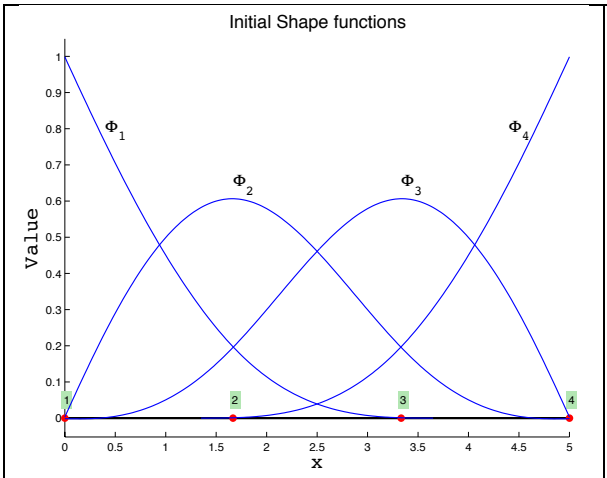



**Fig. 2 - Initial nodes in the problem domain. Shape functions of each node.**

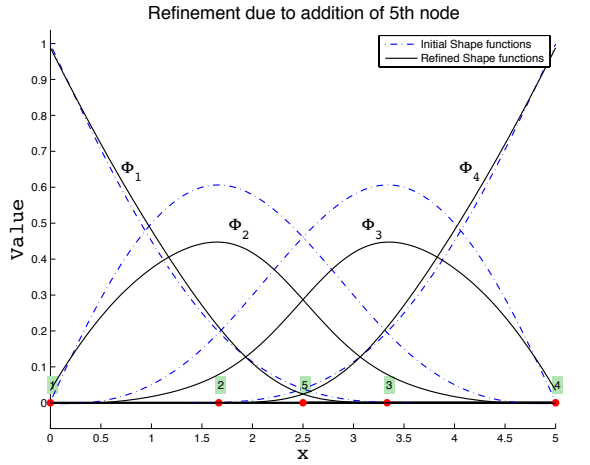



**Fig. 3 - Refinement from four to five nodes. Shape functions of the first four nodes.**

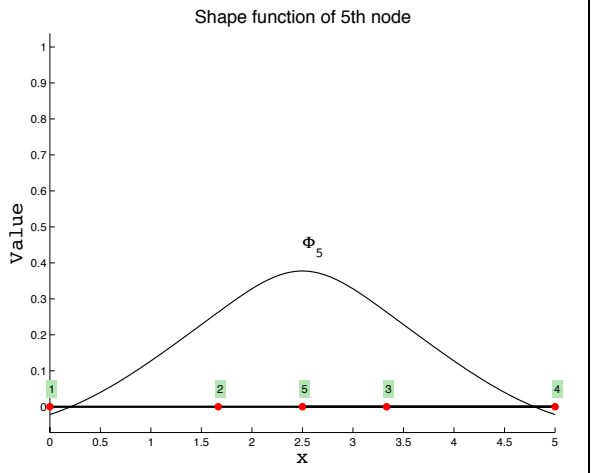



**Fig. 4 - Shape function of the fifth node.**

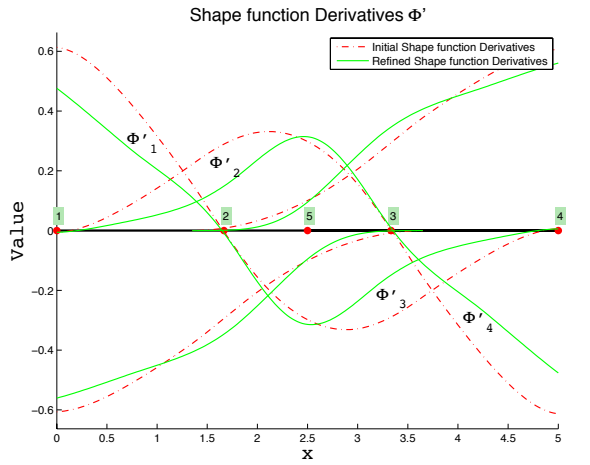



**Fig. 5 - Refinement from four to five nodes. First derivatives of the shape functions of the first four nodes.**

These figures illustrate the degree of modification of the existing shape functions at a specific domain when an extra node is added, since the shape function of the added node affects the values of all nodes of that DoI. The above process is directly extendable to more than one nodes, and in two or three dimensions.

The shape function derivatives which are necessary for the construction of the stiffness matrix (eqs.(13) and (15)) are as follows:

$$\Phi'_{1:(m+n)}(\mathrm{x}_{Gp})^{enriched} = [1 \quad 0][A_{1:(m+n)}(\mathrm{x}_{Gp})]^{-1} W_{1:(m+n)}(\mathrm{x})$$
$$-p^T(\mathrm{x}_{Gp})[A_{1:(m+n)}(\mathrm{x}_{Gp})]^{-1}[A'_{1:(m+n)}(\mathrm{x}_{Gp})][A_{1:(m+n)}(\mathrm{x}_{Gp})]^{-1} W_{1:(m+n)}(\mathrm{x}) \tag{37}$$
$$+p^T(\mathrm{x}_{Gp})[A_{1:(m+n)}(\mathrm{x}_{Gp})]^{-1} W'_{1:(m+n)}(\mathrm{x})$$

By employing the updated moment matrix of eq.(20) the derivatives can be calculated by parts. In order to illustrate the effect of the proposed refinement implementation to the shape function derivatives we consider again the problem shown in Fig. 1. The first part of the derivatives corresponds to the first four nodes and describes the values for the first derivative of the shape functions when the fifth node is not added in the problem domain. The second term modifies the initial part when the fifth node is added into the problem domain.

Fig. 5 illustrates the first derivative of the shape functions of the first four nodes, before and after the addition of the fifth node in the problem domain.

## 4 EXPLICIT ANALYTICAL FORM FOR TYPICAL ARRANGEMENT OF NODES

The case of uniform arrangement of nodes, though restricting, is used in a number of meshless analyses, especially in the main areas of even geometrically irregular bodies. For this case an analytical form of the derived shape functions could be produced, diminishing the computational effort of the shape function construction stage.

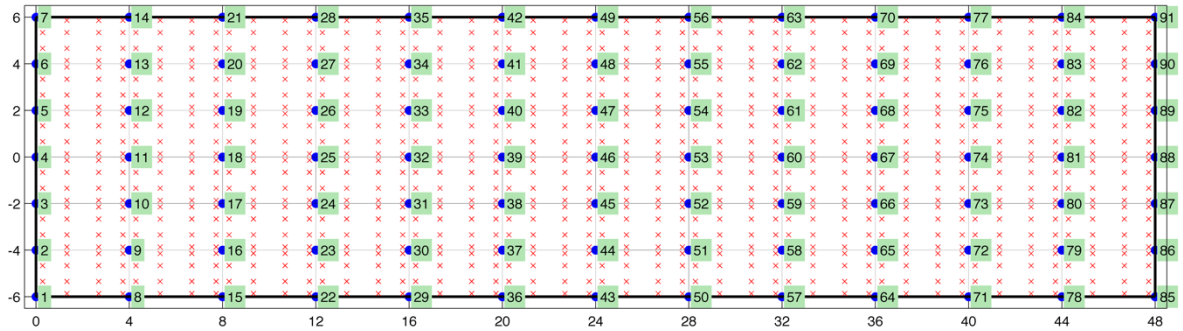



**Fig. 6 – Typical Cantilever with uniform arrangement of nodes**

In such cases, the moment matrix $\mathrm{A}(\mathrm{x}) = \sum_{i=1}^{n} w(\mathrm{x}-\mathrm{x}_i)\mathrm{p}(\mathrm{x}_i)\,\mathrm{p}^T(\mathrm{x}_i)$ depends only on the distance between the nodes of each domain of influence, therefore it could be written with respect to the unknown x. In cases where the above-mentioned moment matrix is of adequate condition, which is a prerequisite for the calculation of the shape functions in general, an analytical (not numerical) inversion follows which leads to the subsequent analytical formulation of the shape

functions $\Phi_i(x) = p^T(x)[A(x)]^{-1}W(x_i)$ at every point x. The same apply for the calculation of the shape functions derivatives.

It should be noted that the analytical form of the referred shape functions is quite complex and lengthy, depending on the arrangement of nodes. For example, for the cantilever of Fig. 6 one of the resulting shape functions for a point x is:

$\boldsymbol{\Phi_1(x)}=$ -(2*(abs(L) - abs(x))^3*(24*x*abs(L - x)^3 - 6*L*abs(L/2 - x)^3 + 12*x*abs(L/2 - x)^3 - 5*L*abs(L)^3 + 6*x*abs(L)^3 - 24*L*abs(L - x)^3 - 12*x*abs(L/2 - x)^2*abs(L) + 24*L*abs(L)*abs(L - x)^2 - 24*x*abs(L)*abs(L - x)^2 + 6*L*abs(L/2 - x)^2*abs(L)))/(L*(11*abs(L)^6 - 18*abs(L/2 - x)^2*abs(L)^4 + 18*abs(L/2 - x)^3*abs(L)^3 - 12*abs(L/2 - x)^3*abs(x)^3 - 30*abs(L)^5*abs(x) + 36*abs(L/2 - x)^3*abs(L - x)^3 - 10*abs(L)^3*abs(x)^3 + 30*abs(L)^4*abs(x)^2 + 54*abs(L)^3*abs(L - x)^3 - 54*abs(L)^4*abs(L - x)^2 - 48*abs(x)^3*abs(L - x)^3 - 36*abs(L/2 - x)^2*abs(L)^2*abs(x)^2 + 36*abs(L/2 - x)^2*abs(L)^2*abs(L - x)^2 + 144*abs(L)*abs(x)^2*abs(L - x)^3 + 48*abs(L)*abs(x)^3*abs(L - x)^2 - 144*abs(L)^2*abs(x)*abs(L - x)^3 + 144*abs(L)^3*abs(x)*abs(L - x)^2 + 12*abs(L/2 - x)^2*abs(L)*abs(x)^3 + 36*abs(L/2 - x)^2*abs(L)^3*abs(x) + 36*abs(L/2 - x)^3*abs(L)*abs(x)^2 - 36*abs(L/2 - x)^3*abs(L)^2*abs(x) - 144*abs(L)^2*abs(x)^2*abs(L - x)^2 - 36*abs(L/2 - x)^2*abs(L)*abs(L - x)^3 - 36*abs(L/2 - x)^3*abs(L)*abs(L - x)^2))

where L is the average length of the domain of influence.

## 5 HIERARCHICAL TYPE 1 REFINEMENT OF THE STIFFNESS MATRIX

The derivation of the shape functions and their derivatives in two parts leads to the adaptive formulation of the stiffness matrix when extra nodes are added in the problem domain. Therefore the shape functions of the initial m and the additional n nodes in the problem domain are given by eq. (27). Similarly, the partial derivatives of the shape functions are expressed by:

$$\Phi'_{1:m}{}^{enriched}(x) = \Phi'_{1:m}{}^{initial}(x) + \delta\Phi'_{1:m}(x) \tag{38}$$

$$\Phi'_{(m+1):(m+n)}{}^{additional}(x) = \Phi'_{(m+1):(m+n)}(x) \tag{39}$$

The enriched global stiffness matrix components is obtained from eq.(13):

$$K_{ij}{}^{enriched} = \int_{\Omega} B_i^{enriched\,T} \mathrm{E}\, B_j^{enriched} d\Omega \tag{40}$$

where for 2D elasticity problems

$$B_i^{enriched} = \begin{bmatrix} \Phi_{i,x}^{enriched} & 0 \\ 0 & \Phi_{i,y}^{enriched} \\ \Phi_{i,y}^{enriched} & \Phi_{i,x}^{enriched} \end{bmatrix}$$

$$= \begin{cases} \begin{bmatrix} \Phi_{i,x}^{initial} + \delta\Phi_{i,x} & 0 \\ 0 & \Phi_{i,y}^{initial} + \delta\Phi_{i,y} \\ \Phi_{i,y}^{initial} + \delta\Phi_{i,y} & \Phi_{i,x}^{initial} + \delta\Phi_{i,x} \end{bmatrix}, & i \le m \\ \begin{bmatrix} \Phi_{i,x}^{additional} & 0 \\ 0 & \Phi_{i,y}^{additional} \\ \Phi_{i,y}^{additional} & \Phi_{i,x}^{additional} \end{bmatrix}, & m \le i \le (m+n) \end{cases} \tag{41}$$

Further elaboration on the expression of the strain matrix $B_i^{enriched}$ leads to:

$$B_i^{enriched} = \begin{cases} \underbrace{\begin{bmatrix} \Phi_{i,x}^{initial} & 0 \\ 0 & \Phi_{i,y}^{initial} \\ \Phi_{i,y}^{initial} & \Phi_{i,x}^{initial} \end{bmatrix}}_{B_i^{initial}} + \underbrace{\begin{bmatrix} \delta\Phi_{i,x} & 0 \\ 0 & \delta\Phi_{i,y} \\ \delta\Phi_{i,y} & \delta\Phi_{i,x} \end{bmatrix}}_{\delta B_i}, & i \le m \\ \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}}_{B_i^{initial}} + \underbrace{\begin{bmatrix} \Phi_{i,x}^{additional} & 0 \\ 0 & \Phi_{i,y}^{additional} \\ \Phi_{i,y}^{additional} & \Phi_{i,x}^{additional} \end{bmatrix}}_{\delta B_i}, & m \le i \le (m+n) \end{cases} \tag{42}$$

$$= \left[B_i^{initial}\right] + \left[\delta B_i\right]$$

A similar expression can be derived for 3D elasticity problems.

After substituting eq.(42) into eq.(40) the stiffness matrix can be written as follows:

$$K_{ij}{}^{enriched} = \int_{\Omega} \left[\left[B_i^{initial}\right] + \left[\delta B_i\right]\right]^T \mathrm{E} \left[\left[B_i^{initial}\right] + \left[\delta B_i\right]\right] d\Omega$$

$$= \underbrace{\int_{\Omega} \left[\left[B_i^{initial}\right]\right]^T \mathrm{E} \left[\left[B_i^{initial}\right]\right] d\Omega}_{K_{ij}{}^{initial}} + \tag{43}$$

$$+ \underbrace{\int_{\Omega} \left[\left[B_i^{initial}\right]\right]^T \mathrm{E} \left[\left[\delta B_i\right]\right] d\Omega + \int_{\Omega} \left[\left[\delta B_i\right]\right]^T \mathrm{E} \left[\left[\delta B_i\right]\right] d\Omega + \int_{\Omega} \left[\left[\delta B_i\right]\right]^T \mathrm{E} \left[\left[B_i^{initial}\right]\right] d\Omega}_{\delta K_{ij}}$$

$$\boldsymbol{K}^{enriched} = \boldsymbol{K}^{initial} + \boldsymbol{\delta K} \tag{44}$$

The first part of $\boldsymbol{K}^{enriched}$ corresponds to the first m nodes, while the second term $\boldsymbol{\delta K}$ contains the contribution of the additional nodes and their interactions with the existing terms. In this respect, only $\boldsymbol{\delta K}$ matrix needs to be calculated at the refinement step, instead of recalculating the enriched matrix from the beginning, a process that saves substantial computational effort.

In order to demonstrate the stiffness matrix topology of the hierarchical h-type refinement, the 2D cantilever of Fig. 7 subjected to a parabolic load at the free end is considered. Problem details are given in Table 1.

| | |
|---|---|
| Loading | P=1000 N |
| Young's modulus | E=3 x $10^7$ N/$m^2$ |
| Poisson's ratio | v=0.30 |
| Height of the beam | D=6m |
| Length of the beam | L=24m |

**Table 1 – 2D cantilever problem details**

The beam is analyzed with a 11x4 node distribution as shown in Fig. 7. The set of nodes with initial and added nodes is illustrated in Fig. 8. Background cells are considered for the numerical integration of the weak form. In each Gauss cell a 2x2 Gauss quadrature is used. A linear basis and a cubic spline weight function are used for the MLS approximation. The support domain is rectangular with dimension 2 times the nodal spacing. The penalty method was chosen for imposing the boundary conditions with a penalty factor of $\alpha = 10^{10}$.

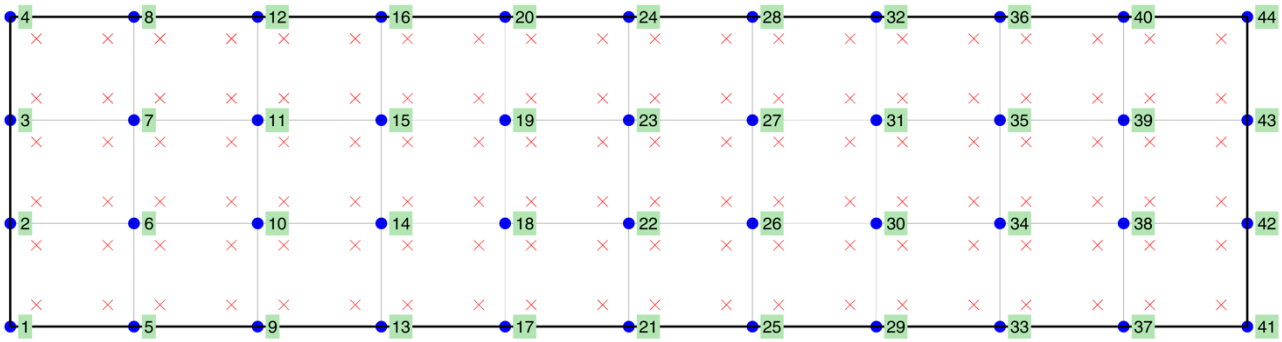



**Fig. 7 - Initial arrangement of nodes ● and integration points ×**

A node enrichment is subsequently performed with the addition of 8 new nodes in the problem domain indicated with the symbol **O** as seen in Fig. 8.

The calculation of the stiffness matrix is performed with standard procedure and the proposed two hierarchical schemes. For the second hierarchical procedure, only the additional matrix $\boldsymbol{\delta K}$ is computed and then added to the existing initial stiffness matrix.

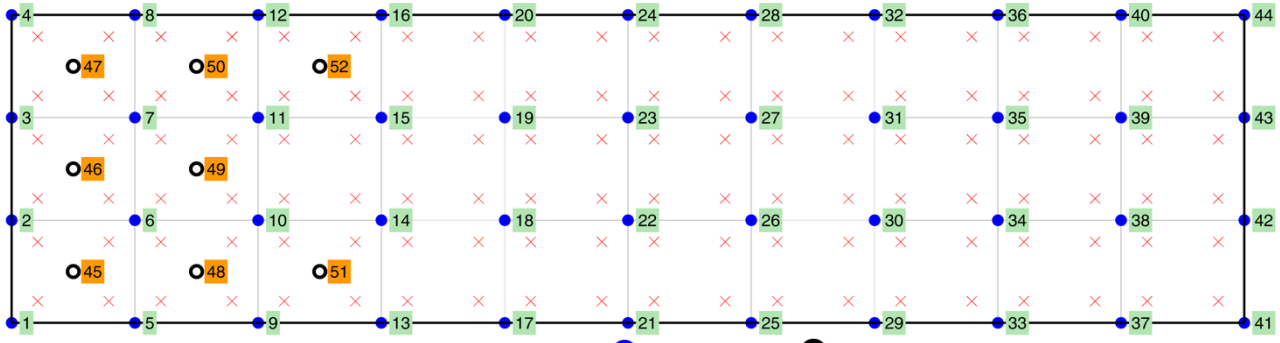
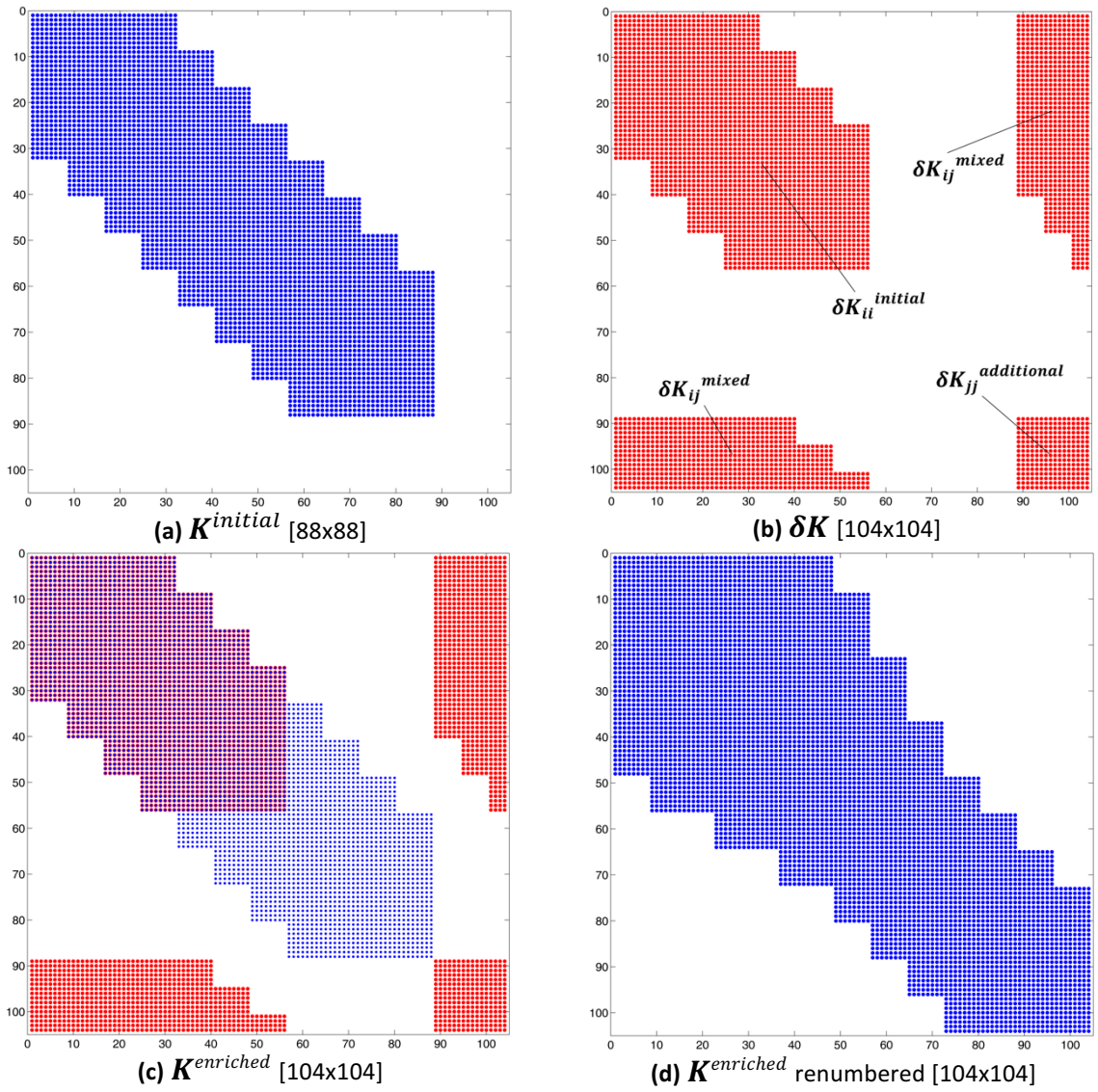



**Fig. 8 – Initial nodes ●, added nodes ○**

Fig. 9 illustrates the topology of the non-zero elements of the stiffness matrices before and after refinement. Matrix $\boldsymbol{\delta K}$ expresses the influence of the additional nodes. Fig. 9(d) depicts $\boldsymbol{K}^{enriched}$ after renumbering.




(a) $\boldsymbol{K}^{initial}$ [88x88] (b) $\boldsymbol{\delta K}$ [104x104]

(c) $\boldsymbol{K}^{enriched}$ [104x104] (d) $\boldsymbol{K}^{enriched}$ renumbered [104x104]

**Fig. 9 – Non-zero elements of the stiffness matrices**

## 6 HIERARCHICAL TYPE 2 REFINEMENT OF THE STIFFNESS MATRIX

The addition of extra nodes and their corresponding influence can be treated slightly different, with respect to the existing nodes, leading to a purely hierarchical formulation of the shape functions and derivatives and subsequently of the stiffness matrix. In this context the additional refining shape functions are treated hierarchically without altering the previously calculated shape functions and their derivatives.

The addition of the n nodes creates n shape functions according to eq.(36) and alters the m previous ones as follows:

$$\Phi_{1:m}(\mathrm{x}) = \underbrace{p^T(\mathrm{x})[A_{1:m}(\mathrm{x})]^{-1}W_{1:m}(\mathrm{x})}_{Initial\ \Phi_{1:m}} + \underbrace{p^T(\mathrm{x})[X]W_{1:m}(x)}_{(m+n)^{th}node\ influence\ \delta\Phi_{1:(m+n)}} \tag{45}$$

We consider the (m+n) node influence as additional refining functions that enrich the existing shape function field and choose to retain $\Phi_{1:m}(\mathrm{x})$ unchanged. In order to achieve this type of hierarchical formulation the last term of eq.(45) is omitted leading to:

$$\Phi_{1:m}(\mathrm{x}) = \underbrace{p^T(\mathrm{x})[A_{1:m}(\mathrm{x})]^{-1}W_{1:m}(\mathrm{x})}_{Initial\ \Phi_{1:m}} \tag{46}$$

which introduces an approximation since it violates the partition of unity.

The enriched shape function values is now applied for the extra nodes only, without the need for modification of the existing shape functions. Therefore, for the existing nodes:

$$\Phi_{1:m}(\mathrm{x})^{enriched} = \Phi_{1:m}(\mathrm{x})^{initial} \tag{47}$$

and for the additional nodes (refinement functions):

$$\Phi_{(m+1):(m+n)}(\mathrm{x})^{enriched} = \Phi_{(m+1):(m+n)}(\mathrm{x})^{hier} \tag{48}$$

where $\Phi_{(m+1):(m+n)}(\mathrm{x})^{hier}$ indicates the values of the shape functions of the extra nodes that are added hierarchically and are computed according to eq.(36). The partial derivatives of the shape functions that are used for the construction of the stiffness matrix are subsequently given by:

$$\Phi'_{1:m}(\mathrm{x})^{enriched} = \Phi'_{1:m}(\mathrm{x})^{initial} \tag{49}$$

$$\Phi'_{(m+1):(m+n)}(\mathrm{x})^{enriched} = \Phi'_{(m+1):(m+n)}(\mathrm{x})^{hier} \tag{50}$$

The global stiffness matrix component $K_{ij}$ is assembled using eq.(13):

$$K_{ij}{}^{enriched} = \int_{\Omega} B_i^{enriched^T} \mathrm{E}\, B_j^{enriched} d\Omega \tag{51}$$

where

$$B_i^{enriched} = \begin{bmatrix} \Phi_{i,x}^{enriched} & 0 \\ 0 & \Phi_{i,y}^{enriched} \\ \Phi_{i,y}^{enriched} & \Phi_{i,x}^{enriched} \end{bmatrix} =$$

$$= \begin{cases} \begin{bmatrix} \Phi_{i,x}^{initial} + \delta\Phi_{i,x} & 0 \\ 0 & \Phi_{i,y}^{initial} + \delta\Phi_{i,y} \\ \Phi_{i,y}^{initial} + \delta\Phi_{i,y} & \Phi_{i,x}^{initial} + \delta\Phi_{i,x} \end{bmatrix} = \left[\boldsymbol{B_i^{initial}}\right], \ i \le m \\ \begin{bmatrix} \Phi_{i,x}^{additional} & 0 \\ 0 & \Phi_{i,y}^{additional} \\ \Phi_{i,y}^{additional} & \Phi_{i,x}^{additional} \end{bmatrix} = \left[\boldsymbol{B_i^{hier}}\right], \ m \le i \le (m+n) \end{cases} \tag{52}$$

Substituting eq.(52) into eq.(51), the strictly hierarchical stiffness matrix can be written in parts as follows :

$$\boldsymbol{K^{hier}} = \begin{bmatrix} \underbrace{K_{ii}{}^{\boldsymbol{initial}}}_{2m\ x\ 2m} & \delta K_{ij}{}^{mixed} \\ \delta K_{ji}{}^{mixed} & \underbrace{\delta K_{jj}{}^{additional}}_{2n\ x\ 2n} \end{bmatrix} \tag{53}$$

where

$$K_{ii}{}^{\boldsymbol{initial}} = \int_{\Omega} B_i^{\boldsymbol{initial}^T} \mathrm{E}\, B_i^{\boldsymbol{initial}} d\Omega \tag{54}$$

$$\delta K_{ij}{}^{mixed} = \delta K_{ji}{}^{mixed} = \int_{\Omega} B_i^{\boldsymbol{initial}^T} \mathrm{E}\, B_j^{hier} d\Omega \tag{55}$$

$$\delta K_{jj}{}^{additional} = \int_{\Omega} B_j^{hier^T} \mathrm{E}\, B_j^{hier} d\Omega \tag{56}$$

Through the procedure described above, the addition of extra nodes in the problem domain is taken into consideration without modification of the existing stiffness matrix. In eq.(53) $\delta K_{jj}{}^{additional}$ is the stiffness matrix of the additional forms, while $\delta K_{ij}{}^{mixed}$ expresses the interaction between the initial and the newly added nodes. This type of purely hierarchical formulation reduces further the computational effort for the assembly of the enriched stiffness matrix.

Moreover, specially tailored solution procedures can be applied that avoid refactorization of the initial stiffness matrix in a similar manner as applied in hierarchical FEM, which further reduces the solution effort of the resulting algebraic equations.

The proposed hierarchical type 2 refinement formulation is demonstrated using the 2D cantilever of Fig. 7 with the refinement scheme depicted in Fig. 8. The addition of the extra nodes is performed with the hierarchical concept described previously. The already calculated values of the shape functions of the first 44 nodes, are kept unaltered, and the influence of the shape functions of the 8 additional nodes to the initial nodes is omitted. In Fig. 10 the additional shape function of node 48 and the influence on the initial nodes is demonstrated.

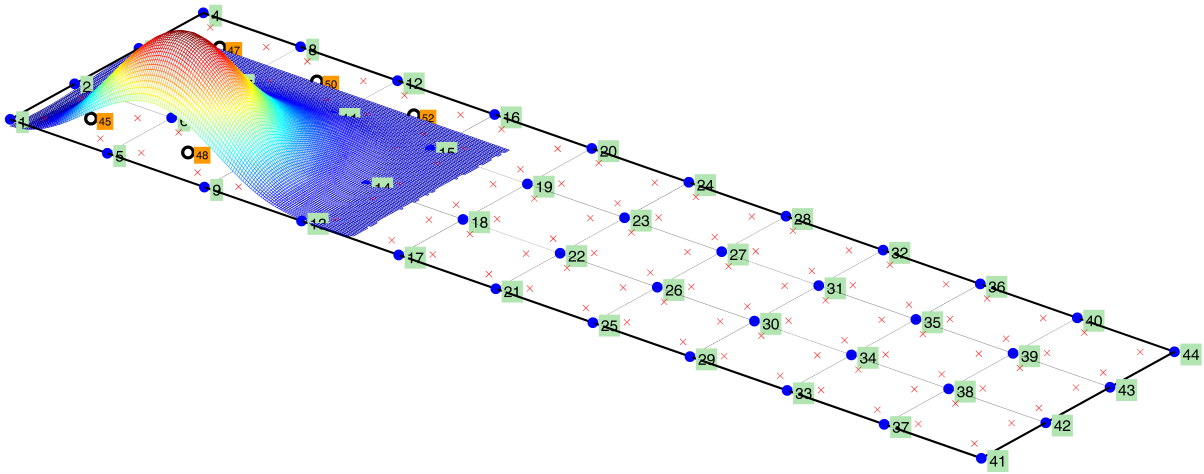



**Fig. 10 - Additional shape function of the 48th node and its DoI**

The stiffness matrix of the initial 44 nodes $K_{ii}^{\boldsymbol{initial}}$, is not affected by the additional nodes. To perform the hierarchical refinement only matrices $\delta K_{ij}^{mixed}$ and $\delta K_{jj}^{additional}$ from eq.(55) and eq.(56) need to be computed. Fig. 11 illustrates the sparsity pattern of the matrices, used to assemble the final stiffness matrix of the cantilever in which $\delta K_{jj}^{additional}$ is (16x16) and $\delta K_{ij}^{mixed}$ is (16x56). The final $K^{enriched}$ has the same pattern of non-zero terms as depicted in Fig. 9c and 9d of the type-1 hierarchical refinement. However, the stiffness terms corresponding to the initial nodes are not modified as per the additional terms illustrated in Fig. 9b.

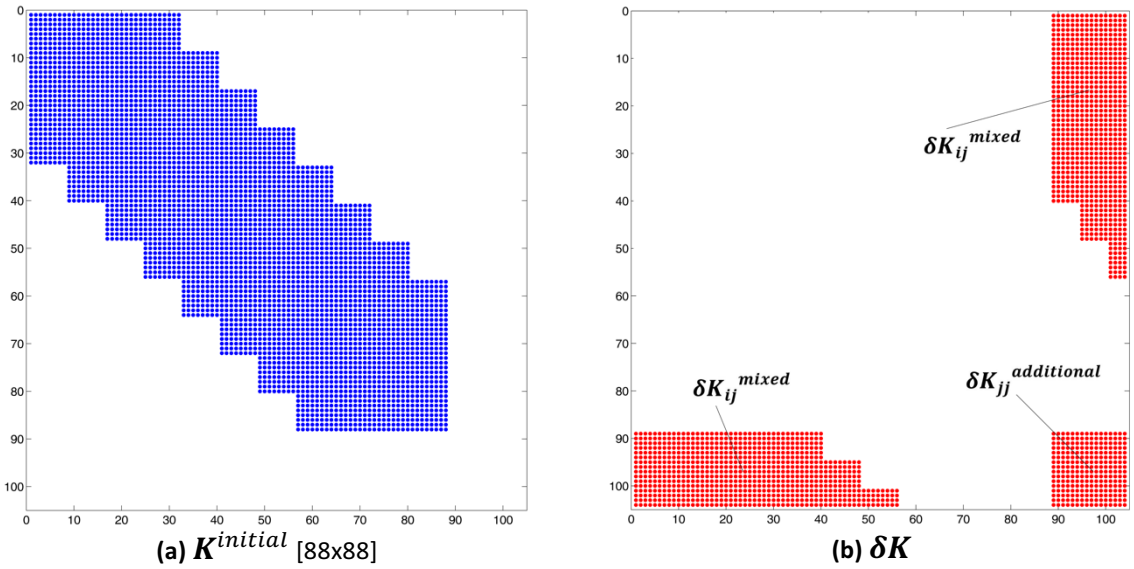



**Fig. 11 - Non-zero elements of the stiffness matrices**

In order to compare the two hierarchical formulations proposed, the energy and displacement norms are calculated before and after the addition of new nodes. For illustration purposes, only the shear stress field for each case is plotted in Fig. 12.

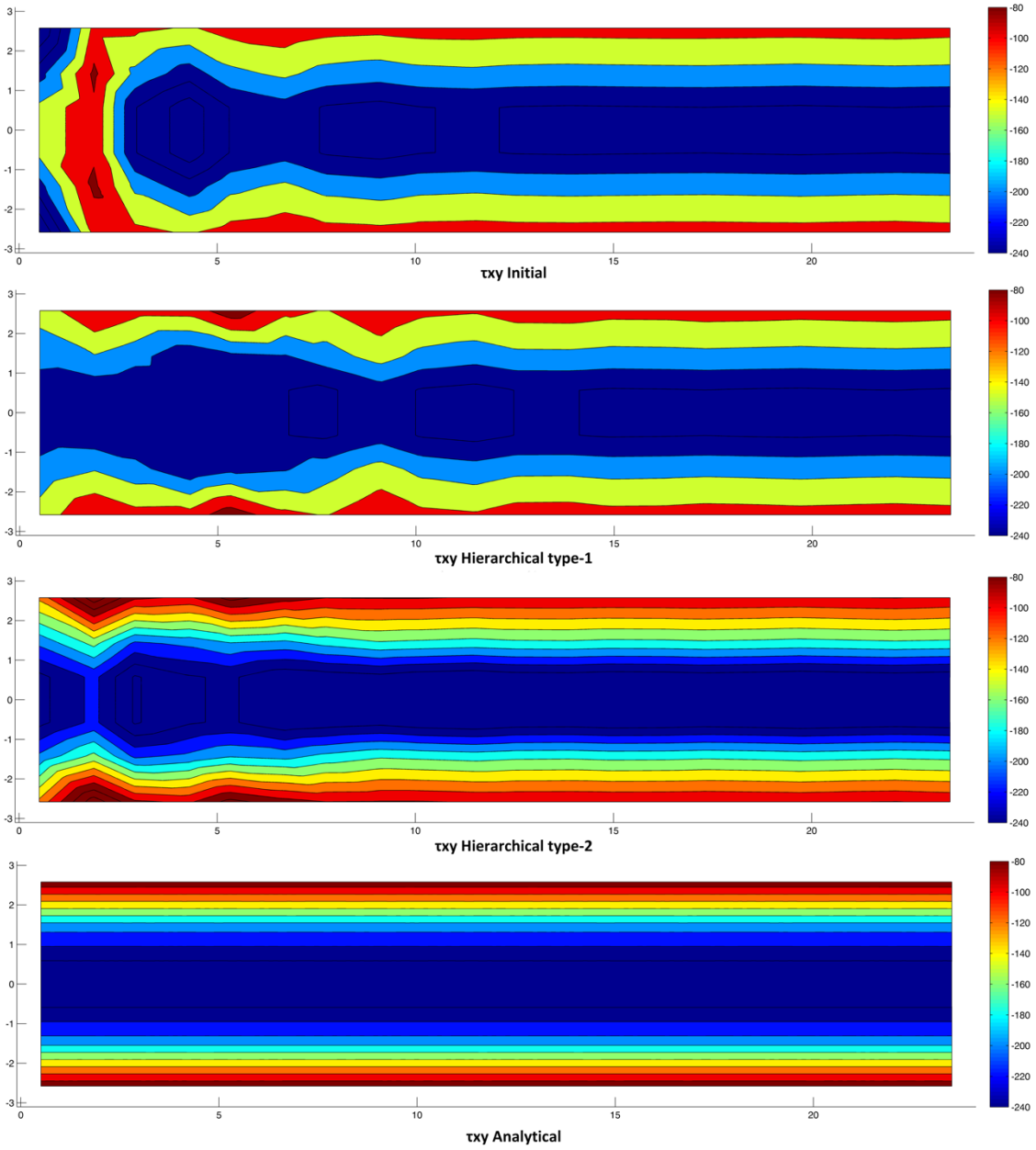



**Fig. 12 - Shear stress field for the cantilever beam**

| | Initial vs Analytical | Hierarchical **type 1** vs Analytical | Hierarchical **type 2** vs Analytical |
|---|---|---|---|
| Energy norm | 0.2362 | 0.0977 | 0.0688 |
| Displacement norm | 5,98E-05 | 7,37E-06 | 1,47E-06 |

**Table 2 - Energy and Displacement norms for the cantilever beam**

As can be seen from Fig. 12 and Table 2, type-2 hierarchical formulation gives results close to type-1 solution. The calculated displacement and stress fields are improved and overall much higher accuracy is obtained.

## CONCLUSION

Two h-type hierarchically refinement schemes of the element free Galerkin method are proposed in which the expressions of the shape functions have been separated in two parts leading to a hierarchical decomposition of the stiffness matrix. In the first refinement scheme the contribution of the additional nodes in a problem domain can be achieved with the computation of an additional matrix **δK**, which is combined with the existing stiffness matrix describing the initial arrangement of nodes in the problem domain, to give the final stiffness matrix. In the second refinement scheme, the addition of extra nodes is taken into consideration, without the need for recalculation or modification of the existing stiffness matrix.

**This work has been supported by the European Research Council Advanced Grant "MASTER—Mastering the computational challenges in numerical modeling and optimum design of CNT reinforced composites" (ERC-2011- ADG 20110209).**

## REFERENCES

[1] T. Belytschko, Y.Y. Lu and L. Gu, "Element-free Galerkin methods", *International Journal for Numerical Methods in Engineering*, vol. 37, 1994, pp. 229-256.

[2] Y.Y. Lu, T. Belytschko, L. Gu, "A new implementation of the element free Galerkin method", *Computer Methods in Applied Mechanics and Engineering*, vol. 113 (3-4), 1996, pp. 397-414.

[3] T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, and P. Krysl, "Meshless methods: An overview and recent developments", *Computer Methods in Applied Mechanics and Engineering*, vol. 139, 1996, pp. 3–47.

[4] H.-J. Chung, T. Belytschko, "An error estimate in the EFG method", *Computational Mechanics*, vol. 21, 1998, pp. 91-100.

[5] S. Li, W.K. Liu, "Meshfree Particle Methods", Springer, 2004.

[6] S. Li, W.K. Liu, "Meshfree and particle methods and their applications", *Applied Mechanics Reviews*, vol. 55, 2002, pp. 1–34.

[7] V.P. Nguyen, T. Rabczuk, S. Bordas, and M. Duflot, "Meshless methods: A review and computer implementation aspects", *Mathematics and Computers in Simulation*, vol. 79, 2008, pp. 763–813.

[8] J.G. Wang, G.R. Liu, "A point interpolation meshless method based on radial basis functions", *International Journal for Numerical Methods in Engineering*, vol. 54, 2002, pp. 1623–1648.

[9] P. Metsis and M. Papadrakakis, "Overlapping and non-overlapping domain decomposition

methods for large-scale meshless EFG simulations", *Computer Methods in Applied Mechanics and Engineering*, vol. 229-232, 2012, pp. 128–141.

[10] A. Karatarakis, P. Metsis, M. Papadrakakis, "GPU-acceleration of stiffness matrix calculation and efficient initialization of EFG meshless methods", *Computer Methods in Applied Mechanics and Engineering*, vol. 258, 2013, pp. 63–80.

[11] K. S. Miller, "On the inverse of the Sum of Matrices", *Mathematics Magazine*, vol. 54 No. 2, 1981, pp 67-72.

[12] S. Fernández-Méndez, A. Huerta, "Imposing essential boundary conditions in mesh-free methods", *Computer Methods in Applied Mechanics and Engineering*, vol. 193 (12-14), 2004, pp. 1257-1275.

[13] C.K. Lee, C.E. Zhou, "On error estimation and adaptive refinement for element free Galerkin method: Part I: Stress recovery and a posteriori error estimation", *Computers and Structures*, vol. 82 (4-5), 2004, pp. 413-428.

[14] C. Wenterodt, O. von Estorff, "Optimized meshfree methods for acoustics", *Computer Methods in Applied Mechanics and Engineering*, vol. 200 (25-28), 2011, pp. 2223-2236.

[15] C.V. Le, H. Askes, M. Gilbert, "Adaptive element-free Galerkin method applied to the limit analysis of plates", *Computer Methods in Applied Mechanics and Engineering*, vol. 199 (37-40), 2010, pp. 2487-2496.

[16] K.T. Danielson, S. Hao, W.K. Liu, R.A. Uras, S. Li, "Parallel computation of meshless methods for explicit dynamic analysis", *International Journal for Numerical Methods in Engineering*, vol. 47, 2000, pp. 1323–1341.

[17] W.-R. Yuan, P. Chen, K.-X. Liu, "High performance sparse solver for unsymmetrical linear equations with out-of-core strategies and its application on meshless methods", *Applied Mathematics and Mechanics (English Edition)*, vol. 27, 2006, pp. 1339–1348.

[18] S.C. Wu, H.O. Zhang, C. Zheng, J.H. Zhang, "A high performance large sparse symmetric solver for the meshfree Galerkin method", *International Journal of Computational Methods*, vol. 5, 2008, pp. 533–550.

[19] O.C. Zienkiewicz, J.Z. Zhu, "The superconvergent patch recovery and a posteriori error estimates. Part 2: Error estimates and adaptivity", *International Journal for Numerical Methods in Engineering,* vol. 33, 1992, 1365-1382.

[20] M. Tabbara, T. Blacker, T. Belytschko, "Finite element derivative recovery by moving least square interpolants", *Computer Methods in Applied Mechanics and Engineering*, vol. 117(1–2), 1994, pp. 211–223.

[21] Szabó B, Babuška I. Finite Element Analysis. Wiley: New York, 1991.

[22] P. Metsis, N. Lantzounis, M. Papadrakakis, A new hierarchical partition of unity formulation of EFG meshless methods, Comput. Methods Appl. Mech. Engrg. 283 (2015) 782-805

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MODERNIZING SOFTWARE IN SCIENCE AND ENGINEERING: FROM C/C++ APPLICATIONS TO MOBILE PLATFORMS

**Liliana Favre**[1,2]**, Claudia Pereria**[1]**, and Liliana Martinez**[1]

[1] Universidad Nacional del Centro de la Provincia de Buenos Aires
Tandil, Argentina
[2] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires,
Argentina

{lfavre, cpereira, lmartine}@exa.unicen.edu.ar

**Keywords:** Software Modernization, Mobile Development, Model Driven Development, Model Driven Architecture, Migration, HAXE, C++.

**Abstract.** *The evolution of software technology leads to continuous migration of software components and applications. Particularly, most software applications in applied science and engineering are for desktop computers and there is a need to migrate them to mobile technologies. This kind of migration faces many challenges due to the proliferation of different mobile platforms. New programming languages are thus emerging to integrate the native behaviors of the different platforms targeted in development projects. In this direction, the HAXE language allows writing mobile applications that target all major mobile platforms. Novel technical frameworks for information integration and tool interoperability such as the Model Driven Development (MDD) can help to manage a huge diversity of mobile technologies. A specific realization of MDD is the Model Driven Architecture (MDA) proposed by the Object Management Group (OMG). In this work, we propose a migration process from C/C++ software to different mobile platforms that integrates MDA standards with HAXE. C/C++ is one of the most commonly used programming language in science and engineering domains and numerous legacy software components written in C++ require to be modernized. On the one hand, the proposed process follows model-driven principles: all artifacts involved in the process can be viewed as models that conform a particular metamodel, the process itself can be viewed as a sequence of model-to-model transformations and all the extracted information is represented in a standard way through metamodels. On the other hand, HAXE easily adapts the native behaviors of the different platforms targeted in development projects enabling extremely efficient cross-platform development, ultimately saving time and resources. The proposal was validated in Eclipse Modeling Framework considering that some of its tools and run-time environments are aligned with MDA standards. The paper includes a simple case study, the migration of a C++ application, "the Set of Mandelbrot", that allow us to exemplify the different steps of the process.*

## 1 INTRODUCTION

Mobile devices such as smartphones or tablets come with engineers and researchers all the time and everywhere. Among other novel features, mobile devices contain global positioning sensors, wireless connectivity, built-in web browsers and photo/video/voice capabilities. These features offer the possibility to adapt functionality to user needs and preferences and can be used to build context-aware applications [29].

The evolution of software technology leads to continuous migration of software components and applications. Particularly, most software applications in applied science and engineering are for desktop computers and there is a need to migrate them to mobile technologies. This kind of migration faces many challenges due to the proliferation of different mobile platforms. The high cost and technical complexity of targeting development to a wide spectrum of platforms, has forced developers to make applications tailored for each type of device. New programming languages are thus emerging to integrate the native behaviors of the different platforms targeted in development projects. In this direction, the HAXE [10] language allows writing mobile applications that target all major mobile platforms, such as Android, iOS and BlackBerry, in a straightforward way.

Novel technical frameworks for information integration and tool interoperability such as the Model Driven Development (MDD) can help to manage a huge diversity of mobile technologies [6]. MDD provides principles and techniques to represent software through models at different abstraction levels. A specific realization of MDD is the Model Driven Architecture (MDA) proposed by the Object Management Group (OMG) [23]. The outstanding ideas behind MDA are separating the specification of the system functionality from its implementation on specific platforms, managing the software evolution from abstract models to implementations. Models play a major role in MDA which distinguishes Computation Independent Model (CIM), Platform Independent Model (PIM) and Platform Specific Model (PSM). Some authors also distinguish Implementation Specific Model (ISM) as a description (specification) of the system in source code.

The essence of MDA is MOF (Meta Object Facility), an OMG standard for defining metamodels that provides the ability to design and integrate semantically different languages such as general-purpose languages, domain specific languages and modeling languages in a unified way. Significant advantages can be made of this unification to construct powerful mobile design environments. The modeling concepts of MOF are classes, which model MOF meta-objects; associations, which model binary relations between meta-objects; Data Types, which model other data; and Packages, which modularize the models [21]. Consistency rules are attached to metamodel components by using OCL [22]. MOF provides two metamodels EMOF (Essential MOF) and CMOF (Complete MOF). EMOF favors simplicity of implementation over expressiveness. CMOF is a metamodel used to specify more sophisticated metamodels. The MOF 2.0 Query, View, Transformation (QVT) metamodel is the standard for expressing transformations [25].

The Architecture Driven Modernization (ADM) approach has established a set of solutions for information system modernization [2]. ADM is defined as “the process of understand and evolve existing software assets for the purpose of software improvement, modifications, interoperability, refactoring, restructuring, reuse, porting, migration, translation, integration, service-oriented architecture deployment”. The OMG ADM Task Force (ADMTF) is developing a set of standards (metamodels) to facilitate interoperability between modernization tools.

The success of approaches such as ADM and MDA depend on the existence of CASE tools that make a significant impact on software processes such as forward engineering and reverse engineering processes [8]. The Eclipse Modeling Framework (EMF) was created for facilitating system modeling and the automatic generation of Java code [26]. EMF started as an implementation of MOF resulting ECORE, the EMF metamodel comparable to EMOF. EMF has evolved starting from the experience of the Eclipse community to implement a variety of tools and to date is highly related to MDD [15]. The subproject M2M supports model-to-model transformations that take one or more models as input to produce one or more models as output. For instance, ATL (Atlas Transformation Language) is a model transformation language that is developed on top of the Eclipse platform [4]. Another subproject is ACCELEO, which is an implementation of the M2T (Model-to-Text) transformation standard of the OMG for EMF-based models [1]. It is used in forward engineering processes.

In the Eclipse environment, MDA is integrated with Java language but it is weakly supported for other programming languages such as C++ [26]. In particular, C++ is one of the most commonly used programming language in science and engineering domains. Numerous legacy software components written in C++ require to be modernized. EMF4CPP is the first step at providing a set of tools for MDD in C++ as an alternative to the Eclipse tools for Java [16]. It is a C++ implementation and type mapping for the Eclipse Modeling Framework core, the ECORE metamodel. The main facilities provided by EMF4CPP are to generate C++ code from ECORE metamodels and to parse and serialize models and metamodels from and into XMI documents [30]. However, an implementation of a MOF-compliant C++ metamodel would be necessary for other MDD processes (for example, reverse engineering or software modernization).

In this work, we propose a migration process from C/C++ software to different mobile platforms that integrates MDA standards with HAXE. An ECORE metamodel for the C++ language is provided. On the one hand, the process follows model-driven principles: all artifacts involved in the process can be viewed as models that conform a particular metamodel, the process itself can be viewed as a sequence of model-to-model transformations and all the extracted information is represented in a standard way through metamodels. On the other hand, HAXE easily adapts the native behaviors of the different platforms targeted in development projects enabling extremely efficient cross-platform development, ultimately saving time and resources.

The paper includes a simple case study, the migration of a C++ application “the Set of Mandelbrot” that allow us to exemplify the different steps of the process. We believe that our approach provides benefits with respect to processes based only on traditional ad-hoc migration techniques increasing productivity due to the automation introduced in the generation of the new software.

The paper is organized as follows. Section 2 presents related work. In Section 3, we summarizes our contribution: a migration process from C++ to mobile platforms. Section 4 describes the HAXE language and its metamodel. Section 5 describes the C++ metamodel. Section 6 details the different stages of the migration process through of a simple case study. Finally, in Section 7 and 8 we present discussion and conclusions respectively.

## 2 RELATED WORK

In this section, we describe existing approaches for the development of mobile applications related to MDD.

Reference [7] proposes a new software architecture with the objective of providing the same service as mobile Web service as well as mobile application. The authors report on the feasibility study that they conducted in order to evaluate whether to use model driven software development for developing mobile applications. They argue that the architecture is flexible enough to support mobile Web services and mobile applications at the same time. They have develop a metamodel to describe mobile application and have shown how to generate mobile application from that model.

The project BAMOS and an architecture designed and implemented for the generic and flexible development of mobile applications is described in [14]. A declarative description of the available services supports the architecture. The authors describe a model driven approach for generating almost the complete source code of mobile services.

Reference [19] goes through mobile development process and architectural structures and their analysis with empirical mobile application development. The architecture and architecture role on the development has been studied in mobile application and multiplatform service development.

Various authors describe challenges of mobile software development, for example, in [11] authors highlight creating user interfaces for different kinds of mobile devices, providing reusable applications across multiple mobile platforms, designing context aware applications and handling their complexity and, specifying requirements uncertainty. Issues related to ensuring that the application provides sufficient performance while maximizing battery life are remarked in [28].

A proposal for supporting mobile application development by using models as inputs to an emulator is outlined at [5]. The authors describe an MDD-based emulator for using in the design of graphical interfaces and interactions. They propose transform functional behavior and requirement models with design restrictions into emulated applications.

Reference [20] describes a DSL (Domain Specific Language), named MobDSL, to generate applications for multiple mobile platforms. They perform the domain analysis on two cases in the Android and iPhone platforms. This analysis allows inferring the basic requirements of the language defined by MobDSL.

The proliferation of mobile devices generated the need to adapt desktop applications to mobile platforms. A reengineering process that integrates traditional reverse engineering techniques such as static and dynamic analysis with MDA is presented at [3]. The article describes a case study that shows how to move CRM (Customer Relationship Management) applications from desktop to mobile platforms. The proposal was validated in the open source application platform Eclipse, EMF, EMP, ATL and Android platform. Reference [13] describes a migration process from Java to mobile platforms through the multiplatform language HAXE.

ANDRIU, a reverse engineering tool based on static analysis of source code for transforming user interface tiers from desktop application to Android, is described in [24]. ANDRIU has been developed for migrating traditional systems to Android applications although it was designed to be extended for different migrations to others mobile platforms.

Reference [18] describes six major trends affecting future smartphone design and use: personal computers, internet of things, multimedia delivery, low power operation, wearable computing and context awareness.

In this paper, we describe an original model-driven migration process that includes automatic analysis of existing code, reverse engineering of abstract high-level models, model

transformation to target models in the HAXE platform and code generation. Our approach offers an opportunity for increasing the automation in software migration. In the following section, we detail the main steps of the proposed migration process.

## 3 OUR CONTRIBUTION : THE MIGRATION PROCESS

The migration process follows model driven development principles: all artifacts involved in the process can be viewed as models that conforms to ECORE meta-metamodels, the process itself can be viewed as a sequence of model-to-model transformations and the extracted information is represented in a standard way through ECORE metamodels. For each transformation, source and target metamodels are specified. A source metamodel defines the family of source models to which transformations can be applied. A target metamodel characterizes the generated models.

Figure 1 summarizes the proposed process. The first step is the reverse engineering of the C/C++ code to build a high-level model of it by using a model injector. Then, the obtained model is refactored to improve the design and eliminate dependencies of C/C++ language. The following step is related to model-to-model transformations that allow integrating C/C++ models with HAXE models. Specifically, this transformation is an ATL transformation that generates a model of the HAXE platform from a C/C++ model. Next, it is possible to generate source code in HAXE from the HAXE model by using M2T transformations expressed in ACCELEO. Considering that HAXE has one cross-platform standard library and various platform specific APIs used to natively access platform features, it is possible to write a mobile application once and have this application instantly available to different mobile devices.

The proposal was validated in the open source application platform Eclipse considering that some of its tools and run-time environments are aligned with MDA. C++ and HAXE metamodels were specified as ECORE instances and model-to-model transformations were implemented in ATL.

The HAXE Language and, HAXE and C/C++ metamodels are described in the following sections. The HAXE metamodel was previously described in a related work [12]. It is worth mentioning that our research also contributes the definition of an ECORE metamodel for the C/C++ language.

## 4 THE HAXE LANGUAGE

HAXE is an open-source high-level multiplatform programming language and compiler that can produce applications and source code for many different platforms from a single code-base [10].

Reference [9] summarizes the HAXE principles as follows: “support mainstream platforms”, “write once, reuse everywhere”, “always native, no wrapper”, “generated but readable” and “trust the developer”. Some languages allow cross-platform development, but neither their features nor their standard libraries are designed to run on multiple platforms. HAXE was designed from scratch to run and compile for many different platforms.

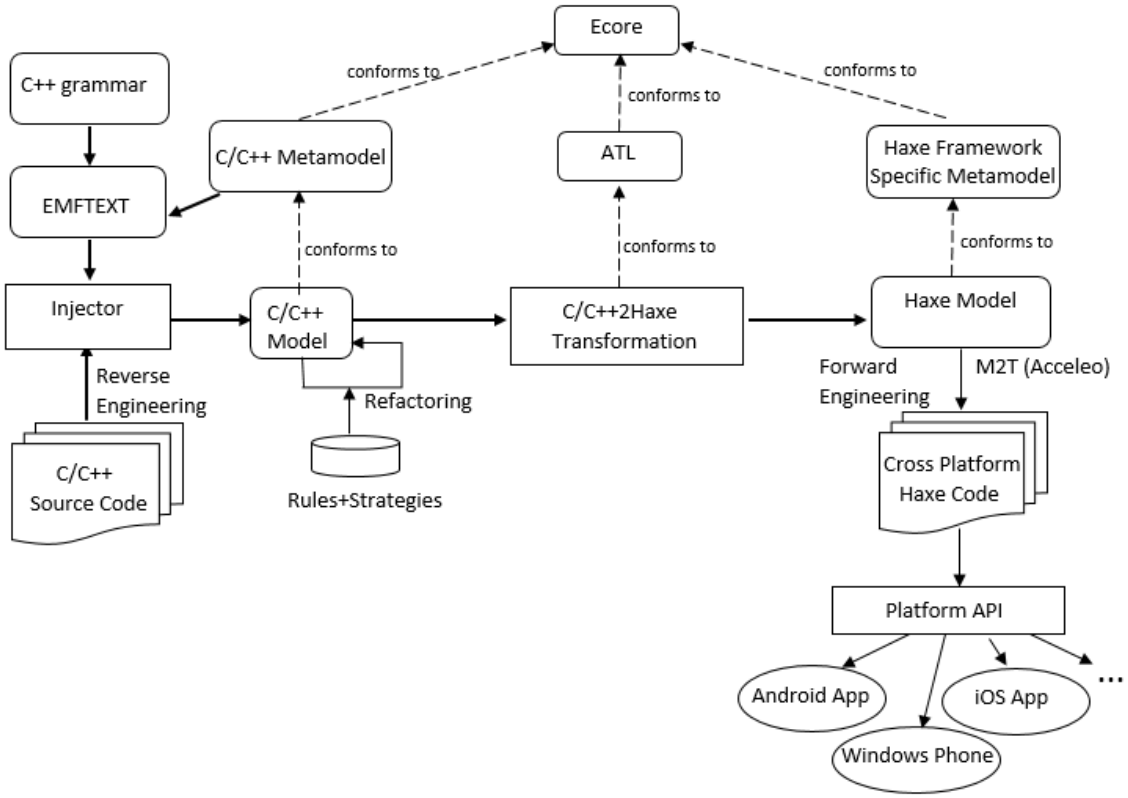



Figure 1. Our contribution: A Migration Process

The HAXE programming language is a high level programming language that mixes features of object-oriented languages and functional ones. It is similar (but not pure) to object-oriented languages. The compiler supports novel features such as type inference, enforcing strict type safety at compile time.

Since language can be compiled for different platforms, it is useful for a wide variety of applications such as games, web and mobile. In previous work, we show an integration the HAXE with MDD defining an ECORE metamodel of the Cross-Platform Framework HAXE. This metamodel corresponds to the HAXE version 3.1.3. In the context of MDA, the instances of this metamodel are platform specific models (PSM). This metamodel allowed us to integrate HAXE with MDA migration process from Java or C/C++ to mobile platform.

HAXE easily adapts the native behaviors of the different platforms targeted in development projects enabling extremely efficient cross-platform development, ultimately saving time and resources. Currently there are nine supported target languages: Javascript, Neko, PHP, Python, C++, Actionscript3, Flash, Java and, C#. In the context of Mobile App Development, HAXE allows writing mobile apps that target all major mobile platforms and run at native speed. The C++ target allows us to target Android or iOS and OpenFL (www.openfl.org) provides support for creating interfaces using a Flash-like API. OpenFL is a free and open source software framework and platform for the creation of multiplatform applications and video games. OpenFL programs are written in HAXE and may be published to Flash movies, or standalone applications for Microsoft Windows, Mac OS X, Linux, iOS, Android, BlackBerry OS, Firefox OS, HTML5 and Tizen.

Next, we summarizes the main characteristics of the HAXE metamodel.

### 4.1 The HAXE Metamodel

The HAXE metamodel conforms to ECORE and is partially shown in Figure 2. The main metaclasses of the HAXE metamodel are those that allow specifying an application using HAXE as language. One of the main metaclasses of the metamodel is *HAXEModel*, that serves as element container used to describe an application and store additional information on it, for example, some options of compilation and different metaclasses for modeling such as modules, classes and packages. *HAXEModel* owns *HAXEModule* and *HAXEPathReferentiable*. Starting from the relations *haxeModules*, *referenced* and *elements*, the class *HAXEModel* allows storing different information. Relation *haxeModules* allows accessing the different HAXE modules used in the project. Through relation *elements*, it is possible to access the different elements of the package tree. Relation *referenced* provides access to elements, which are referenced in the project but are not defined completely. In the case of relations and referenced elements, the type used is *HAXEPathReferentiable*, which is the parent type of metaclasses such as *HAXEType* and *HAXEPackage*. The HAXE language includes different kind of types such as class (the types class and interface), function, abstract type, enumeration, and anonymous structures. The full HAXE metamodel can be found in [12].

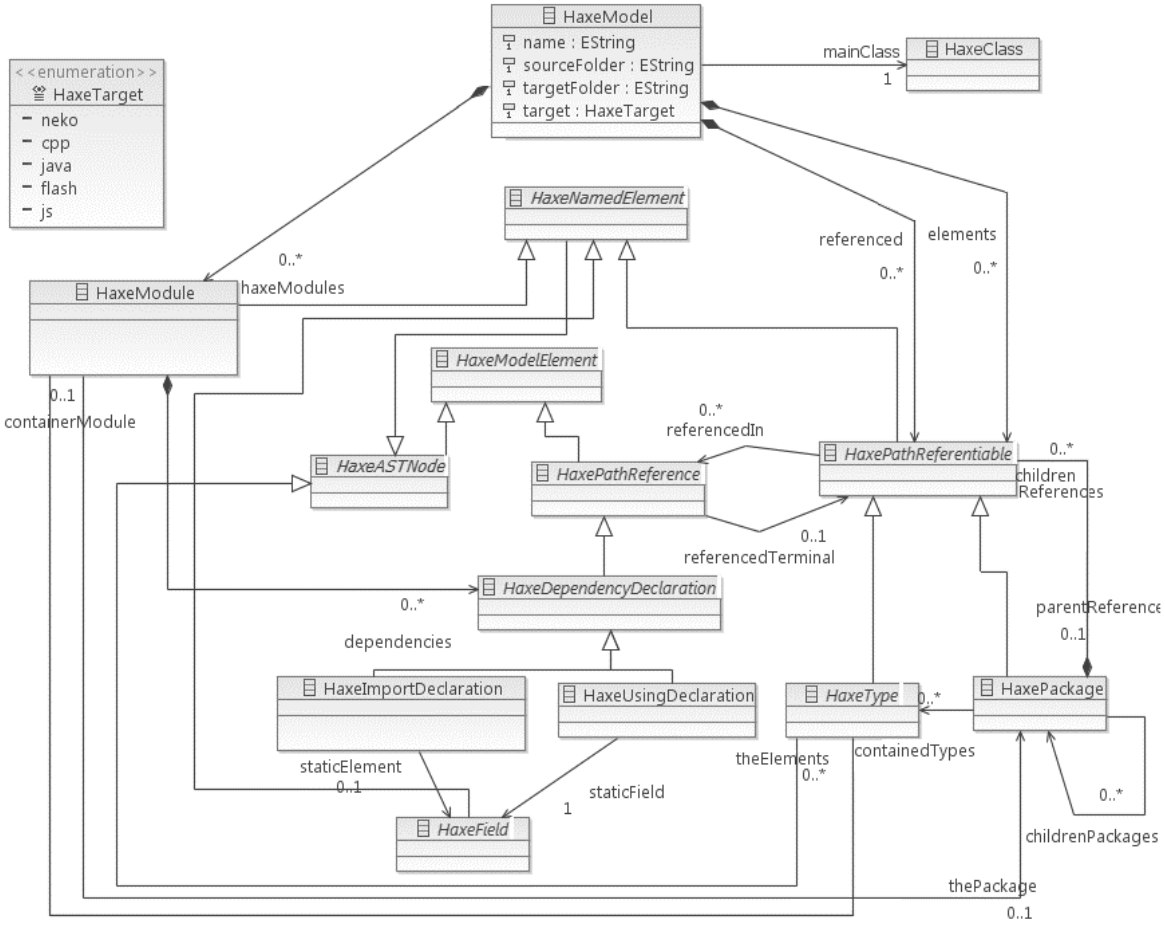



Figure 2. HAXE metamodel

## 5 THE C++ METAMODEL

The C++ metamodel conforms to ECORE and is partially shown in Figure 3. The root metaclass is *Program* that represents a C++ program, which owns source files, instances of *TranslationUnit*. A translation unit contains declarations such as block declaration, function definitions, template declarations, among others. A *SimpleDeclaration*, instance of *Block-Declaration*, has a *DeclSpecifierSeq* that is a sequence of *DeclSpecifiers* which refers to a declaration specifiers and a type specifier. In addition, a simple declaration has an *InitDeclaratorList* containing a variable declaration list that is a list of specifiers and the name of a variable and its corresponding initialization. A *FunctionDefinition* has a *Declarator* containing the function identifier and the parameter list. *Function* and *CtorOrDestFunction*, instances of *FunctionDefinition*, have a body that contains compound statements such as declarations, iterations, and selections. In addition, a *Function* has a *DeclSpecifierSeq* that is a sequence of *DeclSpecifiers* such as function specifiers and a type specifier. *TypeSpecifier* subclasses are *SimpleTypeSpecifier*, *ClassSpecifier* and *EnumSpecifier* among others. A *ClassSpecifier* has a *ClassHead* containing the class key (class or struct) and a *MemberSpecification* that contains *MemberDeclarations* such as variables, function declarations, function definitions, constructors, destructor, template members, etc.

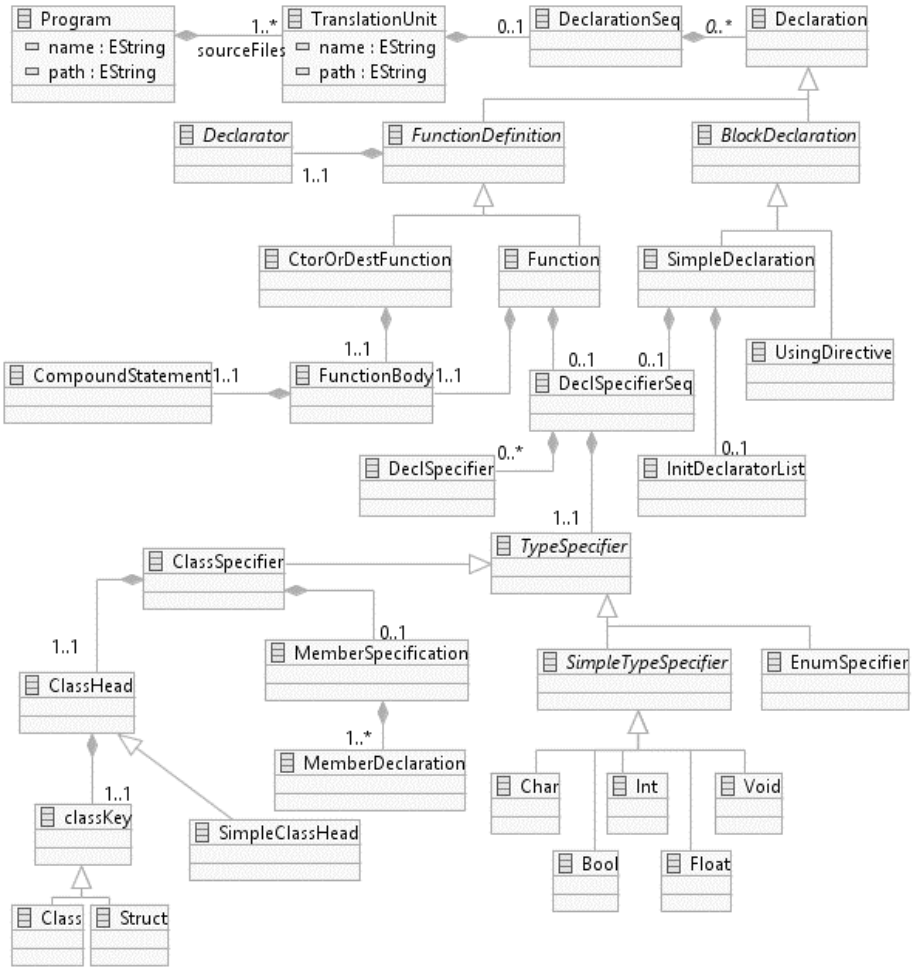



Figure 3. C++ metamodel

## 6 MIGRATING C/C++ TO MOBILE PLATFORMS

In this section, we describe a migration process from C++ code to different mobile platforms through HAXE. This process starts extracting models from the C++ legacy system. Next, these models are transformed into HAXE models that allow generating HAXE code which can be compiled to multiple target languages in a straightforward way.

To illustrate the migration process we describe a simple case study, how to migrate the C++ code of "the Set of Mandelbrot" to HAXE code. The original application consists of a main class, called Mandelbrot, that is responsible for the calculation of the set of Mandelbrot and serves as entry point for the application. It also includes several classes that collaborate in the realization of the task. The class Mandelbrot is responsible for generating the set and displaying it as image. To perform these tasks, the class depends on Picture and Complex classes, the first is used as a data type that supports the manipulation of digital images. The second class is a data type used to model complex number with their respective operations.

The following subsections describes the steps of the migration process.

### 6.1 Injection

This first step extracts a complete model of code from C++ code. To carry out this task, we constructed a model injector by using EMFText [17]. To generate this injector, EMFText requires the language metamodel and the concrete syntax specification. In our approach, to generate the injector we first specified the C++ metamodel based on the C++ grammar [27]. Then, we specified the concrete syntax that defines the textual representation of all metamodel concepts. Taking these specifications, the EMFText generator derives an advanced textual editor that uses a parser and printer to parse language expressions to EMF models or to print EMF models to languages expressions respectively.

Figure 4 exemplifies the first step of the process. It partially shows C++ code of Mandelbrot Set that is the input of the model injector and the C++ model of the application in XMI format [30].

### 6.2 C++ Model Refactoring

This refactoring reorganizes and modifies the syntactic elements to improve the design and to adapt own behavior of the C++ language in order to be directly translate to HAXE, such as remove multiple inheritance, remove multiple class constructors.

The refactoring is implemented as a model-to-model transformation whose source and target models are instances of C++ metamodel. In our case study, refactoring is applied to remove the multiple class constructors of Mandelbrot class.

### 6.3 Transformation of C++ models into HAXE models

Model-to-model transformations were implemented in ATL that is a model transformation language in the field of MDE developed on top of the Eclipse platform. ATL is a hybrid language that provides a mix of declarative and imperative constructs. ATL mainly focuses on the model-to-model transformations which can be specified by means of ATL modules. An ATL module is composed of the following elements:

- A header section that defines the names of the transformation module and the variables of the source and target metamodels.
- An optional import section that enables to import some existing ATL libraries.
- A set of helpers that can be used to define variables and functions.

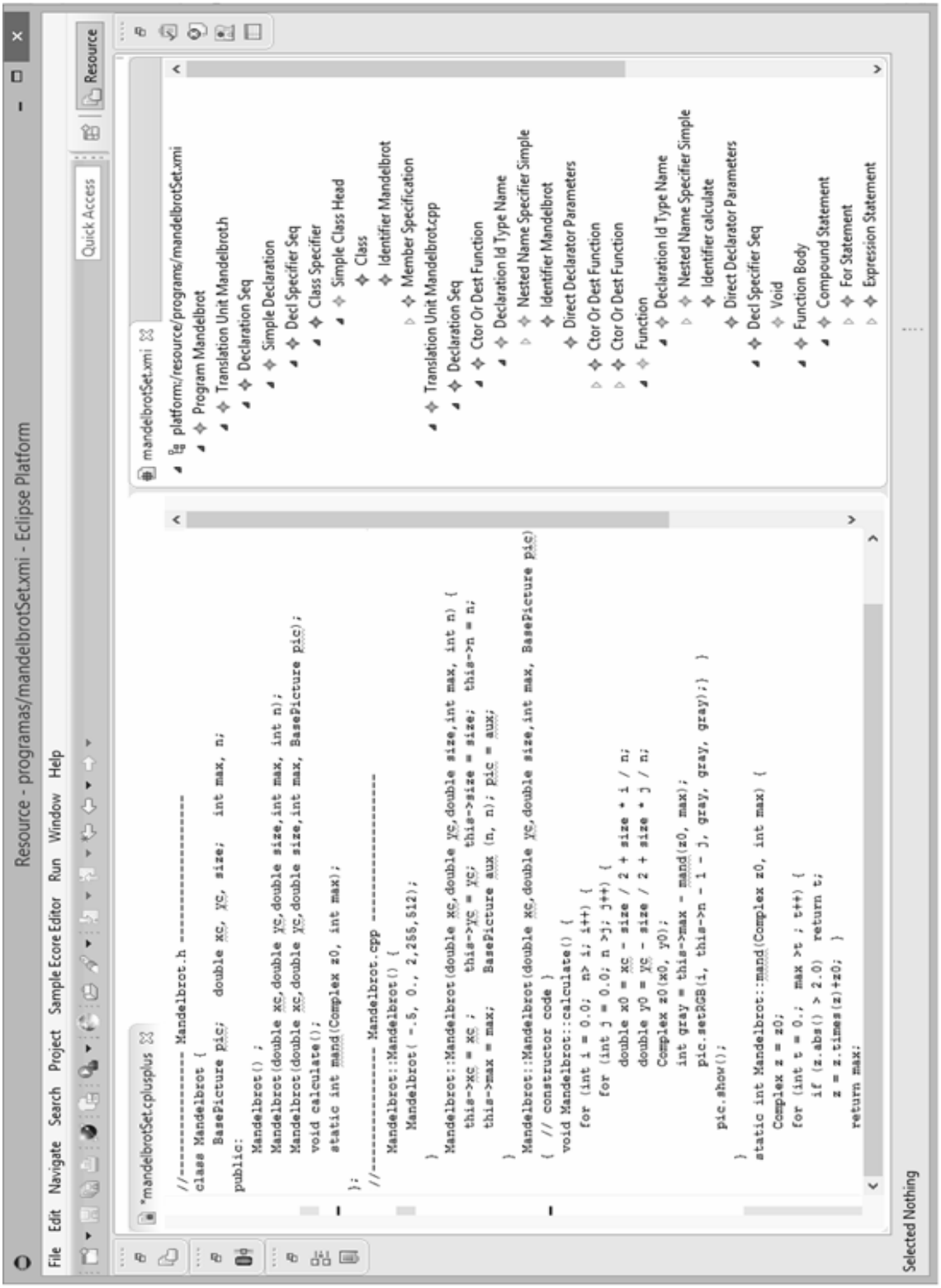



Figure 4.Mandelbrot Class: Code and Model

- A set of rules that defines how source model elements are matched and navigated to create and initialize the elements of the target models.

In our case study, the model-to-model transformation from C++ to HAXE takes as input the model obtained in the previous step and release an HAXE model conforming to the HAXE metamodel. This ATL transformation is partially depicted in Figure 5. The module *CPP2HAXE* that corresponds to the transformation specifies the way to produce HAXE models (target) from C++ models (source). Both source and target models must conform to their respective metamodels. The main rules that carry out the transformation are the followings:

- The rule Program2Model transforms each C++ program into a HAXE model that contains, among others, elements such as *HAXETypes* and *HAXEPackages*.
- The rule TranslationUnit2Module transforms each C++ translation unit into a HAXE module that contains a set of *HAXEDependencyDeclarations*, references a set of elements instances of *HAXETypes*, etc.
- The rule *Classifier2HAXEClass* transforms each C++ classifier, such as class and structure, into a HAXE class whose fields are derived from the classifier members.

All models obtained in this chain of transformations are saved in the interchange format XMI, an OMG standard that combines XML, MOF and UML for integrating tools, repositories, and applications in distributed heterogeneous environments [30].

### 6.4 HAXE Code Generation

From a model HAXE, it is possible to generate a source code in HAXE by using ACCELEO. HAXE allows writing mobile applications that target all major mobile platforms in a straightforward way. The generated code is syntactically correct, although, it does not compile on other platforms without doing changes due to the code refers to proprietary technologies of C++. To run on mobile environments, these technologies can be replaced with OpenFL and HAXEUI (that is an open source, multi-platform application-centric user interface framework designed for HAXE and OpenFL). The code obtained is partially shown in Figure 6.

## 7 DISCUSSION

Our work shows the viability of semi-automatic migration processes based on MDD (MDA in particular). Due the fact that the objective of the migration is not only “compile” an application in a mobile platform but also to create a modified version of the application using quality criteria, the process can not be fully automated. Next, we informally compare the model-driven migration process with brute-force re-development migration.

A crucial limitation of our approach is to require preliminary activities that requires time and cost, for instance we need to define metamodels if they do not exist. It is assumed that using a brute-force redevelopment, developers do not need training to write model transformations, however the programming interfaces of these languages generally restrict the kind of transformation that can be performed. In addition, general-purpose languages do not provide a sufficient level of abstraction to specify them. Changes will be difficult to write and understand and, therefore their maintenance is hard. On the contrary, model driven transformations are expressed in specialized languages for that purpose. Model transformations allow developers to concentrate on conceptual aspects of the relations between models and then to delegate the production of the transformation rules. We can consider that the generation of models by model transformations in ATL, aims to generate

models "Correct-by-Construction" with respect to metamodel specifications. Even with these issues, there is still activities done by hand and the migrated application has to be tested.

A general limitation on both processes is the cost of testing due to the fact these activities in general are handled manually. In the context of model driven approaches there is a need to reduce the cost of testing by defining semiautomatic process based on metamodels. Beyond the previous issues, we consider that mobile developers need frequently adapt software components and applications developed in Java or C/C++. Then, model driven migration processes could be reused and the cost of preliminary activities is recovered.

```
CPP=/CPP2HAXE/CPP.ecore
-- @path HAXE=/CPP2HAXE/HAXE.ecore
module CPP2HAXE;
create OUT: HAXE from IN : CPP;
--  HELPERS
helper context CPP!Program def: get_Types(): Set(HAXE!HAXEType)= ... ;
helper context CPP!Program def: get_Packages(): Set(HAXE!HAXEPackage)= ... ;
...
-- RULES
rule Program2Model {
      from   s : CPP!Program
      to     t : HAXE!HAXEModel (
             name <- s.name,
             elements <- s.get_Types(),
             elements <- s.get_Packages(),
             HAXEModules <- s.get_sourceFiles(),
             mainClass <- s.get_mainClass()
             ...
             )
}
rule TranslationUnit2Module {
      from   s : CPP!TranslationUnit
      to     t : HAXE!HAXEModule (
                    name <- s.name,
                    dependencies <- s.get_Dependencies(),
                    theElements <- s.get_Declarations()
                    ...
             )
}
rule Classifier2HAXEClass {
      from   s : CPP!CPPClassifier
      to     t : HAXE!HAXEClass (
                    name <- s.get_Name()
                    ,HAXEFields <- s.getMemberVariables()
                    ,HAXEFields <- s.getMemberMethods()
                    ,generalization <- s.get_superClass()
                    ,implementation <- s.get_superAbstractSuperClasses()
                    -- ...
             )
}
rule EnumSpecifier2HAXEenum {
      from   s : CPP!CPPClassifier
      to     t : HAXE!HAXEEnum ( ...    )
}
...}
```

Figure 5. *CPP2HAXE* Transformation

```
class Mandelbrot
{
   public static function new_Mandelbrot_9 (xc : Float, yc : Float, size : Float,
                                max : Int, pic : BasePicture)  : Mandelbrot {
      var tmp : Mandelbrot = new Mandelbrot();
      tmp.ctor_Mandelbrot_9(xc, yc, size, max, pic);
      return tmp;
   }
   public static function mand (z0 : Complex, max : Int)  : Int {
      ...
   }
   public function calculate ()  : Void
   {
      {var i : Int = 0;
       while (this.n > i){  {  {
         var j : Int = 0;
         while (this.n > j) {  {
            var x0 : Float = this.xc - this.size / 2 + this.size * i / this.n;
            var y0 : Float = this.yc - this.size / 2 + this.size * j / this.n;
            var z0 : Complex = new Complex(x0, y0);
            var gray : Int = this.max - mand(z0, this.max);
            pic.setRGB(i, this.n - 1 - j, gray, gray, gray);
            }  j++;  }
         }; }  i++;  }  };
      pic.show();
   }
   function new () {
   }
   function ctor_Mandelbrot_9 (xc : Float, yc : Float, size : Float, max :
                                Int, pic : BasePicture)
   {
      if (pic.getHeight() != pic.getWidth())
         throw new IllegalImageSize() ;
      this.xc = xc;
      this.yc = yc;
      this.size = size;
      this.n = pic.getWidth();
      this.max = max;
      this.pic = pic;
   }
   public function getPic()  : BasePicture
   {
      return pic;
   }
   public var pic: BasePicture;
   public var xc: Float;
   public var yc: Float;
   public var size: Float;
   public var max: Int;
   public var n: Int;
}
```

Figure 6. HAXE Code

## 8 CONCLUSIONS

This paper describes an approach for adapting object-oriented software in C/C++ to mobile platforms. A migration process, based on the integration of MDA and the HAXE platform, has been proposed. The main contributions of our approach are the definition of a metamodel for C/C++, the specification of a metamodel transformation between a source C++ metamodel and a target HAXE metamodel, and a generic and extensible migration process for the implementation of cross-platform, multi-device mobile applications from C/C++ code.

The proposal was validated in the open source application platform Eclipse considering that some of its tools and run-time environments are aligned with MDA standards. We believe that our approach provides benefits with respect to processes based only on traditional migration techniques.The migration process can be divided in smaller steps focusing in specific activities, and be automated thanks to the chaining of model transformations. All the involved artifacts can be reused, modified for evolution purposes or extended for other purposes. The metamodel approach enables covering different levels of abstraction and

satisfying several degrees of detail depending on the needs of the migration and is the key for interoperability. All artifacts can be actually represented as models so that there is no information loss during the migration process. Model transformations allow developers to concentrate on the conceptual aspects of the relations between models and delegate the implementation of the transformation.

Our approach has already shown to work on real applications of medium size. We foresee to apply our approach in real industrial projects.

## REFERENCES


[1] ACCELEO.*Obeo.* ACCELEO Generator. http://www.eclipse.org/ACCELEO/, 2016.

[2] ADM. Architecture-driven modernization task force.

http://www.adm.org, 2016.

[3] F. Améndola, L. Favre. Adapting CRM systems for mobile platforms: An MDA perspective. *14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD´13)* (pp. 323-328), Los Alamitos:IEEE Computer Society, 2013.

[4] ATL. Atlas Transformation Language Documentation. http://www.eclipse.org/atl/documentation/, 2015.

[5] J. Bowen, A. Hinze. Supporting mobile application development with model-driven emulation. *Journal of the ECEASST*, *45*, 1–5.

[6] M. Brambilla, J. Cabot, M. Wimmer. *Model-Driven Software Enginneering in Practice*, Synthesis Lectures on Software Engineering. Morgan & Claypool Publishers, 2012.

[7] P. Braun, R. Eckhaus. Experiences on model-driven software development for mobile applications. *Proceedings of Engineering of Computer-Based Systems, IEEE International Conference andWorkshop on the Engineering of computer Base Systems* (pp. 490-493), Los Alamitos: IEEE Computer Society, 2008.

[8] H. Bruneliere, J. Cabot, G. Dupé, F. Madiot. MoDisco: A Model Driven Reverse Engineering Framework. *Information and Software Technology*, *56*(8), 1012–1032, 2014.

[9] N. Cannasse. *HAXE. Too Good to be True?* GameDuell Tech Talk. http://www.techtalk-berlin.de/news/read/nicolas-cannasse-introducing-HAXE/ , 2014.

[10] B. Dasnois. *HAXE 2 Beginner's Guide*. Packt Publishing, 2011.

[11] J. Dehlinger, J. Dixon. Mobile application software engineering: Challenges and research directions. *Proceedings of the Workshop on Mobile Software Engineering* (pp. 29-32). Berlin, Springer-Verlag, 2011.

[12] P. Diaz Bilotto. *Software development for mobile applications through an integration of MDA an HAXE*. (Undergraduate Thesis). Computer Science Department, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina, 2015.

[13] P. Diaz Bilotto, L. Favre. Migrating Java to Mobile Platforms through HAXE: An MDD Approach. Chapter 13. *Modern Software Engineering Methodologies for Mobile and Cloud Environments.* Antonio Miguel Rosado da Cruz, Sara Paiva, eds. (pp.240-268), IGI GLOBAL, 2016.

[14] J. Dunkel, R. Bruns. Model-driven architecture for mobile applications. *Business Information Systems* (LNCS), (vol. 4439, pp. 464-477). Berlin: Springer-Verlag, 2007.

[15] EMF.Eclipse Modeling Framework (EMF). http://www.eclipse.org/modeling/emf/ 2016.

[16] EMF4CPP :What is EMF4CPP? https://code.google.com/archive/p/emf4cpp/ , 2016.

[17] EMFText, www.emftext.org , 2016.

[18] N. Islam, R. Want. Smarthphones: Past, present and future. *Pervasive Computing*, 13(4), 82-92, 2014.

[19] H. K. Kim. Frameworks of process improvement for mobile applications. *Engineering Letters, 16*(4), 550-555, 2008.

[20] D. Kramer, T. Clark, S. Oussena,. MobDSL: A domain specific language for multiple mobile platform deployment. In *Networked Embedded Systems for Enterprise Applications (NESEA), 2010 IEEE International Conference* (pp. 1-7), Los Alamitos: IEEE Press. doi:10.1109/NESEA.2010.5678062, 2010.

[21] MOF. Meta Object Facility (MOF) Core Specification Version 2.4.2, OMG Document Number: formal/2014.

[22] OCL. *OMG Object constraint language (OCL), version 2.4*. http://www.omg.org/spec/OCL/2.4, 2014.

[23] MDA. The Model-Driven Architecture. http://www.omg.org/mda/, 2016.

[24] R. Pérez Castillo, I. García Rodriguez, R. Gómez Cornejo, M. Fernández Ropero, M. Piattini. ANDRIU. A Technique for Migrating Graphical User Interfaces to Android. In *Proceedings of The 25th International Conference on Software Engineering and Knowledge Engineering (SEKE 2013)*(pp. 516-519) Boston: Knowledge Systems Institute, 2013.

[25] QVT. QVT: MOF 2.0 query, view, transformation:Version 1.1. OMG Document Number: formal/2011-01-01. http://www.omg.org/spec/QVT/1.1/SMM, 2012.

[26] D. Steinberg, F. Budinsky, M. Paternostro, E.Merks. EMF: Eclipse Modeling Framework (2nd ed.). Addison-Wesley, 2009.

[27] B. Stroustrup, *The C++ Programming Language.* Addison-Wesley, Third Edition.1997

[28] C. Thompson, D. Schmidt, H. Turner,J. White. Analyzing Mobile Application Software Power Consumption via Model-Driven Engineering. *Proceedings of PECCS, 2011*, 101–113, 2011.

[29] A. I. Wasserman . Software engineering issues for mobile application development. In *Proceedings of the FSE/SDP workshop on Future of software engineering research* (FoSER '10), (pp. 397-400).New York, NY: ACM, 2010.

[30] XMI. XML Metadata Interchange (XMI) Specification. OMG Document Number: formal/2014-04-04. http://www.omg.org/spec/XMI/2.4.2/PDF/, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# NEW EDUCATIONAL MATLAB TOOL TO EXPLAIN TWO DIMENSIONAL FINITE ELEMENT METHOD TO GRADUATE STUDENTS

**Rafet Sisman**[1] **and Abdurrahman Sahin**[2]

[1] Yildiz Technical University
Istanbul
e-mail: rafet@yildiz.edu.tr

[2] Yildiz Technical University
Istanbul
abdsahin@yildiz.edu.tr

**Keywords:** Educational Tool, 2D Finite Element Method, 2D Rectangular Finite Elements, Gauss Quadrature Integration, Lagrangian Shape Functions

**Abstract.** *In this study a simple code is developed in powerful programming software MATLAB. Graduate engineering students are generally expected to cope with Finite Element Method in two and three dimensions during their Finite Element Courses. However textbooks of that courses, sometimes, do not have enough solved examples for students, because in two and three dimensions calculations are difficult to do manually. At this step a computer program which includes infinite number of examples, as user defines them, and their solution steps, is of crucial importance. The produced MATLAB code has these features as it can display stiffness matrix with or without boundary conditions, deformed shape of the body, displacement, strain and stress distributions over the surface. This code also makes it possible to analyze bodies as it is a plane stress or plane strain body. Besides, this MATLAB code can generate a step by step solution report for students to comprehend the Finite Element procedure. Since this code is a simple and educational tool, only one-piece rectangular bodies can be analyzed. The body to be analyzed can be meshed into various number of rectangular elements. These rectangular elements utilize bilinear Lagrangian shape functions, and Gauss Quadrature technique is used to deal with integrations. In this study four Gauss integration points are used as two in X direction and two in Y direction, which yields the exact integration as approximation functions are bilinear.*

# 1 INTRODUCTION

Finite Element Method (FEM) courses are essential for most engineering students, especially mechanical engineering and civil engineering students. FEM courses in graduate level generally include two dimensional problems. However, textbooks of those courses are short of solved problems and solutions are not explanatory, for example only necessary terms of stiffness matrix are determined because of its difficulty to do manually. Those deficiencies become motivation for this study. As an outcome of this study, a simple software is developed via MATLAB [1] to create two dimensional FEM examples. There are a number of important advantages of this software. First of all, this software makes it possible for user to create infinite number of examples with their FEM solutions. FEM solutions of those problems are written to a text file as a solution report which includes all necessary matrices. Also those matrices are written to a Microsoft Excel file for convenience of user. Deformed shape of the system, stress and strain distributions are also among the visual outputs.

Two dimensional FEM problems includes plane stress problems and plane strain problems [2]. The developed software can handle both of these two kinds of problems. Planar problems to be solved with the developed software must include one-part rectangular geometries. Since the motivation of this study emerges from educational necessities, that level of simplicity is adopted.

Meshing algorithm of the developed software is also simple, only rectangular meshing is available. Four node rectangular elements are created during meshing process. To determine element stiffness matrix for those four node rectangular elements numerical integration is used. Using Lagrangian shape functions for rectangular elements with four nodes, and utilizing Gauss Quadrature technique for numerical integrations yields the exact solution when using four Gauss points per element [3].

# 2 TWO DIMENSIONAL FEM FUNDAMENTALS

## 2.1 Constitutive Matrix

In two dimensional FEM, problems can be either plane stress problems or plane strain problems. The difference between these two kinds of problems can be illustrated as shown in Figure 1. If the planar geometry, which is considered in FEM problem, has a small thickness with respect to other two dimensions then it can be said that this problem is a plane stress problem. If that thickness is large enough, which causes zero strain through that direction, with respect to other two dimensions then it can be said that this problem is a plane strain problem [3].

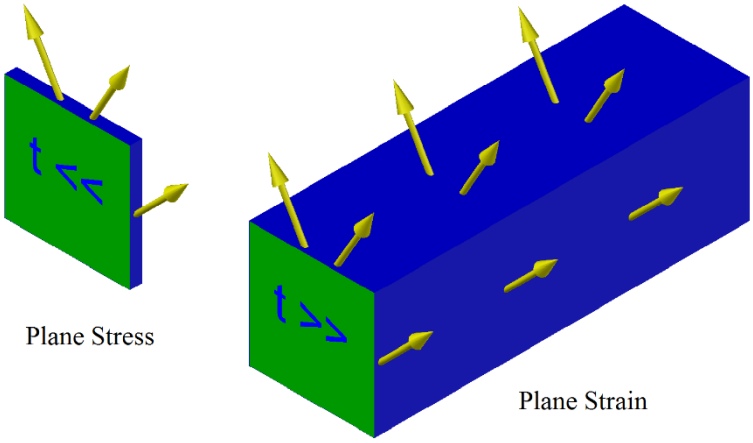



Figure 1 – Plane stress and plane strain geometries

During Finite Element procedures that difference is defined by the constitutive matrix or stress-strain relation matrix, denoted by $[D]$. The constitutive matrix for plane stress problems is given in equation (1), and constitutive matrix for plane strain problem is given in equation (2) where $E$ is Young's modulus and $\upsilon$ is Poisson's ratio [3].

$$[D]=\frac{E}{(1-\upsilon^2)}.\begin{bmatrix}1 & \upsilon & 0\\ \upsilon & 1 & 0\\ 0 & 0 & \frac{1-\upsilon}{2}\end{bmatrix} \tag{1}$$

$$[D]=\frac{E}{(1+\upsilon).(1-2\upsilon)}.\begin{bmatrix}1-\upsilon & \upsilon & 0\\ \upsilon & 1-\upsilon & 0\\ 0 & 0 & \frac{1-2\upsilon}{2}\end{bmatrix} \tag{2}$$

In the developed software, user is expected to define material properties. For two dimensional elastic isotropic materials, Young's modulus and Poisson's ratio are necessary and sufficient parameters to define material. In the software user defines these two parameters, besides user defines problem state as plane stress or plane strain.

## 2.2 Element Stiffness Matrix

As stated before rectangular elements with four nodes are used for the FEM analysis. Those rectangular elements utilize Lagragian shape functions and uses Gauss Quadrature technique for numerical integrations. Lagragian shape functions are given in Figure 2, and Gauss point coordinates and weights are given in Figure 3 [3]. Numerical integration with those shape functions and four Gauss points yields exact integration result.

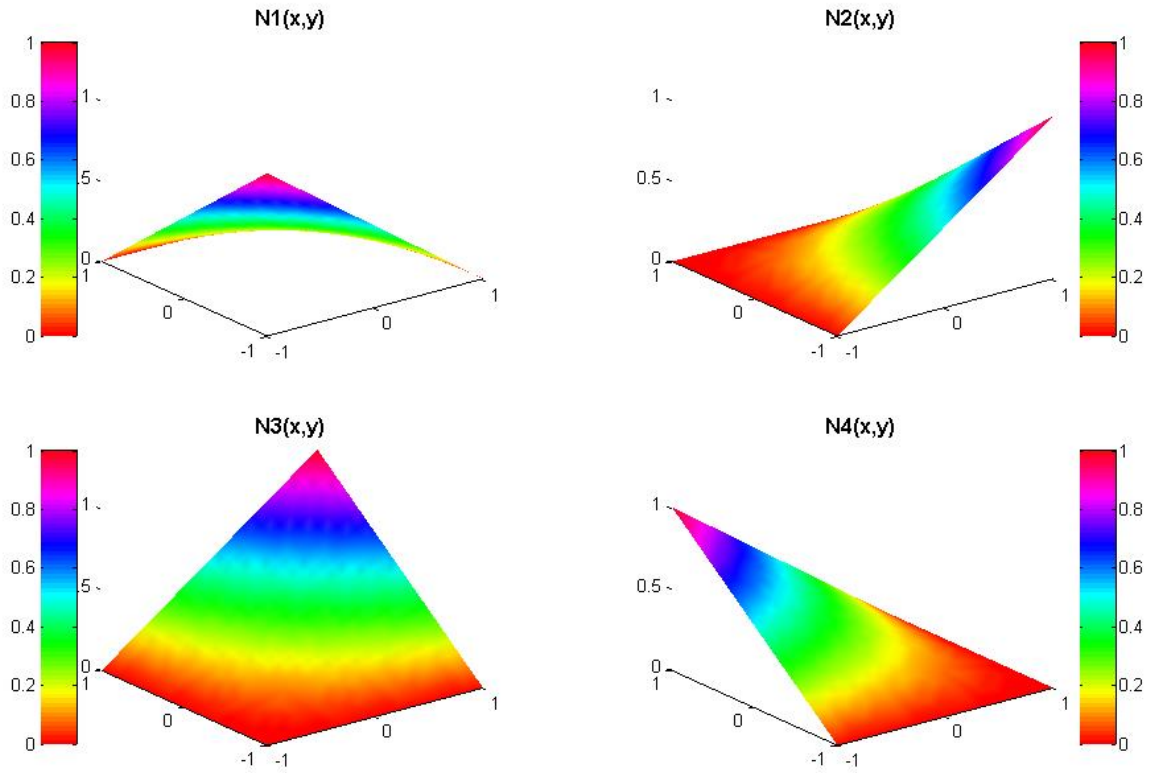



Figure 2 – Lagragian shape functions for rectangular element with 4 nodes

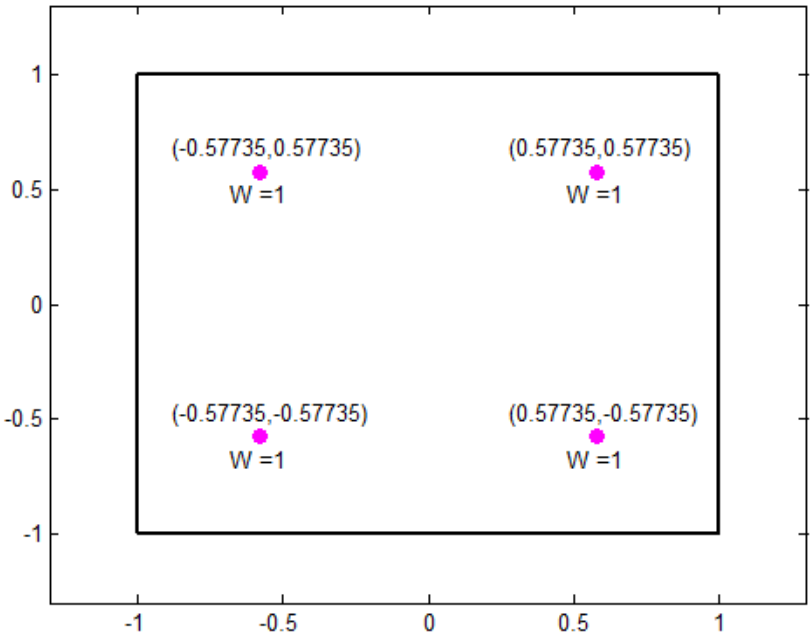



Figure 3 – Gauss quadrature point coordinates and weights

Considering Figure 2 and Figure 3 together, it can be seen that node numbering starts from bottom left node and goes on in counter clockwise direction.

Element stiffness matrix for a rectangular element with four nodes can be determined by the formulae given in equation (3) where $[D]$ is constitutive matrix and $[B]$ is strain-displacement

transformation matrix. Strain-displacement transformation matrix is defined as shown in (4) [3].

$$[k] = t.\iint_{x,y} [B]^T .[D].[B].dxdy \tag{3}$$

$$[B] = \begin{bmatrix} \frac{\partial N_1}{\partial x} & 0 & \frac{\partial N_2}{\partial x} & 0 & \frac{\partial N_3}{\partial x} & 0 & \frac{\partial N_4}{\partial x} & 0 \\ 0 & \frac{\partial N_1}{\partial y} & 0 & \frac{\partial N_2}{\partial y} & 0 & \frac{\partial N_3}{\partial y} & 0 & \frac{\partial N_4}{\partial y} \\ \frac{\partial N_1}{\partial y} & \frac{\partial N_1}{\partial x} & \frac{\partial N_2}{\partial y} & \frac{\partial N_2}{\partial x} & \frac{\partial N_3}{\partial y} & \frac{\partial N_3}{\partial x} & \frac{\partial N_4}{\partial y} & \frac{\partial N_4}{\partial x} \end{bmatrix} \qquad [\varepsilon] = [B].[u] \tag{4}$$

## 2.3 Topology Matrix and Global Stiffness Matrix

To form global stiffness matrix in FEM topology matrices can be used, despite requiring large memory. Since this study introduces an educational tool for FEM, it is necessary to determine topology matrix. Topology matrix relates nodal degrees of freedoms with each other briefly, by utilizing topology matrix global stiffness matrix can be determined easily as shown in equation (5) where $[C]$ is topology matrix and $diag[k]$ is a matrix which have element stiffness matrices on the diagonal [3].

$$[K] = [C].diag[k].[C]^T \tag{5}$$

## 2.4 Boundary Conditions and Solution

In two dimensional FEM problems, nodes have two degrees of freedom as translation in X direction and translation in Y direction. Therefore natural boundary conditions can include restrains in X direction and restrains in Y direction. Also the external forces at nodes can be defined in X direction and Y direction. In the developed software, a user friendly interface makes it possible to define nodal forces and nodal restrains in two directions. For simplicity, each direction represented with a color, X direction is represented with red, Y direction is represented with blue and both directions represented with green. For example, if user defines a nodal force in X direction than it is displayed as a red dot. Those symbols and their meanings are given in Table 1.

| Function of Button | Symbol of Button |
|---|---|
| Nodal force in X direction | |
| Nodal force in Y direction | |
| Nodal force in both directions | |
| Restrain in X direction | |
| Restrain in Y direction | |
| Restrain in both directions | |

Table 1 – Symbols for supports and nodal loads

After applying boundary conditions to global stiffness matrix and load vector, nodal displacements can be determined by using the equation (6) where $\{F\}$ is load vector.

$$\{U\} = [K]^{-1}.\{F\} \tag{6}$$

Determined nodal displacements can be used to calculate element strain and stress values by utilizing equation (4). Using the Lagragian shape functions given in Figure 2 results in linear strains and linear stresses [3]. However, strain and stress values at element centroid are considered during visualization of strains and stresses.

## 3 EDUCATIONAL TOOL DESCRIPTION

### 3.1 Introducing the Program Interface

The graphical user interface of the developed software has three menus as "Analysis", "Output" and "Report" and has eight buttons, the interface and dialog boxes which are used during problem definition are show in Figure 4. A brief explanation of problem definition is given below.

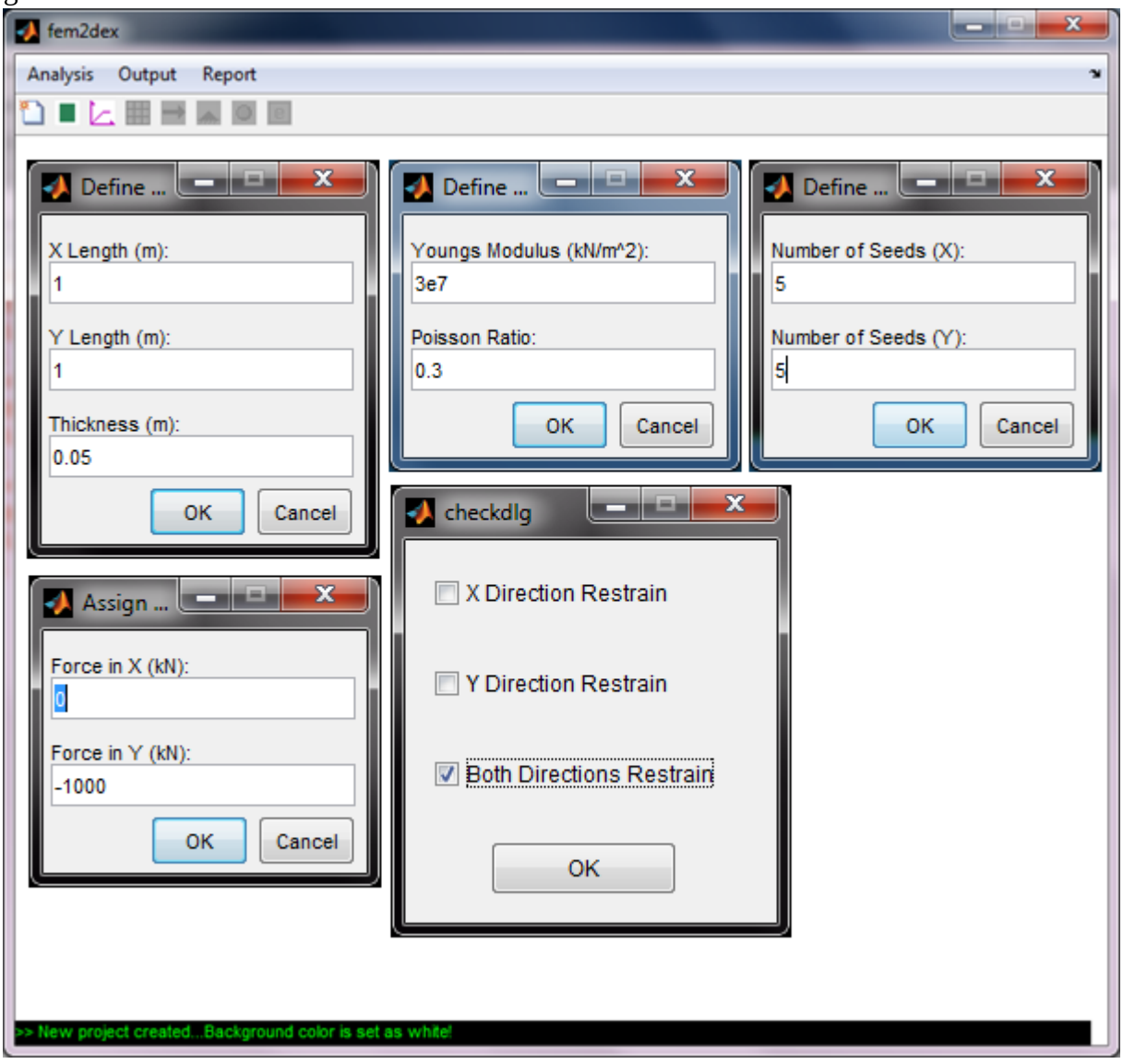



Figure 4 – Graphical user interface of the developed software

By clicking "Define Geometry" button, user can define two lateral dimensions and thickness of the rectangular part. Then, by clicking "Define Material" button, user can define material properties. When geometry definition completed, meshing tool becomes enabled. By using "Define Mesh" button user can input number of seeds in X directions and number of seeds in

Y direction. After defining mesh, user can assign nodal forces and restrains to nodes generated after meshing. The dialog windows for geometry definition, material definition, mesh definition, force assignment and restrain assignment are also shown in Figure 4. Figure 5 shows the functions and icons of the buttons mentioned above.

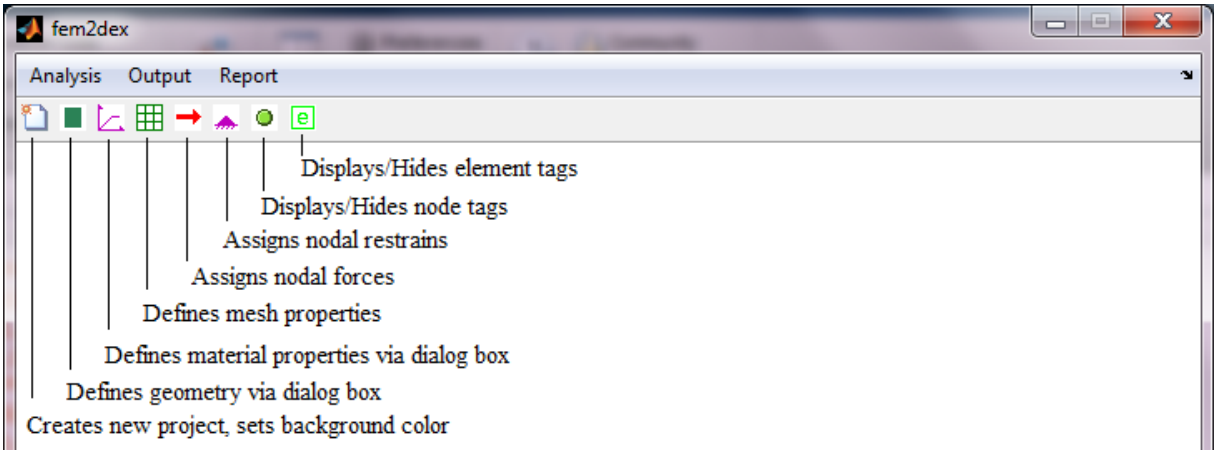



Figure 5 – Button icons and functions

### 3.2 Example Problem Definition and Solution

A simple rectangular plate problem and a simply supported beam problem are solved to illustrate capabilities of the developed software. In the first problem a coarse mesh preferred to explain the software briefly. In the second problem a fine mesh preferred to display visual capabilities of the software. For the plate problem, a rectangular plate with dimensions $1\,\text{m}\times1\,\text{m}\times0.05\,\text{m}$ is meshed into 4 elements as two in X direction and two in Y direction. A material with Young's modulus $E=3.10^7\ \text{kN/m}^2$ and Poisson's ratio $\upsilon=0.3$ is defined. The bottom corner nodes of the rectangular plate are restrained in both X and Y directions and two equal nodal forces with a value of $1000\ \text{kN}$ are assigned to the top corner nodes in X direction. For the beam problem, a beam with dimensions $5\,\text{m}\times0.5\,\text{m}\times0.25\,\text{m}$ is meshed into 40 elements as 10 in X direction and 4 in Y direction. The same material with plate problem is used. The bottom corner nodes of the beam are restrained in both X and Y directions and a $1000\ \text{kN}$ nodal load assigned to the top-mid node of the beam in Y direction. Problem geometries, mesh properties and boundary conditions of these two problems are shown in Figure 6 and Figure 7 respectively.

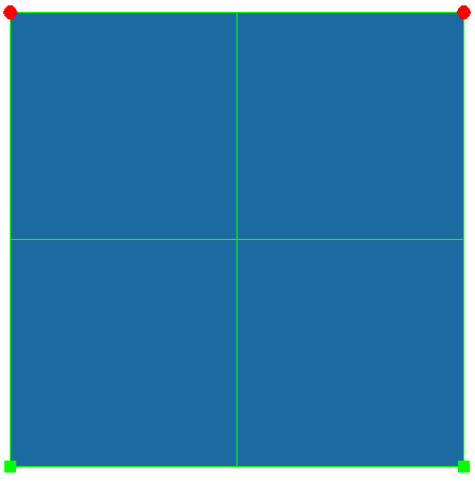


Figure 6 – Plate problem geometry, mesh and boundary conditions

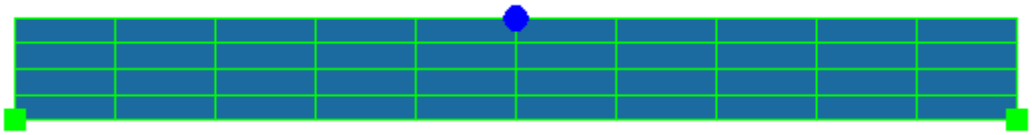


Figure 7 – Beam problem geometry, mesh and boundary conditions

As thicknesses of these two geometries are smaller with respect to other two dimensions of the geometries, both of the problems are solved as plane stress problems. Deformed shape of both geometries are shown in Figure 8 and Figure 9 respectively. Also stress distributions for beam problem are shown in Figure 10. Stress distributions of the plate problem are not given here, since it would not be satisfying because of the coarse mesh defined.

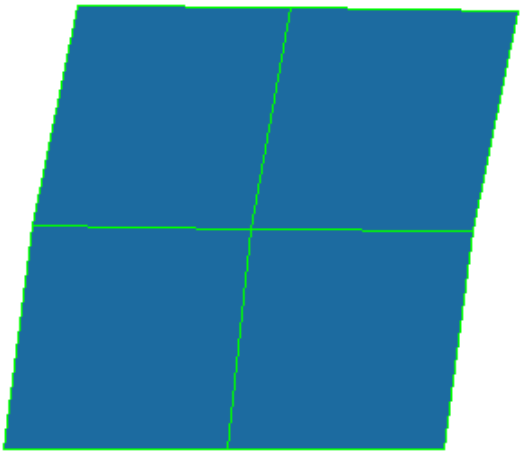


Figure 8 – Deformed shape of the plate

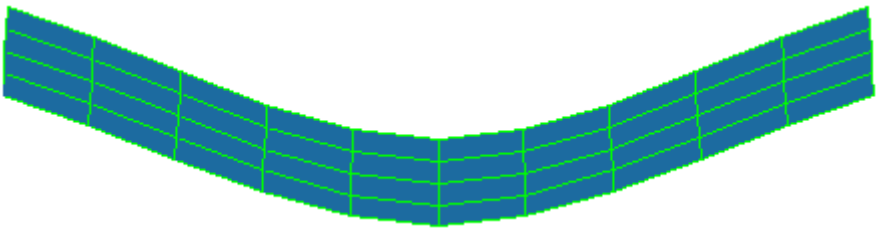


Figure 9 – Deformed shape of the beam

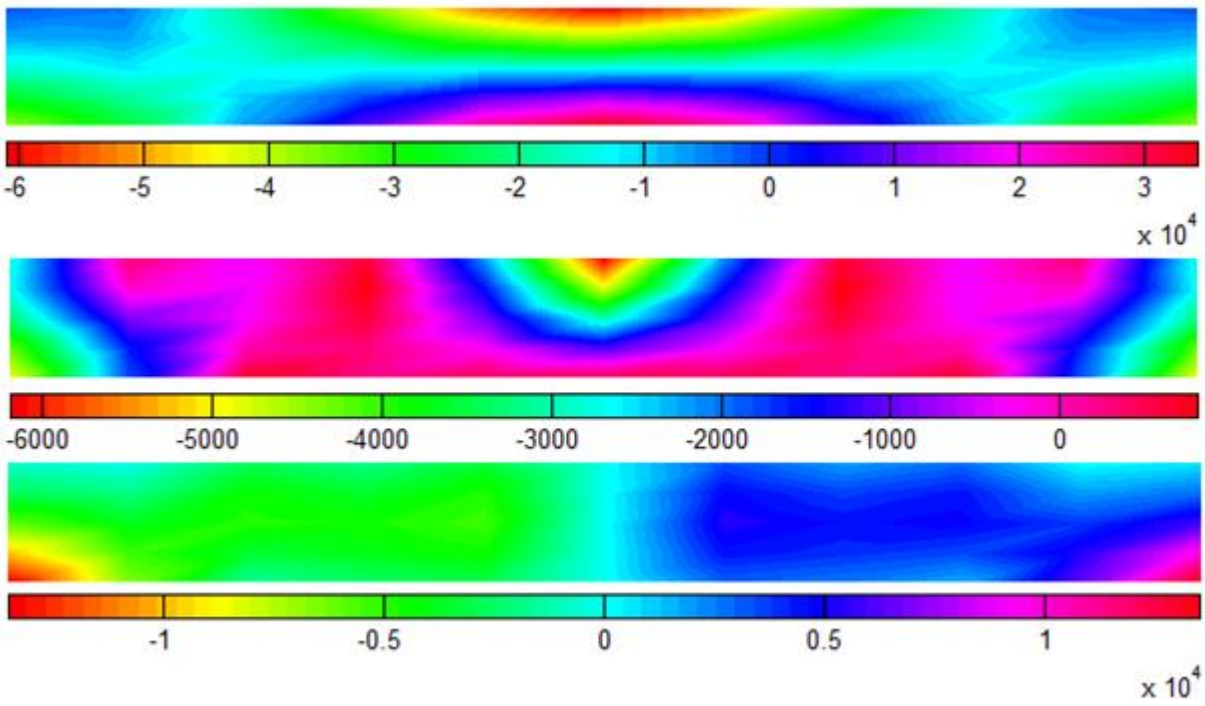



Figure 10 – Stress distributions (kN/m$^2$) for $\sigma_X$, $\sigma_Y$ and $\tau_{XY}$ respectively

### 3.3 Example Problem Outputs

As stated before the developed software can generate a solution report which includes material properties, problem state and all matrices of FEM procedure. Those matrices can also be exported to Microsoft Excel sheets. To illustrate these capabilities, constitutive matrix and element stiffness matrix of plate problem are given in equations (7) and (8).

$$[D] = \begin{bmatrix} 32967032,97 & 9890109,89 & 0,00 \\ 9890109,89 & 32967032,97 & 0,00 \\ 0,00 & 0,00 & 11538461,54 \end{bmatrix} \tag{7}$$

$$[k] = \begin{bmatrix} 741758,24 & 267857,14 & -453296,70 & -20604,40 & -370879,12 & -267857,14 & 82417,58 & 20604,40 \\ 267857,14 & 741758,24 & 20604,40 & 82417,58 & -267857,14 & -370879,12 & -20604,40 & -453296,70 \\ -453296,70 & 20604,40 & 741758,24 & -267857,14 & 82417,58 & -20604,40 & -370879,12 & 267857,14 \\ -20604,40 & 82417,58 & -267857,14 & 741758,24 & 20604,40 & -453296,70 & 267857,14 & -370879,12 \\ -370879,12 & -267857,14 & 82417,58 & 20604,40 & 741758,24 & 267857,14 & -453296,70 & -20604,40 \\ -267857,14 & -370879,12 & -20604,40 & -453296,70 & 267857,14 & 741758,24 & 20604,40 & 82417,58 \\ 82417,58 & -20604,40 & -370879,12 & 267857,14 & -453296,70 & 20604,40 & 741758,24 & -267857,14 \\ 20604,40 & -453296,70 & 267857,14 & -370879,12 & -20604,40 & 82417,58 & -267857,14 & 741758,24 \end{bmatrix} \tag{8}$$

Topology matrix, load vector, displacement vector and global stiffness matrices can also be illustrated via the developed software.

## 4 CONCLUSION

The developed software meets the aim of this study, which is to prepare an educational software to create two dimensional FEM examples and solutions easily. The product of this study can easily be used for FEM courses in graduate level. A large variety of examples from basic examples to complicated examples and their solutions can be prepared with the developed software. Outputs of the software are also very impressive. A solution report in text format can

be generated after completing the analysis. Visual outputs also meet demands, visual capabilities of the software suppress many commercial FEM software.
Results given by the software are compared with a reliable FEM software which is a well-known earthquake engineering simulation tool named as OpenSees [4]. The comparison ended up with a good harmony of the results which verifies the developed software.

## REFERENCES


[1] MATLAB 2014a, The MathWorks, Inc., Natick, Massachusetts, 2014.

[2] Mase, G. T., Mase, G. E., *Continuum Mechanics for Engineers*, CRC Press, 1999.

[3] Desai, C. S., Kundu, T., *Introductory Finite Element Method*, CRC Press, 2001.

[4] McKenna, F., Fenves, G. L., Scott, M. H., and Jeremic, B., *Open System for Earthquake Engineering Simulation (OpenSees)*. Pacific Earthquake Engineering Research Center, University of California, Berkeley, CA, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# CONTRIBUTION OF NUMERICAL THERMAL SIMULATION OF MASONRY TO THE GLOBAL OBJECTIVE OF BUILDINGS ENERGY EFFICIENCY

**Hipólito C. Sousa[1], Rui M. Sousa[1], Luís C. Silva[2] and Gonçalo N. Sousa[3]**

[1] Construct - Faculty of Engineering of the University of Porto
Rua Dr. Roberto Frias, s/n 4200-465 Porto, Portugal
e-mail: hipolito@fe.up.pt, ruysousa@fe.up.pt

[2] Saint-Gobain Weber S.A. - Portugal
Zona Industrial de Taboeira 3800-055 Aveiro, Portugal
e-mail: Luis.Silva@saint-gobain.com

[3] SOPSEC
Rua do Emissor, 110 4400-436 V.N. Gaia, Portugal
e-mail: goncalo.sousa@sopsec.pt

**Keywords:** Energy savings, Lightweight concrete masonry, Thermal performance, Numerical analysis.

**Abstract.** *The challenge of reduce energy consumption in buildings is fundamental for the near future, considering the impact of buildings in the overall energy needs. Buildings enclosure is a constructive element with a high impact on buildings thermal behavior. Masonry is yet the most popular and used solution to produce buildings external walls. Consequently, the improvement of the thermal resistance of those walls is fundamental. The masonry materials industry, including masonry units, mortars, insulation and finishes is very concerned about this problem, which is complex because there are several constraints namely the wall thickness, enough mechanical resistance and the price.*

*The development of new materials and systems must be very well supported for a good level of knowledge of its behavior, including thermal. Experimental testing in laboratory conditions is expensive for several characteristics and numerical simulations based on the 2D or 3D finite element models are implemented.*

*In this paper a numerical sensibility analysis of the thermal performance of a new masonry system is presented. This analysis was applied to a lightweight concrete masonry system tested in laboratory in order to evaluate the influence of the thermal properties of the constituent materials. The numerical simulations are based on 3D FE model. The main results are presented and discussed.*

## 1 INTRODUCTION

The reduction of the energy consumption and efficient use of energy in buildings have become increasingly important in recent decades. The growing energy demands caused high impact on the building sector since it is estimated that this sector is responsible for consuming 40% of the total energy usage in the European Union. In order to reduce this energy consumption the new directive on buildings energy performance [1] established that all buildings constructed after 2018 to 2020 should be nearly zero energy buildings.

To achieve this demanding goal it is necessary to minimize energy consumptions without affecting the thermal comfort requirements. In the case of the construction sector, the performance of the materials/components used in buildings has to be improved, such as enclosure walls with lower thermal transmittance.

Masonry is yet the most popular and used solution used for external/enclosure walls of buildings. Consequently, the reduction of the thermal transmittance of these walls is an important factor to be considered, along with an appropriate combination with the thermal mass, especially in moderate climates and if passive solar design principles are adequately used [2] (building solar orientation and level of shading, surface properties of the glazing elements, amongst others). The masonry materials industry, including masonry units, mortars, insulation and finishes is very concerned about this energy problem, which is complex because there are several constraints to be considered simultaneously, such as the wall thickness, enough mechanical resistance and a competitive price.

Focusing on the reduction of thermal transmittance of masonry, this can be achieved through the use of raw/base materials with lower thermal conductivity and with the use of units whose topology/geometry is conceived to optimize the thermal resistance. For more demanding performances, complementary thermal insulation layers (thermal renderings, external insulation systems) and the filling of the unit voids with thermal insulation materials can also be used, in particular when the increase of wall thickness isn't possible due to practical or economic reasons.

The impact on the energy consumption of building by reducing the thermal transmittance of enclosure walls has been estimated in some scientific studies. For example, according to a numerical study [3] performed on building with four floors and two exposed facades (masonry walls made with vertical hollow units made with thermally enhanced clay) with 33 % of glazing area, annual energy savings near 8% was estimated, when decreasing the equivalent thermal transmittance of the enclosure walls by 43%.

## 2 METHOLOGY

The main objective of this paper is to study the influence of the thermal properties of the constituent materials on the thermal performance of a new single wall system made with lightweight concrete masonry and thermal enhanced rendering. The thermal properties studied were the thermal conductivity of the base materials, namely 3 new types of lightweight concrete to be used in the units and one thermal enhanced mortar to be used in the bed joints. It is mentioned that this wall system was previously tested in laboratory for other properties that are outside the scope of this paper, such as mechanical resistance of masonry, acoustics behavior and watertightness.

Therefore, the work methodology followed in this study was:

- Experimental determination of the thermal conductivity for the 3 types of concrete used in masonry units (latter to be inserted in the numerical model);

- Experimental determination of the compressive strength of units made with the 3 types of concrete (to better evaluate the viability of application of these concretes in masonry products)
- Numerical characterization of the thermal performance of the wall system made with masonry units and joints with lower thermal conductivity materials (main focus of the paper), through a 3D FE model calibrated and validated according to a standard methodology and experimental studies;

## 3 NUMERICAL SIMULATIONS

### 3.1 Construction aspects of the wall system

The single-leaf wall system in analysis is to be used as external enclosure/facade of buildings and was considered in this study according to the constructive/detailing provisions given by the manufacturer/producer:

- Masonry unit/block made from lightweight concrete (concrete with a dry density of 1200 kg/m$^3$ produced with lightweight expanded clay aggregates), vertical perforations (49%) and with external dimensions of 490x250x190mm (length x wide x height), Figure 1;

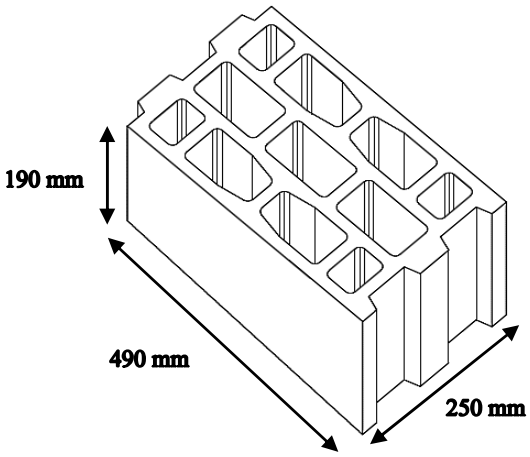



Figure 1: Schematic representation of the masonry unit.

- Vertical joints are interlocking joints (without mortar or dry joints) and bed joints are made with general purpose mortar laid with a central air gap with 65mm wide and 10mm thick (these bed joints are laid on the blind side of the unit, which is made during the production of the unit with 5mm of concrete, thus avoiding mortar penetration into the voids);
- External thermal rendering system applied with 47mm of total thickness, i.e., 40mm of thermal mortar and 7mm of bonding/finishing mortar layers;
- Internal coating/rendering system made with a mineral mortar applied with 20mm of total thickness.

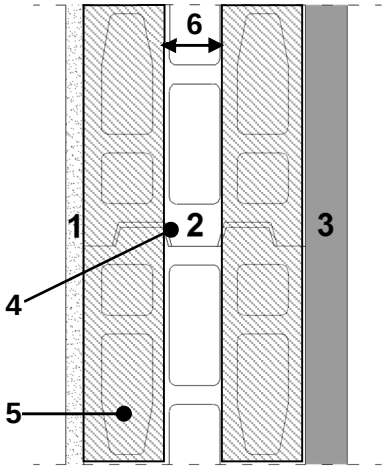



Figure 2: Horizontal cut of the wall system in analysis (1- Internal rendering; 2- Masonry units; 3- External thermal rendering; 4- Shell bedded mortar joints; 5- Dry vertical joints; 6- Central airgap).

### 3.2 Thermal properties of the base materials

Three new types of lightweight concrete (LWC) with different dosages of cement, lightweight expanded clay aggregates, sand and additives were developed by the wall system manufacturer/producer. These new LWC were named in this study according to their dry density ($\rho_{g,dry}$):

- LWC670 (lightweight concrete with 670kg/m$^3$);
- LWC810 (lightweight concrete with 810 kg/m$^3$);
- LWC1100 (lightweight concrete with 1100kg/m$^3$).

Laboratory tests were performed according to EN12664 [4] (Guarded hot plate) to determine the thermal conductivity in dry state ($\lambda_{10,dry,mat}$) of these new LWC. Since these types of LWC are molded into masonry units by an industrial vibration/compression process, 6 test samples were cut off directly from units produced in factory conditions (2 shells/webs for each type of concrete, Figure 3), therefore reproducing the real manufacturing conditions of the LWC masonry units.



Figure 3: Example of test samples removed from the masonry units (shells/webs).

Also 2 types of mortar were considered for the bed joints: a thermally improved lightweight mortar and a general purpose mortar (mortar used in the original wall system).

The properties of the base materials used in the joints and in the renderings were given by technical documents provided by the wall system manufacturer.

In order to consider the base materials affected by the moisture content, a design thermal conductivity ($\lambda_{design,mat}$) was calculated according to standard methodology (EN 1745 [5] and EN ISO 10456 [6]) for an environment with relative humidity and temperature of 23°C and 80%. The results obtained are presented in table 1 and were used in the numerical simulations.

| Layer / element | Base material | $\rho_{g,dry}$ (kg/m$^3$) | $\lambda_{10,dry,mat}$ (W/mK) | $\lambda_{design,mat}$ (W/mK) |
|---|---|---|---|---|
| Internal rendering system | Mineral based mortar | 1500 | 0.610 | 0.775 |
| Masonry wall (unit) | Reference LW Concrete (LWC1200) | 1200 | 0.410 | 0.462 |
| Masonry wall (unit) | New LW Concrete (LWC1100) | 1100 | 0.215 | 0.242 |
| Masonry wall (unit) | New LW Concrete (LWC 810) | 810 | 0.160 | 0.180 |
| Masonry wall (unit) | New LW Concrete (LWC 670) | 670 | 0.122 | 0.132 |
| Masonry wall (bed joints) | Therm. improved lightweight mortar | 1500 | 0.610 | 0.775 |
| Masonry wall (bed joints) | General purpose mortar | 1800 | 0.930 | 1.182 |
| External thermal rendering system | Thermal mortar | 150 | 0.042 | 0.053 |
| | Bonding mortar | 1250 | 0.450 | 0.572 |
| | Finishing mortar | 1750 | 0.820 | 1.042 |

Table 1: Thermal properties of the base materials

For the experimental results obtained its highlighted that the dry thermal conductivity obtained for the new developed LW concretes (LWC1100, LWC810 and LWC670) are lower than the tabulated values given in EN 1745 [5].

### 3.3 Mechanical Strength of units

Laboratory tests were performed according to EN772-1 [7] to determine the compressive strength (perpendicular to the bed joints) of the unit made with the new developed concretes (Figure 4). Only 9 blocks were used (3 units per type of concrete) due to the availability of test samples. The final results obtained are succinctly presented on Table 2.

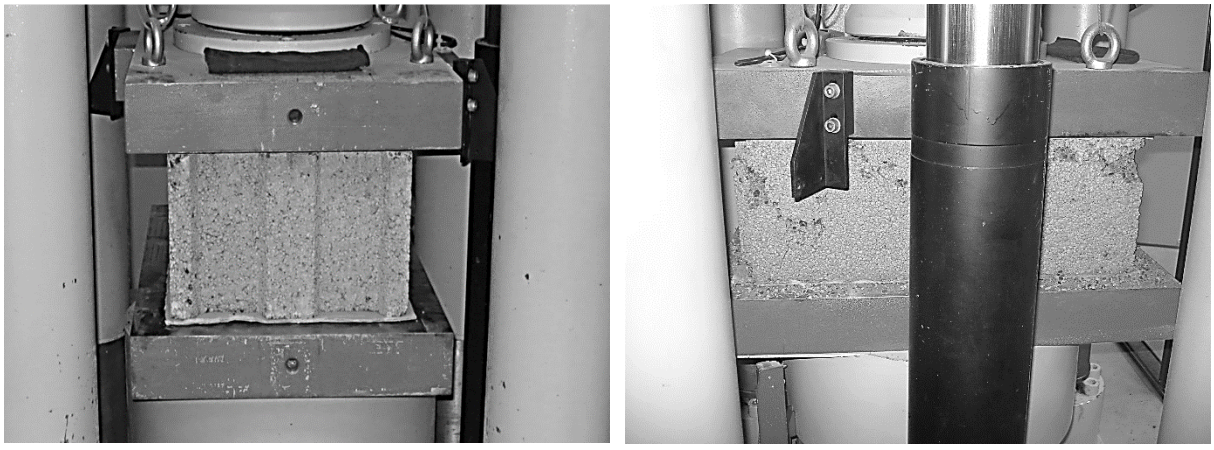


Figure 4: Example of compressive strength tests performed on the masonry units (before and after testing).

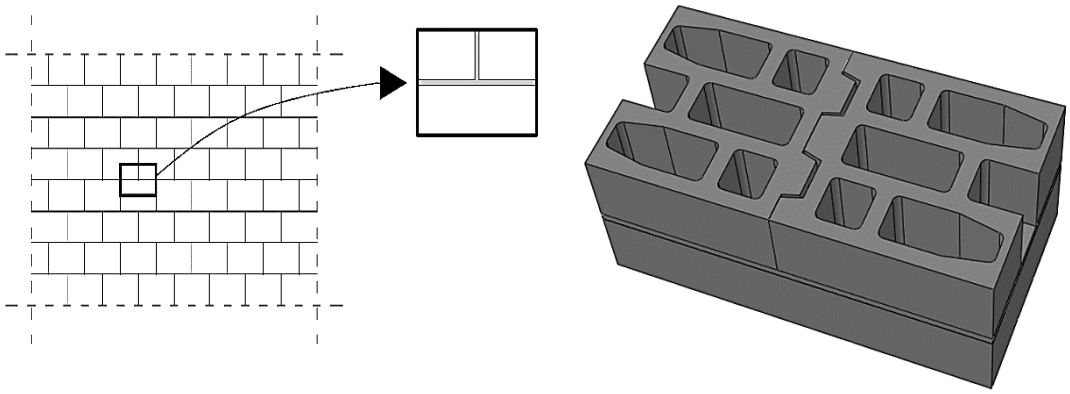
| Type of concrete used in the units/blocks | Average compression strength ($N/mm^2$) | C.V. (%) |
|---|---|---|
| LWC670 | 1.4 | 16.6 |
| LWC810 | 2.3 | 16.3 |
| LWC1100 | 3.1 | 5.8 |

Table 2: Experimental results obtained for the compressive strength of the units

Considering the results obtained, the units made with LWC 810 and LWC1100 concrete has showed the highest compressive strengths, demonstrating viability to have enough mechanical strength to be used as masonry material in real applications (e.g. near the reference limit of 2,5 $N/mm^2$ defined in a French standard for LWC units to be used in non-structural walls [8]). The unit made with LWC 670 has shown the lowest compressive strength and must be improved.

### 3.4 Main aspects of the numerical model

A micromodel based on FEM was used to simulate the thermal behavior of a wall part, detailing the geometry of the blocks, joints (Figure 5) and rendering/coating layers.



a) b)

Figure 5: Geometrical model used in thermal simulations (a) Considered part of the masonry wall; b) Detailed 3D modelling of the masonry wall part).

The simulations were performed by considering a stationary thermal/heat flow obtained from a temperature gradient of 20ºC, the internal and external surface resistances of the wall (Rsi and Rse), the equivalent thermal transmittance of the voids (calculated according to EN ISO 6946[9]), the thermal transmittance of the base materials (see Table 1) and the existence of adiabatic boundaries on the considered part of the wall.

The model accuracy was evaluated and validated according to standards EN 1745[5] e EN ISO 10211[10], i.e., calculation deviations/errors lower than the maximum allowed (1%) were obtained for heat flux in the simulation of standard reference cases of masonry walls. Moreover, this numerical model was previously used and validated in other scientific studies against experimental results obtained from standard laboratory tests (hot box method using heat flow meters and large wall samples). In these studies the numerical and experimental results obtained exhibited differences near ± 4% [11, 12], which are considered acceptable for masonry.

### 3.5 Presentation and discussion of results

As mentioned before, the thermal performance of the wall system was analyzed with the following variations in the masonry base materials:

- Variation on the type of concrete originally used in masonry units (reference situation), i.e., using units made with the developed concretes (LWC670, LWC810 and LWC1100);
- Variation on the type of mortar originally used in bed joints (reference situation), i.e., using bed joints made with thermally improved lightweight mortar.

This performance was evaluated through the thermal resistance of the masonry in design conditions without renderings ($R_{design,mas}$) and through the thermal transmittance of the wall system in design conditions ($U_{design}$). Its highlighted that these design conditions were established by considering that all base materials were affected by moisture content (see 3.2) and had no ageing effects nor other environmental attacks (initial conditions of use were considered). The obtained results are presented in Table 3.

| Type of concrete used in the units/blocks | $R_{design, mas}$ Design thermal resistance of masonry without renderings ($m^2K/W$) | | $U_{design}$ Design thermal transmittance of the wall system ($W/m^2K$) | |
|---|---|---|---|---|
| | Bed joint mortar | | Bed joint mortar | |
| | Lightweight improved | General purpose (reference) | Lightweight improved | General purpose (reference) |
| LWC670 | 1.29 | 1.21 | 0.44 | 0.46 |
| LWC810 | 1.12 | 1.06 | 0.48 | 0.50 |
| LWC1100 | 0.98 | 0.94 | 0.52 | 0.53 |
| LWC1200 (reference) | 0.73 | 0.70 | 0.60 | 0.60 |

Table 3: Results obtained in numerical simulations

An example of the thermal/heat flow distribution obtained for masonry walls is presented in Figure 6 and the thermal transmittance in design conditions ($U_{design}$) of the wall system is presented in Figure 7 according to the type of concrete and type of mortar joints.

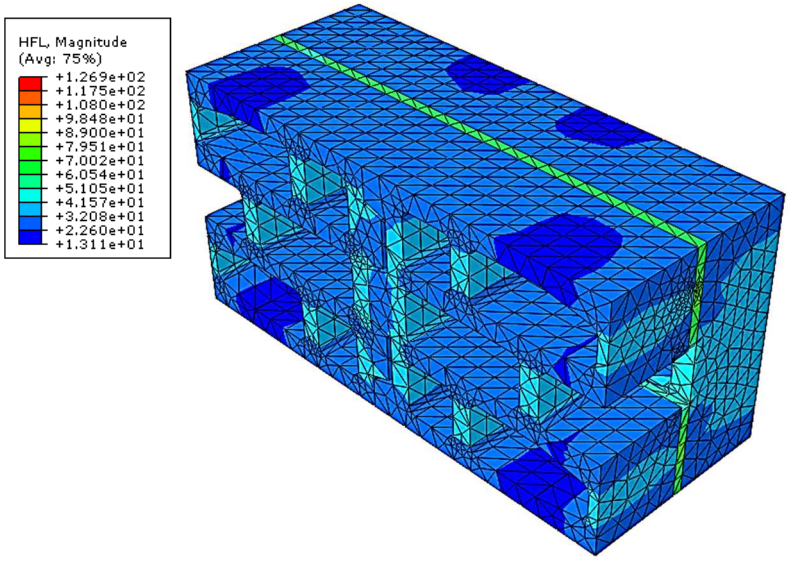



Figure 6: Example of the heat flow in the masonry wall (without renderings)

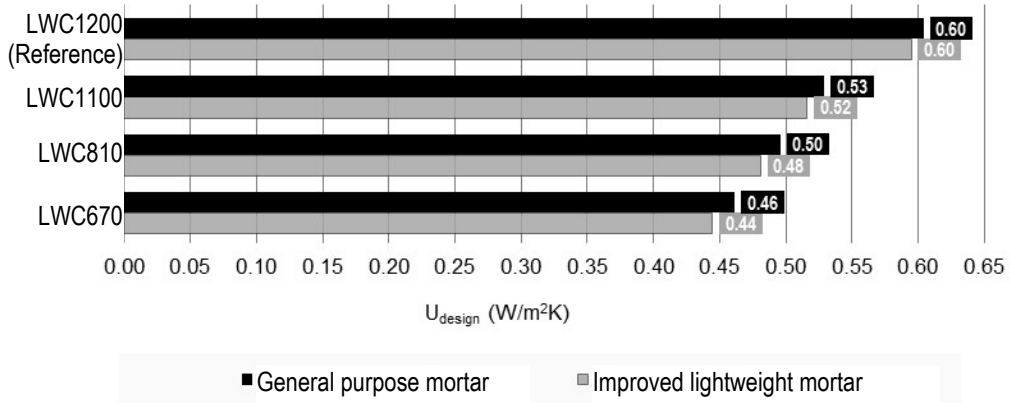



Figure 7: Design thermal transmittance obtained for the wall system made with different base materials

From the obtained results the following aspects are highlighted:

- The units made with the new lightweight concretes (LWC670, LWC810 and LWC1100) improved the thermal resistance of masonry (increase of 34 to 77%) and the thermal transmittance of the wall systems (decrease of 12% a 27%), although in last case with lower impact due to the use of external rendering system with high thermal insulation capability; Moreover, this improvement is higher when using both mortar joints and units with the highest thermal resistances;
- The use of bed joints made with thermally improved mortar only had slight improvement on the masonry thermal resistance (increase of 4 to 7%) or on the thermal transmittance of the wall systems (reduction of almost 0 to 4%), since there is a low quantity of mortar applied on the masonry joints (use of discontinuous and dry joints). Nevertheless, this slight improvement is higher when using both mortar joints and units with the highest thermal resistances;

## 4 GENERAL CONCLUSIONS

The experimental results revealed the development of new lightweight concretes with a lower thermal conductivity and enough mechanical strength to be used in masonry applications, in particular for the LWC810 and LWC1100 concretes. However, the LWC670 concrete should be improved given the lower mechanical strength obtained.

The numerical simulations demonstrated that the use of these concretes on the units, together with the use of thermal enhanced mortar in the bed joints, can significantly reduce the thermal transmittance of the wall system.

Therefore, the use of these new lightweight concretes in masonry applications, especially in the case of LWC810 and LWC1100, have a high potential to produce units and walls systems with enough mechanical strength and lower thermal transmittance, the latter contributing for the reduction of energy consumption of buildings.

## 5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge Saint-Gobain Weber S.A. - Portugal for the funding and support provided in the realization of this study.

## REFERENCES

[1] EBPD 2010 - Directive 2010/31/EU of the European Parliament and of the Council of 19th May 2010 on Energy Performance of Buildings (recast).

[2] D. Alterman, A.W. Page, T. Moffiet, B. Moghtaderi, A measure for the dynamic thermal performance of walling systems incorporating the combined effect of thermal mass and thermal resistance. *9th International Masonry Conference (9th IMC)*, Guimarães, Portugal, July 7-9, 2014.

[3] M. Morales, P. Muñoz, M. Juárez, M. Mendívil, L. Muñoz, Energy efficiency in buildings: Study of single-leaf walls made with clay bricks. *Journal of Energy Engineering*, 142 (1), 2016.

[4] CEN, EN-12664 - Thermal performance of building materials and products: Determination of thermal resistance by means of guarded hot plate and heat flow meter methods - Dry and moist products of medium and low thermal resistance. Brussels, 2001.

[5] CEN, EN 1745 - Masonry and masonry products: Methods for determining thermal properties. Brussels, 2012.

[6] CEN, EN ISO 10456 - Building materials and products: Hygrothermal properties - Tabulated design values and procedures for determining declared and design thermal values. Brussels, 2007.

[7] VEN, EN772-1 - Methods of test for masonry units: Part 1 - Determination of compressive strength. Brussels, 2011.

[8] CSTB, NF P14 304 - Agglomérés - Blocs en béton de granulats légers pour murs et cloisons. Paris, 1983.

[9] CEN, EN ISO 6946 - Building components and building elements: Thermal resistance and thermal transmittance - Calculation method. Brussels, 2007.

[10] CEN, EN ISO 10211. Thermal bridges in building construction; Heat flows and surface temperatures; Detailed calculations. Brussels, 2007.

[11] H. Sousa, R. Sousa, Sensibility analysis of the thermal resistance of masonry through numerical simulations of laboratory tests. *9th International Masonry Conference Guimarães (9th IMC)*, Guimarães, Portugal, July 7-9, 2014.

[12] N. Ramos, V.P. Freitas, P. Pereira, H. Sousa, R. Sousa, U-value determination of masonry walls with complex shape units and thermal rendering. *10th Nordic Symposium on Building Physics (10th NSB)*, Lund City Centre, Sweden, June 15-19, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# TOWARDS MULTIPHYSICS SIMULATION OF DEEP PENETRATION LASER WELDING USING SMOOTHED PARTICLE HYDRODYNAMICS

**Haoyue Hu[1], Peter Eberhard[1], Florian Fetzer[2], and Peter Berger[2]**

[1] Institute of Engineering and Computational Mechanics
University of Stuttgart
Pfaffenwaldring 9, 70569 Stuttgart, Germany
{haoyue.hu, peter.eberhard}@itm.uni-stuttgart.de

[2] Institut für Strahlwerkzeuge
University of Stuttgart
Pfaffenwaldring 43, 70569 Stuttgart, Germany
{florian.fetzer, peter.berger}@ifsw.uni-stuttgart.de

**Keywords:** Laser Welding, SPH, Phase Transition, Recoil Pressure, Ray Tracing, Co-Simulation.

**Abstract.** *Multiphysics simulations of deep penetration laser welding are performed with the meshless Lagrangian Smoothed Particle Hydrodynamics (SPH) method. Compared to mesh-based methods, SPH has advantages in handling phase transitions, free-surface melt flow, and fluid-structure interaction. Based on previous work on simulating conduction mode laser welding using SPH, the numerical model is extended to include further physical effects such as evaporation and exertion of recoil pressure on the melt due to evaporation. Particular emphasis is placed on modeling the energy input through the laser beam. A co-simulation approach is developed by coupling an SPH code with a ray tracer that tracks the propagation of the laser beam in the keyhole in order to achieve spatial distributions of energy transferred to the melt layer. A surface detection and reconstruction algorithm is implemented to exchange current surface data. Simulation results of spot welding and seam welding are shown using this co-simulation approach. The developed model serves as a basis to investigate the influence and sensitivity of process parameters on the weld and to better understand transient effects around the keyhole leading to weld imperfections.*

## 1 INTRODUCTION

Laser welding is widely applied in industry nowadays. It offers significant advantages over conventional welding techniques, such as a small heat affected zone resulting in low mechanical and thermal distortions in the workpiece, high heating and cooling rates, and the ability to join dissimilar materials. Depending on the absorbed power density from the laser beam, laser welding is classified in conduction mode and deep penetration welding. During conduction mode welding, evaporation plays no significant role and the energy is transferred to a more or less flat surface. In contrast, during deep penetration welding, a part of the material evaporates and a vapor-filled capillary surrounded by molten material is formed by the vapor pressure exerted on the melt, a so-called keyhole. Multiple reflections of the laser beam on the walls of the keyhole lead to a very high absorbed intensity by the molten material. As the keyhole follows the movements of the laser beam along the weld joint, the melt flows around the keyhole and solidifies, producing a deep and narrow weld seam. Deep penetration welding is a preferred welding technique, distinguished by high efficiency and fast welding speeds. However, the process easily becomes unstable and weld imperfections like spatters, pores, cracks, and humping, may then occur.

To simulate the process of deep penetration welding, the novel contribution of this paper is that we extend a developed SPH model for conduction mode welding [1, 2] to include evaporation and exertion of recoil pressure on the melt due to evaporated material. Concerning the laser-material interaction, knowledge about the spatial intensity distribution as energy input is essential. We propose a co-simulation approach with a ray tracer that tracks the propagation of the laser beam in the keyhole to obtain spatial intensity distributions. Further steps during the co-simulation are the transformation of the intensities to heat sources, and the surface detection and reconstruction to send an updated surface geometry to the ray tracer.

## 2 MODELING WITH SPH

The SPH method was developed by Gingold and Monaghan [3] and Lucy [4]. As a meshless Lagrangian method, it has been successfully applied to problems with large deformations, fluid flow with free surfaces, multiphase flow, etc. [5]. In SPH a continuum is represented by a set of particles acting as discretization points. Material properties and field variables like velocity and acceleration are associated to each particle such that the overall state of the system is characterized properly. The particles interact with each other within a smoothing length $h$ of a kernel function $W$. The evaluation of field variables and their derivatives for a particle is approximated by sums over all neighboring particles. As a general discretization method, all systems that are described by partial differential equations can be simplified to a set of ordinary differential equations.

The modeling approach for laser welding using the SPH method is based on previous work of simulating the conduction mode laser welding process [1, 2]. Therefore, only a short summary is given in the following. The solid material is modeled with solid SPH particles extended to a thermomechanically coupled formulation. For the liquid melt flow, the standard formulation of weakly compressible SPH proposed by Monaghan [5] is used. Additionally, a temperature-dependent surface tension model for thermocapillary flow as shown in [6] is implemented. Heat transfer through conduction is modeled based on [7]. Temperature-dependent material properties are taken into account, as an example, different heat conductivities and heat capacities for the solid and liquid phases are included in the model. For both solid and liquid phases, the conservation laws of mass, momentum, and energy are discretized and used as basis for the

interaction between solid particles, fluid particles, and also for the fluid-structure interaction. The developed model is implemented into the software package Pasimodo [8, 9], which is used for the welding simulations.

### 2.1 Phase transitions

During deep penetration welding, the occurring phase transitions are melting, solidification, evaporation, and condensation. For phase transitions from solid to liquid or vice versa, the latent heat of fusion $H_f$ has to be considered. If the absorbed amount of heat exceeds the latent heat of fusion, the melting has completed. The particle type is then dynamically changed from a solid to a fluid particle during the simulation. On the contrary, during solidification heat is released from the material. If the released amount of heat exceeds the latent heat of fusion, the solidification has completed and the particle type is then changed from a fluid to a solid particle.

Analogously, for phase transitions from liquid to vapor, the latent heat of evaporation $H_v$ has to be taken into account. Currently, the vapor phase is not simulated, instead, the particles are deleted during the simulation after the evaporation has completed. Thus, condensation is neglected.

The specific enthalpy $h$ is used to characterize the thermal energy stored in the system. At constant pressure, the increase in enthalpy is equal to the amount of heat added to the system. Therefore, a change in enthalpy may be converted to a change in temperature with the heat capacity as proportionality factor. The specific enthalpy over temperature is plotted in Fig. 1, in which constant heat capacities $c_s$ and $c_l$ are used for the solid and liquid phase, and the latent heat of fusion $H_f$ and evaporation $H_v$ are considered in a transition range between two phases.

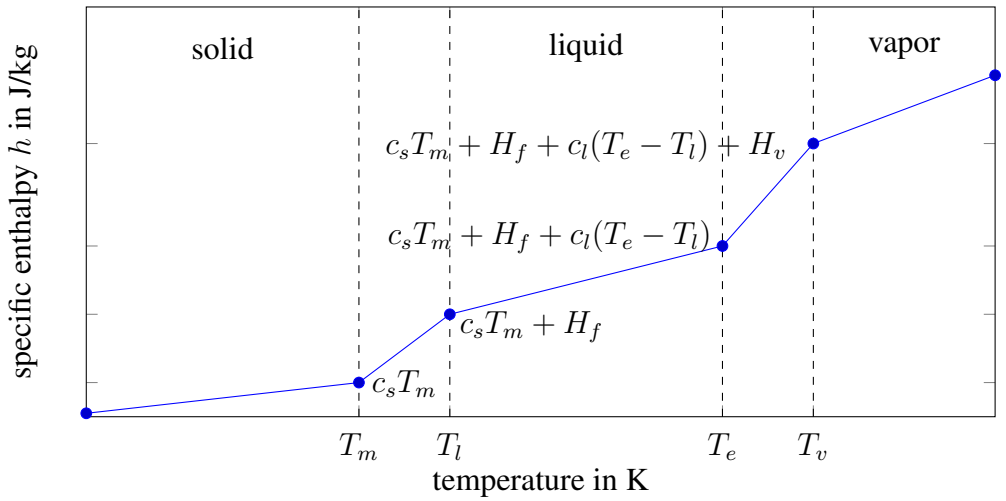



Figure 1: Specific enthalpy over temperature simplified as integrated values $\int c\,\mathrm{d}T$, i.e. the pressure changes are neglected here

### 2.2 Calculation of recoil pressure

Different models exist in literature for the calculation of recoil pressure acting on the surface of an evaporating material due to high-intensity laser irradiation [10, 11]. Here the physical model proposed in [12] is applied, since the predicted recoil pressures using this model are in reasonable agreement with experimental results.

The laser power absorbed by the material surface and spent in evaporation is

$$Q_s = Q_0\alpha(1-r) - Q_{\text{heat}} \ , \tag{1}$$

where $Q_0$ is the incident power of the laser beam, $\alpha$ the transmission coefficient in the laser-induced plasma plume, $r$ the reflection coefficient at the material surface, and $Q_{\text{heat}}$ the energy rate in W spent in heating and melting of the irradiated material. The recoil pressure is then evaluated as

$$p_s = \frac{Q_s}{AH_v}\sqrt{\frac{\pi k_B T_s}{2m_v}} \ , \tag{2}$$

in which $A$ is the area of the laser spot at the surface, $H_v$ is the latent heat of evaporation, $k_B$ the Boltzmann constant, $T_s$ the surface temperature, and $m_v$ the mass of a vapor molecule. The Clausius-Clapeyron equation [13, 14]

$$\frac{p_a + p_s}{p_0} = \exp\left(\frac{m_v H_v}{k_B}\left(\frac{1}{T_v} - \frac{1}{T_s}\right)\right) \tag{3}$$

relates the pressure at the surface to the temperature at the surface. In Eq. (3) $p_a$ is the environmental pressure, $p_0$ the atmospheric pressure, and $T_v$ the evaporation temperature.

In [12] the value of $Q_{\text{heat}}$ is obtained from empirical relations based on experimental results and $Q_s$ is calculated for the whole surface. In the developed model, a local approach is applied instead and $Q_s$ is approximated for each particle using its specific enthalpy

$$Q_s = \{h - [c_s T_m + H_f + c_l(T_e - T_l)]\}\frac{m}{\Delta t} \ , \tag{4}$$

where $m$ is the mass of the particle, and $\Delta t$ the time increment. To obtain both unknowns $T_s$ and $p_s$, first the nonlinear equation

$$f(T_s) := \frac{p_a}{p_0} + \frac{Q_s}{p_0 A H_v}\sqrt{\frac{\pi k_B T_s}{2m_v}} - \exp\left(\frac{m_v H_v}{k_B}\left(\frac{1}{T_v} - \frac{1}{T_s}\right)\right) = 0 \tag{5}$$

is solved with the Newton-Raphson method, where the evaporation temperature is chosen as starting value $T_s^0 = T_v$. The iteration step

$$T_s^{n+1} = T_s^n - \frac{f(T_s)}{f'(T_s)} \tag{6}$$

with the derivative

$$f'(T_s) = \frac{\mathrm{d}f}{\mathrm{d}T_s} = \frac{Q_s}{2p_0 A H_v}\sqrt{\frac{\pi k_B}{2m_v T_s}} - \frac{m_v H_v}{k_B T_s^2}\exp\left(\frac{m_v H_v}{k_B}\left(\frac{1}{T_v} - \frac{1}{T_s}\right)\right) \tag{7}$$

is repeated until a relative or absolute convergence criterion is fulfilled, or if the maximum number of iterations is reached. Afterwards $p_s$ is evaluated using Eq. (2) and the recoil pressure is added to all surrounding particles.

## 3 RAY TRACING ALGORITHM

A ray tracing algorithm introduced in [15] is applied to calculate the local absorbed intensity from the laser beam on a discretized surface. The idea is to use geometrical optics to track the propagation of independent light rays. The geometry of the surface is meshed with triangles and specified in STL format. During the calculation, a large number of rays are generated randomly based on a Monte Carlo method. The overall distribution of the rays corresponds to the intensity profile of the laser beam. When a ray hits a triangle, the local absorbed intensity is evaluated from the absorbed power of the ray divided by the area of that triangle. The absorption is calculated using the Fresnel equations [16]. Then the reflected ray is traced and the calculation of the local absorbed intensity is repeated. After multiple reflections are performed, the total absorbed intensity at each triangle is obtained from the sum of each absorbed intensity from individual rays at the triangle. If no intersection between a ray and the triangles is detected, the ray is deleted during the simulation.

The existing algorithm from [15] is extended to include further physical effects. Temperature-dependent optical material parameters are specified in a data table, and the values are interpolated in between. Furthermore, diffuse reflection according to Lambert's emission law may be chosen in addition to Fresnel absorption.

## 4 CO-SIMULATION APPROACH

To model the laser-material interaction, a co-simulation approach is developed with a server-client architecture, where Pasimodo [8, 9] acts as the server, and the ray tracer as a client. The data exchange is carried out via a TCP/IP network connection.

The procedure of the co-simulation is shown in Fig. 2. The initial surface as input for the ray tracer is taken as a circular plane. The calculated absorbed intensities for each triangle are transferred to Pasimodo as initial energy input through the laser beam. In Pasimodo heat source particles are added at the center of gravity positions of each triangle. To reduce the computational effort, only triangles with an absorbed intensity above a threshold value set as 100 W/m$^2$ are considered. In Pasimodo the welding simulation using SPH is performed with these dynamically added heat sources. After a user-defined time interval of $\Delta t$, the updated geometry of the free-surface melt flow and the temperature of the surface particles are sent to the ray tracer. The co-simulation loop is repeatedly executed until the welding simulation has finished. The additional steps performed during the co-simulation are explained in the following.

### 4.1 Transformation of intensities to heat sources

For the SPH simulation in Pasimodo, the absorbed intensities $I_s$ are transformed to heat source particles at the center of gravity positions of each triangle. The radius of the heat source particle is set to

$$r_s = \sqrt{\frac{A_s}{\pi}} = \sqrt{\frac{\frac{1}{2}\|\mathbf{a} \times \mathbf{b}\|}{\pi}} \,, \tag{8}$$

where $A_s$ is the area of the triangle which may be obtained by half the norm of the cross product of two edge vectors $\mathbf{a}$, $\mathbf{b}$ originating from one vertex, see Fig. 3. All neighboring particles with a distance less than $2r_s$ interact with the heat source particle. To ensure that the total power added to all neighboring SPH particles is equal to the absorbed power from the laser beam at the triangle, a correction factor is determined similarly to [17]. The correction factor is evaluated

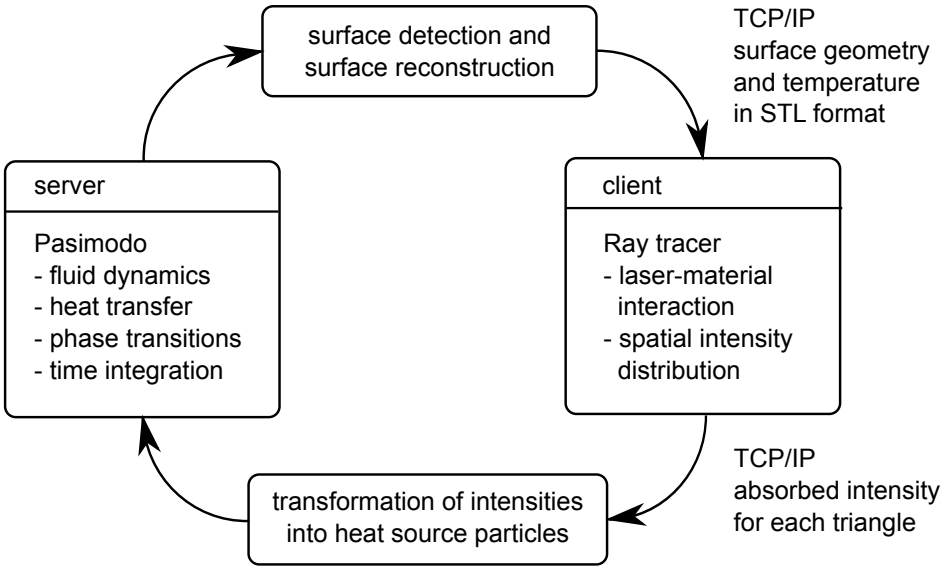



Figure 2: Co-simulation loop

as the total volume of the neighboring particles,

$$\zeta_s = \sum_j \frac{m_j}{\rho_j} = \sum_j V_j \,, \tag{9}$$

where $m_j$ is the mass, $\rho_j$ the density, and $V_j$ the volume of particle $j$. Afterwards, a source term

$$\frac{1}{\rho_j} \frac{I_s A_s}{\zeta_s} \tag{10}$$

in the unit W/kg is added to the right hand side of the energy equation for each SPH particle within a distance of $2r_s$.

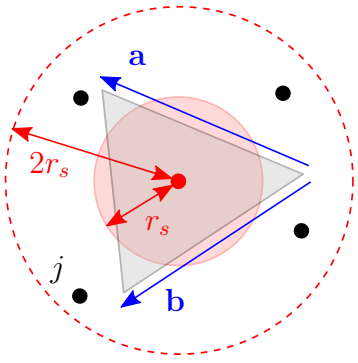



Figure 3: Transformation of a triangle into a heat source particle with radius $r_s$ interacting with its neighbor particles $j$ within a distance of $2r_s$

### 4.2 Surface detection algorithm

The applied algorithm for the detection of surface particles in the vicinity of the laser beam is adopted from [18] which combines the ideas given in [19] and [20]. For each particle $i$ the normal vector of the surface is approximated by a so-called cover vector

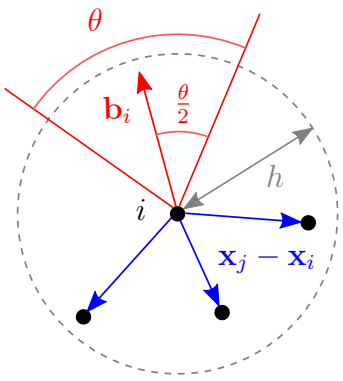
$$\mathbf{b}_i = \sum_j \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|} , \tag{11}$$

which is evaluated based on the distance to its neighbor particles $j$. Around the cover vector, a scan cone is defined with an apex angle of $\theta$, as shown in Fig. 4. In a second step, it is checked if at least another neighbor particle is located in the scan cone created by $\mathbf{b}_i$. An interior particle has to fulfill the condition

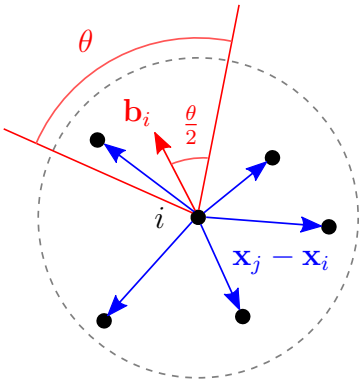
$$\arccos\left(\frac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|} \cdot \frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}\right) \leq \frac{\theta}{2} \tag{12}$$

for at least one neighbor particle $j$, otherwise the considered particle $i$ is regarded as surface particle. In Fig. 4 the dashed circle has the radius of the smoothing length $h$ of particle $i$. All neighboring particles located in the influence range contribute to the cover vector, and Eq. (12) is checked for each of them. In Fig. 4(a) particle $i$ is detected as a surface particle, since there are no other particles located in the scan cone created by $\mathbf{b}_i$. In Fig. 4(b), however, there is a neighbor particle located in the scan cone, for which Eq. (12) is satisfied. Therefore, particle $i$ is identified as an interior particle in this case.

This method is rather simple and fast, but it is not always accurate. Especially for randomly distributed particles, interior particles may be detected as surface particles and vice versa. Another drawback is the detection of different surface particles depending on different apex angles set for the scan cone. Here the angle $\theta$ is set to 60° in accordance with [18].




(a) Particle $i$ is a surface particle, since no other particles are located in the scan cone




(b) Particle $i$ is an interior particle, since a neighbor particle is detected in the scan cone

Figure 4: 2D example of a small set of SPH particles to illustrate the surface detection algorithm

### 4.3 Surface reconstruction algorithm

After all relevant surface particles are detected, the surface is reconstructed and a surface mesh consisting of triangles is created for the next simulation loop with the ray tracer. The

surface mesh is obtained with a Delaunay triangulation [21], in which the surface particles act as vertices of the triangles. Delaunay triangulations tend to avoid badly shaped triangles by maximizing the minimum angle of all triangles. There are several algorithms implementing the Delaunay triangulation method with $O(n \log n)$.

The applied algorithm is from [22, 23] which is based on randomized incremental construction. Currently, a 2D triangulation is performed based on the x- and y-coordinates of detected surface particles. Afterwards, the vertices of the triangles are shifted in z-direction. This approach is inadequate for keyhole geometries with undercuts, but works well for convex geometries.

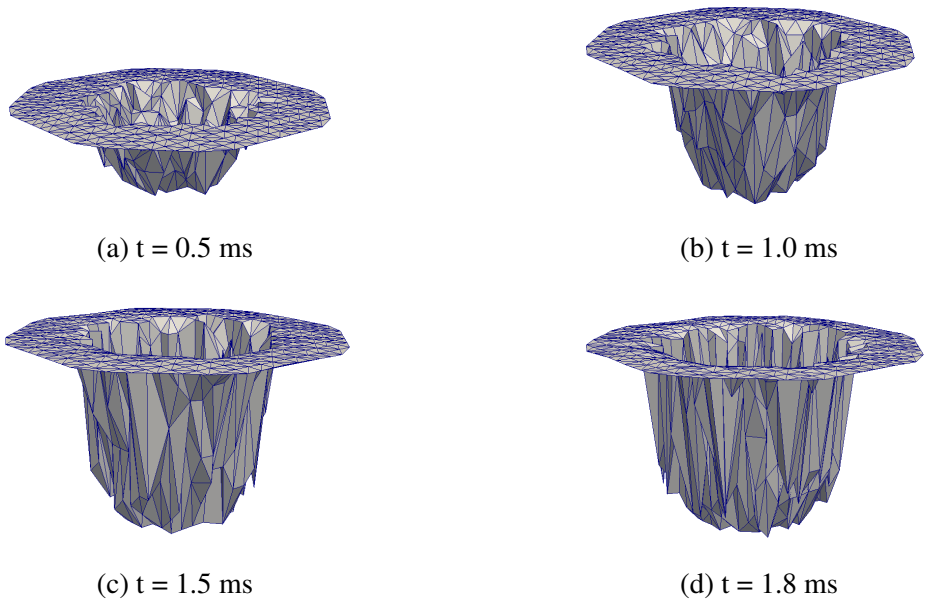
After the triangulation function call which delivers an index list of all triangles, the surface geometry and temperature data are sent to the ray tracer in binary STL format. The surface temperature of each triangle is approximated as mean value of the temperatures at its vertices.

## 5 EXAMPLES

To evaluate the co-simulation approach, two characteristic welding examples are chosen. The first example is spot welding of iron, and the second example is seam welding of aluminum. For both simulation examples the co-simulation loop is repeated every 0.1 ms.

### 5.1 Spot welding of iron

This example shows the formation of a keyhole during spot welding of iron. The applied laser beam has a power of 3 kW and a spot diameter of 0.3 mm. The surface geometry of the melt pool which is sent from Pasimodo to the ray tracer is shown for several time instants in Figure 5. For the given set of process parameters, the aspect ratio of depth to spot radius is approximately one at 0.5 ms, which marks the threshold from conduction mode to deep penetration welding. The aspect ratio increases to two at 1.5 ms.

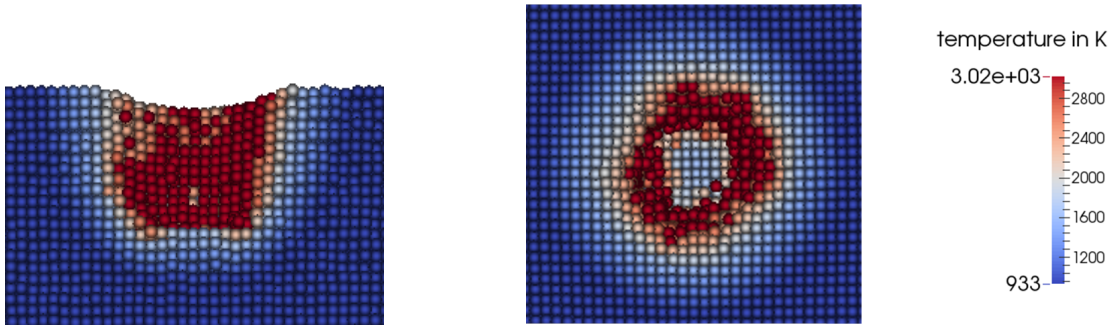


(a) t = 0.5 ms (b) t = 1.0 ms

(c) t = 1.5 ms (d) t = 1.8 ms

Figure 5: STL surface geometries sent from Pasimodo to the ray tracer which illustrate the formation of a keyhole during spot welding of iron

The temperature distribution around the keyhole at $t = 2\,\text{ms}$ is shown in Fig. 6. The evaporated particles are deleted during the simulation and thus, are not displayed. Nearly all surface particles around the keyhole have reached the evaporation temperature, but the absorbed amount of heat is still less than the latent heat of evaporation. The particles at the tip of the keyhole have a lower temperature, which indicates that the particle layer above have evaporated almost instantaneously shortly before the depicted instant of time. Therefore, both the evaporated particles and the laser beam did not have enough time to heat the current surface layer at the tip of the keyhole up to evaporation temperature.

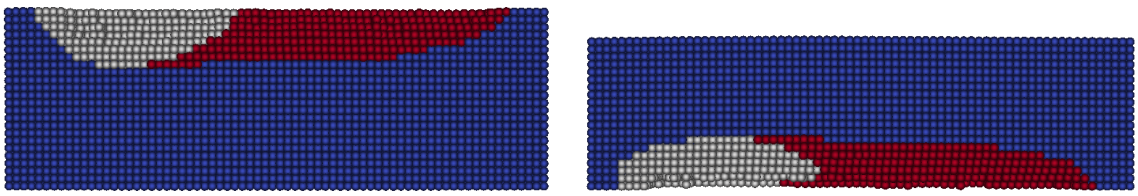



(a) view at mid cross section (b) view at top surface

Figure 6: Temperature distribution in the vicinity of the keyhole at $t = 2\,\text{ms}$

## 5.2 Seam welding of aluminum

As another example, seam welding of aluminum is simulated with a constant welding speed of 25 m/min. The laser power is set to 4.5 kW and the laser spot diameter is 0.6 mm. In Fig. 7 the SPH particles are colored based on the number of phase transitions. The particles in grey form the liquid weld pool, the particles in red were melted and solidified to form the resulting weld, and the particles in blue remained solid during the welding process. Again, we can analyze the aspect ratio of depth to spot radius, which is slightly larger than one, indicating that the threshold for deep penetration welding is reached. This corresponds well with the fact that some particles at the center have evaporated, but the amount of evaporated material is not large enough to form a considerable keyhole that results in a deep weld seam.



(a) view at mid cross section (b) view at half of the top surface

Figure 7: Seam welding of aluminum with weld pool colored in grey and solidified weld colored in red

## 6 CONCLUSIONS

The process of laser welding is simulated using the SPH method, where both the solid and liquid phase are modeled including significant physical effects such as surface tension, heat conduction, and thermal expansion. The phase transitions melting, solidification, and evaporation are taken into account together with the absorption or release of the latent heat of fusion and evaporation. The evaporated material is currently not modeled, but the recoil pressure that affects the melt is implemented based on a physical model in [12].

A co-simulation approach is developed by coupling an SPH code with a ray tracer to capture the laser-material interaction for complicated surface geometries. The data exchange is based on a TCP/IP protocol. Preliminary steps to the data exchange are the transformation of intensities into heat source particles interacting with other SPH particles at the surface of the melt, the surface detection, and the surface reconstruction. Suitable algorithms are chosen from literature and implemented. The obtained results are promising, but there are still some drawbacks and limitations in the implemented versions.

For future work, more complex algorithms for surface detection and reconstruction shall be implemented to better characterize the keyhole during laser welding. The goal is then to investigate instabilities that arise during the process, such as pore formation, spatter, and humping. This will be done numerically and also experimentally.

## ACKNOWLEDGEMENT

The research leading to the presented results has received funding from the German Research Foundation (DFG) under the project EB 195/13-1 and GR 3172/18-1 “Modeling of the capillary in laser beam penetration welding with the Smoothed Particle Hydrodynamics Method”. This support is highly appreciated.

## REFERENCES

[1] H. Hu, P. Eberhard, Two-Phase Modeling of Conduction Mode Laser Welding Using Smoothed Particle Hydrodynamics. In Proceedings of the IV International Conference on Particle-based Methods - Fundamentals and Applications (PARTICLES 2015), Barcelona, Spain, September 28-30, 2015.

[2] H. Hu, P. Eberhard, Thermomechanically Coupled Conduction Mode Laser Welding Simulations Using Smoothed Particle Hydrodynamics. Computational Particle Mechanics, 2016 (submitted for publication).

[3] R.A. Gingold, J.J. Monaghan, Smoothed Particle Hydrodynamics: Theory and Application to Non-Spherical Stars. Monthly Notices of the Royal Astronomic Society, Vol. 181, pp. 375–389, 1977.

[4] L.B. Lucy, A Numerical Approach to the Testing of the Fission Hypothesis. The Astronomical Journal, Vol. 82, No. 12, pp. 1013–1024, 1977.

[5] J.J. Monaghan, Smoothed Particle Hydrodynamics and Its Diverse Applications. Annual Review of Fluid Mechanics, Vol. 44, pp. 323–346, 2012.

[6] M. Tong, D.J. Browne, An Incompressible Multi-Phase Smoothed Particle Hydrodynamics (SPH) Method for Modelling Thermocapillary Flow. International Journal of Heat and Mass Transfer, Vol. 73, pp. 284–292, 2014.

[7] P.W. Cleary, J.J. Monaghan, Conduction Modelling Using Smoothed Particle Hydrodynamics. Journal of Computational Physics, Vol. 148, pp. 227–264, 1999.

[8] F. Fleissner, Parallel Object Oriented Simulation with Lagrangian Particle Methods. Schriften aus dem Institut für Technische und Numerische Mechanik, Vol. 16. Aachen: Shaker Verlag, 2010.

[9] Pasimodo, http://www.itm.uni-stuttgart.de/research/pasimodo. Accessed March 14, 2016.

[10] S.I. Anisimov, Vaporization of Metal Absorbing Radiation. Soviet Physics, Journal of Experimental and Theoretical Physics, Vol. 27, No. 1, pp. 182–183, 1968.

[11] V. Semak, A. Matsunawa, The Role of Recoil Pressure in Energy Balance During Laser Materials Processing. Journal of Physics D: Applied Physics, Vol. 30, pp. 2541–2552, 1997.

[12] X. Chen, H.-X. Wang, A Calculation Model for the Evaporation Recoil Pressure in Laser Material Processing. Journal of Physics D: Applied Physics, Vol. 34, pp. 2637–2642, 2001.

[13] R. Clausius, Über die bewegende Kraft der Wärme und die Gesetze, welche sich daraus für die Wärmelehre selbst ableiten lassen (in German). Annalen der Physik, Vol. 155, pp. 500–524, 1850.

[14] M.C. Clapeyron, Mémoire sur la Puissance Motrice de la Chaleur (in French). Journal de l'École Polytechnique, Vol. 23, pp. 153–190, 1834.

[15] A. Michalowski, Untersuchungen zur Mikrobearbeitung von Stahl mit ultrakurzen Laserpulsen (in German). Doctoral thesis, Institut für Strahlwerkzeuge (IFSW), University of Stuttgart. Munich: Herbert Utz Verlag, 2014.

[16] E. Hecht, Optics, 4th Edition. San Francisco: Addison-Wesley, 2002.

[17] J.J. Monaghan, H. Huppert, M. Worster, Solidification Using Smoothed Particle Hydrodynamics. Journal of Computational Physics, Vol. 205, pp. 684–705, 2005.

[18] A. Barecasco, H. Terissa, C.F. Naa, Simple Free-Surface Detection in Two and Three-Dimensional SPH Solver. Selected Paper from the International Symposium on Computational Science 2013, Kanazawa, Japan, 2013.

[19] H. Aamer, G.A. Dilts, Three-Dimensional Boundary Detection for Particle Methods. Journal of Computational Physics, Vol. 226, pp. 1710–1730, 2007.

[20] S. Marrone, Fast Free-Surface Detection and Level-Set Function Definition on SPH Solvers. Journal of Computational Physics, Vol. 229, pp. 3652–3663, 2010.

[21] B. Delaunay, Sur la Sphère Vide (in French). Bulletin de l'Académie des Sciences de l'URSS, Classe des Sciences Mathématiques et Naturelles, Vol. 6, pp. 793–800, 1934.

[22] K.L. Clarkson, K. Mehlhorn, R. Seidel, Four Results on Randomized Incremental Constructions. Computational Geometry: Theory and Applications, Vol. 3, pp. 185–212, 1993.

[23] K.L. Clarkson, Safe and Effective Determinant Evaluation. In Proceedings of 31st IEEE Symposium on Foundations of Computer Science, 387–395, Pittsburgh, USA, 1992.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MULTIPLE TIMESCALE MODELLING OF PARTICLE SUSPENSIONS IN METAL MELTS SUBJECTED TO EXTERNAL FORCES

**Anton Manoylov[1], Georgi Djambazov[1], Valdis Bojarevics[1] and Koulis Pericleous[1]**

[1] Computational Science and Engineering Group, University of Greenwich
e-mail: A.Manoylov@gre.ac.uk, K.Pericleous@gre.ac.uk,
web: http://cnmpa.gre.ac.uk/group_cseg.html

**Keywords:** CFD-DEM, particle adhesion, collision predictions, electro-magnetic stirring.

***Abstract.*** *Electro-magnetic (EM) fields are widely used in metallurgy in order to stir conducting metals without the risk of contamination or causing an instability or chemical reaction. During the manufacturing of metal matrix composites (MMC), ceramic micro- and nano-particles are added into the metal melt, and ultrasonic (US) processing and EM stirring are used to break the agglomerates and to enhance the dispersion of the particles. EM stirring can also be used to remove the unwanted particles from liquid metal by pushing them towards the walls of the crucible where they adhere and can be easily removed.*

*A model has been developed to account for the complex interaction of the particles with each other, with the walls, as well as with the flow of the metal melt. Particles are modelled as elastic spheres with adhesion. Adhesion is incorporated in the model using the Johnson, Kendal, Robert (JKR) and Derjaguin, Muller, Toporov (DMT) theories. The case of the oblique impact of the particles is modelled according to the Thornton and Yin method based on the partial-slip theory developed by Mindlin & Deresievics. The developed particle model is then coupled with the magneto-hydrodynamics (MHD) code PHYSICA in order to demonstrate the effect of the EM stirring and vibration.*

*Multiple time-scales are used which permits modelling the realistic time range of metal-processing and at the same time capture the individual collisions between particles with sufficient precision. Several methods of predicting the particle collisions are employed and their efficiency is compared for the case of removing contaminating particles from liquid metal.*

# 1 INTRODUCTION

This research is focused on developing the computational model of interacting particles and particle clusters suspended in liquid metals. The application of such model includes manufacturing of metal-matrix composites (MMC), where understanding of the forces acting on particles is required to prevent the formation of clusters and to disperse the reinforcing particles evenly in the volume of metal melt [1], [2]. Another application is removing the contaminants from liquid metal [3]-[5] which can be achieved by combining electro-magnetic (EM) stirring with EM expulsion owing to the fact that the metal is liquid and electrically conducting while the contaminants are solid and non-conducting.

Both applications share the same difficulty: the interaction of the particles with each other and with the walls of the crucible is essential for the outcome of the process, and therefore must be resolved accurately. The elastic forces, friction and cohesion between the particles dictate whether the clusters of particles form or break in the case of MMCs, while in the second case the efficiency of decontamination largely depends on the ability of the particles to adhere to the walls. The EM stirring is typically quite slow (5-10 cm/s, [5], [6]). In order to trace the particles during the stirring, at least one full cycle of stirring should be simulated, which is of the order of 1s for a 10 cm crucible. The large disparity in time-scales associated with particle collisions on one side, and on fluid flow on the other side, calls for multi-scale analysis of the problem, similar to that proposed in [7].

## 1.1 Fluid time-scale

In this paper the fluid flow is a consequence of applying the EM fields to electrically conducting liquid metal. EM fields induce electrical current in liquid metal that gives rise to a Lorentz force, which drives the stirring motion. It is assumed, that fluid flow is too slow to affect the imposed EM fields, therefore the EM fields are solved in the amplitude and frequency domain and no time-scale is associated with the EM fields. The fluid flow is modelled by solving the Navier-Stokes equation with an additional term representing the Lorentz force. The fluid time-scale for a 10 cm crucible is then of the order of 1 s. In the time reference of individual particle collisions the flow field can be assumed steady.

## 1.2 Fluid particle time-scale

If no collisions are happening, the particle motion depends on the fluid-particle interaction forces, where the most prominent contribution is that of the drag force, while other forces include lift and added mass, Magnus effect, pressure gradient force etc. These and other forces are reviewed in e.g. [8]-[13]. Some works involve resolving the fluid particle interaction using fluid mesh elements surrounding the particle [14] [15]. This method is convenient to accurately simulate the local fluid flow around the particle and derive the global fluid particle forces. It is however practical only if particles are larger than fluid flow mesh cells, or in other words, if particle size is comparable to the features of the flow. In this paper, particle's size ranges from 1 nm to 100 μm, while the minimum size of the fluid flow cell is ~3 mm. All the fluid particle interaction forces therefore rely on interpolated values of the fluid velocity and pressure. Under these conditions it is safe to assume that the forces acting on a particle do not significantly change until the particle moves to the distance, comparable to its own size. The fluid-particle time-scale is therefore associated with the particle size and average fluid velocity:

$$T_{fp} = \frac{R_p}{U} \tag{1}$$

To ensure accurate simulation, the fluid-particle time-step then must be

$$\Delta_{fp}t = C_{fp}\frac{R_p}{U} \tag{2}$$

Where $C_{fp}\leq 1$ is a parameter. For particles of 1μm radius immersed in the fluid stirred on average at U=5 cm/s the time-step $\Delta_{fp}t\leq 20$ μs for $C_{fp}=1$. If fluid velocity gradient is large then care should be taken to reflect this in the choice of $C_{fp}$. Taking $C_{fp}=0.01$ makes $\Delta_{fp}t=200$ ns.

### 1.3 Collision time-scale

The average collision time can be estimated based on the Young's modulus ($E_p$), particle radius ($R_p$), density ($\rho_p$) and approach velocity (U) [7]

$$T_c = 2R\left(\frac{\rho_p^2}{UE_p^2}\right)^{\frac{1}{5}} \tag{3}$$

For two SiC particles of 1μm radius approaching each other at U=5 cm/s the collision time $T_c$=2ns. Equation (1) is of course an approximation, as it does not take into account the impact angle, tangential forces, spin of the particles, adhesion force, surrounding fluid and presence of other particles. To ensure the accuracy of force transmission during the collision of particles, the time-step in modelling must be associated with the Rayleigh wave speed $V_r$ [16]:

$$\Delta_c t = C_r\frac{\pi R_p}{2V_r}, \tag{4}$$

Where $R_p$ is the particle radius, and $C_r\leq 1$ is a proportionality coefficient. Approximation of the Rayleigh wave speed $V_r$ using the material Poisson's ratio ν was presented by e.g. [17]:

$$V_r \approx \frac{256}{293} + \nu\left(\frac{60}{307} - \nu\left(\frac{4}{125} + \nu\left(\frac{5}{84} + \frac{4}{237}\nu\right)\right)\right) \tag{5}$$

For SiC particle of 1μm radius and C=1, (2) and (3) give $\Delta_c t$=0.2ns. This is equivalent to $T_c/10$, so C=1, 0.5 and 0.1 corresponds to modelling collision in 10, 50 or 100 time-steps. The ratio between the collision and fluid-particle time-steps is then

$$\frac{\Delta_{fp}t}{\Delta_c t} = 10000 \tag{6}$$

Which totally justifies the multi-scale approach.

## 2 REVIEW OF ADHESION THEORIES

Bradley [18] first described the van der Waals force acting between two rigid spheres in contact and calculated the pull off force as $P_c=4\pi\gamma R$, where $\gamma$ is interfacial energy of the contacting materials [1] and $R$ is the radius of the sphere.

Derjaguin [19] pointed out that elastic deformations of the spheres need to be accounted for as well as the adhesive interactions. He presented the first attempt to consider the problem of adhesion between elastic spheres: calculating the deformations of the spheres using Hertzian contact theory, he evaluated the work of adhesion assuming only the pair-wise interactions of the closest surface elements. The interaction energy between small elements of curved surfaces was assumed the same as for parallel planes which is known as the Derjaguin approximation.

On the other hand, Johnson [20] made an attempt to solve the adhesive contact problem by combining the Hertzian spherical contact problem and the problem of a rigid flat-ended punch. Johnson et al. [21] applied Derjaguin's idea to equate the work done by the surface attractions against the work of deformation in the elastic spheres to Johnson's [20] combined stress superposition. This resulted in the creation of the famous **JKR** (Johnson, Kendall, and Roberts) theory of adhesive contact [21]. According to them the attractive adhesion force is acting only over the contact area and significantly affects the shapes of the contacting spherical bodies. The pull

[1] The formulae for the pull off force of adhered particles are often used with the notation $\Delta\gamma$ which is the *work of adhesion*. For spheres of the same material $\Delta\gamma\approx\gamma/2$, therefore $P_c=2\pi\Delta\gamma R$

off force calculated using JKR model is $P_c=3\pi\gamma R$. The contact area is a circle with radius $a$, defined as follows:

$$a^3 = \frac{3R}{4E}\left[P + 6\pi\gamma R + \sqrt{12P\pi\gamma R + 36\pi^2\gamma^2R^2}\right], \quad (7)$$

where $P$ is the applied normal load and $E$ is the combined Young's modulus. Hertzian theory evaluates the contact radius simply as $a^3=3PR/4E$, therefore JKR theory is reduced to Hertzian if adhesion is neglected, i.e $\gamma$=0.

Derjaguin et al [22] developed a contact theory (**DMT –** Derjaguin, Müller, Toporov) that combined Bradley's adhesion force with Hertz elastic contact theory. The attractive intermolecular force is assumed applicable in the contact area as well as in the surrounding annulus zone. The resulting profile of the deformed spheres remains Hertzian and the pull off force is equal to the one derived by Bradley, $P_c=4\pi\gamma R$. The contact radius is then given by

$$a^3 = \frac{3R}{4E}\left[P + 4\pi\gamma R\right] \quad (8)$$

Qualitative analysis of both JKR and DMT models performed by Tabor [23] as well as more detailed analysis based on the Lennard-Jones potential conducted by Muller et al [24] showed that the contradiction between the models lies in the physical principles of adhesive contact assumed by the authors. Both Tabor and Muller concluded that the adhesive contact of larger, softer bodies with stronger surface interaction can be described by the JKR model, while the DMT model is applicable to the smaller, harder bodies with weaker surface interaction. Parameters $\tau$, $\mu$ were introduced in [23] and [24] to determine which model is more appropriate:

$$\tau \cong \left[\frac{R\gamma^2}{E^2z_0^3}\right]^{1/3}, \ \mu = \frac{32}{3\pi}\left[\frac{2R\gamma^2}{\pi E^2z_0^3}\right]^{1/3}, \quad (9)$$

where $z_0$ is the equilibrium separation distance, typically 0.16-0.4 nm [25]. According to Muller if $\mu$<1 then DMT is applicable whereas if $\mu$>>1 it is JKR.

Maugis [26] suggested a smooth transition model between JKR and DMT approaches which exploits the principles of fracture mechanics. For simplicity, Lennard-Jones interaction potential is replaced by the step-function, which is known as Dugdale approximation. Greenwood and Johnson [27] suggested an alternative model to Maugis based on a combination of two Hertzian profiles that also connect both the JKR and DMT models in one general theory. These two models use a parameter, which defines the area where the adhesion force is applicable. The necessity to evaluate this parameter at every time step during particle collision makes it impractical to use either Maugis [26] or Greenwood and Johnson [27] theories in a DEM solver. Therefore in the present paper the JKR and DMT models are implemented and the Müller parameter $\mu$ is used to determine which one is more applicable.

## 3 CONTACT MECHANICS

### 3.1 Oblique loading without adhesion.

The most commonly used particle contact model was first introduced by Cundall and Strack [28] in attempt to predict the complex behaviour of sand specimens under loading and unloading. They suggested treating sand particles as spheres which can move individually and interact only at the contact spots. The contact model consisted of linear spring elements as well and viscous damping elements in both normal and tangential directions, as shown schematically in Figure 1a. The modifications of this model are reviewed in e.g. [9][10]. The developments of this approach can include addition of rolling and twisting resistance [7] which are neglected in this paper.

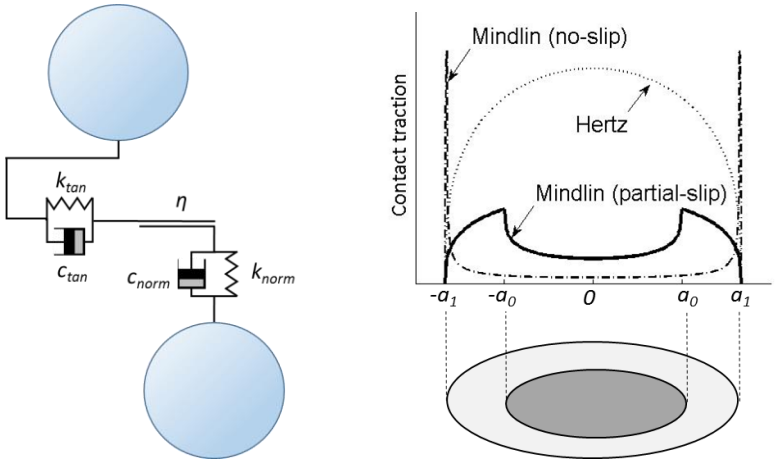



Figure 1 (a) Commonly used spring-dashpot and sliding element model; (b) contact traction distribution of two contacting spherical bodies according to the Mindlin and Deresiewicz model. ▭ - indicates circular zone with radius $a_0$ where elastic tangential force is applicable, □ - indicates the ring-shaped micro-slip area with external radius $a_1$..

While being extensively used in CFD-DEM simulation codes such as developed by Goniva, Kloss, Hager, Wierink and colleagues [12], [13], this model has a number of disadvantages. Firstly, accurate description of the contact between spherical bodies given by Hertz predicts non-linear normal elastic stiffness as $k_n=2E^*a$, where $E^*$ is the combined Young's modulus and $a$ is the radius of the (circular) contact area. It is noted in [7] that for small deformations Cundall and Strack model works well, although it is not obvious how to correlate the constant elastic stiffness values $k_{norm}$, $k_{tan}$ and viscous damping coefficients $c_{norm}$, $c_{tan}$ with properties of the materials involved. In addition to that, this paper considers nano- and micro-particles of sizes 50 nm to 100 μm, and therefore adhesion force must be incorporated. All of the adhesion models mentioned in the Section 2 of this paper are based on Hertz elastic theory. For these reasons, Hertz theory is used in this paper to evaluate the relationships between normal force and displacement as well as contact area.

The tangential contact forces are implemented in this paper by means of the Mindlin and Deresiewicz theory [29]. It is assumed that two elastic spheres in tangential contact experience a partial-slip, where the total force is a combination of elastic tangential force in the circular area in the centre of the contact zone and sliding friction force in the ring shaped exterior of the contact zone. Once the partial-slip tangential force exceeds the total sliding friction force, the bodies slide relative to each other. The tangential force in this case is then equivalent to the sliding friction force $F_s=\eta P$, where $\eta$ is the friction coefficient, $P$ is the normal load. The distribution of contact traction is illustrated in Figure 1b.

Thornton and Yin [30] combined all the major cases of the loading/unloading conditions described by Mindlin & Deresievicz [29] and derived the following expression for the tangential stiffness during oblique loading:

$$k^t = 8G^* a\theta \pm \eta(1-\theta)\frac{\Delta P}{\Delta \delta_t} \tag{10}$$

where $G^*$ is the combined shear modulus, $a$ is the contact radius, $\eta$ is the friction coefficient, $\Delta P$ is the increment of the normal load, $\Delta\delta_t$ is the increment of the tangential displacement and $\theta$ is a parameter defining the ratio of the elastic force to the micro slip friction force. The parameter $\theta$ depends on the loading history and is defined as follows:

$$\theta^3 = 1 - \frac{T + \eta\Delta P}{\eta P}; \qquad \theta^3 = 1 - \frac{T^* - T + 2\eta\Delta P}{2\eta P}; \qquad \theta^3 = 1 - \frac{T - T^{**} + 2\eta\Delta P}{2\eta P}, \tag{11}$$

for loading for unloading for reloading

where $T$ is current value of the tangential force and $T^*$ and $T^{**}$ are the load reversal points. Normal elastic stiffness is defined as $k^n=2E^*a$ according to Hertz theory; see [30] for details.

### 3.2 Oblique contact with JKR adhesion

Savkoor and Briggs [31] extended the JKR contact theory to consider the effect of adhesion in the case of oblique loading. It was suggested that applying the tangential force reduces the potential energy by an amount of $T\delta_t/2$. Adding this term to the JKR energy balance equation modified the contact radius (7) as:

$$a^3 = \frac{3R}{4E}\left[P + 6\pi\gamma R \pm \sqrt{12P\pi\gamma R + 36\pi^2\gamma^2R^2 - \frac{T^2E}{4G}}\right] \tag{12}$$

It was concluded that in the presence of tangential force, the contacting spheres peel off each other thus reducing the contact area. The peeling process continues until $T$ reaches the critical value of

$$T_c = 4\sqrt{(3P\pi\gamma R + 9\pi^2\gamma^2R^2)\mathrm{G/E}}. \tag{13}$$

For the normal load Thornton and Yin [21] have adopted the JKT theory. The stiffness is then evaluated as

$$k^n = 2E^*a\left[3 - 3\left(\frac{a_c}{a}\right)^{\frac{3}{2}}\right] / \left[3 - \left(\frac{a_c}{a}\right)^{\frac{3}{2}}\right] \tag{14}$$

where $a_c=9\pi\gamma R^2/4E$ is the JKR contact radius at the moment of separation (pull off radius).

In the case of oblique loading Thornton and Yin [30] followed [31] in what concerns the peeling process. They however assumed that once the peeling process is complete, the contacting bodies operate in the partial slip regime as described before with the difference that the normal force $P$ is replaced with $P+6\pi\gamma R$.

### 3.3 Oblique contact with DMT adhesion.

In this paper it is suggested to combine the Thornton and Yin [30] partial slip no adhesion model with DMT adhesion. The DMT theory assumes that the deformed shapes of the contacting bodies remain within Hertzian elastic theory. Therefore a no-adhesion model [30] was adopted where the normal force $P$ is replaced with $P+4\pi\gamma R$ to account for the adhesion force. This approach considers instantaneous separation of the particles, as opposed to the JKR theory, where particles stretch elastically prior to pulling off. The maximum stretching in the JKR case is evaluated as $\delta_c = \left(\frac{3\pi^2\gamma^2R}{16E^2}\right)^{1/3}$ whereas $\delta_c=0$ in the DMT case. The effect of the stretching prior to separation is illustrated in [2].

## 4 VISCOUS DRAG

The momentum of the fluid is transferred on the particles via the drag force. Di Felice's [32] theory is used to account for the effect of presence of other particles. Drag force on a single

particle in a flow with relative velocity $v = v_f - v_p$, where $v_f$, $v_p$ are the velocities of the fluid and the particle, can be evaluated as follows:

$$\begin{aligned} F_d &= \frac{1}{2}\rho_f v^2 \mathrm{C}_d \pi R_p^2 \varepsilon^{-\beta} \\ \mathrm{C_d} &= \left(0.63 + \frac{4.8}{\sqrt{\mathrm{Re_p}}}\right)^2 \\ Re_p &= \frac{\rho_f}{\mu_f}\alpha_f R_p |v_f - v_p| \end{aligned} \tag{15}$$

where $Re_p$ is the particle Reynolds number, $\mu_f$ and $\rho_f$ are dynamic viscosity and density of the fluid, $\varepsilon$ is the void fraction value, $C_d$ is the drag coefficient for spherical particles, and function $g(\varepsilon) = \varepsilon^{-\beta}$ is a measure of how much the drag force is affected by the presence of other particles. Empirical parameter β was evaluated to fit the experimental data for a wide range of Reynolds numbers ($10^{-2}$ to $10^4$) and void fraction values (0.4 to 1):

$$\beta = 3.7 - 0.65e^{-0.5\left(1.5 - log_{10} Re_p\right)^2} \tag{16}$$

In the literature, modifications of $g(\varepsilon)$ are used, such as $g(\varepsilon) = \varepsilon^{1-\beta}$ [10], $g(\varepsilon) = \varepsilon^{2-\beta}$, [12], [13], or $g(\varepsilon) = \varepsilon^{-1-\beta}$ [11]. Di Felice noted however that in the case of the flow through random packed spheres ($\varepsilon \approx 0.4$), Ergun's equation predicts $g(0.4) = \frac{14.6}{C_d}\left(1 + \frac{51.4}{Re_p}\right)$. For a wide range of Reynolds numbers $g(0.4)$ is best predicted by $g(\varepsilon) = \varepsilon^{-\beta}$. In e.g. Stokes drag formula is used multiplied by $g(\varepsilon) = \varepsilon^{-\beta}$ [7]. If the void fraction ε is close to unity, which is true for dilute suspensions, the choice of $g(\varepsilon)$ does not significantly affect the resulting drag force value.

The void fraction value ε is typically evaluated based on the density of particles in a mesh cell (see e.g. [12], [13]). In the present model the CFD mesh is not defined, therefore the void fraction is evaluated based on the particles located within $10R_p$ distance of the current particle centre.

## 5 PREDICTING PARTICLE COLLISIONS

### 5.1 Computational model

In order to obtain the positions, orientations, linear and angular velocities of the particles, the linear and angular momentum equations are solved:

$$\begin{aligned} m_p \frac{dv}{dt} &= F_0 + F_{fp} + F_c \\ I_p \frac{d\omega}{dt} &= M_{fp} + M_c \end{aligned} \tag{17}$$

where $m_p$ and $I_p$ are particle mass and moment of inertia, $v$ and $\omega$ are linear and angular velocities, $F_{fp}$ and $M_{fp}$ and $F_c$ and $M_c$ are total force and torque acting on the particle due to fluid particle interaction and collisions respectively, while $F_0$ is a sum of other forces, such as gravity or buoyancy. Equations (17) are discretized up to second order terms for positions and orientations and first order terms for velocities:

$$\begin{aligned} P(t_0 + \Delta t) &= P(t_0) + v(t_0)\Delta t + \frac{1}{2}\Delta_{fp} t^2 \frac{F(t_0, v_0)}{m_p} \\ v(\mathrm{t_0} + \Delta t) &= v(t_0) + \Delta t \frac{F(t_0, v_0)}{m_p} \end{aligned} \tag{18}$$

where $P$ is particle position vector, $F = F_0 + F_c + F_{fp}$ is total sum of forces, $t_0$ – current time and $\Delta t$ time-step. It is not defined at this stage whether $\Delta t$ is a fluid- particle or collision time-step.

For the equations for particle orientation and angular velocity the position must be replaced with orientation, velocity by angular velocity, forces by torques and mass by moment of inertia.

## 5.2 Multi-scale approach

The multi-scale approach is implemented is a similar way to that in [7]. Authors of [7] consider fluid, particle and collision time-steps, where collision time-step is used only for the particles that collide. The workflow of the method is schematically illustrated in Figure 2. The same method with minor modifications is implemented by the authors of this paper and is presented in Section 5.3. Fluid time-step is not applicable in this paper as the fluid flow is assumed steady. One way coupling is implemented from the fluid scale to fluid-particle scale, i.e. the effect of particle motion on the fluid flow is considered negligible in this paper. The term "particle time-step" in [7] is replaced by the authors with "fluid-particle time step" in this paper as it seems to be less confusing. From the description of the method [7] it follows that the collision time-step is applied to the particles as soon as they are identified as colliding, i.e. at the beginning of the fluid-particle time-step. In the case where the ratio between collision and fluid-particle time-steps is $\Delta_{fp}t/\Delta_c t \approx 10000$, modelling the motion of the particle using the collision time-step before the actual collision is an unnecessary waste of CPU time.

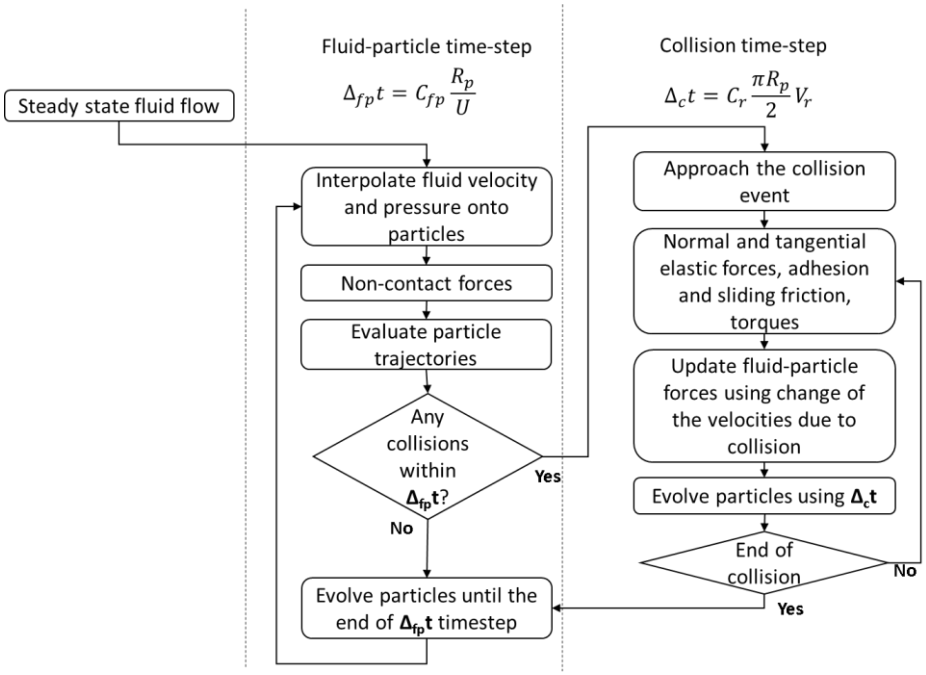



Figure 2 Work-flow of an algorithm with fluid-particle and particle collision timesteps

Section 5.4 presents a method where not only the collision pairs (particle-particle or particle-wall) are identified, but also the time of collision is evaluated. This allows to advance the colliding particles to the moment of their collision using the non-contact forces evaluated at the fluid-particle time-step. Then, the collision is resolved using collision time-step and particles are advanced to the end of the fluid-particle time-step. Most of the fluid-particle interaction forces are based on the relative velocity and spin of particle and fluid. In both methods therefore care should be taken to recalculate these forces using the velocity, direction of motion and spin modified due to collision. Both methods presented in Sections 5.3 and 5.4 are called "dilute" because they operate under the assumption that a colliding pair does not affect any other particles. This assumption holds provided that collisions are well dispersed in space and/or time.

This brings us to the method described in Section 5.5. It combines the benefits of time-step and event-driven methods. Event-driven methods are used in modelling granular gases or molecular dynamics [33], where collisions are rare and collision time is small in comparison to the global time-scale of the problem. The method considers identifying the whole sequence of collisions. Starting from the earliest collision, all the particles are advanced to the time of collision. After collision is resolved using the collision time-step, the trajectories of the colliding and neighbouring particles are re-evaluated. New collisions might be added to or removed from the sequence due to the changes in trajectories. The procedure is then repeated for the next collision in the sequence. The work-flow of the method is illustrated in Figure 3 .

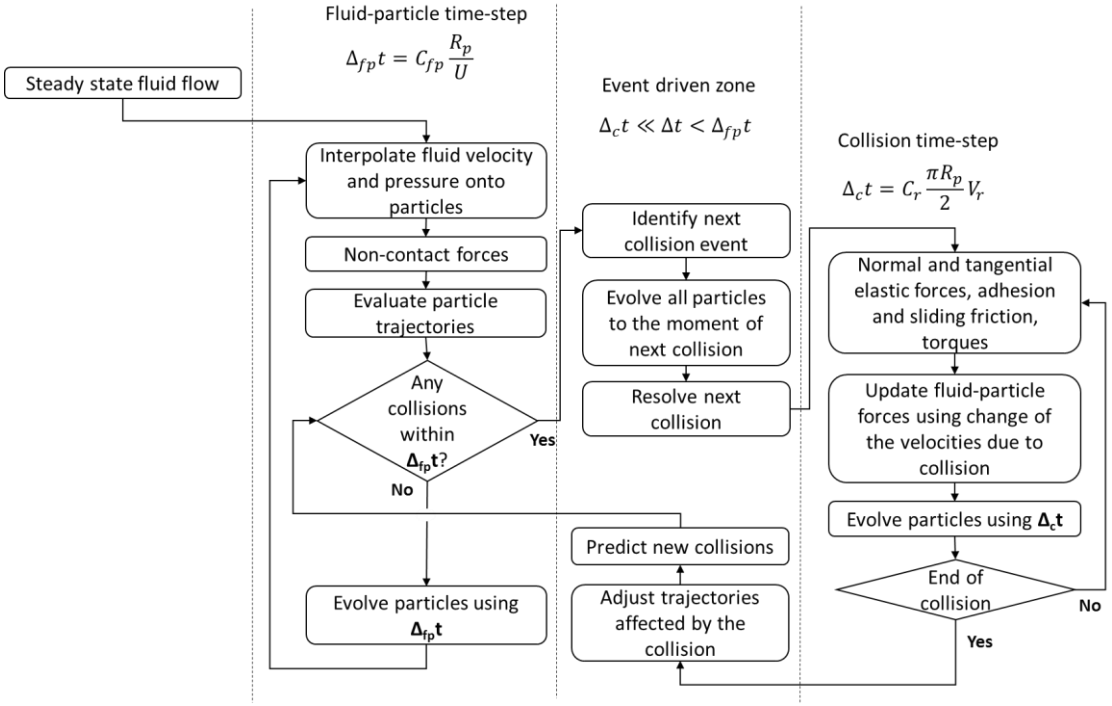



Figure 3 A combined time-step and collision event driven method: fluid particle forces evaluated every fluid-particle time-step are used for predicting particle trajectories; collisions are resolved using collision time-step; when particles are not colliding, the system of particles evolves from one collision event to another until the end of the fluid-particle time-step..

## 5.3 Linear dilute method

Let particles positions at the time $t_0$ at the beginning of the fluid-particle time-step be $P_1(t_0)$ and $P_2(t_0)$. Equation (18) then renders

$$P_i(t_0 + \Delta_{fp}t) = P_i(t_0) + v_i(t_0)\Delta_{fp}t + \frac{1}{2}\Delta_{fp}t^2\frac{F_i(t_0)}{m_i}, i = 1,2 \quad (19)$$

The trajectories of the particles within the fluid-particle time-step $\Delta_{fp}t$ are then parabolic curves connecting $P_i(t_0)$ with $P_i(t_0+\Delta_{fp}t)$. For simplicity, let us assume, that particles move along the straight lines connecting $P_i(t_0)$ with $P_i(t_0+\Delta_{fp}t)$ as shown in Figure 4a.

These straight lines can be parameterized for $t \in [t_0, t_0 + \Delta_{fp}t]$ as:

$$P_i(t) = P_i(t_0) + v_i^* t \quad (20)$$

where $v_i^*$ is given by:

$$v_i^* = \frac{P_i(t_0+\Delta_{fp}t)-P_i(t_0)}{\Delta_{fp}t} = v_i(t_0) + \frac{1}{2}\Delta_{fp}t\frac{F_i(t_0)}{m_i} \quad (21)$$

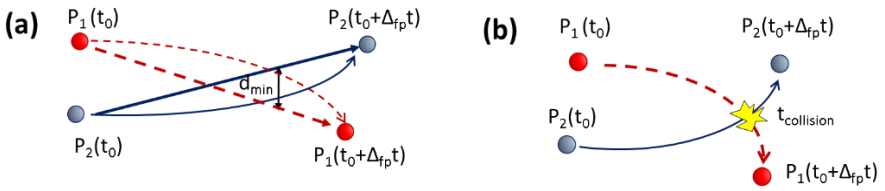



Figure 4 (a) Predicting particle collisions using the closest distance between (a) straight line approximation of particle trajectories; (b) parabolic trajectories of the particles

Note, that this linear approximation of the trajectory is different to that, obtained using first order terms for particle positions, as $v_i^*(t_0)=(v_i(t_0)+ v_i(t_0+\Delta_{fp}t))/2$. Assuming that particles are moving along a parameterized line defined by equation (20), the distance between them can be evaluated as [34]:

$$d(t) = |P_1(t) - P_2(t)| = \left|P_1(t_0) - P_2(t_0) + t\left(v_1^*(t_0) - v_2^*(t_0)\right)\right| \tag{22}$$

The "closest point of approach" as it is described in [34] is reached when $d(t)$ is minimum, therefore $d(t)^2$ is minimum:

$$d(t)^2 = t^2\left(v_1^*(t_0) - v_2^*(t_0)\right)^2 + 2t\left(P_1(t_0) - P_2(t_0)\right)\cdot\left(v_1^*(t_0) - v_2^*(t_0)\right) + \left(P_1(t_0) - P_2(t_0)\right)^2 \tag{23}$$

where "$\cdot$" denotes the dot product, and power of two denotes dot product of a vector with itself. Since $d(t)^2$ is minimum, the derivative must be equal to zero:

$$0 = 2\frac{d}{dt}dt = 2t\left(v_1^*(t_0) - v_2^*(t_0)\right)^2 + 2\left(P_1(t_0) - P_2(t_0)\right)\cdot\left(v_1^*(t_0) - v_2^*(t_0)\right) \tag{24}$$

which gives a solution $t=t_{cpa}$

$$t_{cpa} = -\frac{2\left(P_1(t_0)-P_2(t_0)\right)\cdot\left(v_1^*(t_0)-v_2^*(t_0)\right)}{\left|v_1^*(t_0)-v_2^*(t_0)\right|^2} \tag{25}$$

The minimum distance $d(t_{cpa})$ is then compared to the sum of the particle radii $R_1+R_2$. If $d(t_{cpa})>R_1+R_2$ then there is no collision. If $d(t_{cpa})\leqslant R_1+R_2$ and $t_{cpa}$ is within the fluid-particle time-step then the collision occurs. If $d(t_{cpa})\leqslant R_1+R_2$ and $t_{cpa}$ is out of the range of the fluid-particle time-step, then various scenarios are possible, which are illustrated schematically in Figure 5. Equation (22) defines a parabola facing upwards. If therefore, $t_{cpa}<t_0$ then $d(t)$ is increasing monotonously within the fluid-particle time-step, and it is only necessary to check whether collision occurs at the beginning of the fluid-particle time-step. If however $t_{cpa}>t_0+\Delta_{fp}t$ then $d(t_0+\Delta_{fp}t)$ should be tested for collision.

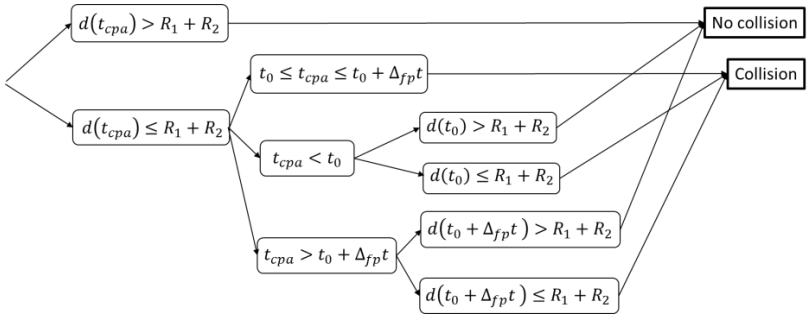



Figure 5 Set of conditions that the closest approach distance between particles is tested against

This method is the easiest for implementation among those considered in this paper. However, as a consequence of approximating the trajectories of the particles by the straight lines, the collision predictions are not accurate. Particles moving along the parabolic trajectories may

collide a time different to the predicted one, or not collide at all. The collision event is approached in a series of collision time-steps. The particles tagged as colliding are then advanced using the collision time-step until they have collided or until the end of the fluid-particle time-step is reached. The particle collisions are identified only at the beginning of the fluid-particle time-step. If collision of the particles changes their trajectories in a way that creates new non-predicted collisions, or renders some of the predicted collisions obsolete, this will not be reflected in the analysis. Reasonable accuracy can then be achieved with the optimal fluid-time particle step chosen by the cost of computation. The method has also been found efficient for a dilute concentration of particles, where collisions of more than two bodies (three particles or two particles and a wall) are negligibly rare.

### 5.4 Quartic dilute method

This method is similar to the linear dilute method described in Section 5.3 and the workflow is the same as represented diagrammatically in Figure 2. The only differences are how the collisions are identified and how the collision events are approached. In this method the trajectories as given by Equation (19) are not approximated by straight lines as shown in Figure 4a but treated as parabolic (Figure 4b). The squared distance between the particles is compared to the squared sum of the radii:

$$|P_1(t) - P_2(t)|^2 = (\mathrm{R}_1 + R_2)^2 \tag{26}$$

Equation (26) is a quartic equation in t:

$$\mathrm{A}_4 t^4 + A_3 t^3 + A_2 t^2 + A_1 t + A_0 = 0 \tag{27}$$

where

$$A_4 = \frac{1}{4}\left(\frac{F_1}{m_1} - \frac{F_2}{m_2}\right) \cdot \left(\frac{F_1}{m_1} - \frac{F_2}{m_2}\right)$$

$$A_3 = (v_1 - v_2) \cdot \left(\frac{F_1}{m_1} - \frac{F_2}{m_2}\right)$$

$$A_2 = (v_1 - v_2) \cdot (v_1 - v_2) + \big(P_1(t_0) - P_2(t_0)\big) \cdot \left(\frac{F_1}{m_1} - \frac{F_2}{m_2}\right)$$

$$A_1 = 2\big(P_1(t_0) - P_2(t_0)\big) \cdot (v_1 - v_2)$$

$$A_0 = \big(P_1(t_0) - P_2(t_0)\big) \cdot \big(P_1(t_0) - P_2(t_0)\big) - (R_1 + R_2)^2$$

Note that particle positions $P_1$ and $P_2$, velocities $v_1$ and $v_2$, and forces $F_1$ and $F_2$ are vectors, and "$\cdot$" denotes dot product.

Equation (27) is then solved analytically for real roots using the algorithm adapted for computations [35]. The method developed in [35] is easy to implement and far less computationally expensive than the classic Cardano-Ferrari formulae for quartic equations. Equation (27) has none, or one to four real roots. If there are no real roots, the trajectories of the particles do not intersect. The smallest real root fitting into the range of the fluid-particle time-step $t \in [t_0, t_0 + \Delta_{fp} t]$ is used as a collision time for a pair of particles. Note that conditions shown in Figure 5 do not have to be checked, as the solution of (27), if exists, represent the moment of collision, rather than the shortest approach distance between trajectories as in Section 5.3

For particle collisions with walls that are represented by planes, Equation (27) is reduced to quadratic equation. If walls and other geometrical features of the problem are represented by (parts of) spheres, cylinders or cones, then a quartic equation different to (27) must be solved, which is not covered in this paper.

This method saves computational time because the colliding particles are advanced to the moment of collision in one step rather than in a long series of collision time-steps. This method however shares the same disadvantage with the linear dilute method: particle collisions are predicted at the beginning of the fluid particle time-step and the predictions are not corrected due to collisions that have occurred previously.

### 5.5 Quartic collision driven method

This method is a development of the quartic dilute method (Sections 5.4) which takes into account the effect of the particle collision on other predicted collisions. For this purpose the collision that occurs first is identified. All the particles are then advanced to the moment of collision and the collision is resolved using the collision time-step. The new trajectories for all the particles are computed and the process is repeated until the end of the fluid-particle time-step. Additional modification allows to re-evaluate only the trajectories of the neighbouring particles.

The work-flow of the method is schematically illustrated in Figure 3. The collision loop stops or interrupts ("end of collision" block in the diagram) if one of three conditions are met: (a) particles have separated, (b) particles have stuck to each other (or to the wall), (c) collision took a certain number of time-steps (parameter) and is not yet resolved. Condition (c) is of particular importance if particles oscillate during the collision until the balance between attractive-adhesive and repulsive-elastic forces is achieved. If oscillation takes a long time, the collision loop must be interrupted in order to re-evaluate the fluid-particle forces and update the collision predictions. Condition (b) is met if a decaying oscillation between particles (or a particle and a wall) is detected. This condition is essential in the case where particles stick to the walls and remain there: particles that are stuck can then be removed from the computation cycle.

## 6 SIMULATION RESULTS

Earlier work by authors [2] considered the effect of the shock waves on dense agglomerates of particles in the context of de-agglomeration of ceramic nano-reinforcements in the production of aluminium based metal-matrix nano-composites [1]. The effect of the choice of adhesion model (JKR or DMT) was studied for a two-dimensional case of 36 densely packed particles as well as the effects of size, material properties and characteristics of the shock wave pulse [2]. The developments presented in the current paper are aimed at up-scaling the time and dimensions of the simulated problem. First, a test case is presented where particles are stirred using electro-magnetic fields (Section 6.1). In Section 6.2 the execution time is compared for the collision prediction methods described in Sections 5.2-5.5.

### 6.1 Electro-Magnetic Stirring of SiC particles

Test case considered in this paper simulates 216 SiC particles suspended in liquid silicon in the context of recycling of PV silicon kerf. The technique that is being developed in the project [3] aims to remove the SiC contaminants with the help of EM stirring. In brief, the solid non-conducting SiC particles approach the crucible walls as they move with the stirring flow of liquid silicon. Particles stick to the walls due to adhesion force and can be easily removed from the solidified silicon ingot by polishing the surface. The EM stirring (both the EM part and the fluid flow) has been modelled using the finite volume method which is described in detail in [36]. The current paper concentrates on the fluid-particle interaction and particle collisions only.

Initially, 216 particles are located in the nodes of a 6×6×6 grid as shown in Figure 6a. The pressure contours and velocity vectors obtained for the case of EM stirring of liquid silicon are shown in Figure 6b. Two distinct stirring poloidal motion patterns are indicated by arrows in

Figure 6b. Figure 7 (a) and (b) present the positions of particles after 10 seconds of stirring: (a) particles radius is 100 μm, (b) 10 μm. Black particles are those stuck to the walls or bottom of the crucible. Two conclusions can be drawn from Figure 7 (a) and (b). First, that larger, i.e. heavier, particles of 100 μm radius (Figure 7 (a)) sink to the bottom of the crucible before they get carried by the stirring motion. Second, that 10 seconds of stirring is clearly not sufficient to push all the particles towards the walls, thus allowing them a chance to stick. In reality, silicon is slowly cooled down, and stirring may last for several hours, giving plenty of time for the particles to stick to the walls. Several hours of processing is however rather large time-scale for modelling. The execution time of the simulation software for 10 second of real time is presented in the next section.

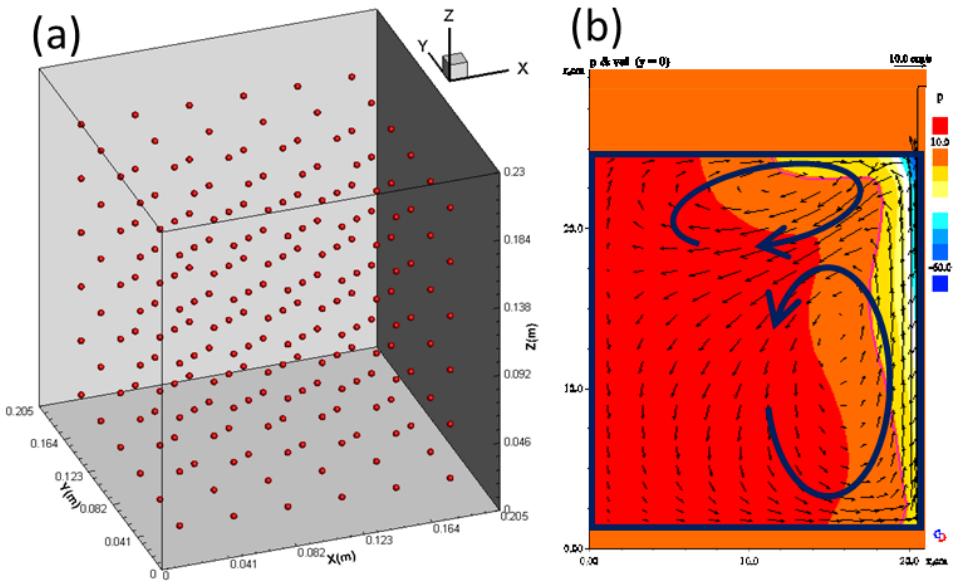



Figure 6 (a) Initial distribution of 216 particle (not to scale) located in the nodes of a 6x6x6 mesh within a cuboid domain of 20.5×20.5×23 $cm^3$ (b) fluid pressure contours and velocity vectors in a 41×41×23 $cm^3$ crucible; quarter of the crucible is shown, assuming symmetry conditions with respect to YZ and XZ planes, large arrows schematically indicate the sense of poloidal motion caused by EM stirring.

## 6.2 Execution time of collision prediction algorithms

In this section the execution time for 10 second of EM stirring of 216 particles suspended in liquid silicon is presented. The developed simulation code has room for significant improvement by utilizing multiple CPUs or a GPU card(s) using message passing interface. Using parallel computing would help to up-scale the model in terms of the number of particles. The algorithms developed in this paper helped to up-scale the model in terms of time and lateral dimensions. The code was tested on a single CPU of a multi-CPU machine based on Intel Xeon, 3.4GHz with 128 GB RAM.

The fluid-particle interaction and collision time-steps are defined as described in Sections 1.2 and 1.3, and therefore proportionally depend on the particle size. It is expected, that provided that all the conditions are similar, including the likelihood and frequency of collisions, the execution time for particles of different sizes should change proportionally with the sizes. This was found to be the case, as shown in the bar chart of Figure 8, where execution times are compared for particles of 100, 10 and 1 μm radii for the three described collision prediction methods, and a direct single-scale method, where no predictions are used and the whole problems is modelled using collision time-step.

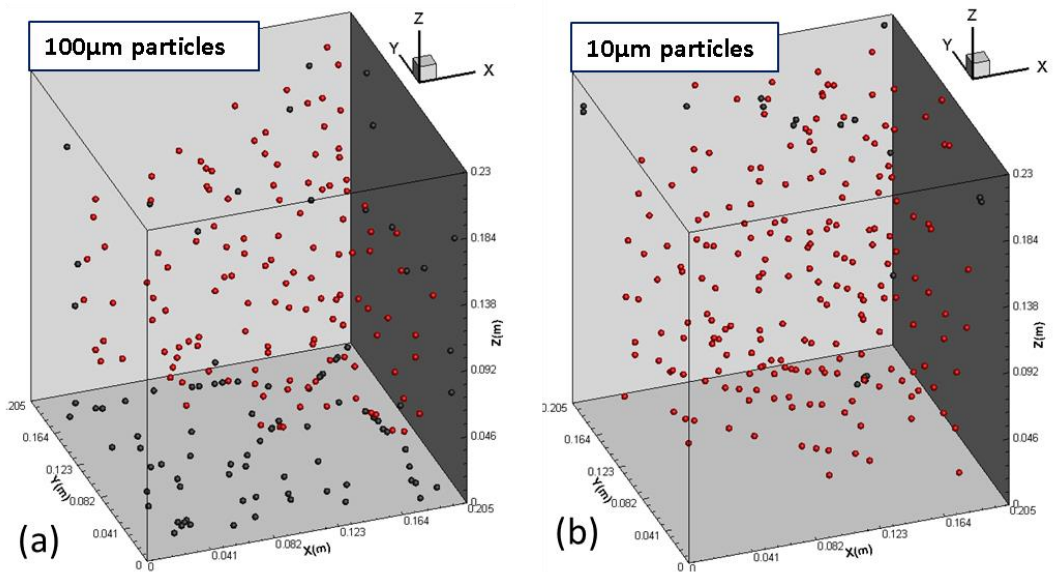



Figure 7 Particles after 10 s of stirring; **black** particles are those stuck to the bottom or walls of the crucible; (a) 100 µm (b) 10 µm; particles are not to scale.

Using the direct method it would take ~ 4 months, ~3.5 years and ~34 years to complete the simulation. The execution was halted after 48h of running and the total execution time was estimated based on the progress made so far. Using collision prediction models the simulation took approximately 45 min, 9 h and 90 h for particles of 100, 10 and 1 µm radii respectively, which is faster than the direct method by a factor of ~3000. Note that time-step coefficients $C_{fp}$ =0.01 (2) and $C_r$ =0.01 (4) were used, i.e. at least 100 time-steps to resolve an average estimated collision event. Increasing the coefficients up to e.g. 0.1 would decrease the simulation time by a factor of 10. There is also room for potential improvement of the code, which includes optimizing loops structure and data storage as well as using parallel computing. Main aim of this test case is to compare the prediction methods.

Owing to a very dilute concentration of particles, there was no noticeable difference in the particle trajectories. In fact, no predictions were compromised by the presence of other particles, i.e. the quartic dilute method gave perfectly accurate result, which might not be the case for a larger number of particles, if the initial positions are more concentrated, or if the fluid flow causes local concentrations of particles. The quartic dilute method was always the fastest due to the fact that particles were advanced to the exact moment of their collision and in spite of the fact that solving quartic equation requires extra algebraic operations and therefore is more computationally expensive. When using the collision driven method, time is lost due to re-evaluating the particles trajectories after each collision, which in general (not dilute) systems is expected to give a more accurate result. Linear dilute is algebraically the easiest, but time is lost when approaching the collision event in collision time-steps. Minor differences in the execution time for the prediction methods were: 45 minutes for linear dilute, 40 min for quartic and 43 min for collision driven methods for 100 µm particles; 10 h, 8 h 30 min and 9 h 20 min respectively for 10 µm particles, and 90 h, 84 h and 91 h for the smallest, 1 µm particles. The execution times are shown in Figure 8. Note that the execution time of a model that uses direct method goes far beyond the chart.

Figure 8 demonstrates that using the direct method results in unreasonably long computational time, while all of the prediction methods take approximately the same time. The quartic collision driven method although somewhat more complex in implementation is an obvious choice as it gives accurate predictions as opposed to the "dilute" methods.

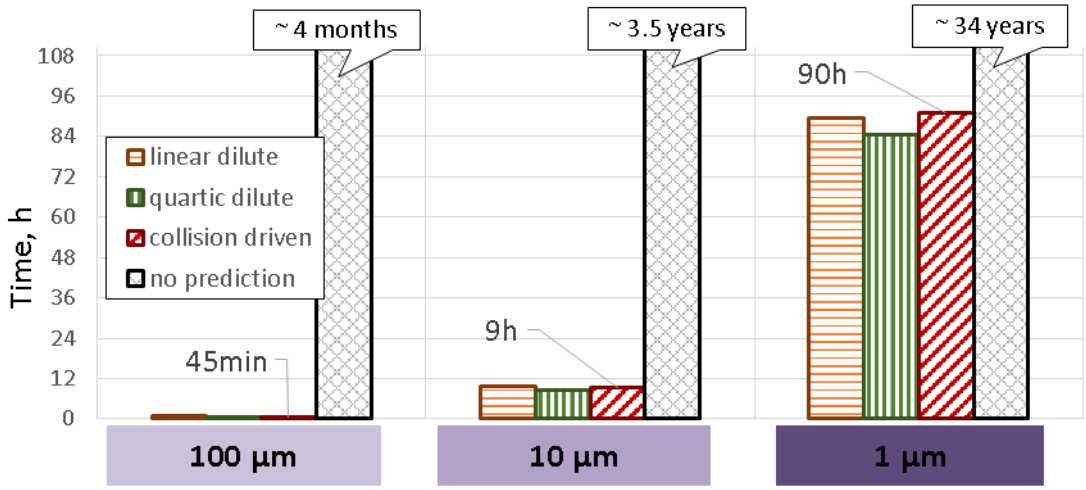



Figure 8 Comparison of the execution times for 10 seconds of real time simulation for linear dilute, quartic dilute, quartic collision driven and direct methods. Bars corresponding to the direct method case go well beyond the chart.

## 7 CONSLUSIONS

This paper considers CFD-DEM modelling of particles suspended in liquid metals. Dense agglomerates of particles are studied in detail in an earlier paper [2], while the current paper concentrates on the problem where concentration of particles is dilute, but the time and dimensions are large in comparison to the particle size and collision time. Multiple time-scale approach was implemented and three collision prediction methods are developed. Two methods are based on resolving the particle collisions independently, assuming that each collision is not interfering with other particle. These methods are called dilute as they are applicable to the dilute systems of particles only. The third method combined the benefits of time-step and event driven modelling. This method is also applicable to dense systems of particles, although the efficiency has not been tested. The execution time for the three prediction methods was compared and it was found that using multi-scale approach and collision prediction methods reduces the computational time by a factor of ~3000 for dilute systems. The differences in computational time for the three methods were minor, which makes the collision driven quartic method the optimal choice.

## ACKNOWLEDGMENTS

The authors acknowledge financial support from the projects ExoMet [1] (co-funded by the European Commission, contract FP7-NMP3-LA-2012-280421 and the European Space Agency) and SIKELOR [3] - funded the European Commission, sub-programme: ENV.2013.6.3–1, project reference 603718.

## REFERENCES

[1] http://www.exomet-project.eu/

[2] Manoylov, A., Bojarevics, V., Pericleous, K. A. Modeling the Break-up of Nano-particle Clusters in Aluminium- and Magnesium-Based Metal Matrix Nano-Composites. *Metallurgical and Materials Transactions. A* (2015) 46(7): 2893-2907

[3] http://www.sikelor.eu/

[4] I. Lombardi, F. Tappa, G. Fragiacomo, S. Pizzini, F. Hugo: "Progress in the recycling of silicon fine powders from fixed abrasive operations on Silicon Ingots. *Proc. 26th European PV Solar Energy Conference and Exhibition*, Hamburg, Germany, 2011

[5] Djambazov, G., Pericleous, K., Bojarevics, V., Forzan, M. and Dughiero F, Modeling the Electromagnetic Processing of Recycled Silicon Dust, *TMS Annual Meeting*, Nashville, Tennessee (2016),.

[6] Bojarevics, V., Djambazov, G. S., Pericleous K. A. (2015) Contactless Ultrasound Generation in a Crucible. *Metallurgical and Materials Transactions A*, 46A. pp. 2884–2892. (doi: 10.1007/s11661-015-2824-5)

[7] Marshall, J. S., Discrete-element modelling of particulate aerosol flows. *J. Comp. Physics* (2009) 228:1541-1561

[8] Marshall, J. S. and Li, S. *Adhesive Particle Flow: A Discrete-Element Approach*, Cambridge University Press, New York, 2014

[9] Zhou, Z. Y., Kuang, S. B., Chu, K. W., Yu, A. B. Discrete particle simulation of particle–fluid flow: model formulations and their applicability. *J. Fluid Mech.* (2010) 661: 482–510.

[10] Zhu, H.P., Zhou, Z.Y., Yang, R.Y., Yu, A.B. Discrete particle simulation of particulate systems: Theoretical developments. *Chemical Engineering Science* (2007) 62:3378 – 3396

[11] Kafui, K.D., Thornton, C., Adams, M. J. Discrete particle-continuum fluid modelling of gas-solid fluidised beds. *Chem. Eng. Science* (2002) 57:2395-2410

[12] Goniva C., et al., A MULTI-PURPOSE OPEN SOURCE CFD-DEM APPROACH 8th International Conference on CFD in Oil & Gas, Metallurgical and Process Industries, SINTEF/NTNU, Trondheim Norway, 21-23 June 2011

[13] Hager, A et al. Parallel Resolved Open Source CFD-DEM: Method, Validation and Application. *The Journal of Computational Multiphase Flows*, (2014) 6(1):13-27.

[14] Hill, R. J., Koch, D. L., Ladd, A. J. C. The first effects of fluid inertia on flows in ordered and random arrays of spheres, *Journal of Fluid Mechanics* 448, (2001) pp. 213-241

[15] Hill, R. J., Koch, D. L., Ladd, A. J. C. Moderate-Reynolds-number flows in ordered and random arrays of spheres, *Journal of Fluid Mechanics* 448, (2001) pp. 243-278

[16] Thornton, C., Yin, K.K., Adams, M. J., Numerical simulation of the impact fracture and fragmentation of agglomerates, *J. Phys. D: Appl. Phys.* 29 (1996) 424-435

[17] Pichugin, A. V. Approximation of the Rayleigh Wave Speed. *Unpublished draft* (2008)

[18] Bradley, R. S. The Cohesive force between Solid Surfaces and the Surface Energy of solids *Phil. Mag.* (1932) 13:853–62,

[19] Derjaguin, B. Untersuchungen über die Reibung und Adhäsion, IV. 1. Theorie des Anhaftens kleiner Teilchen. *Kolloid-Zeitschrift* (1934) 69:155-164.

[20] Johnson, K. L. A note on the adhesion of elastic solids. *British Journal of Applied Physics* (1958) 9:199-200.

[21] Johnson, K.L., Kendall, K. and Roberts, A.D. Surface Energy and the Contact of Elastic Solids. *Proceedings of the Royal Society A.* (1971)324:301–13

[22] Derjaguin, B.V., Muller, V.M. and Toporov, Y.P. Effect of Contact Deformations on the Adhesion of Particles. *Journal Colloid and Interface Science* (1975):53(2)

[23] Tabor, D. Surface Forces and Surface Interactions. *Journal Colloid and Interface Science* (1977) 58(1)

[24] Muller, V.M., Yushenko V. S. and Derjaguin, B.V. On the Influence of Molecular Forces on the Deformation of an Elastic Sphere and Its Sticking to a Rigid Plane. *Journal Colloid and Interface Science*. (1980) 77(1)

[25] Israelachvili, J. N., *Intermolecular and Surface Forces*, Academic Press, New York, (1985).

[26] Maugis, D. Adhesion of Spheres: The JKR-DMT Transition Using a Dugdale Model. *Journal Colloid and Interface Science.* (1992) 150(1)

[27] Greenwood, J. A., Johnson, K. L. An alternative to the Maugis model of adhesion between elastic spheres. *J. Phys. D: Appl. Phys*. (1998) 31:3279–3290

[28] Cundall, P. A., Strack, O. D. L. A discrete numerical model for granular assemblies. *Geotechique* (1979) 20(1): 47-65.

[29] Mindlin R. D. and Deresiewicz H. Elastic Spheres in Contact Under Varying Oblique forces. *J. Appl. Mech., Trans. ASME* (1953) 20:327.

[30] Thornton, C., Yin, K. K. Impact of elastic spheres with and without adhesion. *Powder Technology* (1991) 65:153-166

[31] Savkoor A.R. and Briggs, G. A. D. The Effect of Tangential Force on the Contact of Elastic Solids in Adhesion. *Proc. Roy. Soc. A* (1977) 356:103

[32] Di Felice, R. The voidage function for fluid-particle interaction systems. *Int. J. Multiphase Flow*(1994) 20(I):153-159

[33] Poeschel, T., Schwager, T., *Computational Granular Dynamics. Models and Algorithms*, , Springer, Berlin Hiedelberg New York, Ch. 3, p. 135 (2005)

[34] Schneider, P. J., Eberly, D. H., *Geometric Tools for Computer Graphics*, Morgan Kaufmann Publishers, San Francisco, (2003)

[35] H. E. Salzer, "A Note on the Solution of Quartic Equations", *Trans. AMS* (1959)

[36] Djambazov, G., Bojarevics, V., Pericleous K. A. and Croft, N. Finite volume solutions for electromagnetic induction processing. *Applied Mathematical Modelling* (2015),doi: 10.1016/j.apm.2015.03.059

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# STATISTICAL DISTRIBUTION FUNCTION OF CHARGED PARTICLES IN MAGNETIC FIELD

**B. Chetverushkin**[1], **N. D'Ascenzo**[1,2] **and V. Saveliev**[1,3]

[1]Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Russia
e-mail: office@keldysh.ru, saveliev@mail.desy.de

[2] Deutsches Elektronen Synchrotron, Hamburg, Germany
e-mail: ndasc@mail.desy.de

[3] Immanuel Kant Baltic Federal University, Kaliningrad, Russia
e-mail: saveliev@mail.desy.de

**Keywords:** Statistical distribution function, electromagnetic interaction, kinetic model, magneto gas dynamics

**Abstract.** *The statistical distribution function introduced by Boltzmann and his kinetic equation are the fundamental basis of the kinetic theory of gases and of the basic methods of solution of problems in the gas dynamics. At present time one of the areas of high interest in modern physics is the plasma in fusion processes and astrophysics which requires an extension of the kinetic description to charged particles dynamics, in particular regarding the electromagnetic interactions. We propose a unified distribution function which includes the electromagnetic interactions for charged particles in electromagnetic magnetic field and is suitable for the solution of problems of charged particle dynamics with Boltzmann type equations and kinetic consistent magneto gas dynamic equations.*

## 1 INTRODUCTION

The Boltzmann equation is the fundamental basis of the kinetic theory of gases [1]. The distribution function and the kinetic equations were introduced by Boltzmann for the accurate description of the microscopic properties and the dynamics of gases. The macroscopic properties of gases can be extracted as the moments of the distribution function. Despite the known difficulties in the solution of the Boltzmann equation modern developments are going to explore its solution on powerful high performance computing systems by numerical methods. Two modern examples are Lattice Boltzmann Schemes (LBS) , which use the evolution of the distribution function and calculate the macroscopic gas dynamic parameters as moments of the distribution function [2] and Kinetically Consistent Difference Schemes (KCDS), which derive the gas dynamic equation from the Boltzmann equation by integrating the velocity space with the summational invariants [3]. The resulting equations gives a more reach physical models of the dynamics of gases in comparison to the gas dynamic equations, as the Navie-Stokes equations, which are formulated phenomenologically [4].

Many modern important problems require not only to study the dynamics of neutral particles, as a gas, but also to deal with the dynamics of charged particles, as clouds of electrons, or charged interstellar matter in astrophysics or ionized gas in strong electromagnetic fields as in plasma physics processes. In order to solve these problems one has to take into account the electromagnetic interactions of the charged particles or charged media.

The study of the relations between the statistical gas dynamics and the electromagnetic laws goes in the direction of trying to combine directly the kinetic equations of the charged particles with the electromagnetic equations. One approach was done by Vlasov, who introduced the electromagnetic forces in the kinetic equation - Boltzmann - Vlasov equations. In this approach electromagnetic interactions are introduced phenomenologically as an external Lorentz force [5] in the momentum equation. Few later attempts to implement the electromagnetic terms directly in the distribution function faced the difficulty with the axial vector behaviour of the magnetic interactions [6, 7, 8].

We propose a new approach to define an unified distribution function, which includes the electromagnetic terms in the distribution function for the solution of the magneto gas dynamic problem by BLS or KCDS for charged particles in an electromagnetic field.

The main goal of this paper is to formulate the statistical distribution function with included electromagnetic terms and prove at the first step the validity of the proposed distribution function by the reconstruction of the kinetic consistent magneto gas dynamic equations including the equation of the evolution of the magnetic field in the ideal case for charged particles.

## 2 THEORETICAL ISSUES

The goal of the present study is the definition of the united distribution function describing the dynamics of the charged particles (electron cloud, ionised gas), including conditions with the external electromagnetic field.

### 2.1 Distribution Function of Gas Dynamic Processes

The kinetic theory describes the gas dynamics by the Boltzmann differential equation through the evolution of the distribution function $f(\mathbf{x}, \boldsymbol{\xi}, t)$ [1]:

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi}, t)}{\partial t} + \boldsymbol{\xi} \cdot \nabla f(\mathbf{x}, \boldsymbol{\xi}, t) = C(f), \tag{1}$$

where: $C(f)$ is a nonlinear integral operator which describes the collisions of particles.

The computational interest in the kinetic formulations of gas dynamics is high due to the linearity of the differential operator on the left side of Eq. (1). The nonlinearity is confined by the collision term, which is generally local in $\mathbf{x}$ and $t$.

Another important feature for numerical methods is that the collision integral vanishes in the equilibrium state when the local Boltzmann distribution function is Maxwellian:

$$f_M = \frac{\rho m^{1/2}}{(2\pi kT)^{3/2}} exp\left\{-\frac{m}{2kT}(\boldsymbol{\xi} - \mathbf{u})^2\right\} \tag{2}$$

The dynamic of the gas could be defined as the time evolution of the Maxwellian distribution function with the conditions that the collision processes is occurred during a characteristic time at the end of which the distribution function became Maxwellian [9]. The macroscopic observables such as density, momentum, energy flux as a function of $\mathbf{x}$ and $t$ are defined from the moments of the distribution function with respect to the particle velocity.

### 2.2 Electromagnetic Interactions

Earlier investigations [11] show that the electromagnetic fields do not destroy the validity of the distribution function and Boltzmann equation for the electrons cloud and open the way to the implementation of the electromagnetic processes terms in the distribution function.

From the kinetic relations, in addition to the gas kinetic motions, the motions of the charged particles is defined through nature of the electric and magnetic fields and the forces acting on the charged particles.

The electric field has the polar vector behaviour and the corresponding velocity of charged particles follows the direction of the electric field and can be considered as simple drift of the charged particles in the direction of the electric field.

The magnetic field has the axial vector behaviour and acts in a different way respect to other fields. As well known the motions of the charged particles in magnetic field affects only the direction of the velocity and the motion of the charged particle has a circular trajectory in the plane perpendicular to the magnetic field. The theory of complex variables is ideally suited to solve problems involving two dimensional circulation and in many ways complex variable theory is simpler than real variable theory and much more powerful [12].

As an example let's consider the motions of the charged particle in the uniform magnetic field [13]. The magnetic filed is directed along the $z$ - axis. The equation of the motions is:

$$\frac{d}{dt}\mathbf{p} = \frac{e}{c}\mathbf{v} \times \mathbf{B} \tag{3}$$

can be rewritten in term of momentum:

$$\mathbf{p} = \frac{E\mathbf{v}}{c^2} \tag{4}$$

where: $E$ is the energy of particle, which in the magnetic field is constant,

$$\frac{E}{c^2}\frac{d\mathbf{v}}{dt} = \frac{e}{c}\mathbf{v} \times \mathbf{B} \tag{5}$$

Or in the velocity components:

$$\frac{d}{dt}v_x = \omega v_y, \quad \frac{d}{dt}v_y = -\omega_x, \quad \frac{d}{dt}v_z = 0 \tag{6}$$

where: $\omega = \frac{ec\mathbf{B}}{E}$

Multiplying first eq. of 6 on $i$ and summing with second eq. of 6

$$\frac{d}{dt}(v_x + iv_y) = -i\omega(v_x + v_y) \tag{7}$$

from which:

$$v_x + iv_y = v_{0t}e^{i(\omega t+\alpha)} \tag{8}$$

where the $v_{0t} = \sqrt{v_x^2 + v_y^2}$ and $\alpha$ defined by the initial conditions.

This relation means that $v_{0t}$ is the velocity of the charged particle in plane $xy$, constant during the motions in the magnetic field.

### 2.3 Proposed Distribution Function

Following the above considerations the contribution to the velocity as result of the electromagnetic interaction can be defined as a complex term and the important that the magnetic interaction term in the conditions when the Larmor radius is smaller in comparison to the total volume is part of the statistical kinetic process, the charged particles don't changing the energy, momentum, but just the direction.

Using the above definitions we can obtain the local complex Maxwellian distribution function of charged particles with drift velocity $\mathbf{u}$ including the electric field interaction contribution and statistical term as complex velocity including the magnetic interaction contribution with the characteristic velocity is chosen as Alfven velocity [14] .

$$f_M = \frac{\rho m^{1/2}}{(2\pi kT)^{3/2}} \exp\left\{-\frac{m}{2kT}\left|(\boldsymbol{\xi} - \mathbf{u}) - i\frac{\mathbf{B}}{\sqrt{\mu\rho}}\right|^2\right\}, \tag{9}$$

where:
$\rho$ is the density,
$\boldsymbol{\xi}$ is the particle velocity
$\mathbf{u}$ is the drift velocity,
$T$ is temperature,
$\mu$ is the magnetic permeability, generally is a function of $\rho$.

### 2.4 The Macroscopic Magneto Gas Dynamic Parameters

The macroscopic hydrodynamic and electromagnetic observables can be obtained as a moments of the proposed distribution function respectively as the real and imaginary part of the integral along the line $L$ parallel to the real axis and shifted by $i\mathbf{B}/\sqrt{\mu\rho}$ in the imaginary axis direction.

$$\rho = \int_L m f_M d^3\boldsymbol{\xi} \tag{10}$$

$$\mathbf{u} = \frac{1}{\rho}\int_L m\boldsymbol{\xi} f_M d^3\boldsymbol{\xi} \tag{11}$$

$$E = \int_L \frac{1}{2} m\boldsymbol{\xi}^2 f_M d^3\boldsymbol{\xi} \tag{12}$$

$$\mathbf{B} = -\frac{1}{\sqrt{\mu\rho}}\int_L m\boldsymbol{\xi}^* f_M d^3\boldsymbol{\xi} \tag{13}$$

Thus, by using the proposed distribution function can calculate the mass, momentum, energy fluxes and most of the electromagnetic observables are calculated directly, i.e. one does not have to evaluate the hydrodynamic and magnetic force components separately or differently, as it will be shown below.

## 3 RECONSTRUCTION OF THE IDEAL MAGNETO GAS DYNAMICS SYSTEM OF EQUATIONS

In order to provide the validity of the proposed statistical complex distribution function, we show that the equilibrium condition reproduces correctly the ideal magneto gas dynamics system of equations.

The magneto gas dynamics system of equations is obtained by the integration of Eq. (1) with vanishing collision integral along the line $L$ with respect to the particle velocity $\boldsymbol{\xi}$. The summational invariants $\phi(\xi) = \left(m, m\boldsymbol{\xi}, \frac{1}{2}m\boldsymbol{\xi}^2\right)$ and $\phi(\xi) = m\boldsymbol{\xi}^*$ are used, respectively, for the gas dynamics observables and the magnetic induction:

$$\begin{aligned}&\int_L \phi\left(\boldsymbol{\xi}\right)\frac{\partial}{\partial t} f_M(\mathbf{x},\boldsymbol{\xi},t)d\boldsymbol{\xi}\\ &+\int_L \phi\left(\boldsymbol{\xi}\right) div(\boldsymbol{\xi} f_M(\mathbf{x},\boldsymbol{\xi},t))d\boldsymbol{\xi} = 0\end{aligned} \tag{14}$$

The continuity equation for the density is obtained from the real part of Eq. (14) with the summational invariant $\phi\left(\boldsymbol{\xi}\right) = m$:

$$\begin{aligned}\Re: &\int_L m\frac{\partial}{\partial t} f_M(\mathbf{x},\boldsymbol{\xi},t)d\boldsymbol{\xi}\\ &+\int_L m\frac{\partial}{\partial x_i}(\xi_i f_M(\mathbf{x},\boldsymbol{\xi},t))d\boldsymbol{\xi} = 0;\end{aligned} \tag{15}$$

After substituting the integration variable and inserting the proposed distribution function, we obtain:

$$\begin{aligned}&\frac{\partial}{\partial t}\int_{-\infty}^{+\infty} m\frac{\rho m^{1/2}}{\left(2\pi kT\right)^{3/2}}\exp\left\{-\frac{m}{2kT}\left|\boldsymbol{\xi}'\right|^2\right\}d\boldsymbol{\xi}'\\ &+\frac{\partial}{\partial x_i}\int_{-\infty}^{+\infty} m\left(\xi_i' + u_i\right)\frac{\rho m^{1/2}}{\left(2\pi kT\right)^{3/2}}\exp\left\{-\frac{m}{2kT}\left|\boldsymbol{\xi}'\right|^2\right\}d\boldsymbol{\xi}'\\ &= 0\end{aligned} \tag{16}$$

Finally from the evaluation of the integrals in Eq. (16) we obtain:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u_i}{\partial x_i} = 0. \tag{17}$$

Similarly, using the summational invariants $\phi(\boldsymbol{\xi}) = (m\boldsymbol{\xi}, \frac{1}{2}m\xi^2)$ respectively, we obtain from (14) the conservation laws for the momentum and energy, including the magnetic field contribution:

$$\frac{\partial}{\partial t}\rho u_i + \frac{\partial}{\partial x_k}\left[\left(p + \frac{B^2}{2\mu}\right)\delta_{ik} + \rho u_i u_k - \frac{B_i B_k}{\mu}\right] = 0 \tag{18}$$

$$\frac{\partial}{\partial t}E + \frac{\partial}{\partial x_i}\left[u_i\left(E + p + \frac{B^2}{2\mu}\right) - \frac{B_i}{\mu}\sum_k u_k B_k\right] = 0 \tag{19}$$

The total energy density includes thermal, kinetic and magnetic energies:

$$E = \frac{3}{2}p + \frac{1}{2}\rho u^2 + \frac{B^2}{2\mu} \tag{20}$$

The magnetic induction evolution equation is obtained from the imaginary part of Eq. (14) with the summation invariant $\phi(\boldsymbol{\xi}) = (m\boldsymbol{\xi}^*)$:

$$\begin{aligned}\Im : &\int_L m\xi_i^* \frac{\partial}{\partial t} f_M(\mathbf{x}, \boldsymbol{\xi}, t) d\boldsymbol{\xi} + \int_L m\xi_i^* \frac{\partial}{\partial x_k}(\xi_k f_M(\mathbf{x}, \boldsymbol{\xi}, t)) d\boldsymbol{\xi} \\ &= 0;\end{aligned} \tag{21}$$

After substituting the integration variable and inserting the distribution function, we get the equation for the imaginary part as:

$$\begin{aligned}&-\frac{\partial}{\partial t}\int_{-\infty}^{+\infty} m\frac{B_i}{\sqrt{\mu\rho}}\frac{\rho m^{1/2}}{(2\pi kT)^{3/2}}\exp\left\{-\frac{m}{2kT}\left|\boldsymbol{\xi}'\right|^2\right\}d\boldsymbol{\xi}' \\ &-\frac{\partial}{\partial x_k}\int_{-\infty}^{+\infty}\left(m\frac{1}{\sqrt{\mu\rho}}\left(\xi_k' B_i - \xi_i' B_k + u_k B_i - u_i B_k\right)\right. \\ &\left.\frac{\rho m^{1/2}}{(2\pi kT)^{3/2}}\exp\left\{-\frac{m}{2kT}\left|\boldsymbol{\xi}'\right|^2\right\}\right)d\boldsymbol{\xi}' = 0\end{aligned} \tag{22}$$

The evaluation of the Eq. 22 gives:

$$\frac{\partial}{\partial t}B_i + \frac{\partial}{\partial x_k}\left(u_k B_i - u_i B_k\right) = 0 \tag{23}$$

For the compressible media, we use the linear approximation of the magnetic permeability:

$$\mu = \frac{\mu_v}{\rho} \tag{24}$$

where $\mu_v$ is a constant representing the magnetic permeability per volume.

## 4 CONCLUSIONS

We propose the novel complex Maxwellian distribution function which includes electromagnetic interactions in a natural way and gives the possibility to reconstruct the macroscopic observables and magneto gas dynamic equations through moments of the distribution function.

This approach gives the possibility to use the proposed distribution function for the solution of the magneto gas dynamic problems by kinetic consistent methods and lattice Boltzmann methods, which are now considered as the most perspective ones for the calculation on modern parallel computing systems.

The computational kinetic schemes with the use of the proposed complex distribution function provide a methods derive the full magneto gas dynamic equations for the charged particles in a more physical way.

## 5 ACNOWLEDGEMENT

This work was supported by the Russian Science Foundation, project 14-11-00170

## REFERENCES

[1] Boltzmann L.,Vorlesungen uber Gastheorie, part I, 1896, part II, 1898, "Lectures on Gas Theory" (translation), University Of California press, 1964.

[2] Succi S., The Lattice Boltzmann Equation, Claredon Press, Oxford press, 2001.

[3] Chetverushkin B., Elizarova T., USSR Comput. Math.Phys, 25 (1985) 164.

[4] Batchelor G., An Introduction to Fluid Dynamics, Cambridge University Press,1967.

[5] Vlasov A., J. Experimental and Theoretical Physics 8 (1938) 291.

[6] Croisille J.-P., Khnfir R. and Chanteur G., J. Sci. Comput.10 (1995) 81.

[7] Huba J.D. and Lyon J.G., J. Sci. Comput. 61 (1999) 391.

[8] Dellar P.J., J. Comput. Phys.179 (2002) 391.

[9] Chetverushkin B., Russ. J. Numer. Anal. Math. Modelling, 20 (2005) 337.

[10] Chetverushkin B., Kinetic Schemes and Quasi-Gas Dynamic System of Equations, CIMNE, 2008.

[11] Tones L. and Allis W.P., Phys. Rev. 52 (1937) 117.

[12] Graebel W.P., Advanced Fluid Mechanics, Academic Press, Elsevier (2007).

[13] L.Landau L. and Lifshitz L., The Classical Theory of Fields, Volume 2 of a Course of Theoretical Physics, Pergamon Press, 1971.

[14] H.Alfven, Cosmical electrodynamics, Oxford, 1950.

[15] L.Landau L. and Lifshitz L., The Physics Kinetic , Volume 10 of a Course of Theoretical Physics, Pergamon Press, 1971.

[16] Ahern S. et al., ORNL/TM-2007/238 (2007).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A STABILISED TOTAL LAGRANGIAN CORRECTED SMOOTH PARTICLE HYDRODYNAMICS TECHNIQUE IN LARGE STRAIN EXPLICIT FAST SOLID DYNAMICS

***G. Greto[1], S. Kulasegaram[1], C. H. Lee[2], A. J. Gil[2], and J. Bonet[3]**

[1]Cardiff University
The Parade, Cardiff, CF24 3AA
e-mail: {gretog,kulasegarams}@cardiff.ac.uk

[2] Swansea University
Bay Campus, Swansea, SA1 8EN
{a.j.gil,c.h.lee}@swansea.ac.uk

[3]University of Greenwich
London, SE10 9LS
j.bonet@greenwich.ac.uk

**Keywords:** Conservation laws, SPH, Locking, Incompressibility, Fast dynamics, JST.

**Abstract.** *An explicit Total Lagrangian mixed momentum/strains formulation [1–5], in the form of a system of first order conservation laws, has been recently proposed to overcome the shortcomings posed by the traditional second order displacement-based formulation, namely: (1) bending and volumetric locking difficulties; (2) hydrostatic pressure fluctuations; and (3) reduced order of convergence for derived variables. Following the work of Bonet and Kulasegaram [6, 7], the main objective of this paper is the adaptation of Corrected Smooth Particle Hydrodynamics (CSPH) in the context of Total Lagrangian mixed formulation. Appropriate nodally conservative Jameson-Schmidt-Turkel (JST) stabilisation is introduced by taking advantage of the conservation laws. This mixed linear momentum-deformation gradient technique performs extremely well in nearly incompressible bending dominated scenarios [1, 2] without the appearance of spurious pressure oscillations. Additionally, as both linear momentum and deformation gradient are used as primary variables of the system, equal order of approximation should be achieved in both fields. A series of numerical examples are carried out to assess the applicability and robustness of the proposed algorithm.*

## 1 INTRODUCTION

Dynamic explicit displacement-based finite element codes based on low order finite element technology are commonly used for complex numerical simulations in the aerospace, automotive, biomedical, defence and manufacturing fields. However, difficulties arise in modelling high speed impacts or materials capable of enduring large finite deformations before failing. Most computational codes prefer to employ the 8-noded underintegrated hexahedral element to model solid components. Many practical applications (i.e. crashworthiness and drop-impact modelling), however, experience extremely large solid deformations accompanied by severe mesh distortion. This may lead to poorly shaped elements, unless some form of adaptive remeshing is applied.

From the viewpoint of spatial discretisation, the standard displacement based formulation with low order elements is known to experience locking behaviour in nearly-incompressible, bending dominated scenarios. It is also known that the use of linear interpolation within a finite element leads to second order convergence for the primary variables, but one order less for derived variables. From the time discretisation point of view, the Newmark method has a tendency to introduce high frequency noise in the solution field, especially in the vicinity of sharp spatial gradients.

In order to address these issues, a new mixed-formulation for Total Lagrangian fast solid dynamics has recently been proposed [4]. The new mixed-formulation consists in a system of first order conservation laws, formally akin to Computational Fluid Dynamics (CFD) models, with linear momentum and deformation gradient being the unknown conservation variables. In [2], the technique was applied in the framework of Finite Volume method, incorporating a Jameson-Schmidt-Turkel (JST) stabilising contribution [8]. Adding JST has the dual aim of (a) accurately representing potential discontinuities in the solution, and (b) introducing artificial dissipation to it where continuous. In [3], stabilisation is achieved through the use of Petrov-Galerkin algorithm. All these attempts have yielded promising results in two- and three-dimensional applications, with velocities (or displacements) and stresses (or strains) achieving the same degree of accuracy.

The main goal of this paper is the introduction of a stabilised Corrected Smooth Particle Hydrodynamics (CSPH) methodology, tailor-made for the aforementioned mixed formulation. The adoption of a system of first order conservation laws allows CFD techniques to be introduced within the context of particle methods for solid mechanics, in the form of an adapted nodally conservative JST stabilisation [2]. Incorporation of JST stabilisation will be useful to provide artificial (numerical) dissipation and shock-capturing capabilities. In particular, the second order (harmonic) and fourth order (biharmonic) operators employed in the JST stabilisation can be readily obtained by closed-form differentiation of the interpolating kernel functions. A quintic kernel approximation [6] is suitably used, due to the presence of the biharmonic JST operator.

This paper is structured as follows: section 2 outlines the theoretical framework of the $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation system of equations; section 3 focuses on SPH spatial discretisation of the governing equations. After that, section 4 describes the Total Variation Diminishing Runge-Kutta (TVD-RK) time integration method, used in the numerical computations. A series a numerical examples in section 5 assesses the feasibility and robustness of the proposed formulation. Concluding remarks are summarised in section 6.

## 2 p-F MIXED METHODOLOGY IN REVERSIBLE ELASTODYNAMICS

A deformable continuum body, of initial shape and position described by $\boldsymbol{X} \in V \subset \mathbb{R}^3$, finds itself in a current (or deformed) configuration $\boldsymbol{x} \in V(t) \subset \mathbb{R}^3$ at the end of a motion $\Phi$ that links each material particle in $\boldsymbol{X}$ to its current position $\boldsymbol{x}$ at time instant $t$, $\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{X}, t)$, see fig. 1a.

Neglecting dissipative effects, the dynamics of motion $\boldsymbol{\Phi}$ can be described by a $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation system of first order conservation laws as [1, 2]:

$$\left\{ \begin{array}{c} \dfrac{\partial \boldsymbol{p}}{\partial t} - \boldsymbol{\nabla}_0 \cdot \boldsymbol{P}(\boldsymbol{F}) = \rho_0 \boldsymbol{b} \\ \dfrac{\partial \boldsymbol{F}}{\partial t} - \boldsymbol{\nabla}_0 \cdot \left( \dfrac{1}{\rho_0} \boldsymbol{p} \otimes \boldsymbol{I} \right) = \boldsymbol{0} \end{array} \right. \tag{1}$$

Here, $\boldsymbol{p} \, (= \rho_0 \boldsymbol{v})$ is the linear momentum per unit of undeformed volume, $\rho_0$ is the material density, $\boldsymbol{v}$ represents the velocity field, $\boldsymbol{P}$ is the first Piola-Kirchhoff stress tensor, $\boldsymbol{b}$ is the body force per unit mass, $\boldsymbol{F}$ is the deformation gradient and $\boldsymbol{I}$ is the identity matrix. The notation $\boldsymbol{\nabla}_0$ denotes the material gradient operator (in undeformed space), where $[\boldsymbol{\nabla}_0]_I \equiv \frac{\partial}{\partial X_I}$. The first equation in (1) represents the conservation of linear momentum, while the second equation in (1) is a conservation law for the deformation gradient $\boldsymbol{F}$. The above laws (1) can be written in a more compact form:

$$\frac{\partial \boldsymbol{\mathcal{U}}}{\partial t} + \frac{\partial \boldsymbol{\mathcal{F}}_I}{\partial X_I} = \boldsymbol{\mathcal{S}} \qquad \forall I = 1, 2, 3 \tag{2}$$

where $\boldsymbol{\mathcal{U}}$ is the vector of conservation variables, $\boldsymbol{\mathcal{F}}_I$ is the flux vector in the material direction $I = 1, 2, 3$ and $\boldsymbol{\mathcal{S}}$ is the source term described by:

$$\boldsymbol{\mathcal{U}} = \left[ \begin{array}{c} \boldsymbol{p} \\ \boldsymbol{F} \end{array} \right] \qquad \boldsymbol{\mathcal{F}}_{\boldsymbol{N}} = \boldsymbol{\mathcal{F}}_I N_I = \left[ \begin{array}{c} -\boldsymbol{P}\boldsymbol{N} \\ -\frac{1}{\rho_0} \, \boldsymbol{p} \otimes \boldsymbol{N} \end{array} \right] \qquad \boldsymbol{\mathcal{S}} = \left[ \begin{array}{c} \rho_0 \boldsymbol{b} \\ \boldsymbol{0} \end{array} \right] \tag{3}$$

For closure of the above system (1), a material constitutive law that describes the way in which $\boldsymbol{P}$ depends on $\boldsymbol{F}$ has to be introduced. For this purpose, a standard nearly incompressible hyperelastic NeoHookean model will be used below. The path-independent behaviour of a hyperelastic material enables the direct extraction of the first Piola-Kirchhoff stress $\boldsymbol{P}$ from a strain energy functional $\psi(\boldsymbol{F}, J)$ as

$$\boldsymbol{P} = \frac{\partial \psi}{\partial \boldsymbol{F}} \tag{4}$$

where $J$ is the Jacobian of the deformation gradient, $J = \det \boldsymbol{F}$.

It is useful to decompose the strain energy functional $\psi$ into the summation of a volumetric and a deviatoric component:

$$\psi = \psi^{\text{vol}} + \psi^{\text{dev}} \tag{5}$$

where

$$\psi^{\text{vol}} = \frac{1}{2} \kappa (J - 1)^2 \qquad \psi^{\text{dev}} = \frac{1}{2} \mu \left( J_{\boldsymbol{F}}^{-2/3} \, (\boldsymbol{F} : \boldsymbol{F}) - 3 \right) \tag{6}$$

in (6), $\kappa$ is the bulk modulus of the material, and $\mu$ is its shear modulus.

The first Piola-Kirchhoff stress tensor $\boldsymbol{P}$ can then also be decomposed into the summation of deviatoric, $\boldsymbol{P}_{dev}$, and volumetric, $\boldsymbol{P}_{vol}$ , contributions:

$$\boldsymbol{P} = \boldsymbol{P}_{dev} + \boldsymbol{P}_{vol} \qquad \boldsymbol{P}_{vol} = p\boldsymbol{H} \qquad \boldsymbol{P}_{dev} = \mu J^{-2/3}\left(\boldsymbol{F} - \frac{1}{3}(\boldsymbol{F} : \boldsymbol{F})\ \boldsymbol{F}^{-T}\right) \tag{7}$$

in (7), $p = \kappa(J-1)$ is the hydrostatic pressure, and $\boldsymbol{H} = J\boldsymbol{F}^{-T}$ is the cofactor matrix of $\boldsymbol{F}$.

## 3 SPATIAL DISCRETISATION

In the present work, system (1) will be spatially discretised using the SPH method. Being a mesh-free technique, SPH can be employed effectively in the simulation of high-velocity impacts and high strain-rate deformations. Another advantage of SPH is the locality of the spatial discretisation, with consequent savings in computational power [9]. In a discretised domain, the SPH interpolation at the position $\boldsymbol{x}$ of a quantity $\phi(\boldsymbol{x})$, denoted here as $\langle\phi(\boldsymbol{x})\rangle$, can be approximated by a summation over all neighbouring particles $\boldsymbol{x}_b$:

$$\langle\phi(\boldsymbol{x})\rangle \approx \sum_b V_b\ \phi(\boldsymbol{x}_b)\ W(\boldsymbol{x}_b - \boldsymbol{x}, h) \tag{8}$$

where $h$ is the smoothing length, and $W(\boldsymbol{x} - \boldsymbol{x'}, h)$ is usually a polynomial function with compact support ($W$ vanishes for $\|\boldsymbol{x} - \boldsymbol{x'}\|_{L^2} \geq 2h$) for which the following properties are valid:

$$\int W(\boldsymbol{r}, h)\, d\boldsymbol{r} = 1; \qquad \lim_{h\to 0} W(\boldsymbol{r}, h) \to \delta(\boldsymbol{r})$$

Calculation of the first derivative of a function $\phi$ in SPH involves the derivation of the kernel function $W$:

$$\nabla\phi(\boldsymbol{x}) = \sum_b V_b\ \phi(\boldsymbol{x}_b)\ \nabla W(\boldsymbol{x}_b - \boldsymbol{x}, h) \tag{9}$$

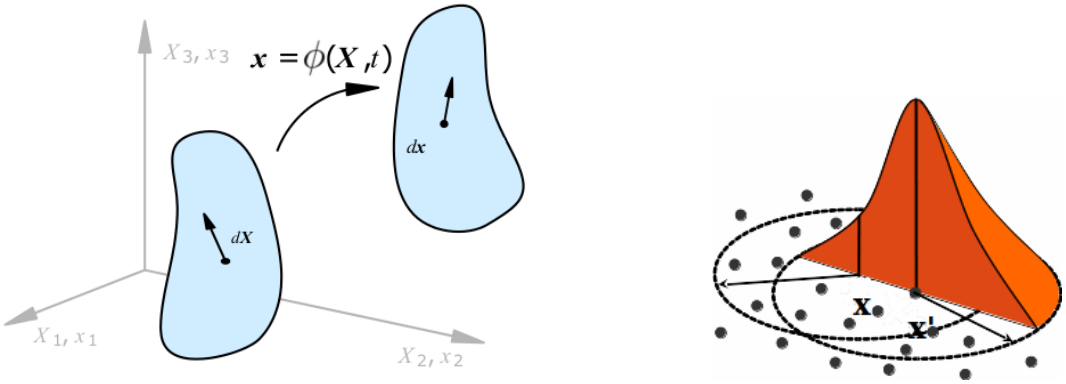
A typical SPH kernel function is presented in fig. 1b.

The SPH interpolation technique described above performs poorly at and near boundary regions. During the spatial discretisation, the border kernel supports are rendered incomplete by boundary edges, leading to kernel interpolations that are not partitions of unity. To improve the accuracy of the SPH interpolation near boundaries, and to exactly preserve momentum, corrections must be introduced on both the kernel and the kernel gradient [6,7,10]. The resulting corrected SPH (CSPH) formulation greatly enhances the accuracy and the consistency of the discretisation.

### 3.1 CSPH mixed formulation equations

The mixed $\boldsymbol{p}$-$\boldsymbol{F}$ system described by equation (1) will be discretised in space using the CSPH particle method with added dissipation from the JST technique ($\boldsymbol{p}$-$\boldsymbol{F}$ *JST-CSPH* scheme). To achieve this, and following the work of Bonet and co-authors [6, 7, 10, 11], the weak statement for the linear momentum evolution must be obtained through the use of work-conjugate principles [1] and integration by parts:

$$\int_V \delta\boldsymbol{v} \cdot \frac{\partial \boldsymbol{p}}{\partial t} dV = \int_V \delta\boldsymbol{v} \cdot \rho_0 \boldsymbol{b}\, dV + \int_{\partial V} \delta\boldsymbol{v} \cdot \boldsymbol{t}^B dA - \int_V \boldsymbol{P} : \boldsymbol{\nabla}_0 \delta\boldsymbol{v}\, dV \tag{10}$$




(a) (b)

Figure 1: (1a): The motion $\Phi$ of a continuous solid from a reference configuration $\boldsymbol{X}$ to the current configuration $\boldsymbol{x}$. (1b): Section of a typical, bell shaped SPH kernel function. The contribution of neighbouring particle in $\boldsymbol{x}'$ to the computed property at target position $\boldsymbol{x}$ is weighted by the value of its kernel at distance $\|\boldsymbol{x}' - \boldsymbol{x}\|$.

Upon application of the particle integration on the above expression (10), and using the kernel approximation $\boldsymbol{\nabla}_0 \delta \boldsymbol{v}(\boldsymbol{X}) = \sum_{b \in \Lambda_{\boldsymbol{X}}^b} V_b \, \delta \boldsymbol{v}_b \otimes \boldsymbol{\nabla}_0 W_b(\boldsymbol{X})$ to express the arbitrary virtual velocity field $\delta \boldsymbol{v}$, one obtains:

$$\frac{d\boldsymbol{p}_a}{dt} = \rho_0 \boldsymbol{b}_a + \frac{A_a}{V_a} \boldsymbol{t}_a^B + \sum_{b \in \Lambda_a^b} V_b \, (\boldsymbol{P}_b + \boldsymbol{P}_a) \, \tilde{\boldsymbol{\nabla}}_0 W_b(\boldsymbol{X}_a) \tag{11}$$

where $V_b$ is the volume fraction assigned to the particle in $\boldsymbol{x}_b$ and $W_b(\boldsymbol{X})$ is the kernel function value there.

Following References [6,7,9–11], the first Piola-Kirchhoff stress tensors $\{\boldsymbol{P}_a, \boldsymbol{P}_b\}$ are further approximated by using values of $\{\boldsymbol{F}_a, \boldsymbol{F}_b\}$ estimated directly at the particles. In (11), corrections are introduced on the kernel gradient to yield $\tilde{\boldsymbol{\nabla}}_0 W_b(\boldsymbol{X}_a)$, ensuring the correct evaluation of a gradient of a general linear function [6, 10]. However, the above semi-discrete formulation still suffers from the appearance of accumulated numerical instabilities over a long term response. To overcome this, in contrast to the dissipative mechanism introduced by Monaghan [9] through viscous fluxes, here a nodally conservative JST stabilisation $\boldsymbol{\mathcal{D}}(\boldsymbol{p}_a)$ will be incorporated into (11), mirroring CFD techniques [2]:

$$\frac{d\boldsymbol{p}_a}{dt} = \rho_0 \boldsymbol{b}_a + \frac{A_a}{V_a} \boldsymbol{t}_a^B + \sum_{b \in \Lambda_a^b} V_b \, (\boldsymbol{P}_b + \boldsymbol{P}_a) \, \tilde{\boldsymbol{\nabla}}_0 W_b(\boldsymbol{X}_a) + \boldsymbol{\mathcal{D}}(\boldsymbol{p}_a) \tag{12}$$

The nodally conservative JST stabilisation is additively decomposed into a second order (harmonic) operator $\boldsymbol{\mathcal{D}}_2(\boldsymbol{p}_a)$ and a fourth order (biharmonic) operator $\boldsymbol{\mathcal{D}}_4(\boldsymbol{p}_a)$:

$$\boldsymbol{\mathcal{D}}(\boldsymbol{p}_a) = \boldsymbol{\mathcal{D}}_2(\boldsymbol{p}_a) + \boldsymbol{\mathcal{D}}_4(\boldsymbol{p}_a) \tag{13}$$

where

$$\boldsymbol{\mathcal{D}}_2(\boldsymbol{p}_a) = \kappa^{(2)} c_p \, h_{\min} \sum_{b \in \Lambda_a^b} V_b \left( \boldsymbol{p}_b - \boldsymbol{p}_a \right) \tilde{\nabla}_0^2 W_b(\boldsymbol{X}_a) \tag{14}$$

$$\boldsymbol{\mathcal{D}}_4(\boldsymbol{p}_a) = - \kappa^{(4)} c_p \, h_{\min}^3 \sum_{b \in \Lambda_a^b} V_b \left( \nabla_0^2 \boldsymbol{p}_b - \nabla_0^2 \boldsymbol{p}_a \right) \tilde{\nabla}_0^2 W_b(\boldsymbol{X}_a) \tag{15}$$

In (14) and (15), $c_p$ is the pressure wave speed, $h_{min}$ is the particle spacing, $\kappa^{(2)}$ and $\kappa^{(4)}$ are user-defined parameters and $\tilde{\nabla}_0^2$ represents the corrected Laplacian operator [6].

In regard to the fibre map evolution (1b), the application of particle collocation method [6] directly on (1b), along with the use of kernel approximation $W$, leads to:

$$\frac{d\boldsymbol{F}_a}{dt} = \boldsymbol{\nabla}_0 \boldsymbol{v}\big|_a \approx \sum_{b \in \Lambda_a^b} \frac{V_b}{\rho_0} \left( \boldsymbol{p}_b - \boldsymbol{p}_a \right) \otimes \tilde{\boldsymbol{\nabla}}_0 W_b(\boldsymbol{X}_a) \tag{16}$$

Finally, the set of stabilised particle equations (12) and (16) can then be explicitly integrated from time step $t^n$ to $t^{n+1}$. As described in [1], the scheme is suitably modified to guarantee the conservation of angular momentum.

## 4 TEMPORAL DISCRETISATION

As the resulting set of semi-discrete equations (12) and (16) is not simple, it will be more appropriate to use an explicit type of time integrator. The spatial discretisation discussed above yields a system of ordinary differential equations of the form:

$$\frac{d\boldsymbol{\mathcal{U}}_a}{dt} = \boldsymbol{\mathcal{R}}_a(\boldsymbol{\mathcal{U}}_a, t) \tag{17}$$

where $\boldsymbol{\mathcal{R}}_a(\boldsymbol{\mathcal{U}}_a, t)$ represents the right-hand-side of SPH spatial discretisation associated with particle $a$. To advance (17) in time from $t_n$ to $t_{n+1}$, an explicit, two-stage TVD-RK time integrator is used:

$$\begin{cases} \boldsymbol{\mathcal{U}}_a^* = \boldsymbol{\mathcal{U}}_a^n + \Delta t \, \boldsymbol{\mathcal{R}}_a(\boldsymbol{\mathcal{U}}_a^n, t^n) \\ \boldsymbol{\mathcal{U}}_a^{**} = \boldsymbol{\mathcal{U}}_a^* + \Delta t \, \boldsymbol{\mathcal{R}}_a(\boldsymbol{\mathcal{U}}_a^*, t^{n+1}) \\ \boldsymbol{\mathcal{U}}_a^{n+1} = \dfrac{1}{2}(\boldsymbol{\mathcal{U}}_a^n + \boldsymbol{\mathcal{U}}_a^{**}) \end{cases} \tag{18}$$

The explicit TVD-RK has been adopted here due to its excellent entropy conserving properties [1, 2]. The time step $\Delta t = t^{n+1} - t^n$ is calculated based on the Courant-Friedrichs-Lewy number $\alpha_{CFL}$, by means of $\Delta t = \alpha_{CFL} \left( h \,/ c_{p,max} \right)$ where $c_{p,max}$ is the maximum $p$-wave speed and $h$ is the smoothing length.

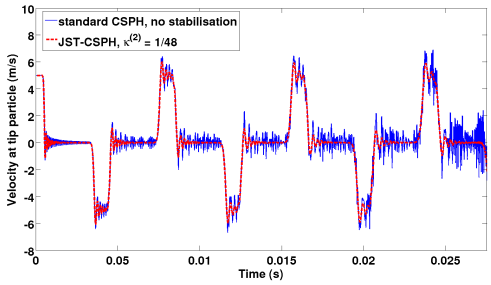
Along with $\boldsymbol{p}$ and $\boldsymbol{F}$, the geometry, $\boldsymbol{x}$, is also advanced in time using the same time integrator in a monolithic manner.

## 5 APPLICATIONS

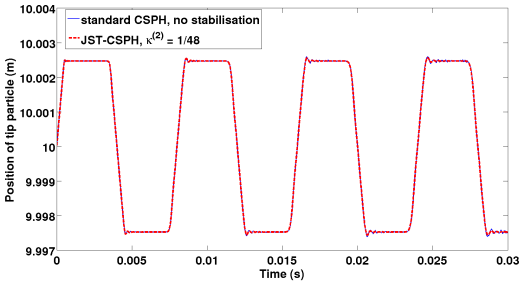
A series of benchmark examples are tested in order to assess the applicability and effectiveness of the $\boldsymbol{p}$-$\boldsymbol{F}$ JST-CSPH mixed algorithm over the Total Lagrangian CSPH displacement based formulation [6, 7, 10].

**1-D rod in tension** A 1D bar is clamped on one end and left free at the other, where is given an initial velocity of $5\,m/s$. The chosen material is linearly elastic, with elastic modulus $E = 200\,GPa$ and density $\rho = 8000\,kg/m^3$. The bar is discretised in space using 200 uniformly distributed particles. Aim of this example is to perform a preliminary assessment of the method on a simple problem.

Fig. 2a shows the velocity history at the free tip, while fig. 2b displays the tip displacement. The JST term here has a decisive effect towards achieving a more accurate solution. The Total Lagrangian setting ensures that the simulation does not suffer from tensile instability, a major issue in solid mechanics SPH simulations (see [7, 12, 13]).




(a) Velocity plot at tip particle with (red) and without (blue) JST dissipation.




(b) Displacement plot at tip particle with (red) and without (blue) JST dissipation.

Figure 2: 1d tensile test, $\boldsymbol{p}$-$\boldsymbol{F}$ CSPH based model.

**Bending column** A bottom end clamp is imposed at $t = 0\,s$ on a 3D column of length $6\,m$, with unit cross section, initially travelling at constant side velocity $v = 10\,m/s$. The column is made of a rubbery material modelled as nearly incompressible NeoHookean, with properties $E = 17\,MPa$, $\nu = 0.45$, $\rho = 1100\,kg/m^3$. The spatial discretisation is obtained through the CSPH-JST $\boldsymbol{p}$-$\boldsymbol{F}$ model, using a total of 3969 particles, while a value of $\alpha_{CFL} = 0.3$ has been chosen to control the time step size. Fig. 3 demonstrates the performance of the model in a bending dominated scenario.

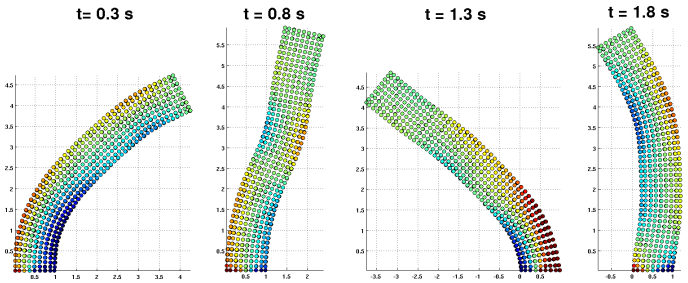
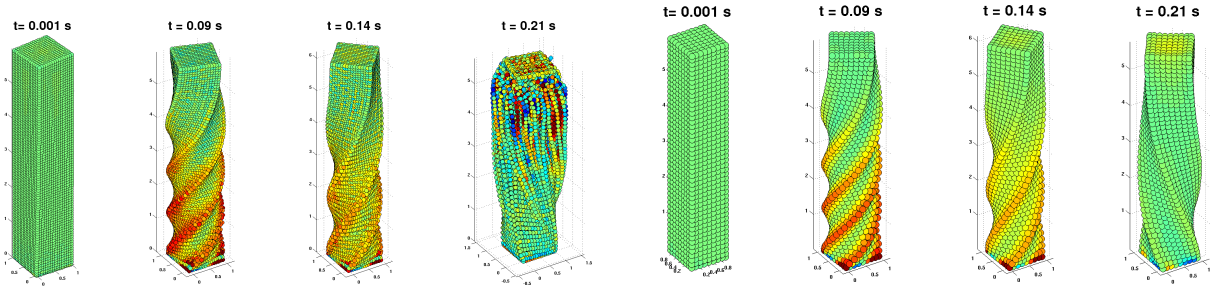



Figure 3: Bending simulation, JST-CSPH $\boldsymbol{p}$-$\boldsymbol{F}$ based model. NeoHookean material ($E = 17\,MPa$, $\nu = 0.49$, $\rho = 1100\,kg/m^3$), 3969 particles. Pressure plot captured on a cross section of the column at different simulation times.

**Twisting column** A column with the same geometric and material properties as the one in the previous example is now subjected to a sinusoidal angular velocity of $\boldsymbol{\omega} = 100 \sin\left(\frac{\pi z}{2L}\right) \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\, rad/s$

and its bottom is clamped. A comparison between fig. 4a (standard CSPH) and fig. 4b (***p*-*F*** JST-CSPH) illustrates the robustness of the mixed Total Lagrangian method: while the simulation in fig. 4a fails after a few time steps, the one in fig. 4b accurately captures the structure behaviour.




(a) standard CSPH model (b) JST-CSPH ***p*-*F*** mixed formulation model

Figure 4: Twisting column, NeoHookean material ($E = 17\,MPa$, $\nu = 0.49$, $\rho = 1100\,kg/m^3$), 3969 particles. Pressure plot captured at different simulation times.

**Punch test** A rectangular region of $3\,cm \times 1\,cm$ is punched with a constant vertical velocity of $-10\,m/s$ on one-third of the top surface. The billet is made of a NeoHookean hyperelastic material with parameters $E = 1\,MPa$, $\nu = 0.4$, $\rho = 1000\,kg/m^3$. The particle discretisation has been realised with 225 uniformly distributed particles; the CFL constant has been chosen as $\alpha_{CFL} = 0.3$. As reported in [11], the standard CSPH method [6, 7, 10] is not capable of solving this problem, which involves extremely large distortions. This deficiency can be alleviated using the proposed JST-CSPH ***p*-*F*** mixed algorithm, as demonstrated in fig. 5.

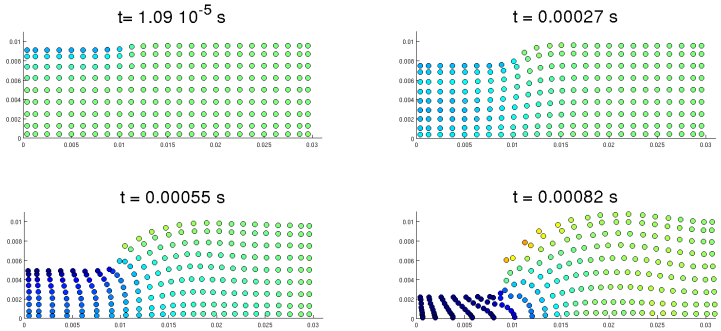



Figure 5: Pressure contour plot for a punch test on a rectangular, 2D billet, captured at various instants of simulation times. Material properties: NeoHookean hyperelastic material, $E = 1\,MPa$, $\nu = 0.4$, $\rho = 1000\,kg/m^3$. JST-CSPH ***p*-*F*** mixed formulation model, 225 particles employed.

**Tensile cube** A 3-D cube of unit side length is constrained at the bottom face and left free on the other sides. The cube is made of NeoHookean material ($E = 21\,GPa$, $\nu = 0.3$, $\rho = 7000\,kg/m^3$) and is discretised using 11 particles per edge. It is subjected to a prescribed initial sinusoidal velocity field of $\mathbf{w} = 500 \sin\left(\frac{\pi z}{2}\right) \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}\,m/s$. In the top row of images in fig. 6, JST

dissipation has not been added to the solution, and chequerboard pressure behaviour develops as the simulation advances in time. To the contrary, as can be observed in the bottom row of fig. 6, the use of JST-CSPH $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation with $\kappa^{(4)} = 1/1024$ eliminates the appearance of non-physical pressure fluctuations in near incompressibility regime, stabilising the solution.

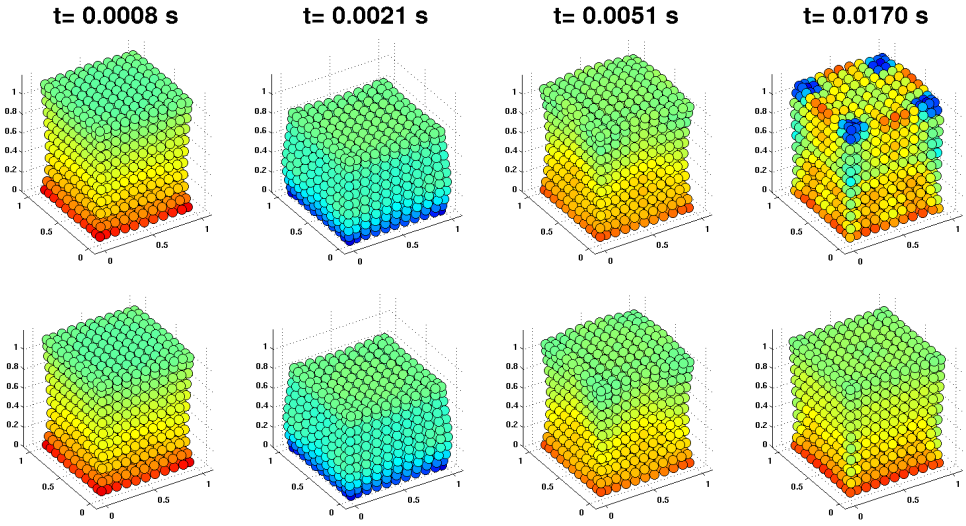



Figure 6: Tensile cube, (top) CSPH $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation, (bottom) JST-CSPH $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation. NeoHookean material ($E = 21\,GPa$, $\nu = 0.3$, $\rho = 7000\,kg/m^3$), 1331 particles. Pressure plot captured at different simulation times.

## 6 CONCLUSIONS

In this work, a stabilised Total Lagrangian mixed formulation $\boldsymbol{p}$-$\boldsymbol{F}$ JST-CSPH algorithm is presented for the solution of large strain solid dynamics problems. The methodology is based on a system of first order conservation laws, where the linear momentum $\boldsymbol{p}$ and the deformation gradient $\boldsymbol{F}$ are regarded as the primary variables. The well known SPH particle formulation has been modified to accommodate corrections to the kernel interpolating function, resulting in improved behaviour at or near the boundaries of the problem. A nodally conservative JST stabilisation is incorporated into the CSPH $\boldsymbol{p}$-$\boldsymbol{F}$ mixed formulation algorithm, taking advantage of the conservation laws. The obtained spatial discretisation is combined with a two step TVD Runge-Kutta algorithm for the time integration. Once the numerical model has been so defined, results from numerical simulations of benchmark problems are presented in order to demonstrate the applicability and robustness of the proposed methodology.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support provided by the Sêr Cymru National Research Network for Advanced Engineering and Materials, United Kingdom.

## REFERENCES

[1] A.J. Gil, C.H. Lee, J. Bonet and R. Ortigosa. A first order hyperbolic framework for large strain computational solid dynamics. Part II: Total Lagrangian compressible, nearly incompressible and truly incompressible elasticity. *CMAME*, **300**, 146–181, 2016.

[2] M. Aguirre, A.J. Gil, J. Bonet, A.A. Carreño. A vertex centred Finite Volume Jameson-Schmidt-Turkel algorithm for a mixed conservation formulation in solid dynamics. *JCP*, **259**, 672–699, 2014.

[3] C.H. Lee, A.J. Gil, J. Bonet. Development of a stabilised Petrov-Galerkin formulation for conservation laws formulation in Lagrangian fast solid dynamics. *CMAME*, **268**, 40–64, 2014.

[4] C.H. Lee, A.J. Gil and J. Bonet. Development of a cell centred upwind finite volume algorithm for a new conservation law formulation in structural dynamics. *C&S*, **118**, 13–38, 2013.

[5] I.A. Karim, C.H. Lee, A.J. Gil and J. Bonet. A Two-Step Taylor Galerkin formulation for fast dynamics. *EC*, **31**, 366–387, 2014.

[6] J. Bonet, S. Kulasegaram. Correction and stabilization of smooth particle hydrodynamics methods with applications in metal forming simulations. *IJNME*, **47**, 1189–1214, 2000.

[7] J. Bonet, S. Kulasegaram. Remarks on tension instability of Eulerian and Lagrangian Corrected Smooth Particle Hydrodynamics (CSPH) methods. *IJNME*, **52**, 1203–1220, 2001.

[8] A. Jameson, W.Schmidt, E. Turkel. Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time-Stepping Schemes. *AIAA 14th Fluid and Plasma Dynamic Conference*, AIAA paper **81**-1259, Palo Alto, USA, June 1981.

[9] J.J. Monaghan. Smoothed particle hydrodynamics. *ARA&A*, **30**, 543–574, 1992.

[10] J. Bonet, T.-S.L. Lok. Variational and momentum preservation aspects of Smooth Particle Hydrodynamic formulations. *CMAME*, **180**, 97–115, 1999.

[11] Y. Vidal, J. Bonet and A. Huerta. Stabilized updated Lagrangian corrected SPH for explicit dynamic problems. *IJNME*, **69**, 2687–2710, 2007.

[12] L.D. Libersky, A.G. Petschek, T.C. Carney, J.R. Hipp, F.A. Allahdadi. High strain Lagrangian hydrodynamics. *JCP*, **109**, 67–75, 1993.

[13] J.W. Swegle, D.L. Hicks, S.W. Attaway. Smoothed particle hydrodynamics stability analysis. *JCP*, **116**, 123–134, 1995.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# BLOOD FLOW SIMULATION USING SMOOTHED PARTICLE HYDRODYNAMICS

**Mohammed K. AL-SAAD[1], Sivakumar Kulasegaram[2]** and **Stephane P.A. Bordas[2]**

[1]School of Engineering, Cardiff University
Cardiff University, Queen's Buildings, the Parade, CARDIFF
CF24AA, Wales, UK
Al-SaadMK@Cardiff.ac.uk

[2]School of Engineering, Cardiff University
Cardiff University, Queen's Buildings, the Parade, CARDIFF
CF24AA, Wales, UK
KulasegaramS@cardiff.ac.uk, stephane.bordas@alum.northwestern.edu,

**Keywords:** smooth particle hydrodynamics (SPH); blood flow; thrombus; arteries

**ABSTRACT:** *To understand the characteristics of blood flow, it is important to identify key parameters that influence the flow of blood. The characterisation of blood flow will also enable us to understand the flow parameters associated with physiological conditions such as atherosclerosis. Thrombosis plays a crucial role in atherosclerosis and it also helps to stop bleeding when a blood vessel is injured. This article focuses on using a meshless particle-based Lagrangian numerical technique, named the smoothed particles hydrodynamic (SPH) method, to study the flow behaviour of blood and to explore the flow conditions that induce the formation of thrombus in a blood vessel. Due to its simplicity and effectiveness, the SPH method is employed here to simulate the process of thrombogenesis for various blood flow parameters. In the present SPH simulation, blood is modelled by particles that have the characteristics of plasma and of platelets. To simulate the coagulation of platelets which forms thrombus, the adhesion and aggregation processes of the platelets are modelled by an effective inter-particle attraction force model. With these models, the motion of platelets in flowing blood, and their adhesion and aggregation are effectively coupled with viscous blood flow. In this study, the adhesion and aggregation of blood particles are analysed on a (straight vessel) under various velocities of blood scenarios. The results are compared with experimental results, and a good agreement is found between the simulated and experimental results.*

## 1 INTRODUCTION

A thrombus is considered to be one of the most important causes of many diseases in human body. On the other hand, A blood clot anchored to a damaged vascular wall can stop bleeding or it can prevent atherosclerosis in arteries. The danger is that a thrombus can affect the blood flow in the vessels and this can cause potentially deadly accidents, such as cardiac infarction (or heart attack) or ischemic stroke when the damage occurs in the coronary or the carotid arteries, respectively[1, 2]. The formation of a thrombus depends on platelet flow; for example, the transport to denuded subendothelium, formation of membrane tethers, adhesion to the subendothelium, and aggregation[3]. Many experimental studies have provided information on the biochemical effects of fluid forces on thrombogenesis. In recent years, due to the availability of vast computational power, research on computer simulation of thrombosis has become a field of deep interest. Although fluids can be simulated in either the Eulerian or Lagrangian method, the Lagrangian method is considered to be more suitable for this type of simulation due to their obvious advantages in tracking movement of particles similar to platelets[4]. The purpose of this study is to analyse flow parameters that influence the formation of thrombosis inside arteries. A Lagrangian smoothed particles hydrodynamics (SPH) is used for numerical simulations of the blood flow consisting plasma and platelets.

## 2 NUMERICAL METHODOLOGY

The governing equations for solving incompressible or weakly compressible isothermal fluid flow using SPH are mass and momentum conservation equations given by,

$$\frac{1}{\rho}\frac{D\rho}{Dt} + \nabla.\mathbf{v} = 0 \tag{1}$$

$$\frac{D\mathbf{v}}{Dt} = -\frac{1}{\rho}\nabla \mathrm{p} + v\nabla^2\mathbf{v} + \mathbf{F} \tag{2}$$

where $\rho$, $t$, $v$, $\mathbf{v}$, and p represent the density, time, kinematic viscosity, velocity and pressure of the fluid particles and, **F** represents the external force acting on fluid particles. The fluid pressure for weakly compressible SPH formulation is obtained by an equation of state as presented in [5]**.** The numerical procedure to calculate fluid velocity is derived from the momentum equation (2) as,

$$\mathbf{v}^{n+1} = \mathbf{v}^n + \left(-\frac{1}{\rho}\nabla P + v\nabla^2\mathbf{v}^n + \mathbf{F}\right)\Delta t \tag{3}$$

Where superscript $n$ and $n+1$ refer to current and next time steps, respectively, and $\Delta t$ is the numerical time step. The position and density of the fluid can be updated respectively at every time step by,

$$\mathbf{x}^{n+1} = \mathbf{x} + \mathbf{v}^{n+1}\Delta t, \tag{4}$$

And (from the continuity equation (1)),

$$\rho^{n+1} = \rho^n - \rho^n(\nabla.\mathbf{v}^{n+1})\Delta t. \tag{5}$$

The pressure is then estimated from the updated density.

## 3 MODELLING PLATELET MOTION

The platelets tend to adhere and aggregate when the blood vessel is damaged. This can lead to formation of a primary thrombus. Inside the primary thrombus the neighbour platelets link together, which are then bound by vWF fibrinogen in plasma and collagen in the sub-endothelial tissue [6]. This process takes place by making a link between neighbouring platelets and bound by vWF fibrinogen in plasma and collagen in the sub-endothelial tissue. To numerically model such platelet motion, an algorithm based on a penalty or spring force mechanism [7] is adopted. This model dictates the interactions between platelets and plasma inside the blood vessel. When the platelets are within a distance $d_{ad}$ from the damaged area, the platelets are attracted towards the damaged wall by an adhesive force given by eq. (6). The platelets adhering to the wall are then activated and attract other platelets which are within a distance of $d_{ag}$ from them. This attractive force is called an aggregation force which is given by eq. (7). The aggregation force takes the same form as that of the adhesive force but has a different spring constant.

$$\boldsymbol{F}_{ad} = \begin{cases} K_{ad}(|\boldsymbol{r}_{ij}| - r_o)\boldsymbol{n}_{ij} & (|\boldsymbol{r}_{ij}| \le d_{ad}) \\ 0 & (|\boldsymbol{r}_{ij}| > d_{ad}) \end{cases} \tag{6}$$

$$\boldsymbol{F}_{ag} = \begin{cases} K_{ag}(|\boldsymbol{r}_{ij}| - r_o)\boldsymbol{n}_{ij} & (|\boldsymbol{r}_{ij}| \le d_{ag}) \\ 0 & (|\boldsymbol{r}_{ij}| > d_{ag}) \end{cases} \tag{7}$$

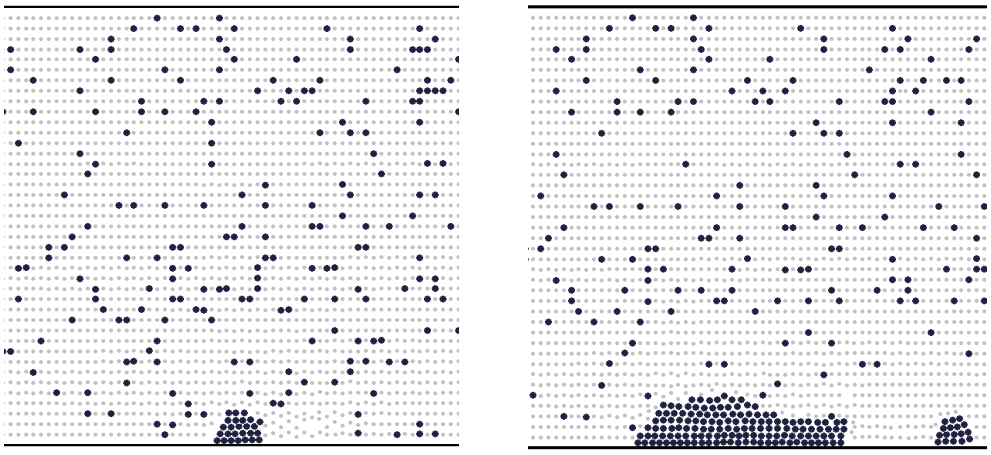
In the above equations $\boldsymbol{F}_{ad}, \boldsymbol{F}_{ag}$ are the adhesive and aggregate forces and $K_{ad}, K_{ag}$ are the corresponding spring constants. The $\boldsymbol{r}_{ij}$ here is distance between activated platelet and vessel wall (or other non-activated platelets), $r_o$ is the original or natural length of the spring and $\boldsymbol{n}_{ij}$ is a unit vector linking platelet and damaged wall (or linking activated platelet and other surrounding platelets). These two forces are introduced in equation (2) for platelet particles which are influenced by adhesion and aggregation.

## 4 BLOOD FLOW MODEL

In this work, the blood flow simulations were performed inside a straight blood vessel with velocity range of blood flow between100-700 $\mu m/s$ defined at the inlet of vessel. The total length of the vessel ($L$) and the width between two walls ($D$) are respectively $130\mu$m and 40 $\mu$m. The dimensions of the damaged wall ($L_i$) is $30\mu$m (refer to the length of the wall damage) and the distance from the inlet to the damaged wall ($L_o$) is $40\mu$m (see Fig.1). The total number of particles used in the simulation was 5371. Four layers of boundary dummy particles were also used. The initial distance between particles is 1.0 $\mu$m. The density $\rho$ and kinematic viscosity $v$ of the plasma and platelets, were set as $\rho = 1\text{x}10^3 \text{kg/m}^3$ and $v = 1\text{x}10^{-6}\text{m}^2/\text{s}$. The boundary conditions were; a uniform velocity at the inlet, zero pressure at the outlet and, non-slip condition at the walls enforced by dummy boundary particles. The amount of the platelet particles used in the simulation is approximately 8.8% of the plasma to resemble normal physiological condition. The time step was set to $5\text{x}10^{-7}\text{s}$ to ensure the stability of numerical integration scheme. In the reported numerical simulations, the spring constants $K_{ad}$ and the $K_{ag}$ are $9.0\text{x}10^9\, N/\text{m}$ and $4.5\text{x}10^9\, N/m$ respectively, while $d_{ad}$= 3.0μm = $d_{ag}$, and $r_o$ = 2.0μm.

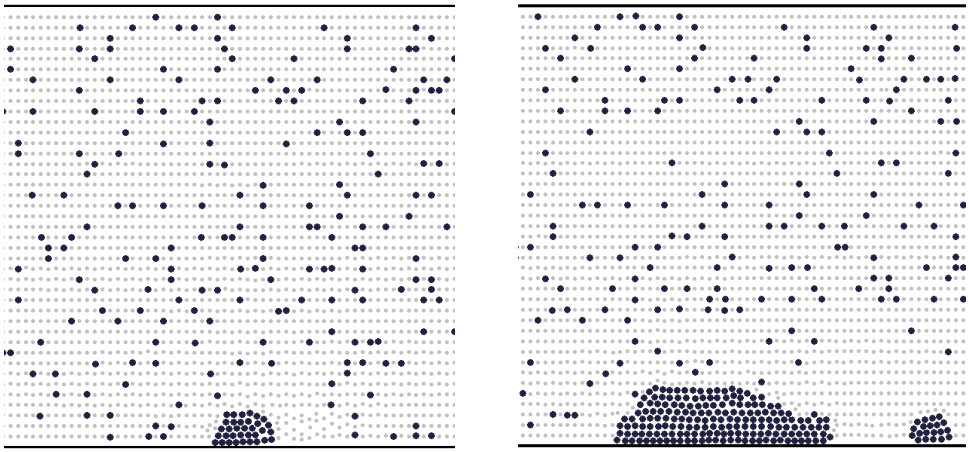
## 5 NUMERICAL RESULTS

The purpose of this study is to demonstrate the formation of thrombus and to investigate the applicability of SPH in modelling such process. The corrected SPH is used to improve the accuracy [8] of the simulation. Normally, a thrombus is formed by adhesion and aggregation of platelets which are transported by the blood flow in different geometries of arteries or vessels, where the growth rate of thrombus formation varies with the stenosis and the flow rate of blood. Figure 1 illustrates the formation of thrombus at two different stages of the flow. In these figures, for clarity, plasma and platelet particles are shown by two different the plasma and platelets are denoted by light and the dark grey respectively. The platelets are activated when they are within $d_{ad}$ distance from the damaged region and form a primary thrombus. During the course of time, a primary thrombus is developed to cover the whole damage area by forming several layers of platelets. When thrombus grows to a certain volume, part of the thrombus is separated and transported downstream by the blood flow. Figures 1, 2 and 3 depict the growth of thrombus at different times for velocities 100 ,500 and 700 $\mu m/s$ of the blood flow. From the figures below, various stages of thrombus growth on the damaged area of the wall are clearly evident. It can be noted from Fig. 1(b), 2(b) and 3(b) that, part of the thrombus is separated from primary thrombus once the primary thrombus grows to a substantial volume. It is interesting to observe that the volume of the primary thrombus and the time at which separation of the thrombus takes place are affected by the flow rate. From these figures it can also be noted that, when the flow rate was 700 $\mu m/s$ the thrombus growth was thinner compare to the cases where the blood flow rate was 100 and 500$\mu m/s$. Further, it was observed that with higher flow rates the separation of thrombus takes place quicker. In Figure 4, the growth rate of thrombus against various blood flow velocities are plotted. It transpires from Fig. 4 that the growth rate of thrombus gradually increases with blood velocity until approximately 500 $\mu m/s$. Beyond 500 $\mu m/s$, the thrombus growth rate drops to a lower level. The results illustrated in Figure 4 qualitatively agree with experimental observation made in [9]. As results reported here are from 2-dimensional simulations, direct comparisons could not be made at this point. It is evident from all the results listed below that the blood flow rate plays a crucial role in the build-up and separation of thrombus. The results show that the growth rate of the thrombus, its thickness, and formation/separation vary according to the blood flow rate and these results are consistent with experimental observations reported in [9].



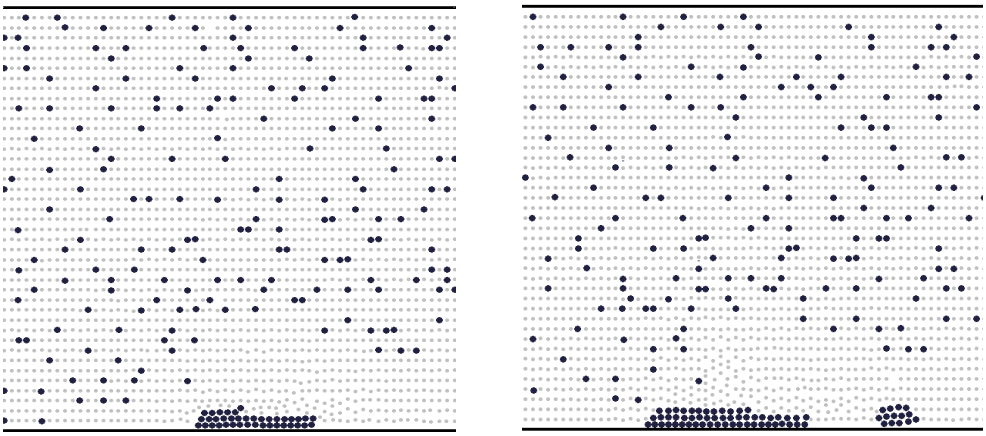
(a) (b)

Figure1: The platelet aggregation in the velocity=100 $\mu m/s$ at (a) t=0.2s; (b) t=0.6s



(a) (b)

Figure2: The platelet aggregation in the velocity=500 μm/s at (a) t=0.2s; (b) t=0.6s



(a) (b)

Figure 3: The platelet aggregation in the velocity=700 μm/s at (a) t=0.2s; (b) t=0.6s

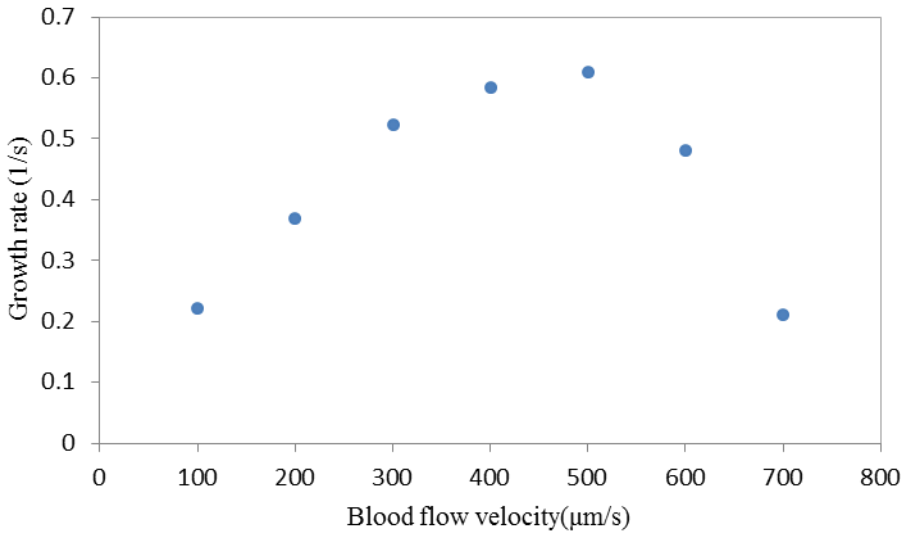



Figure 4: Effect of blood flow velocity with thrombus growth rate

## 7 CONCLUSIONS

This work has focused on the simulation of the thrombogenesis process using the SPH method by considering platelet adhesion/aggregation and the influence of blood flow rates on thrombus growth. In the numerical simulations, blood inside a straight vessel is discretised by particles which are assumed to have the characteristics of blood constituents, such as plasma and platelets. The platelet adhesion and aggregation process during the blood flow is modelled by adopting an inter-particle penalty force method. The aforementioned model proved efficient in simulating adhesion and aggregation process without rigorous computational efforts. The potential of SPH method to simulate thrombogenisis process is demonstrated via numerical examples. The numerical simulations were able to indicate how blood flow velocity influenced thrombus growth rate in a straight vessel. Further, the numerical results also qualitatively agree with experimental observations reported in literature. This study also demonstrates the ability and accuracy of the SPH method in modelling blood flow with low Reynolds numbers.

It is essential to further investigate the accuracy of the methods developed here in a 3-dimensional context. The 3-dimensional simulations will enable the results to be directly and quantitatively compared with the experimental observations reported in [9]. In addition, the assumed penalty values (or spring constants) used in the analysis discussed above can be more accurately estimated or calibrated by 3-dimensional numerical models.

## 8 REFERENCES


[1] Epstein, F.H., et al., *The pathogenesis of coronary artery disease and the acute coronary syndromes.* New England journal of medicine, 1992. **326**(4): p. 242-250.

[2] Fuster, V., et al., *Insights into the pathogenesis of acute ischemic syndromes.* Circulation, 1988. **77**(6): p. 1213-1220.

[3] Ruggeri, Z., *Von Willebrand factor, platelets and endothelial cell interactions.* Journal of Thrombosis and Haemostasis, 2003. **1**(7): p. 1335-1342.

[4] Sulsky, D. and H.L. Schreyer, *Axisymmetric form of the material point method with applications to upsetting and Taylor impact problems.* Computer Methods in Applied Mechanics and Engineering, 1996. **139**(1): p. 409-429.

[5] Batchelor, G., *K. 1967 An Introduction to Fluid Dynamics*. 1970, Cambridge University Press.

[6] Savage, B., E. Saldívar, and Z.M. Ruggeri, *Initiation of platelet adhesion by arrest onto fibrinogen or translocation on von Willebrand factor.* Cell, 1996. **84**(2): p. 289-297.

[7] Kamada, H., et al., *A three-dimensional particle simulation of the formation and collapse of a primary thrombus.* International Journal for Numerical Methods in Biomedical Engineering, 2010. **26**(3-4): p. 488-500.

[8] Bonet, J. and S. Kulasegaram, *Correction and stabilization of smooth particle hydrodynamics methods with applications in metal forming simulations.* International journal for numerical methods in engineering, 2000. **47**(6): p. 1189-1214.

[9] Begent, N. and G. Born, *Growth rate in vivo of platelet thrombi, produced by iontophoresis of ADP, as a function of mean blood flow velocity.* Nature, 1970. **227**: p. 926-930.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE VALIDITY OF TIDAL TURBINE ARRAY CONFIGURATIONS OBTAINED FROM STEADY-STATE ADJOINT OPTIMISATION

**Christian T. Jacobs[1], Matthew D. Piggott[1], Stephan C. Kramer[1] and Simon W. Funke[2]**

[1]Department of Earth Science and Engineering, Imperial College London
London SW7 2AZ, United Kingdom
e-mail: {c.jacobs10,m.d.piggott,s.kramer}@imperial.ac.uk

[2] Biomedical Computing Department, Simula Research Laboratory
P.O. Box 134, 1325 Lysaker, Norway
e-mail: simon@simula.no

**Keywords:** Shallow Water Equations, Turbines, Adjoint Optimisation, Large Eddy Simulation.

**Abstract.** *Extracting the optimal amount of power from an array of tidal turbines requires an intricate understanding of tidal dynamics and the effects of turbine placement on the local and regional scale flow. Numerical models have contributed significantly towards this understanding, and more recently, adjoint-based modelling has been employed to optimise the positioning of the turbines in an array in an automated way and improve on simple man-made configurations (e.g. structured grids of turbines) [1].*

*Adjoint-based optimisation of high-resolution and ideally 3D transient models is generally a very computationally expensive problem. Multiple approaches are therefore used in practice to obtain feasible runtimes: using high viscosity values to obtain a steady-state solution, or a sequence of steady-state solutions for "time-varying" setups; limiting the number of adjoint computations; or reformulating the problem to allow for coarser mesh resolution to make it feasible for resources assessment (e.g. [2], [3]). However, such compromises may affect the reliability of the modelled turbines, their wakes and interactions, and thus bring into question the validity of the computed optimal turbine positions.*

*This work considers a suite of idealised simulations of flow past tidal turbine arrays in a two-dimensional channel. It compares four regular array configurations, detailed by Divett et al. [4], with the configuration found through adjoint optimisation in a steady-state, high-viscosity setup. The optimised configuration produces considerably more power than the other configurations (approximately 40% more than the best man-made configuration). The same configurations are then used to produce a suite of transient simulations that do not use constant high-viscosity, and instead use large eddy simulation (LES) to parameterise the resulting turbulent structures. All simulations are performed using OpenTidalFarm [1].*

*It is shown that the 'low background viscosity'/LES simulations produce less power than that predicted by the constant high-viscosity runs. Nevertheless, they still follow the same trends in the power curve throughout time, with optimised layouts continuing to perform significantly better than simplified configurations.*

## 1 INTRODUCTION

Tidal stream turbines are a promising new source of renewable energy. In order for the deployment of turbines to be effective and economically viable, their placement in a turbine array/farm must be carefully considered. Knowledge of the tidal flow conditions, the effect the turbines will have on the flow and any natural habitats, and how any resulting changes will affect turbine power generation downstream (e.g. through blockage effects), all need to be investigated. Experimental studies are often limited to small-scale domains with idealised flow conditions, which may not scale well to realistic marine environments. As more computational resources become available to solve bigger problems, numerical modelling and adjoint-based optimisation of tidal turbine positions is becoming an increasingly popular alternative.

Current work on adjoint-based optimisation has shown very promising results with regards to improving the efficiency of turbine farms over simple man-made configurations, such as a structured grid of turbines or staggering them in some regular way. For example, Funke et al. [1] optimised an array in the Inner Sound of the Pentland Firth, comprising 256 turbines, which resulted in a 33% increase in power extraction. Similar array optimisation studies by Culley et al. [5] and Barnett et al. [6] in idealised scenarios and once again in the Pentland Firth yielded increased power generation of up to 25%. However, despite the power that adjoint-based optimisation offers, it is extremely computationally-demanding. Such optimisation requires both the forward run and the adjoint run (i.e. the simulation run backwards in time) to be performed at every iteration of an optimisation procedure, based on numerical algorithms such as L-BFGS-B (the limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with bound constraints) [7, 8]. High-resolution transient simulations of flow past marine power turbines, particularly those that feature turbulent wakes and thus require even smaller time-steps to capture the physics, are often prohibitively expensive.

As a result, current work on the adjoint optimisation of tidal turbine placement in the literature has been mostly limited to steady-state simulations in which very high, uniform and constant, non-physical values of the background viscosity are required to ensure that a steady-state solution exists (e.g. 0.5 $\mathrm{m^2 s^{-1}}$ [2], 3 $\mathrm{m^2 s^{-1}}$ [9, 6], 30 $\mathrm{m^2 s^{-1}}$ [5, 3, 1], 90 $\mathrm{m^2 s^{-1}}$ [10]). Such simulations only require one solve of the governing forward equations (with the time derivatives removed) and one solve of the adjoint equations[1], per optimisation iteration, making the problem tractable with available resources. However, the wake downstream of each turbine is likely to be diffused out in such simulations, and no turbulence is modelled. The validity of the results from the optimisation process is therefore questionable. The size of the wake and the presence of any turbulent eddies can potentially affect the power generated by the turbines further downstream, and any blockage effects from neighbouring turbines may not be realistic. When the optimised turbine positions from the steady-state computations are applied to a realistic, transient flow comprising a much lower background viscosity and higher Reynolds number, the amount of power generated by the flow may be substantially different and the benefits of the optimisation process may not be as pronounced compared to the steady-state, high-viscosity case.

The work presented in this paper first considers four simple man-made turbine array configurations by Divett et al. [4]. Steady-state simulations are run with each of these configurations, and the configuration that provides the highest amount of generated power is then optimised using adjoint-based optimisation. The resulting optimised configuration, and also the four man-

[1]The simulation can also be divided up into $n$ steady-states, in which case $2n$ solves are required per optimisation iteration.

made configurations, are then each considered in a transient setup in which a more realistic background viscosity is used and Large Eddy Simulation (LES) is applied to parameterise the turbulence. Section 2 presents the shallow water equations that are solved by the numerical model, along with details of the Smagorinsky LES turbulence parameterisation that is enabled in the transient flow simulations. Setup details are provided in Section 3, followed by an analysis of the steady-state and transient simulation results in Section 4. The paper closes with some concluding remarks in Section 5.

## 2 NUMERICAL MODEL

This work uses the OpenTidalFarm numerical modelling package [1] for solving the shallow water equations. It uses FEniCS [11] to discretise the equations with the finite element method, and the Dolfin-Adjoint framework [12, 13] to automatically annotate the forward model system and compute the adjoint model based on that.

### 2.1 Momentum equation

The equation governing conservation of momentum is given by

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -g\nabla\eta + \nabla \cdot \mathbb{T} - \frac{c_b + c_t}{H}\|\mathbf{u}\|\mathbf{u}, \tag{1}$$

where $\mathbf{u}$ is the depth-averaged velocity, $g$ is the acceleration due to gravity, $\eta$ is the free-surface displacement, $h$ is the free-surface height at rest, $H$ is the total free-surface height (i.e. $h + \eta$), $c_b$ is the (dimensionless) bottom drag coefficient, and $c_t$ is the (dimensionless) tidal turbine drag coefficient. The Euclidean norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ is used here. The stress tensor $\mathbb{T}$ is given by

$$\mathbb{T} = \nu \left( \nabla \mathbf{u} + \nabla \mathbf{u}^{\mathrm{T}} \right), \tag{2}$$

where $\nu$ is the kinematic viscosity.

Each of the $N$ turbines is represented by a two-dimensional Gaussian profile, such that for a turbine $i$ of radius $r_i$ centred at $(x_i, y_i)$,

$$c_t = \sum_{i=1}^{N} K_i \phi_{x_i,r_i}(x)\phi_{y_i,r_i}(y), \tag{3}$$

where $K_i$ is a dimensionless friction coefficient and

$$\phi_{p,r}(x) = \begin{cases} \exp(1 - 1/(1 - |\frac{x-p}{r}|^2)) & \text{if } |\frac{x-p}{r}| < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

### 2.2 Continuity equation

The continuity equation is given by

$$\frac{\partial \eta}{\partial t} + \nabla \cdot (H\mathbf{u}) = 0. \tag{5}$$

### 2.3 Turbulence parameterisation

Large eddy simulation (LES) is used to parameterise the turbulence generated in the transient flow simulations with a low background viscosity. The crux of LES is to resolve the large-scale eddies by using a fine enough mesh spacing, whilst modelling any smaller, sub-grid scale eddies

with an eddy viscosity term which acts as a diffusive agent; it essentially describes the energy transfer from the larger (resolved) eddies down to the smaller eddies. The particular LES model employed in this work is the standard Smagorinsky model [14, 15]. The eddy viscosity term is given by

$$\nu' = (c_s \Delta)^2 \, |\mathbb{S}|, \tag{6}$$

where $c_s$ is the Smagorinsky coefficient, set to 0.2 in this work (within the range of typical $c_s$ values [16]). The filter width $\Delta$ is an estimate of the local element size which is defined here as the square root of the area of each element. The strain rate tensor $\mathbb{S}$ is defined as

$$\mathbb{S} = \frac{1}{2} \left( \nabla \mathbf{u} + \nabla \mathbf{u}^{\mathrm{T}} \right), \tag{7}$$

The modulus of $\mathbb{S}$ is given by

$$|\mathbb{S}| = \sqrt{2 \sum_i \sum_j \mathbb{S}_{ij} \mathbb{S}_{ij}}. \tag{8}$$

where $\mathbb{S}_{ij}$ is the $(i,j)$-th component of $\mathbb{S}$. The eddy viscosity $\nu'$ is added to the background viscosity $\nu$ in (2) and is updated at the beginning of each time-step.

Turbulence is a truly three-dimensional phenomenon. Furthermore, LES requires a substantial amount of resolution to sufficiently resolve the large-scale eddies. The validity of applying an LES model in this work to a two-dimensional domain is therefore questionable. Nevertheless, 2D turbulence modelling in the context of tidal turbine simulation has shown good agreement with experimental data [17], and similar LES schemes such as MILES have generated promising results in resolving turbine wakes [18, 19].

## 3 SIMULATION SETUP

All the simulations presented here are two-dimensional and were performed in a $3 \times 1$ km$^2$ rectangular domain (in the $x$-$y$ plane), following a similar setup to the work by Divett et al. [4]. A total of 15 turbines were deployed in the four man-made configurations as described in [4], and a suite of steady-state shallow water simulations was first performed for each of the configurations. The physical parameters are provided in Table 1.

### 3.1 Mesh

The computational mesh that discretised the $3 \times 1$ km$^2$ domain, comprising triangular elements, was generated using Gmsh [20]. The inner region defined by $500 \leq x \leq 2{,}500$ m and $125 \leq y \leq 875$ m, where all the turbines are situated, contained a structured grid of solution nodes with a characteristic element length of $\Delta x = 2$ m. The outer section of the domain's mesh, which is outside the region of interest and therefore does not require high numerical resolution, was unstructured with $\Delta x = 100$ m.

### 3.2 Initial and boundary conditions

The initial conditions $\mathbf{u}(x, t = 0) = [0, 0]^{\mathrm{T}}$ ms$^{-1}$ and $\eta(x, t = 0) = 0$ m were applied at $t = 0$ s. Throughout the simulations a Dirichlet velocity boundary condition of 2 ms$^{-1}$ was strongly imposed to allow inflow from the left. For the steady-state simulations, outflow was

| Parameter | Value |
|---|---|
| Free-surface height at rest | $h$ = 50 m |
| Kinematic viscosity (steady-state) | $\eta$ = 1 m$^2$s$^{-1}$ |
| Kinematic viscosity (transient) | $\eta$ = $10^{-6}$ m$^2$s$^{-1}$ |
| Smagorinsky coefficient | $c_s$ = 0.2 |
| Gravitational acceleration | $g$ = 9.81 ms$^{-2}$ |
| Water density | $\rho$ = 1,000 kgm$^{-3}$ |
| Bottom drag coefficient | $c_b$ = 0.0025 |
| Number of turbines | $N$ = 15 |
| Turbine radii | $r_i$ = 10 m $\forall$ $i$ = 1, . . ., $N$ |
| Turbine friction coefficient | $K_i$ = 12 $\forall$ $i$ = 1, . . ., $N$ |

Table 1: The common parameter values used in all simulations. The Smagorinsky coefficient $c_s$ is only required in the transient simulations where the Smagorinsky LES model is used. Note also the difference in the background kinematic viscosity between the steady-state and transient simulations.

modelled by a zero-value Dirichlet free-surface condition strongly imposed along the right-hand boundary. For the transient simulations, a Flather [21] condition was used to minimise spurious reflections at the outlet. No-normal flow conditions were weakly imposed along the lateral walls.

### 3.3 Spatial and temporal discretisation

The Galerkin finite element method was used to discretise the model equations in space. Continuous piecewise quadratic Lagrange basis functions were used to represent the velocity solution field, whereas continuous piecewise linear Lagrange basis functions were used for the free-surface $\eta$ (thereby forming the P2-P1, or Taylor-Hood, element pair [22]) as well as all other fields (including $\nu$ and $\nu'$).

Unlike the steady-state simulations, the transient simulations required temporal discretisation due to the presence of time derivatives. The implicit backward Euler method was used for this purpose so that larger time-steps could be taken. All transient simulations were run until $T$ = 2,000 s with a time-step of $\Delta t$ = 4 s, which gave ample opportunity for the turbulent flow to become fully developed and for any spurious reflections from the outflow to dissipate. Newton iteration dealt with the non-linearity introduced through the advection term in (1); for each time-step, $k$ iterations were performed such that $\|\mathbf{u}^{k+1} - \mathbf{u}^k\| \leq 10^{-9}$, where $k$ is typically between 3 and 10. Once both (1) and (5) are discretised and assembled to form a fully block-coupled system, the system was solved directly with LU decomposition.

### 3.4 Man-made turbine configurations

#### 3.4.1 Centred

As per the work in [4], the centred configuration illustrated in Figure 1 considered a regular $5 \times 3$ grid of turbines positioned in the centre of the channel. The turbines in each row are $7.5d$ away from each other, and each row is $10d$ apart, where $d$=$2r$ is the turbine diameter.

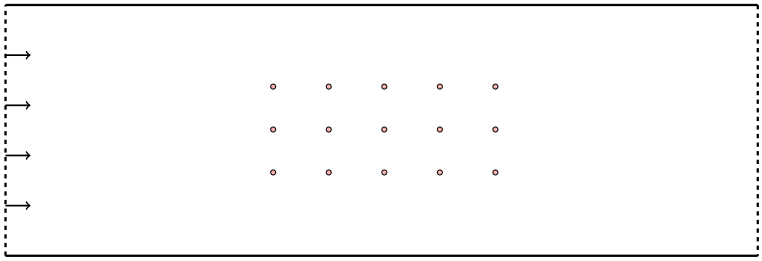


Figure 1: The centred turbine configuration. The whole $3 \times 1$ km$^2$ channel is shown.

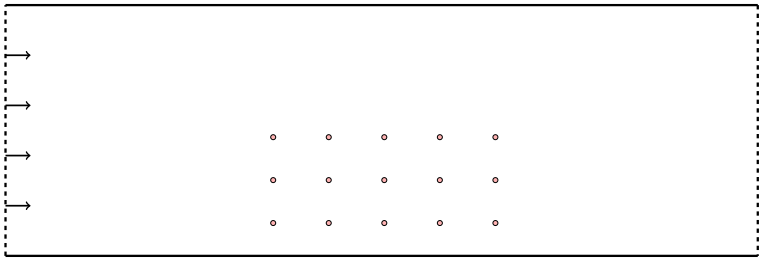


Figure 2: The offset turbine configuration. The whole $3 \times 1$ km$^2$ channel is shown.

### 3.4.2 Offset

The offset case (see Figure 2) considers the same regular grid of turbines as the centred case, but instead of being placed in the centre of the domain, the grid is positioned closer to one of the side walls. Such a configuration may be required to, for example, allow cargo to be transported via ship down one side of the channel [4].

### 3.4.3 Staggered

This is similar to the centred configuration, but each row is staggered such that the turbines in one row are aligned between the turbines of the adjacent row. This is illustrated in Figure 3.

### 3.4.4 Larger spacing

In the larger spacing case (see Figure 4), the spacing between each turbine in a given row remains the same as the centred case, but the spacing between each row is now $20d$ instead of $10d$.

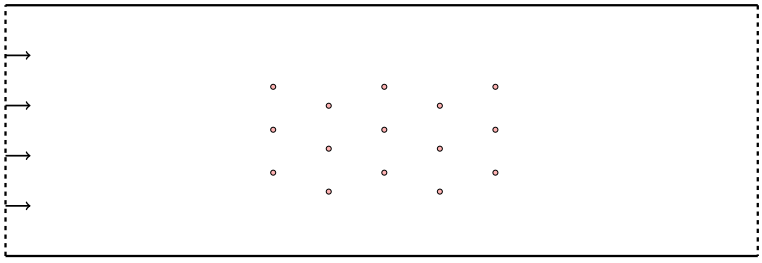


Figure 3: The staggered turbine configuration. The whole $3 \times 1$ km$^2$ channel is shown.

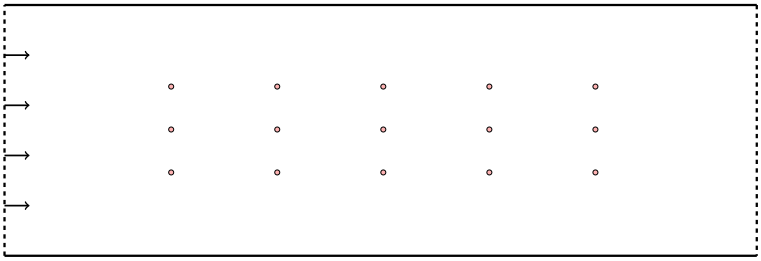


Figure 4: The larger spacing turbine configuration. The whole $3 \times 1$ km$^2$ channel is shown.

### 3.5 Adjoint optimisation

Adjoint-based optimisation was employed to maximise the amount of power extracted by the turbines with respect to the individual turbine positions. Such optimisation of high-resolution transient simulations is computationally expensive. Therefore, in order for the adjoint problem to be feasibly solved with available resources, only the steady-state problem was optimised in this work.

The Dolfin-Adjoint library [12, 13] was used to annotate the governing equations and automatically differentiate the forward model with respect to the optimisation variables, chosen to be the individual turbine positions here. The adjoint system was solved to obtain gradient information, which was then passed to the L-BFGS-B algorithm [7, 8] to optimise the functional describing the total power extracted by the turbine array, defined by

$$P(\mathbf{u}, \mathbf{m}) = \int_{\Omega} \rho c_t \|\mathbf{u}\|^3 \, \mathrm{d}x, \tag{9}$$

where $\rho$ is the density of water (1,000 kgm$^{-3}$), $\mathbf{m}$ is the vector of optimisation variables (in this case, the coordinates of each turbine), and $\Omega$ denotes the whole domain [1].

The staggered array configuration was used as the initial condition for the optimisation process, since this was found to be the best man-made configuration (see Section 4). The optimisation was constrained such that turbines must be at least $1.5d$ apart, where $d$ is the turbine diameter, and can only be placed in the region defined by $500 \leq x \leq 2{,}500$ m and $125 \leq y \leq 875$ m (i.e. the area of highest mesh resolution). The optimisation was terminated after 100 iterations, resulting in the configuration shown in Figure 5.

## 4 RESULTS

### 4.1 Steady-state simulations

The results from the steady-state simulations in Figure 6 show that the centred array was the poorest choice of configuration for this setup. The power extracted by such a setup was limited because, apart from the turbines in the first row, all the turbines downstream were situated in the wakes/separated flow regions of the ones upstream (as shown in Figure 7). By definition there is a velocity deficit in these regions and thus, from equation (9), the power extracted was reduced. When a larger turbine spacing was used, the wake was able to recover before interacting with the downstream turbines (as shown in Figure 8), and thus 23% more power could be extracted from the flow. By offsetting the centred array configuration such that the turbines are closer to a side wall, the power extraction was increased, albeit fairly insignificantly, relative to the centred configuration. The staggered array configuration yielded the best power extraction potential,

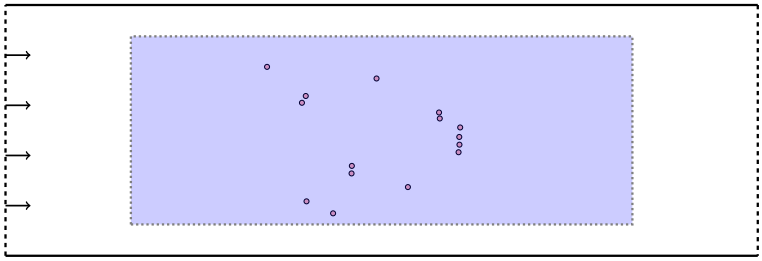


Figure 5: The optimised turbine configuration from the steady-state simulation after 100 optimisation iterations. The whole $3 \times 1$ km$^2$ channel is shown. The optimisation process was constrained such that turbines were not allowed to be placed outside of the dotted region shaded in blue. For the steady channel flow considered here a symmetric configuration would be expected, but this is not the case since the optimisation process was terminated after 100 iterations and had not fully converged (see Figure 10). Asymmetry may also be expected if the optimisation gets stuck in a local minimum.

generating 75% more power than the centred configuration. As Figure 9 shows, the accelerated region of flow around the first row of turbines interacts with the second (staggered) row behind it, resulting in more energy being removed from the flow and thus more power extraction. Each of these findings agrees qualitatively with the transient simulations of [4] which showed that, relative to the centred array, only a 4% increase in power extraction could be obtained with the offset array, whereas a 31% and 54% increase could be obtained with the larger-spaced and staggered configurations, respectively.

Since the staggered array yielded the most power, it was used as the initial condition and improved on throughout the adjoint optimisation process. This yielded a configuration that could extract 48.12 MW of power — a 156% increase relative to the centred array, as shown in Figure 10. This is greater than any difference in power extraction between the simple man-made configurations. Relative to the staggered array (i.e. the optimisation's initial condition and the best man-made configuration considered here), 46% more power could be extracted.

Flow around a turbine results in acceleration around its edges, but also in a velocity deficit immediately behind the turbine (i.e. in its wake). It would be unwise to place a turbine in the wake of another since the power is related to the velocity. It therefore makes sense that the optimisation process re-positioned the turbines such that the effect of wakes from other turbines is minimised. Furthermore, regions of accelerated flow can be used to harness more power which is why some turbines are packed closer together in the cross-flow direction (e.g. the row of four turbines on the right-hand side of Figure 5). This is highlighted in Figure 11.

### 4.2 Transient simulations

Unlike the steady-state simulations which featured highly dissipated wakes due to the presence of a uniformly high background viscosity, the transient simulations exhibited longer, sharper regions of separation with some downstream turbulent eddy shedding as Figure 12 demonstrates. From the power profiles in Figure 6, the oscillations caused by turbulent eddies were relatively small. However, considerably less power was extracted compared to the steady-state setups, after the initial spin-up period; the initial peak was caused by the inflow condition of 2 ms$^{-1}$ superimposing itself with a region of 2 ms$^{-1}$ outflow generated by the Flather boundary condition, which eventually exited the domain.

With respect to the differences in power between the configurations, qualitatively similar behaviour was observed. For example, in the centred configuration the turbine rows were still

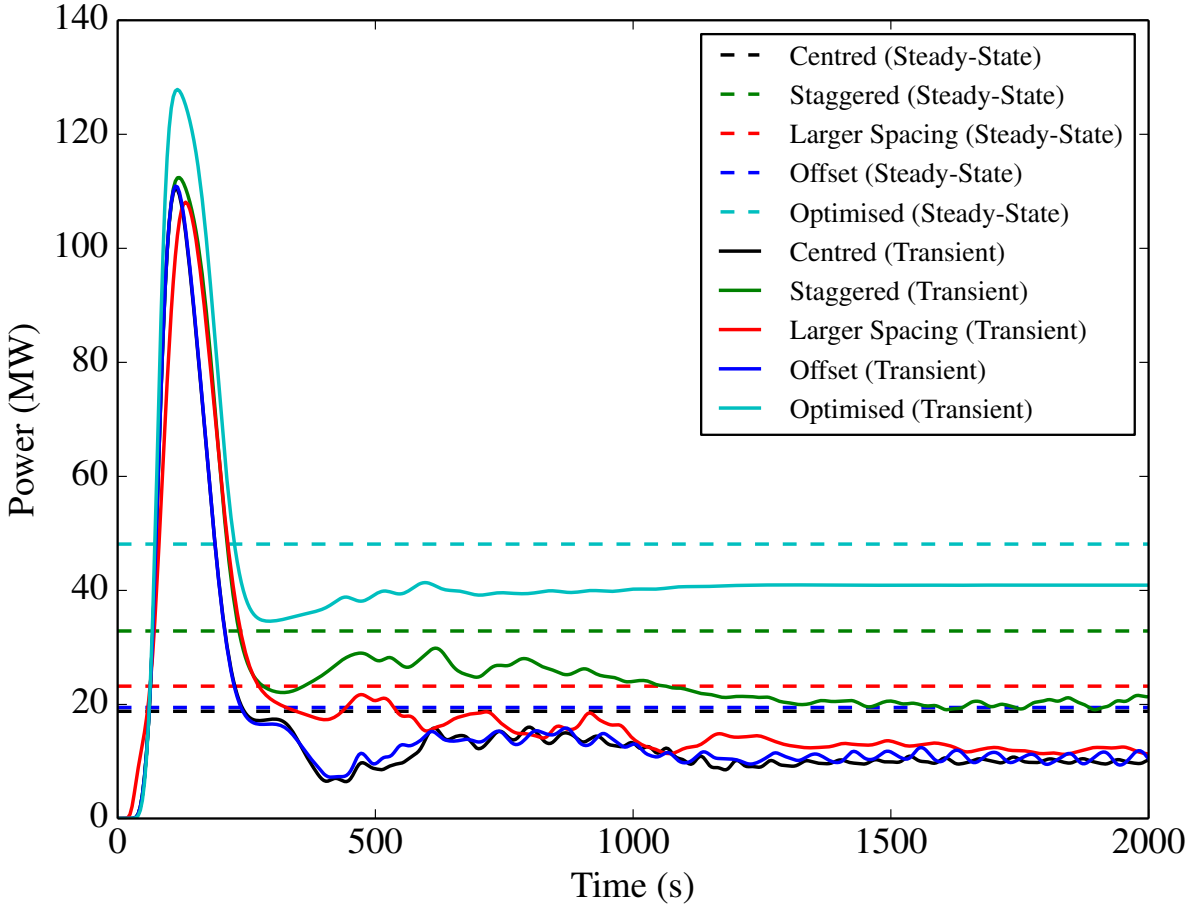



Figure 6: The total power generated against time, for each of the steady-state and transient simulations.

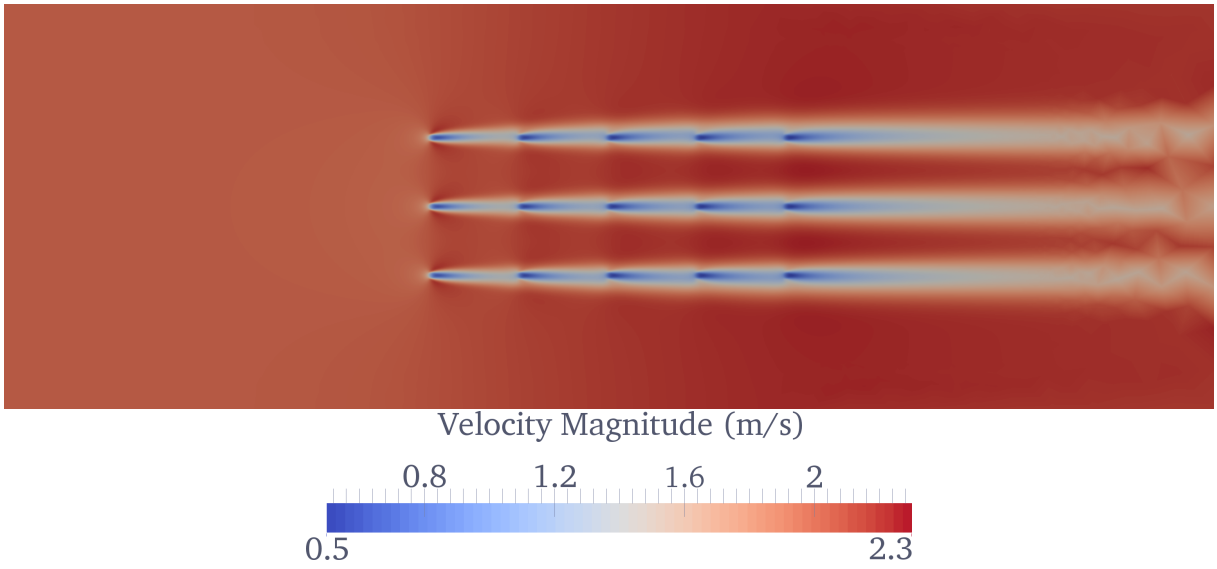



Figure 7: The flow speed around the centred turbine array in the steady-state simulation.

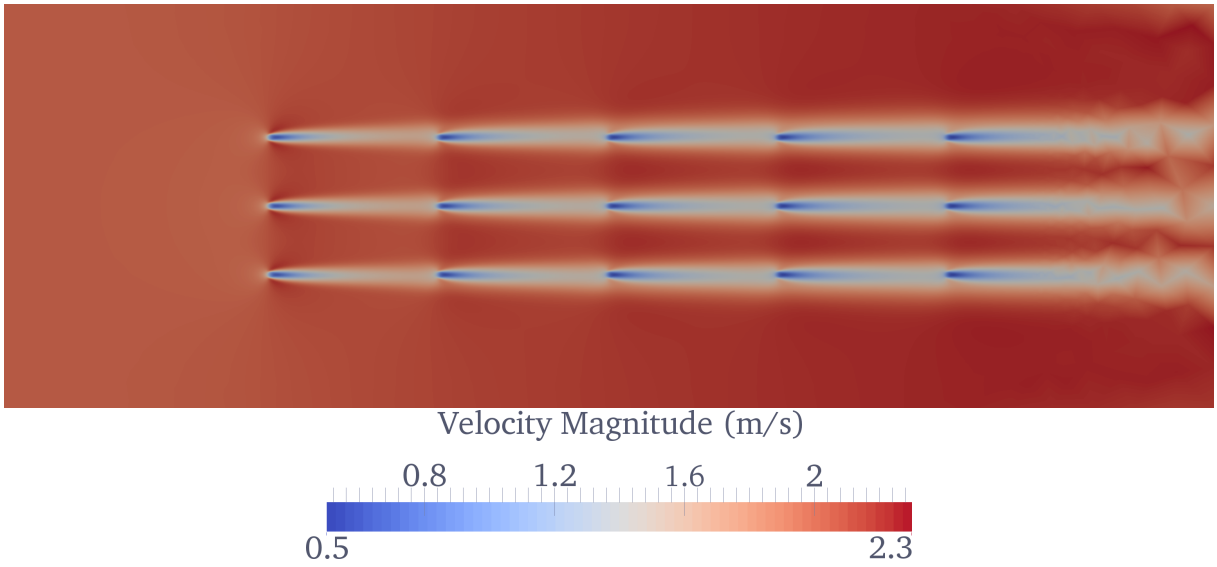



Figure 8: The flow speed around the array with larger turbine spacing in the steady-state simulation.

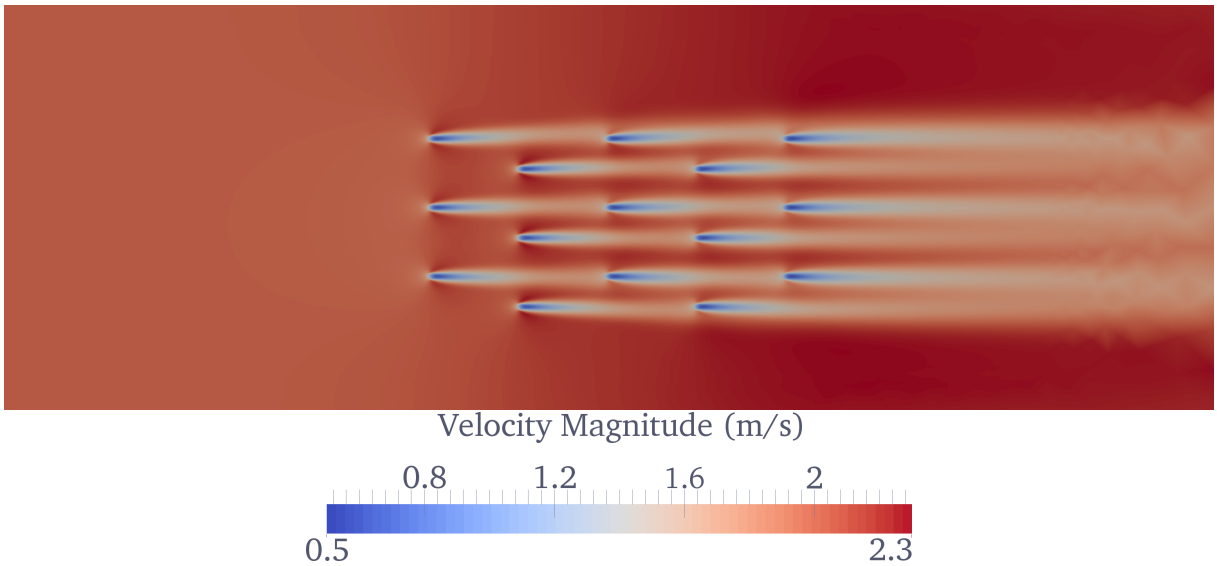



Figure 9: The flow speed around the staggered turbine array in the steady-state simulation.

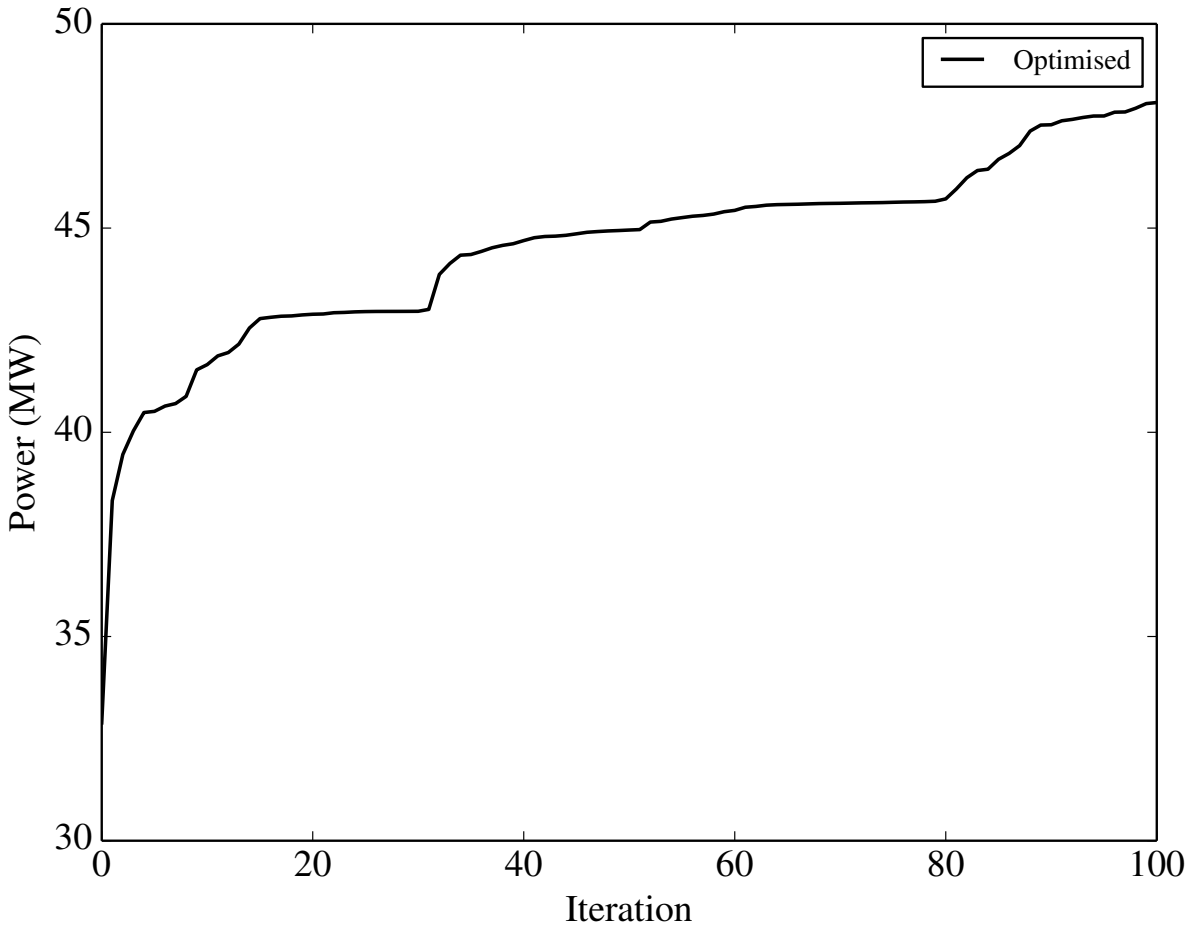



Figure 10: Total power generated at each adjoint optimisation iteration. Iteration 0 represents the power from the initial staggered configuration.

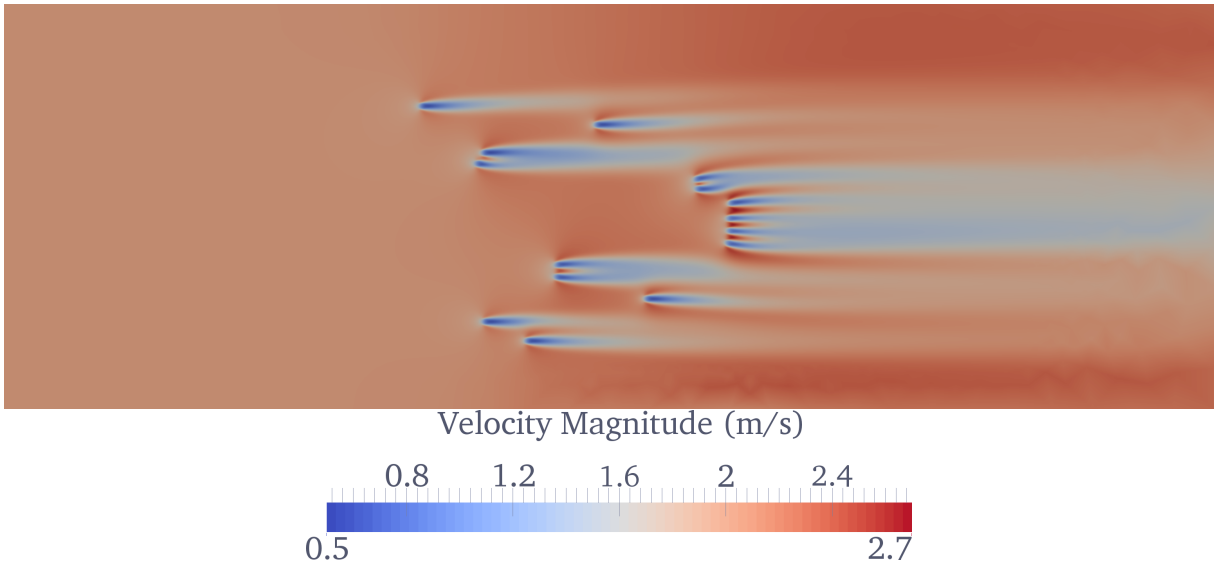



Figure 11: The flow speed around the optimised turbine array in the steady-state simulation. Note the increased upper limit of the colour bar.

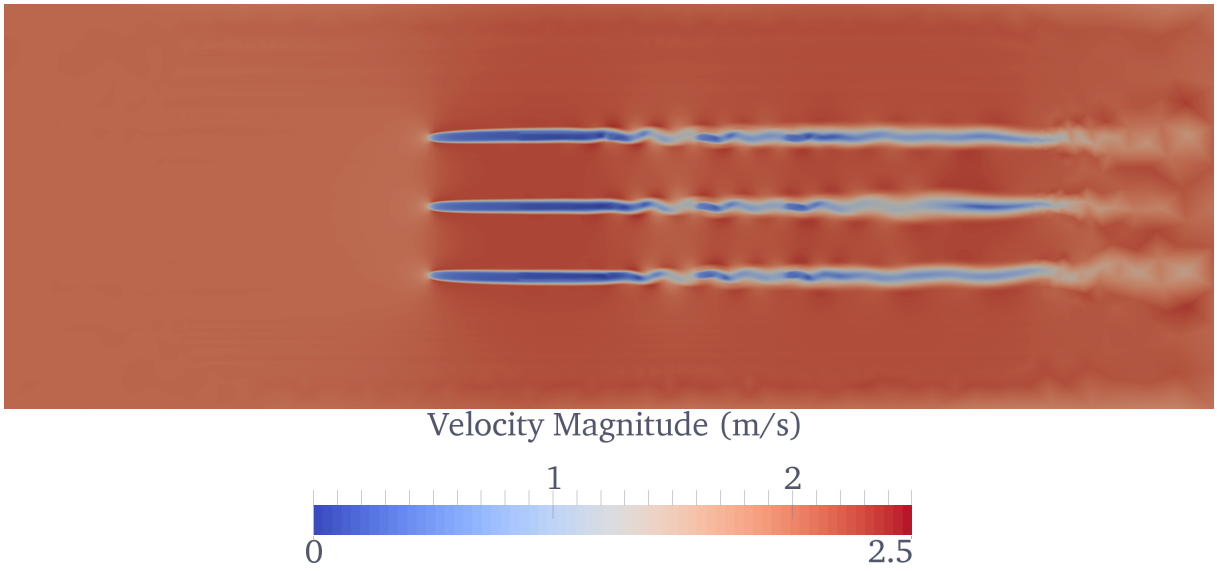



Figure 12: The flow speed around the centred turbine array in the transient simulation. Note the decreased lower limit of the colour bar.

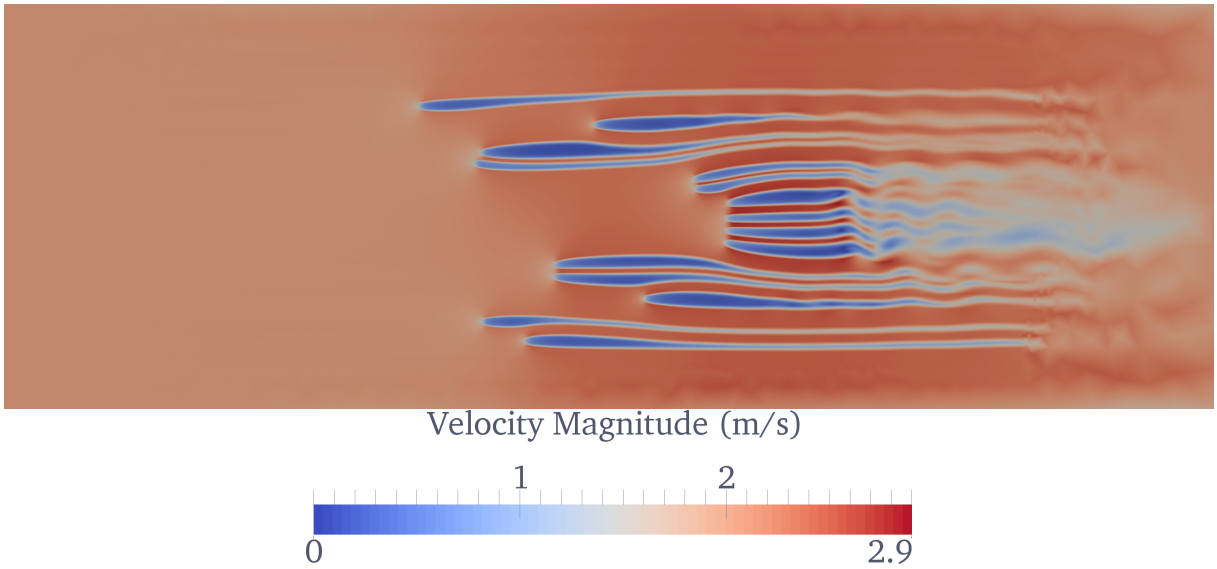



Figure 13: The flow speed around the optimised turbine array in the transient simulation. Note the increased upper limit of the colour bar as a result of increased flow speed around the turbines on the right-hand side that are situated very close together.

sufficiently close to be in the wakes of the upstream turbines. However, in the transient runs the velocity deficit was much stronger, dropping to a near-zero flow speed immediately behind the turbines, resulting in decreased power generation. In the simulation with larger turbine spacing, the improvement in power extraction relative to the centred array was not as significant compared to the steady-state setup, because the wakes were longer and sharper. The turbines should therefore be placed even further apart to allow the wake to sufficiently recover.

Once again, the optimised array configuration out-performed all four man-made configurations under consideration. Interestingly, many of the wakes were entrained into the regions of acceleration and flowed around the downstream turbines as Figure 13 shows, instead of immersing them in the region of low flow speed and limiting power generation. Unlike the other configurations, turbulent eddies were only shed downstream of all the turbines, which explains why the power in Figure 6 remained constant once the flow became fully established.

## 5 CONCLUSIONS

- Adjoint optimisation can be used to greatly improve the power extraction potential of tidal turbine arrays/farms.
- The use of a lower background viscosity and turbulence modelling shows that the velocity deficit is much stronger in reality, resulting in decreased power generation relative to the high viscosity steady-state runs. Furthermore, turbulent eddies have the potential to affect the amount of power extracted, to a small degree in the configurations considered here.
- The power curves of the transient, turbulent runs are qualitatively similar to the steady-state runs, in that the best (staggered) man-made setup is still out-performed by the optimised configuration from adjoint modelling.
- Horizontal LES (HLES) [23] or 3D LES turbulence models, coupled with more realistic turbine parameterisations (e.g. rectangular blocks of drag as opposed to smooth Gaussian functions), should be adopted in future work.

## 6 ACKNOWLEDGEMENTS

The authors acknowledge the support and use of the Imperial College High Performance Computing Service and the UK National Supercomputing Service (ARCHER). CTJ was supported by ARCHER eCSE grant 03-7. SWF is supported by The Research Council of Norway through a Centres of Excellence grant to the Center for Biomedical Computing at Simula Research Laboratory, project number 179578. The authors would also like to acknowledge the support of the UK Engineering and Physical Sciences Research Council under projects EP/J010065/1, EP/L000407/1 and EP/M011054/1.

## REFERENCES

[1] S. W. Funke, P. E. Farrell, and M. D. Piggott. Tidal turbine array optimisation using the adjoint approach. *Renewable Energy*, 63:658–673, 2014.

[2] S. W. Funke, S. C. Kramer, and M. D. Piggott. Design optimisation and resource assessment for tidal-stream renewable energy farms using a new continuous turbine approach. Submitted.

[3] D. M. Culley, S. W. Funke, S. C. Kramer, and M. D. Piggott. Tidal stream resource assessment through optimisation of array design with quantification of uncertainty. In *Proceeedings of the European Wave and Tidal Energy Conference (EWTEC) 2015*, 2015.

[4] T. Divett, R. Vennell, and C. Stevens. Optimization of multiple turbine arrays in a channel with tidally reversing flow by numerical modelling with adaptive mesh. *Philosophical Transactions of the Royal Society A*, 371(1985):1471–2962, 2013.

[5] D. M. Culley, S. W. Funke, S. C. Kramer, and M. D. Piggott. A hierarchy of approaches for the optimal design of tidal turbine farms. In *Proceeedings of the 5th International Conference on Ocean Energy*, 2014.

[6] G. L. Barnett, S. W. Funke, and M. D. Piggott. Hybrid global-local optimisation algorithms for the layout design of tidal turbine arrays. *Renewable Energy*, Submitted.

[7] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[8] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.

[9] S. W. Funke. *The automation of PDE-constrained optimisation and its applications*. PhD thesis, Imperial College London, 2012.

[10] D. M. Culley, S. W. Funke, S. C. Kramer, and M. D. Piggott. Integration of cost modelling within the micro-siting design optimisation of tidal turbine arrays. *Renewable Energy*, 85:215–227, 2016.

[11] A. Logg, K.-A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2012.

[12] P. E. Farrell, D. A. Ham, S. W. Funke, and M. E. Rognes. Automated derivation of the adjoint of high-level transient finite element programs. *SIAM Journal on Scientific Computing*, 35(4):C369–C393, 2013.

[13] S. W. Funke and P. E. Farrell. A framework for automated PDE-constrained optimisation. *ACM Transactions on Mathematical Software*, Submitted.

[14] J. Smagorinsky. General Circulation Experiments with the Primitive Equations. *Monthly Weather Review*, 91(3):99–164, 1963.

[15] J. Deardorff. A numerical study of three-dimensional turbulent channel flow at large Reynolds numbers. *Journal of Fluid Mechanics*, 41(2):453–480, 1970.

[16] J. W. Deardorff. On the magnitude of the subgrid scale eddy coefficient. *Journal of Computational Physics*, 7(1):120–133, 1971.

[17] S. Nash, A. I. Olbert, and M. Hartnett. Towards a Low-Cost Modelling System for Optimising the Layout of Tidal Turbine Arrays. *Energies*, 8:13521–13539, 2015.

[18] T. Divett, R. Vennell, and C. Stevens. Channel scale optimisation of large tidal turbine arrays in packed rows using large eddy simulations with adaptive mesh. In *Proceeedings of the 2nd Asian wave tidal energy conference, Tokyo*, 2014.

[19] T. Divett, R. Vennell, and C. Stevens. Channel-scale optimisation and tuning of large tidal turbine arrays using LES with adaptive mesh. *Renewable Energy*, 86:1394–1405, 2016.

[20] C. Geuzaine and J.-F. Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[21] R. A. Flather. A tidal model of the northwest European continental shelf. *Memoires de la Société Royale des Sciences de Liège*, 10(6):141–164, 1976.

[22] M. G. Larson and F. Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*. Springer, 2013.

[23] G. A. Liek. Horizontal Large Eddy Simulation using Delft2D-MOR: the influence of large horizontal eddies on the shape, depth and extent of scour holes. Master's thesis, TU Delft, 2000.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# DEVELOPMENT OF ACCURATE PNEUMATIC TYRE FINITE ELEMENT MODELS BASED ON AN OPTIMIZATION PROCEDURE

**Chrysostomos-Alexandros Bekakos*[1], George Papazafeiropoulos*[2], Dan J. O'Boy[1], Jan Prins[3]**

[1] Department of Aeronautical and Automotive Engineering, Loughborough University
Epinal Way, LE11 3TU, United Kingdom
c.bekakos@lboro.ac.uk

[2] National Technical University of Athens, Zografou Campus
9, Iroon Polytechniou St., 15780 Zografou, Athens, Greece
gpapazafeiropoulos@yahoo.gr

[3] Jaguar Land Rover
Banbury Road, Gaydon, Warwickshire CV35 0RR, United Kingdom

**Keywords:** Optimization, Eigenfrequency Analysis, Tyre Finite Element Method, ABAQUS-MATLAB linking, tyre model calibration.

**Abstract.** *A novel method for extracting the geometric and constitutive material properties of pneumatic tyres from available numerical or experimental data for the development of realistic and reliable tyre numerical models is proposed. This method involves an optimization procedure, which incorporates a finite element model as a solver (ABAQUS) properly coupled with an optimiser function (MATLAB). Following that, an initial tyre model (P235/75R17) is developed, and then its properties are suitably adjusted via the optimization process, in order for the former to best fit a target model available in the literature, with respect to eigenfrequency analysis results. After the termination of the algorithm, the "optimum" tyre model (i.e. the model which best conforms to the target model) is obtained, the response of which is further investigated to ensure its realistic behaviour, which warrants its use for various numerical simulations. The results of this study show clearly the efficiency of the optimization procedure proposed, as well as the realistic response of the tyre model developed.*

## 1 INTRODUCTION

Handling low frequency interior noise and vibrations which transmits through subframe components on vehicles is a main issue regarding their design. The importance of this aspect is apparent from the related legislation which limits the level of noise a vehicle is allowed to produce. The main source of the vehicle noise is the vibrations induced by the tyres. These, after being transmitted from the tyre to the wheel axle, and through that to the passengers in the vehicle, can have various undesirable effects, some of which are the passengers' inconvenience or body distress, the low performance of the vehicle and its suspension system, etc.

A tyre is subjected to dynamic forces mainly from two main sources: (a) road surface irregularities, potholes, bumps and various other obstacles which impose dynamic loads to the tyre, and (b) dynamic loads originating from various nonuniformities of the tyre, such as slight imbalances or asymmetric tread pattern designs.

It is essential to consider the dynamic characteristics of the tyres of a vehicle, to minimise the aforementioned negative consequences. For this purpose, there is need for detailed knowledge of the dynamic response of a tyre, which is associated with the energy that is being transmitted to the vehicle from various external dynamic events. The dynamic response of a tyre is characterized by its vibration modes, or eigenmodes, namely the natural frequencies of the tyre and the corresponding mode shapes. These, apart from their significance for the design process and troubleshooting of various problems, can constitute a basis for the computational efficiency of the various numerical models of tyres used by both tyre and automotive industries for prediction of performance.

In this study the eigenmodes at the low frequency range are considered for the development of a realistic tyre model, based on numerical data published in the literature. This is achieved through an optimization process which efficiently adjusts various tyre parameters, so that the eigenmodes of the final tyre model reach the corresponding data as close as possible. After the optimum tyre properties are found, the configured model is then used to study the effect of inflation pressure and vertical load (i.e. stationary tyre loaded with vertical force) on its eigenmodes.

## 2 LITERATURE REVIEW

Tyre vibration modes are widely used over the years to represent dynamics in tyre models. The dynamic response of tyre models has been studied analytically, experimentally or semi-empirically, and numerically, however due to the limitations of the analytical and experimental studies, many studies in the literature employ numerical (often finite element) models, which can simulate complex geometries as well as material, geometric and boundary nonlinearities. In this study, the effect of the various parameters on the tyre response is incorporated into an optimization procedure, which ultimately determines the optimum values of these parameters, in order to minimise the error between the numerical model and the available data. Relevant studies about tyre dynamics, as well as optimization procedures are mentioned in the next.

### 2.1 Experimental studies

Experimental studies about the eigenmodes analysis of tyres can be found in [1-3]. In [1], the dynamic response of the vehicle in terms of accelerations was monitored at the wheel axis and the passenger compartment. The tyre vibration modes were identified from the peaks in the response. In [2] a lumped parameter model was developed to study the behaviour of a tyre running on a road surface with irregularities characterized by short wave-length spectrum components. However, the parameters of the lumped model are given by empirical relations, which have resulted from an experimental methodology. In [3] an experimental modal parameter estimation method is presented, in which the frequency response function (FRF) of a tyre is decomposed into the components of individual modes based on the Fourier transform algorithm.

### 2.2 Analytical studies

The analytical models developed for the estimation of the eigenproperties of a tyre range from simple mass/spring systems to various forms of idealized, spring supported, flexible

rings. Representative studies are these in [4], where a rotating and vibrating tyre coupled at its spindle to a secondary structure is simulated. A model of a membrane on an elastic foundation is used for the description of the vibration of a rolling tyre, as well as models for the calculation of the forces at the spindle of a tyre rolling over a small cleat. In [5], the tyre is modelled as a shell structure in contact with the road surface. The contact patch is simulated as a prescribed deformation, and the coupled tyre-cavity governing equation of motion is solved analytically to obtain the tyre structural and acoustic responses.

### 2.3 Numerical studies

Numerical studies regarding the modal analysis of tyres can be found in [6-9]. In [6] the vibration modes of radial tyres on a fixed spindle can be seen and the effect of the tyre components and their contribution in the mode shapes is investigated. Following that, the corresponding tyre model under rolling conditions was considered in [7], and it was shown that non-rolling tyre models are subordinate to their rolling counterparts, as they do not take into account the proper kinematics. In [8], the finite element commercial software ANSYS was used to study the effects of the inflation pressure, the ply angle, the tread pattern and the thickness of the belt on the natural frequencies of the tyre. A basic assumption in this study was that the rubber was simulated as a linear elastic material. Another commercial finite element software (ABAQUS) has been used in [9], where by using various capabilities of ABAQUS, the footprint under purely vertical load was obtained for a vertically loaded tyre. Afterwards, the nodes (node coordinates) being in contact with the road were maintained in contact by applying an equivalent distributed vertical load, whereas the centre of the wheel was set free in all degrees of freedom. In this condition of the model, a frequency analysis was performed and it was found that the boundary conditions on the tyre model can have large impact on its eigenmode response.

### 2.4 Optimization

In general, the methods used to optimise a model (optimization methods) range from relatively simple mathematical programming based (exact) methods to novel heuristic search techniques. The methods belonging to the first category are very efficient for cases with a few design variables. Methods belonging to this category are those using the sequential quadratic programming procedure for nonlinear optimization (used in this study), as well as others. More details regarding these methods are presented in [10]. More robust optimization techniques, which are capable of searching effectively the whole design variable domain and not being trapped into local optima, can be used for increased number of design variables, or non-differentiable functions. Recently developed heuristic methods, such as genetic algorithms, simulated annealing, threshold accepting, tabu search, ant colonies, particle swarm, etc. provide more attractive alternatives.

## 3 NUMERICAL MODELING

### 3.1 Introduction

The tyre considered in this study was modelled in the commercial finite element code ABAQUS 6.13. Implicit integration was performed using ABAQUS/Standard, which was also used for the eigenfrequency and eigenmode extraction of the tyre. The optimization procedure, as well as the necessary coupling with ABAQUS, was implemented in MATLAB programming language.

### 3.2 Tyre modeling

The cross section of the tyre, P235/75R17, is shown in Figure 1. The tyre is comprised of the belt region, the tread region and the side walls which are being modelled with a hyperelastic material, representative of rubber. The hyperelastic material is simulated by the one term polynomial strain energy potential (Mooney-Rivlin model) with one term Prony series to account for viscoelasticity [11]. The belt region contains reinforcement of two layers (illustrated as Belt layer 1 & 2 in the Figure), and the reinforcement of carcass. The last extends over the belt region and it covers the side walls. Both belt layers and the carcass are discretized with surface elements with twist (SMFGAX1). The rim is discretized with 2-node, linear links for axisymmetric planar geometries (RAX2), and the belt, bead, sidewall and tread regions are discretized with 4-node bi-linear, reduced integration elements with hourglass control (CGAX4R). The nodes of the surface elements of the carcass share the same nodes with those of the belt region elements. If separate nodes are used for these two sections (which have the same coordinates) numerical instabilities may occur during the analysis.

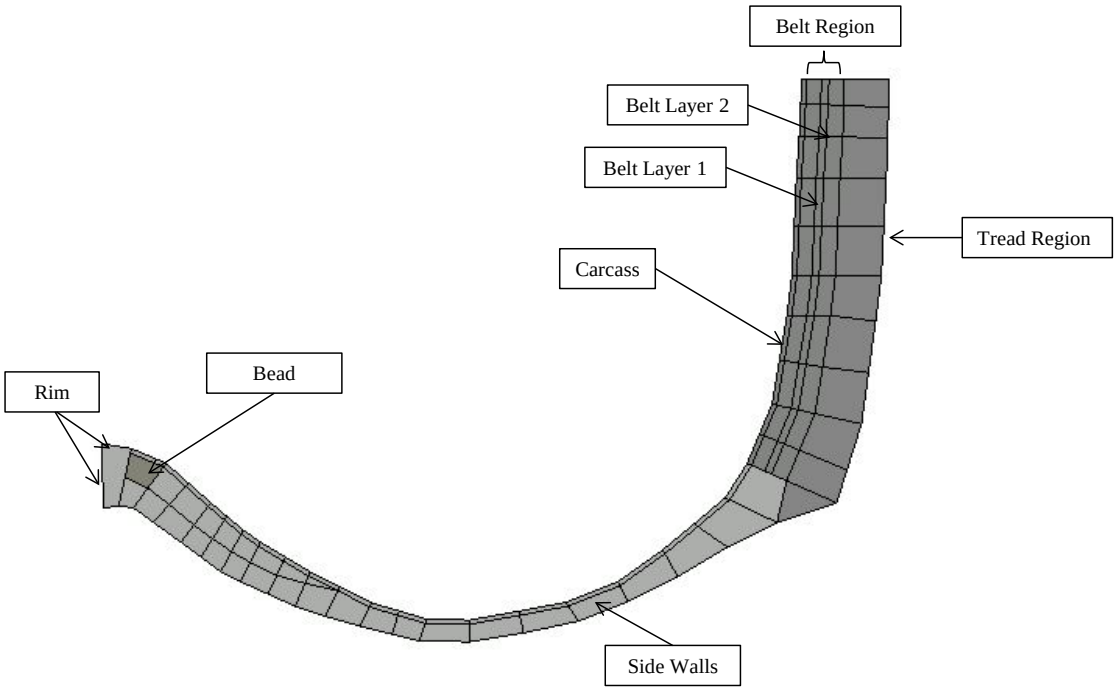



Figure 1: Tyre half-cross section geometry.

By utilizing the capabilities of ABAQUS with regard to symmetric model generation (SMG), symmetric results transfer (SRT) and restart option, the full 3d numerical model of the tyre is developed, as shown in Figure 2. Inflation pressure is imposed on the inner surface of the tyre as a distributed load. Regarding the boundary conditions, two cases can be distinguished: (a) for the unloaded tyre, the boundary conditions are imposed on the six degrees of freedom of the wheel centre (fixed-spindle), and (b) for the loaded tyre, the road is considered to be fixed and the tyre centre is constrained along all degrees of freedom except for the degree of freedom along which the vertical load is imposed. The rim is rigidly constrained to the tyre centre. The friction between the tyre and the road (in the case of the loaded tyre) is assumed to be of Coulomb type, with coefficient equal to 0.5.

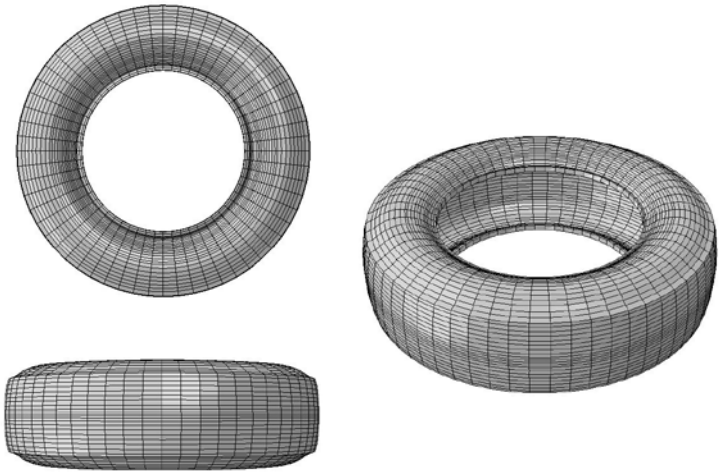


Figure 2: Illustration of the tyre model.

### 3.3 Formulation of the optimization problem

In this section, the optimization procedure is outlined, and its various aspects are presented (design variables, parameters, constraints and objective function).

- Design variables

The geometric properties of the belts and carcass reinforcement, as well as the hyperelastic Mooney-Rivlin $C_{10}$ constant are selected as design variables. The reinforcement layers are defined in ABAQUS as smeared layers with a thickness equal to the ratio of the area of each reinforcing bar to the reinforcing bar spacing. This calculated thickness is assumed to remain constant all over the extent of the layer. This consideration has a considerable effect on the selection of the design variables, since the stiffness of each reinforcement layer contributes to the eigenproperties of the tyre. Due to the fact that the rebar stiffness is given by a fraction of two separate input parameters, for constant layer stiffness they become dependent on each other. Therefore, it is objective that only one of the two parameters for each layer is selected as an independent design variable, and the other remains fixed. The variable to remain fixed is the easier to be measured, in terms of order of magnitude. Another point to be mentioned is that, because the two belt layers have symmetric orientation with respect to the plane of the tyre, and the tyre is a centre symmetric structure, its eigenmodes are expected to be also symmetric; this means that the cross section areas of the two belt reinforcements have to be equal, and therefore the belt reinforcement cross sectional area was considered as a single design variable. The design variables of the optimization problem, along with their upper and lower bounds are shown in Table 1.

| Design variable | Lower bound | Upper bound |
|---|---|---|
| $A_{belt}$ | $10^{-7}$ | $10^{-5}$ |
| $A_{carcass}$ | $10^{-8}$ | $10^{-5}$ |
| $C_{10}$ | $10^{5}$ | $10^{7}$ |

Table 1: Design variables of the optimization problem and their lower and upper bounds.

- Parameters

The parameters of the optimization problem are the design input data that remain fixed during the optimization process. These include, as aforementioned in the previous section, the spacing of the rebar layers, which is set to be equal to 0.00116m for the belts and 0.001m for the carcass. Furthermore, the constants of the Mooney-Rivlin strain energy potential are $C_{01}$=0 and $D_1$=5.085*$10^{-8}$$Pa^{-1}$. The cord angles are 70 and 110 degrees for the two belt layers, and 0 degrees for the carcass. The material properties of the belts and the carcass are also held fixed during the optimization process. More details about these properties can be found in [11]. The inflation pressure with which the tyre is inflated is 240kPa.

- Constraints

No constraints are imposed to the model being optimised, apart from the upper and lower limits of the design variables. The latter require some experience to be specified, because large upper bounds or small lower bounds can lead to numerical instabilities in the solver, such as excessive element distortion, etc, which result in the premature termination of the optimization procedure.

- Objective function

The objective function for the optimization problem has to be of an appropriate form, so that it becomes minimum if the numerically calculated eigenfrequencies coincide with the ones available from the literature. The first 16 eigenfrequencies of the tyre are considered in the objective function, which is given by the equation:

$$obj = \sqrt{\sum_{i=1}^{16} \left( f_{i,num} - f_{i,lit} \right)^2} \tag{1}$$

where $f_{i,num}$ is the $i^{th}$ eigenfrequency calculated by the numerical model in every iteration of the algorithm and $f_{i,lit}$ is the corresponding $i^{th}$ eigenfrequency available in the literature. The correspondence between the various eigenfrequencies is made by taking into account the deformed configurations of the various eigenmodes.

### 3.4 Algorithm used for the optimization problem

The optimization algorithm used in this study is a sequential quadratic programming (SQP) method. In this method, a quadratic programming (QP) subproblem is solved at each iteration. For this purpose the MATLAB built in function fmincon is used. This function used an active set strategy and updates an estimate of the Hessian of the Lagrangian at each iteration using the BFGS formula. An active-set method initializes by making a guess of the optimal active set, and if this guess is incorrect, it repeatedly uses gradient and Lagrange multiplier information to proceed towards the optimum solution.

The fmincon optimiser (MATLAB) is properly coupled with the analysis solver (ABAQUS) in order to take the frequency analysis results. This is done inside the objective function in which ABAQUS is called to perform the necessary analyses. Except for this, the necessary input (*.inp) files for the ABAQUS runs are created by suitable MATLAB functions. To read the results of the analyses from the corresponding ABAQUS results (*.fil) files, special MATLAB functions are used. While the analysis solver is running the optimiser is halted and its execution is continued after the lock (*.lck) file has been deleted.

## 4 OPTIMIZATION RESULTS

The results of the optimization process as presented in the previous sections are shown in Table 2.

| | **Initial model** | **Optimised model** | **Wheeler et al. [6]** | **Deviation (%)** |
|---|---|---|---|---|
| Design Variables | | | | |
| $A_{belt}$ ($m^2$) | $2.11868*10^{-7}$ | $3.64826*10^{-7}$ | N/A | - |
| $A_{carcass}$ ($m^2$) | $4.20835*10^{-7}$ | $8.01133*10^{-8}$ | N/A | - |
| $C_{10}$ (Pa) | $10^6$ | $10^6$ +0.01489 | N/A | - |
| Eigenfrequencies | | | | |
| $f_1$ [0,0] (Hz) | 36.85 | 30.86 | 31.7 | 2.66 |
| $f_2$ [0,0] (Hz) | 37.17 | 35.85 | 35 | 2.43 |
| $f_3$ [1,1] (Hz) | 43.85 | 36.92 | 37.8 | 2.33 |
| $f_4$ [1,1] (Hz) | 43.85 | 36.92 | 37.8 | 2.33 |
| $f_5$ [1,0] (Hz) | 65.07 | 58.75 | 58.5 | 0.43 |
| $f_6$ [1,0] (Hz) | 65.07 | 58.75 | 58.5 | 0.43 |
| $f_7$ [2,1] (Hz) | 76.33 | 68.41 | 66.1 | 3.49 |
| $f_8$ [2,1] (Hz) | 76.33 | 68.41 | 66.1 | 3.49 |
| $f_9$ [2,0] (Hz) | 86.65 | 78.67 | 79.5 | 1.04 |
| $f_{10}$ [2,0] (Hz) | 86.65 | 78.67 | 79.5 | 1.04 |
| $f_{11}$ [3,0] (Hz) | 104.36 | 96.42 | 97.6 | 1.21 |
| $f_{12}$ [3,0] (Hz) | 104.36 | 96.42 | 97.6 | 1.21 |
| $f_{13}$ [3,1] (Hz) | 117.07 | 107.9 | 102.7 | 5.06 |
| $f_{14}$ [3,1] (Hz) | 117.07 | 107.9 | 102.7 | 5.06 |
| $f_{15}$ [4,0] (Hz) | 122.65 | 114.9 | 115.9 | 0.83 |
| $f_{16}$ [4,0] (Hz) | 122.65 | 114.9 | 115.9 | 0.83 |
| Algorithm Details | | | | |
| Min. value of obj. function | - | 8.59 | - | - |
| Number of obj. function evaluations | - | 25 | - | - |

Table 2: Results of the optimization procedure of the tyre frequency analysis considered in this study.

It is noted that each natural frequency corresponds to a pair of integers enclosed in brackets ([c,m]). The first integer denotes the number of sinusoidal waves in the circumferential direction of the wheel, whereas the second integer shows the number of waves in the meridional direction at a specific location, where the deformation of the eigenmode shape is maximum. In addition, only the first 16 eigenmodes were considered for the development of the realistic tyre model, in order to reduce the computational cost.

The first column of Table 2 shows the data of the initial model, used as the starting point of the optimization process. It is evident that the eigenfrequencies of the initial model have large difference from the eigenfrequencies of the model published in [6]. In the second column, the parameters of the optimum model are shown, as well as the values of the design variables leading to it. Regarding the eigenfrequencies, it is observed that they are much closer than those of the initial model, leading thus to a numerical model that conforms more to the available numerical data, and therefore it is more realistic. The maximum deviation of the eigenfrequencies is noted to be roughly 5%. The optimum model has higher cross section of the reinforcement of the belts, and lower cross section area of the reinforcement of the carcass

than the initial model. The hyperelastic constant $C_{10}$ is only slightly increased after the optimization. Regarding the algorithm output, the minimum value of the objective function is equal to approximately 8.59Hz, and the algorithm converged after 25 objective function evaluations. The reason for the termination of the algorithm is that the magnitude of the search direction was less than the corresponding tolerance. The most important factor affecting the tyre modal behaviour during the optimization procedure is proved to be the cross section area of the carcass ($A_{carcass}$). Due to the fact that the initial model has generally higher eigenfrequencies than those of the target model [6], its stiffness had to be decreased, in order for the model to approach the latter. The decrease in stiffness is achieved with a relatively large decrease in the cross sectional area of the carcass, although the cross section area of the belt reinforcement increases.

In Figure 3 the various eigenmodes of the optimised tyre model are shown. The figure is divided into 9 subfigures, each of which shows a tyre eigenmode shape viewed from 4 different perspectives. The fundamental eigenmode is the axial or lateral mode, and after this the torsional, pitch, diametric, and higher modes follow. There is total correspondence between the integer pairs which appear in the bottom of each subfigure, and the ones shown in the first column of Table 2.

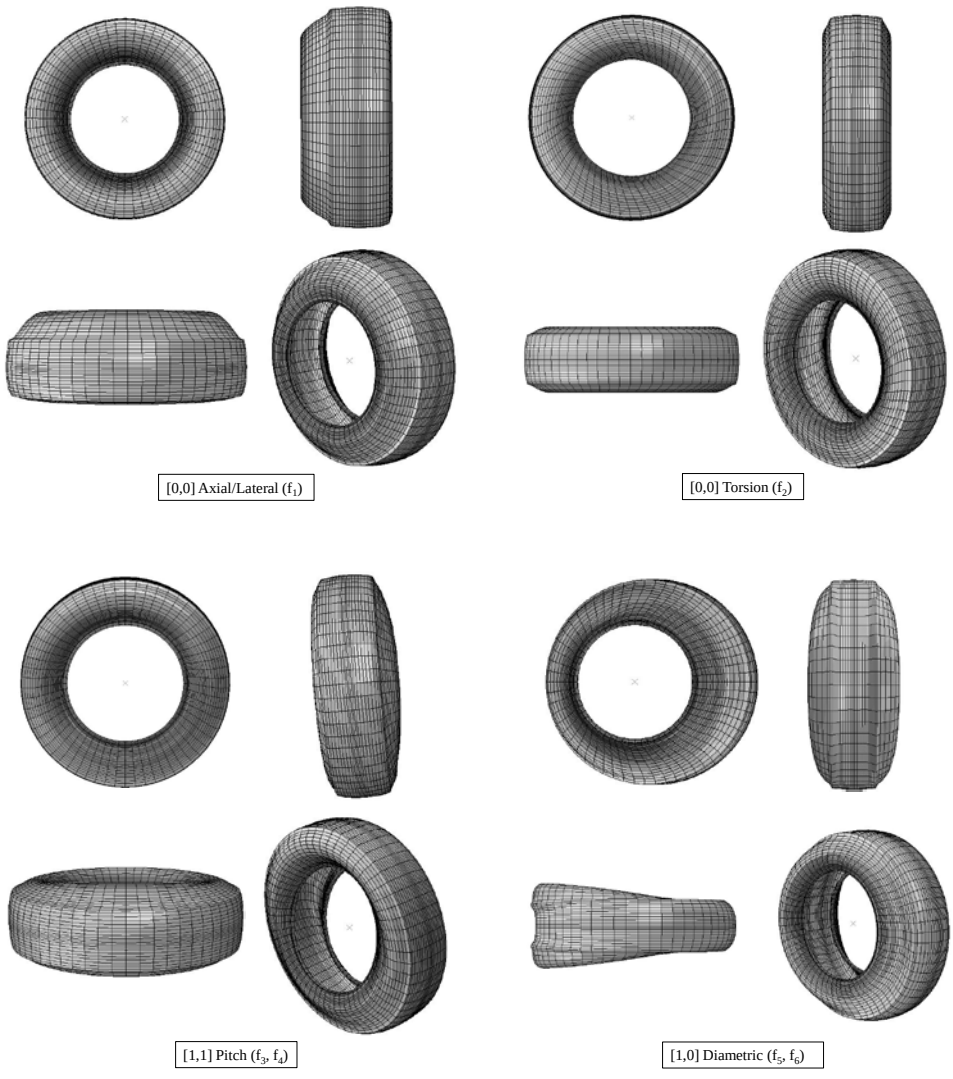



Figure 3: Eigenmode shapes of the optimised tyre model (continued in the next page).

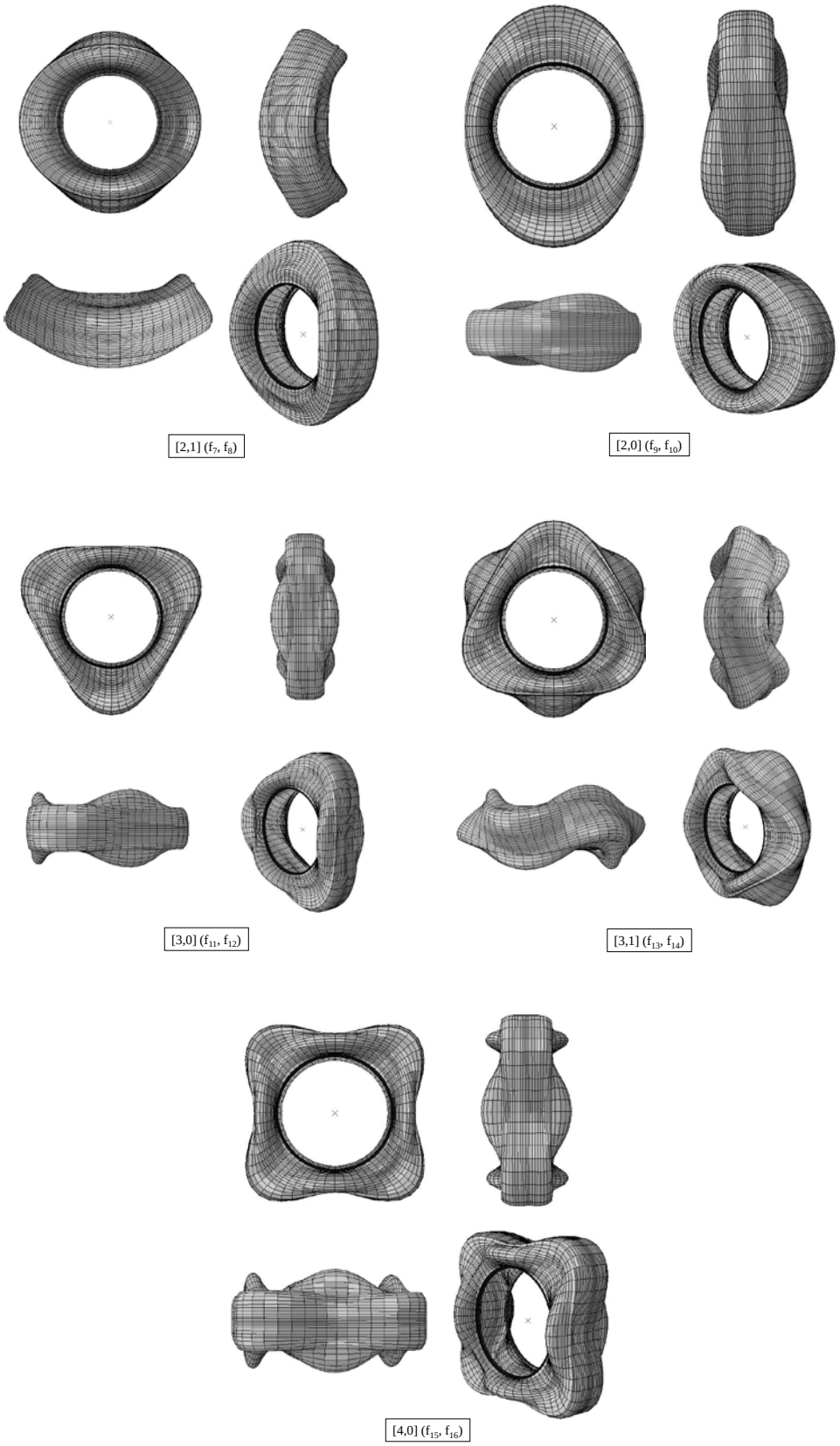



Figure 3: Eigenmode shapes of the optimised tyre model (continued from previous page).

## 5 EFFECT OF INFLATION PRESSURE ON MODAL RESPONSE OF OPTIMUM TYRE

The optimised model is the most realistic version of the selected tyre type (P235/75R17) with respect to its modal response. It is close to the modal data available in [6], and for this reason, it allows its use for dynamic response analyses. In an attempt to further valildate the optimised model, the variation of its eigenmodes and eigenfrequencies is studied for varying inflation pressure.

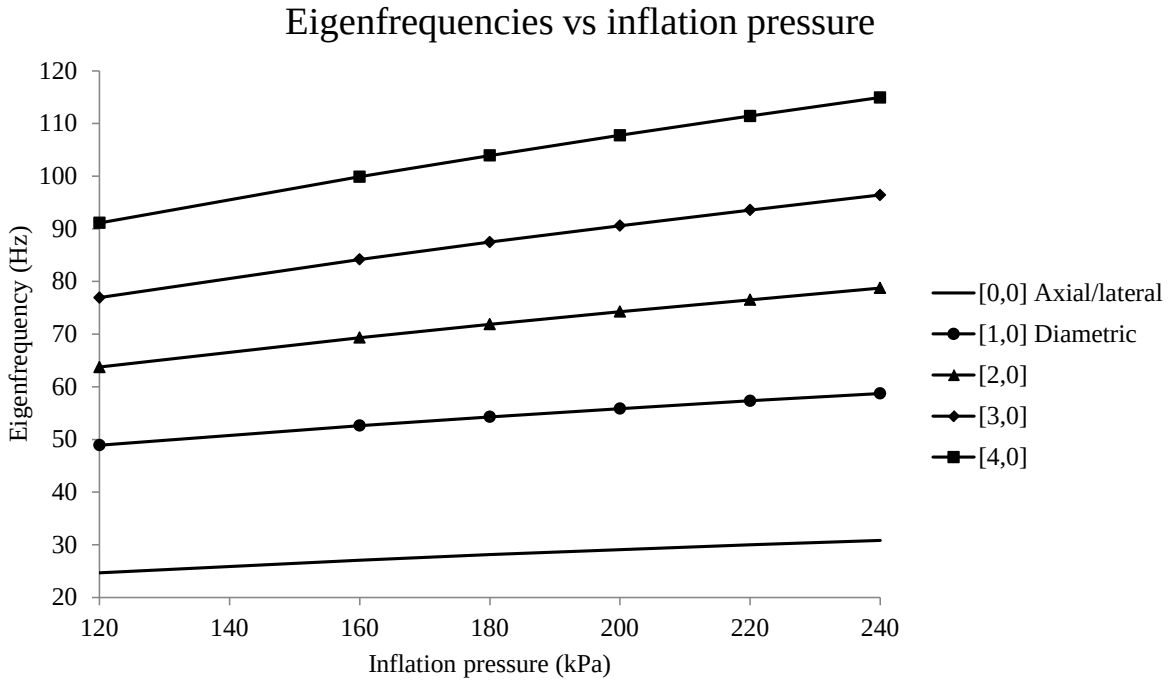



Figure 4: Eigenfrequencies of various mode shapes of the optimised tyre model for various inflation pressures.

In Figure 4 the effect of the inflation pressures on the eigenfrequencies of the tyre can be observed. As it can be expected, as the inflation pressure rises, the eigenfrequency of a specific eigenmode increases, as the increased inflation pressure makes the tyre stiffer. This is a trend widely observed in the literature and once again corroborates the realistic behaviour of the optimum tyre. Moreover, it is apparent that the increase of the eigenfrequency of each mode for increasing inflation pressure is nonlinear. Specifically, for lower values of the inflation pressure, the rate of increase in the eigenvalues becomes higher than that for higher values of the inflation pressure. Finally, for the higher order eigenmodes, the increase in the eigenfrequency for the same difference in the inflation pressure is larger than that for the lower eigenmodes, which is in agreement with relevant results found in [8].

## 6 COMPARISON BETWEEN OPTIMUM AND INITIAL TYRE MODELS

In this section the static response of the initial tyre model and the optimised tyre model is considered with a view to make a comparison between the two models in terms of their response. For this purpose, the contact area of a vertically loaded tyre on a rigid surface is calculated for the two tyre models. The results are shown in Figure 5. It seems that the contact area of the optimised tyre model is generally larger than that of the initial tyre model, for the same values of vertical load, with an exception in the interval of vertical load around 2 kN,

where the two areas are roughly equal. This is a direct consequence of the fact that, the eigenfrequencies of the optimum tyre model are lower than the corresponding ones of the initial tyre model, which indicates lower stiffness, and therefore larger contact area. Although the difference seems to be relatively small for vertical loads lower than 2 kN, the discrepancy is expected to be quite significant for vertical loads higher than 4 kN. The values of the contact area calculated for the optimum model are close to numerical [12] and experimental [13] results found in the literature which proves that this tyre model exhibits realistic behaviour.

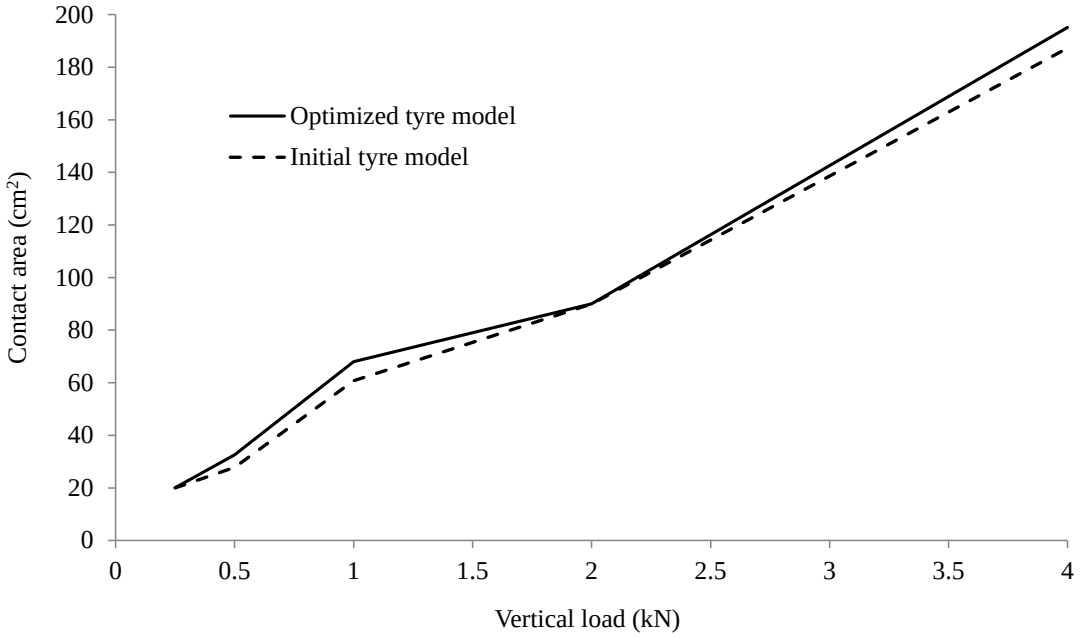



Figure 5: Contact area versus vertical load for the two tyre models considered in this study.

## 7 CONCLUSIONS

- An optimization procedure has been implemented which successfully adjusted the geometric properties of a tyre (P235/75R17), so that its modal properties match closely those available in the literature [6].
- The optimization procedure proposed in this study can be extended for the calibration of various tyre material constitutive models, which in most cases are unknown, based on experimental data, such as contact area and modal analysis results. For instance, the parameters of the Mooney-Rivlin hyperelastic material model can be extracted from tyre response data through the optimization procedure. In general, this procedure can be followed for the development of any realistic tyre model.
- Successful coupling between the finite element software ABAQUS (used as the solver) and MATLAB (used as the optimiser) was performed, in a way that optimises the running speed of the whole optimization process.

- The deviation of the eigenfrequencies of the optimised tyre model from the corresponding eigenfrequencies of the target tyre model [6] (i.e. the model to which the initial model is fitted) was generally small, not larger than roughly 5%.
- The response of the optimised tyre model was found to be in close agreement with available numerical and experimental data.

## ACKNOWLEDGEMENTS

This work was supported by Jaguar Land Rover and the UK-EPSRC grant EP/K014102/1 as part of the jointly funded Programme for Simulation Innovation.

## REFERENCES

[1] R.W. Scavuzzo, T.R. Richards, L.T. Charek, Tire Vibration Modes and Effects on Vehicle Ride Quality. *Tire Science and Technology, TSTCA*, **21**(1), pp.23-39, 1993.

[2] P. Bandel, C. Monguzzi, Simulation model of the dynamic behavior of a tire running over an obstacle. *2nd annual meeting of The Tire Society,* The University of Akron, Akron, Ohio, March 23-24, 1983.

[3] H. Matsuoka, M. Okuma, A New Experimental Modal Parameter Estimation Method for Cylindrical Structures by Standing Wave Decomposition and Application to Tire Modelling. *Proceedings of ISMA at The International Conference on Noise and Vibration Engineering*, V1913 – 1919, 2002.

[4] E.P. Vinesse, Substructuring method and eigenfunction representation for a rolling tyre coupled to a secondary mechanical system. *Proc Instn Mach Engrs*, **210**, 313-325, 1996.

[5] L. Molisani, A coupled tire structure-acoustic cavity model, *PhD Thesis, Virginia Polytechnic Institute and State University*, Blacksburg, Virginia, US, 2004.

[6] R.L. Wheeler, H.R. Dorfi, B.B. Keum, Vibration modes of radial tires: measurement, prediction, and categorization under different boundary and operating conditions. *SAE Noise and Vibration Conference and Exhibition*, Traverse City, Michigan, May 16-19, 2005.

[7] H.R. Dorfi, R.L. Wheeler, B.B. Keum, Vibration Modes of Radial Tires: Application to Non-rolling and Rolling Events. *SAE Noise and Vibration Conference and Exhibition*, Traverse City, Michigan, May 16-19, 2005.

[8] A. Chatterjee, V. Ranjan, Free vibration analysis of radial pneumatic tire using FEM. *Int. J. Emerging Technology & Advanced Engineering*, **2**(8), 319-324, 2012)

[9] E. Bolarinwa, O. Olatunbosun, On Finite Element Tyre Modal Analysis, *SAE Technical Paper 2015-01-1518*, 2015

[10] J. Nocedal, S.J. Wright, *Numerical Optimization, 2nd Edition*. Springer Series in Operations Research, Springer Verlag, 2006.

[11] C.A. Bekakos, G. Papazafeiropoulos, D.J. O'Boy, J. Prins, Pneumatic tyres interacting with deformable terrains. *13th International Conference on Motion and Vibration Control*, University of Southampton, July 3-6, 2016.

[12] J.H. Lee, Finite element modelling of interfacial forces and contact stresses of pneumatic tire on fresh snow for combined longitudinal and lateral slips, *Journal of Terramechanics*, **48**, 171-197, 2011.

[13] S.A. Shoop, Finite Element Modelling of Tire-Terrain Interaction, *Report number: ERDC/CRREL TR-01-16*, University of Michigan, 2001.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# METEOROLOGICAL DATA ASSIMILATION USING AN ADJOINT PROGRAM GENERATED BY AUTOMATIC DIFFERENTIATION

**Yasuyoshi Horibata**

Hosei University
3-7-2 kajino-cho, Koganei, Tokyo, Japan
e-mail: horibata@hosei.ac.jp

**Keywords:** Adjoint program, Automatic differentiation, Atmospheric flow simulation, TAPENADE, modified L-BFGS-B.

**Abstract.** *Meteorological data assimilation is formulated as a large-scale nonlinear optimization problem. In order to solve the problem, modified L-BFGS-B algorithm is used. At each search, the algorithm requires the gradient of a cost function. Automatic differentiation transforms a source computer program that computes a mathematical vector function into a new source program that computes derivatives of this function. A source computer program that simulates the atmospheric flow is differentiated in reverse mode by automatic differentiation tool TAPENADE, and the adjoint program is generated. The gradient is computed by using the adjoint program. Numerical experiments are presented for a three-dimensional downburst.*

## 1 INTRODUCTION

Numerical simulation of the atmospheric flow is an initial/boundary value problem: given an initial condition of the atmosphere, and appropriate surface and lateral boundary conditions, the model simulates the atmospheric evolution. Assimilation of meteorological observations is the process through which all the available information is used in order to estimate as accurately as possible the state of the atmospheric flow (an initial condition) [1]. The available information consists of the observation proper, and of the physical laws that govern the evolution of flow. In this paper, the assimilation is formulated as a nonlinear optimization problem; a cost function is minimized. At each iteration, an optimization algorithm requires the gradient of the cost function.

We developed a method for gradient computation by perturbing the initial condition for the discretized model equation [2, 3]. In this paper, however, we compute the gradient using an adjoint program.

The adjoint program is obtained by automatic differentiation (AD) tool. Automatic differentiation transforms a source computer program that computes a mathematical vector function into a new source program that computes derivatives of this function. A source computer program that simulates the atmospheric flow is differentiated by the AD tool TAPENADE [4]. The generated adjoint program is employed to compute the gradient of the cost function.

## 2 AUTOMATIC DIFFERENTIATION

Given a vector argument $X \in R^n$, a source computer program P computes some vector function $Y = F(X) \in R^m$. The AD tool generates a new source program that, given the argument $X$, computes some derivatives of $F$. P represents a sequence of instructions, which is identified with a composition of vector functions. Thus

$$P \quad is \quad \{I_1; I_2; \ldots I_p\} \tag{1}$$

$$F \quad = \quad f_p \circ f_{p-1} \circ \ldots \circ f_1 \tag{2}$$

Here each $f_k$ is the elementary function implemented by instruction $I_k$.

The chain rule gives the Jacobian $F'$ of $F$. Using the Jacobian, for a small perturbation $\delta X$ in $X$, the corresponding perturbation $\delta Y = F(X + \delta X) - F(X)$ in $Y$ is computed:

$$\delta Y = F'\delta X = f'_p(X_{p-1}) f'_{p-1}(X_{p-2}) \ldots f'_2(X_1) f'_1(X_0)\delta X \tag{3}$$

Here, $X_k$ are the values of all variables after each instruction $I_k$; $X_0 = X$ and $X_k = f_k(X_{k-1})$.

The tangent program computes this perturbation; the program is generated by differentiating the source program in tangent mode. On the other hand, we define a scalar linear combination $Y^T W$ as the new result of the source program; $W$ is the weighting vector. The gradient of $Y^T W$ is

$$F'^T(X)W = f'^T_1(X_0) f'^T_2(X_1) \ldots f'^T_{p-1}(X_{p-2}) f'^T_p(X_{p-1})W \tag{4}$$

The adjoint program is generated by differentiating the source program in reverse mode, and computes the gradient.

## 3 SOURCE PROGRAM

The model equation for the atmospheric flow [5] is written as

$$\frac{\partial \mathbf{x}}{\partial t} \quad = \quad \mathbf{F}(\mathbf{x}, \mathbf{y}) + \mathbf{D}(\mathbf{x}, \mathbf{y}) \tag{5}$$

$$\frac{\partial \mathbf{y}}{\partial t} = \mathbf{C}(\mathbf{x}, \mathbf{y}) \tag{6}$$

where

$$\mathbf{x} = (\theta, q_v, q_c, q_r, K_m) \qquad \mathbf{y} = (u, v, w, \pi) \tag{7}$$

Here $q_v$, $q_c$, and $q_r$ are the mixing ratios of water vapor, cloud water, and rainwater, respectively; $\theta$ and $K_m$ are the potential temperature and the SGS eddy coefficient for momentum, respectively; $u$, $v$, and $w$ are the $x$ , $y$, and $z$ components of the wind velocity, respectively; and $\pi$ is the deviation of the nondimensional pressure.

This is the model in differential form. Once it has been discretized in space using finite differences, the model is written as a set of nonlinear-coupled ordinary differential equations. Next, once a time-difference scheme is used, it becomes a set of nonlinear-coupled difference equations.

Let $\mathbf{x}(t_0)$, $\mathbf{y}(t_0)$ denote the initial condition. The time-difference scheme integrates the model equation by marching outward from the initial condition. The initial condition defines a unique solution $\mathbf{x}(t)$, $\mathbf{y}(t)$ to the model equation.

## 4 COST FUNCTION

The cost function is taken as

$$J = \sum_{i=0}^{N} H[\mathbf{x}(t_i), \mathbf{y}(t_i)]\, dt \tag{8}$$

where $H[\mathbf{x}(t_i), \mathbf{y}(t_i)]$ is a scalar measuring the distance between $\mathbf{x}(t_i)$, $\mathbf{y}(t_i)$ and its observations available at time $t_i = i\Delta t$; $\Delta t$ is the time interval. The available observations are assumed to be distributed over a limited time interval $[t_0, t_N]$. The constraint is the model equation. For a given initial condition and for the corresponding solution $\mathbf{x}(t_i)$, $\mathbf{y}(t_i)$ of the model, the cost function is evaluated. Thus, the cost function is regarded as a function of $\mathbf{x}(t_0)$, $\mathbf{y}(t_0)$.

Hence, assimilation of meteorological observations is formulated as a large-scale nonlinear optimization problem. This problem has thousands or millions of variables. Modified L-BFGS-B algorithm [6, 7] is employed in order to solve this problem.

## 5 NUMERICAL EXPERIMENTS

Numerical experiments are presented for a three-dimensional downburst. A downburst is a strong downdraft which induces an outburst of damaging winds on or near the ground. A downburst is simulated using the source program [5]. Figure 1 and Figure 2 show the evolution of the rainwater and flow, respectively. Out of the simulation result, the x component of the wind velocity and the rainwater are assumed to be observations from a Doppler radar.

In order to estimate the initial condition used for the simulation, the modified L-BFGS-B algorithm solves the large-scale optimization problem. Starting from the initial guess, the L-BFGS-B algorithm repeats search until it reaches the optimal point. At each search, the gradient of the cost function is computed using the adjoint program [8]. Figure 3 shows the rainwater and flow fields of the starting point used in the optimization.

Figure 4 shows the convergence history of the optimization.

Figure 5 compares the rainwater field recovered by the optimization and the corresponding one of the initial condition in the $x - z$ cross section at $y = 0$. Figure 6 compares the flow field recovered by optimization and the corresponding one of the initial condition in the $x - z$ cross section at $y = 0$.

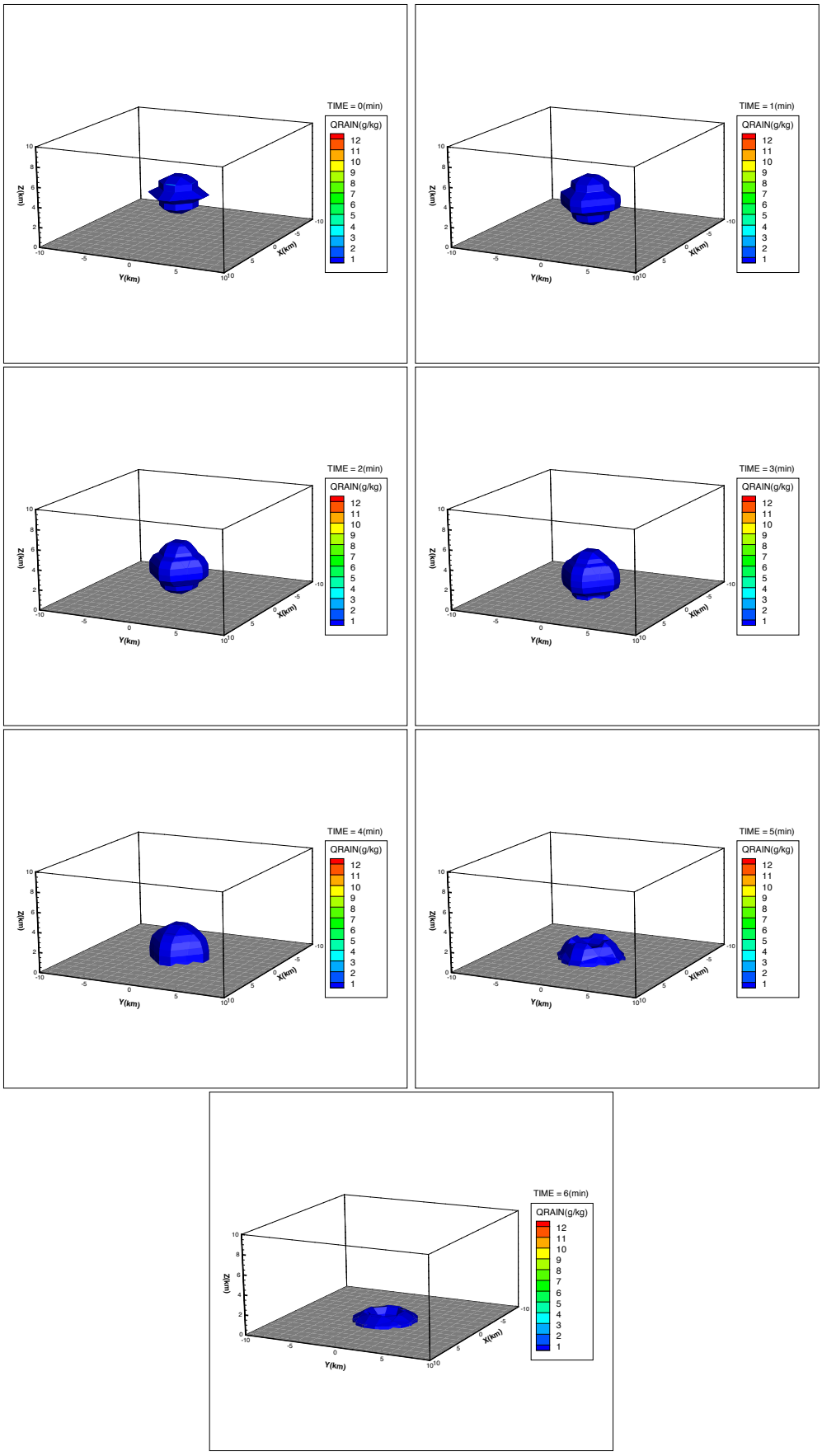



Figure 1: Evolution of the rainwater field after the downburst initiation.

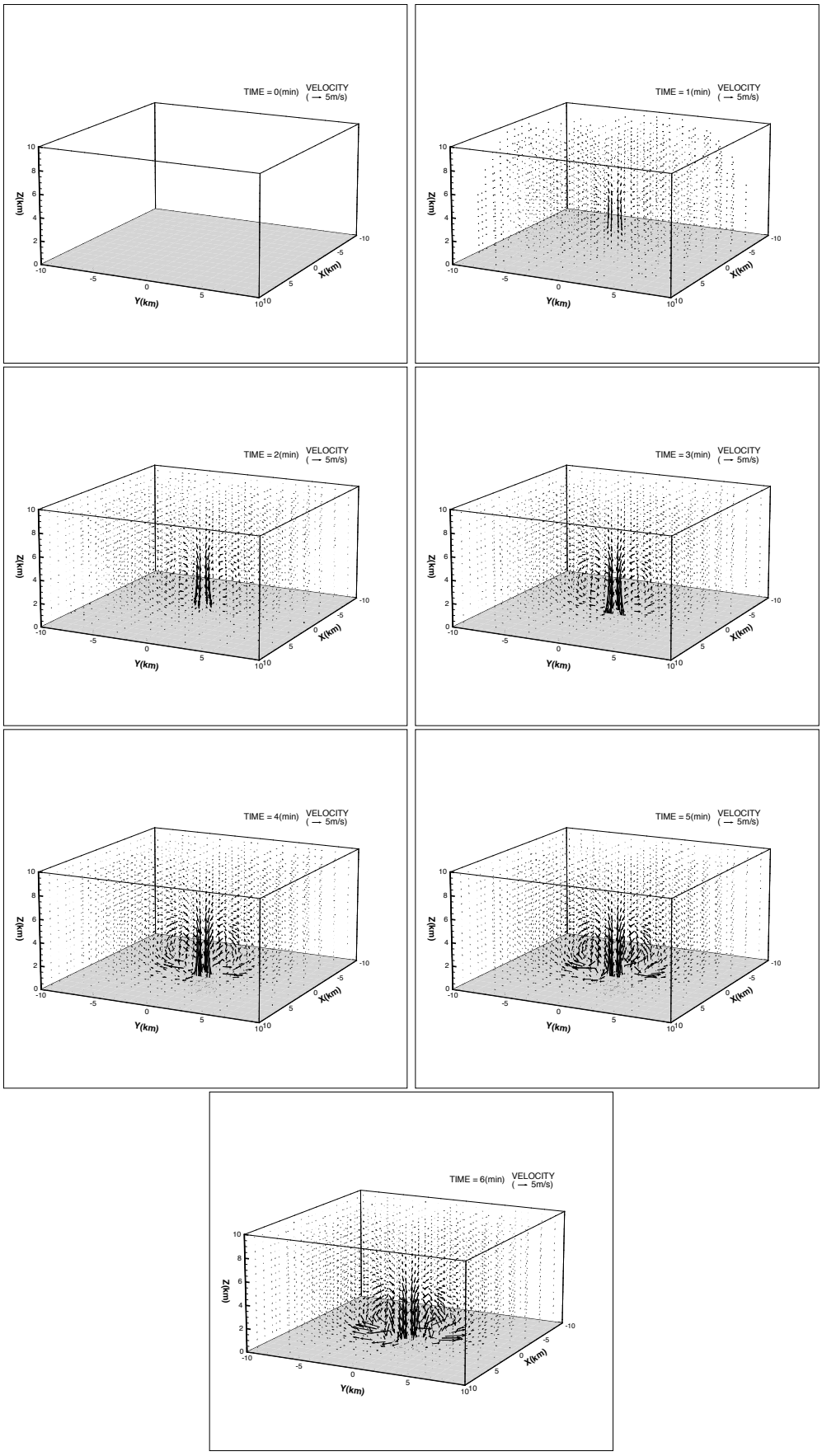



Figure 2: Evolution of the flow field after the downburst initiation.

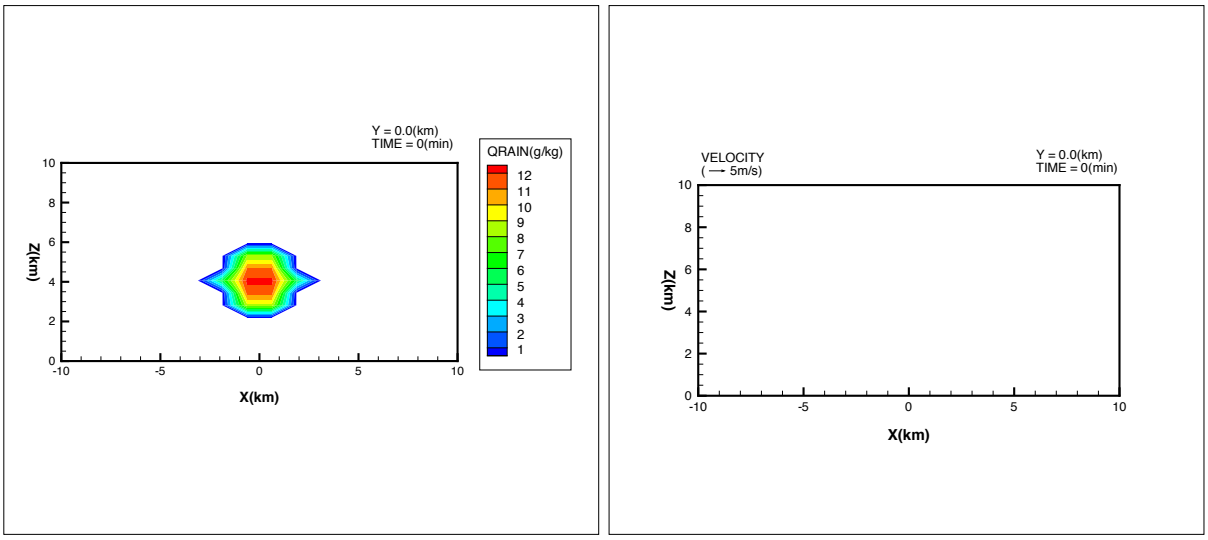



Figure 3: Starting point for the optimization; the rainwater (left) and flow (right) fields in the $x - z$ cross section at $y = 0$.

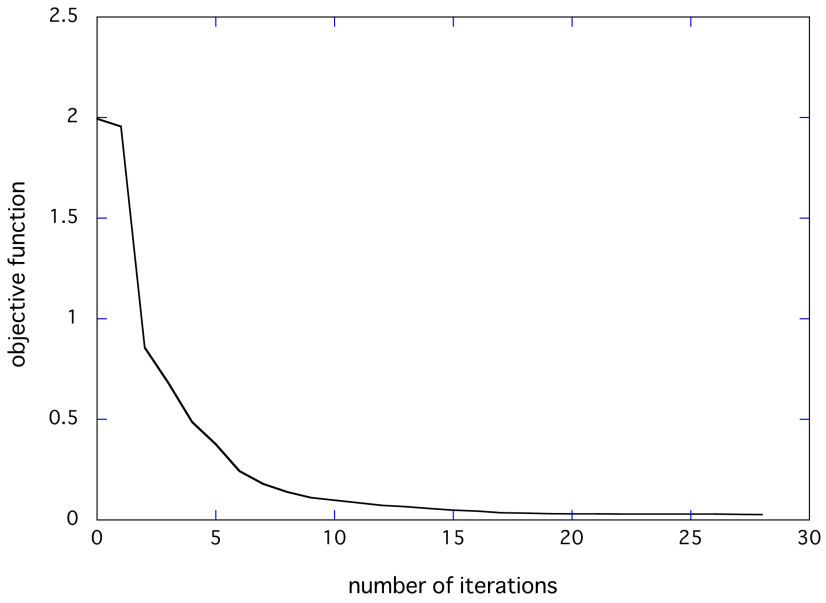



Figure 4: Convergence history of the optimizatiot.

## 6 CONCLUSIONS

- Meteorological data assimilation is formulated as a non-linear optimization problem.
- A source computer program that simulates the atmospheric flow is differentiated by automatic differentiation tool TAPENADE.
- The adjoint program generated by TAPENADE is used to compute the gradient of a cost function
- Numerical experiments are done for a three-dimensional downburst.

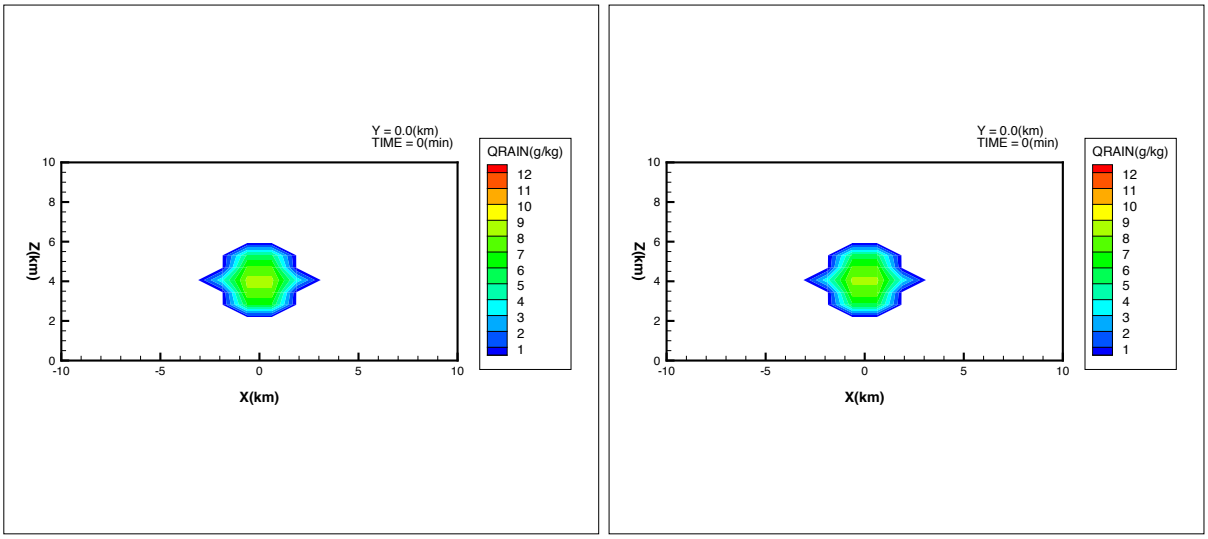



Figure 5: The rainwater field recovered by the optimization (left) and the corresponding one of the initial condition for the simulation (right) in the $x - z$ cross section at $y = 0$.

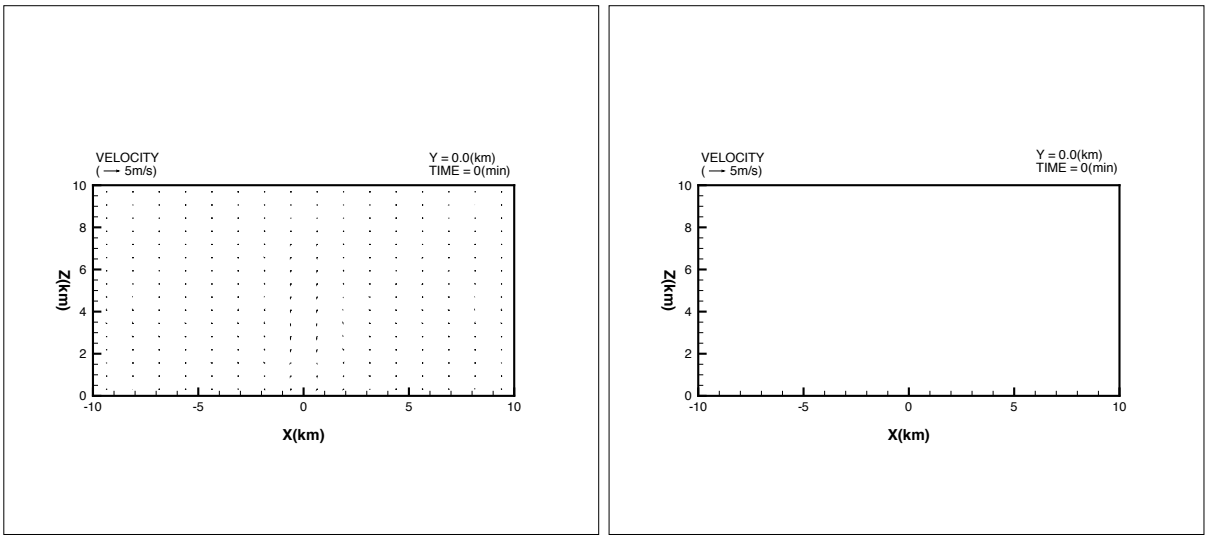



Figure 6: The flow field recovered by the optimization (left) and the corresponding one of the initial condition for the simulation (right) in the $x - z$ cross section at $y = 0$.

- The recovered fields agree reasonably with the initial condition used for observation generation.

## REFERENCES

[1] O. Talagrand, Assimilation of Observations, An Introduction, *J. Met. Soc. Japan*, Special Issue **75**, 1B, pp. 191-209, 1997.

[2] Y. Horibata, Gradient Computation of a Nonlinear Programming Problem for Meteorological Assimilation, *Proc. of European Conference on Computational Fluid Dynamics*, ECCOMAS CFD 2006, Egmond aan Zee, Netherlands, Paper 405, 2006.

[3] Y. Horibata, Comparison of algorithms for solving large-scale optimization problems with bound constraints for meteorological data assimilation, *Proceedings of the 6th European Congress on Computational Methods in Applied Sciences and Engineering*, ECCOMAS 2012, Vienna, Austria, Paper 2811, 2012.

[4] L. Hascoet, V. Pascual, TAPENADE 2.1 user's guide, INRIA, France, 2004.

[5] Y. Horibata, H. Oikawa, Numerical Simulation of Downbursts Using a Terrain-following Coordinate Transformation, *Proc. of the Second European Computational Fluid Dynamics Conference*, Stuttgart, Germany, pp. 1002-1009, 1994.

[6] C. Zhu, R. H. Byrd, P. Lu, J. Nocedal, Algorithm 778: L-BFGS-B, FORTRAN subroutines for large scale bound constrained optimization, *ACM Transactions on Mathematical Software*, **23**, pp. 550-560, 1997.

[7] J.L. Morales, J. Nocedal, L-BFGS-B: Remark on Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization, *ACM Transactions on Mathematical Software*, **23**, Article 7, 2011.

[8] Y. Horibata, Adjoint program generated by automatic differentiation of a meteorological simulation program, and its application to gradient computation, *Proceedings of the 6th International Conference on Computational Methods*, 14th - 17th July 2015, Auckland, G.R. Liu and Raj Das, Paper ID 861, ScienTech Publisher, New Zealand, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EVOLUTIONARY TOPOLOGY OPIMIZATION USING PARAMETERIZED B-SPLINE SURFACE

**Igor Pehnec**[1], **Damir Vučina**[1], **Frane Vlak**[1]

[1] Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture
Rudjera Boškovića 32, Split, Croatia
e-mail: {ipehnec, vucina, fvlak}@fesb.hr

**Keywords:** Topology optimization, parameterization, structural mechanics, topology description function – Level set function.

**Abstract.** *In solid mechanics, the topology optimization process involves two separate fields: displacement fields as well as density fields (material distribution). The level set topology optimization method is based on implicit functions whose iso-lines describe the respective geometry. The traditional level set method is performed by using the Hamilton-Jacobi transport equation, which implies the use of gradient optimization methods that are prone to getting stuck in local extremes. Furthermore, resulting optimal shapes are strongly dependent on the initial guess.*

*In order to provide a robust topology optimization method insensitive to local optima, parameterized level set functions and evolutionary optimization methods (genetic algorithms) are used. Using B-spline parameterized level set functions enables the number of variables to be reduced and the use of genetic algorithms possible.*

*The method is presented on a simple 2D cantilever plate. The objective function is minimum of compliance in one case and minimum of volume in the second case. Due the geometric change, the finite element method requires remeshing. To avoid remeshing, the "Ersatz material" technique is applied. It changes the material properties by setting Young's module of elasticity to zero for elements of the FE mesh where the material should be removed.*

*The proposed method of topology optimization using genetic algorithms is feasible, but a very time-consuming process. For more complex problems, the proposed procedure with a smaller number of generations can be used as a generator of the initial solutions for subsequent faster procedures.*

## 1 INTRODUCTION

The task of topology optimization is to distribute a material in order to optimize the respective response. In structural mechanics the response is usually displacements. Topology optimization as the inverse process of synthesis is a very complex and creative process, which involves many aspects such as experience, ideas, rules, methods, analysis, interaction and other conceptual elements.In literature several methods are developed for topology optimization :

- Bubble method, where in order to optimize e.g. plate with holes, the optimizer is changing the size of the gaps in the observed domain [1], [2], [3], [4], [5]
- Evolutionary structural optimization method (ESO) [6] and [7], where the material is removed (nibbling on) using a material property (modulus of elasticity) where the finite element method gives a smaller stress response. The main problem using this method for topology optimization is that after each optimization step, the geometry is irreversible.
- Solid Isotropic Material with Penalization method (SIMP) [2], [3], [8], [9], [10] using the compliance minimization problem to determine the best material distribution that provides a better stiffness/weight ratio. The variable that describes the distribution of material is the material stiffness, which can have the value of solid material or zero (no material).
- Level set methods [6], [11], [12], [13], [14], [15], [16], [17], [18], [19] based on implicit functions whose iso-lines describe the geometry. Thus, the level set function is able to describe the topological changes of the geometry during optimization process.

A topology description function method [20] was developed from the Level set method. Some researchers studied non gradient methods for topology optimization [21], but in that case the problem is time consuming. A possible solution lies in the parallelization of process [22].

Topology optimization problems that are discussed in the literature can include several areas: parameterization of geometry of the domain, interpolation methods, solving equilibrium and optimization algorithm to optimize shape and topology.
Topology optimization is nearest to the abstract conceptual phase in design. In the phase of shape optimization, effective parameterization of shape becomes crucial for efficient optimization of geometry. Parameterization can be done using some of the functions: NURBS, B-spline, Bezier surface, Radial basis function (RBF) [14], [16], [23], [24].

## 2 MATHEMATICAL FORMULATION

Topology optimization is a complex process that involves a series of numerical procedures. In the next flowchart, the numerical procedure for topology optimization in structural mechanics is presented.

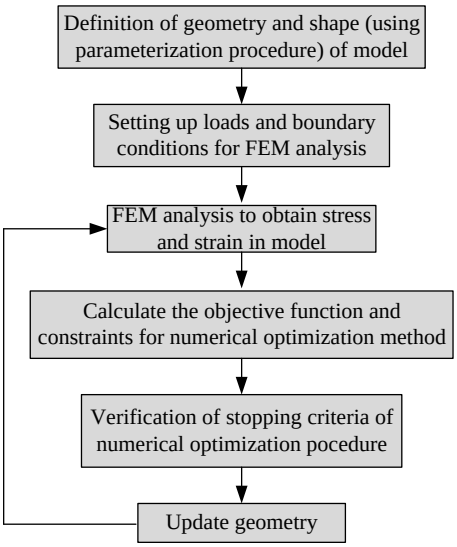



Figure 1: Flowchart of topology optimization in structural mechanics

## 2.1 Level set function (Shape definition)

Level set is an implicit function that describes the geometry. The Level set function lies in $n$+1 dimension with respect to geometry that describes, and its iso-line defines the geometry. Where $n$ is dimension of the observed domain. This feature of the level set function provides the description of the topological changes of the geometry during the optimization process (creating new holes in the domain), without requiring a new parameterization of the geometry. This feature is very important because otherwise the optimization process will have a variable number of unknowns (Figure 2).

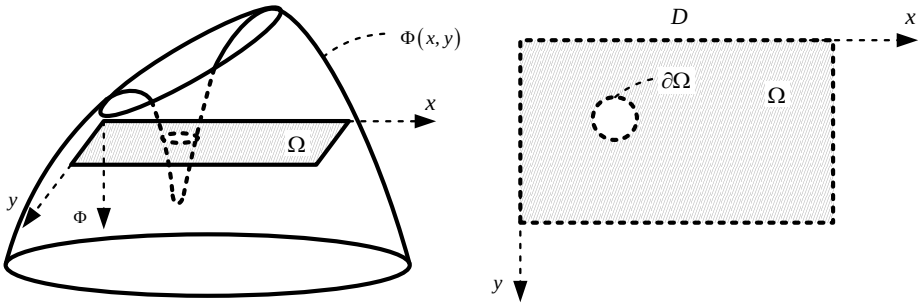



Figure 2: The general implicit function Φ which penetrates the domain D

A rectangular domain $D$, shown in Figure 2, can be partitioned in the following way:

- Material (inside) which is defined as $\Phi(\mathbf{x},t)<0 \quad \forall x \in \Omega \backslash \partial\Omega$,
- Void (outside) which is defined as $\Phi(\mathbf{x},t)>0 \quad \forall x \in D \backslash \Omega$, and
- Boundary $\Phi(\mathbf{x},t)=0 \quad \forall x \in \partial\Omega \cap D$.

where $\Omega$ is the area filled with the material, $\partial\Omega$ is the border of the area $\Omega$, and $D$ is the fixed domain within which the shape and topology of the observed area $\Omega$ are defined.
The information whether the observed point lies inside or outside the area can be provided using the Heaviside step function $H$ [12]- [20]:

$$H\big(\phi(\boldsymbol{x})\big) = \begin{cases} 1 & if \ \ \phi(\boldsymbol{x}) \leq 0, \\ 0 & if \ \ \phi(\boldsymbol{x}) > 0, \end{cases} \tag{1}$$

At a time $t$, the contour function is:

$$\Phi(\boldsymbol{x}, t) = 0 \tag{2}$$

### 2.2 Parametric shape B spline surface

Polynomial expressions, such as the NURBS and B spline surfaces, can be used to describe more complex geometries. In this paper, level set functions are described using B-spline surfaces. In that case, a representive point of the B spline surface or approximation point $\boldsymbol{P}(x,y)$, , is defined in the following way [24]:

$$\boldsymbol{P}(x,y) = \sum_{i_n=0}^{n} \sum_{i_m=0}^{m} \boldsymbol{N}_{i_n,j_n}(x) \cdot \boldsymbol{N}_{i_m,j_m}(y) \cdot \boldsymbol{p}_{i_n i_m} \tag{3}$$

wherw $0 \leq i_n \leq n$ and $0 \leq i_m \leq m$ are rectangular patch. The polynomial components of $\boldsymbol{P}(x,y)$ are defined piecewise on the sets $[x_i , x_{i+1}] \times [y_i , y_{i+1}]$. The vector $x \in [x_0, x_n]$ and $y \in [y_0, y_m]$ is called a knot vector, and it is refined in $x$ and $y$ direction, and $\boldsymbol{p}_{i_n i_m}$ are control points of surface.

The shape function of zero order in x direction is:

$$\boldsymbol{N}_{i,0}(x) = \begin{cases} 1, & x_i \leq x < x_{i+1} \\ 0, & otherwise \end{cases}, \tag{4}$$

and for other orders ($0 \leq i \leq n + j$),

$$\boldsymbol{N}_{i,j}(x) = \frac{x - x_i}{x_{i+j-1} - x_i} \boldsymbol{N}_{i,j-1}(x) + \frac{x_{i+j} - x}{x_{i+j} - x_{i+1}} \boldsymbol{N}_{i+1,j-1}(x) \tag{5}$$

The spline order is $1 \leq j_x \leq n$ and $1 \leq j_y \leq m$, and for single node is presented uniformly as:

$$x_i = \begin{cases} 0, & 0 \leq i \leq j \\ \dfrac{i-j}{n+1-j}, & j+1 \leq i \leq n \\ 1 & n+1 \leq i \leq n+j+1 \end{cases} \tag{6}$$

Analog are able to introduce the shape functions for second direction (y).

B splines have the property of local control since basic functions have non zero values only locally for a few adjacent control points $\boldsymbol{p}$.

### 2.3 Objective function and constraints

In structural mechanics, the objective function for topology optimization problems is usually represented as minimization of compliance with constraints related to displacements and available volume of material [2]:

$$\min_{\substack{\boldsymbol{v} \in U \\ E \in E_{ad}}} l(\boldsymbol{v}, \Phi) \tag{7}$$

$$a_B(\boldsymbol{u}, \boldsymbol{v}, \Phi) = l(\boldsymbol{v}, \Phi) \quad \Rightarrow$$

$$\int_D \boldsymbol{\varepsilon}(\boldsymbol{u}) \cdot \boldsymbol{E}(\boldsymbol{X}) \cdot \boldsymbol{\varepsilon}(\boldsymbol{v}) H(\Phi) d\Omega = \int_\Omega \boldsymbol{f}_v \cdot \boldsymbol{v} \cdot H(\Phi) d\Omega + \int_\Gamma \boldsymbol{f}_s \cdot \boldsymbol{v} \cdot \delta(\Phi) |\nabla \Phi| d\Gamma \tag{8}$$

The Previous expression represents a weak variational form. The index $B$ indicates that bi-linear form $a_B$ depends on the design variables. $U$ is respective virtual (kinematic admissible) displacement field, $\boldsymbol{f}_v$ are the body forces, $\boldsymbol{f}_s$ is boundary traction on the traction part $\Gamma$ of the boundary, $E_{ad}$ is admissible stiffness, $\boldsymbol{E}$ is stiffness matrix, $\boldsymbol{u}$ is equilibrium displacement field, $\boldsymbol{v}$ is virtual displacement field (kinematically permissible), $a(\boldsymbol{u},\boldsymbol{v})$ is virtual work of internal forces, $l(\boldsymbol{v}, \boldsymbol{\Phi})$ is virtual work of external load, $\boldsymbol{X}$ are design variables of shape and topology (represented distribution of material), and $\delta(\Phi)$ is partial derivation of the Heaviside function $H(\Phi)$ (Dirac function).

The linear form of external loads is equal to the sum of the external load $\boldsymbol{f}_v$ per unit volume $\Omega$ and external load $\boldsymbol{f}s$ per unit contour $\Gamma$.
From the above follows that for steady state bilinear forms optimal allocation of material is seeking to provide a maximum stiffness [3]:

$$\max_{E\in E_{ad}} \left( \min_{u\epsilon U,E} \{a_B(\boldsymbol{u},\boldsymbol{v}) - l(\boldsymbol{v})\} \right) \tag{9}$$

In this case, the displacement field and allowable distribution of material are constraints. The function to be minimized is compliance, which provides the optimal distribution of material that will satisfy equilibrium (nested formulation).
The topology optimization problem can be defined as the minimum of volume with respect to the stress constraints. If the domain is divided into finite elements (FE), the minimum volume formulation is:

$$\min_{\mathrm{E}} \mathrm{V}$$
$$\boldsymbol{K}(E_e)\boldsymbol{u} = \boldsymbol{f}_v \quad E_e \in E_{ad} \tag{10}$$

with respect to the stress constraint:

$$\sigma < \sigma_{allowed} \tag{11}$$

where $\boldsymbol{K}$ is stiffness matrix which depends on the material properties, $\boldsymbol{E}_e$ represents the material properties of elements.

## 2.4 Level set method

The objective function, using a Lagrange multiplier method, now looks like [14]:

$$\underbrace{min}_{\Phi(\mathrm{x})}: \qquad L(\boldsymbol{u},\Phi,\lambda) = \frac{1}{2}\int_D \left(\varepsilon(\boldsymbol{u})\right)^T \boldsymbol{E}\varepsilon(\boldsymbol{v})H(\Phi)d\,\Omega + \lambda\left[\int_D H(\Phi)d\,\Omega - \zeta V_0\right] \tag{12}$$

If the level set function (2) is dynamically changed in time, then a continuous velocity field $v_n=dx/dt$ is introduced and time-dependent change of the level set surface is obtained (Hamilton-Jacobi partial differential equation H-J) [15], [16]:

$$\frac{\partial\Phi(\boldsymbol{x},t)}{\partial t} + \nabla\Phi(\boldsymbol{x},t)\cdot\frac{d\boldsymbol{x}}{dt} = 0, \tag{13}$$

where $t$ is the virtual time step.
The H-J partial differential equation can be solved using the "upwinding" procedure (finite differences method in the spatial domain and time stepping scheme), or using a parameterization function to describe level set surface $\Phi(\mathbf{x},t)$.
The upwinding procedure in a domain that is optimized cannot create a new hole, which is its major disadvantage.
The H-J equation can be parameterized using some of the parametric equations (e.g. B-spline) (3), so that separation of space and time variables is achieved. That procedure provides an uncoupled system of ordinary differential equations [14], [15], from (3) and (13):

$$\boldsymbol{\phi}^T(\boldsymbol{x})\frac{d\boldsymbol{P}(t)}{dt} + v_n\left|\left(\nabla\boldsymbol{\phi}(\boldsymbol{x})\right)^T\boldsymbol{P}(t)\right| = 0 \tag{14}$$

where $\boldsymbol{P}(t)$ are B-spline approximation function.
The velocity ($v_n$) can be obtained by considering the minimum of the potential energy (imposition of "external" velocity field).
Using a permutation of boundaries of domain on the objective function (12), the velocity of all points of level set function [25] is obtained:

$$v_n = \boldsymbol{\varepsilon}^T\boldsymbol{E}\boldsymbol{\varepsilon} - \lambda \tag{15}$$

From previous equations can be seen that the level set method is based on a gradient optimization method. Therefore, the main disadvantage of the level set method as well as all gradient methods of topology optimization is dependence on the initial solution.
A topology optimization method using B-spline parameterization of the level set surface is proposed. The optimization variables are control points of the B-spline. Moving the B-spline nodes causes a change of Level set function, as well as the observed domain (iso-line of level set function). Such a formulation allows a usage of genetic algorithms as optimization method. Genetic algorithms are global optimization methods, but numerically highly expensive. B-spline parameterization methods are reducing the required number of variables, which allows application of genetic algorithm optimization in reasonable time.
In the presented method, optimization variables are changing the geometry by using the level set function (2). The binary code of one individual, *z* coordinate of parameter $\boldsymbol{p}$ from equation (3) with resolution of 8 bits, is presented in Figure 3:

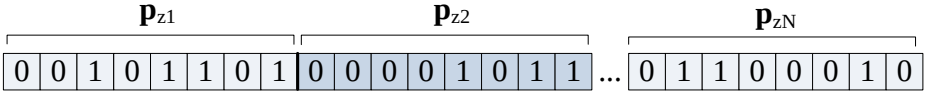


Figure 3 Optimization variables in binary code

Using the operators of selection, crossover and mutation, the genetic algorithm changes the population (bits in binary code) of individuals in the next generations. Some of binary coded phenotypes are shown in Figure 4.

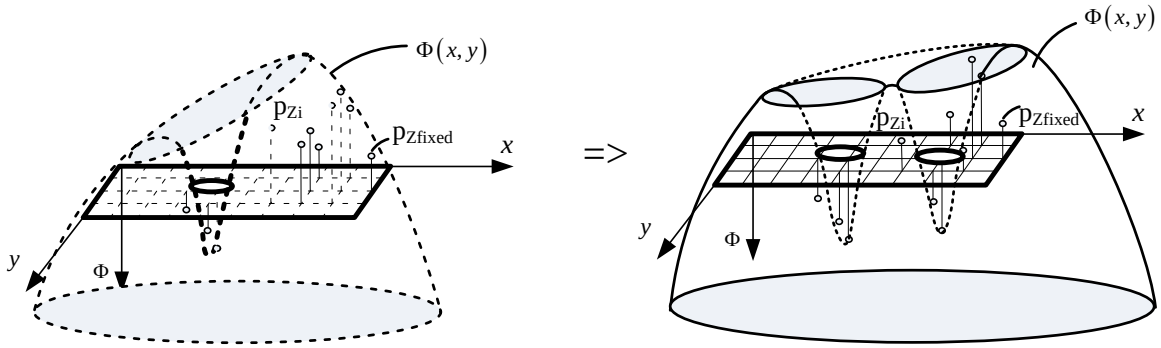


Figure 4 Geometry change with B-spline parametrized Level set implicit function

Such a description of the geometry also allows topological changes of the model.
The objective function (12) in each time step is calculated utilizing FEM analysis and simple structurally active elements counter.
Also, the finite element method requires considerable time for executing calculations, especially in situations when the change of geometry occurs in each step of the optimization procedure, which results in a high time-cost re-meshing method. To avoid re-meshing, each element of stiffness matrix is represented with appropriate value (1):

- null – if level set function larger than zero, or
- elasticity modulus of material ($E$) – if level set function equal or smaller than zero.

This procedure is called „Ersatz material“, and it is not numerically expensive.

## 3 TEST CASE AND RESULTS

As a test case for the proposed topology optimization procedure, a 2D cantilever plate is defined as shown below:

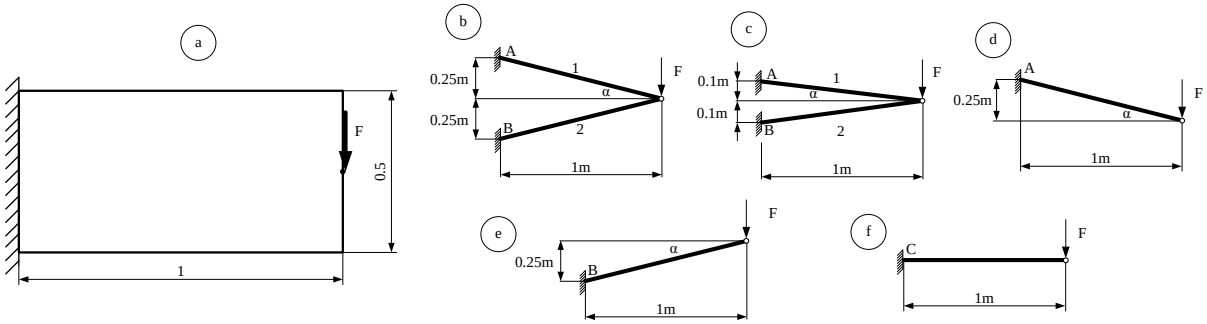



Figure 5: The geometry, loads and boundary conditions of cantilever plate for topology optimization (a) and some of discrete solutions (b, c, d, e)

Some of the expected potential discrete solutions are shown in Figure 5 b, c, d, e, f which can be found as optimal local extremes.

For structural topology optimization of the cantilevered beam defined in Figure 5, a B-spline parameterized level set surface was used. Optimization variables are the $z$ coordinates of B-spline control points $\boldsymbol{p}_{ij}$ in Eq. (3). Changing of the B-spline control points has the effect to change the geometry of model. As the geometry of the model is discretized by finite elements, each element of discrete domain has its own material properties, respecting the Heaviside function (1) of the FE geometric center point. If the coordinate of the center point of the observed FE takes a negative value of the approximated level set function, the value of Young's modulus of elasticity $E$= 70[GPa] is assigned to the element. Otherwise, if the coordinate of the center point of the observed FE takes a positive value of the approximated level set function, then zero value is assigned to the element material property ( $E$= 0, i.e. that element is non-load-bearing part of the structure).
In this paper, a genetic algorithm for structural topology optimization is used. Genetic algorithms (GA), as global search methods for nonlinear optimization, are not sensitive to local extremes of the objective function.
All numerical procedures were performed using in-house MATLAB code, inspired by [8], [9]. The number of finite elements of the discretized 2D plate (Figure 5) is 3200. If the objective function is defined as the minimum of the volume with stress constraint, and 576 B-spline control points i.e. 576 optimization variables are chosen to describe a level set function the proposed procedure provides:

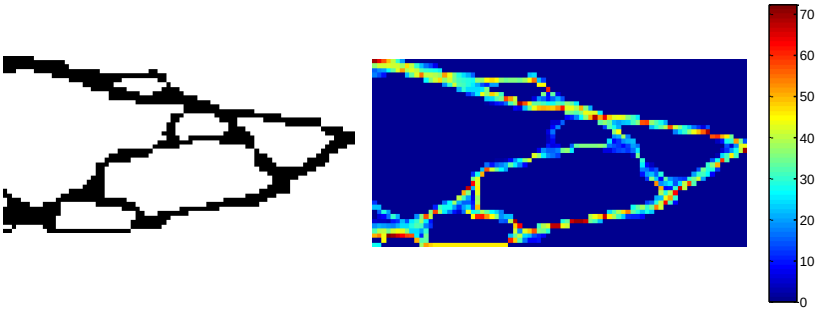



Figure 6: Topology optimization of 2D cantilever plate with 576 variables of optimization for minimum volume problem

Figure 6 shows the solution of topology optimization 2D cantilever plate using a genetic algorithm after 200 generations. The number of the population is 2304, and allowable stress is 70 GPa. The time required for the represented numerical procedure was 35 hours on Intel 3.10 GHz 4 core CPU. From Figure 6 it can be seen that convergence is not achieved due to large number of optimization variables. In order to achieve convergence, the number of generations should be increased, but in that case the numerical procedure might be too long. However, an expected solution is obtained. The reduction of volume is 20.3% (648 elements of the initial 3200 elements).

The numerical procedure can be accelerated by using parallelization of FE processes on a cluster [22], or by reducing the number of variables. If 144 B-spline control points, i.e. optimization variables, are chosen to describe a level set function, the time required for represented numerical procedure was reduced to 14 hours (Figure 7):

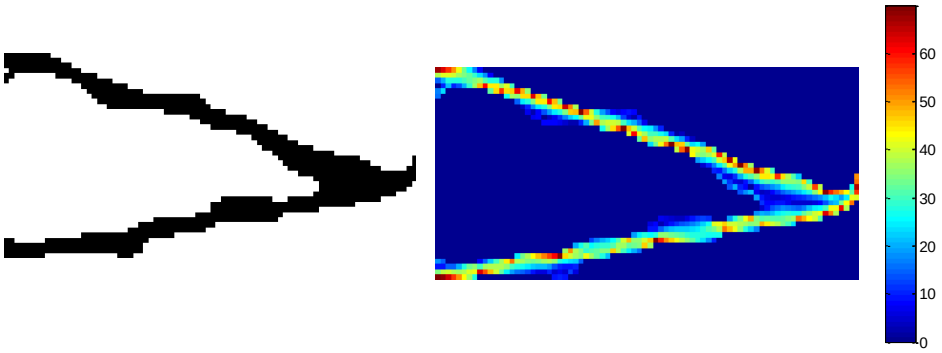



Figure 7 Topology optimization of 2D cantilever plate with 144 variables of optimization

Number of population is 1296 over 200 generations, which was sufficient for the procedure to converge to the global optimum. The reduction of volume is 20% (640 elements). Better optimization results were achieved using a reduced numbers of variables (144) due to a lower search field. A better, final solution can be provided using numerical filters.

The next example provides solution for a minimum compliance problem with volume constraint ($\zeta$ = 0.3, Eq. 11) –Figure 8:

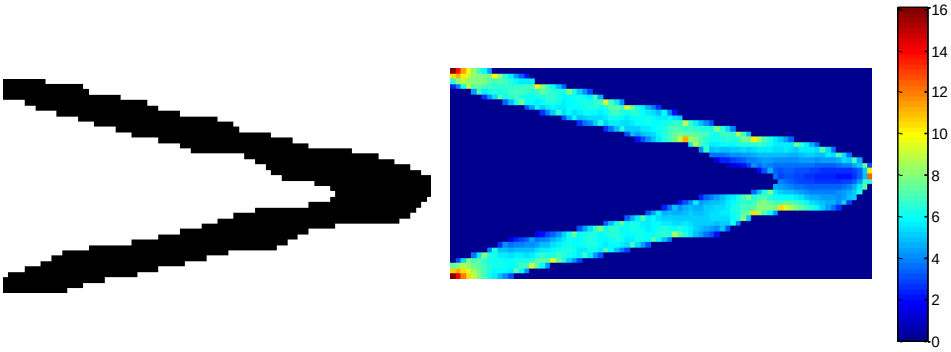



Figure 8: Topology optimization of 2D cantilever plate with 144 optimization variables for minimum compliance problem

Figure 8 shows that the parametrized level set function and the genetic algorithm are able to provide a good solution of topology optimization. The number of the population is 2304 with 200 generations. The time required for represented numerical procedure was 14 hours on Intel 3.10 GHz 4 core CPU.

The solutions presented in Figure 7 and Figure **8** are very similar to the structural type shown in Figure 5b. Simple structural analysis, considering four types of the truss or beam structures shown in Figure 5b-5f, can confirm that this structural type is optimal one with respect to the previously defined objectives and constraints. This conclusion can be used for justification of the quality of the solutions obtained by the method presented in this paper. Moreover, proposed method may serve as a generator of initial solutions for the traditional gradient topology optimization methods.

## 4 CONCLUSIONS

Topology optimization as a process of synthesis of shape and topology in structural mechanics is very complex and demanding numerical procedure. In this work a method for topology optimization of a 2D structure (cantilever plate) using a paremetrized level set surface for geometric description was proposed. The optimization variables in the proposed procedure are *z*-coordinates of B-spline control points $\boldsymbol{p}_{ij}$ defined in Eq. (3). The parameters of the B-spline surface that describe the level set surface produce a change of 2D plate geometry. The procedure avoids the Hamilton-Jacobi transport equation and allows the application of non-gradient and evolutionary optimization methods. Parametrization of level set surface using B-spline results in reduced number of variables of optimization process, as well as good shape representation of model. Moreover, reducing the number of the optimization variables decreases the required numerical analysis time. A reduced number of variables allows application of numerically demanding genetic algorithm method. The global optimization method (GA) gives a great advantage over conventional topology optimization methods mostly based on gradient methods. Also, the proposed method allows the application of the minimum volume as objective function, which is, due to the discrete character of the objective function and the complexity of the problem, impossible for gradient optimization methods without additional numerical procedures. Instead of remeshing FE during the geometry changes, the "Ersatz material" techniques are used as computation ally cheaper procedures.

The results show the applicability of the proposed topology optimization method by using a parameterized level set function and genetic algorithm. The main application of the proposed method of topology optimization is to obtain good initial solutions for the traditional level set method, which has a large dependence on the selected initial solution. The acceleration of the proposed procedure is possible by using a cluster of computers, or additional numerical algorithms for parameterization and optimization.

**FUNDING:**
This work was supported by the Croatian Science Foundation [grant number IP-2014-09-6130].

## REFERENCES


[1] H.A. Eschenauer, V.V. Kobelev, A. Schumacher, Bubble method for topology and shape optimization of structures, *Structural and Multidisciplinary Optimization*, 8(1), pp. 42-51, 1994.

[2] M.P. Bendsoe, O. Sigmund, Topology Optimization Theory, Methods and Applications, *Springer*, Berlin, Germany, 2004.

[3] G.I.N. Rozvany, Topology optimization in structural mechanics, *Springer, CISM*, Udine Italy, 1997.

[4] H.A. Eschenauer, N. Olhoff, Topology optimization of continuum structures: a review, *Applied Mechanics Reviews* 54(4), pp: 331–390, 2001.

[5] H.A. Eschenauer, N. Olhoff, Topology optimization of continuum structures: A review, *American Society of Mechanical Engineers ASME*, 54(4), pp: 331-390, 2001.

[6] H. Jia, H. G. Beom, Y. Wang, S. Lin, B. Lui, Evolutionary level set method for structural topology optimization, *Computer & structures* 89(5-6), pp: 445-454, 2011.

[7] Y.M. Xie, G.P. Steven, A simple evolutionary procedure for structural optimization, *Computers & Structures* 49(5), pp: 885–896, 1993.

[8] E. Andreassen, A. Clausen, M. Schevenels, B. S. Lazarov, O. Sigmund, Efficient topology optimization in MATLAB using 88 lines of code, *Structural and Multidisciplinary Optimization* 43, pp: 1-16, DOI 10.1007/s00158-010-0594-7, 2010.

[9] O. Sigmund, A 99 line topology optimization code written in Matlab, *Structural and Multidisciplinary Optimization* 21(2), pp: 120-127, 2001.

[10] G.I.N., Rozvany, Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics, *Structural and Multidisciplinary Optimization* 21, pp: 90-108, 2001.

[11] S. Osher, J.A. Sethian, Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations, *Journal of Computational Physics* 79(1), pp: 12-49, 1988.

[12] S. Osher, R.P. Fedkiw, Level Set Methods: An Overview and Some Recent Results, *Journal of Computational Physics* 169, pp: 463-502, 2001.

[13] S. Osher, R. Fedkiw, Level Set Methods and Dynamic Implicit Surfaces, Springer, *Applied Mathematical Sciences* 153, USA, 2003.

[14] Z. Luo, L. Tong, Z. Kang, A level set method for structural shape and topology optimization using radial basic functions, *Journal of Computers and Structures* 87, pp: 425-434, 2009.

[15] M. Y. Wang, X. Wang, D. Guo, A level set method for structural topology optimization, *Computer methods in applied mechanics and engineering* 192, pp: 227–246, 2003.

[16] S. Wang, M.Y. Wang, Radial Basis Functions and Level Set Method for Structural Topology Optimization, *International Journal for Numerical Methods in Engineering* 65, pp 2060-2090, 2006.

[17] T. Belytschko, S. P. Xiao, C. Parimi, Topology optimization with implicit functions and regularization, *International Journal for Numerical Methods in Engineering* 57, pp: 1177–1196, 2003.

[18] X. Guo, K. Zhao, M. Y. Wang, A new approach for simultaneous shape and topology optimization based on dynamic implicit surface function, *Control and Cybernetics* 34, pp: 255-282, 2005.

[19] N. P. van Dijk, K. Maute, M. Langelaar, F. van Keulen, Level-set methods for structural topology optimization: a review, *Structural and Multidisciplinary Optimization* 48 (3), pp: 437-472, 2013.

[20] M.J. de Ruiter, F. van Keulen, Topology optimization using a topology description function, *Structural and Multidisciplinary Optimization* 26, pp: 406-416, 2004.

[21] O. Sigmund, On the usefulness of non-gradient approaches in topology optimization, *Structural and Multidisciplinary Optimization* 43, pp: 589-596, 2011.

[22] D. Vučina, Ž. Lozina, I. Pehnec, Ad-hoc cluster and workflow for parallel implementation of initial-stage evolutionary optimum design, *Structural and Multidisciplinary Optimization*, 45(2), pp: 197-222, DOI: 10.1007/s00158-011-0687-y, 2011.

[23] P.W. Christensen, A. Klarbring, „An Introduction to Structural Optimization“, Springer, 2009.

[24] P.J. Schneider; D.H. Eberly, „Geometric Tools for Computer Graphics“, Morgan Kaufmann, San Francisco, USA, 2002

[25] G. Allaire, F. Jouve, A-M. Toader, Structural optimization using sensitivity analysis and a level-set method, *Journal of Computational Physics* 194, pp: 363-393, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# GEARBOX DESIGN VIA MIXED-INTEGER PROGRAMMING

**Bastian Dörig[1], Thorsten Ederer[1,2], Peter F. Pelz[2], Marc E. Pfetsch[1], and Jan Wolf[3]**

[1] Department of Mathematics, Technische Universität Darmstadt
Dolivostr. 15, 64293 Darmstadt, Germany
e-mail: {doerig, ederer, pfetsch}@mathematik.tu-darmstadt.de

[2] Department of Mechanical Engineering, Technische Universität Darmstadt
Otto-Berndt-Str. 2, 64287 Darmstadt, Germany
e-mail: {thorsten.ederer, peter.pelz}@fst.tu-darmstadt.de

[3] Department of Economics, Universität Siegen
Hölderlinstr. 3, 57076 Siegen, Germany
e-mail: jan.wolf@uni-siegen.de

**Keywords:** Mixed-Integer Nonlinear Programming, Technical Operations Research, Global Optimization, Gearbox Design, Mechanical Transmission System.

**Abstract.** *Gearboxes are mechanical transmission systems that consist of multiple gear wheels and shafts. The transmission ratios of the gears are determined by the size and interconnection of these components. Gearboxes have to adhere to very strict requirements regarding weight, production costs, and available space. Moreover, the load cases as a result of motor-vehicle-pairings are often not clear a priori. Therefore, automobile manufacturers are confronted with a multicriteria design problem under uncertainty.*

*In this work, we present an approach on how to formulate the gearbox design problem as a mixed-integer nonlinear program. This enables us to compute provably globally optimal gearbox designs. We show how different degrees of freedom, input parameters and the numerical accuracy influence the computation time and the quality of solutions.*

**Acknowledgement.** The authors thank the German Research Foundation DFG for funding this research within the Collaborative Research Center SFB 805 “Control of Uncertainties in Load-Carrying Structures in Mechanical Engineering”.

## 1 INTRODUCTION

Gearboxes are mechanical transmission systems that consist of multiple gear wheels and shafts, see Figure 1. The transmission ratios of the gears are determined by the size and interconnection of these components. As central element of the drive train, gearboxes are highly relevant for the efficiency and durability of motor vehicles. They have to adhere to very strict requirements regarding weight, production costs, and available space. Moreover, the load cases as a result of motor-vehicle-pairings are often not clear a priori. Hence, automobile manufacturers are confronted with a multicriteria design problem under uncertainty.

A traditional approach to system design problems is to manually identify and compare a selection of promising proposals. However, the gearbox is a highly complex system. A comparative analysis of every imaginable gearbox topology is doomed to failure in view of the combinatorial explosion. Engineers work around this problem through a bottom-up design approach: Individual components are optimized and combined to subsystems, which in turn are optimized and combined to larger subsystems. This approach usually leads to good designs, but one cannot make any statement about the objective quality of the resulting systems.

Optimization problems in engineering are often tackled with probabilistic methods, e.g. swarm intelligence or genetic algorithms. They are easy to adapt to new problems and often find good solutions in a short amount of time. However, it is unclear how long such heuristics need to keep running. We do not know whether they will be able to find even better solutions or not. In contrast, mathematical optimization methods are capable of finding *provably globally optimal* solutions by quantifying how much potential for improvement remains at any time.

In this paper, we present a mixed-integer nonlinear program (MINLP) that generates gearbox system designs which are provably globally optimal with respect to a given objective. As example gearboxes, we focus on dual-clutch transmission systems. Technical details on the topic of gearboxes can be found e.g. in [1] and [2].

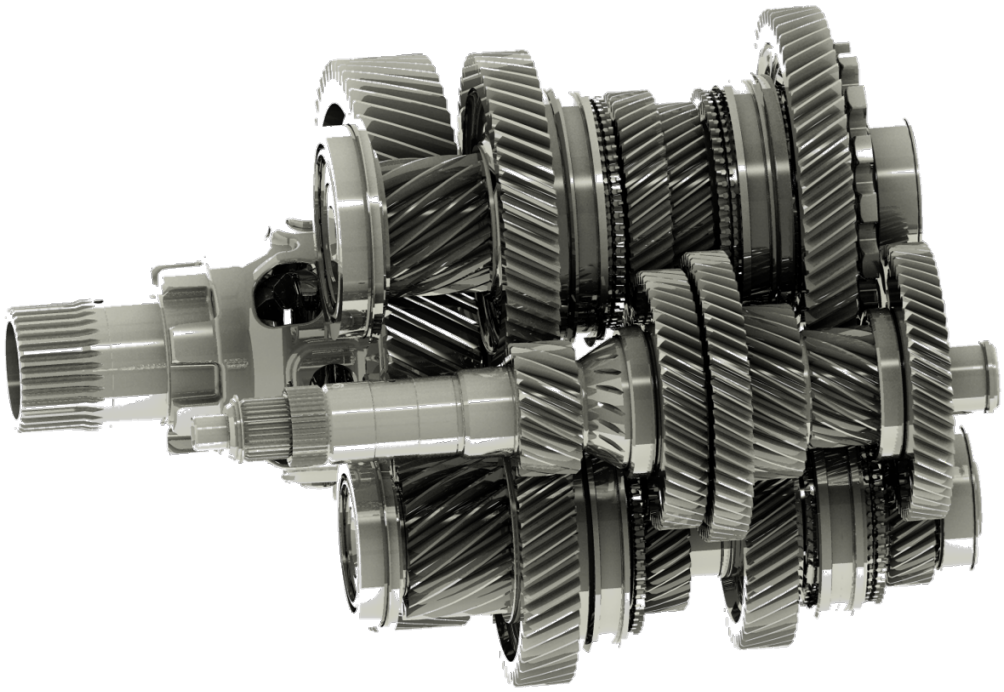


Figure 1: Rendered image of a dual-clutch transmission (Source: Volkswagen AG).

The paper is organized as follows: In Section 2, we introduce the principle of dual-clutch transmissions. In Section 3, we present the mathematical model and discuss possible variations. In Section 4, we show some computed gearboxes and compare runtimes of two MINLP solvers.

## 2 DUAL-CLUTCH TRANSMISSIONS

The abstract task of a gearbox is to transfer power (given by a torque and an angular velocity) from the input shaft (which is driven by the motor) to the output shaft (which is connected to the differential and moves the wheels). The power flow is established by engaging suitable gear wheels. The transformation ratio of the gear can be adjusted by changing the ratio of the gear wheel radii: an increase of torque causes a reduction of angular velocity and vice versa. Large gear wheels (as a result of large demanded gear ratios) are avoided by introducing a countershaft and realizing the gear as a series of two gear wheel pairs (pre-transmission and post-transmission) with smaller gear wheel radii each. Two or three countershafts instead of one can be used to reduce the axial expansion of the gearbox.

Many modern gearboxes are so-called dual-clutch transmission systems. That is, there are two input shafts which can separately be rotated and clutched to the motor. Even numbered gears are assigned to one input shaft and odd numbered gears are assigned to the other input shaft. Thereby, changing gears becomes possible without interruption of traction. To save space, the two input shafts are realized as a long full shaft fitted inside a shorter hollow shaft.

With respect to gear wear and noise impact, it is advantageous to let the gear wheels always engage with each other. In that case, a design modification is required in order to prevent blockage of the transmission: The gear wheels are pivot-mounted on the shafts, i.e., they can rotate independently from each other. A power flow is established by coupling some gear wheels to their respective shafts using synchronization systems. Due to their high complexity, using as few of these systems as possible has a high priority. For instance, a sliding sleeve can be placed between two gear wheels and synchronize either one of them, but not simultaneously. Alternatively, selected gear wheels can be mounted fix onto a shaft to omit a synchronizer.

To simplify the placement and interaction of components, we can assume that gear wheels and sliding sleeves need approximately the same axial space: In this way, we can infer a discrete number of alignment planes in a gearbox of given expansion. A shaft and an plane together determine a position where a component can be placed. Gears that are placed on the same plane can potentially interact with each other.

### The manual design procedure

The conventional design procedure of a gearbox is a multi-phase process chain. The planning department is confronted with the task to design a new gearbox for a vehicle. There may be early assessments of the available space and of the forces the gears will have to bear, but these plans may change in later development stages. In an abstract view, the planner has to pass through the following three planning stages:

**Topology** Some initial design choices like the number of shafts, the number of gears and the desired gear transmission ratios can be based on experience, empirical values or quick calculations. These decisions become more complex, if the same topology must be shared by multiple vehicle models: While being economically favorable, this approach may force the planner to make compromises on the performance of the gearbox in later development stages. For instance, fixed shaft distances are a restriction on the gear wheel radii and therefore on the possible transmission ratios.

**Transmission** In this stage, the requested transmission ratios have to be realized by placing gear wheels onto shafts. The planner has to choose gear wheel radii that not only lead to the right transmission ratio for each gear, but that also fit well (i.e. in a space-saving way) into the gearbox topology. An additional degree of freedom is the balance between pre-transmission and post-transmission. The height of the gearbox is a consequence of the gearbox topology as a whole. As a simplification for this stage, the gear wheels are considered as toothless, i.e., the planner expects that the wheels can be geared in a subsequent stage without too many issues.

**Gearing, Efficiency and Misalignment** After the topology has been fixed, a range of extended feasibility criteria has to be checked. Firstly, the gear wheels have to be endowed with teeth. The tooth number along the diameter of each wheel is an integer, so not all gear wheel combinations engage equally well. Secondly, the gear realizations and the gearing of the wheels influence the transmission efficiency of the gearbox. Thirdly, shafts bend as a side effect of the power flow. If they bend to much, the teeth become misaligned and cause increased wear and acustic noise.

In each stage, two things can happen: (a) the planner realizes that his decisions in earlier stages are not compatible with the requirements of the next stage, or (b) the outer requirements for the design are updated. In both cases, the planner has to go back to an earlier stage and incorporate the new information into the current design proposal. This process is iterated until the planner and the company can settle on a final design.

### An automated design procedure

To improve the manual design process, we have to view the system as a whole. However, the extended feasibility criteria in the third stage of the design procedure are highly complex and cannot be modeled within the scope of this work. Therefore, we focus on the first two stages of the design process: the topology and the transmission. Our aim is to find optimal gearbox proposals in a short amount of time. On the one hand, this enables the planner to consider a large number of possible requirement changes in advance. On the other hand, the planner can focus on technical aspects in stage three instead of repeating the work in earlier stages.

## 3 A GEARBOX DESIGN MODEL

The aim of the following model is to find an optimal gearbox design within a given axial space and for demanded total transmission ratios. In the following, capital letters denote sets or parameters (i.e. the problem input), small letters denote indices or continuous decision variables, and greek letters denote binary variables. An overview of all decision variables is given in Table 1. The full model is shown on the next page. The constraints will be explained in order of appearance. Parameters are explained together with the constraints in which they occur.

Three sets characterize the topology of the gearbox. Let $G = \{-1, 1, 2, \dots\}$ denote the set of gears (including one reverse gear $-1$), let $P = \{1, 2, \dots\}$ denote the set of planes that fit into the axial space, and let $D = \{1, 2\}$ denote the set of exactly two drive shafts. A plane and a shaft together determine a position where a component can be placed.

The objective function combines three objective criteria: (i) minimize the height of the gearbox, (ii) minimize the number of sliding sleeves, and (iii) minimize the number of gear wheels on the input shaft, cf. Eq. (1).

Table 1: Decision variables of the mixed-integer nonlinear program.

| Var. | Description | Domain |
|---|---|---|
| $i_g$ | pre-transmission of gear $g$ | $[I^{\min}, I^{\max}]$ |
| $j_d$ | post-transmission from drive shaft $d$ to the output shaft | $[J^{\min}, J^{\max}]$ |
| $r_g$ | gear wheel radius on the input shaft belonging to gear $g$ | $[R^{\min}, R^{\max}]$ |
| $s_g$ | gear wheel radius on the drive shaft belonging to gear $g$ | $[S^{\min}, S^{\max}]$ |
| $t_p$ | gear wheel radius on the input shaft in plane $p$ | $[R^{\min}, R^{\max}]$ |
| $u_{p,d}$ | gear wheel radius on drive shaft $d$ in plane $p$ | $[S^{\min}, S^{\max}]$ |
| $v_d$ | maximum gear wheel radius on drive shaft $d$ | $[S^{\min}, S^{\max}]$ |
| $y_d$ | gear wheel radius on drive shaft $d$ of the final drive | $[Y^{\min}, Y^{\max}]$ |
| $z$ | gear wheel radius on the output shaft of the final drive | $[Z^{\min}, Z^{\max}]$ |
| $\xi_{g,p,d}$ | indicator whether gear $g$ is realized on drive shaft $d$ in plane $p$ | $\{0,1\}$ |
| $\gamma_{p,d}$ | indicator whether any gear is realized on drive shaft $d$ in plane $p$ | $\{0,1\}$ |
| $\delta_{g,p}$ | indicator whether gear $g$ is realized on any drive shaft in plane $p$ | $\{0,1\}$ |
| $\zeta_p$ | indicator whether any gear is realized on any drive shaft in plane $p$ | $\{0,1\}$ |
| $\sigma_{p,d}$ | indicator whether a sleeve is located on drive shaft $d$ in plane $p$ | $\{0,1\}$ |
| $\phi_p$ | indicator whether the input shaft is a full shaft in plane $p$ | $\{0,1\}$ |
| $a_d$ | distance between input shaft and drive shaft $d$ | $[A^{\min}, A^{\max}]$ |
| $b_d$ | distance between drive shaft $d$ and output shaft | $[B^{\min}, B^{\max}]$ |
| $c$ | distance between the drive shafts | $[C^{\min}, C^{\max}]$ |
| $h$ | pseudo-height of the gearbox | $[H^{\min}, H^{\max}]$ |

The exact height along the vertical vehicle axis is quite difficult to express, because the two drive shafts need not be horizontally aligned. Therefore, we use the pseudo-height $h$ defined by Eq. (24) as approximation of the real height. The weighing parameters $W_1$ and $W_2$ control how the objective criteria are balanced. This decision is at the discretion of the planner. For our benchmarks, we used $W_1 \gg 1$ and $W_2 \ll 1$, i.e., we tolerate a larger gearbox to save a sliding sleeve, but not to save a gear wheel on the input shaft.

Eq. (2) determines the overall transmission ratio for each gear. Since these overall ratios are empirical values, we allow for small deviations of $5\,\%$ from the reference values in hope of achieving a more compact design. The interval $[K_g^{\min}, K_g^{\max}]$ gives the range of acceptable overall ratios of gear $g$. The sum expression identifies the matching post-transmission $j_d$ for the pre-transmission $i_g$ of gear $g$. Eqs. (3) and (4) relate the gear wheel radii with the pre-transmission and respectively post-transmission ratios. The sign function is needed to allow for the negative pre-transmission ratio $i_{-1}$ of the reverse gear. Note that we do not need to include the radius $s_{\text{int}}$ of the intermediate wheel that engages with the input shaft gear wheel and the reverse gear wheel, since $i_{-1} = (s_{-1}/s_{\text{int}}) \cdot (s_{\text{int}}/r_{-1})$.

Eqs. (5) to (8) specify relations between various gear assignment indicators. The most general indicator $\xi_{g,p,d}$ becomes active (equal to one) if and only if the gear $g$ is established along the position given by plane $p$ and drive shaft $d$. This position specifies the gear wheel that needs to be synchronized to establish the power flow. The auxiliary variables $\gamma_{p,d}$, $\delta_{g,p}$ and $\zeta_p$ each indicate if at least one $\xi_{g,p,d}$ over some set is active. Additionally, Eq. (6) ensures that each gear $g$ is realized exactly once.

minimize
$$h + W_1 \cdot \sum_{p\in P}\sum_{d\in D} \sigma_{p,d} + W_2 \cdot \sum_{p\in P} \zeta_p \tag{1}$$

subject to
$$K_g^{\min} \le i_g \cdot \sum_{d\in D} j_d \cdot \sum_{p\in P} \xi_{g,p,d} \le K_g^{\max} \qquad \forall g \in G \tag{2}$$
$$r_g \cdot i_g = s_g \cdot \operatorname{sign}(g) \qquad \forall g \in G \tag{3}$$
$$y_d \cdot j_d = z \qquad \forall d \in D \tag{4}$$
$$\delta_{g,p} = \sum_{d\in D} \xi_{g,p,d} \qquad \forall g \in G,\, p \in P \tag{5}$$
$$1 = \sum_{p\in P} \delta_{g,p} \qquad \forall g \in G \tag{6}$$
$$\gamma_{p,d} = \sum_{g\in G} \xi_{g,p,d}\,, \quad \zeta_p \ge \gamma_{p,d} \qquad \forall p \in P,\, d \in D \tag{7}$$
$$\zeta_p \le \sum_{d\in D} \gamma_{p,d} \qquad \forall p \in P \tag{8}$$
$$\gamma_{p,d} + \sigma_{p,d} \le 1\,, \quad \gamma_{p,d} \le \sum_{\substack{p'\in P \\ |p'-p|=1}} \sigma_{p',d}\,, \quad \xi_{-1,p,d} \le \gamma_{p,3-d} \qquad \forall p \in P,\, d \in D \tag{9}$$
$$\phi_p \ge \phi_{p+1} \qquad \forall p \in P,\, p < |P| \tag{10}$$
$$\delta_{g,p} \le \begin{cases} \phi_g & \text{if } g \text{ is odd} \\ 1-\phi_g & \text{if } g \text{ is even} \end{cases} \qquad \forall p \in P,\, g \in G, g > 0 \tag{11}$$
$$\delta_{-1,p} \le 1 - \phi_p \qquad \forall p \in P \tag{12}$$
$$r_g = \sum_{p\in P} t_p \cdot \delta_{g,p}\,, \quad s_g = \sum_{p\in P}\sum_{d\in D} u_{p,d} \cdot \xi_{g,p,d} \qquad \forall g \in G \tag{13}$$
$$t_p \ge R^{\text{full}} \cdot \phi_p + R^{\text{hollow}} \cdot (1-\phi_p) \qquad \forall p \in P \tag{14}$$
$$t_p \le R^{\text{full}} \cdot \phi_p + R^{\text{hollow}} \cdot (1-\phi_p) + R^{\max} \cdot \zeta_p \qquad \forall p \in P \tag{15}$$
$$u_{p,d} \ge S^{\min} \cdot (1-\sigma_{p,d}-\gamma_{p,d}) + S^{\text{sync}} \cdot \sigma_{p,d} + S^{\min} \cdot \gamma_{p,d} \qquad \forall p \in P,\, d \in D \tag{16}$$
$$u_{p,d} \le S^{\min} \cdot (1-\sigma_{p,d}-\gamma_{p,d}) + S^{\text{sync}} \cdot \sigma_{p,d} + S^{\max} \cdot \gamma_{p,d} \qquad \forall p \in P,\, d \in D \tag{17}$$
$$v_d \ge u_{p,d} \qquad \forall p \in P, d \in D \tag{18}$$
$$a_d \ge t_p + u_{p,d} + Q \cdot (1-\gamma_{p,d}) + 2 \cdot Q \cdot \xi_{-1,p,d} \qquad \forall p \in P,\, d \in D \tag{19}$$
$$a_d \le t_p + u_{p,d} + A^{\max} \cdot (1-\gamma_{p,d}) \qquad \forall p \in P,\, d \in D \tag{20}$$
$$a_d \ge R^{\text{hollow}} + y_d + Q\,, \qquad b_d = y_d + z \qquad \forall d \in D \tag{21}$$
$$c \ge \sum_{d\in D} u_{p,d} + Q \cdot \big(\sum_{d\in D} \gamma_{p,d} - \delta_{-1,p}\big) \qquad \forall p \in P \tag{22}$$
$$c \le \sum_{d\in D} u_{p,d} + C^{\max} \cdot (1-\delta_{-1,p}) \qquad \forall p \in P \tag{23}$$
$$h \ge c + \sum_{d\in D} v_d\,, \quad h \ge 2 \cdot z \tag{24}$$

Eq. (9) (left) demands that at most one component may be placed on a position and Eq. (9) (middle) ensures that each gear can be synchronized by a neighbouring sleeve. If the reverse gear wheel is placed on a position, another gear wheel must be placed on the other drive shaft in the same plane, cf. Eq. (9) (right).

The input shaft is divided into a full shaft and a hollow shaft, signalled by the fullness indicator $\phi_p$ for each plane $p$. There must only be one transition from full shaft to hollow shaft, so these indicators need to have a monotoneous behaviour, cf. Eq. (10). Odd numbered gears are established on full shaft planes and even numbered gears are established on hollow shaft planes, cf. Eq. (11). The reverse gear is also placed on the hollow shaft, which allows for quick shunting movements without interruption of traction, cf. Eq. (12).

The variables $r_g$ and $s_g$ represent the gear wheel radii from the point of view of the gears. The variables $t_p$ and $u_{p,d}$ represent the radii from the point of view of the positions inside the gearbox. Eq. (13) identifies variables with the same meaning. If there is no gear wheel on plane $p$, Eqs. (14) and (15) set the radius at this position to the input shaft radius (concretely, they select the radius $R^{\text{full}}$ of the full shaft or the radius $R^{\text{hollow}}$ of the hollow shaft).

Eqs. (16) and (17) determine the drive shaft radii $u_{p,d}$: If the synchronizer indicator $\sigma_{p,d}$ is active, the radius must equal the synchronizer radius $S^{\text{sync}}$ (which we assume as fixed), and if no indicator is active, the radius at this position must equal the radius $S^{\text{min}}$ of the drive shaft. The maximum radius $v_d$ of gear wheels on the drive shaft $d$ in Eq. (18) is needed later for computing the pseudo-height $h$.

Eqs. (19) to (23) determine the pairwise distances between shafts. The distance $a_d$ between the input shaft and drive shaft $d$ equates the sum of engaging gear wheel radii. When multiple gears share the same drive shaft, all engaging gear wheel pairs have to agree on the same shaft distance. If only one component is a gear wheel, there must be a gap of the tooth height $Q$ between the gear and the sleeve or shaft. Between the input shaft gear wheel and a not engaging reverse gear, the gap must be at least two tooth heights. Finally, the distance $a_d$ must be large enough to separate the final drive wheel from the hollow shaft, cf. Eq. (21) (left).

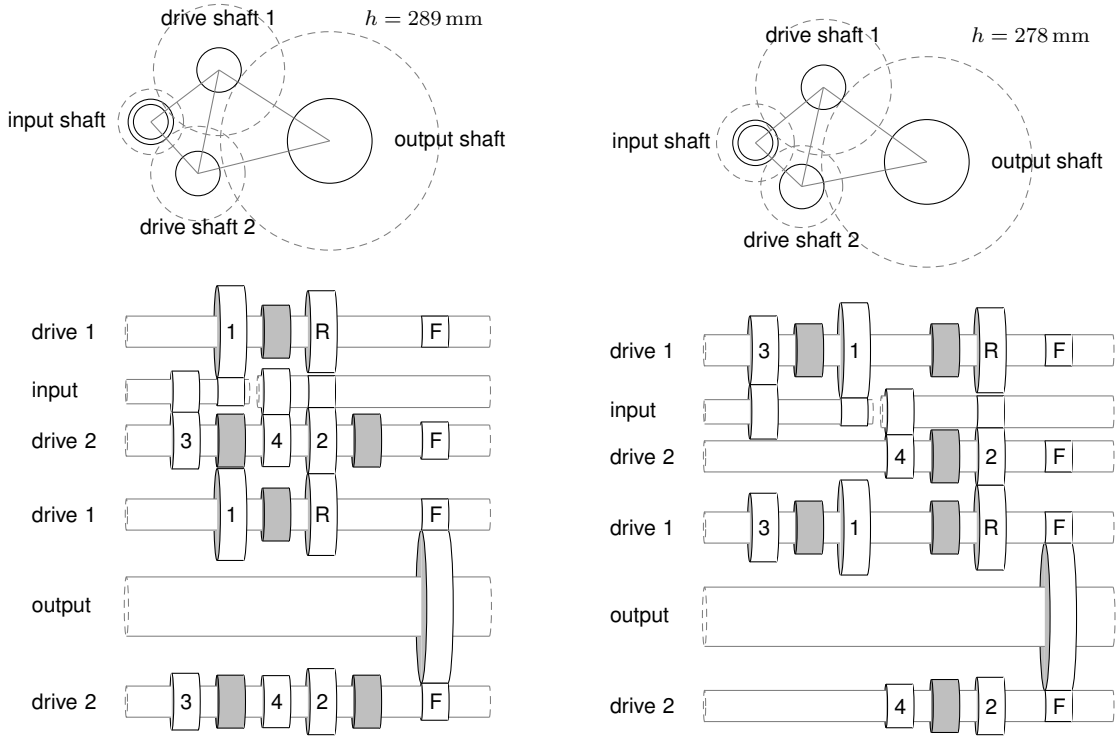
The distance $b_d$ between the drive shaft $d$ and output shaft depends on the final drive radii $y_d$ and $z$. Between the two drive shafts, no gears wheels are allowed to engage except for the reverse gear wheel.

The pseudo-height $h$ that enters the objective is the maximum of two values: Firstly, the drive shaft distance $c$ plus the respective maximum radii on both shafts. Secondly, the diameter of the final drive gear wheel on the output shaft. The pseudo-height is a good approximation of the gearbox height if the two drive shafts are horizontally not too far apart.

## 4 BENCHMARKS AND RESULTS

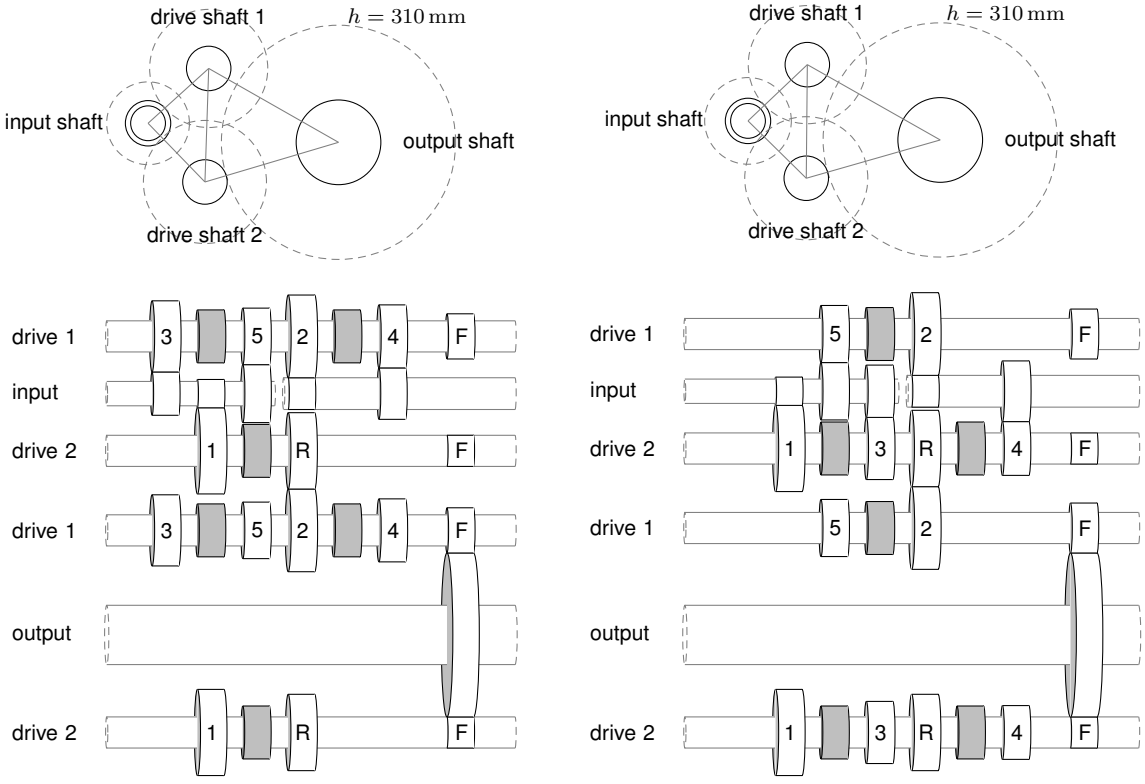
The mixed-integer nonlinear program (MINLP) has been implemented in the mathematical modeling framework *JuMP [3] (version 0.12.0)* in the programming language *Julia (version 0.4.3)*. We have chosen *Couenne [4] (version 0.5.6)*, a free global solver for nonconvex MINLP based on branch-and-bound and convex envelopes, and *SCIP [5] (version 3.2.1)*, a commercial open-source framework for Constraint Integer Programming, to solve the following instances.

As first example, we design a gearbox with four gears and one reverse gear. If we make only four planes available, the solvers return with the status *infeasible*, i.e, there is no imaginable gearbox design in four planes under the stated rule set. It turns out that there are possible designs for five and six available planes. Couenne needed $T^{\text{couenne}}_{4+1,\,5} = 13\,\text{s}$ and $T^{\text{couenne}}_{4+1,\,6} = 21\,\text{s}$ to prove optimality. SCIP was done in $T^{\text{scip}}_{4+1,\,5} = 5\,\text{s}$ and $T^{\text{scip}}_{4+1,\,6} = 9\,\text{s}$.



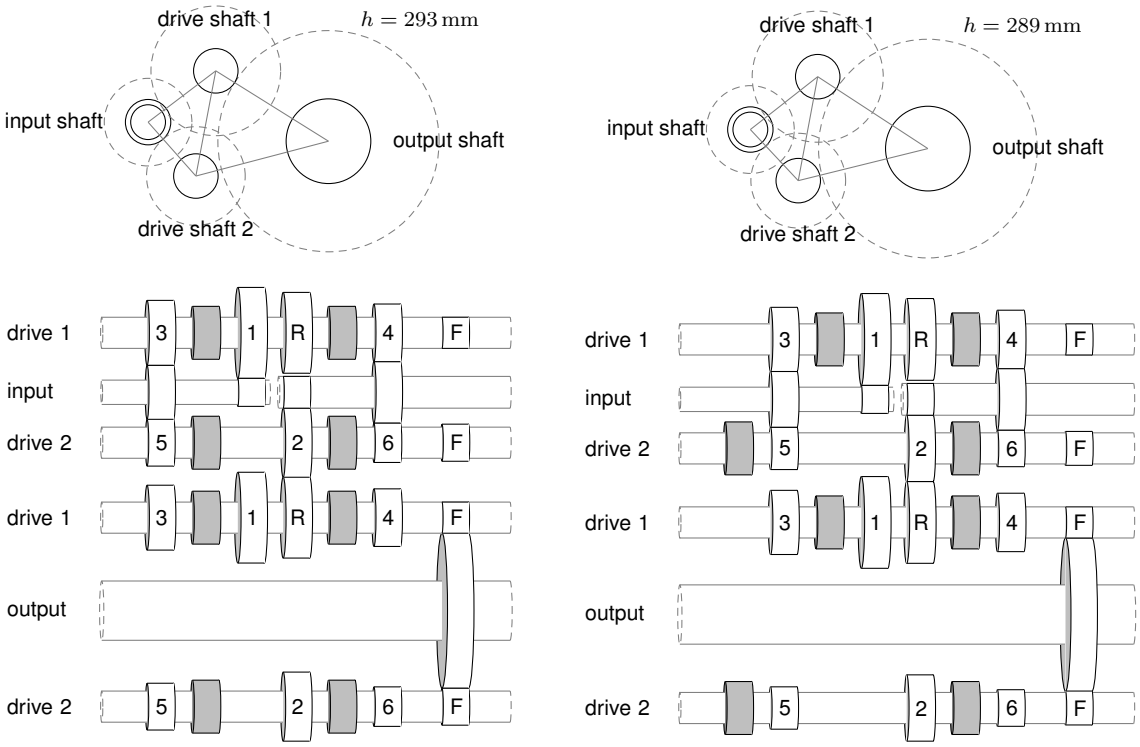

(a) Minimum axial expansion: five planes. (b) Additional axial space: six planes.

Figure 2: Optimal gearboxes with four plus one gears.



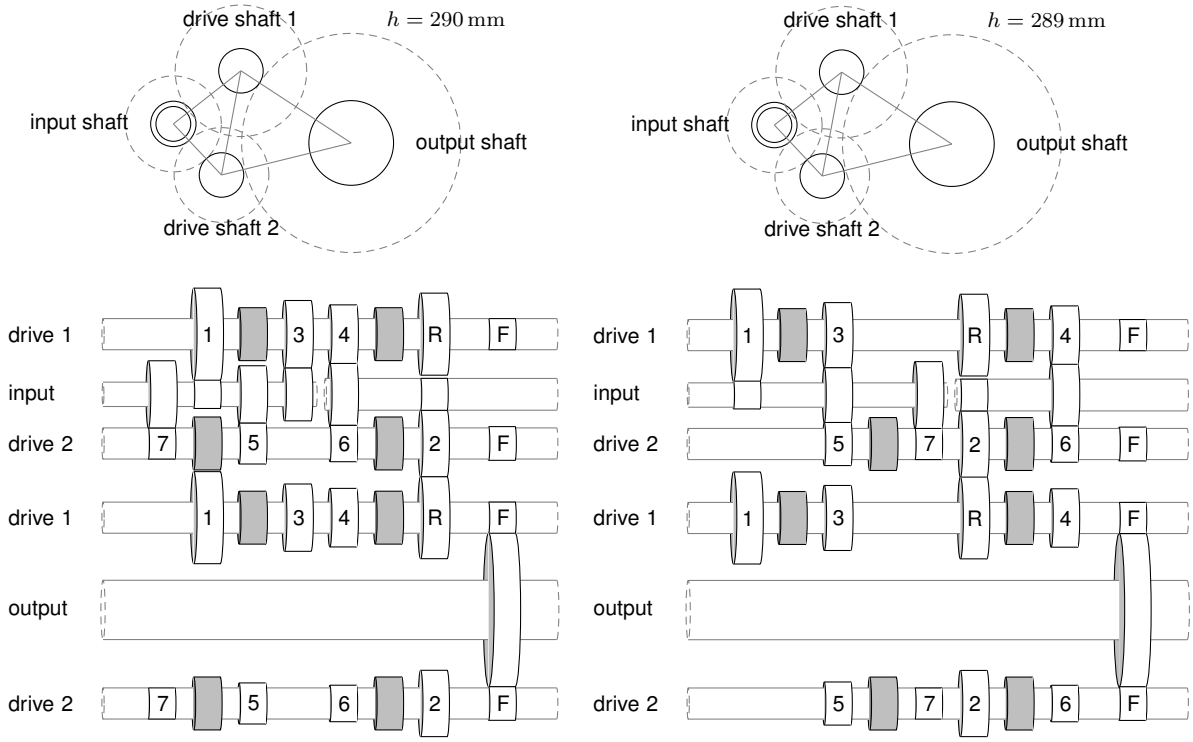

(a) Minimum axial expansion: six planes. (b) Additional axial space: seven planes.

Figure 3: Optimal gearboxes with five plus one gears.




(a) Minimum axial expansion: six planes. (b) Additional axial space: seven planes.

Figure 4: Optimal gearboxes with six plus one gears.




(a) Minimum axial expansion: seven planes. (b) Additional axial space: eight planes.

Figure 5: Optimal gearboxes with seven plus one gears.

Figure 2 shows a representation of the results. On the top, the gearbox is viewed in axial direction – relating to the gearbox in Figure 1, the viewpoint is on the right. The black circles depict the cross sections of the shafts and the gray lines represent distances that are given by decision variables ($a$, $b$ and $c$). The dashed circles indicate the maximum gear wheel radius on each shaft. (They are allowed to overlap as long as these gear wheels do not share a plane.) On the top right, the optimal pseudo-height $h$ is given.

On the bottom of each figure, we see a folding pattern of the gearbox topology. It shows the size and position of the gear wheels and sliding sleeves, the pairwise distances between the shafts, and the gear wheel engagements. The final drive gear wheels coupling drive shafts and output shaft are positioned on the far right and labeled with the letter F. The input shaft is separated into the full shaft on the left and the hollow shaft on the right.

In the presented model, we have not considered power flows along arbitrary many gear wheel pairs: Normal gears use two gear wheel pairs (pre-transmission and post-transmission) and the reverse gear uses three gear wheel pairs (the pre-transmission of a normal gear, the transmission to the reverse gear wheel and the post-transmission). Therefore, in these solutions the power flow of each gear takes the shortest route from the input shaft over the accordingly labeled gear wheel to the output shaft. In the folding pattern, the power flow *appears to jumps* from a drive shaft engaging with the input shaft to its identical copy engaging with the output shaft.

Next, we design gearboxes with five gears and one reverse gear. There is no design with only five planes available. The solutions for six and seven planes are depicted in Figure 3. The runtimes were $T^{\text{couenne}}_{5+1,\,6} = 28\,\text{s}$ and $T^{\text{coenne}}_{5+1,\,7} = 73\,\text{s}$ for Coeunne, and $T^{\text{scip}}_{5+1,\,6} = 5\,\text{s}$ and $T^{\text{scip}}_{5+1,\,7} = 11\,\text{s}$ for SCIP.

The effect of additional axial space is underwhelming: Both gearboxes have exactly the same pseudo-height and the additional plane is not even used. The gear wheels of both solutions are slightly permutated. This is a random effect from the solution procedure, because neither topology is *better* regarding our objective function.

Optimal gearboxes with six plus one and seven plus one gears are shown in Figure 4 and Figure 5. There is no design for a six plus one gearbox in five planes and neither a design for a seven plus one gearbox in six planes. The computations with Couenne required $T^{\text{couenne}}_{6+1,\,6} = 35\,\text{s}$ and $T^{\text{couenne}}_{6+1,\,7} = 168\,\text{s}$ for the six plus one gearboxes and $T^{\text{couenne}}_{7+1,\,7} = 436\,\text{s}$ and $T^{\text{couenne}}_{7+1,\,8} = 803\,\text{s}$ for the seven plus one gearboxes. The computations with SCIP required $T^{\text{scip}}_{6+1,\,6} = 21\,\text{s}$ and $T^{\text{scip}}_{6+1,\,7} = 37\,\text{s}$ for the six plus one gearboxes and $T^{\text{scip}}_{7+1,\,7} = 117\,\text{s}$ and $T^{\text{scip}}_{7+1,\,8} = 116\,\text{s}$ for the seven plus one gearboxes.

Again, the additional plane does not help in achieving a more space-efficient design. The difference of the millimeter values between the left and the right can be explained as numerical artifact. In the case of Couenne, we needed to increase the allowable fractional gap and the feasibility tolerance to compute the presented results. SCIP has stricter default tolerances than Couenne but did warn of a constraint violation in the order of $10^{-6}$. These large instances seem to be numerically challenging. Instead of increasing solver tolerances even further, we think it would be more useful to search for improved model formulations.

Note that the gearboxes with five plus one gears (cf. Fig. 3) need *more* space than the gearboxes with one additional gear (cf. Fig. 4). This is an intended effect of the objective criteria weighing: Five plus one gears can be realized with only three sliding sleeves, but a fourth sleeve would have allowed for a more efficient packing.

## 5 CONCLUSIONS

- We presented a mixed-integer nonlinear program for the optimal design of gearboxes on the example of a dual-clutch transmissions.
- Optimal designs for four plus one gearboxes and up to seven plus one gearboxes have been presented. The computation times of both solvers in the range of seconds to minutes are quite reasonable.
- This paper only shows a first step to a practically useful gearbox design framework. For real-world use, we need to take further details, e.g. the gear teeth design and shaft bending, into account. We plan to integrate these factors into our model in further work.

## REFERENCES

[1] R. Fischer, G. Jürgens, F. Küçükay, R. Najork, B. Pollak, Das Getriebebuch. Springer-Verlag, 2012.

[2] H. Naunheimer, B. Bertsche, G. Lechner, Fahrzeuggetriebe: Grundlagen, Auswahl, Auslegung und Konstruktion, 2007.

[3] M. Lubin, I. Dunning, Computing in Operations Research Using Julia, *INFORMS Journal on Computing*, **27**, No. 2, pages 238–248, 2015.

[4] P. Belotti, J. Lee, L. Liberti, F. Margot, A. Wächter, Branching and bounds tightening techniques for non-convex MINLP, *Optimization Methods and Software*, **24**, No. 4–5, pages 597–634, 2009.

[5] T. Achterberg, SCIP: Solving constraint integer programs, *Mathematical Programming Computation*, **1**, No. 1, pages 1–41, 2009.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A FRAMEWORK FOR THE DESIGN BY OPTIMIZATION OF HYDROFOILS UNDER CAVITATING CONDITIONS

**Paolo Olivucci**[1] **and Stefano Gaggero**[2]

[1] University of Genoa
Via Montallegro, 1, 16145 Genoa, Italy
e-mail: 3529870@studenti.unige.it

[2] University of Genoa
Via Montallegro, 1, 16145 Genoa, Italy
stefano.gaggero@unige.it

**Keywords:** Optimization, Cavitation, Hydrofoils, RANSE, OpenFOAM.

**Abstract.** *Hydrodynamic shape optimization based on CFD calculations can dramatically improve the design of marine devices (i.e. propellers, rudders and appendages) by simultaneously considering opposite objectives and by modeling phenomena that well-established and still widely adopted design approaches (i.e. lifting line and lifting surface) cannot accurately deal with. Cavitation on propellers, for instance and among the others, is one of the most dangerous phenomena. It causes vibration, erosion and it is a source of radiated noise, consequently resulting incompatible with modern propeller design, continuously aimed for higher efficiency, comfort and environmentally safe operations. An accurate selection, firstly, of the most appropriate blade sections is, consequently, of crucial importance at least to limit the side effects of cavitation. In the present work, therefore, a numerical framework for the design by optimization of marine hydrofoils under cavitating conditions is proposed. By combining a parametric description of the hydrofoil shape, the NSGA-II multi-objective genetic algorithm and appropriate flow solvers, new hydrofoil shapes are derived. Objectives of the design are blade sections with enlarged cavitation buckets to increase the cavitation inception speed and to reduce the cavity volume (under the constraint of unchanged delivered lift) with respect to widely accepted NACA66 profiles. Boundary element methods and RANSE solvers (a proprietary Hess & Smith implementation and the open-source RANSE solver OpenFOAM) are applied in succession in order to verify the influence of the inviscid/viscous nature of the flow on the final optimal hydrofoil shape and of the additional maximum lift/drag ratio objective required in the case of viscous calculations.*

## 1 INTRODUCTION

In recent years, the design of marine propellers has consistently evolved. From Betz optimum criteria to lifting line and lifting surface computer codes, the definition of the optimal (under certain constraints and objectives) blade geometry results as the massive application of numerical approaches the more sophisticated the more computational resources became available.

Efforts have been made to optimize the propeller geometry increasing efficiency [1] or optimizing the cavitation behavior [2]. The final consequence of this trend was the completely numerical design by optimization approach proposed in [3] for both conventional and unconventional propellers [4][5]. The so-designed propellers showed superior performance with respect to traditional design methodologies thanks to the application of computationally efficient and accurate potential flow solvers based on panel methods and global convergence optimization algorithms to deal with multi-objective and constrained designs, even in the case of high-demanding controllable pitch propellers [3][6]. Efficiency and cavitation, as a matter of fact, are contradictory demands. In order to fulfill the demand of high efficiency and stable sheet cavitation on the suction side of the blade the loading at tip (the pitch, consequently) cannot be too small: tip vortex cavitation, consequently, may occur and the "bursting" of the vortex in the propeller wake represents a source of broadband excitation to the ship structures. On the other hand, an excessive unloading of the tip may cause leading edge separating vortexes (or pressure side cavitation) which erosive action on the propeller tip can be substantial. The optimization of the main geometrical parameters of the propellers (radial distribution of pitch, camber and chord, for instance) based on accurate potential flow solver calculations allowed for a well-balanced choice of the optimal geometry.

The application of potential panel methods, however, was bottlenecked by the inviscid assumption for the flow. If the influence of viscosity can be partially accounted via simplified formulations when global parameters such pitch or chord were modified, instead, when local geometrical variations (blade leading edge radius or blade sectional thickness distribution, for instance) have to be investigated, it is mandatory to make use of more accurate approaches to avoid unphysical solutions. The designs by optimization already proposed [3][4], for instance, dealt only with global variations of the geometry, avoiding excessive modifications of the base profile shape in the light of limiting the influence (only partially controlled by panel methods and empirical formulations) of the viscosity on local flow phenomena. On the other hand, the reduction of cavitation induced pressure pulses and the prevention of cavitation erosion can be efficiently achieved exactly by using better design of the blade sections instead of the simple geometries of B-Series or MAU series, making optimization also of blade sections worth of investigation.

The use of thin and wide sections as the results of increased blade area ratio and reduced loading per unit area is, probably, the simpler way to improve the cavitation behavior of marine propellers. Thinner sections means larger margins against bubble cavitation on the suction side at a cost of a relatively narrow cavitation bucket that increase the risk of sheet cavitation on both blade sides and dramatically reduces the sheet cavitation inception speed when the propeller operates in a spatial non-uniform inflow. Moreover, larger blade area means higher friction losses and lower propulsive efficiency, opposite to the design objectives. The erosive nature of bubble and pressure side cavitation, in addition, influence the design of the propeller blade. A certain pitch and camber combination is always selected in order to always have an enough margin against pressure side cavitation under the worst unloaded condition. The consequence of this design philosophy is that the design operation point in the average ship wake field can no longer be placed at the angle of attack of shock-free entry of

the profile. The required average angle of attack is, thus, shifted towards the suction side cavitation with an increased risk of large cavitation volumes and hull excitations.

Hence, the development of robust tools to accurately deal also with a better design of blade sections is indispensable to pursue efficient designs with improved cavitation performance. Instead than designing propellers with higher values of blade area, the application of new blade sections, in fact, may overcome many of the critical points of the usual hydrofoil shapes adopted in the propeller design. By using sections with wider cavitation buckets the design point can be brought back close to the shock free angle of attack of the profile, preventing high suction peaks on the suction side leading edge of the blade and increasing the cavity inception speed. NACA profiles [7], from this point of view, represent a substantial improvement with respect to traditional section series and, despite they were designed for wing sections of airplanes, they are widely accepted in propeller design. The almost flat pressure distribution on the suction side, achieved with the design assumption of the largely adopted a = 0.80 camber distributions, allows for a relatively wide cavitation bucket, partially limited by the non-optimal response to negative (lower than the ideal) angles of attack. Even better cavitation performance could be achieved by employing different camber and thickness distributions for the evaluation of which, however, advanced design tools are required.

The first attempts to wider the cavitation bucket of hydrofoils for marine propellers date back to the eighties when Shen [8] applied the Eppler's method [9] for the design of improved hydrofoil shapes. Successful application were those by Kuiper [2] and Dang [10], both in the case of hydrofoils and propellers. A non-linear design of the so-called "new blade section" was achieved by Dang [11]. The newly designed hydrofoils were characterized by an unloading of the leading edge to withstand large variations of the angle of attack together with a maximum thickness toward the leading edge as the application of a boundary element method coupled with a thin boundary layer solver. The importance of this design approach was to have pointed out the importance in the design process of the viscous phenomena, in particular the turbulent separation of the boundary layer, which should be accurately taken into account for reliable design and performance improvements. Further advances in wing sections design were those of Amoignon [12], which proposed a gradient-based optimization environment for the definition of optimal wing profiles aimed to the minimization of the drag at constant lift. The solution of the Euler equations with appropriate corrections to account for the laminar flow allowed for reliable designs while by using a RANS equations solver it was possible to consider in the optimization process the turbulent viscosity that may dominate the flow in the particular case of high-lift systems. Even if limited to lift/drag optimization, the application of fully viscous solvers to the design of aeronautical wing profiles showed the obvious advantages that could be expected by using a RANSE solver and that, reasonably, can be expected also in the more demanding case of hydrofoils subjected to the cavitation avoidance constraint.

In the present work, as the first step towards the application of fully viscous approaches to the complete design of marine propellers, a design by optimization of marine hydrofoils is proposed. The design framework consists in a robust parametric description of the hydrofoil shape by using B-Spline curves, a multi-objective optimization based on the NSGA-II algorithm and two different flow solvers, a Boundary element method and the OpenFOAM RANSE tool. Similarly to what proposed by Dang [11], the objectives of the optimization are the widening of the cavitation bucket of a parent hydrofoil (the NACA66 Mod, a=0.80 mean camber line, with a maximum camber/chord ratio equal to 0.02 and a maximum thickness/chord ratio equal to 0.04) and the maximization of lift/drag ratio. If, of course for its inherent assumption, the prediction of the frictional drag is pointless when the Boundary element method is applied, with the OpenFOAM RANSE [13] solver it is possible to account

for the viscosity influence. Two different optimization, therefore, are proposed. The first considers as objective only the widening of the bucket and it is carried out by employing only the Boundary element method. Results are finally checked with the RANSE in order to verify that the optimal solutions under the potential flow hypotheses are still valid when the viscosity is taken into account. The second one is directly based on viscous calculations and, as highlighted above, in addition considers as objective the maximization of the lift/drag ratio at a given delivered lift. Besides, a preliminary validation of the flow solvers is carried out in order to highlight possible limitations and their influence on the predicted cavitation buckets of both the reference and the optimized hydrofoils.

## 2 THE DESIGN BY OPTIMIZATION

The design by optimization consists, essentially, in a reverse design process based on an automatic try-and-error procedure: the hydrofoil shape optimization problem can be formulated as the iterative finding of the set of real valued bounded free variables that control the geometry of the hydrofoil able to minimizes an objective function (cavitation) subject to a set of constraints (lift, for instance). By using any of the available flow solvers, the optimal shape is found by the iterative modification (based on predicted performance) of an initial set of geometries: the design is indirect, rising form the analysis of hundreds (or thousands) of different possible configurations among which identify the better candidates against selected constraints and objectives. This design process, in principle, can be summarized in three steps:

a. A parametric description of the geometry, able to automatically generate suitable variations of the reference design,
b. A computationally efficient CFD method (the Boundary element method or the RANSE solver) for the hydrodynamic analysis of the effects of the shape variations of the thousands of the different candidates under investigation,
c. A reliable and robust derivative-free global convergence optimization algorithm able to find the path to the overall optimum by automatically changing (a) the hydrofoil shape and by computing (b) its performance.

### 2.1 Parametric description of the hydrofoil geometry

The parametric description of the hydrofoil geometry is one of the key point of the optimization process. It is necessary to define automatically each new shape whose hydrodynamic characteristics have to be evaluated in the design process, on the basis of a rather limited number of parameters that turn to be the free variables of the optimization, i.e. the set of values free to be changed in order to identify the optimal design. At the same time, this description has to comply with some “geometrical” constraints: the hydrofoil shape, at least in the case of partially cavitating applications, has to be “faired”, avoiding abrupt changes in curvature and slope. A rather well-established way to achieve these results consists in the adoption of B-Spline curves. As proposed in [3][4][5], B-Splines were successfully adopted to model the main geometrical characteristics of conventional and unconventional propellers, providing a robust approach for the definition and the smooth modification, through their control points, of the blade shape. Similarly, also in the hydrofoil case, B-Spline curves were adopted. Rather than directly describing the suction and the pressure side curves, however, it was preferred to model the thickness and camber distributions, following a usually adopted convention in the design of marine propellers. An example of the B-Spline approximation of the parent NACA66Mod section is proposed in Figure 1. The fifteen free parameters considered in the optimization, together with their allowed range of variation expressed as a percentage of the reference NACA66Mod value, are summarized in Table 1.

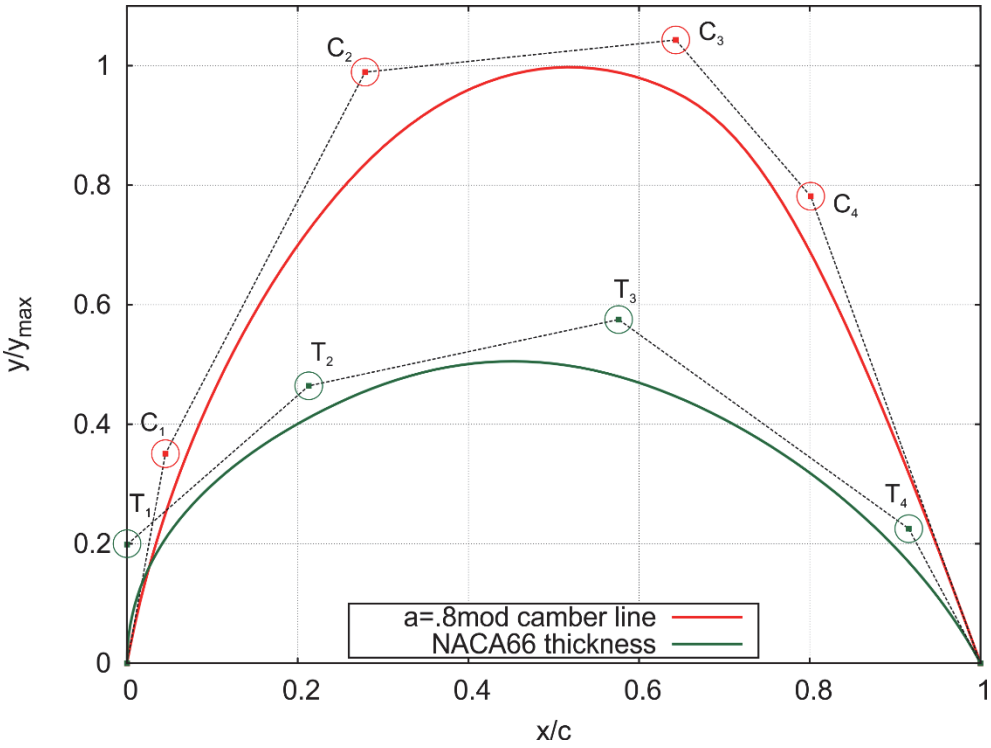



Figure 1: B-Spline description of the reference NACA66Mod non-dimensional half thickness and non-dimensional camber distribution and relative control polygons.

| Point | Δx/c | Δy/c | Point | Δx/c | Δy/c |
|---|---|---|---|---|---|
| $C_1$ | ± 40% | ± 10% | $T_1$ | - | + 50%/-20% |
| $C_2$ | ± 40% | ± 10% | $T_2$ | + 50%/-20% | + 50%/-20% |
| $C_3$ | ± 40% | ± 10% | $T_3$ | + 50%/20% | ± 30% |
| $C_4$ | ± 40% | ± 10% | $T_4$ | ± 40% | ± 20% |

Table 1: Ranges of variation of the free parameters adopted for the hydrofoil shape definition. The x/c position of point $T_1$ was fixed equal to 0 (no variations allowed) to avoid cusps at the hydrofoil leading edge.

## 2.2 Boundary Element Method and RANSE solvers for hydrofoils hydrodynamic analysis

For the solution of the flow around bi-dimensional hydrofoils, two different solvers were adopted. The first one is a potential Boundary element method for incompressible, inviscid and irrotational flows. The current implementation is based on the low-order Smith and Hess [14] approximations. The Laplace equation for the perturbation velocity potential $\phi$ is solved by a superposition of piecewise constant sources and vortexes, distributed only over the hydrofoil surface discretized by using straight segments, which intensity is defined based on the kinematic and the Kutta boundary conditions:

$$\begin{cases} \nabla^2\phi = 0 \\ \partial\phi/\partial n + \mathbf{V}\cdot\mathbf{n} = 0 \\ \Delta p_{trailing\ edge} = 0 \end{cases} \quad (1)$$

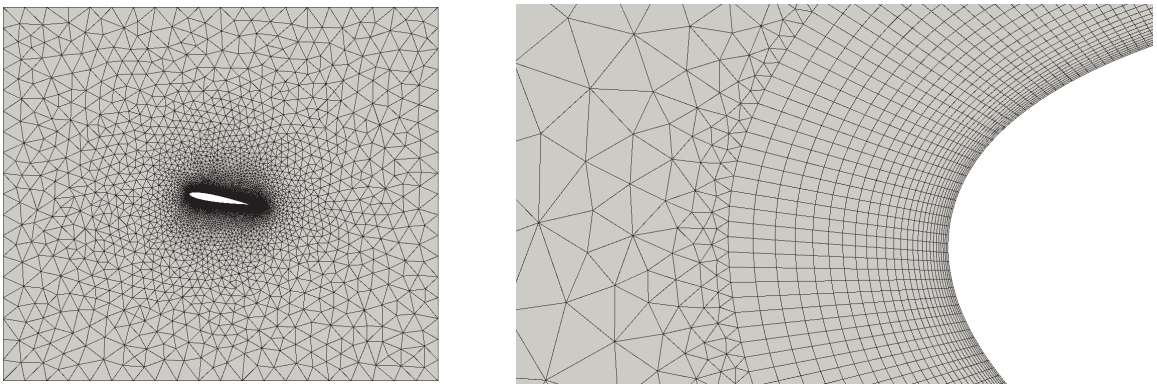
The corresponding pressure distribution over the hydrofoil, from which forces are derived, is calculated with the Bernoulli's theorem. The second solver is based on the solution of the RANSE equations for an incompressible fluid:

$$\begin{cases} \nabla \cdot \mathbf{V} = 0 \\ \dfrac{\mathcal{D}}{\mathcal{D}t}(\mathbf{V}) = \dfrac{1}{\rho}(-\nabla p + \nabla \cdot \boldsymbol{T}_{RE} + \boldsymbol{S}) \end{cases} \tag{2}$$

where $\boldsymbol{T}_{RE}$ is the tensor of Reynolds stresses (computed according to Boussinesq hypothesis) and $\boldsymbol{S}$ is a tensor representative of additional momentum sources. In the present work, the set of equations (2) is solved, on an unstructured finite-volume mesh, by using OpenFOAM [13] and the *SST* $k-\omega$ turbulence model to account for the additional stresses due to turbulence.

### 2.3 Validation analysis

A preliminary validation of the both the flow solvers has been carried out in order to assess their accuracy and to highlight the possible limitations that could affect the design of optimal hydrofoils in the optimization process. A suitable test case is represented by the prediction of the pressure distribution around a NACA0012 hydrofoil for which accurate measurements are available in [15]. Boundary element method calculations are based on a surface mesh of 1000 panels arranged in a cosinusoidal spacing at leading and trailing edge that is necessary to have a better description of the local high curvature surfaces. RANSE calculations were carried out on a hybrid unstructured mesh arranged with the open source meshing tools *GMSH* [16]. An example of the computational mesh, with a close-up view at the hydrofoil leading edge, is proposed in Figure 2. The hybrid mesh (about 35000 cells) consists in an inner, near-wall structured part, made up of prism cells to control the boundary layer phenomena, and an external part filled by triangular elements. An appropriate spacing of the cells was adopted in order to comply with non-dimensional wall distances compatible with the *high-* or *low-*Reynolds formulation of the turbulence model equations. The results of the analysis are summarized in Figure 3, where the predicted pressure distributions, calculated with different second-order accurate discretization schemes (the OpenFOAM *limiteLinearV* and *linearUpwind*) and turbulence modeling, are compared with measurements and with dedicated numerical calculations carried out with the NASA CFL3D code [17] and available in [18].

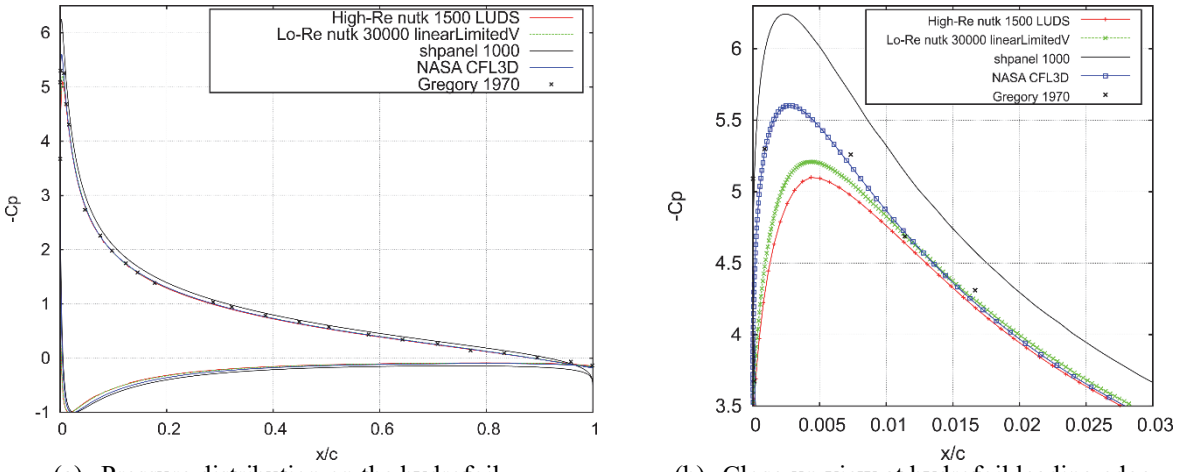


(a) Computational domain (b) Detail of the mesh at the hydrofoil leading edge

Figure 2: Mesh arrangement for viscous flow calculations – NACA0012 hydrofoil.

The agreement is satisfactory. Both the Boundary element method and the OpenFOAM RANSE solver predict pressure distributions on the suction side close to measurements. At

the leading edge some differences, however, can be appreciated and have to be considered in the optimization process. Except the NACA CFL3D calculations, that effectively were carried out on a rather different (fully structured, low-Reynolds assumption with y+<<1) mesh arrangement, OpenFOAM results, regardless the wall bounded turbulence modeling and the spatial discretization schemes, slightly underestimate the suction pressure peak that, instead, is reasonably foreseen by the CFL3D code. On the contrary, the suction peak predicted by the Boundary element method is higher than both the measurements and the viscous calculations. The cavitation buckets predicted by the Boundary element method are expected, consequently, to be "narrow" (i.e. slightly anticipating, at a given angle, the occurrence of cavitation) than those predicted by viscous calculations.




(a) Pressure distribution on the hydrofoil (b) Close-up view at hydrofoil leading edge.

Figure 3: Predicted pressure distribution ($-C_P$) on the NACA0012 hydrofoil at 10 deg. of angle of attack. Comparison between measurements and numerical calculations.

However, what really matters and drives the convergence of the optimization algorithm is the relative merit and not the absolute value of the predicted performance: only the ability of the numerical approaches to rank correctly different hydrofoil shapes is necessary to perform a design by optimization. The results achieved with this numerical setup (mesh, turbulence assumption and numerical schemes), consequently, may be considered satisfactory, especially in the light of need of computationally fast (few seconds for Boundary element calculations, a couple of minutes for the OpenFOAM RANSE results) approaches to deal with the thousands of cases to be analysed.

## 2.4 Optimization process with constraints and objectives

The optimization process proposed for the design of hydrofoils has a multi-objective nature. Improving the cavitating performance of a hydrofoil presupposes the necessity to simultaneously monitor different and opposite phenomena: suction side sheet cavitation may be reduced by using high values of camber that, in turn, determine an increase of the risk of sheet cavitation on the pressure side with the change of the angle of attack. An increase of the thickness generally turns out to be an effective way to mitigate the suction peaks at leading edge at the cost, however, of an anticipated inception of the bubble cavitation at the hydrofoil mid-chord. In addition, a constraint on the delivered lift has to be forces in order to design hydrofoils which performance are in line with those provided by the reference geometry. A multi-objective optimization algorithm is, consequently, required. In the present calculation, the NSGA-II [19] – Nondominated Sorting Genetic Algorithm was selected.

Similarly to what proposed by Dang [10], the cavitation bucket represents a useful representation of the cavitating performance for the optimization process. By simply assuming as the cavitation inception the condition in correspondence of which the minimum value of the pressure coefficient on the hydrofoil equals the cavitation number ($-C_P = \sigma$), any cavitation bucket may be simply computed, regardless the flow solver, by collecting, in correspondence to a sufficient number of different angles of attack, the minimum value of the pressure coefficient. Exactly the same computational idea is under the devised optimization approach. As exemplified in the diagram of Figure 4, by simultaneously requiring the minimization of the value of the inception points for a set of angles of attack it is possible to drive the selection of any new hydrofoil shape towards the widening of the cavitation bucket, i.e. towards the postponing of the cavitation phenomena.

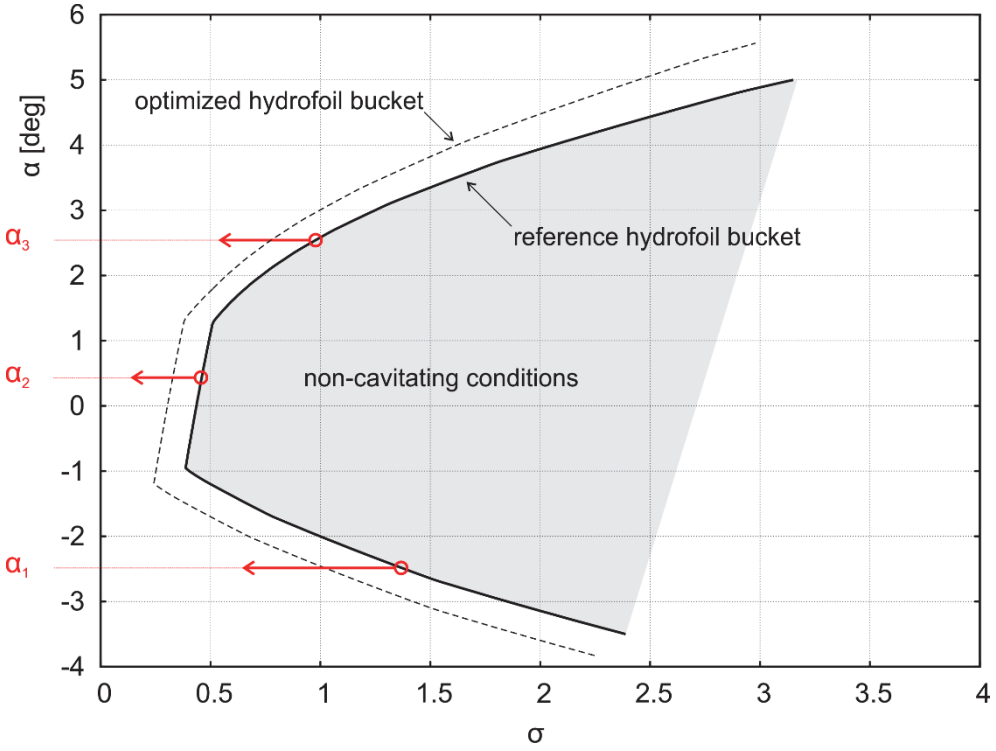



Figure 4: Example of widening the hydrofoil cavitation bucket by simultaneously controlling the cavitation inception at different angles of attack.

The (minimum) number of angles of attack to be monitored in the design process, of course, has to be carefully selected. Having in mind the different type of cavitation a hydrofoil may be subjected to, angles have to be selected in order to let the algorithm monitoring suction and pressure side sheet as well as mid-chord cavitation on the back side. The cavitation bucket of the reference geometry serves as guideline for this selection. The proposed optimization algorithm, in the end, is structured to:

- Minimize the inception value of pressure side sheet cavitation ($-C_P$ at $\alpha_1$),
- Minimize the inception value of mid-chord suction side bubble cavitation ($-C_P$ at $\alpha_2$),
- Minimize the inception value of suction side sheet cavitation ($-C_P$ at $\alpha_3$),
- Deliver a Lift coefficient at a given angle of attack $\alpha_{Lift}$ equal to that of the reference hydrofoil (with a ±2% tolerance to speed-up the Pareto convergence),

- Minimize, only in the case when the OpenFOAM RANSE solver is adopted in the optimization process, the frictional drag.

# 3 RESULTS

## 3.1 Design of the optimal hydrofoil – inviscid approach

The first design was obtained by the application of the Boundary element method as the flow solver in the optimization run. Thus, only a constraint on the lift computed by the Boundary element method itself in the case of the reference NACA66Mod hydrofoil ($C_L$ = 0.26 ± 2% at $\alpha = 0°$) was forced. Two different cases were addressed. The first one (case I-inv) was carried out considering as a geometrical constraint the maximum thickness of the hydrofoil which was forced to be equal to the value of the reference NACA66Mod profile. The second run (case II-inv) allows unconstrained variations of the free parameters, resulting in optimized hydrofoils with a maximum thickness and camber different from those of the reference geometry.

The optimization process, as the consequence of the computational efficiency of the Boundary element method, solved 60 thousands different candidate geometries (a population of 100 and 600 generations) for case I-inv. Case II-inv, instead, solved only 10 thousands of different geometries as the combination of a population of 100 members and 100 generations. The performance of case I-inv are summarized in the Pareto diagram of Figure 5. Results are collected in terms of reduction of the inception point of sheet and back cavitation on the *x*- and *y*- axis while bubbles inception is monitored by the size of the markers. Green points define a subset of geometries having better performance (leading edge sheet cavitation only) with respect to the reference shape that is marked in blue.

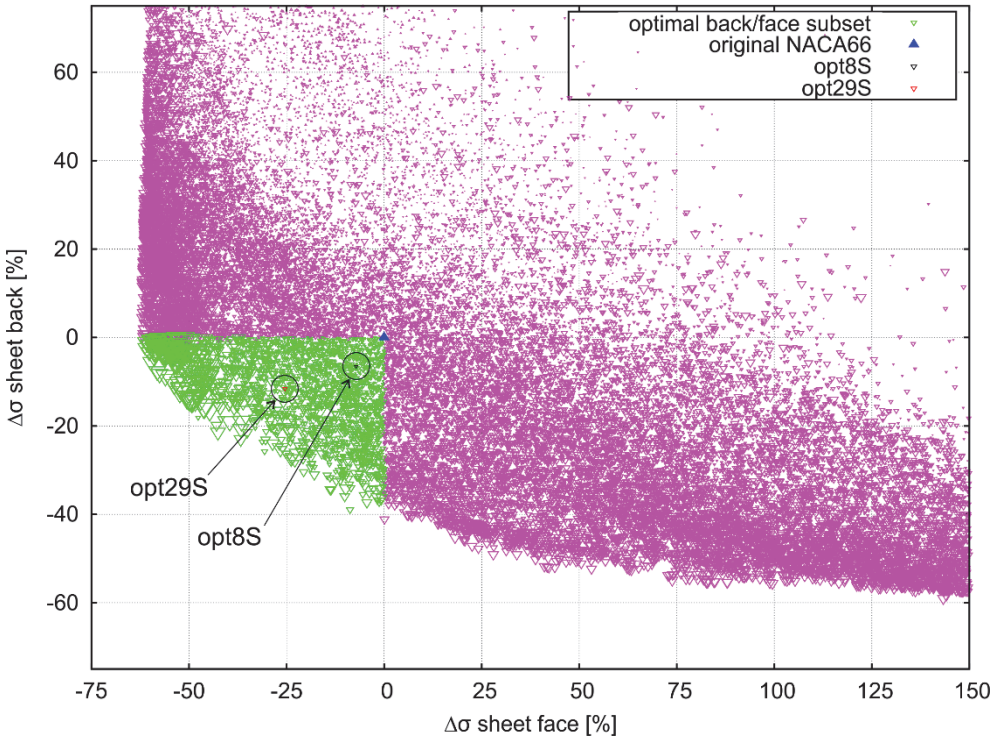



Figure 5: Pareto diagram of the inviscid optimization – case I-inv.

The improvements (Boundary element method results) achieved through the optimization are remarkable, in particular for what regards the reduction of the suction and pressure side sheet cavitation inception. Bubble cavitation is only slightly enhanced, as better highlighted by the bucket diagrams of Figure 6. Two are optimal hydrofoils extracted from the design by optimization: *opt8S* is, among the optimal back/face subset, the geometry with the highest improvement in terms bubble cavitation postponing while *opt29S* is a better-balanced geometry that achieve higher improvements for what regards back and face cavitation at the cost of negligible improvements for what regards mid-chord phenomena.

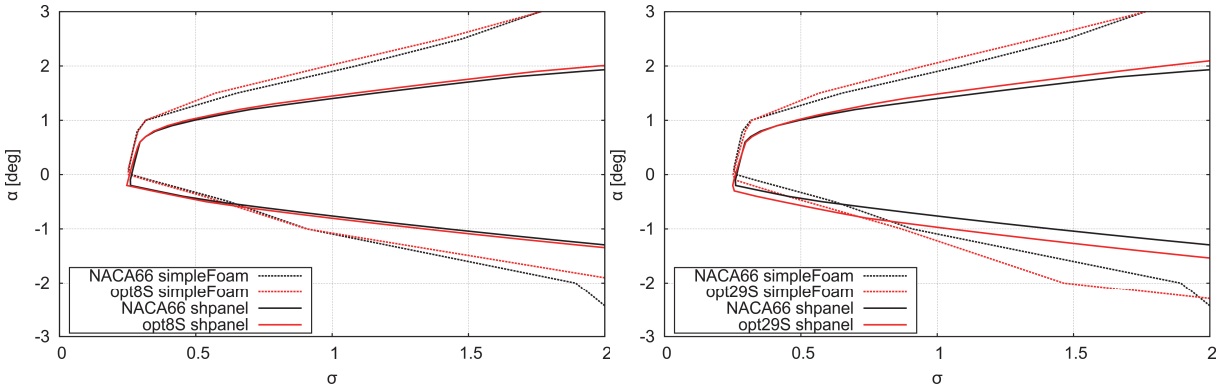



Figure 6: Comparison between the case I-inv optimal (*opt8S* on the left, *opt29S* on the right) and the reference (NACA66Mod) cavitation buckets. Boundary element method (shpanel) and RANSE (simpleFoam) calculations.

Essentially the improvements in bubble phenomena, monitored by the panel method, are almost negligible. In the case of the *opt8S* geometry the enhancement regards the entire range of angle of attack originally subjected (for the reference NACA66Mod) to bubble cavitation while in the case of the *opt29S* only in a narrow range around the $\alpha_2$ angle (that adopted in the optimization to monitor bubble cavitation) the optimized hydrofoil provides better performance. The improvements in sheet cavitation inception, instead, are clearly evidenced by the comparison of the bucket diagrams. For the entire range of angle of attack under investigation the optimized geometries provide better performance that result particularly improved in the case of *opt29S* hydrofoil.

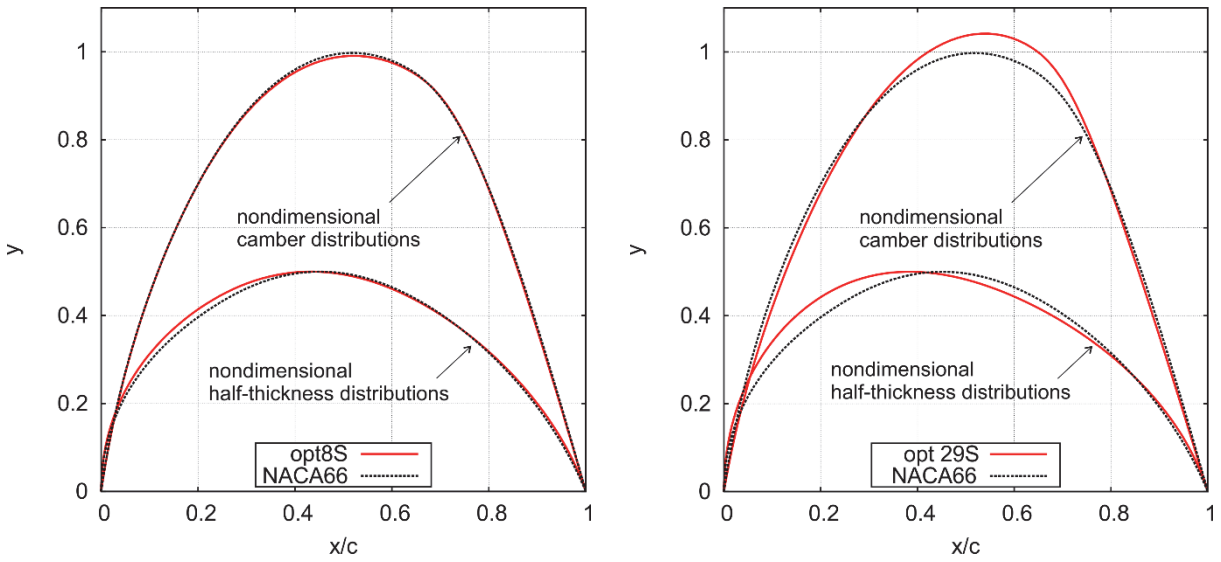



Figure 7: Comparison between the optimal (*opt8S* and *opt29S*) and the reference (NACA66Mod) hydrofoil geometries.

Pressure side inception is significantly postponed, with an appreciable improvement in correspondence of the sharp point that characterize the transition bubble/pressure side cavitation of the reference NACA66Mod shape. These results have been achieved, as expected, by the shift, towards the leading edge, of the maximum thickness, that is particularly evident in the case of the *opt29S* geometry (Figure 7). This optimized hydrofoil, indeed, effectively produces the higher improvements in terms of leading edge cavitation that are obviously mitigated by locally thicker sections. The improvements evidenced by the optimization are furthermore important from two additional point of view. If the shift of the maximum thickness towards the leading edge is an obvious conclusion for the mitigation of the sheet cavitation, the resulting geometry, as the outcome of the constrained optimization, automatically provides the required performance (lift, for instance) without any worsening, related to the thickness distribution, in terms of bubble cavitation. In addition, the calculations carried out with the OpenFOAM RANSE solver confirm the trends and the ranking of the geometries under investigation evidenced by the panel method. Both the analyzed geometries are overall better than the reference hydrofoil, confirming the outcomes of the design by optimization carried out with the "low" fidelity Boundary element method. In addition, for both the *opt8S* and the *opt29S* hydrofoils, viscous calculations foresee a (very) slightly better performance for what regards bubble cavitation: *opt8S*, also when analyzed with OpenFOAM, shows a slightly higher increase of the mid-chord cavitation inception speed. Similarly, the cavitation bucket predicted by using viscous calculations in the case of *opt29S* is characterized by the substantial improvement in terms of pressure side phenomena already foresee by the inviscid calculations. Even if the absolute values of the cavitation buckets calculated by using the Boundary element method or the RANSE solver are slightly different, the relative merits of one geometry with respect to another are very similar. The Boundary element method, consequently, can be considered a reliable and suitable approach to design hydrofoils under cavitating conditions with the only limitation of the "efficiency" (i.e. the lift/drag ratio) of the newly designed hydrofoils due to the inherent limitations of the potential flow assumption (Table 2).

The second optimization run with the inviscid approach, case II-inv, was carried out by unconstraining the maximum of the thickness distribution. The reference profile under investigation, effectively, is relatively thin: bubble phenomena are naturally postponed (indeed the bubble inception cavitation index of about 0.30, as from Figure 6, is very low) at the cost of a leading edge sheet cavitation sensibly worsened. A released maximum thickness should further improve the hydrofoil performance.

The results of case II-inv optimization are summarized in the Pareto diagram of Figure 8 by using the same convention (axis and markers size) adopted for case I-inv. Among the optimal suction and pressure side cavitation subset, two candidates, namely *opt28* and *opt469*, were selected based on bubble cavitation improvements in order to assess the influence of the thickness distribution by analysing, through Boundary element method and RANSE calculations, their complete cavitation buckets.

The complete cavitation buckets of *opt28* and *opt469* (Figure 9) are in line with the results of the optimization: both the hydrofoils provide, for what regards leading edge sheet cavitation, better performance with respect to the reference NACA66Mod. Once again, the relative merits of on geometry with respect to another are identically foresee also by the RANSE calculations, indirectly proving the reliability of the Boundary element method in the optimization process.

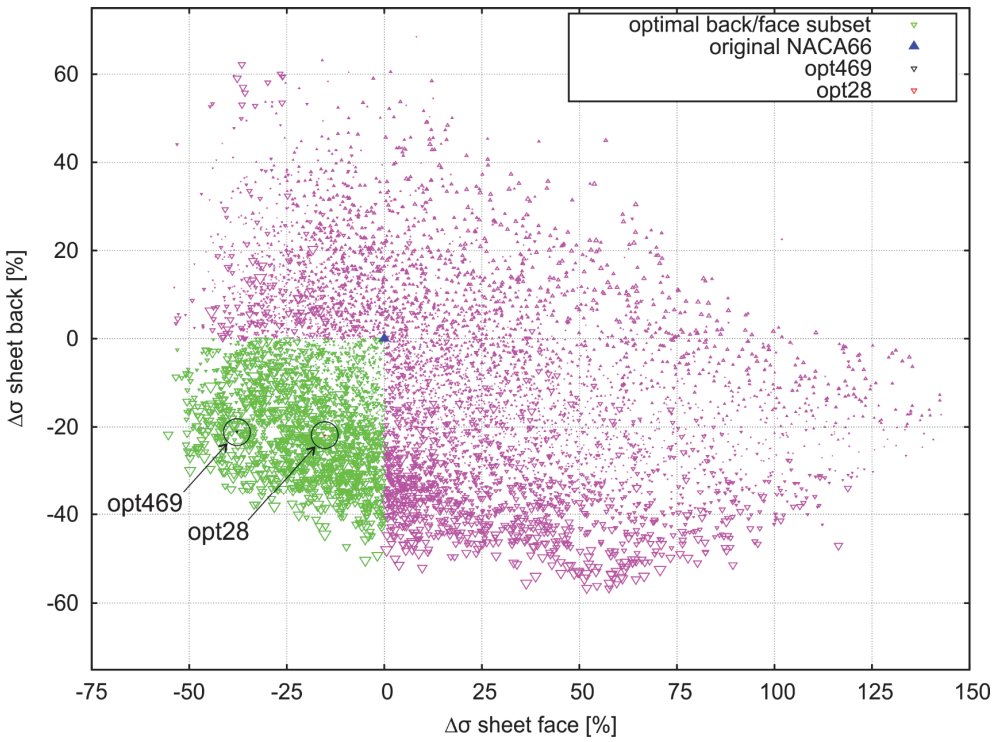



Figure 8: Pareto diagram of the inviscid optimization – case II-inv.

The improvements in terms of sheet cavitation (especially on the face side) of *opt469*, as it could be expected by comparing the thickness distributions of the hydrofoils shown in Figure 10, are the higher among those of the analyzed geometries (*opt8S*, *opt29S*, *opt28*): *opt469* has, indeed, a maximum thickness that is about the 20% higher than that of the NACA66Mod.

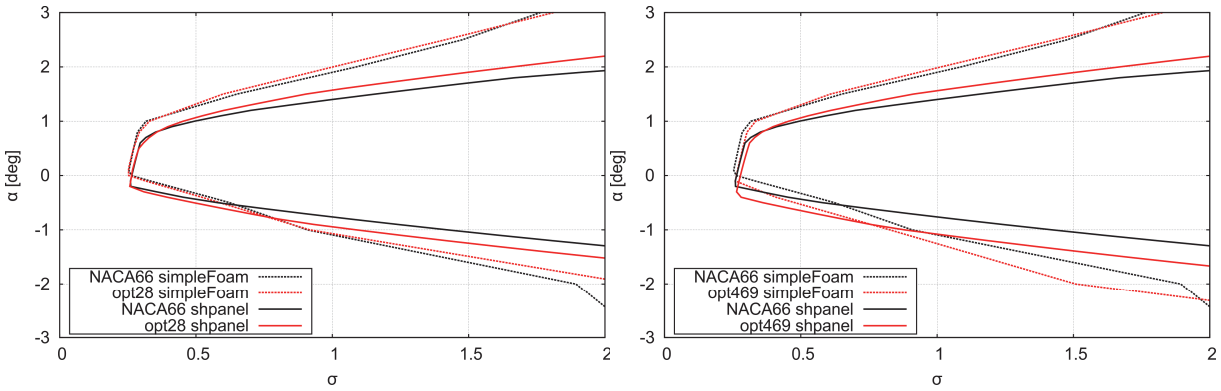



Figure 9: Comparison between the case II-inv optimal (*opt28* on the left, *opt469* on the right) and the reference (NACA66Mod) cavitation buckets. Boundary element method (shpanel) and RANSE (simpleFoam) calculations.

Moreover, the increased maximum thickness is shifted, as in case I-inv optimization, towards the leading edge. The resulting worsening of the bubble phenomena is exactly evidenced by both the viscous and the inviscid calculations. On the contrary, hydrofoil *opt28* was selected with stricter margin on bubble cavitation. The increase in maximum thickness, that positively acts to widen the cavitation bucket, is limited by the avoidance of bubble cavitation in excess to that of the reference geometry. As pointed out by the cavitation diagrams, the slight increase of thickness that characterize the *opt28* geometry thanks to which its leading edge cavitating performance are improved, is counterbalanced by a slightly lower value of

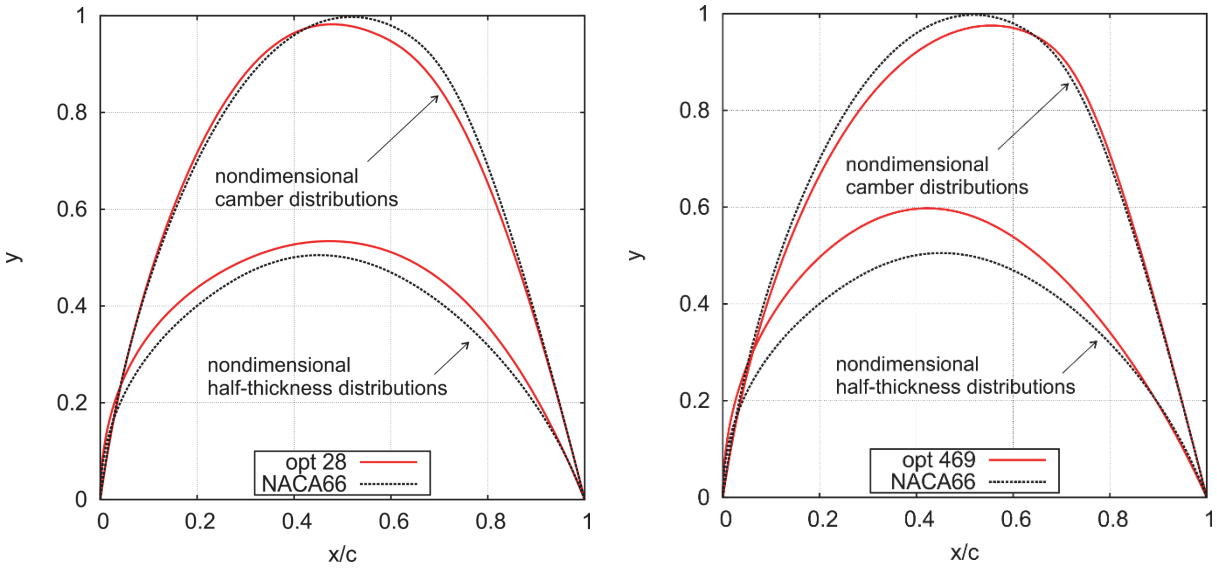



Figure 10: Comparison between the optimal (*opt28* and *opt469*) and the reference (NACA66Mod) hydrofoil geometries.

maximum camber that allows a mid-chord cavitating behavior similar, by both the numerical predictions, to that of the reference profile. As for the case I-inv optimization run, the lift over drag ratios (Table 2), computed a posteriori by using the RANSE solver, for the selected optimal geometries is always lower than that of the reference hydrofoil. The direct application, in the next section, of the viscous solver into the optimization process may definitely guide the design towards even efficient configurations.

| Geometry | Lift / Drag |
|---|---|
| NACA66Mod - a = 0.80, $t_{max}/c = 0.04$, $f_{max}/c = 0.02$ | 28.060 |
| Opt8S | 26.932 |
| Opt29S | 26.457 |
| Opt28 | 27.268 |
| Opt469 | 27.910 |

Table 2: Lift over drag ratios (OpenFOAM RANSE calculations) for the selected hydrofoils at the angle of attack of 0°.

### 3.2 Design of the optimal hydrofoil – viscous approach

An even accurate design by optimization was obtained by the application, directly in the design and not only as a verification tool, of the OpenFOAM viscous RANSE solver. As extensively pointed out in Dang [10][11], viscosity may play an important role. If in the aeronautical field, of course, the main goal of any new designs mainly concerns with the control of the boundary layer and, in turn, of the frictional forces [12], for maritime applications it is important to account for viscosity even if the attention is focused on cavitation. At first, obviously, to have more efficient designs requiring, as an additional objective, the maximization of the lift/drag ratio, for instance. Secondly to have more accurate designs (on the basis, of course, of a high-fidelity flow solver) and account or even exploit any interaction with the boundary layer to widen the cavitation bucket.

Similarly to what already proposed in the case of the inviscid calculations, two different optimization runs have been proposed. Case I-visc was carried out with a constraint on the maximum hydrofoil thickness (equal to that of the reference NACA66Mod) for a total of 3000 geometries selected by the optimization algorithm while in case II-visc, in analogy with case II-inv, this constraint was neglected and only 1500 geometries were tested. Even if restricted to 2D calculations, the computational efficiency of the viscous solution is orders of magnitude lower than that of the Boundary element methods, forcing the solution only of a fraction of the geometries analyzed by the inviscid approach to comply with a reasonable design time. The amount of collected data is only sufficient to highlight some tendencies and some preliminary optimal results, as evidenced by the Pareto frontiers clearly identified in Figure 11 and 12. In addition to the objectives related to cavitation, monitored on the axes and by the size of the markers, the color scale evidences the lift/drag ratio.

As it can be appreciated by the analysis of the Pareto diagrams and the cavitation buckets of two extracted geometries (*opt33SV* for case I-visc and *opt8V* for case II-visc) the inclusion of an additional objective (the lift over drag ratio) and the relatively limited number of evolved generations strongly condition the designs. Both the geometries were selected in the light of a balanced design (simultaneous improvement of the entire set of objectives) trying to limit as much as possible the mid-chord bubble cavitation, that is recognized to be one of the most dangerous cavitating phenomena. As in the case of inviscid calculations, the maximum thickness is shifted towards the leading edge, as shown in Figure 14. An evident difference between *opt33SV* and *opt8V* can be, however, appreciated: the latter is characterized by a slightly higher value of maximum thickness, allowed in the optimization process of case II-visc, and a substantial tapering at the trailing edge.

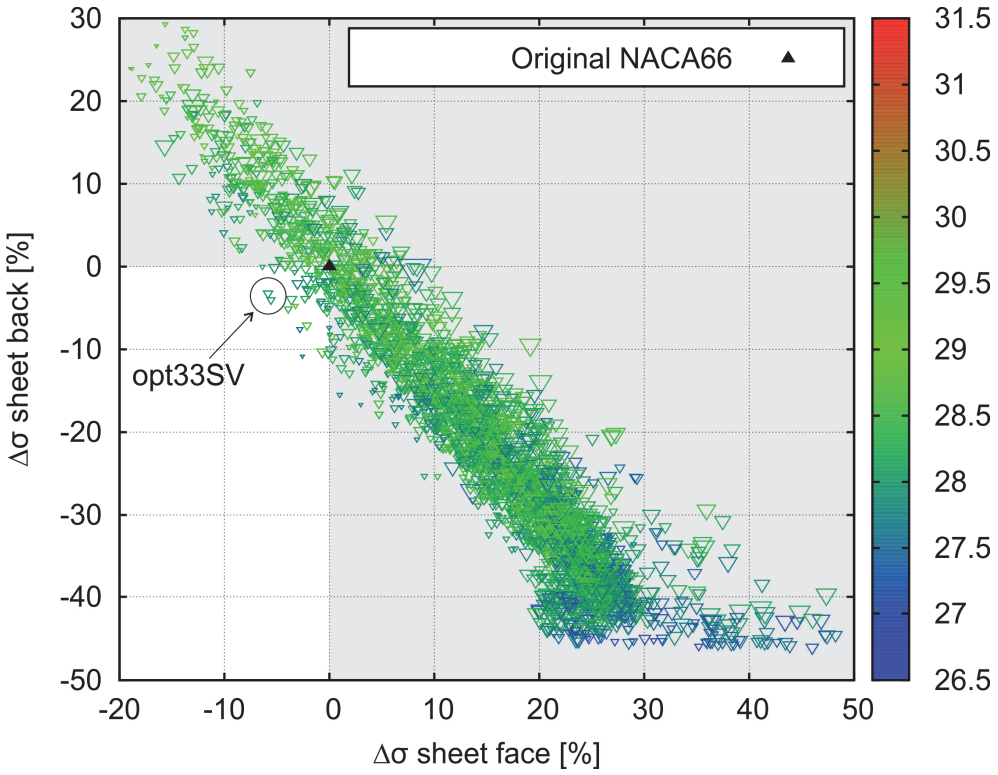



Figure 11: Pareto diagram of the viscous optimization – case I-visc.

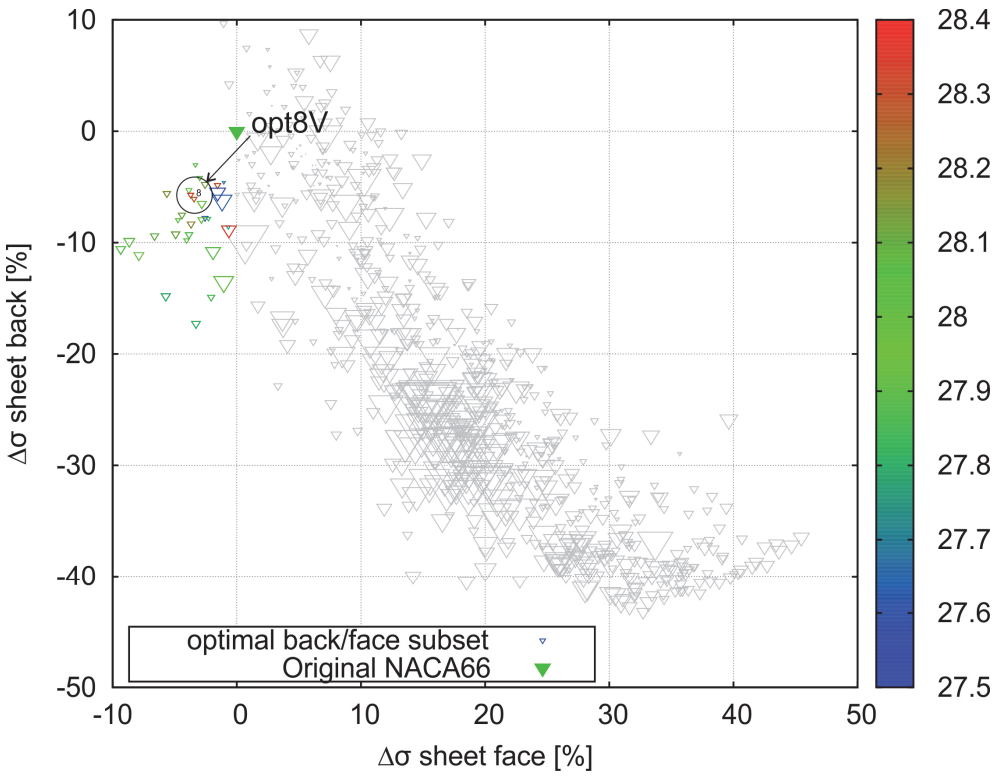



Figure 12: Pareto diagram of the viscous optimization – case II-visc.

The improvements achieved by both the optimizations are highlighted by the bucket diagrams of Figure 13 and by the comparison of the lift over drag ratio of Table 3. All the cavitating analyses were carried out also with the panel method in order to further verify the reliability of the panel method, eventually only from a comparative point of view, in a case where also viscous were considered for the design.

Optimization of case I-visc (*opt33SV*), despite the constraint on the maximum thickness, provides slightly better cavitating performance (with respect to both the reference NACA66Mod and the unconstrained thickness optimal shape *opt8V*) on both the suction and the pressure sides with a negligible worsening, however, of the mid-chord phenomena.

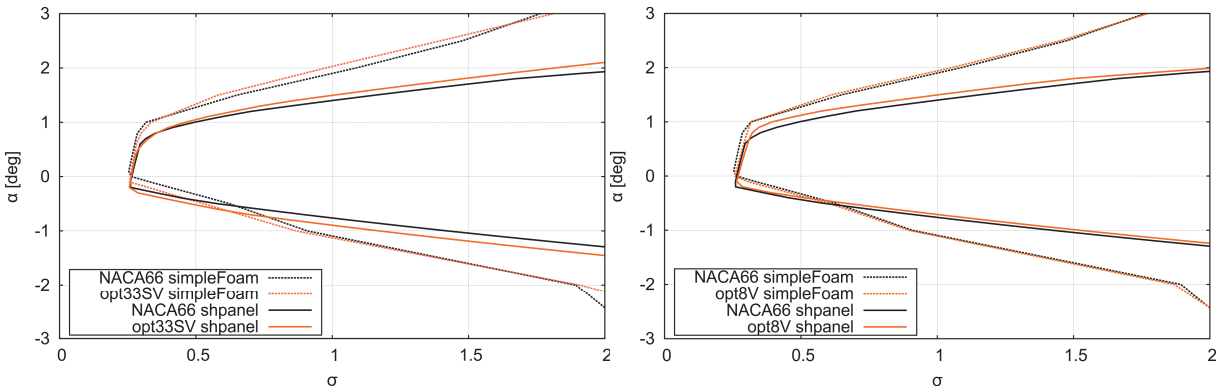



Figure 13: Comparison between the case I-visc optimal (*opt33SV* on the left), case II-visc optimal (*opt8V*) and the reference (NACA66Mod) cavitation buckets. Boundary element method (shpanel) and RANSE (simpleFoam) calculations.

| Geometry | Lift / Drag |
|---|---|
| NACA66Mod - a = 0.80, $t_{max}/c$ = 0.04, $f_{max}/c$ = 0.02 | 28.060 |
| Opt33SV | 28.150 |
| Opt8V | 28.265 |

Table 3: Lift over drag ratios (OpenFOAM RANSE calculations) for the selected hydrofoils at the angle of attack of 0°.

A single point in correspondence of which the mid-chord phenomena are monitored, as well as only one "positive" and one "negative" angles in correspondence of which the leading edge cavitation phenomena are controlled in the optimization process, may impair the design, producing geometries (i.e. *opt8V*) only "locally" better. Only the analysis of the entire cavitation bucket may effectively shows the overall fitness of the design that, however, is verified in the case of the *opt33SV* hydrofoil. With the exception of the higher values of the angle of attack subjected to bubble cavitation, evidenced by the panel method as well, both the suction and the pressure side cavitation performance and the bubble cavitation close to the monitored point are improved with respect to reference geometry. Differently from the inviscid optimizations, that achieved optimal geometries with better cavitating performance but at the cost of a reduced lift/drag ratio (all the analyzed geometries from cases I-inv and II-inv), the *opt33SV* provides, in addition, a slightly reduced value of frictional resistance that results in a slightly improved lift/drag ratio. *Opt8V*, on the contrary, from the cavitation point of view is only negligibly improved. Back and face cavitation is postponed only locally while the overall performance of the hydrofoil are exactly comparable with those of the reference NACA. For values of the angle of attack close to 0° the pressure side cavitation is even slightly worsened, as well as the mid-chord bubbles that are slightly anticipated with respect to the original geometry. This behavior, foreseen by the Boundary element method too, is balanced by the higher lift over drag ratio that can be achieved with this highly tapered geometry.

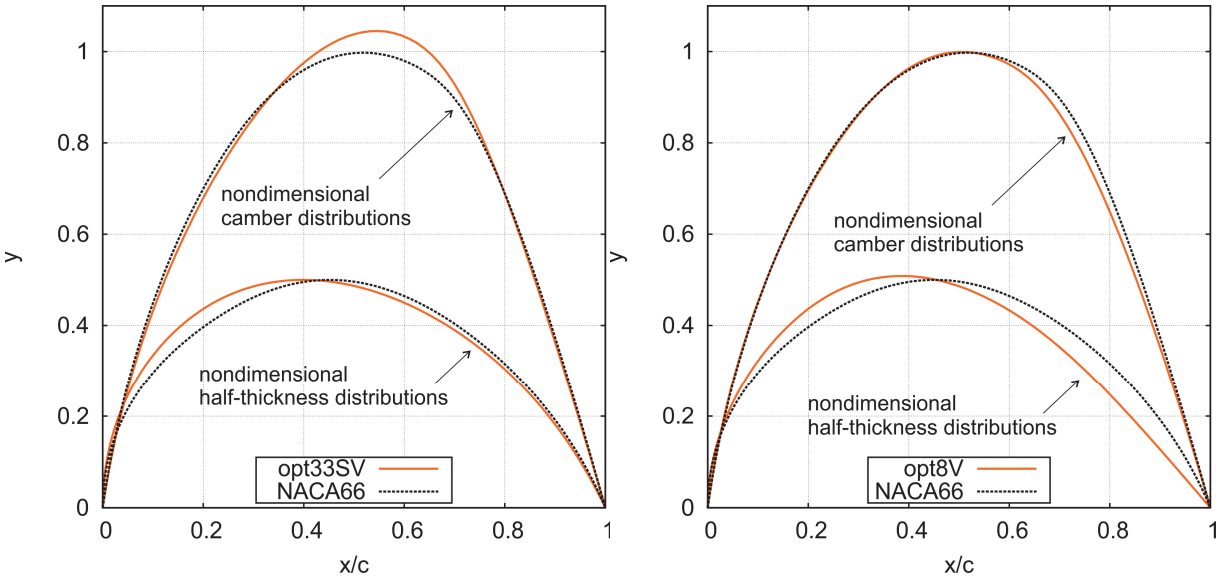



Figure 14: Comparison between the optimal case I-visc (*opt33SV* on the left), case II-visc optimal (*opt8V*) and the reference (NACA66Mod) hydrofoil geometries.

## 4 CONCLUSIONS

A computational framework for the design by optimization of marine hydrofoils under cavitating conditions has been proposed. The design approach relies on a robust parametric description of the hydrofoils geometry, to derive fair and smooth shapes appropriate for subcavitating operations, a multi-objective genetic algorithm and adequate flow solvers (a dedicated Boundary element method and the OpenFOAM RANSE solver) to obtain fast and reliable prediction of the hydrodynamic performance of the bi-dimensional sections under investigation.

Preliminary designs by using the inviscid approach have been proposed, in order to investigate trends and the possible improvements achievable with a very fast computational approach. High fidelity RANSE analyses have been furthermore carried out for the shapes devised by the inviscid calculations in order to verify the robustness of the BEM predictions, also in the light of the validation analysis proposed for a NACA0012 profile. The optimized geometries shown an encouraging widening of the cavitation bucket of the reference NACA66Mod hydrofoil, not only limited to pressure and suction side cavitation. Despite the thinness, that naturally postpones mid-chord bubble phenomena, of the geometries under investigation, a certain improvement of the cavitating performances for angles of attack close to the ideal one has been achieved as well. These results were exactly confirmed by the RANSE calculations, which identically ranked the devised geometries and, consequently, validated the design procedure based on the relative merit of a geometry with respect to the others rather than on absolute values. Only a slightly worsening of the lift of drag ratio, verified a posteriori with the viscous calculations, was observed in the case of the newly designed geometries.

The viscous optimization, further to provide even accurate predictions, was aimed to the definition of optimal hydrofoil shapes also from the “efficiency” point of view, by including as an additional objective of the optimization, the maximization of the lift over drag ratio. The designs by optimization obtained with the viscous flow solver, even if bottlenecked by the relatively small number of geometries analyzed, demonstrated, indeed, the possibility to define geometries that provide better lift/drag ratios. The Pareto diagrams obtained by using the OpenFOAM solver showed rather limited improvements in terms, for instance, of both back and face cavitation with respect to the remarkable improvements evidenced by pure Boundary element calculations. These outcomes are, however, in line with the improvements calculated by the viscous approach when applied to the designs devised by the inviscid solver. The optimization of hydrofoils under cavitating conditions, consequently, results a challenging problem, especially if the parent geometry is an already well-designed hydrofoil. Only small improvements, once calculated with the high fidelity RANSE solver were achieved, but it was possible to obtain an almost overall geometry (*opt33SV*, for instance) able to postpone suction side and pressure side sheet cavitation and mid-chord bubble phenomena as well (at least for some angles), with an higher lift over drag ratio.

Some points, however, are worth of further analyses and investigations. The relatively small number of geometries tested in the viscous optimization runs, of course, was not adequate to accurately populate the Pareto diagram and clearly identify optimal candidates in a complex, and with contradictory objectives, design. In addition, wider variations of the free parameters should be allowed in order to explore, by using the accurate flow solver represented by the OpenFOAM RANSE tool, even very non-conventional designs, in principle generally avoided when Boundary element methods are applied: out of their range of applications, their inherent assumptions turn into unreliable predictions. Wider free parameters variations, in turn, mean the necessity of a robust finite volume meshing tools able to automatically handle unconventional shapes, especially in present case, for what regards the structured, near-

wall, part of the mesh. Finally, a more accurate monitoring of the cavitation bucket (by using more than the three angles considered in present calculations) may help in devising geometries with globally (instead than locally) wider bucket diagrams. If the analysis, in the optimization process, of a further set of angles by using the Boundary element method is computationally insignificant, in the case of a viscous driven optimization it additionally reduces the computational efficiency of the design. At the end of the analyses and as the results of all the cross-checked calculations, thanks to comparative nature of the optimization process, the Boundary element method results an efficient tool for a preliminary design of hydrofoils with widened cavitation bucket, restricting viscous calculations only in the case of lift/drag optimizations.

## REFERENCES


[1] S.V. Andersen, P. Andersen, 1986, Hydrodynamic Design of Propellers with Unconventional Geometry, *Transaction of the Royal Institution of Naval Architects*, **129**, 201-221, The Royal Institution of Naval Architects, London, 1986.

[2] G. Kuiper, S.D. Jessup, A Propeller Design Method for Unsteady Conditions, *Proceedings of the SNAME, SNAME Centennial Meeting*, New York, 1993

[3] D. Bertetta, S. Brizzolara, S. Gaggero, M. Viviani, L. Savio, CPP propeller cavitation and noise optimization at different pitches with panel code and validation by cavitation tunnel measurements, *Ocean Engineering*, **53**, 177-195, 2012.

[4] S. Gaggero, J.G. Adalid, M. Perez-Sobrino, Design of contracted and tip loaded propellers by using boundary element methods and optimization algorithms, *Applied Ocean Research*, **55**, 102-129, 2016.

[5] S. Gaggero, C.M. Rizzo, G. Tani, M. Viviani, EFD and CFD design and analysis of a propeller in decelerating duct, *International Journal of Rotating Machinery*, 2012.

[6] M. Martelli, M. Figari, M. Altosole, S. Vignolo, Controllable pitch propeller actuating mechanism, modelling and simulation, *Proceedings of the Institution of Mechanical Engineers Part M: Journal of Engineering for the Maritime Environment*, **228**, 29-43, 2014.

[7] T. Brockett, Minimum pressure envelopes for Modified NACA-66 Sections with NACA a=0.8 camber and Buship Type I and Type II sections, *David Taylor Model Basin*, Report n. 1780, 1966.

[8] Y.T. Shen, Wing Sections for Hydrofoils – Part 3: Experimental Verification, *Journal of Ship Research*, **29**, 39-50, 1985.

[9] R. Eppler, Direct Calculation of Airfoils from Pressure Distribution, *NASA technical translation, NASA TT F-15 417*, 1974.

[10] J. Dang, New Blade Section Design and Its Application, *Proceedings of the Third International Symposium on Cavitation*, Grenoble, France, pp. 301-308, April 1998.

[11] J. Dang, Improving Cavitation Performance with New Blade Sections for Marine Propellers, *International Shipbuilding Progress*, **51**, 353-376, 2004.

[12] O. Amoignon, AESOP – A numerical platform for aerodynamic shape optimization, *Journal of Optimization in Engineering*, **11**, 555-581, 2010.

[13] The OpenFOAM Foundation. OpenFOAM 2.3.0 users guide, 2014.

[14] J.L. Hess, The Problem of Three-Dimensional Lifting Potential Flow and Its Solution by Means of Surface Singularity Distributions, *Computer Methods in Applied Mechanics and Engineering*, **4**, 283-319, 1974.

[15] N. Gregory, C.L. O'Reilly, Low-Speed Aerodynamic Characteristics of NACA 0012 Aerofoil Sections, including the Effects of Upper-Surface Roughness Simulation Hoar Frost, *NASA R&M 3726*, 1970.

[16] C. Geuzaine, J.-F. Remacle, Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities, *International Journal for Numerical Methods in Engineering,* **79** (11), 1309-1331, 2009.

[17] S.L. Krist, R.T. Biedron, C.L. Rumsey, CFL3D User's Manual (Version 5.0), *NASA Langley Research Center, NASA/TM-1998-208444*, 1998.

[18] C.L. Rumsey, Turbulence Modeling Resource, *NASA Langley Research Center*, http://turbmodels.larc.nasa.gov/index.html

[19] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computational*, **6**, 2, 2002.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A COMPUTATIONAL STRATEGY FOR TRAJECTORY OPTIMIZATION OF UNDERACTUATED MULTIBODY SYSTEMS WITH CONTACTS

**Silvia Manara**[1], **Alessio Artoni**[1], **and Marco Gabiccini**[1,2,3]

[1]Dipartimento di Ingegneria Civile e Industriale
Largo Lucio Lazzarino 1, 56122 Pisa, Italy
e-mail: silvia.manara@for.unipi.it
{a.artoni,m.gabiccini}@ing.unipi.it

[2] Research Center "E. Piaggio", University of Pisa
Largo Lucio Lazzarino 1, 56122 Pisa, Italy

[3] Department of Advanced Robotics, Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy

**Keywords:** Optimization, Trajectory Planning, Optimal Control, Computational Methods.

**Abstract.** *In this paper we propose a strategy for improving the computational efficiency of direct methods for trajectory optimization of multibody systems. We particularly focus on those applications where the system necessarily has to interact with the surrounding environment through intermittent contacts. The problem is hereby formulated such that just the initial and final states of the system over a given time interval are prescribed, so as to let the solver automatically synthesize the best contact sequence to accomplish the considered task. The proposed computational strategy consists in: (i) solving a preliminary optimization problem that roughly approximates the original one, but differs from it by one or more conveniently chosen parameters and is faster to solve; (ii) using the obtained solution as an initial guess for the actual (full-fledged) optimal control problem. The performance of the method is evaluated in a simulated planar system, whose peculiarity is to be trivially underactuated. An extensive investigation is presented which shows how a proper choice of the parameters in the preliminary optimization can lead to a significant reduction in the computational effort required to solve the problem. The results we present assess both the validity and the robustness of the proposed method.*

## 1 INTRODUCTION

The analysis of multibody systems interacting with the surrounding environment through intermittent contacts gives rise to non-trivial issues, mainly caused by the discontinuities due to friction and impulsive dynamics. However, in contexts such as robotic manipulation and locomotion, interactions between the system and parts of the environment through unilateral contact forces are essential in order to accomplish a given task.

The complexity of the trajectory planning problem in such contexts is due to the fact that, over time, the system may be in an enormous number of possible configurations, separated by collision events occurring when a new contact is established or a prior one is broken. However, reducing the search space dimension by *a priori* scheduling the contact sequence to be followed by the system may sometimes be difficult and, more notably, so restrictive to lead to the planning of unnatural behaviors. On the other hand, the combinatorial explosion of possible configurations caused by unspecified contacts can hardly be faced by sampling-based planners [1, 2]. Even the recently proposed multi-modal planners [3] can cope with just a limited number of contact modes.

Although promising methods for reducing the search space have been recently proposed [4], the robotics community is showing an increasing interest in optimization-based methods [5]. In fact, such methods, also known as *direct transcription* methods, are not significantly affected by the presence of *a priori* unknown interaction forces, since the formulation of the problem as an optimal control one allows the contact forces to be simply embedded in the structure of the problem itself, without altering it. Different approaches have been proposed to represent unilateral contact forces. In [6], where the contact invariant optimization is presented, a penalty term in the cost function accounts for the feasibility of the contact forces. In other works [7, 8], contact forces were added to the problem variables and their constitutive equations were listed among the constraints of the optimization problem. Since gradient-based methods can, in general, very hardly handle discontinuities, special care must be devoted to modeling contact forces. In [7], a complementarity formulation of contact (resulting in a Linear Complementarity Problem, LCP) is employed, whereas in [8] two different contact models, respectively based on complementarity and on a continuous penalty-based formulation, are evaluated. In [9], where an optimization framework for trajectory planning is proposed and applied to a simplified biped system performing a one-step walking motion, smoothing functions are introduced to relax the complementarity conditions involving normal and tangential contact forces.

Other investigations about the potential of gradient-based methods have mainly focused on trajectory planning of underactuated systems. The particular significance of this kind of systems is that they are paradigmatic of locomotion problems, where at least the degrees of freedom (DoF) of the floating base are not actuated. In [10], where a very complex humanoid model is considered, the problem is approached by considering just the centroidal dynamics of the system, as the resulting equations describe the effect of external wrenches (due to contact and gravitational forces) on the underactuated degrees of freedom. Instead in [11], the full system dynamics is considered. Here, a very detailed model of human lower limbs is used and contact invariant optimization is employed to make it discover complex motions, including walking. Different ways of ensuring dynamic feasibility have been proposed as well. In [11], consistently with the contact invariant approach, the violation of the equations of motion is heavily penalized in the cost function, whereas in other works the dynamics of the systems is added as a set of nonlinear constraints to the optimal control problem. Among these last mentioned works, an interesting contribution is given in [12], where an extremely simplified model of a

floating base legged robot is employed and, without specifying the contact sequence, different gaits are identified through direct trajectory optimization. Other assumptions are made, such as periodicity, which make this work focus more on the optimization of the stride itself rather than on finding more general walking behaviors.

Recent research efforts in robotics [13] have been devoted to provide further insight into the performance of direct optimization methods. In this paper, we present a study that aims to assess the performance through extensive numerical tests of different direct trajectory optimization schemes in terms of computational time and sensitivity to parameters. Additionally, an original strategy is devised that substantially reduces the computational time to convergence. Both the validity and the robustness of the proposed method are evaluated through simulation of a 3-DoF underactuated 2D system.

## 2 MODEL DESCRIPTION

### 2.1 System structure and task definition

The computational performance of different direct trajectory optimization schemes, which is the main object of the present paper, was evaluated for planning the trajectory of a benchmark system, represented in Fig. 1.

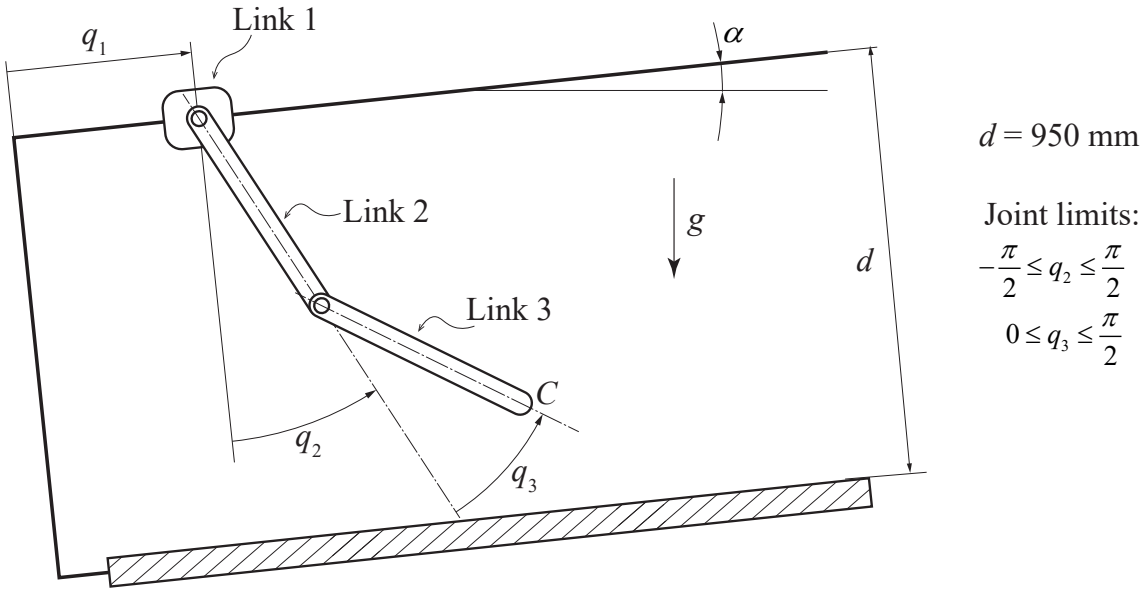



Figure 1: PRR model.

| Link 1 | Link 2 | Link 3 |
|---|---|---|
| $m_1 = 5$ kg | $m_2 = 3$ kg | $m_3 = 3$ kg |
| | $l_2 = 500$ mm | $l_3 = 500$ mm |

Table 1: Geometric and inertial properties of the system.

This is a 3-DoF planar PRR system, consisting of three links connected together by two actuated revolute (R) joints and attached to a supporting frictionless track through a non-actuated

prismatic (P) joint.

From a given initial state, the system has to reach a specified final position, which is 3 m away along the direction of the track, within a fixed time horizon, and stop there. It is worth pointing out that this problem is not as simple as it could appear, as the system under investigation is trivially underactuated, and therefore it necessarily has to exploit the interaction with the underlying floor in order to produce the thrust force required to climb up the slope until the target final position is reached.

### 2.2 Formulation of the system dynamics

For the description of the system dynamics, we employ the Euler-Lagrange equations for holonomic systems:

$$B(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) = Q(q,\dot{q},u) \tag{1}$$

where $q \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ collect, respectively, the configuration variables and the input torques at the revolute joints, $B(q) \in \mathbb{R}^{n\times n}$ is the inertia matrix, $C(q,\dot{q}) \in \mathbb{R}^{n\times n}$ is the matrix of Coriolis and centrifugal terms and $G(q) \in \mathbb{R}^n$ is the vector of gravitational forces. $Q(q,\dot{q},u) \in \mathbb{R}^n$ is the vector of generalized forces, whose $j$-th component can be expressed as:

$$Q_j(q,\dot{q},u) = \begin{cases} J_{v_C}(q)[:,j]^\top f_C(q,\dot{q}) & \text{if } j \in \bar{\mathscr{A}}, \\ J_{v_C}(q)[:,j]^\top f_C(q,\dot{q}) + u_j & \text{if } j \in \mathscr{A} \end{cases} \tag{2}$$

where $J_{v_C}$ is the linear velocity Jacobian of the contact point $C$, $f_C$ is a vector that represents the force acting at $C$ because of the contact with the ground, and $u_j$ is the input torque applied to the $j$-th joint. $\mathscr{A}$ and $\bar{\mathscr{A}}$ are, respectively, the set of actuated and non-actuated degrees of freedom, having cardinalities $|\mathscr{A}| = m$ and $\left|\bar{\mathscr{A}}\right| = n - m$.

### 2.3 Contact force model

The optimization of trajectories for systems with intermittent contacts requires a careful modeling of contact forces. In this paper we employ a penalty-based model, as it provides smoothness which conveniently fits with the Newton-type algorithms used to solve the optimization problem.

We approximate the unilateral linear contact model by the following differentiable constitutive relation (Fig. 2) between the normal component of the contact force $f_n$ and the normal gap $g_n$ separating the candidate contact point $C$ from the ground (it depends on the internal configuration of the system, i.e. $g_n = g_n(q)$)

$$f_n(y) = f_0 \log_2\left(1 + 2^{-\frac{\kappa}{f_0} g_n}\right) \tag{3}$$

where $f_0$ is the force value at $g_n = 0$, and $-\kappa$ is the contact stiffness as $g_n \to -\infty$.
The parameters $f_0$ and $\kappa$ provide a straightforward way to properly tune the model. For the problem under investigation, we chose to set $f_0 = 20$ N and $\kappa = 10^4$ N/m, as this tuning results in a realistic modeling of contact actions, without exacerbating the numerical stiffness of the problem. With these values, in fact, a very small penetration ($g_n = -5$ mm) yields a value of the normal force ($f_n \approx 50$ N) of the same order of magnitude as the forces actually required in our problem. Also, with this choice of the parameters, the effect of contact forces acting at a distance — which is one of the main drawbacks of the penalty-based formulation of contact —

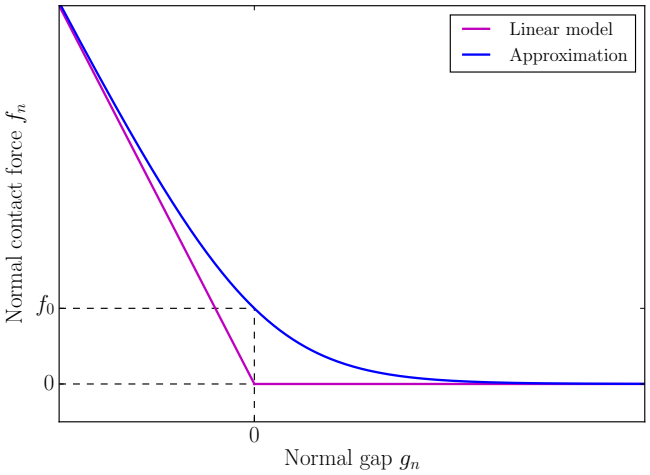



Figure 2: Normal contact force model.

is so small to be negligible: its value at $g_n = 10$ mm is less than 1 N ($f_n \approx 0.9$ N) and rapidly vanishes as the normal gap increases.

Also the tangential component of the contact model is approximated, according to a regularization strategy of the stick-slip behavior, as in [14]. For a 2D problem, the relation between the tangential component of the contact force $f_t$ and the tangential velocity $V_t$ of the contact point can be modeled as follows:

$$f_t(g_N, V_t) = \mu f_n(g_N)\gamma(V_t), \quad \text{with} \quad \gamma(V_t) = -\tanh\left(\frac{V_t}{\widehat{V}_t}\right) \tag{4}$$

Here, $\mu$ is the coefficient of friction, $\gamma$ is a smooth function that approximates the Coulomb model (Fig. 3), and $\widehat{V}_t$ is a reference sliding velocity at which the tangential force $f_t$ is 76% of its asymptotic value. This tangential force model requires tuning as well. We set $\widehat{V}_t = 5 \cdot 10^{-3}$ m/s,

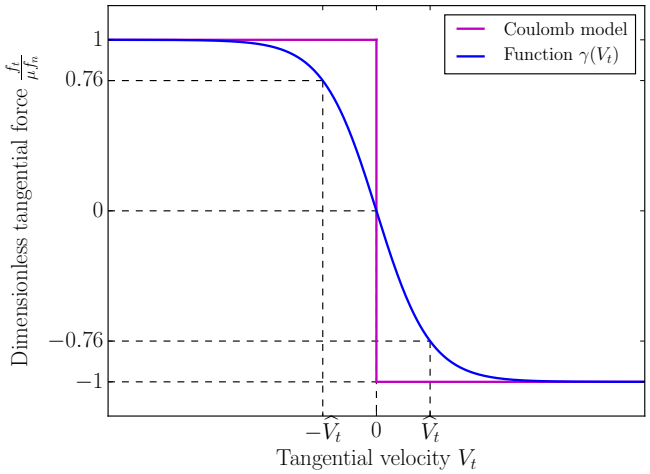



Figure 3: Coulomb friction model and its smooth approximation.

as the corresponding model appears to be a fair balance between physical plausibility and numerical tractability. The friction coefficient was fixed to $\mu = 1$.

## 3 FORMULATION OF THE OPTIMAL CONTROL PROBLEM

### 3.1 Discretization

The solution of the optimal control problem through a direct transcription method requires the parameterization of the continuous optimization problem into a finite dimensional one, through discretization of the time horizon $[0, T]$ into $N$ time intervals $[t_k, t_{k+1}]$, so that $t_0 = 0$, $t_N = T$ and $k \in \{0, 1, \ldots, N-1\}$. It is worth emphasizing that problems involving intermittent contacts require very small time steps, because of the typical stiffness of the associated differential equations. For this reason, in our problem we chose to discretize the time horizon $T = 6$ s into $N = 200$ uniform steps, each having duration $h = T/N = 30$ ms.

Hence, discrete time versions of both the dynamics of the system and the contact force model are needed. We parameterize these equations by a collocation scheme based on a single collocation point, chosen in the middle of each time interval. Over each discretization interval, we assume the states $(q, \dot{q})$ to vary linearly, and the controls (which include the torques at the revolute joints $u$, together with the components of the contact force $f_C$) to be constant. According to this scheme, *kinematic reconstruction* and *dynamic equations* of the system can be written as follows:

$$q_{k+1} - q_k - h\bar{\dot{q}}_k = 0 \tag{5}$$

$$\bar{B}_k(\dot{q}_{k+1} - \dot{q}_k) + h(\bar{C}_k\bar{\dot{q}}_k + \bar{G}_k - \bar{Q}_k) = 0 \tag{6}$$

where, for simplicity of notation, we denoted $q_k = q(t_k)$, $\dot{q}_k = \dot{q}(t_k)$, $\bar{q}_k = (q_{k+1} + q_k)/2$ and $\bar{\dot{q}}_k = (\dot{q}_{k+1} + \dot{q}_k)/2$. Moreover, $\bar{B}_k = B(\bar{q}_k)$, $\bar{C}_k = C(\bar{q}_k, \bar{\dot{q}}_k)$, $\bar{G}_k = G(\bar{q}_k)$, and $\bar{Q}_k = Q(\bar{q}_k, \bar{\dot{q}}_k, u_k)$.

### 3.2 Optimization

Let us define a vector of decision variables $w \in \mathbb{R}^\nu$, which collects the sequence of unknowns $(q_0, \dot{q}_0, u_0, f_{C_0}, \ldots, q_k, \dot{q}_k, u_k, f_{C_k}, \ldots, q_N, \dot{q}_N)$. Since we have $2n(N+1)$ discretized state variables, $mN$ discretized control actions, and $N$ variables for each component of the contact force, the total number of optimization variables is:

$$\nu = 2n(N+1) + (m+2)N \tag{7}$$

The nonlinear program (NLP) we need to solve in order to obtain the optimal trajectory — and the corresponding sequence of control actions to be taken — can be formalized as follows:

$$\begin{aligned} \min_{w} \quad & f(w) \\ \text{subject to} \quad & g_{\min} \leq g(w) \leq g_{\max} \\ & w_{\min} \leq w \leq w_{\max} \end{aligned} \tag{8}$$

$g(w)$ is a set of nonlinear constraints, including Eqs. (5)–(6), together with the discretized versions of the normal and tangential contact force models (Eqs. (3)–(4)). In addition, by properly enforcing $w_{\min}$ and $w_{\max}$, we can assign the initial and final configurations and impose joint and torque limits (torques are constrained in the range $[-50, 50]$ Nm).
$f : \mathbb{R}^\nu \to \mathbb{R}$ is the objective function to be minimized. In our case, it is made up of five different contributions, aimed at minimizing, respectively: input torques, accelerations, dissipation due to sliding, input torque variations and contact force variations. The objective function was

calibrated by properly scaling and weighting the different contributions, so as to obtain natural and smooth behaviors.

Problem (8) results in a large-scale but very sparse NLP, which we solve through the interior-point solver IPOPT [15]. For the calculation of derivatives, we benefit from the Algorithmic Differentiation (AD) implemented in the CasADi framework [16].

## 4 COMPUTATIONAL STRATEGY: TWO-STAGE OPTIMIZATION

The computational strategy we propose in this paper aims to reduce the computational effort required for the solution of the trajectory optimization problem. The NLP we consider as the reference problem we want to solve is characterized by the following choice of the parameters:

$$\begin{aligned} \text{Discretization} \quad & N = 200 \\ \text{Normal contact} \quad & f_0 = 20 \text{ N} \\ & \kappa = 10^4 \text{ N/m} \\ \text{Tangential contact} \quad & \widehat{V}_t = 5 \cdot 10^{-3} \text{ m/s} \end{aligned} \tag{9}$$

The computational strategy we devised is to solve this problem through a *two-stage* optimization. The first stage consists in solving a preliminary optimization problem which roughly approximates the original one, but is simplified by slight changes in some parameters, so as to speed up the solution. In order to suitably choose the values to assign to these parameters, we investigated the beneficial effects of:

- reducing the number of time steps $N$. This simplifies the problem by decreasing its dimension, as clearly appears from (7). We will refer to the number of time steps used in the first optimization as $N'$;
- relaxing the contact parameters, as this reduces the numerical stiffness of the problem. More in details, we smooth the tangential contact model by increasing the value of the reference tangential velocity $\widehat{V}_t$ (see Eq. (4)) in the first stage. We will indicate this value using the symbol $\widehat{V}_t'$. Regarding the normal contact model, we chose to relax it by reducing the value of the asymptotic stiffness $\kappa$ to the value $\kappa'$ in the preliminary optimization, while leaving the parameter $f_0$ unchanged.

Once this preliminary optimization is solved, we proceed to the second stage, using the just computed solution to warm-start the final optimization problem, whose parameters are summarized in (9). The results we present show how this two-stage strategy, with a proper choice of the parameters in the preliminary optimization, can lead to a significant reduction in the computational time required to obtain the solution.

## 5 RESULTS

In this section, we present the results of our analysis. All of them were obtained on a notebook computer with a 2.40 GHz Intel(R) Core(TM) i7-5500U processor and 8 GB of RAM. The performance is evaluated here in terms of CPU time to convergence.

### 5.1 Reference solution

The reference solution was obtained in a single-stage optimization. Namely, this solution was computed by solving the optimal control problem, characterized by the parameters detailed

in (9), without exploiting the advantage of solving a preliminary optimization with a coarser time grid, nor a relaxed contact model. The slope of both the track and the ground was set to $\alpha = \pi/30$.

The initial guess was selected so that all of the variables were zero except for the configuration variables $q$: the internal configuration angles were set to their initial state values (equal to those in the final state), whereas the prismatic coordinate $q_1$, representing the position of the PRR system along the track, was initialized using a cubic interpolation between the initial and final conditions. This choice was made with the precise intention of providing the solver with the minimum domain knowledge on how to accomplish the task, since we would like it to discover the optimal gait by itself.

The solution was obtained after $300$ s CPU time ($671$ iterations), and it is depicted in Fig. 4. The corresponding values of some of the optimization variables over the time horizon are plotted in Fig. 5.

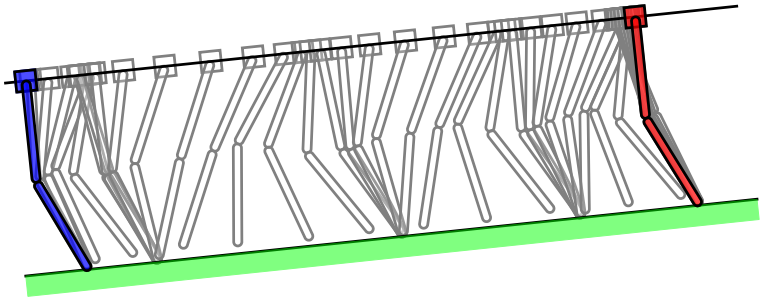


Figure 4: In gray, frames of the reference optimal trajectory. In blue and red, respectively, the initial and final configurations prescribed to the system.

### 5.2 Two-stage optimization strategy

Let us now compare the performance of the proposed two-stage optimization strategy against the reference single-stage solution. The strategy described in Sec. 4 consists in solving a preliminary approximated optimization, and using its solution to warm-start the full-fledged optimal control problem. The preliminary optimization differs from the reference problem by one or more parameters: in particular, we investigated the effect of enlarging the time steps and relaxing the parameters of the contact model. The first stage of optimization is initialized in the same naive way as illustrated in Sec. 5.1. The obtained solution is then interpolated on a finer grid and used to initialize the solver of the second optimization, whose parameters are the reference ones listed in (9).

The results are presented in tables, which are organized as follows:

- each table refers to a specific value of $\kappa'$, that is the value adopted for the normal contact stiffness in the preliminary optimization: Table 2 reports the results obtained when the normal contact model is left unchanged through the two stages, whereas Tables 3 and 4 refer to results obtained by dividing the normal contact stiffness respectively by 2 and by 4 in the first stage;

- in each table, the results obtained for different levels of relaxation of the tangential contact force model in the first optimization are organized in columns (labelled by the corre-

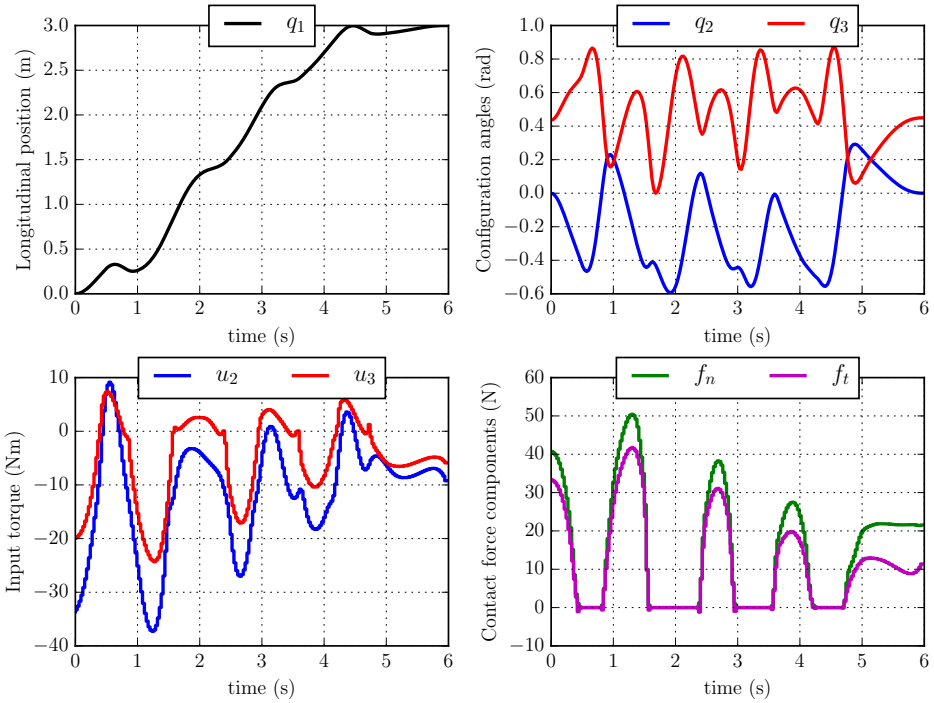



Figure 5: Reference solution: configuration variables, input torques and contact force as functions of time.

sponding value of $\widehat{V}_t'$);

- in each table, the results obtained for different numbers of time steps $N'$ in the first optimization are organized in rows;
- in each cell of the table, the performance of the corresponding two-step optimization in terms of CPU time to convergence is reported (in seconds).
  Specifically, the first (second) term stands for the computational time required to solve the first (second) optimal control problem. The word FAILED indicates that the solver was not able to find any solution to the preliminary optimization within the maximum number of iterations imposed (5000).

All of these results refer to the case $\alpha = \pi/30$.

### 5.3 Remarks

Let us now focus on what we can deduce from our analysis. First of all, we notice that in most cases, compared to the reference solution (obtained after 300 s CPU time), the two-stage strategy leads to a significant reduction in the computational time required to obtain a sensibly similar solution: despite direct optimization methods suffer the risk of being trapped in local minima, all of the optimal solutions found are comparable in terms of value of the objective function. The mean computational time saving over all those cases in which the solver found a solution to the first stage is of 163.5 s (54.5% of the reference CPU time). In the most favourable case, which corresponds to adopting $N' = 100$, $\kappa' = \kappa/2$ and $\widehat{V}_t' = 2\widehat{V}_t$ in the preliminary phase, this strategy led to an 85.3% reduction in the CPU time required to compute a solution, as can be seen from Table 3.

| | $\widehat{V}_t' = \widehat{V}_t$ | $\widehat{V}_t' = 2\widehat{V}_t$ | $\widehat{V}_t' = 4\widehat{V}_t$ |
|---|---|---|---|
| $N' = 75$ | $32 + 110 = 142$ | $34 + 33 = 67$ | $20 + 125 = 145$ |
| $N' = 100$ | $97 + 82 = 179$ | $87 + 21 = 108$ | $19 + 51 = 70$ |
| $N' = 125$ | $63 + 27 = 90$ | $74 + 50 = 124$ | $76 + 74 = 120$ |

Table 2: Results obtained for $\kappa' = \kappa$ (CPU time in seconds).

| | $\widehat{V}_t' = \widehat{V}_t$ | $\widehat{V}_t' = 2\widehat{V}_t$ | $\widehat{V}_t' = 4\widehat{V}_t$ |
|---|---|---|---|
| $N' = 75$ | $37 + 372 = 409$ | $19 + 57 = 76$ | $16 + 51 = 67$ |
| $N' = 100$ | $230 + 60 = 290$ | $28 + 16 = 44$ | $15 + 68 = 83$ |
| $N' = 125$ | FAILED | $82 + 22 = 104$ | $20 + 65 = 85$ |

Table 3: Results obtained for $\kappa' = \kappa/2$ (CPU time in seconds).

| | $\widehat{V}_t' = \widehat{V}_t$ | $\widehat{V}_t' = 2\widehat{V}_t$ | $\widehat{V}_t' = 4\widehat{V}_t$ |
|---|---|---|---|
| $N' = 75$ | $55 + 41 = 96$ | $73 + 28 = 101$ | $11 + 71 = 82$ |
| $N' = 100$ | $194 + 35 = 229$ | $35 + 120 = 155$ | $46 + 60 = 106$ |
| $N' = 125$ | FAILED | $72 + 28 = 100$ | $22 + 317 = 339$ |

Table 4: Results obtained for $\kappa' = \kappa/4$ (CPU time in seconds).

We could observe that in those cases in which the strategy was not effective — the solver was not able to find a solution to the preliminary optimization or the total CPU time exceeded $300$ s — this was plausibly due to an undesirable mix of a smoothed normal contact ($\kappa' = \kappa/2$ or $\kappa' = \kappa/4$) with a stiff tangential contact ($\widehat{V}_t' = \widehat{V}_t$), see first column of Table 3 and Table 4. In fact, in these conditions, because of both the layout of the system and the fact that sliding behaviors are penalized in the objective function, the solver finds some difficulties in computing solutions which satisfy the constraints while minimizing the objective function. In fact, reducing the normal contact stiffness has the side effect of increasing the intensity of the normal force acting at a distance. Without relaxing the tangential contact model, even small values of the tangential velocity $V_t$ at a distance from the surface of the ground result in a sensible contribution in the term of the objective function which penalizes the dissipation due to sliding. Even in the event that the solver succeeds, the solution could represent an unnatural trajectory (the foot has to move very far from the ground), constituting an inadequate initial guess for the second optimization, which may require a large amount of time to convergence (see for example the case $N' = 75, \kappa' = \kappa/2, \widehat{V}_t' = \widehat{V}_t$ in Table 3).

Concerning the number of steps $N'$ to be used in the first stage, we notice that in those cases where the contact force model is accurate (first column in Table 2), the best choice is to use a high number of steps in the first optimization: since the first problem to be solved, although complex, is very similar to the actual one, its solution is a very good initial guess for the second stage, and the solver requires a short time to find a solution to the second optimization. On the other hand, the more the contact model is smoothed in the first stage, the more it is convenient to use fewer integration steps: since the first problem is a very simplified version of the original one, performing a consistent coarse discretization of it seems to be the best choice (last column in Table 4).

### 5.4 Robustness of the method

In order to test the validity and robustness of the proposed method, we picked the best performing mix of parameters from our analysis and employed it in the two-stage optimization strategy for a slightly different problem, where the system and the task to be accomplished (move 3 m forward along the track in $T = 6$ s, then stop) are the same, but the slope angle $\alpha$ is different. Without performing any analogous analysis for the modified problem, we exploit the results from the study conducted for the reference problem to tune up a two-stage optimization strategy and apply it in the new scenarios. In the first stage, we solved a preliminary optimization where the parameters were modified as follows:

$$\begin{aligned} \text{Discretization} \quad & N' = N/2 \\ \text{Normal contact} \quad & \kappa' = \kappa/2 \\ \text{Tangential contact} \quad & \widehat{V}_t' = 2\widehat{V}_t \end{aligned} \tag{10}$$

as this combination of parameters led to the best performance for the reference problem, and it seems to be a reasonable choice according to the outlined remarks. Of course, this may not be the best choice for the modified problem, yet our aim is here to prove that, from a computational perspective, solving the direct optimization problem in a two-stage fashion, with a reasonable slackening of the parameters in the first stage, is convenient compared to a single-stage optimization.

Also in these modified cases, the proposed strategy proved to be very efficient in improving the computational performance of the trajectory optimization, therefore assessing the robustness of the method. The results we present are arranged in tables, comparing the CPU time (in seconds) required to obtain the optimal solution for the single and two-stage optimization. Table 5 refers to a problem where the slope angle is half the reference value ($\alpha = \pi/60$), whereas Table 6 reports the results obtained in a problem where the slope angle is doubled ($\alpha = \pi/15$).

| Reference solution (single-stage optimization) | Two-stage optimization $N' = N/2$, $\kappa' = \kappa/2$, $\widehat{V}_t' = 2\widehat{V}_t$ |
|---|---|
| 403 | $31 + 60 = 91$ |

Table 5: Results obtained for $\alpha = \pi/60$ (CPU time in seconds).

| Reference solution (single-stage optimization) | Two-stage optimization $N' = N/2$, $\kappa' = \kappa/2$, $\widehat{V}_t' = 2\widehat{V}_t$ |
|---|---|
| 494 | $32 + 43 = 75$ |

Table 6: Results obtained for $\alpha = \pi/15$ (CPU time in seconds).

## 6 CONCLUSIONS

In this paper, we have devised and numerically assessed a computational strategy to improve the performance of direct trajectory optimization of systems interacting with the surrounding environment through intermittent contacts. This computational strategy consists in subsequently solving two optimal control problems of increasing complexity. A penalty-based

formulation for modeling the contact forces is employed throughout the paper. A straightforward method for simplifying the problem in the preliminary optimization was proposed and several tests were performed on a 2D underactuated multibody system. The results we presented prove the effectiveness and the robustness of the proposed strategy.

Our current research effort is focused on the application of similar strategies for Model Predictive Control of underactuated multibody systems with contacts, where computational efficiency is essential in order to cope with the strict requirements of replanning the trajectory online.

## REFERENCES

[1] S. M. LaValle. *Planning algorithms*. Cambridge University Press, 2006.

[2] S. Karaman and E. Frazzoli. Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research (IJRR)*, 30(7):846–894, 2011.

[3] K. Hauser and V. Ng-Thow-Hing. Randomized multi-modal motion planning for a humanoid robot manipulation task. *The International Journal of Robotics Research (IJRR)*, 30(6):678–698, 2011.

[4] S. Tonneau, N. Mansard, C. Park, D. Manocha, F. Multon, and J. Pettré. A reachability-based planner for sequences of acyclic contacts in cluttered environments. In *International Symposium on Robotics Research (ISRR 2015)*, 2015.

[5] J. T. Betts. *Practical methods for optimal control and estimation using nonlinear programming*. SIAM, 2nd edition, 2010.

[6] I. Mordatch, E. Todorov, and Z. Popović. Discovery of complex behaviors through contact-invariant optimization. *ACM Transactions on Graphics (TOG)*, 31(4):43–50, 2012.

[7] M. Posa, C. Cantu, and R. Tedrake. A direct method for trajectory optimization of rigid bodies through contact. *The International Journal of Robotics Research (IJRR)*, 33(1):69–81, 2014.

[8] M. Gabiccini, A. Artoni, G. Pannocchia, and J. Gillis. A computational framework for environment-aware robotic manipulation planning. In *International Symposium on Robotics Research (ISRR 2015)*, 2015.

[9] C. Mummolo, L. Mangialardi, and J.H. Kim. Concurrent contact planning and trajectory optimization in one step walking motion. In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2015)*, 2015.

[10] H. Dai, A. Valenzuela, and R. Tedrake. Whole-body motion planning with centroidal dynamics and full kinematics. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2014.

[11] I. Mordatch, J. M. Wang, E. Todorov, and V. Koltun. Animating human lower limbs using contact-invariant optimization. *ACM Transactions on Graphics (TOG)*, 32(6):203–210, 2013.

[12] W. Xi and C. D. Remy. Optimal gaits and motions for legged robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, 2014.

[13] D. Pardo, L. Möller, M. Neunert, A. Winkler, and J. Buchli. Evaluating direct transcription and nonlinear optimization methods for robot motion planning. *arXiv:1504.05803v1 [math.OC]*, 2015.

[14] P. Wriggers. *Computational Contact Mechanics*. Springer, New York, 1999.

[15] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[16] J. Andersson. *A General-Purpose Software Framework for Dynamic Optimization*. PhD thesis, Arenberg Doctoral School, KU Leuven, Department of Electrical Engineering (ESAT/SCD) and Optimization in Engineering Center, Kasteelpark Arenberg 10, 3001-Heverlee, Belgium, October 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ANALYSIS OF NON-LINEAR ENVIRONMENTAL LOAD COMBINATIONS BY EXTENDED CONTOUR-LINE ALGORITHMS

**Bernt J. Leira**[1]

[1] NTNU, Dept. Marine Technology
MTS, Otto Nilsens vei 10, N-7491 Trondheim, Norway
e-mail: Bernt.Leira@ntnu.no

**Keywords:** Environmental loads, Structural response, Non-linear combination, Algorithms.

**Abstract.** *For both ships and offshore structures, load effects caused by a combination of wind, current and waves need to be accounted for during the design process. It will easily become a very comprehensive task to consider all possible combinations of such load effects, and efficient methods to deal with this issue is generally in demand. The present paper deals with application of extended contour-line methods for this purpose. General methods for identification of load combinations in relation to continuous stochastic processes are first highlighted. The particular case of multiple FBC-processes with given amplitude distribution is subsequently addressed. (The FBC-process is named with reference to the work by Ferry-Borges&Castanheta (1971)). The case where all process components have identical basic time intervals is first considered. The relationship between the FBC-process and the so-called environmental contour methods (based on Inverse FORM techniques) which are presently being applied for various design purposes is elaborated. FBC-processes with widely different basic time intervals are next investigated.The present paper illustrates how a FORM search can be applied along the limit state surface in order to identify the relevant "load combination point" for such cases. This requires that a particular linear or non-linear combination of the load effects is specified. Such a combination will typically be based on a particular mechanical limit state.*

*The motivation for the present work is to highlight and extend the methodology related to load combination rules to be applied for engineering design of marine structures which are subjected to stochastic environmental processes with multiple components.*

*Development of the relevant tools based on application of "translation processes" (which are transformed Gaussian processes) are outlined. Examples are given which represent cases with both two and three simultaneous loading components. For both examples, uniform as well as non-uniform time intervals for the process components are considered.*

## 1 INTRODUCTION

In the present paper, general methods for identification of load combinations in relation to continuous stochastic processes are first highlighted. The particular case of multiple FBC-processes with known probability amplitude distribution is subsequently addressed. (The FBC-process is named with reference to the work by FerryBorges&Castanheta [1]). The case where all process components have identical basic time intervals is first considered. The relationship between the FBC-process and the so-called environmental contour methods which are presently applied for various design purposes is elaborated.

FBC-processes with widely different basic time intervals are next investigated. A methodology is outlined which enables to establish the environmental design contour also for this case.

Relationships between environmental parameters and structural load effects are frequently avail-able once the particular characteristics of a specific structure to be installed are defined. This is most straightforward for the static type of response, while dynamic response (e.g. due to stochastic loading) generally requires significantly more effort. In some cases, simplified dynamic response analyses (e.g. based on regular wave excitation models) can further be performed in order to identify gross features of the dynamic response.

The present paper illustrates how a FORM search can be applied along the limit state surface in order to identify the relevant "load combination point" for such cases. This requires that a particular type of linear or non-linear combination of the load effects is also specified. This combination will typically be based on a particular mechanical limit state function which is relevant for the whole structure or one of its components.

The motivation for the present work is to highlight and further extend the methodology behind load combination rules which are to be applied for engineering design of structures which are subjected to stochastic environmental processes with multiple components. Examples of application to the combination of wind, wave and current loading are considered. Such load combinations are relevant e.g. for long-span marine bridges which are planned for crossing of the widest Norwegian fjords as part of the so-called "Ferry-free E39 Project". One of the existing bridges with floating pontoons, i.e. the Norhordaland Bridge, is shown in Figure 1. An example of a future bridge concept, i.e. the submerged floating tunnel is shown in Figure 2. For the latter, first- and second-order wave loads as well as current loading (including the possibility of Vortex-Induced-Vibrations) are the most relevant ones.



Fig. 1. The Norhordaland Bridge with floating pontoons (located at the West Coast of Norway).

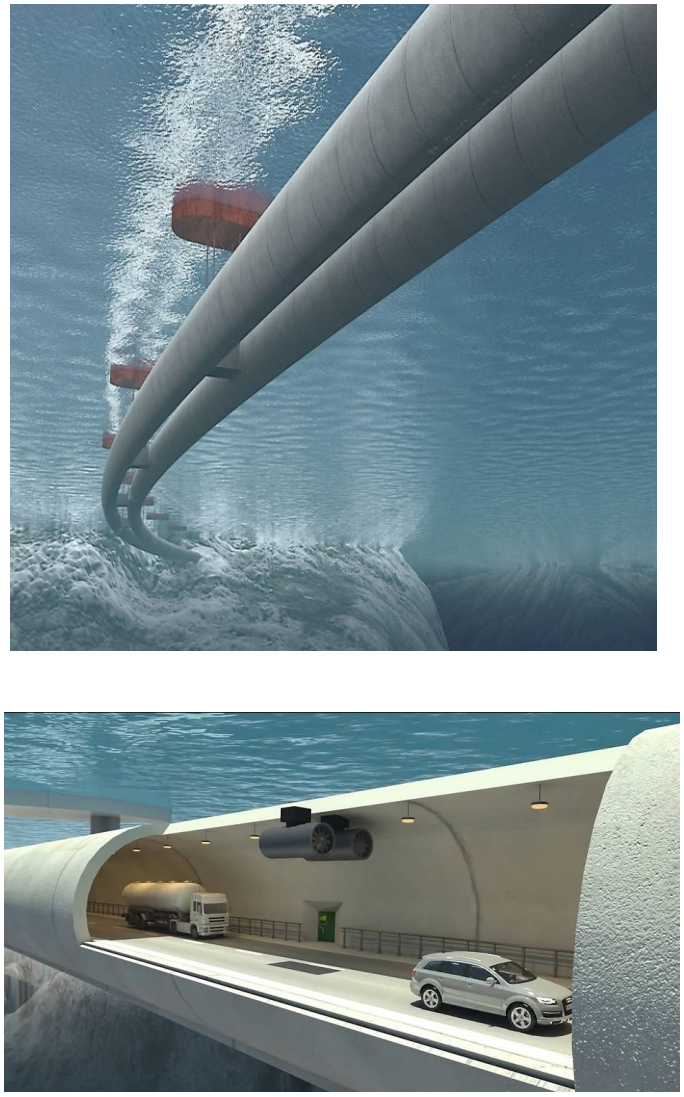


Fig. 2 Example of submerged tunnel concept (Source: Norwegian Public Road Authorities).

## 2 LOAD COMINATIONS FOR CONTINUOUS PROCESSES

A distinction should be made between combination of loads versus combination of load effects. Clearly, the combination of external loads with given magnitudes will in general imply different relative magnitudes between the associated load effects. In codified design, combination of different types of loading are typically specified in terms of return periods for the different environmental processes. As an example, for offshore structures the dominant load component is specified to have a return period of 100 years, while the secondary component is frequently specified to have a return period of 10 years (when the process components are assumed to be uncorrelated).

A further distinction should be made between situations where the relationship between the load-effects and environmental parameters are known and cases where these relationships have not yet been obtained. Even if the load effects are known, a further differentiation can be made between whether the capacity surface (mechanical limit state function) which corresponds to failure of a given structural component is available or not.

If a limit state function is specified, a load-effect combination needs to be analysed which frequently involve dynamic effects. In the present paper, focus is on continuous–time processes. For cases where the process components are also continuous-valued and in addition the limit state function is known, the combined load effect is frequently analyzed by means of the so-called up-crossing rate (or more generally the out-crossing rate for multidimensional formulations). For linear combinations, upper bound expressions can be derived which involve the up-crossing rate for each of the component processes, see e.g. Madsen et. al. [2] and Melchers [3].

Simplified methods for definition of relevant "point values" which are assumed to cover the most critical load combinations have also been introduced. One of these is the celebrated Turkstra rule which selects the expected extreme value of one component, which is then combined with the expected instantaneous values of the others, see Turkstra [4]. A sequence of such combinations (which is equal to the number of components) is then required.

A second type of simplified method is the so-called "Square-root-of-sum-of-squares" rule (SRSS-rule). The expected extreme values for all the components are then squared and added together, and the square-root of the resulting sum is then computed.

The simplified load combination methods do not explicitly take into account the particular distribu-tion functions which apply for the involved components (except for computation of the associated expected values).

For process components which are discrete instead of continuous-valued the up-crossing rate can still be applied. However, for this case the analysis can be made somewhat simpler by utilization of the step-wise behavior of the sample functions. This is achieved by means of the FBC-process representation which was mentioned above. The particular type of distribution functions and the characteristic time interval for each process component can then be taken into account in a proper way.

In general the time interval will be different for the different process components. In some cases the lengths of the time scales are widely different as for example in connection with the joint representation of wind and snow parameters.

## 3 CONTOUR METHODS FOR PROCESS COMPONENTS WITH IDENTICAL BASIC TIME INTERVALS

### 3.1 General

For cases where the limit state function is not specified, a range of environmental conditions that are relevant for analysis can still be identified. This is based on consideration of the multi-dimensional joint probability density and distribution functions which define the long-term statistical properties of the process components. Iso-probability surfaces can then be computed which correspond to a specified exceedance probability (or equivalently a specified return period) for the components.

In the following, a brief review of the much applied contour methods for identification of such relevant design events is first given. The connection with the FBC process is also highlighted. Subsequently, load effect combinations for cases with known limit state functions are considered. Identification of the associated "load combination point" for such cases is addressed.

Having obtained the design contour and the associated critical point, calculation of so-called "long-term" load effects can be avoided. Such calculations would require that the load effect is evaluated for each combination of the environmental parameters. An integration (or more typically a weighted summation) across the entire variation range for each parameter is

then required. This can be quite laborious, especially for a large number of environmental parameters.

### 3.2 Design contours

Environmental processes such as wind and wave characteristics (e.g. mean wind velocity and significant wave height) are generally of a non-stationary character. A simplified representation is typically applied where these processes are modeled according to the step-wise representation as mentioned above. The "step-levels" of the basic components are generally non-Gaussian distributed. However, they can still be represented as being transformations of processes which have Gaussian distributed step levels. Such transformed processes are frequently referred to as "translation processes".

The transformation between these basic processes and the auxiliary normalized Gaussian processes is in that case provided by the Rosenblatt transformation, see e.g. Madsen et. al. [2], Melchers [3]. For two process components this transformation is expressed as:

$$\begin{aligned} \Phi\left(u_1(t)\right) &= F_{x_1}\left(x_1(t)\right) \\ \Phi\left(u_2(t)\right) &= F_{x_2|x_1}\left(x_2(t)\,|\,x_1(t)\right) \end{aligned} \tag{1}$$

where the second line involves the conditional distribution function of $x_2$ given $x_1$.

For the case of uncorrelated basic components only the diagonal terms will be non-zero, and the elements of the Jacobian matrix simplify into the following expressions:

$$\partial u_i/\partial x_i = J_{ij}(x_i) = f_i(x_i)\,/\varphi(u_i(x_i)) \tag{2}$$

In the case of correlated basic components more complex expression apply although they can in principle be evaluated in a straightforward way.

Other possible types of transformations also exist, somewhat depending on the type of statistical information which is available. As an example, the Nataf transformation can be applied if only the marginal distributions and the pairwise correlation coefficients are known, see Nataf [5], Der Kiureghian and Liu [6].

Having performed the transformation into normalized components, the corresponding cumulative distribution for the distance from the origin to a specific point will be independent of the direction in the transformed space. This is due to the isotropic properties of the transformed processes. Hence, only the length of the radius vector will be of significance. This implies that the iso-probability levels correspond to concentric circles. The probability of exceeding a given value of the radius (R) in any direction is hence given by the following expression

$$p_f(R) = 1-\Phi(R) = \Phi(-R) \tag{3}$$

This probability of exceedance can also be interpreted in terms of a specific return period in the following manner: Designating the number of events (i.e. number of repetitions of the basic time interval) which corresponds to the given return period by N, the probability of exceeding the corresponding radius value is expressed as:

$$p_f(R) = \Phi(-R) = 1 - (1/N) \tag{4}$$

Examples of 2D and 3D contours corresponding to given return periods are given in Sections 5 and 6 of the present paper.

### 3.3 Load effects and identification of "Load Combination Point"

For each point along the design contour (in the basic parameter space) the corresponding load effect can be computed once a sufficient number of structural properties are given. Static load effects can (at least in principle) be expressed directly as functions of the environmental parameters. For stochastic dynamic load effects such relationships can usually only be established for the parameters of the probability distributions of the response processes. However, by specifying a given fractile of the response distribution, functional relationships which are similar to the static case can be obtained.

Introducing a limit state function for a specific structural member, a linear or non-linear combination of the load effects will result. The associated limit state surface can also be transformed into the space of the normalized Gaussian processes. The most critical combination of the associated load parameters will then correspond to the point on the limit state surface which has the minimum distance to the origin. This point can e.g. be identified based on FORM/SORM algorithms, see e.g. Madsen et. al. [2], Melchers [3].

Having obtained the point on the limit state surface which is closest to the origin, a scaling can be performed in the direction of the origin. This scaling is performed such that the resulting new point is located on the environmental contour surface (i.e. the contour which corresponds to the given return period). This scaled point then represents the relevant "load combination point".

Comparison of this load combination point can then be made with the points which correspond to other procedures such as the Turkstra and SRSS rules which do not utilize the distribution functions explicitly.

## 4 ENVIRONMENTAL CONTOURS AND LOAD COMBINATIONS FOR PROCESSES WITH NON-UNIFORM BASIC TIME INTERVALS

### 4.1 General

Two main options for analysis of the case with non-uniform time intervals for the basic components are considered in the following:

Option (i) Redefinition of the cumulative distribution functions for all the components except the one with the longest time interval. Introducing the notation $n_i = (\Delta T / \Delta T_i)$ for the ratio between the longest time interval and the interval length for component number i, the modified distribution function then reads:

$$F_{X_{i,\Delta T}}(x) = \left\{F_{X_i}(x_i)\right\}^{n_i} \tag{5}$$

where $F_{X_i}(x_i)$ is the cumulative distribution function for the short time interval, while $F_{X_{i,\Delta T}}(x)$ is the corresponding cumulative distribution corresponding to the longest time interval (which now becomes the common interval for all the components). The transformation into the normalized components can be performed in the same way as before.

Option (ii) Direct transformation into normalized components and subsequently accounting for the differences in reference time intervals. This procedure is based on re-scaling of the components with the shortest time intervals. Some further details of the direct transformation approach are outlined below.

### 4.2 Direct transformation

Also for the case on non-uniform time intervals, the same type of transformation as for FBC processes with identical basic time intervals can clearly be applied for each component. The "radius value" of normalized component number *i* which corresponds to the given return period is denoted by $R_i$. This quantity is obtained by solving Equation (4) when inserting the number of load interval repetitions which corresponds to that particular component.

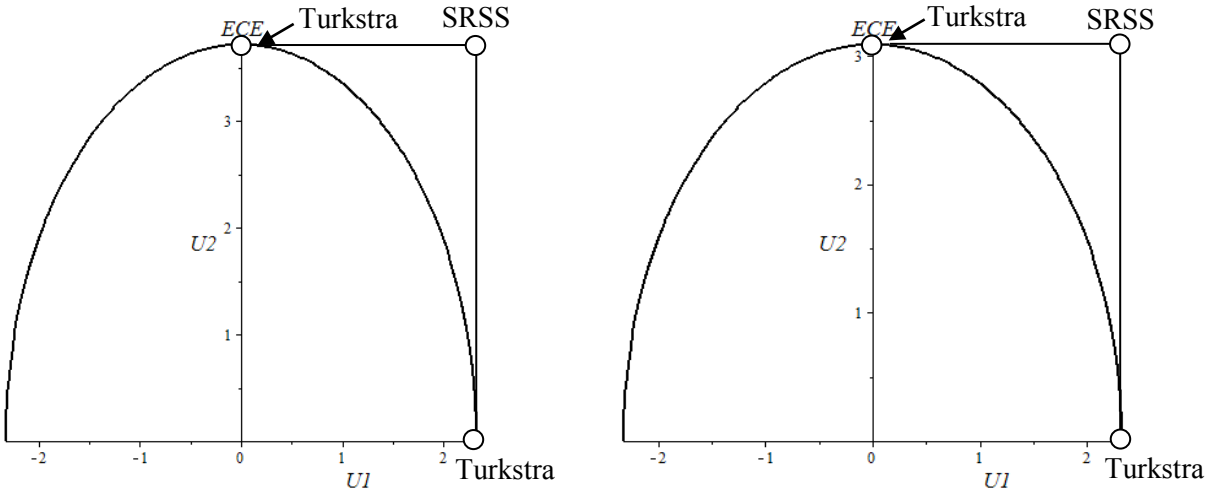
To simplify the description, the two-dimensional case is considered as an example. Denoting the direction angle in the normalized plane by $\theta$, the two components are now expressed by decomposing the respective component values to the $U_1$ and $U_2$ axis. This gives on component form:

$$u_1(\theta) = R_1 \cos(\theta)$$
$$u_2(\theta) = R_2 \sin(\theta) \tag{6}$$

This can also be expressed in terms of the equation for a corresponding ellipse as

$$\left(\frac{u_1}{R_1}\right)^2 + \left(\frac{u_2}{R_2}\right)^2 = 1 \tag{7}$$

Examples of such an extreme contour ellipse (ECE) in the "normalized" plane with $N_1$=100 and $N_2$=10000 is shown in Figure 1(a). The corresponding result for $N_1$=100 and $N_2$=1000 is shown in Figure 1(b). As observed, there is a marked downwards shift of the maximum point along the vertical axis for case (b) as compared to case (a). The points which correspond to the Turkstra and SRSS rules are also shown in the figure.




(a) $N_1$=100, $N_2$=10000 (b) $N_1$=100, $N_2$=1000

Figure 1. Comparison of contours in the "normalized" plane.

### 4.3 Search for load combination point

For option (1) the search for the "load combination point" is performed in the same manner as for the case with identical time intervals. For option (2), a re-scaling of the component axes corresponding to all but the longest basic time interval is performed in the normalized space. The scaling is expressed by:

$$u_{i,\mathrm{mod}} = u_i \cdot \left( \frac{R}{R_i} \right) \tag{8}$$

where R is the "radius value" which corresponds to the given return period for the component with the longest time interval.

Having performed such a scaling, the search for the design point proceeds in the same way as for the first option.

## 5 EXAMPLE OF A TWO-DIMENSIONAL LOAD COMBINATION

### 5.1 General Description

"Environmental parameter processes" which correspond to mean wind and significant wave height are first considered. Both the mean wind velocity and the significant wave height are assumed to be characterized by their respective Weibull distributions. Both processes are first assumed to have the same basic time interval which is taken to be 3 hours. For a return period of 100 years, the values of $R_1$ and $R_2$ will then both have values of 4.5. The effect of applying different basic time intervals for the two components is subsequently investigated.

The two processes are presently assumed to be uncorrelated. Furthermore, *normalized processes are applied* for the two environmental parameters, which means that both of them have scale factors which are equal to unity. This implies that in order to obtain the physical magnitudes of the environmental parameters, they need to be multiplied by the respective Weibull scale parameters for each specific case.

The shape parameter of the distribution for the mean wind velocity is set to $\eta_u = 2.2$, while for the significant wave height a value of $\eta_w = 1.6$ is applied. This means that the cumulative distribution function for the process "step levels" in both cases is given by

$$F_X(x) = 1 - \exp\left[-(x)^{\eta}\right] \tag{9}$$

where $\eta$ is the particular shape parameter that applies for each component, i.e. $\eta = \eta_u = 2.2$ or $\eta = \eta_w = 1.6$, respectively. As mentioned, the variable x represents the physical quantity divided by the corresponding Weibull scale parameter.

For each of the independent load components which are Weibull distributed, the transformation into standard Gaussian variables is then expressed as follows:

$$u_i(x) = \Phi^{-1}\left\{1 - \exp\left[-(x)^{\eta}\right]\right\} \tag{10}$$

where $\Phi^{-1}\{\ \}$ is the inverse of the standard Gaussian distribution function.

Subsequently, the combined static load effect due to these two environmental processes is

considered and a limiting capacity value for this combined load effect is introduced. The relevant "load combination point" identified in the response plane by means of the FORM algorithm is described in the next section.

### 5.2 Contour and design point for identical basic time intervals

The two-dimensional contour which corresponds to the (dimensionless) wind-wave environmental processes with a return period of 100 years (i.e. N = 292 000) is shown in both Figure 2 and Figure 3 below. In Figure 3, the contour corresponding to non-uniform time scales is also included for the purpose of comparison.

Application of the present contour in connection with identification of the load combination "point" based on a particular limit state function is next considered. The static response of the structure due to the mean wind, $r_{s,u}$ is assumed to be given by an expression of the following type:

$$r_{s,u} = C_u \cdot U^2 \tag{11}$$

where $C_u$ is a coefficient which depends on the geometry of the part of the structure which is subjected to the wind, in addition to the stiffness and material properties of the structure itself. U is the *normalized wind velocity* with a probability distribution function of the type given in Equation (10) with a shape factor of 2.2. In the present example the coefficient $C_u$ has a value of (1/(3.6*3.6)).

Similarly, the static response of the structure due to the (second order) action of the waves is represented by the following expression:

$$r_{s,w} = C_w \cdot W^2 \tag{12}$$

where the proportionality factor for this load effect is equal to $C_w$=(1/(5.5*5.5)). The design limit now corresponds to the normalized stress being equal to 1.0 which gives

$$r_{s,u} + r_{s,w} = 1 \quad \Rightarrow C_u U^2 + C_w \cdot W^2 = 1 \tag{13}$$

or

$$g(U,W) = 1 - \left(C_u U^2 + C_w \cdot W^2\right) \tag{14}$$

where g(U,W) designates the limit state function. The corresponding limit state corresponds to this function being equal to zero. This constitutes an ellipse in the (non-dimensional) wind-wave-parameter plane.

The associated 100-year contour and the limit state surface are shown in Figure 2. The point with the minimum distance to the origin in the normalized plane is identified by the FORM algorithm which was referred to above. The point obtained by scaling this point to the design contour represents the "load combination point".

In the standard Gaussian plane, the point on the failure surface has coordinates (4.5, 1.5) which corresponds to (4.9, 1.6) in the non-dimensional wave-wind plane. The corresponding point on the contour has the following coordinates: (4.6,1.5). The physical values will depend on the Weibull scale parameters. As an example, consider a case for which the scale parameter for the wave height is 5 m and 10 m/s for the wind velocity which gives combination point coordinates of (23 m,15 m/s).

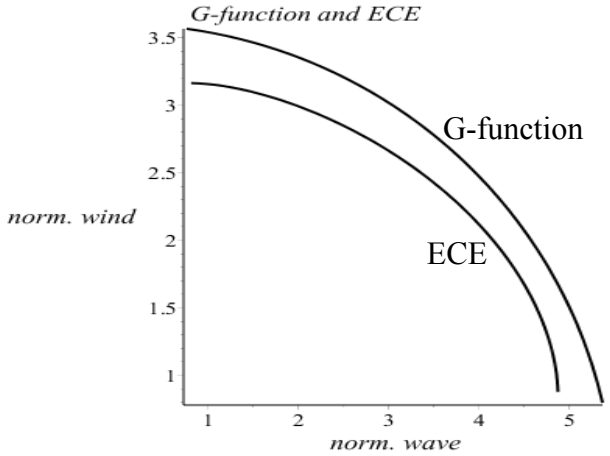



Figure 2. Contour and limit state surface for the case with uniform time intervals for both wind and wave components (N = 292 000).

In the standard Gaussian plane, the load "combination point" has coordinates (4.3,1.4). By dividing these coordinates with the 100-year return value (i.e. $R_1$=$R_2$=R=4.5) for both components, we obtain the following ratios (4.3/4.5,1.5/4.5) = (0.95, 0.3). The corresponding return periods for these down-scaled environmental parameters can then also be determined.

### 5.3 Contour and design point for non-uniform time intervals

The time interval for the wind process is next taken to correspond to 10 minutes, which is 1/18 of the time interval for the wave process (which is 3 hours).

For the wind process, the transition between different time intervals can be performed by application of proper conversion factors for the mean wind velocities (corresponding to different averaging times). Particular expressions for such a conversion are summarized e.g. in Ghiocel and Lungu [7]. As an example, the conversion factor from the 10-minute average value to the 1-hour average value is (1/1.20) = 0.83 for city areas, while it is (1/1.05) = 0.95 for the seacoast. The corresponding values for conversion from 10-minutes to 3 hours are respectively (1/1.35) = 0.74 and (1/1.07) = 0.93 for city areas and the seacoast. The latter value (i.e. (1/1.07) for the seacoast) is applied in the present study.

For the wind velocity there will hence be a two-fold effect related to application of shorter basic time intervals: (i) The probability distribution for the average wind velocity is shifted to higher values and (ii) The number of basic time intervals (i.e. number of repetitions) is significantly increased which also serves to shift the probability distribution upwards.

Clearly, the assumption of independence between the 18 repeated 10-minute average sequences is in general highly questionable. The presence of correlation would imply that instead of 18 independent repetitions a reduced "equivalent" number could be applied.

The contours which correspond to a 3hr versus 10 minute time interval for the wind process are compared in Figure 3. There is a very strong difference between the contour shapes for the two cases.

The design point in the non-dimensional wave wind plane now has coordinates (2.2, 3.3). The corresponding point in the transformed Gaussian plane has coordinates (1.9,3.7). Scaling this to the environmental contour, the load combination point in the wave-wind plane has coordinates (2.2,3.2) and (1.9,3.5) in the Gaussian plane. By dividing these values with the corresponding 100-year return values we obtain (1.9/4.5, 3.5/5.08) = (0.4, 0.7) for the present case. This implies that the relative influence of the wind load has increased significantly as compared to the uniform case.

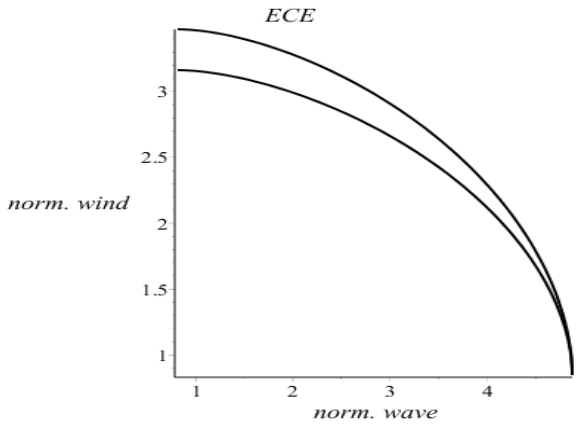



Figure 3. Comparison of contours for the case with (a) uniform time intervals ($N_1 = N_2 = 292\,000$) (b) non-uniform time intervals for the wave-wind process. ($N_1 = 292\,000$, $N_2 = 5\,256\,000$).

## 6 EXAMPLE OF A THREE-DIMENSIONAL LOAD COMBINATION

### 6.1 General

An example with three different environmental components is next considered. The two first components are the same as in the previous example, while the third component corresponds to the water current velocity.

Until now it has been common practice to use averaging periods of 10 minutes and longer (e.g. 30 minutes) when recording the current flow directly. However, much shorter intervals have also been considered as relevant, Yttervik [8].

It seems that studies on conversion factors between velocities for different averaging times are not available and will probably be very site dependent. In the present analysis a basic time interval length of 10 minutes is applied.

The cumulative probability distribution of the (dimensionless) mean current velocity is also assumed to be given by a Weibull distribution. The current velocity is normalized such that the scale factor is equal to unity. The corresponding shape factor is $\eta_c = 1.2$.

The static load effect due to the current is given by an expression which is similar to those for the static wind and wave loads:

$$r_{s,c} = C_c \cdot C^2 \tag{15}$$

The values of the three constants for the three-dimensional case are now set equal to $C_u$ = 1/(10.*10.), $C_w$ = 1/(5.5*5.5) and $C_c$ = (1/(6*6)). The total static load effect is expressed as the sum of all the three contributions, and the resulting limit state function then becomes:

$$g(U, W, C) = 1 - \left(C_u U^2 + C_w \cdot W^2 + C_c \cdot C^2\right) \tag{16}$$

which represents the surface of an ellipsoid in the three-dimensional wave-wind-current space (when the limit state function is equal to zero).

### 6.2 Contour and design point for identical basic time intervals

The contour surface which corresponds to non-dimensional wave, wind and current values is shown in Figure 4 for the case with $N_1=N_2=N_3=N=$ 292 000 (which corresponds to a return period of 100 years).

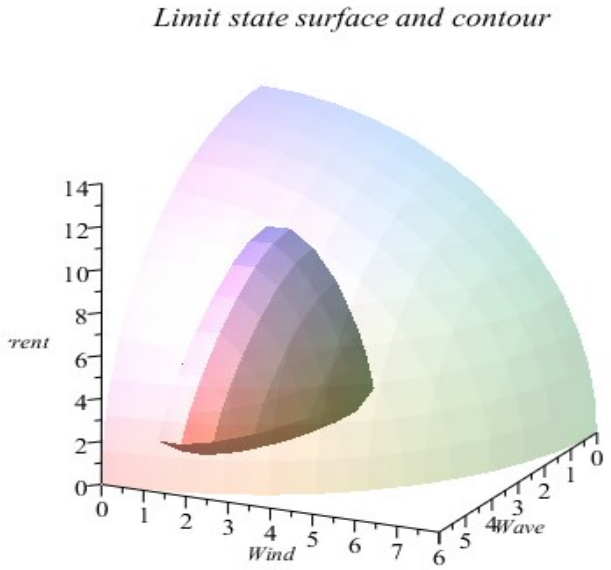



Figure 4. Contour and failure surface in the non-dimensional "wave-wind-current" space for the case with identical time intervals ($N_1=N_2=N_3=N=$292 000, corresponding to a return period of 100 years).

The point on the failure surface which is identified to be closest to the origin based on application of the FORM algorithm is found to have coordinates (4.0, 1.0, 4.5) in the normalized wave-wind-current space. The physical values can subsequently be obtained by multiplying with the respective scale parameters. It is seen that the wave and the current are the dominant load parameters for the present combination.

### 6.3 Contour and design point for identical basic time intervals

As the next step, a different FBC-model is applied where the basic time interval for both the average wind and current velocity are 10 minutes, while that for the significant wave height is 3 hours. The contour surface for this case is shown in Figure 5. It is observed that it is widely different from the contour surface in Figure 4.

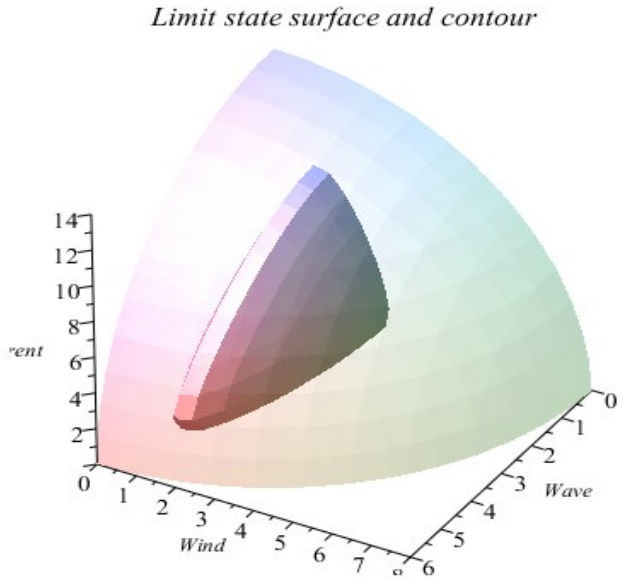



Figure 5. Contour and failure surface in the non-dimensional wind-wave-current space with non-uniform basic time intervals. ($N_1$= 292 000, $N_2$= 5 256 000, $N_3$=5 256 000)

The coordinates of the "load combination point" in the normalized wave-wind-current space are now computed as (2.5, 3.4, 5.7). This implies that the current is the dominating load parameter also for this case, while the wind now is the second most parameter as opposed to the wave for the previous case.

## 7 CONCLUDING REMARKS

The role of contours in relation to calculation of design load effects and the associated proper load combinations was highlighted. Application examples were given, both for the case with identical basic time intervals and for the case with non-uniform intervals. A non-linear combination of load effects was applied to illustrate the implications of the analysis procedure.

For the mean wind velocity (and also for the mean current velocity) the issue of averaging period seems to have a strong influence on the resulting contour shape. Accordingly, relevant conversion formulas between different averaging periods for the environmental processes that are involved should be readily available for user of design codes in order to achieve a transparent formulation. The assumption of independence between values for reduced averaging periods also needs further clarification.

## 8 REFERENCES

[1] Ferry Borges, J. and Castanheta, M. (1971): *Structural Safety*, 2nd. Ed., National Civil Engineering Laboratory, Lisbon, Portugal.

[2] Madsen, H; Krenk, S. and Lind, N.C. (1986): *Methods of Structural Safety*. Prentice-Hall, Englewood Cliffs, Ney Jersey.

[3] Melchers, R.E. (1999): *Structural Reliability Analysis and Prediction*, John Wiley & Sons, Chichester, England.

[4] Turkstra, CJ (1970). ”*Theory of Structural Design Decisions*”, Study No 2, Solid Mechanics Division, Universi-ty of Waterloo, Waterloo, Canada.

[5] Nataf, A. (1962): ”Determination des distributions dont les marges sont donnees”, *Comptes Rendus de l’Academie des Sciences*, Paris, 225,42-43.

[6] Der Kiureghian, A. and Liu,P.L. (1986):”Structural Reliability Under Incomplete Probability Information”, ASCE, *Journal of Engineering Mechanics*, Vol. 112, No 1, pp 85-104.

[7] Ghiocel, D. and Lungu, D. (1975): *Wind, Snow and Temperature Effects on Structures Based on Probability*, Abacus Press, Kent, England.

[8] Yttervik, R. (2004): *Ocean current variability in relation to offshore engineering*. Phd thesis, Dept. Marine Technology, NTNU, Trondheim, Norway.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ADJOINT-BASED ROBUST OPTIMIZATION USING POLYNOMIAL CHAOS EXPANSIONS

**Joao Miranda, Dinesh Kumar, and Chris Lacor**

Vrije Universiteit Brussel (VUB), Department of Mechanical Engineering, Fluid Mechanics and Thermodynamics Research Group
Pleinlaan 2, 1050 Brussels, Belgium
e-mail: Joao.Miranda@vub.ac.be, Dinesh.Kumar@vub.ac.be, Chris.Lacor@vub.ac.be

**Keywords:**

**Abstract.** *Adjoint methods are nowadays widely used to efficiently perform optimization for problems with a large number of design variables. However, in reality, the problem at hand might be subjected to uncertainties in the operational conditions or, in case of optimizing geometries, the design variables itself might be uncertain due to manufacturing tolerances. For such applications, the optimum obtained using deterministic methods might be very sensitive to small variations in the uncertainties, i.e. it lacks robustness. In a robust optimization, the uncertainties are taken directly into account during the optimization process by introducing, next to the mean objective, its variance as a second objective. This implies that, when using gradient based optimization methods, the gradients of both objectives (mean and variance) must be known. In this work the Polynomial Chaos Expansion (PCE) is used in combination with adjoint methods to efficiently obtain both gradients. A non intrusive, regression based PCE is used, requiring a new adjoint solution for each sampling point in order to build the PCE of the gradient. A PCE for the objective is also built (at no extra cost) in order to compute the gradient of the variance.*

*A weighted average of both gradients is then used to find an optimum. By changing the weighting factors the solution can be found favouring either of the two objectives. The developed approach is applied to relevant engineering problems, such as geometrical optimization of pipe flows and flow over airfoils. The design variables are the shape coordinates and no parameterization is used. In this work, the design variables were considered deterministic with the uncertainties coming from operational conditions.*

## 1 INTRODUCTION

With the increase in available computational power, CFD has evolved from an analysis tool towards a design tool, see [1]. In the design process, CFD simulations are used in order to find the value of the objective(s) for a given set of design variables. In order to find the best design configuration, an optimization algorithm is then used to drive the design variables to their optimal value, minimizing the objective(s), or, in case of conflicting objectives, leading to a Pareto front.

The optimization techniques can be classified as gradient free or gradient based. Gradient based approaches make use of the gradient (with respect to the design variables) of the objective function, which provides the optimal search direction in the design space, see [2] and [3]. The gradient free methods, essentially sample the design space and try to find the optimum solely based on the value of the objective(s), see [4].

Depending on the parameterization, optimization problems often have a large number of design variables, see [5] and [6], especially in the case of shape optimization. This makes traditional techniques for gradient calculation, such as finite differences, which require an extra flow simulation for every design variable, too expensive. An adjoint method offers a cheap alternative as it allows to calculate the gradient, at roughly the cost of an extra CFD simulation, independently of the number of design variables, see [7]. Adjoint solvers can be classified as discrete or continuous, depending on how the adjoint equations are derived, see [8] and [9]. Both approaches have their positive and negative aspects and the choice of one or the other method is usually related to a personal preference.

While advances in computational capabilities and numerical techniques lead to a widespread use of optimization algorithms to design engineering components, these are mainly employed on deterministic problems. This means that all the design conditions are well defined, immutable and are assumed to be an exact representation of the reality. However, in real world applications this is seldom the case as the operational conditions may significantly differ from the design conditions. Moreover, due to manufacturing tolerances, the finished component may be slightly different than the optimal design. Due to the high number of uncertainties in the manufacture and operating life of any engineering component, a deterministic optimization approach may result in an optimum that is very sensitive to all those small uncertainties, leading to a lower performance than desired. Consequently, an optimization methodology that takes into account the robustness of the design is of special interest for real world engineering applications.

In a robust optimization setting the objective is to reduce both the value of the mean objective and its sensitivity to operating and manufacture uncertainties. To this end, when using a gradient based optimization method, the gradients of both mean objective and of its variance must be known. These gradients are then combined, using weighted averaging, to find the best solution for each problem. In the case of conflicting objectives, e.g. the design corresponding to the minimum objective is not robust (has high variance), a Pareto front, or part of it, can be built by varying the weights given to each gradient. In such a case is the design engineer responsibility to chose a suitable compromise.

The success of the robust optimization algorithms is intrinsically connected to the performance of the chosen uncertainty quantification (UQ) method. In intrusive PC approaches, the PC expansion is directly embedded in the code leading to a less expensive methodology than the non-intrusive approaches. However, their implementation is cumbersome and prone to errors and involves deep knowledge of the CFD code. An example of an intrusive approach can be

found in [10]. On the other hand, non-intrusive methods such as Monte-Carlo and Non-Intrusive Polynomial Chaos (NIPC) have the need for an higher number of function evaluations, but the implementation is straightforward and independent of the problem being studied, see [11]. However, the number of required samples will increase exponentially with the number of uncertainties and the order of the polynomials. This is known as the curse of dimensionality.

In order to calculate the gradient of the mean objective and of its variance, PCEs of the objective and of its gradient are required. To this end, both the primal and adjoint solutions are needed in each sample point. As the samples are independent of each other they can be evaluated in parallel reducing the overall simulation time.

In order to compute the gradient of the variance, a PCE is also built for the objective. This is done at no extra cost since the objective samples are calculated form the primal solution at each sample.

In this work the uncertainties are restricted to finite physical quantities. A uniform distribution for the stochastic variables is chosen. Consequently the polynomials, used for building the PCE expansion, are the Legendre polynomials, see [12]. In order to find the polynomial coefficients an overdetermined system is solved using regression.

## 2 GRADIENT BASED ROBUST OPTIMIZATION

A general minimization problem can be written as:

$$\begin{aligned} &\text{minimize} \quad J(U,\alpha) \\ &\text{subject } to \quad R(U,\alpha) = 0 \end{aligned} \tag{1}$$

where $J$ is a cost function to be minimized, $R$ is the set of governing equations, $U$ the state variables of the primal problem and $\alpha$ the design variables. In a gradient based optimization, the gradient of the cost function with respect to the design variables, i.e. $\frac{dJ}{d\alpha}$, has to be calculated.

$$\frac{dJ}{d\alpha} = \frac{\partial J}{\partial \alpha} + \frac{\partial J}{\partial U}\frac{\partial U}{\partial \alpha} \tag{2}$$

Where the term $\partial U/\partial \alpha$ depends on the primal solution, and so it is expensive to find. Taking into account that

$$\frac{dN}{d\alpha} = \frac{\partial R}{\partial U}\frac{\partial U}{\partial \alpha} + \frac{\partial R}{\partial \alpha} = 0 \tag{3}$$

eq. 2 becomes:

$$\frac{dJ}{d\alpha} = \frac{\partial J}{\partial \alpha} - \frac{\partial J}{\partial U}\left(\frac{\partial R}{\partial U}\right)^{-1}\frac{\partial R}{\partial \alpha} \tag{4}$$

Introducing $\lambda$, the vector of the adjoint variables:

$$\left(\frac{\partial R}{\partial U}\right)^{T}\lambda = \left(\frac{\partial J}{\partial U}\right)^{T} \tag{5}$$

eq. 4 can be rewritten as:

$$G = \frac{dJ}{d\alpha} = \frac{\partial J}{\partial \alpha} - \lambda^{T}\frac{\partial R}{\partial \alpha} \tag{6}$$

Eq. 6 does not depend on the primal solution. Hence the calculation of the gradient with respect to any design variable $\alpha$ is reduced to the calculation of simple partial derivatives.

### 2.1 Stochastic Adjoint Based Gradient

The stochastic gradient $G(\xi)$ can be written as PCE of order $p$ for $n$ random non-design variables $\xi \equiv \{\xi_i\}_{i=1}^{n}$ as:

$$G(\xi) = \sum_{i=0}^{P} \hat{G}_i \psi_i(\xi) \tag{7}$$

where the total number of terms on the expansion is given by $P+1 = \frac{(p+n)!}{p!n!}$ and $G_i$ represents the i$^{\text{th}}$ random mode of $G(\xi)$.

In order to find the components of the PCE ($G_i$), a regression approach is chosen. An overdetermined system is built:

$$\begin{pmatrix} \psi_0(\xi^1) & \psi_1(\xi^1) & \cdots & \psi_P(\xi^1) \\ \psi_0(\xi^2) & \psi_1(\xi^2) & \cdots & \psi_P(\xi^2) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_0(\xi^{n_s}) & \psi_1(\xi^{n_s}) & \cdots & \psi_P(\xi^{n_s}) \end{pmatrix} \begin{pmatrix} \hat{G}_0 \\ \hat{G}_1) \\ \vdots \\ \hat{G}_P \end{pmatrix} = \begin{pmatrix} G(\xi^1) \\ G(\xi^2) \\ \vdots \\ G(\xi^{n_s}) \end{pmatrix} \tag{8}$$

where $G(\xi^1), ...G(\xi^{n_s})$ are the gradients calculated at each sample, $\hat{G}_0, ...\hat{G}_P$ are the components of the polynomial of order $p$ and $\xi^0, ...\xi^{n_s}$ are the $n_s$ samples mapped in the chosen polynomial domain.

When using regression to find the PCE components, the advised number of samples is two times the total number of terms in the PC expansion, $n_s = 2 * (P + 1)$, [13].

### 2.2 Gradient of the Variance

Although both the mean ($\hat{G}_0$) and variance ($\sigma_G$) of the gradient can directly be calculated from eq.(7), in order to do robust optimization the gradient of the variance of the objective ($\nabla\sigma_J$) is also required.

From Eq. 6 and eq.(7) it comes directly:

$$G(\xi) = \frac{dJ}{d\alpha}(\xi) = \sum_{i=0}^{P} \widehat{G_i} \psi_i(\xi) \tag{9}$$

On the other hand, a PCE of the objective functional gives:

$$J(\xi) = \sum_{i=0}^{P} \widehat{J}_i \psi_i(\xi) \tag{10}$$

Taking the gradient one can write:

$$\frac{dJ(\xi)}{d\alpha} = \sum_{i=0}^{P} \frac{d\widehat{J}_i}{d\alpha} \psi_i(\xi) \tag{11}$$

Comparison with equation 9 gives:

$$\widehat{G_i} = \frac{d\widehat{J}_i}{d\alpha} \tag{12}$$

The first term in the PCE of eq.(10), $\widehat{J_0}$, is the mean objective. According to eq.(12), the gradient of the mean objective, follows from the PCE of the gradient:

$$\hat{G}_0 = \frac{d\widehat{J_0}}{d\alpha}. \tag{13}$$

Furthermore:

$$\hat{G}_i = \frac{d\widehat{J_i}}{d\alpha} \tag{14}$$

For uniform distribution the variance of the objective can be written as:

$$\sigma_J^2 = \frac{\hat{J}_1^2}{3} + \frac{\hat{J}_2^2}{5} + \cdots \tag{15}$$

Taking the gradient:

$$\nabla\sigma_J^2 = \frac{d}{d\alpha}\left(\frac{\hat{J}_1^2}{3} + \frac{\hat{J}_2^2}{5} + ...\right) = \left(\frac{1}{3}\frac{d\hat{J}_1}{d\alpha}\hat{J}_1 + \frac{1}{5}\frac{d\hat{J}_2}{d\alpha}\hat{J}_2 + ...\right) = 2\left(\frac{1}{3}\hat{G}_1\hat{J}_1 + \frac{1}{5}\hat{G}_2\hat{J}_2 + ...\right) \tag{16}$$

As an example, for a second order polynomial one obtains:

$$\nabla\sigma_J^2 = 2\left(\frac{1}{3}\hat{G}_1\hat{J}_1 + \frac{1}{5}\hat{G}_2\hat{J}_2\right) \tag{17}$$

### 2.3 Optimization

For the optimization process, a weighted combination of the two gradients is used as:

$$G_t = \omega\hat{G}_0 + (1-\omega)\nabla\sigma_J^2 \tag{18}$$

Where, $G_t$ is the effective gradient to be used during the optimization and $\omega$ is a weighting coefficient to be chosen by the user.

$$\alpha^{n+1} = \alpha^n - \beta * \frac{dJ}{d\alpha^n} \tag{19}$$

where $\alpha$ is a vector containing the design variables and $\alpha^{n+1}$, $\alpha^n$ are respectively the updated and the original values. $\beta$ is a scalar that is chosen iteratively, i.e. it can change at each iteration, in order to guarantee a smooth optimization process.

As no parameterization is used, every mesh point on the surfaces to optimize is a design variable. This gives a lot of freedom in the shape modifications which can lead to unrealistic oscillatory shapes. An implicit smoothing algorithm is employed to make sure the optimized surface remains smooth. In order to maintain the quality of the mesh in the laminar test cases the mesh is smoothed by solving a Laplacian equation, see [14]. For the turbulent test cases, in order to maintain a proper boundary layer resolution, the problem is remeshed.

A optimization problem with deterministic design variables and using a steepest descent optimization is described in algorithm 1.

**Algorithm 1** Non-Design Stochastic Variables

Set $tol$, $U_{initial}$
**while** $\langle G_0 \rangle > tol$ **do**
    {
    Sample $\xi^1, \xi^2 ... \xi^{n_s}$
    **for** $i = 1 \rightarrow n_s$ **do**
        {
        Solve primal problem, $U(\alpha, \xi^i)$
        Compute Objective $J(\alpha, \xi^i)$
        Solve adjoint problem, $\lambda(\alpha, \xi^i)$
        Compute gradient $G(\alpha, \xi^i))$
        }
    **end for**
    Use regression to find $\hat{G}_0, ..., \hat{G}_P$ and $\hat{J}_0, ..., \hat{J}_P$
    Compute gradient of the variance $\nabla \sigma_J^2$
    Compute $G_t$ using Eq. 18
    $\alpha = \alpha - \beta G_t$
    }
**end while**
End

## 3 NUMERICAL IMPLEMENTATION

Both the primal and adjoint problems are solved using the the **SIMPLE** algorithm in OpenFOAM. The steady state solution is achieved using an Euler implicit time scheme. Bounded central schemes are used for the numerical discretization of both the primal and the adjoint problems, see [15]. In the turbulent test cases, the turbulence in the primal problem is modeled using a RANS approach. The Spalart-allmaras model is chosen and the boundary layer is modeled by means of wall functions. For the adjoint solver, a frozen turbulence approach is used.

The sampling needed for the Latin Hypercube scheme, ensuring a good spread of the samples in the stochastic space. The random number generator is connected to the time and date of the machine in order to guarantee randomness. The samples are all obtained in $[0, 1]^n$ and then mapped into the stochastic variable distribution. For a uniform distribution $U[x_a, x_b]$, the mapping is simply a projection, $[0, 1] \rightarrow [x_a, x_b]$. The PCE terms are found by solving the overdetermined problem by using a regression approach.

## 4 TEST CASES

### 4.1 2D U-bend

A common fluid mechanics problem found in engineering applications is the minimization of total pressure loss in flows through piping systems.

For an arbitrary pipe shape, the total pressure loss (the objective to minimize) can be written as:

$$J = \int_{\partial\Omega} P_t d\dot{m} \tag{20}$$

where, $P_t$ is the total pressure, $\partial\Omega$ the domain's boundary, and d$\dot{m}$ the elementary mass flux

through the boundary.

In this test case we consider the robust optimization of a U-bend under operating uncertainties. This type of bend is commonly found as part of a larger serpentine in applications such as heat exchangers. For serpentines with a high number of U-bends, the accumulated pressure drop through all the U-bends may severely impair the total efficiency of the system. As a result it is crucial to the reduce the pressure drop at each bend.

In a robust optimization setting the problem can be written as:

$$minimize : \mu(J), \sigma(J) \tag{21}$$

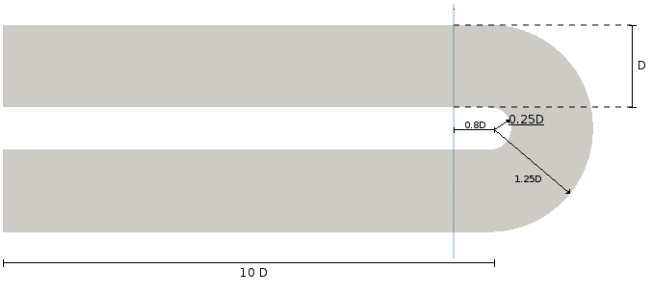



Figure 1: U-bend pipe.

The initial shape and dimensions of the U-bend are shown in Figure 1. For this test case $D = 0.075m$ and the deterministic Reynolds number is $Re = 400$. The laminar Navier-Stokes equations are therefore considered for the primal and the adjoint. In this test case the optimization is restricted to the outer wall of the U-Bend. The inlet velocity is considered as uncertain with a uniform distribution $U_{inlet} = [0.066, 0.084]m/s$. A $2^{nd}$ order polynomial chaos method is used in $J$, requiring six samples at each optimization step to get the PCEs of the objective and its gradient.

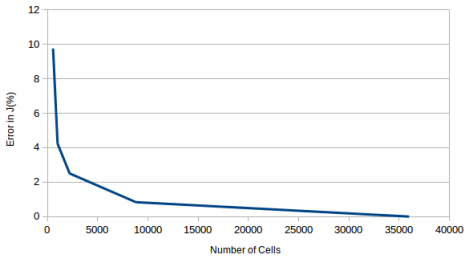



Figure 2: Mesh independence test.

Figure 2 shows the error in the calculated total pressure drop on meshes of different size. The solution on a very fine mesh of $36000$ cells is considered as a reference solution. It is observed that for a mesh with $\approx 9000$ cells, the error on the pressure drop is less than $1\%$. This mesh is therefore considered as a realistic mesh for the problem at hand, suitable for use in the optimization process.

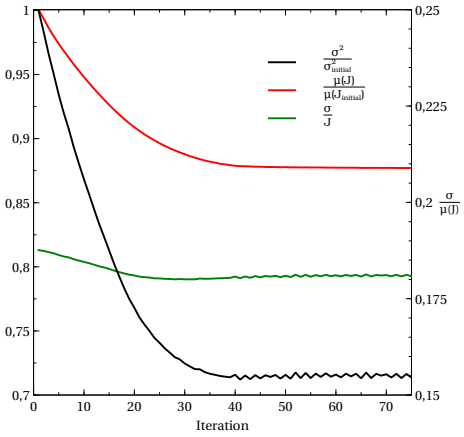



Figure 3: $\mu(J)$ and $\sigma^2$ evolution during optimization using $\omega = 1$.

In Figure 3 the evolution of the mean objective and the variance during the optimization are shown for $\omega = 1$. Both the variance and the mean objective are nondimensionalized by their values at the beginning of the optimization process ($\mu(J_{initial})$, $\sigma^2_{initial}$). It can be observed that the optimization of the mean objective also leads to a lower variance, hence the optimal solution (lowest pressure drop) is additionally a robust one. This conclusion is further supported by observing that $\frac{\sigma}{J}$ is almost constant through the optimization process.

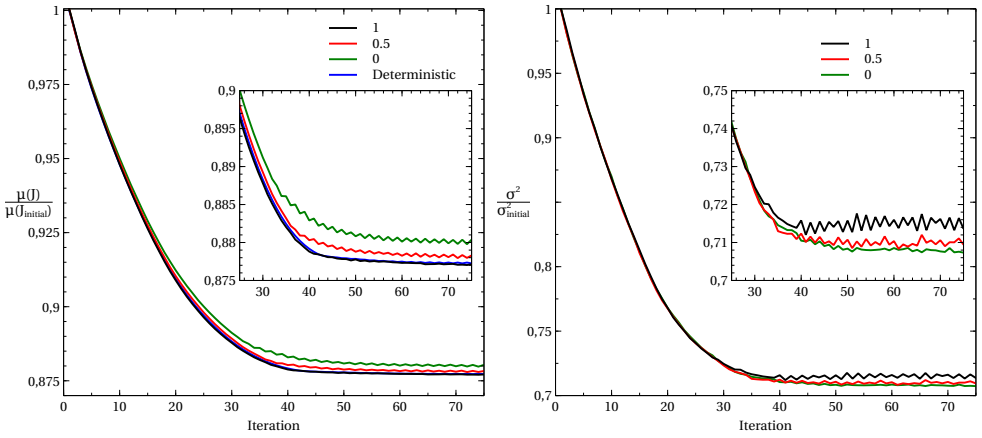



Figure 4: Deterministic and Robust optimization using $\omega = 1,\ 0.5$ and $0$.

Figure 4 shows the evolution of the variance and the mean objective for different values of $\omega$. Taking into account that the optimal is a robust one, optimizing the mean objective, $\omega = 1$, or the variance, $\omega = 0$, will lead to very similar results. However, it can be observed that there are small differences between the different cases. For $\omega = 1$, the optimal mean objective reaches its lowest value and the variance is the highest. On the other hand, for $\omega = 0$, the variance reaches its lowest value whereas the objective is at its highest. Furthermore, for $\omega = 1$ the mean objective presents a smooth optimization but the variance is oscillating, which indicates that a minimum has not been reached for the latest. The opposite happens for $\omega = 0$.

It can be observed that for an intermediate value of $\omega = 0.5$, both the mean objective and the variance take intermediate values. This shows that combining both gradients can be used to obtain a mix between optimal and robust results.

Table 1: Deterministic operational conditions.

| | |
|---|---|
| U | $102m/s$ |
| Re | $6 \times 10^6$ |
| AoA | $1^\circ$ |

Table 2: Lift and Drag coefficients for NACA 0012 at AoA = $1^\circ$ and Re $6 \times 10^6$.

| | Cl | Cd | L/D |
|---|---|---|---|
| Numerical | 0.109 | 0.0089 | 12.25 |
| Reference NASA-SA | 0.110 | 0.0085 | 12.94 |
| error [%] | 0.9 | 4.7 | 5.3 |

Finally it is important to notice that the optimization of the mean objective, $\omega = 1$, shows an almost perfect correspondence with the deterministic optimization.

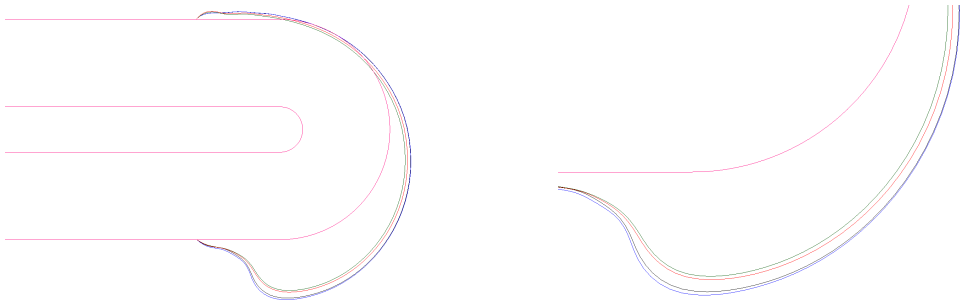


Figure 5: Initial(Pink) and optimal shapes. Deterministic(Blue), $\omega = 1$(Black), $0.5$(Red) and $0$(Green) .

Figure 5 depicts the optimal shapes obtained with the different approaches. As expected from the previous results analysis, all the obtained optimal shapes are similar. In all cases, there is an asymmetrical swelling of the U-bend more pronounced on the side where the flow exits the U-bend. Furthermore, the optimal shape obtained using $\omega = 1$ is the closest to the one obtained in the deterministic problem whereas the shape obtained using $\omega = 0$ shows the largest differences.

### 4.2 2D Airfoil

In this section the robust optimization of a subsonic airfoil with uncertainties in the operating conditions is considered. The deterministic objective is to maximize the Lift-to-Drag ratio (L/D). The robust optimization problem is the defined as:

$$minimize : \mu(-L/D), \sigma(-L/D) \tag{22}$$

The initial airfoil is a NACA 0012 and the deterministic operational conditions are given in Table 1.

The chosen turbulence model is the Spalart-Almaras with wall functions. A C-type mesh of $80000$ cells is used. The numerical results are compared to the reference results of [16] in Table 2. A good comparison is observed confirming that the size and the quality of the mesh is appropriate.

Table 3: Random variable distributions.

| Random Variable | Probability distribution |
|---|---|
| AoA ° | Uniform [0.5,1.5] |
| U $[m/s]$ | Uniform [98.84,105.2] |

The random variables for this test case are the freestream velocity and the angle of attack with uniform distributions as listed in Table 3.

The number of samples required to find the PCE coefficients dramatically increases with the order of the PCE expansion. As each sample involves a primal and an adjoint CFD solutions, the methodology becomes prohibitively costly for high order PCEs.

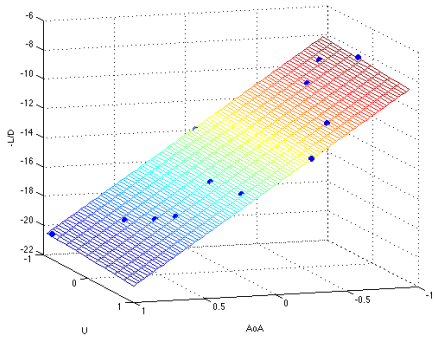



Figure 6: Response surface for $-L/D$.

Figure 6 shows the response surface for $-L/D$ in relation to the stochastic variables (mapped to [-1 1]). It can be observed that all the samples lie in a plane. This indicates that a first order PCE is sufficient to accurately capture the response.

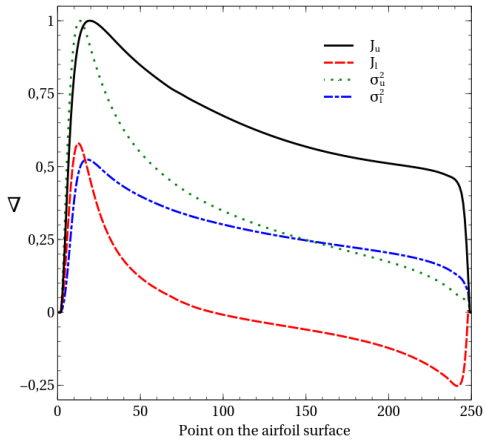


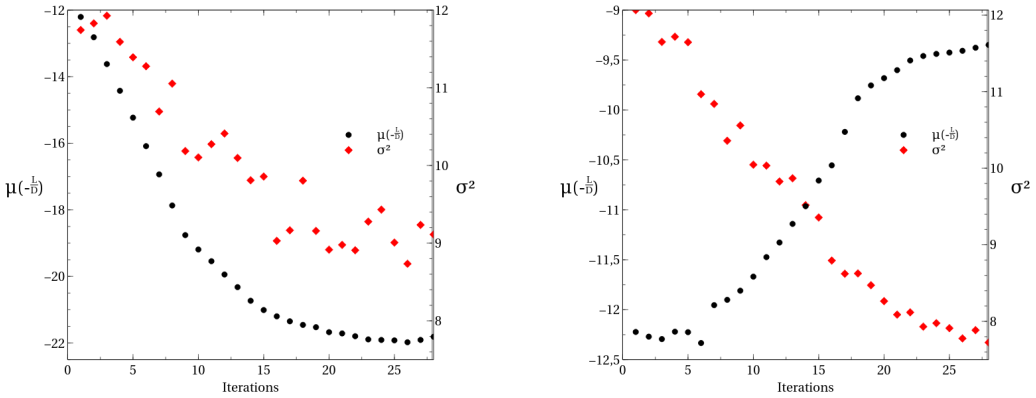

Figure 7: Sensitivities of the mean objective ($\mu(J)$) and variance ($\sigma^2$). $u$ and $l$ refer respectively to the upper and lower surfaces of the airfoil.

In Figure 7 the initial surface sensitivities for the mean objective and variance are shown. It can be noted that for the upper surface both sensitivities are positive corresponding to an outward movement of the nodes. On the lower surface, the variance sensitivities are also positive but the mean objective sensitivities change sign, meaning that towards the trailing edge the

Table 4: Optimal Cl and Cd at AoA = 1°.

| | Cl | Cd | $\mu(L/D)$ | $\sigma^2$ | $\mu(L/D)-\sigma$ | $\mu(L/D)+\sigma$ |
|---|---|---|---|---|---|---|
| Initial | 0.109 | 0.0089 | 12.25 | 12 | 8.79 | 15.71 |
| $\omega = 1$ | 0.204 | 0.0094 | 21.17 | 9.19 | 18.14 | 24.20 |
| $\omega = 0$ | 0.099 | 0.0109 | 9.34 | 7.72 | 6.59 | 12.12 |

nodes will move inwards leading to slightly cambered airfoil. It is also to be mentioned that the leading and trailing edges are fixed.




(a) $\omega = 1$ (b) $\omega = 0$

Figure 8: Evolution of $\mu(-L/D)$ and $\sigma^2$ during the optimization.

The evolution of the variance and the mean objective during the optimization using $\omega = 1$ and $\omega = 0$ are shown on Figure 8 and the final results are summarized in Table 4. For $\omega = 1$ the evolution of the $\mu(J)$ and $\sigma^2$ are similar, indicating that the airfoil corresponding to the optimal mean value is also more robust than the initial airfoil. On the other hand, when $\omega = 1$, $\mu(J)$ and $\sigma^2$ have opposite behaviors indicating that a very robust airfoil will have a poor mean performance. It can also be observed that although the variance is reduced in both cases, the lowest value is obtained when the gradient of the variance is used. Moreover, the evolution of the variance during optimization with $\omega = 0$ is much smoother than when optimizing with $\omega = 1$. Finally, it is interesting to notice that even in a pessimistic scenario, $L/D = \mu(L/D) - \sigma$ on the airfoil obtained when optimizing the mean of the objective, is still better than the $L/D$ of original airfoil or $\mu(L/D)$ of the airfoil obtained using $\omega = 0$.

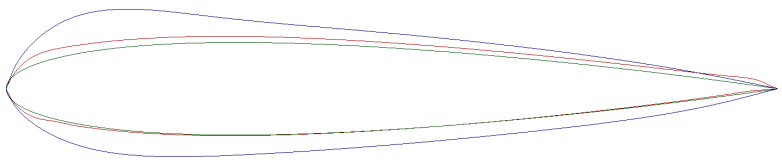
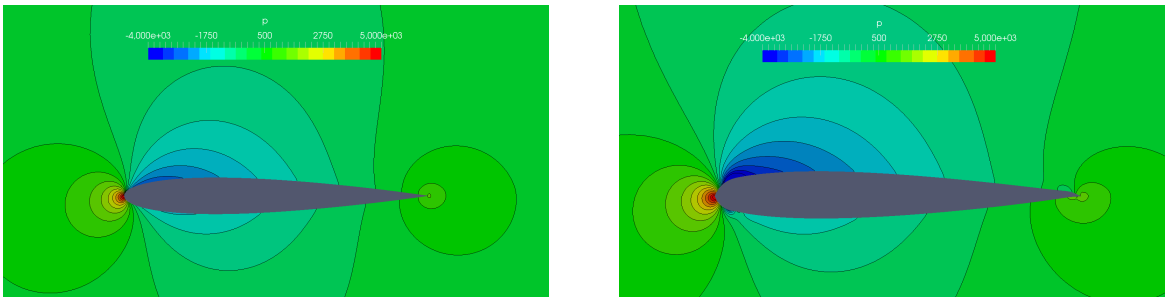


Figure 9: Initial(Green) and optimal shapes: $\omega = 1$(Red) and 0(Blue) .

The optimal airfoil shapes are shown in Figure 9. For the airfoil obtained using $\omega = 1$, the most noticeable differences can be observed on the upper surface and the leading edge.

The increase in thickness on the leading edge results in a higher acceleration of the flow over the airfoil creating a region of low pressure and increasing lift. Moreover, the upper surface experiences a much larger deformation than the lower surface, thus, the resulting airfoil will have a small curvature which additionally contributes to the increase of lift.

On the other hand, for the case with $\omega = 0$, the optimal airfoil is overall much thicker than the initial configuration. This explains the increase in drag and suggests a lower sensitivity to the angle of attack explaining the smaller variance. It is important notice that no constraints where imposed in the optimization.




(a) (b)

Figure 10: Initial (a) and final (b) pressure distribution for $\omega = 1$.

Figure 10 depicts the pressure contours for the initial and optimal airfoil using $\omega = 1$. It can be observed that the low pressure region in upper surface on the optimized airfoil is much more intense than on the initial one. This explains the large increase in the lift on the optimized airfoil when compared to the initial configuration.

As a last comment, given the relativity small change of the airfoil shape, especially if $\omega = 1$, it seems that the current solution is only a local optimum. Other shapes might be found starting from another initial airfoil. In order to retrieve the Pareto front it seems advisable to combine the gradient based approach with an evolutionary algorithm.

## 5 CONCLUSIONS

A robust gradient based shape optimization using PCE has been developed. The gradient used in the steepest descent is a weighted average of the gradient of the mean objective and the gradient of its variance. by changing the weights, solutions favoring either mean or variance can be found.

Two test cases are presented.

In the first test case, a U-bend, the optimal solution obtained using the gradient of mean objective was found to be also a very robust one. Therefore, and as expected, the optimization of the variance led to a similar optimal solution. The small differences found on the optimal solutions agreed with the chosen value for $\omega$.

In the second test case, a 2D airfoil under turbulent subsonic flow, the optimization of the mean objective and the variance led to two very distinct configurations with somewhat conflicting objectives. The observed changes in the airfoil shape are rather small, especially if a high weight was given to to optimizing the mean objective. This suggests a local optimum, and a dependency on the initial airfoil configuration. In order to retrieve a global optimum and the Pareto front, a combination with evolutionary algorithms seems indicated. This will be the subject of future research.

## 6 ACKNOWLEDGEMENTS

This research was partly funded by the Flemish Science Foundation FWO under a grant G0B7113N. This support is gratefully acknowledged.

## REFERENCES

[1] Deb, Kalyanmoy *Optimization for engineering design: Algorithms and examples*. PHI Learning Pvt. Ltd., 2012.

[2] Snyman, Jan *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms, Vol. 97*. Springer, 2005.

[3] Papadimitriou, Dimitrios and Giannakoglou, Kyriakos, Aerodynamic Shape Optimization Using First and Second Order Adjoint and Direct Approaches. *Archives of Computational Methods in Engineering, Vol. 15* 447–488, 2008.

[4] Deb, Kalyanmoy *Multi-objective optimization using evolutionary algorithms, Vol. 16*. John Wiley & Sons, 2001.

[5] Samareh, Jamshid A., Survey of shape parameterization techniques for high-fidelity multidisciplinary shape optimization *AIAA journal, Vol. 39* 877–884, 2001.

[6] Castonguay, Patrice and Nadarajah, Siva K, Effect of shape parameterization on aerodynamic shape optimization *45th AIAA Aerospace Sciences Meeting and Exhibit* 1–20, 2007.

[7] Jameson, A. Aerodynamic design via control theory. *Journal Scientific Computating Vol. 3* 233–260, 1988.

[8] G. van Bloemen Waanders, et Al. Sensitivity technologies for large scale simulation *Sandia Nationa Laboratories*, 2005.

[9] Nadarajah, Siva Kumaran The discrete adjoint approach to aerodynamic shape optimization *Stanford University*, 2003.

[10] Lacor, C. and Smirnov, S., Non-Deterministic Compressible Navier-Stokes Simulations using Polynomial Chaos *Proc. ECCOMAS Conf*, 2008.

[11] Dinescu, Cristian, Smirnov, Sergey, Hirsch, Charles and Lacor, Chris. Assessment of intrusive and non-intrusive non-deterministic CFD methodologies based on polynomial chaos expansions *International Journal of Engineering Systems Modelling and Simulation Vol. 2,* Inderscience Publishers, 87–98, 2010.

[12] Xiu, Dongbin and Karniadakis, George Em, The Wiener–Askey polynomial chaos for stochastic differential equations *SIAM journal on scientific computing Vol. 24,* 619–644, 2002.

[13] Hosder, Serhat and Walters, Robert W and Balch, Michael, Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables. *Proceedings of the 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, AIAA paper*, 2007.

[14] Jasak, HRVOJE and Tukovic, Zeljko, Dynamic mesh handling in openfoam applied to fluid-structure interaction simulations. *Proceedings of the V European Conference Computational Fluid Dynamics, Lisbon, Portugal, June*, 14–17 2010.

[15] Weller, H. G. and Tabor, G. and Jasak, H. and Fureby, C., A tensorial approach to computational continuum mechanics using object-oriented techniques *Computers in Physics, Vol. 12*, 620–631, 1998.

[16] Rumsey, Chris Turbulence modeling resource. *NASA, 2D NACA 0012 Airfoil Validation Case* 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# APPLICATION OF THE DECONVOLUTION METHODS FOR PROCESSING OF MEASUREMENT SIGNALS IN THE FAST PROCESSES

**Marat A. Goldfeld, Valery V. Pickalov**

Khristianovich Institute of Theoretical and Applied Mechanics SB RAS,
4/1 Institutskaya Str., Novosibirsk, Russia,
{gold, pickalov}@itam.nsc.ru

**Keywords:** Temperature Reconstruction, Thermocouples, High Speed Flow, Inverse Problems, Deconvolution.

**Abstract.** *In practice of experiment, one had to be faced with the distortions caused by measuring equipment, especially on a limit of its resolution. Sometimes not the equipment but the investigated process itself gives the grounds for distortion of the registered data, for example, at photography of quickly moving objects. Other examples of distortion of data are measurements of temperature and forces in wind tunnels. Problems of elimination of instrument function (or response function) of the device are most often reduced to the solution of the integral equations, usually, with convolution kernels. The solution of these equations is called problem of deconvolution. In present paper, methods and algorithms of the solution of some tasks, which appear at processing the signals obtained in aerodynamic experimental researches, are presented. The regularization algorithm of the solution of deconvolution problem, adapted to the level of experimental noise, is described. Results of numerical simulation of a task of temperature correction of thermocouples (temperature gages) lag at restrictions of temporary resolution of the thermocouples and short-term process are given.*

*The numerical solution of a non-stationary problem of heat conductivity in two-dimensional statement for determination of average temperature of the thermocouples at three values of their diameter (d=50, 100, 200 microns) was performed. Within numerical modeling, the two-dimensional equations of heat conductivity by the relaxation method were solved. Experimental definition of instrument function of the thermocouples was carried in experiment with constant step thermal loading. The same type of thermal loading was used at numerical simulation. Direct comparison of the calculated and experimental data for the thermocouple with diameter of 50 microns shows that by means of calculation it is possible to receive the temperature step function close to the experimental one. Therefore, it is possible to estimate theoretical impulse response function of the thermocouple by differentiation of the calculated response function of the thermocouple to the step loading.*

*The calculation results agree well with the temperature measurements by the thermocouple of the same size obtained by its immersion in molten aluminum. Instrument function, defined from the real experiment with immersion of thermocouples in aluminum melt, appeared to be close to the function obtained theoretically.*

## 1 INTRODUCTION

Ground-based and flight researches play an important role in creating new hypersonic vehicles. At hypersonic speeds, aerodynamic heating leads to extremely high heat flow and temperature on the outer surface of the machine and in the channel of the hypersonic propulsion air-jet engine [1]. A well-designed thermal protection system must maintain the integrity of the aircraft, preventing the thermal loads. Aerodynamic heating and aerodynamic resistance determine the choice of structural materials in the design of devices for hypersonic flight. Security, viability and stability are key considerations in achieving the goals of the flight. For their achievement, the knowledge of temperatures and heat flows for creation of effective thermal protection system [1, 2] is required.

Over the past decade, numerous experimental and numerical study of temperature and heat flow on aircraft elements were conducted to estimate and improve methods and models which are used for a prediction of operating conditions of the hypersonic vehicle [3-5].

Temperature measurement is important and therefore quite well studied direction in metrology. The application of traditional contact methods (thermocouples, resistance thermometers) [6, 7] and optical techniques [8] (Rayleigh scattering, pyrometers, laser-induced fluorescence, etc.) provides the acceptable accuracy of measurements. These methods have their advantages, disadvantages and limitations, and therefore are applied selectively in various conditions. Application of the contact methods for the measurement of high temperatures (more than 2000K) is limited to temperatures of melting and a high inertia of sensitive elements, resulting in a thermocouple condition to be far away from the stationary state. Optical methods, which allow to measure high temperatures, have limited access of optical diagnostics to the system [8, 9].

Implementation of optical temperature measurement in channel often becomes impossible for this reason. Research in high-speed wind tunnels with short operating time from 5 to 100 msec appears difficult solvable problem because of measuring system inertia [10]. The problem becomes especially actual when determining temperature fields inside the models channel [11].

Heat flow sensors and thermocouples based on them is one of the most widely used devices for temperature measurement in view of their simplicity, low price, simplicity of production and reliability and broad application. However, their applicability is not always possible due to the relatively low melting point and limited bandwidth of the sensor [6]. Therefore, in many applications, methods of compensation signal needed to establish the true temperature in small or extremely short measuring times. In this case, definition of the thermocouple time constant, which characterizes its inertia [10, 12], is required.

In [12] a method based on the technique of the two thermocouples to evaluate the average time constants has been proposed. The method is based on mutual power spectra of temperature fluctuations. However, this method is very sensitive to measurement noise and does not allow receiving high accuracy that at determining of a time constant was approximately 25 %.

Measurement of pulsating temperatures in the combustion chambers and the propulsion is necessary to determine the entropy and noise in these systems. It is known that most of the thermocouples, suitable for use in such an unfavorable flow have response times considerably more than required in the working frequency range. The main problem of compensation is a measurement of the thermocouple response time, as the response depends on the environmental conditions in which thermocouples are located, and is defined by the convective and radiating components [13].

Previously developed methods for measuring the time constant are usually based on a quick immersion of thermocouple into the hot medium or switching of a gas flow from cold to hot. Mechanical methods are suitable only for time constants from 100 msec or more. However, for ‘fast’ thermocouple with a response of less than 10 msec mechanical methods of switching cannot usually be applied. Electrical heating methods also are not satisfactory, especially if the thermocouple is working near the limit of its working capacity. Therefore methods to determine the thermocouple

time constant based on mutual power spectra processing [13] or solution of inverse problems [14] are developed.

Among several approaches to assessing the relationship of the time constants two basic assumptions were used: (1) the time constant is a function of the size of thermocouple based on empirical equations for the heat transfer coefficient of thin wires [15]; (2) the ratio of the time constants is defined as the ratio of the temperature derivatives, which is valid at equal temperatures of both thermocouples [16, 17]. The validity of such approaches in high-enthalpy flow and combustion remains uncertain, as the thermophysical properties can be changed in accordance with the conditions in the high-temperature region.

Serious problems in the process of temperature measurements arise in the study of fast processes in a short duration measurements, since in such cases it is necessary to take into account the inertia of the of time measurements. Such account inevitably leads to the necessity of solving inverse heat conduction problems [18], as well as to the problems of eliminating (correction) the impact of measuring system instrumental function. It should be noted that methods of inverse problems solution for processing signals of the dynamic weight and pressure measurements are used for a long time and more often than in the interpretation of thermocouple signals. For example, in [19, 20] the recovery of the temporal aerodynamic characteristics (pressure and forces) by solution of the convolution integral equation (deconvolution method) are described, usually by reducing the problem to a system of linear algebraic equations, or by solution of the problem in Fourier space [21].

In this regard a new modification of the method of two thermocouples was developed [13, 22], and the methods of deconvolution with parametric specification of the thermocouples instrumental functions in the space of the Laplace transform domain [23-25]. The resulting final integral equation containing the measured and actual flow temperature is solved using the original regularization parameter related to the grid in time [26]. This approach is further developed to take into account the dependence of the thermophysical properties of materials on the temperature, which leads to the nonlinear inverse problems of heat conduction, which are reduced by linearization to the known linear problems [27].

The analysis of known researches show that when interpreting the signals of temperature sensors and heat flow there is a trend towards more complex mathematical models, the transition from the time constant to the variable function of time.

The purpose of this paper is the following.

Study of the possibilities for improving the accuracy of temperature measurements over short time intervals, characteristic of fast running processes, particularly when tested in pulse wind tunnels.

Development of mathematical methods, algorithms and programs for improving the accuracy of temperatures and heat flows in the framework of modern methods for solving ill-posed inverse problems of gas dynamics.

Verification of methods and models for temperature and heat flow measurement, based on application of the method of two thermocouples and the method of a deconvolution.

## 2 FACILITY AND MEASUREMENTS

Experiments were performed at the hot-shot wind tunnel IT-302M [28] with arc heating in direct connect supersonic test mode. Such mode of investigation allows an effective use of the advantages of the hot-shot wind tunnel prechamber as a source of a high-enthalpy test gas (air).

The choice of the initial values of air pressure in the discharge chamber and the voltage of capacitors allowed obtaining the required flow temperature. This approach ensures not only the necessary Mach number but also the required pressure and temperature before the model entrance. High parameters may be reached due to the absence of technological problems related to the temperature

strength of the sharp edges of the model and instrumentation. In addition, the setup has large quartz windows for flow visualization and optical measurements.

The experiments were carried out at the following conditions at the duct entrance: Mach numbers $M_{en}$=5-8, total temperature $T_t$=2000-3000K, static pressure $P_{en}$=8-50 kPa. Operation time of wind tunnel at these conditions was 80-120ms. Peculiarity of this wind tunnel is a decrease of flow parameter during operation time. Character of change of total pressure and total temperature are presented in Fig. 1 and 2. One can see that total pressure is drops more than in four times and total temperature decreases approximately by 40% during the wind tunnel operation.

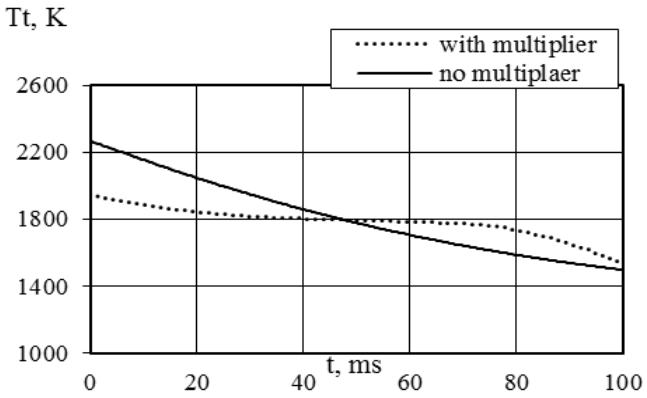



Figure 1: Time history of total temperature in free stream flow at Mach number 7.

Therefore some numbers of runs were carried out with the pressure multiplicator for maintenance of constant value of pressure and temperature as it can be seen in Figs. 1, 2 (point lines). This allows carrying out a measurement at fixed temperature (heat flux) of external flow. At the same time presented data demonstrate that real time of measurements does not exceed 50-60 ms even at Mach number of 7. At a decrease of Mach number, this time drops up to 20-25 ms.

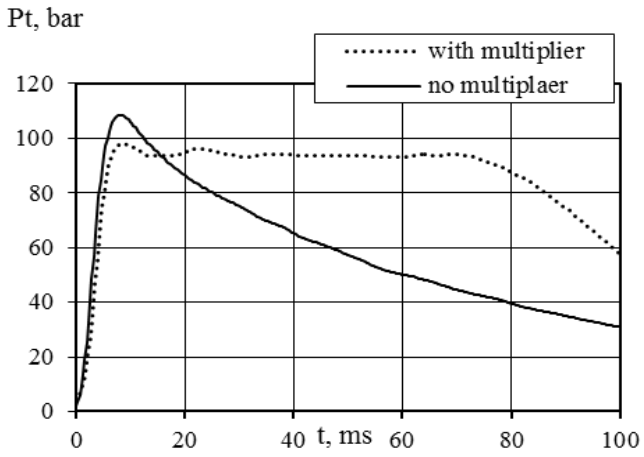



Figure 2: Time history of total pressure in free stream flow at Mach number 7.

The total temperature in the channel exit was measured by means of the rake consisting of 5 gas-flow hromel-alumel thermocouples with the sizes of measuring junction of 0.05, 0.1, 0.2 and 0.3 mm. The minimum sizes of thermo-junction are chosen for decrease of the response time at measurement under the conditions of short-time operation of wind tunnel. Thermocouples were installed in tubes made of stainless steel with external diameter 2.5 mm which functions as heat shield. The

device construction (Fig. 3) was manufactured so that identical dynamic and thermal conditions were provided before the entrance in the each thermocouple channel.

During the tests, next parameters were measured: the total flow parameters in first and second prechambers; distributions of static pressure and heat flux in the model channel; Pitot pressure and temperature before model by different technique. This allows obtaining flow condition in the measuring region.



Figure 3: Device for installation of thermocouple sensors.

The typical form of a response of thermocouples with the various size thermojunction is presented in Figure 4 together with change of total temperature in free stream of a wind tunnel. It can be seen that the achievable maximum of temperature depends on the size of junction of thermocouple and lies in the range from 40 to 100 ms while the real maximum of temperature in wind tunnel operation section was reached approximately at 10th ms. In all cases, this maximum essentially was lower than total temperature in a free stream. Its value and time of achievement substantially depends on initial temperature in a free stream and rate of its reduction.

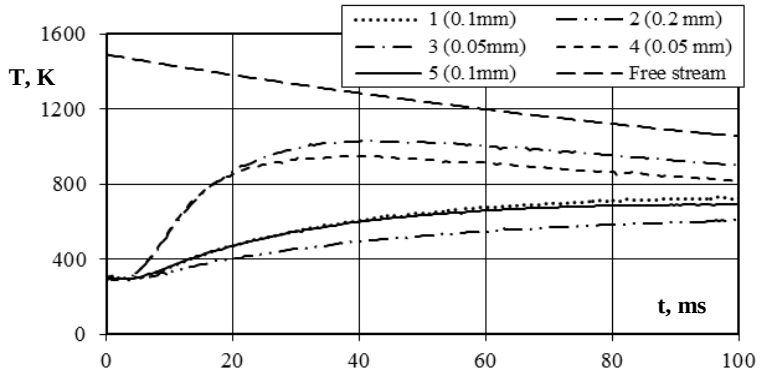



Figure 4: Measurement of total temperature in operation section of wind tunnel.

This data shows that all used thermocouples have a time delay of different duration. This data demonstrates also that measurement by thermocouples of the identical size, but they, nevertheless, does not always coincide owing to a deviation of the thermocouple junction size from the nominal value.

## 3 DECONVOLUTION METHOD

For processing the received experimental data and restoration of actual temperature, a deconvolution approach has been used. This procedure of stagnation temperature restoration is based on the usage of convolution integral equation solution. Such procedure is called deconvolution [29]. This

approach implies the need for information on the transfer function, which should be obtained based on calibration measurements of temperature step for thermocouples with the various size of junction.

Often in diagnostic experimental practice one can face with distortions caused by the measuring apparatus, especially working at the limit of its resolution. For example, the resolution of flow images are limited by so-called point spread function (PSF), arising due to diffraction effects on the receiving aperture, as well as various types of aberrations. Sometimes it is not the equipment, but the process under study itself gives rise to distortion of the recorded data, for example, when photographing fast-moving particles with insufficient temporal resolution. Another example of a data corruption is temperature and balance measurements in wind tunnels. Problems of the apparatus functions corrections are usually reduced to the solution of integral equations, most often of convolution type. Treatment of these equations and their solution are called deconvolution problem.

In general, the integral equation for linear distortions can be written as:

$$f(x) = \int g(x')K(x,x')dx' \tag{1}$$

Here, the vector x may include spatial variables as well as others - time, radiation frequency, etc.; $f(x)$ is the measured signal, $g(x)$ - unknown function, $K(x,x')$ - known kernel of the integral equation describing the instrumental response to the impulse action (δ-function). In the case of time invariant distortion equation (1) simplifies to

$$f(x) = \int g(x')K(x-x')dx' \tag{2}$$

the kernel of which now depends only on the difference of the arguments. Equation (6) is called the convolution equation. This equation describes the optical image blur due to subject movement (or recording device) as well as the problem of reconstruction of signals from sensors that have lag time [19]. Solution of integral equations of the first kind of the form (1), (2) refers to the number of ill-posed problems and requires the development of special regularization methods [8, 30-32].

The standard approach to the solution of convolution is to use the known relationship between Fourier transforms of three functions in equation (2):

$$\tilde{f}(\nu) = \tilde{g}(\nu)\widetilde{K}(\nu), \tag{3}$$

where the Fourier transform of the unknown function $\tilde{g}(\nu)$ is obtained by a simple algebraic transformation However, this approach is strongly susceptible to the random experimental noise and its regularization requires special procedures, to suppress high frequency noise components. Formula for the Tikhonov regularization of convolution equation is as follows:

$$g_\alpha(x) = F^{-1}\left(\tilde{f}(\nu)\frac{\widetilde{K}^*(\nu)}{\widetilde{K}(\nu)\widetilde{K}^*(\nu)+\alpha\nu^2}\right). \tag{4}$$

Here $F^{-1}$ – operator of inverse Fourier transform, symbol * is complex conjugation, regularization parameter α is restored from the equality of the residual norm to the norm of noise $\eta^2$ (residual criterion):

$$\left\|\tilde{f}(\nu) - \tilde{g}_\alpha(\nu)\widetilde{K}(\nu)\right\|^2 = \eta^2.$$

Particular attention is paid to the so-called algebraic reconstruction technique (ART), allowing to find the solution of large systems of linear algebraic equations (SLAE) by iterations, including a highly underdetermined SLAE, using a priori information at each iteration step. The system of linear algebraic equations corresponding to the integral equation (1) is obtained by its discretization:

$$f = Kg, \tag{5}$$

where $f$ and $g$ are vectors, and $K$ is matrix.

The algorithm ART is based on an additive correction of the solution at each step, using the solution in the previous step by the following formula [33]:

$$g_j^{k+1} = g_j^k + \lambda \frac{f_{i(k)} - \sum_j a_{i(k)j} g_j}{\sum_j a_{i(k)j}^2}, i(k) = k \cdot mod(M) + 1. \quad (6)$$

Here $k$ is the iteration number, $\lambda$ – relaxation parameter, $g$ - restored signal from the measured signal $f$, $a_{ij}$ – an element of the matrix $K$, approximating the PSF function (kernel of the convolution equation), $M$ – number of the matrix $K$ rows, $j=1,...,N$, $i=1,...,M$. Iterative process is stopped by the residual criterion [30] or by achieving a minimum residual norm.

Since inverse problems are very sensitive to the measurement noise, the measured signals $f$ are filtered by smoothing splines.

## 4 NUMERICAL SIMULATIONS

For calibration, we used the method of rapid drop of thermocouple in the melt aluminum to make a temperature step in the signal of the thermocouple.

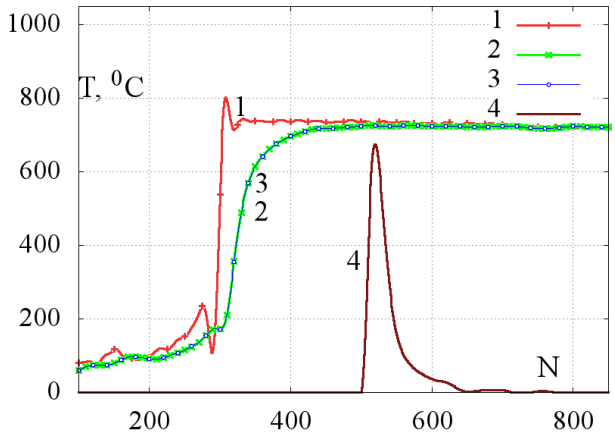



Figure 5: Reconstruction of the thermocouple PSF from the temperature step.

Figure 5 shows the smoothed experimental curve of the thermocouple (green line (2), noise estimation - 2%), and its transfer function, PSF (brown (4)), obtained by differentiating. In this section, ***N*** is a number of the time grid point. Use this function as a kernel of the integral equation allows to solve this equation for the true values of the temperature (red curve (1)). Technical features of the introduction of the thermocouple into the melt led to some transition process when jumping from room temperature to melt temperature. To improve initial part of the curve, it was partly approximated by a straight line with a continuation to zero. Consequently, the reconstruction before the start of the real temperature jump is not enough reliable, the step itself is restored with good accuracy.

Figures 6-7 demonstrate developed algorithm properties in numerical simulations of temperature step reconstruction. Here we investigate influence of the experimental noise on resulting reconstruction and spline smoothing of the noise. PSF of the thermocouple was taken from the real experiment (Figure 5). In Figures 6-8 black curves (3) are exact model temperature functions, green curves (2) present 'measured' temperatures, red ones (1) are reconstructed solutions of the equation (6). Also blue curves (4) show which 'measured' temperature would be like with reconstructed temperature (results of direct problem solutions) and brown curves (5) represent model PSF close to the experimental one.

The initial portion of the reconstructed step contains a small splash, similar to the Gibbs phenomenon, often displayed in problems of signal processing. In our numerical simulation, it was ver-

ified that this effect does arise for the stepwise model temperature distribution using real thermocouple transfer function. Figure 6 shows the 'measured' model temperature curve (arbitrary units, green curve) to which was added a random Gaussian noise of 2%, and the resulted reconstruction without smoothing (red curve). In reconstruction noticeable strong fluctuations around the exact model step (blue curve) aroused, errors level RMS=10.7%.

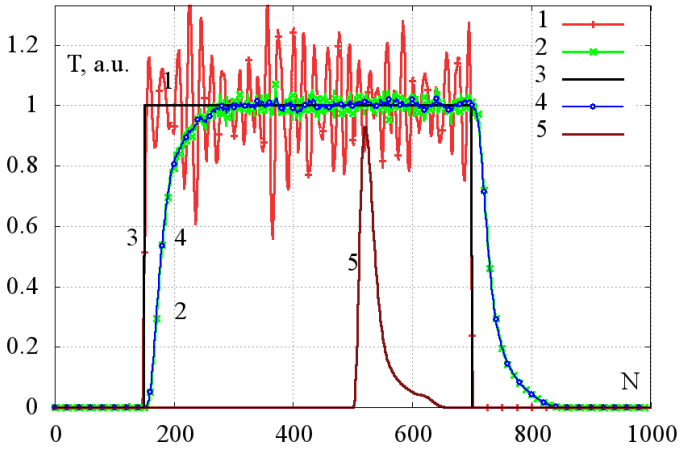



Figure 6: Temperature step reconstruction with experimental PSF, without smoothing, RMS=10.7%.

Figure 7 shows recovery of the model temperature step with the experimental noise (2%) in the thermocouple signal, when it is smoothed by regularized splines. Total step reconstruction error (RMS) was 4.65%.

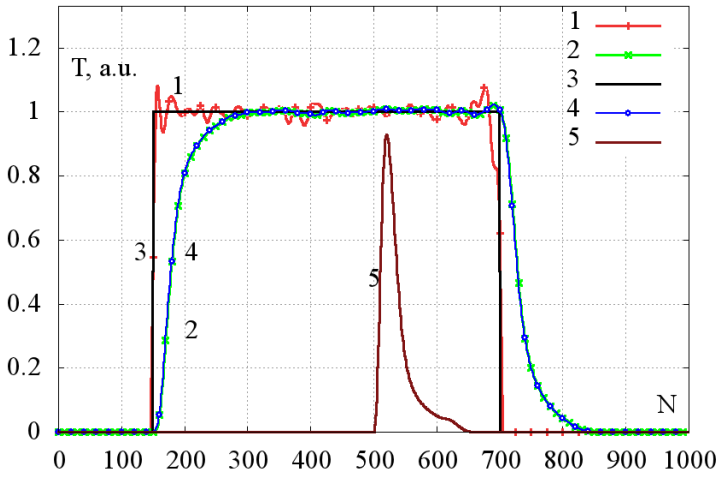



Figure 7: Temperature step reconstruction with experimental PSF; noisy data (2%), with spline smoothing, RMS=4.65%.

Qualitatively another case is presented in Figure 8. Here there is decreasing temperature function with noise level 2%. Here data with noise gave RMS=2.64% (a) after 140 iterations, while signal smoothing gives error much less, RMS=0.3% (b).

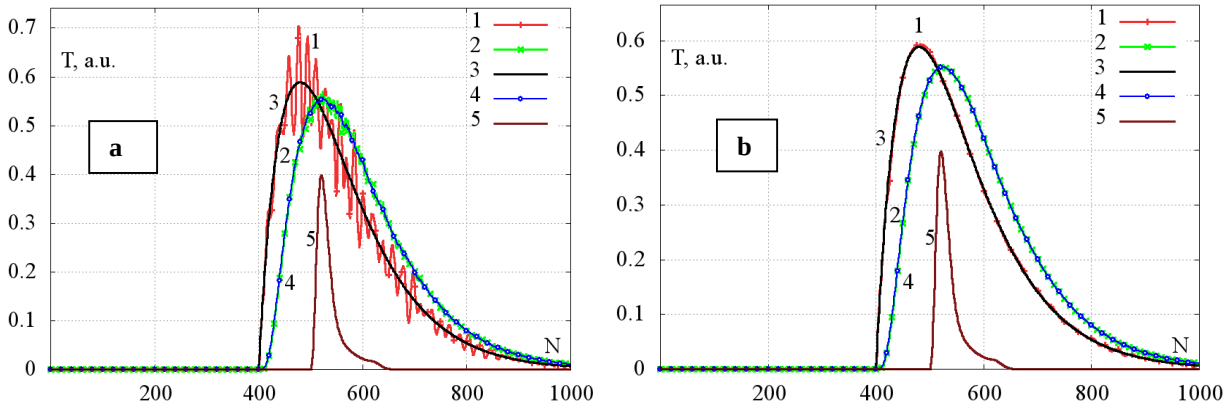



Figure 8: Temperature splash reconstruction with experimental PSF; noisy data (2%); a) without noise smoothing, RMS=2.64%; b) with spline smoothing, RMS=0.3%.

We can see good quality of the reconstruction observed in the initial part of the decreasing signal, but in the wings of the distribution reconstruction quality is not so good.

## 5 EXPERIMENTAL RESULTS

To perform comparative measurements from a variety of sources, it is necessary to estimate the error for each type of sensor. Usually two main sources of errors for the thermocouples are identified. Firstly, it is an error of a signal conversion (voltage) to the temperature because of use standard translation tables which errors may reach 1.7% change in temperature, which correspond directly to errors in the heat flow for the same 1.7%. Secondly, the accuracy of the thermophysical properties of the materials of the thermocouple junction can cause errors up to 8% of the heat flux [23, 34]. Since the same type of sensors is used, the physical error sources are the same, but the values are different. For the evaluation of measurement errors calibration of all of sensors was carried out in all estimated range of measurements. The weighted average difference between the calibration point and the calibration curve was used as an error in the temperature measurement. The error from the voltage conversion in the temperature is found to be no more than 0.2% and an error of thermophysical properties reached 5.7%.

One of the method of stagnation temperature restoration is based on the results of the solution convolution integral equation (deconvolution). This requires information on the transfer function, which can be obtained on the basis of determination of the temperature calibration step (Figure 9).

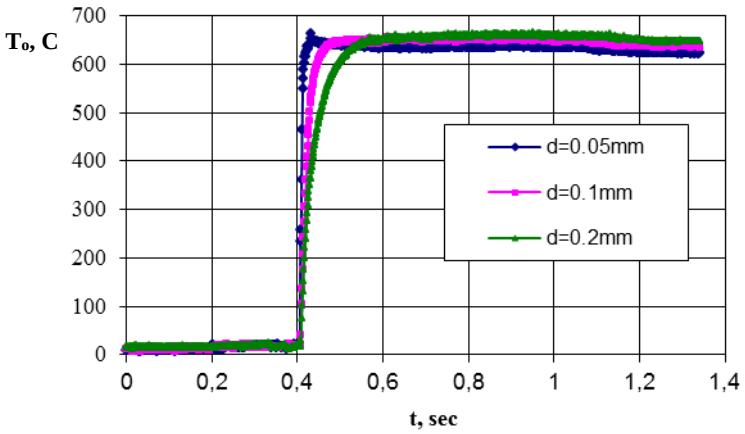



Figure 9: Thermocouples response at fast immersion in molten aluminium.

In this case, the transfer function is obtained by differentiating the response of the measuring system on the step. To get a step signal thermocouple was immersed fast in the molten aluminum (99%). The special device for simultaneous input of three thermocouples was for this purpose used.

Figure 9 shows the step responses for the thermocouples of different size, which depends on the size of the thermocouple junction. The temperature of molten aluminum (643$^{o}$C) was measured with an average error of no more than 1.54% for the entire time span, which is typical within the range of preliminary calibration of thermocouples.

For processing of the obtained experimental data and to implement reconstruction of real temperature of flow in impulse wind tunnel, two different approaches, described above, have been employed. The first method of signal processing is based on use of readings of two thermocouples with diameters of 0.1 mm and 0.2 mm, and the data on change of Reynolds number during an mode of operation of a impulse wind tunnel. The results obtained have shown in Figure 10.

One can see that initial temperature (without correction) was lower than theoretical value and maximum was reached at 55 ms and 90 ms. These moments are close to finish of operation of wind tunnel.

Application of developed technique of two thermocouples allows with sufficient accuracy to define total temperature in wind tunnel at short-duration action. Results of reconstruction of total temperature by means of this technique are presented in Figure 10 (red line). Maximal temperature is reached approximately at 27 ms and it is close to theoretical value. At the same time, one should notice that restored temperature remains a little bit lower than theoretical value. Nevertheless, results obtained confirmed possibility of determination of maximal temperature and applicability of such approach for a wide class of the tasks in view of simplicity and accessibility.

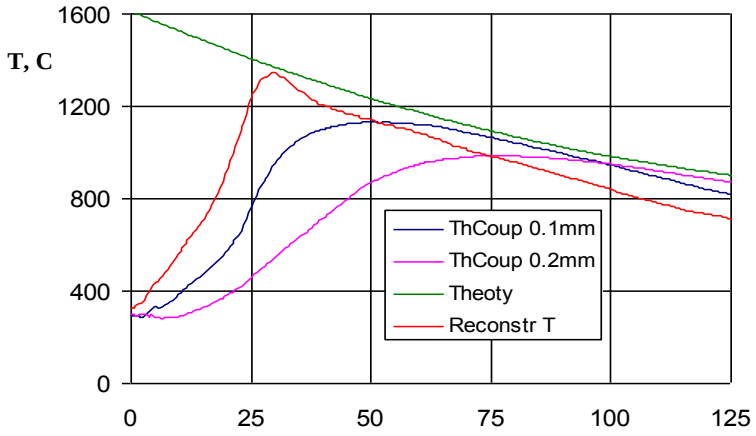



Figure 10: Total temperature reconstruction: method of two thermocouples

Second method of reconstruction allows to exclude errors denoted above. This method is based on deconvolution of integral equation. Kernel of this equation can be restored by differentiation of the experimental output response of thermocouple to the temperature step of input signal. For this purpose the information about the transfer function is necessary. This function has been obtained from results of measurements of a temperature step for two different diameters (Figure 11). In both cases, the same signals of thermocouples were used, which were used at the processing by means of technique of two thermocouples. Before deconvolution, experimental signal was smoothed by splines. Convolution kernels were chosen as decreasing exponents with a single unknown parameter, namely, half-width. During deconvolution the regularization procedure in the Fourier domain was employd. Half-width parameter searching was done by minimization of residual norm, and resulted in the next values: the first half-width was approximately 20 ms and the second one was 40 ms.

Curves of “ThCoup 0.1mm” and of “ThCoup 0.2mm” in Figure 11 coincide with temperature measurements by two thin thermocouples with diameters 0.1 and 0.2 mm, and curves of “Reconstr 0.1mm” and “Reconstr 0.2mm” are result of solution (deconvolution) of integral equation for curves 1 and 2. As opposed to previous method here were obtained two values of maximal temperatures, which correspond well to each other in magnitude but have different position depending on time.

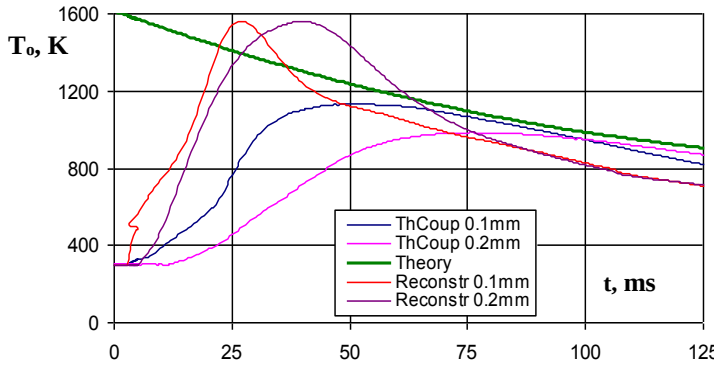



Figure 11: Total temperature reconstruction: deconvolution approach

One can see that application of deconvolution method allows determining the temperature, conformable to temperature of flow in wind tunnel better than method of two thermocouples. Apparently, best result was obtained at the use of thermocouple with smaller size (0.1 mm).

Error estimation of restoration procedure using the suggested approaches has shown that temperature can be recovered with the accuracy not worse than 8%.

## 6 CONCLUSIONS

The estimation of an error of definition of total temperature by means of the offered approaches has demonstrated that the temperature can be reconstructed with accuracy not worse than 8%.

An iterative reconstruction technique ART gives some promising results for the deconvolution problem, which take into account noise estimation for adaptive regularized cubic splines. The important effect is found. Results of true temperature reconstruction at solution of the integral equation strongly depend on noise level in a signal and quality of its smoothing. In addition, noise in point spread function, which in these numerical experiments has reached 10%, influences results weakly.

## 7 ACKNOWLEDGEMENTS

This research was partially supported by the Russian Fund for Basic Research, project No. 15-08-04804.

## REFERENCES

[1] N. Lamorte, P.P. Friedman, Hypersonic aeroelastic and aerothermoelastic studies using computational fluid dynamics. *AIAA J.*, **52**, No.9, 263-277, 2014.

[2] T. Wadhams, E. Mundy, M. MacLean, M. Holden, Ground test studies of the HI-FiRE-1 transition experiment. Part 1: Experimental results. *J. of Spacecraft and Rockets*, **45**, No.6, 1134-1148, 2008.

[3] J.R. Micol, Hypersonic aerodynamic /aerothermodynamics testing capabilities at Langley Research Center: Aerothermodynamic facilities complex. *AIAA Paper*, 95-2107, 1995.

[4] J.S. Hodge, S.F. Harvin, Test capabilities and recent experiences in the NASA Langley 8-foothigh temperature tunnel. *AIAA Paper* 2000-2646, 2000.

[5] M.S. Holden, T.P. Wadhams, A database of aerothermal measurements in hypersonic flows in 'Building Block' experiments for CFD validation. *AIAA Paper* 2003-1137, 2003.

[6] O.A. Geraschenko, A.N. Gordov, A.K. Eremina, et al. *Temperature Measurements*, Kiev, Naukova Dumka, 1989 (in Russian).

[7] R.S. Okojie, P.M. Danehy, A.N. Watkins, A.F. Mielke, J.H. Grinstead, An overview of NASA hypersonic experimental diagnostic and instrumentation technologies for ground and flight testing. *AIAA Paper* 2009-7279, 2009.

[8] V.M. Boiko, A.M. Orishich, A.A. Pavlov, V.V. Pickalov *Methods of Optical Diagnostics in Aerophysical Research*. Novosibirsk, NSU, 2009 (in Russian).

[9] T.E. Diller, Advances in heat flux measurement. *Advances in Heat Transfer,* **23**, Academic Press, New York, 279–368, 1993.

[10] U. Hegde, M.Y. Bahadori, D.P. Stocker, Oscillatory temperature measurements in a pulsed microgravity diffusion flame. *AIAA Journal*, **38**, No.7, 1219-1229, 2000.

[11] P.J. Kennedy, J.M. Donbar, J.R. Trelewicz, C. Gouldstone, J.P. Longtin, Heat flux measurements in a scramjet combustor using direct write technology. *AIAA Paper* 2011-2330, 2011.

[12] W.C. Strahle, M. Muthukrishnan, Thermocouple time constant measurement by cross power spectra. *AIAA J.*, **14**, No.11, 1642-1645, 1976.

[13] M. Tagawa, Y. Ohta, Two-thermocouple probe for fluctuating temperature measurement in combustion – rational estimation of mean and fluctuating time constants. *Combustion and Flame*, **109**, No.4, 549–560, 1997.

[14] J. Zhou, Y. Zhang, J.K. Chen, Z.C. Feng, Inverse estimation of spatially and temporally varying heating boundary conditions of a two-dimensional object. *International Journal of Thermal Sciences*, **49**, No.9, 1669-1679, 2010.

[15] L.J. Forney, G.C. Fralick, Two-wire thermocouple: Frequency response in constant flow. *Rev. Sci. Instrum.*, **65**, No.10, 3252–3257, 1994.

[16] P. Cambray, Measuring thermocouple time constants: A new method. *Combust. Sci. Technol.*, **45**, No.4, 221–224, 1986.

[17] I. Warshavsky, On-line dynamic gas pyrometry using two-thermocouple probe. *Rev. Sci. Instr*., **66**, No.3, 2619-2624, 1995.

[18] O.M. Alifanov, *Inverse Heat Transfer Problems*. Berlin, Springer, 1994.

[19] D.J. Mee, Dynamic calibration of force balances for impulse hypersonic facilities. *Shock Waves*, **12**, No.6, 443-455, 2003.

[20] A.L. Smith, D.J. Mee, W.J.T. Daniel, T. Shimoda, Design, modelling and analysis of a six component force balance for hypervelocity wind tunnel testing. *Computers & Structures*, **79**, No.11, 1077-1088, 2001.

[21] R. Semaan, P. Scholz, Pressure correction schemes and the use of the Wiener deconvolution method in pneumatic systems with short tubes. *Experiments in Fluids*, **53**, No.3, 829-837, 2012.

[22] P. Hung, S. McLoone, G. Irwin, R. Kee, C. Brown, In situ two-thermocouple sensor characterisation using cross-relation blind deconvolution with signal conditioning for improved robustness. *Lecture Notes in Electrical Engineering*, **24**, 273-286, 2009.

[23] S. Loehle, U. Fuchs, Theoretical approach to surface heat flux distribution measurement from in-depth temperature sensors. *Journal of Thermophysics and Heat Transfer*, **26**, No.2, 352-356, 2012.

[24] S. Löhle, J.-L. Battaglia, P. Jullien, B. van Ootegem, J.-P. Lasserre, J. Couzi, Improvement of high heat flux measurement using a null-point calorimeter. *Journal of Spacecraft and Rockets*, **45**, No.1, 76-81, 2008.

[25] S. Loehle, U. Fuchs, P. Digel, T. Hermann, J.-L. Battaglia, Analysing inverse heat conduction problems by the analysis of the system impulse response. *Inverse Problems in Science and Engineering*, **22**, No.2, 297-308, 2014.

[26] J.I. Frankel, M. Keyhani, Theoretical development of a new surface heat flux calibration method for thin-film resistive temperature gauges and co-axial thermocouples. *Shock Waves*, **23**, No.2, 177-188, 2012.

[27] J.I. Frankel, M. Keyhani, Nonlinear inverse calibration heat conduction through property physics. *Journal of Thermophysics and Heat Transfer*, **28**, No.2, 203-217, 2014.

[28] L.N. Pusyrev, M.I. Yaroslavtsev, Stabilization of gas parameters in prechamber of hypersonic hot-shot wind tunnel. *Izvestia Academii Nauk*, No.5, 86-93, 1990 (in Russian).

[29] S.M. Riad, The deconvolution problem: An overview. *Proceedings of the IEEE*, **74**, No.1, 82-85, 1986.

[30] A.N. Tikhonov, V.A. Arsenin, *Solutions of Ill-posed Problems*. Washington, Winston & Sons, 1977.

[31] V.V. Pickalov, Flow diagnostics and inverse problems: deconvolution and tomography. *15th International Conference on Methods of Aerophysical Research* (ICMAR-2010, Novosibirsk, November 1-5, 2010). Abstracts. Pt. II. Novosibirsk, Parallel, 200-201, 2010.

[32] V.V. Pickalov, M.A. Goldfeld, Temperature reconstruction in high-speed flow. *Proceedings of the 12th International Symposium on Experimental Computational Aerothermodynamics of Internal Flows* (13-16 July 2015, Lerici, Italy). No.ISAIF12-40, 1-10, 2015.

[33] R. Gordon, A tutorial on ART (Algebraic Reconstruction Techniques), *IEEE Trans. Nucl. Sci.*, **21**, No.3, 78-94, 1974.

[34] W. Flaherty, J.M. Austin, Comparative surface heat transfer measurements in hypervelocity flow, *Journal of Thermophysics and Heat Transfer*, **25**, No.1, 180-183, 2011

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# INVERSION METHODS BASED ON KALMAN FILTERING FOR IDENTIFICATION IN TUNNELING PROBLEMS

**Tamara Nestorović**[1] **and Luan T. Nguyen**[1]

[1]Ruhr-Universität Bochum, Mechanics of Adaptive Systems
Universitätsstr. 150, 44801 Bochum, Germany
e-mail: {Tamara.Nestorovic,Thanh.Nguyen-w7t}@rub.de

**Keywords:** Inversion methods, Kalman filter, Parameter estimation.

**Abstract.** *In this study we present an overview of the inversion methods based on the Kalman filtering technique, which was implemented for the identification of geological structure ahead of the underground tunnel. The methodology is particularly aimed at solving for model parameters even in the presence of noisy environment (measurements). One of the main goals of this investigation is related to the advance of the tunnel boring machine (TBM) during excavation in unknown environment in order to mitigate the excavation risks due to unpredicted obstacles and reduce costs caused by the TBM stoppage during tunneling under uncertain soil conditions. The inversion methods are implemented for the purpose of the reconnaissance in mechanized tunneling.*

*Several inversion methods are investigated and implemented. The Kalman filter as parameter estimator, already having been successfully implemented in control systems, has been introduced here also for the purpose of geotechnical parameters estimation. Since the Kalman filter was originally developed for linear systems primarily represented in the state space form, the implementation with the geotechnical soil models was limited by the nonlinearities. Therefore modifications of the Kalman filter have been introduced and implemented, such as extended Kalman filter (EKF), sigma-point Kalman filter (SPKF) and unscented Kalman filter (UKF). The extended Kalman filter local iteration procedure incorporated with finite element analysis software has been used for identification of the soil parameters using tunneling induced deformations under assumptions of existing surface pressure load and an obstacle ahead of the tunnel face. The identification is performed based on the numerically generated noisy measurements. The inherent linearization in the extended Kalman filter makes it difficult to implement and it is reliable only for slightly nonlinear system models. Therefore the further improvement of the inversion method has been done by combining the EKF with the derivativeless deterministic sampling-based approximation where a set of deterministically sampled points, called sigma points, can be used to represent the mean and covariance of the estimated quantities. The update mechanisms are inherited from the linear Kalman filter. For the purpose of the full seismic waves inversion for predicting ahead of the underground tunnel a new hybridized global optimization method that combines the simulated annealing global search with unscented Kalman filter minimization has been proposed. The implementation of the inversion methods has been shown on several numerical examples.*

## 1 INTRODUCTION

From the optimization point of view, inverse methods can generally be considered as a solution of the optimization problem, mainly through minimization of certain objective function. Implementation of the inversion methods in different fields of research and engineering plays an important role. One demanding implementation field is the reconnaissance in mechanized tunneling. Identification of the soil parameters ahead of the tunnel front, in order to predict possible obstacles, layer changes or similar anomalies can contribute to considerable savings by avoiding the damages or unplanned stoppage of the tunnel boring machine (TBM). Inversion methods for parameter estimation based on Kalman filters represent a promising tool for such applications.

The Kalman filter method was developed by Rudolf E. Kalman in early 1960s [1] and was originally aimed at implementations in control of dynamics systems and signal processing. In recent years, some applications of the Kalman filter method have also arisen in the field of geotechnical engineering for estimation of material parameters [2, 3, 4, 5]. Originally the Kalman filter was developed as dynamic least square states estimator for systems represented by linear models in the state space form. A modification of the approach with application in geotechnical engineering uses quasi-static models since the geotechnical parameters are static over the considered estimation time frame. Thus, a stationary transition of state estimates is suitable for this purpose [2].The procedure starts from a priori estimates and utilizes a set of measurement data to calculate a posteriori estimates. Means and covariances of the quantities being estimated are repeatedly updated so that the variance of the estimation error is reduced until convergence of the estimated quantities to unbiased true state and model parameters is reached [1, 6]. In this paper several techniques relying on modifications of Kalman filter are presented and implemented for the estimation of parameters in geotechnical models. The methodology is documented by several numerical examples.

## 2 KALMAN FILTER BASED METHODS

Quasi-stationary representation of a geotechnical model can be generally expressed in the form of a nonlinear state space matrix equation

$$\mathbf{d} = \mathbf{h}(\mathbf{x}, \mathbf{m}, \mathbf{f}) \tag{1}$$

where $\mathbf{d}$ stores modeling outputs obtained from the model $\mathbf{h}(\cdot)$ (usually a finite element – FE model), $\mathbf{x}$ is the current physical state of the model, $\mathbf{m}$ contains the model parameters (for example material parameters), and $\mathbf{f}$ represents external loads including the boundary conditions. When it is assumed that the applied loads and the initial and boundary conditions are properly determined, the identification problem consists of determining the actual model parameters $\mathbf{m}$. A frequent problem is the model calibration by inverse analysis in order to match the model outputs $\mathbf{d}$ to the in-situ measurement or laboratory tests measurement data $\mathbf{d}^{obs}$. This matching can be expressed in terms of optimization of a certain objective function, i.e. minimization of the required misfit. Arising problems such as high nonlinearity of the geotechnical models, time consuming forward calculations of the models and uncertainty of the measurement data require special treatment and adaptation of the identification methods to these requirements. The Kalman filter based methods which meet those requirements are presented in the following sections.

### 2.1 Extended Kalman Filter

The quasi-stationary model of Eq. 1 can be represented in terms of the state transition 2a and the output 2b equations:

$$\mathbf{m}_k = \mathbf{m}_{k-1} + \mathbf{w}_{k-1}, \tag{2a}$$

$$\mathbf{d}_k = \mathbf{h}\left(\mathbf{m}_k\right) + \mathbf{v}_k, \tag{2b}$$

where vector $\mathbf{m}_k$ contains the model parameters and $\mathbf{x}_k$ is the physical state of the model at recursive step $t_k$. Vector $\mathbf{d}$ extracts modeling outputs at predefined positions in the model from the modeling observation Eq. 2b. In Eq. 2a, the stationary process transition is added by a modest amount of pseudo-noise $\mathbf{w}_k$ to act like parameter updating from one iterative step to the next. Inexactness of the modeling observation function $\mathbf{h}(\cdot)$ in Eq. 2b is characterized by an amount of uncertainty $\mathbf{v}_k$. Additive uncertainties $\mathbf{w}_k$ and $\mathbf{v}_k$ are assumed to be uncorrelated and white Gaussian having zero-mean and time-invariant covariance matrices $\mathbf{Q}$ and $\mathbf{R^m}$ respectively.

In the Kalman filter algorithm the mean and error covariance of the state are propagated through time. These mean and error covariance of the state are updated every time new measurements arise [7]. The time update equations of the mean and covariance of the state are calculated in the following way:

$$\hat{m}_k^- = \hat{m}_{k-1}^+; P_k^- = P_{k-1}^+ + Q, \tag{3}$$

where $\hat{m}_k = E(m_k)$ represents the mean (mathematical expectation) of the state and $P_k = E\left[(m_k - \hat{m}_k)(m_k - \hat{m}_k)^T\right]$ is the covariance matrix at *k*-th iteration. The superscript '–' denotes *a priori* estimate, i.e. the estimate before measurement is available, whereas the superscript '+' denotes *a posteriori* estimate after the measurement is taken into account. At the time $t_k$ in the local iterative loop $i$ of the extended Kalman filter, the Kalman gain $K_{k,i}$, the measurement updates of the state estimate $\hat{m}_{k,i+1}^+$ and the estimation error covariance matrix $P_{k,i+1}^+$ are iterated according to the following equations.

$$\begin{aligned}
H_{k,i} &= \left.\frac{\partial h}{\partial m}\right|_{\hat{m}_{k,i}^+} \\
K_{k,i}^g &= P_k^- H_{k,i}^T (H_{k,i} P_k^- H_{k,i}^T + R)^{-1} \\
P_{k,i+1}^+ &= (I - K_{k,i}^g) P_k^- \\
\hat{m}_{k,i+1}^+ &= \hat{m}_k^- + K_{k,i}^g \left[d_{exp} - h(\hat{m}_{k,i}^+, 0) - H_{k,i}(\hat{m}_k^- - \hat{m}_{k,i}^+)\right]
\end{aligned} \tag{4}$$

Here $H$ represents the sensitivity matrix (observation derivative matrix with respect to the state estimate), $K^g$ is the Kalman gain and $d_{exp}$ is expected measurement. The sensitivity matrix is calculated for each global iteration $k$ and local iteration $i$ as

$$H_{m,n} = \frac{\partial h_m}{\partial m_n} = \frac{h_m(\mathbf{m} + \Delta m_n \mathbf{e}_n) - h_m(\mathbf{m})}{\Delta m_n} \tag{5}$$

where $h_m$ is the *m*-th element of observation vector, $m_n$ is the *n*-th element of the parameter vector and $\Delta m_n$ is a small increment of the *n*-th element of parameter vector. The unit vector $\mathbf{e}_n$ has value one in *n*-th element and zeros elsewhere. The estimation algorithm is coupled with the finite element analysis (FEA) software. An independent treatment of FEA program represents a significant advantage of the proposed method, in the sense that any FEA program can be

incorporated into the process when the communicative interface between the main algorithm flow and the FEA program has been created. In order to begin the filter process, an initial value for state vector $x_0$ and its corresponding estimation error covariance matrix $P_0$ are assigned. Usually, initial state vector contains the most likely values of parameters of the system that can be inferred from prior knowledge or preliminary geomechanical tests. The estimation error covariance matrix gives information about the confidence of the initial parameter values. The procedure is to locally iterate the filtering process on the first set of data in order to assess the information inherent to the data. Then, the second set of data is processed to filtering and is similarly iterated, and so on. The iterative process will continue until a stop criterion is met, which may be *i*) convergence of the cost function to he predefined minimal values, *ii*) absolute value of the two successive estimates is less than a predefined tolerance or *iii*) the specified number of iterations. In this method, the error covariance matrix of estimation $P_k$ is enlarged with a modification weight in every global iteration to obtain fast convergence.

### 2.2 Sigma-Point Kalman Filter

Due to inherent linearization the EKF is related with implementation difficulties, and in addition this state estimation method has limitations in implementation with highly nonlinear models. Therefore alternative types of the Kalman filter for nonlinear models have been developed, which obviate the need for linearization of the state-space model, like the unscented Kalman filter [8]. Later, [9] has combined and systematized this new filter together with some other independent works that derived the derivativeless deterministic sampling based approximation of the Gaussian statistics to belong to the family of SPKF. This family of the Kalman filter is based on the principle that a set of deterministically sampled points, called sigma-points, can be used to represent the mean and covariance of the estimated quantities. Update mechanisms are inherited from the linear Kalman filter [1].

The sigma-points consist of $2n + 1$ discretely distributed points around the current estimate on the $n$-dimension space. The distribution of these sigma-points around the current estimate is determined by square-root decomposition of the prior covariance and the spread of them can be assigned by setting a scaling parameter $\zeta$ as follows.

$$\boldsymbol{\sigma} = \zeta\sqrt{\mathbf{P}_{k-1}}, \tag{6}$$

$$(\boldsymbol{\mathcal{M}}_{k-1})_0 = \hat{\mathbf{m}}_{k-1}, \tag{7}$$

$$(\boldsymbol{\mathcal{M}}_{k-1})_i = \hat{\mathbf{m}}_{k-1} + (\boldsymbol{\sigma})_i, \text{ for i = 1:n} \tag{8}$$

$$(\boldsymbol{\mathcal{M}}_{k-1})_{i+n} = \hat{\mathbf{m}}_{k-1} - (\boldsymbol{\sigma})_i, \text{ for i = 1:n} \tag{9}$$

where the scaling parameter is calculated as $\zeta = \sqrt{n + \lambda}$. The parameter $\lambda$ remains free to tune. It is suggested to select $n + \lambda = 3$ when $\mathbf{m}$ is assumed Gaussian [8]. The notation $(\cdot)_i$ denotes the $i$-column of the matrix within parentheses.

Each sigma-point is associated with a weight. The weights are defined as in Eq. 16 such that the sum of weights is unity.

$$W_0 = \frac{\lambda}{n+\lambda}, W_i = W_{i+n} = \frac{1}{2(n+\lambda)}, \text{ for i = 1:n} \tag{10}$$

Direct nonlinear transfer of the sigma-points through the nonlinear state-space functions helps reserve second order accuracy of the Gaussian approximate estimated quantities as opposed to first order truncation of the Taylor series employed by the EKF. In addition, the minimization

of the objective misfit function Eq. 11 should be provided.

$$S(\mathbf{m}_k) = \frac{1}{2}\left((\mathbf{h}(\mathbf{m}_k) - \mathbf{d}^{obs})^t \mathbf{R}^{-1}(\mathbf{h}(\mathbf{m}_k) - \mathbf{d}^{obs}) + (\mathbf{m}_k - \hat{\mathbf{m}}_{k|k-1})^t \mathbf{P}^{-1}_{k|k-1}(\mathbf{m}_k - \hat{\mathbf{m}}_{k|k-1})\right) \tag{11}$$

The recursive prediction step and correction step of the SPKF are summarized in Algorithm 1.

**Algorithm 1:** The SPKF algorithm for parameter identification

**Initialization:**

$$\begin{aligned} \hat{\mathbf{m}}_0 &= \mathbf{m}^{prior} \\ \mathbf{P}^m_0 &= \mathbf{P}^{prior} \end{aligned}$$

$k \leftarrow 1$

**while** $S$ *(Eq. 11)* $< TOLERANCE$ **do**

**Prediction step:**

Generate the sigma-points Eq. 6-9:$(\boldsymbol{\mathcal{M}}_{k-1})_i$

$$(\boldsymbol{\mathcal{M}}_{k|k-1})_i = (\boldsymbol{\mathcal{M}}_{k-1})_i$$

$$\hat{\mathbf{m}}_{k|k-1} = \sum_{i=0}^{2n} W_i (\boldsymbol{\mathcal{M}}_{k|k-1})_i$$

$$\mathbf{P}^m_{k|k-1} = \sum_{i=0}^{2n} W_i \left((\boldsymbol{\mathcal{M}}_{k|k-1})_i - \hat{\mathbf{m}}_{k|k-1}\right)\left((\boldsymbol{\mathcal{M}}_{k|k-1})_i - \hat{\mathbf{m}}_{k|k-1}\right)^t + \mathbf{Q}$$

$$(\boldsymbol{\mathcal{D}}_{k|k-1})_i = \mathbf{h}\left((\boldsymbol{\mathcal{M}}_{k|k-1})_i\right)$$

$$\hat{\mathbf{d}}_{k|k-1} = \sum_{i=0}^{2n} W_i (\boldsymbol{\mathcal{D}}_{k|k-1})_i$$

$$\mathbf{P}^d_{k|k-1} = \sum_{i=0}^{2n} W_i \left((\boldsymbol{\mathcal{D}}_{k|k-1})_i - \hat{\mathbf{d}}_{k|k-1}\right)\left((\boldsymbol{\mathcal{D}}_{k|k-1})_i - \hat{\mathbf{d}}_{k|k-1}\right)^t + \mathbf{R}$$

$$\mathbf{P}^{md}_{k|k-1} = \sum_{i=0}^{2n} W_i \left((\boldsymbol{\mathcal{M}}_{k|k-1})_i - \hat{\mathbf{m}}_{k|k-1}\right)\left((\boldsymbol{\mathcal{D}}_{k|k-1})_i - \hat{\mathbf{d}}_{k|k-1}\right)^t$$

**Correction step:**

$$\begin{aligned} \mathbf{K}_k &= \mathbf{P}^{md}_{k|k-1}(\mathbf{P}^d_{k|k-1})^{-1} \\ \hat{\mathbf{m}}_k &= \hat{\mathbf{m}}_{k|k-1} + \mathbf{K}_k\left(\mathbf{d}^{obs} - \hat{\mathbf{d}}_{k|k-1}\right) \\ \mathbf{P}^m_k &= \mathbf{P}^m_{k|k-1} - \mathbf{K}_k \mathbf{P}^d_{k|k-1}\mathbf{K}^t_k \end{aligned}$$

$k \leftarrow k + 1$

**end**

In this Algorithm, the means of predicted parameters and model outputs are calculated as weighted summation over the sigma-points and the corresponding predicted model outputs.

The predicted model outputs $\boldsymbol{\mathcal{D}}_{k|k-1}$ are the results of direct nonlinear transfers of the representative sigma-points $\boldsymbol{\mathcal{M}}_{k|k-1}$ through the nonlinear model function $\mathbf{h}(\cdot)$. The predicted estimation error covariances of the estimated parameters $\mathbf{P}^m$, of the model outputs $\mathbf{P}^d$, and of the cross-covariance between them $\mathbf{P}^{md}$ are calculated as weighted summation of the squares of distances between each sigma-point and respective model output to the corresponding mean estimates. The correction equations for the SPKF take into account the means and covariances approximated using the sigma-points and the nonlinear model outputs in the prediction step without need for calculation of the Jacobians.

For application in calibration of FE model in geomechanics, value for each parameter can be limited to a certain range depending on initial categorization of geomaterial samples. Therefore, it is advantageous to apply a feasible range for each parameter in the identification process. The advantage of giving constraints to the parameter space is threefold: *i*) it helps to exclude the unreasonable convergence points, which are local minima, of the objective function, *ii*) it allows the FE code to run uninterruptedly during the iterative estimation process as the parameters are clipped in the reasonable regions, and *iii*) it may help to accelerate the estimation process by filtering algorithms since the estimation bias is not allowed to explode. Here, the 'clipping' constraints handling for both the EKF and the SPKF described in [10] is applied.

### 2.3 Unscented Kalman Filter

The UKF is a very successful state and parameter estimator for nonlinear dynamic models in which the posterior estimates are inferred from noisy observation data. It aims to overcome the inherent limitations of the traditional extended Kalman filter that works only for slightly nonlinear models and requires linearization of the nonlinear models. This filter belongs to the sigma-point Kalman filter family, and represents a derivative-free estimator for nonlinear state-space models that can be efficiently adapted for solving parameter identification of static and dynamic models [9].

The forward model is described in a similar form as in Eq. 2b. In this case **d** stores the model outputs at $r$ observation points resulting from the nonlinear model **h**. For time-dependent problem such as wave propagation, the model $\mathbf{h}(\cdot)$ is solved in time steps for the simulation period; and the resulting batch data are stored in **d**. The modeling error **v** is caused by assumptions made in building the mathematical model and numerical approximations. This error can be assumed to be zero-mean Guassian distribution with covariance $\boldsymbol{R}$. The misfit $s$ to be minimized is constructed for each output measurement $j$ by squares of differences between the measured data $d^{obs}(t)$ and the modeled outputs $d(t)$ in the time window of interest $[0, \tau]$:

$$s_j = \frac{1}{2}\int_0^{\tau} \|d_j^{obs}(t) - d_j(\boldsymbol{m}, t)\|^2 dt, j \in \{1, 2, .., r\}. \tag{12}$$

The predicted mean and covariance before measurement update are denoted as $\hat{\boldsymbol{m}}$ and $\boldsymbol{P}^m$ respectively. If $n$ model parameters are to be estimated, $2n + 1$ sigma-points are defined as follows for the unscented transformation to approximate a Gaussian distribution of the predicted estimate centered at $\hat{\boldsymbol{m}}$ with covariance $\boldsymbol{P}^m$:

$$\boldsymbol{\mathcal{M}}_0 = \hat{\boldsymbol{m}}, \tag{13}$$

$$\boldsymbol{\mathcal{M}}_i = \hat{\boldsymbol{m}} + \left(\sqrt{(n+\lambda)\boldsymbol{P}^m}\right)_i, \text{for } i = 1 : n \tag{14}$$

$$\boldsymbol{\mathcal{M}}_{n+i} = \hat{\boldsymbol{m}} - \left(\sqrt{(n+\lambda)\boldsymbol{P}^m}\right)_i, \text{for } i = 1 : n. \tag{15}$$

The parameter $\lambda$ remains free for tuning in order to adjust the spread of the sigma-points about the predicted mean estimate in directions of principle variances. It is suggested to select $n+\lambda = 3$ if $\boldsymbol{m}$ is Gaussian [8]. The notation $(\cdot)_i$ denotes the $i$-column of the matrix within parentheses. Each sigma-point is associated with a weight defined as in Eq. 16.

$$W_0 = \frac{\lambda}{n+\lambda}, W_i = W_{i+n} = \frac{1}{2(n+\lambda)}, \text{for } i = 1:n. \tag{16}$$

The misfit measures for all $r$ output measurements are obtained by mapping the sigma-points through the misfit functional of Eq. 12. The mapping for $r$ multiple receivers $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)$, $\boldsymbol{s} : \mathbb{R}^n \mapsto \mathbb{R}^r$, is done for each sigma-point as written in Eq. 17.

$$\boldsymbol{\mathcal{S}}_i = \boldsymbol{s}\left(\mathcal{M}_i\right), \text{for } i = 0:2n. \tag{17}$$

After the nonlinear mapping of the sigma-points through the nonlinear model, the mean of the predicted misfit $\hat{\boldsymbol{s}}$, covariance of the predicted parameters $\boldsymbol{P}^m$, covariance of the misfit at the predicted model $\boldsymbol{P}^s$, and cross-covariance between them $\boldsymbol{P}^{ms}$ are approximated by weighted summation rule as follows:

$$\hat{\boldsymbol{s}} = \sum_{i=0}^{2n} W_i \boldsymbol{\mathcal{S}}_i, \tag{18}$$

$$\boldsymbol{P}^m = \sum_{i=0}^{2n} W_i \left(\boldsymbol{\mathcal{M}}_i - \hat{\boldsymbol{m}}\right)\left(\boldsymbol{\mathcal{M}}_i - \hat{\boldsymbol{m}}\right)^t + \boldsymbol{Q}, \tag{19}$$

$$\boldsymbol{P}^s = \sum_{i=0}^{2n} W_i \left(\boldsymbol{\mathcal{S}}_i - \hat{\boldsymbol{s}}\right)\left(\boldsymbol{\mathcal{S}}_i - \hat{\boldsymbol{s}}\right)^t + \boldsymbol{R}, \tag{20}$$

$$\boldsymbol{P}^{ms} = \sum_{i=0}^{2n} W_i \left(\boldsymbol{\mathcal{M}}_i - \hat{\boldsymbol{m}}\right)\left(\boldsymbol{\mathcal{S}}_i - \hat{\boldsymbol{s}}\right)^t. \tag{21}$$

Posterior mean $\hat{\boldsymbol{m}}_+$ and covariance $\boldsymbol{P}^m_+$ of the estimated model parameters are updated following the Kalman filter's measurement update step:

$$\hat{\boldsymbol{m}}_+ = \hat{\boldsymbol{m}} + \boldsymbol{K}\left(\boldsymbol{0} - \hat{\boldsymbol{s}}\right) \tag{22}$$

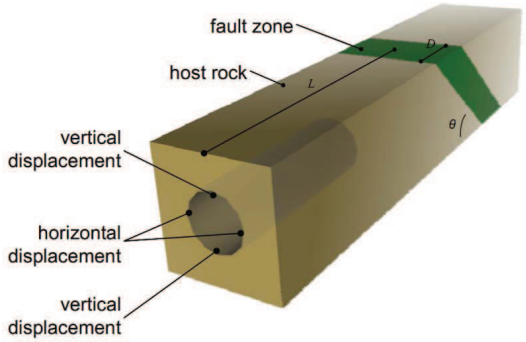
$$\boldsymbol{P}^m_+ = \boldsymbol{P}^m - \boldsymbol{K}\boldsymbol{P}^s\boldsymbol{K}^t \tag{23}$$

with the Kalman gain $\boldsymbol{K}$ calculated as

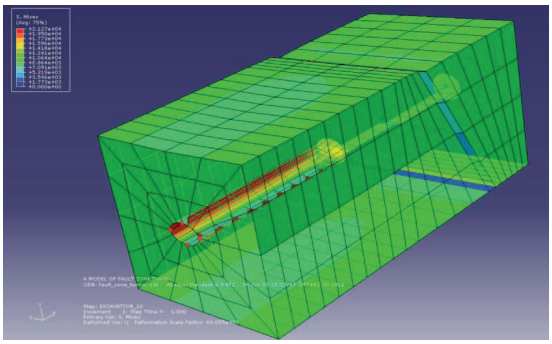
$$\boldsymbol{K} = \boldsymbol{P}^{ms}\left(\boldsymbol{P}^s\right)^{-1}. \tag{24}$$

To initialize the UKF procedure, some prior knowledge about the model is provided in form of Gaussian distribution $\rho_0(\boldsymbol{m}) \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{P}_0)$. In the first iteration of the UKF, this prior information is assigned to the predicted mean and covariance: $\hat{\boldsymbol{m}} = \boldsymbol{m}_0$, and $\boldsymbol{P} = \boldsymbol{P}_0$. In the successive iterations, the predicted estimated mean and covariance are updated by taking the respective values from the posterior estimate, i.e. $\hat{\boldsymbol{m}} = \boldsymbol{m}_+$ and $\boldsymbol{P} = \boldsymbol{P}_+$.

Described UKF procedure for minimizing the misfit functional locally has been successfully implemented within the hybrid algorithm combined with simulated annealing in the global optimization procedure with implementation in full waves inversion [11].




a)



b)

Figure 1: a) 3D tunnel excavation with a fault zone ahead of tunnel face [12] and b) FE model

| | Host rock | Fault zone |
|---|---|---|
| $E[kN/m^2]$ | $2 \times 10^8$ | $5 \times 10^7$ |
| $\nu$ | 0.2 | 0.35 |
| $\rho[kg/m^3]$ | 2500 | 2200 |

Table 1: Material parameters

## 3 IMPLEMENTATION OF KALMAN FILTERS

Following numeric examples demonstrate efficient implementation of the inversion methods based on Kalman filtering techniques for the parmeter estimation in geotechnical and tunneling problems.

### 3.1 3D tunnel excavation: EKF identification of geometrical and elastic material parameters

In this numeric example the EKF is used for estimation of both material and geometrical parameters of a 3D tunnel excavation model [12] depicted in Figure 1 a). The model represents a mountain cut-out with overburden height of 500 meters. Dimensions of the model are $100m \times 100m \times 200m$. A fault zone is located at $L$ meters from the model frontal face measured from its top edge having a thickness of $D$ meters. The fault dip direction and the angle of $66^\circ$ of inclination are supposedly known from an engineering-geology investigation.

The material parameters are given in Table 1. The initial geostatic stresses are imposed with prescribed horizontal stress to vertical stress ratio equal to 1, i.e. $\sigma_h = \sigma_v$. To account for this model feature, a surface pressure of $1.23 \times 10^4\ kN/m^2$ is applied on the top side of the model. This value is calculated assuming the overburden rock mass to be homogeneous, with a density of 2500 $kg/m^3$. The displacement history data are collected during the simulation of 10 tunnel excavations, each of 10 meters length. Observation points are located at the tunnel crown, bottom and at the side walls. A 3D finite element tunnel excavation model built with ABAQUS is presented in Figure 1 b). The elements in the tunnel are 6-node triangular prism elements (5-sided), and rest of the soil body elements are 8-node brick elements (6-sided). The state vector contains three components: one material parameter – Young's modulus *E* and two geometrical parameters – *L* and *D* as depicted in Figure 1.

$$x = \left[ \begin{array}{ccc} L[m] & D[m] & E[kN/m^2] \end{array} \right]$$

Convergence of the algorithm was investigated for different initial state estimate vectors and it

was shown that the convergence rate varies with the initial guess. Results for two cases of initial state vectors are presented in Tables 2 and 3.

| Parameter | Initial state | Estimated state | Deviation from true state |
|---|---|---|---|
| L[m] | 130 | 119.70 | 0.30 |
| D[m] | 10 | 6.01 | 0.01 |
| E[$kN/m^2$] | $4.0 \times 10^7$ | $5.03 \times 10^7$ | $0.03 \times 10^7$ |

Table 2: Converged result after 15 iterations - Case 1

| Parameter | Initial state | Estimated state | Deviation from true state |
|---|---|---|---|
| L[m] | 115 | 120.32 | 0.32 |
| D[m] | 4 | 5.99 | 0.01 |
| E[$kN/m^2$] | $4.0 \times 10^7$ | $4.96 \times 10^7$ | $0.04 \times 10^7$ |

Table 3: Converged result after 20 iterations - Case 2

Graphical results of the progressing Kalman filter for Case 1 are shown in Figure 2.

### 3.2 Implementation of SPKF for model calibration of soil parameters in mechanized tunneling

Geomechanical analysis of tunnel excavation in this example can be considered as a quasi-static non-Markov process, i.e. the state of the model is about equilibrium along the analysis and the current state depends not only on the previous state but also on the history of analysis steps. The model of a shield tunnel boring machine (TBM) advancing in homogeneous soil, Fig. 3, obtained using the FEA software PLAXIS 3D was provided by the C2-subproject within the Collaborative Research Center SFB 837 at the Ruhr University Bochum [14]. The model domain is 60 m long, 40 m wide, and 45 m deep. The tunnel of diameter D = 8.5 m and under an overburden depth equal to 1D. Except that the ground surface is stress-free, all other boundaries of the FE model are constrained by zero displacements. Only a half of the model is analyzed because of the symmetric geometrical, loading and material conditions about the vertical plane along the tunnel axis. The homogeneous soil behaves according to the Hardening Soil constitutive model in [13].

The set of parameters chosen for identification, with their assumed 'true' values for the forward simulation includes stiffness for un-/reloading $G_{ur} = 41600[kN/m^2]$, secant stiffness in standard drained triaxial test $E_{50} = 35000[kN/m^2]$, and friction angle $\phi = 35[°]$. Observation positions and orientations at observation surface at 39 m from beginning position of the tunnel head (before excavation) and perpendicular to the tunnel axis are depicted in Fig. 3. The model is calculated for 30 excavation steps, each step advances the TBM 1.5 m forward. Calculated outputs obtained from the FE model are registered at every excavation step. Synthetic measurement data are assumed with additive Gaussian noise with standard deviations of 3% and 5% the mean value of the noise-free FE calculated outputs for each measurement point. The feasible ranges for the considered soil parameters are defined such that the lower bounds and upper bounds are considerably distant from true parameters: $20833 \leq G_{ur}$[kN/m$^2$] $\leq 62499$, $17500 \leq E_{50}$[kN/m$^2$] $\leq 52500$ and $32 \leq \phi[°] \leq 52.5$. Taking into account the assumption

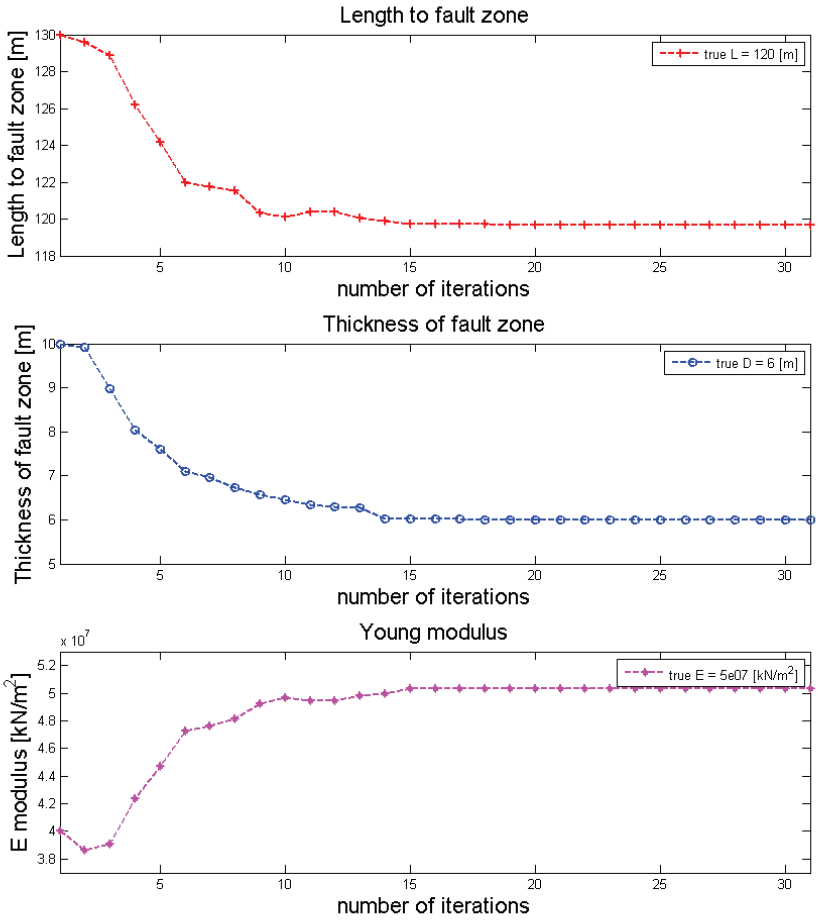



Figure 2: Graphical result convergence of the estimated parameters - initial case 1

that the parameters are uncorrelated, the initial state error covariance matrix for the parameters $\mathbf{m} = [G_{ur}; \phi; E_{50}]$ is set: $\mathbf{P}(0|0) = \text{diag}\left(4000^2, 5^2, 5000^2\right)$. The authors' practice for setting the covariance matrix of the process noise $\mathbf{Q}$ is that it can be set in such a way that its square roots are ranging from five percent to ten percent the values of initial estimation error standard deviations. In this work, it is chosen to be around seven percent, i.e. the covariance matrix of the process noise can be set as $\mathbf{Q} = 0.005\mathbf{P}_0$. For setting up the nonlinear Kalman filters in this study, every measurement datum is assigned to have the same covariance, i.e. $\mathbf{R} = 0.01^2 * \mathbf{I}(m, m)$, with $\mathbf{I}(m, m)$ the identity matrix having dimension equal to the size of the measurement data, $m$.

Selected comparison results of the EKF and SPKF with initial parameters chosen at the extremes of feasible ranges are presented in Table 4 and Table 5 for cases of 3% and 5% measurement noise, respectively.

Graphical plots of the converging soil parameters in Fig. 4 demonstrate fast and robust convergence. The investigation has shown, that depending on the initially selected parameter values, the convergence of the SPKF can require less itteration steps than for the EKF.

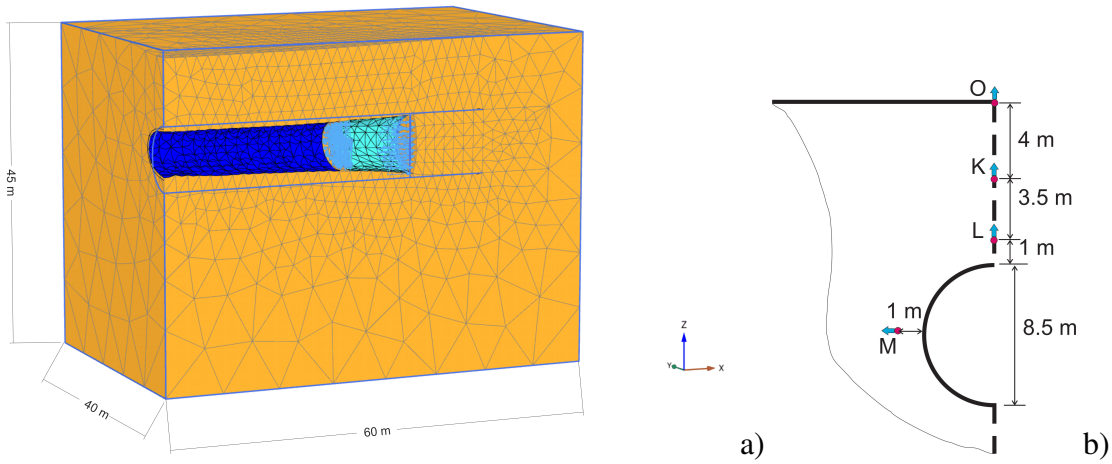



Figure 3: a) FE model of the 3D tunnel excavation and b) Measurement points at observation section

| Parameter | Initial | | EKF | | SPKF | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| $G_{ur}$[kN/m$^2$] | 20833 | 4000 | 41588.02 | 552.30 | 41476.61 | 296.64 |
| $E_{50}$[kN/m$^2$] | 17500 | 5000 | 34856.67 | 958.45 | 34643.40 | 387.50 |
| $\phi$[°] | 32 | 5 | 34.96 | 0.25 | 34.91 | 0.35 |

Table 4: Identified soil parameters with 3% measurement noise

| Parameter | Initial | | EKF | | SPKF | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| $G_{ur}$[kN/m$^2$] | 20833 | 4000 | 42483.47 | 6317.98 | 42288.64 | 392.62 |
| $E_{50}$[kN/m$^2$] | 17500 | 5000 | 37036.86 | 1130.50 | 36646.55 | 600.89 |
| $\phi$[°] | 32 | 5 | 35.52 | 2.82 | 35.42 | 0.37 |

Table 5: Identified soil parameters with 5% measurement noise

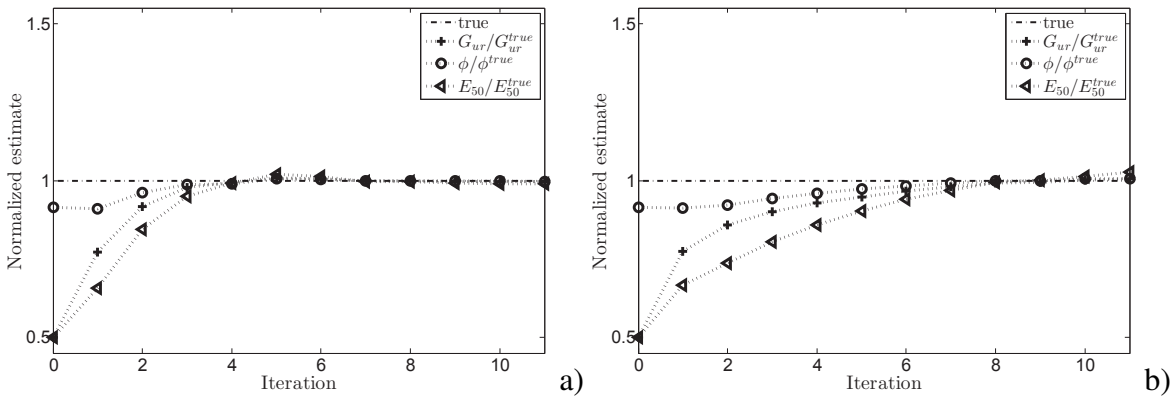



Figure 4: Iterative development of soil parameters with a) 3% and b) 5% measurement noise SPKF

## 4 Conclusions

Good convergence of the parameter estimation shown through examples of geotechnical problems presented in this work has proven that the Kalman filtering techniques have prospects to be a reliable method for parameter identification in geotechnical engineering. An important advantage of the Kalman filters, utilization of the measurement uncertainties, which cannot be avoided in large scale geotechnical structures, makes it a very attractive and powerful estimation method, which performs well even in the environment corrupted by noise. Another prominent advantage is that the choice of initial model parameters must not be very close to the true parameter set as for other gradient-based optimization methods. Implementation of the Kalman filters, relating the implementation effort in comparison with the resulting outcome, represents relatively simple, but still a very powerful tool, since using modern programming languages the Kalman filter algorithm can be realized within several code lines including vector and matrix manipulations. However, the tuning and adjusting the filter's configurations plays a decisive role regarding its successful implementation and therefore requires a special attention. Good understanding of the model and the influence of each parameter on the model is crucial for the configuration of the Kalman filter in order to achieve its successful convergence.

### Acknowledgements

The authors gratefully acknowledge the funding by the German Research Foundation (DFG) within Collaborative Research Center under grant SFB-837/A2.

## REFERENCES

[1] R.E. Kalman, A New Approach to Linear Filtering and Prediction Problems, *Trans. ASME*, Ser. D: J. Basic Eng. **82**, 35-45, 1960.

[2] M. Hoshiya, A. Suto, Kalman Filter - Finite Element Method in Identification, *Journal of Engineering Mechanics* **119**:2, 197-210, 1993.

[3] A. Murakami, *Studies on the Application of Kalman Filtering to some Geomechanical Problems Related to Safety Assessment*, doctoral dissertation, Kyoto University, 1991.

[4] Wang Yang, *Extended Kalman Filter Method for Parameter Identification and its Application in Tunneling*, doctoral dissertation, Universität Innsbruck, 2004.

[5] L.T. Nguyen, M. Datcheva, T. Nestorović, Identification of a fault zone ahead of the tunnel excavation face using the extended Kalman filter, *Mechanics Research Communications*, **53**, 47-52, 2003.

[6] A. Gelb (ed.), *Applied Optimal Estimation*, MIT Press, Cambridge, Massachusetts, 1974.

[7] D. Simon, *Optimal State Estimation - Kalman, $H_\infty$ and Nonlinear Approaches*. John Wiley & Sons, 2006.

[8] S.J. Julier, J.K. Uhlmann, A new extension of the Kalman filter to nonlinear systems, *Proc. AeroSense: 11th Int. Symp. Aerospace/Defense Sensing, Simulation and Controls*, 182-193, 1997.

[9] R. Van Der Merwe, *Sigma-point Kalman filters for probabilistic inference in dynamic state-space models*, doctoral dissertation, Oregon Health & Science University, 2004.

[10] R. Kandepu, L. Imsland, B.A. Foss, Constrained state estimation using the unscented Kalman filter, *Proceedings of the 16th Mediterranean Conference on Control and Automation*, 1453-1458, 2008.

[11] L.T. Nguyen, T. Nestorović, Unscented hybrid simulated annealing for fast inversion of tunnel seismic waves, *Computer Methods in Applied Mechanics and Engineering*, **31**, 281-299, 2016.

[12] J. Meier, M. Datcheva, T. Schanz, Identification of Constitutive and Geometrical Parameters of Numerical Models with Application in Tunneling, *EURO:TUN, Vienna*, 2007.

[13] T. Schanz, P.A. Vermeer, P.G. Bonnier, The hardening-soil model: Formulation and verification, *Beyond 2000 in Computational Geotechnics*, Balkema, Rotterdam, 281-290, 1999.

[14] V. Zarev, T. Schanz, M. Datcheva, Three-dimensional finite-element analysis of mechanized shield tunneling in urban areas, *Proceedings of The Thirteenth International Conference on Civil, Structural and Environmental Engineering Computing*, Greece 1-10, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EVALUATION OF DIFFERENT APPROACHES FOR THE OPTIMIZATION OF LAYOUT AND CONTROL OF BOOSTER STATIONS

**Philipp Pöttgen, and Peter F. Pelz**

Technische Universität Darmstadt, Chair of Fluid Systems
Otto-Berndt-Straße 2, 64283 Darmstadt, Germany
{philipp.poettgen, peter.pelz}@fst.tu-darmstadt.de

**Keywords:** Booster Station, Booster System, Energy Efficiency, Optimization, Technical Operations Research.

**Abstract.** *Pumps and pumping systems consume about 12 % of the annual electricity production in Europe. A parallel arrangement of two or more pumps is called "booster station". Booster stations meet a varying pressure demand with high energy-efficiency by deactivating individual pumps at smaller loads. Hence, one major difference from single pump units to booster stations is the diversity of control options. Due to the non-linear characteristics of the machines and physical laws we are facing a Mixed-Integer Nonlinear Problem (MINLP). The global optimization of a MINLP requires a specialized solver and a huge amount of time. To reduce the complexity of the problem, we evaluate two possible ways: i) We apply piecewise linearization techniques to the problem in order to gain a Mixed Integer Linear Problem (MILP). The advantage is the vast variety of available solvers and a significant reduction of calculation time. The disadvantage of this method is being less accurate. ii) The enumeration of the integer decisions and the fitting of the characteristic curves by algebraic functions enable us to reduce the complexity of the problem to a simple Nonlinear Problem (NLP). This technique allows us to keep the accuracy of the physical laws, but has the big drawback of the enumeration time. The consecutive use of two global optimization solvers combines the advantages of the formerly mentioned approaches, but we lose the guarantee for global optimality.*

# 1 INTRODUCTION

Pumps and pumping systems consume about 10 % of the annual electricity production in Europe [1]. The European Union aims for the usage of energy efficient machines to gain energy savings. These energy efficient machines are the first step towards reducing the energy consumption of pumping systems. Various eco-design directives have been and will be carried out in order to ban machines with low efficiency from the market [2].

The second step must be the efficient use of these excellent machines. A parallel arrangement of two or more pumps is called “booster station”. Booster stations meet a varying pressure demand with high energy-efficiency by deactivating individual pumps at smaller loads. They cover a wide performance map and are flexible machines with widespread possibilities of usage in industrial and residential applications, e.g. chemical processing plants or water supply in skyscrapers.

The design and control of a booster station offer manifold degrees of freedom. The power input of a booster station in one point of duty depends on more than one design or control variable. Global optimization programs offer help for tuning the mixture of continuous and discrete variables, so that the power input of the booster station becomes minimal.

The aim of our research is to make optimization programs available for engineering use. A modeling framework and exchangeable input data enable us to quickly gain a control strategy and design information from measured characteristics.

# 2 TECHNICAL SPECIFICATIONS

## 2.1 Working Principle of a Booster Station

The technical task of a booster station is to promote water in a piping network. Therefore the pressure is increased to overcome geodesic differences or pressure loss by dissipation. Figure 1 shows the connection scheme of the booster station. Multiple pumps, in most applications of the same type, are connected in parallel. The incoming water flows from the suction pipe, through the single pump units, into the pressure pipe. The pressure at the outlet is always higher than at the inlet. The load of the booster station is given by the total volume flow $Q_{tot}$ and the pressure head $\Delta H$. To avoid reverse flow, a check valve is installed behind each pump.

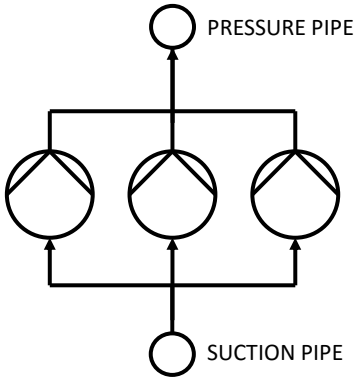



Figure 1: Connection scheme of a booster station.

The parallel arrangement has two major reasons: Firstly, one is able to avoid heavy part load in the pumps by deactivating single pumps. Secondly, if one pump fails, it is still possible to use the others for promoting water. Other topologies, especially the direct connection of pumps in series, are generally possible and useful [3], but the application of a control strategy in between the considered points of duty is very difficult. Thus, we do not consider this topological option in our paper.

## 2.2 Pump Characteristics

To describe a pump's field of operation, four parameters are relevant: The volume flow $Q$, the pressure head $\Delta H$, the rotational speed $n$ and the power input $P$. The actual point of duty is determined exactly by any two of these parameters for strictly monotonic characteristics. This condition is fulfilled for the relevant operating range for any pump and in this paper for the whole operating range. Figure 2 shows the characteristics of a single pump.

The manufacturer describes the characteristics in his catalogue based on measurement data. For the reference rotational speed $n_{ref}$ the pressure head and volume flow of the pump are varied in a test rig and the values as well as the corresponding power input are measured. Following industrial standards the unit to measure the pressure head is meter water column ($mWC$). Table 1 gives the measurement points for the pump type A considered in this paper.

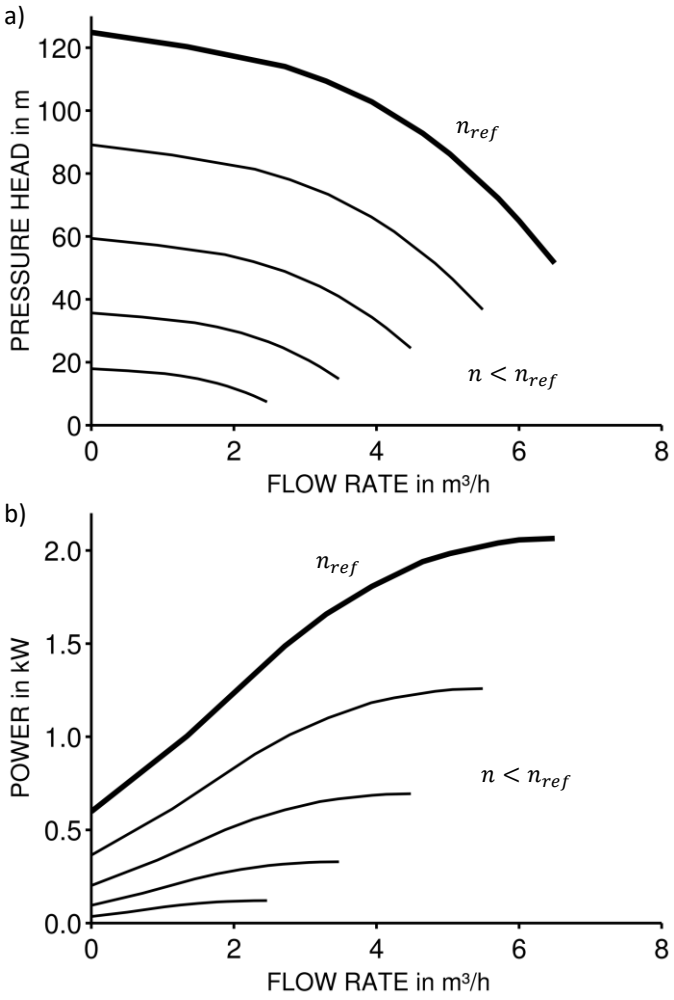



Figure 2: Characteristics of pump type A.

| Flow Rate in $m^3\ h^{-1}$ | Pressure head in mWC | Power Input in kW |
|---|---|---|
| 0.0000 | 124.87 | 0.5981 |
| 1.3421 | 120.31 | 1.0031 |
| 2.7176 | 114.00 | 1.4878 |
| 3.2971 | 109.33 | 1.6588 |
| 3.9364 | 102.76 | 1.8072 |
| 4.6499 | 92.78 | 1.9402 |
| 5.0308 | 86.19 | 1.9832 |
| 5.7083 | 72.11 | 2.0403 |
| 6.0010 | 64.99 | 2.0567 |
| 6.5000 | 51.49 | 2.0649 |

Table 1: Measurement points for pump type A.

Many market available pumps are able to alter their rotational speed in the interval $n_{min} \leq n \leq n_{max}$. To calculate the characteristics for altering rotational speed, the scaling laws hold:

$$Q(n) = \left(\frac{n}{n_{ref}}\right) Q_{ref}, \tag{1}$$

$$\Delta H(n) = \left(\frac{n}{n_{ref}}\right)^2 \Delta H_{ref}, \tag{2}$$

$$P(n) = \left(\frac{n}{n_{ref}}\right)^3 P_{ref} \frac{1 - \left(1 - \eta_{opt}\right)\left(\frac{n}{n_{ref}}\right)^{-0.1}}{\eta_{opt}}. \tag{3}$$

The scaling law for the power characteristics is expanded by an empirical correction term in order to account for the efficiency decrease due to decreasing rotational speed, wherein $\eta_{opt}$ stands for the maximum hydraulic efficiency at reference speed. Usually, but not necessarily, the maximal rotational speed of a pump is used as reference value.

## 2.3 Operational Mode

The control variable for the operation of the booster station is the pressure in the pressure pipe. The controller adjusts the rotational speed of the pumps and switches pumps on and off. While the rotational speed is a continuous variable which could be controlled, e.g. by a PID controller, the operational status is a discrete variable that causes a discontinuous transition. To react on this discontinuity the designer implements a control strategy.

This strategy may be devised by many different approaches and with various objectives. A simple approach takes the maximum rotational speed of a pump into account. All active pumps are running at the same rotational speed. Whenever the active pumps cannot fulfil the load at maximum rotational speed the controller turns on an additional pump and adjusts the rotational speed of all pumps. This simple approach guarantees the functionality of the system, but does not consider the power input.

Another option considers the power input. A pump's efficiency is not constant over the volume flow. The degrees of freedom for the control of the booster station are the number of active pumps and the rotational speed of any active pump. The pressure difference over all active pumps must be equal. The total volume flow is distributed between the pumps depend-

ing on the pump type and the rotational speeds. We will optimize the setting of the control variables, so that the total power input $P_{tot}$ of the system becomes minimal.

## 3 MATHEMATICAL MODEL

We present different approaches for calculating a design as well as a control strategy for booster stations from the characteristics of the machines. While the rotational speed of a pump is controlled continuously within the technical available range, the topological layout and number of activated pumps are discrete decisions. Due to the non-linear characteristics of the machines and physical laws we are facing a Mixed-Integer Nonlinear Problem (MINLP).

To simplify the MINLP in order to gain a quicker solution we examine two different approaches: We apply piecewise linearization techniques to the problem in order to gain a Mixed-Integer-Linear-Problem (MILP). In the second approach the fixation of the integer decisions and the fitting of the characteristic curves by algebraic functions enable us to reduce the complexity of the problem to a simple Nonlinear Problem (NLP).

### 3.1 Pump Characteristics

Appropriate modelling of pump characteristics is very important for the optimization program. From the characteristics measurement we gain several discrete data points for the operation of the pump at one specific rotational speed. Thus, we have to estimate values in between these points and to expand the data to create the field of operation. In general we have two different possibilities for this estimation: (i) A fitting function or (ii) piecewise linear interpolation. Both options are described in detail in the following:

(i) We use a polynomial function 2nd to 5th degree to fit the reference curve of the pump. The higher the degree of the polynomial function, the better the measured points and the shape of the curve is met. The lower the degree of the polynomial function, the simpler becomes the resulting optimization program. For the pump in our study, we decided for a 3rd degree polynomial to model the pressure characteristics and a 4th degree polynomial to model the power characteristics. The polynomial coefficients are given in Table 2:

$$\Delta H_{ref}(Q_{ref}) = a_H {Q_{ref}}^3 + b_H {Q_{ref}}^2 + c_H Q_{ref} + d_H, \tag{4}$$

$$P_{ref}(Q_{ref}) = a_P {Q_{ref}}^4 + b_P {Q_{ref}}^3 + c_P {Q_{ref}}^2 + d_P Q_{ref} + e_P. \tag{5}$$

| Coefficient | | Value | Coefficient | | Value |
|---|---|---|---|---|---|
| $a_H$ | in mWC $h^3$ $m^{-9}$ | -0.2448 | $a_P$ | in kW $h^4$ $m^{-12}$ | 0.001357 |
| $b_H$ | in mWC $h^2$ $m^{-6}$ | 0.3421 | $b_P$ | in kW $h^3$ $m^{-9}$ | -0.02259 |
| $c_H$ | in mWC h $m^{-3}$ | -3.197 | $c_P$ | in kW $h^2$ $m^{-6}$ | 0.09047 |
| $d_H$ | in mWC | 124.9 | $d_P$ | in kW $h^1$ $m^{-3}$ | 0.2196 |
| | | | $e_P$ | in kW | 0.59811 |

Table 2: Polynomial coefficients for pump type A.

For altering the rotational speed the scaling laws (equations (2) and (3)) hold. By using additionally equation (1) we derive:

$$\Delta H(Q,n) = \left(\frac{n_{ref}}{n}\right) a_H Q^3 + b_H Q^2 + \left(\frac{n}{n_{ref}}\right) c_H Q + \left(\frac{n}{n_{ref}}\right)^2 d_H, \tag{6}$$

$$P(Q,n) = \left[\left(\frac{n_{ref}}{n}\right) a_P Q^4 + b_P Q^3 + \left(\frac{n}{n_{ref}}\right) c_P Q^2 + \left(\frac{n}{n_{ref}}\right)^2 d_P Q + \left(\frac{n}{n_{ref}}\right)^3 e_P\right] \frac{1-\left(1-\eta_{opt}\right)\left(\frac{n}{n_{ref}}\right)^{-0.1}}{\eta_{opt}}. \quad (7)$$

Due to the shape of the characteristic curves of a pump (cf. Figure 2) and the nonlinear scaling laws this technique necessarily leads to a nonlinear model (MINLP or NLP).

(ii) The reference measurements for the pump are given as triples $(Q_{ref}, \Delta H_{ref}, P_{ref})$ for the reference rotational speed. For altering the rotational speed the scaling laws hold once again. But in this case we discretize the rotational speed on the permitted interval and explicitly scale the triples of measurement points for these discrete values. As stated in Chapter 2.2, only two of the four values for volume flow, pressure head, power input and rotational speed are independent from each other. If the point of duty is fixed in the $Q$-$\Delta H$-plane, the values for rotational speed and power input are directly dependent.

To interpolate between the points we use the aggregated convex combination (ACC) described by Vielma et al. [4]. The closer the intervals between the discrete values for the rotational speed are, the smaller is the possible linearization error, but the larger becomes the optimization problem.

The choice for a pump model directly influences the possible options for the final optimization program: Linear interpolation in an optimization program requires the use of binary or integer variables. Hence the resulting problem will include discrete variables (MINLP or MILP).

### 3.2 Flow Model

We model fluid systems as a mathematical flow graph $G(V, E)$. The edges $E$ of the graph represent technical components, in this particular case the pumps $K$, or simple connections. The vertices $V$ are used to connect the edges [5]. Each edge transfers a volume flow $Q$. In each vertex the pressure $H$ is calculated. For every pump edge a power input $P$ and the rotational speed $n$ is calculated additionally.

Depending on the optimization scope and the program type, two different formulations for the load case are applicable: (i) each pump type is represented as one edge of the graph or (ii) each pump is represented by an edge of the graph.

(i) If one edge of the flow graph represents a number of pumps of the same type, one further premise applies: The rotational speed for all pumps of this type is the same. We introduce an integer variable $x$ for each pump type that represents the number of active pumps for this pump type and modify the requirement for the load case:

$$Q_{tot} = \sum_{\{i \text{ in } K\}} x_i Q_i. \quad (8)$$

Due to the parallel arrangement the volume flow for all pumps of one type equals the volume flow for a single pump multiplied by $x$. The required pressure head of the booster following from the given input and output pressure is not influenced by the variable $x$ and still accounts for all pumps. Parallel pumps must have the same pressure head.

$$\forall i \in K: H_{i,input} + \Delta H_i = H_{i,output} \tag{9}$$

If the range of $x$ includes 0, the pump type can be deactivated. The pressure head and the rotational speed, following from the pump characteristics model, will still be calculated, but the energy consumption of the pump vanishes.

The correlation of the power input for multiple pumps is the same as of the volume flow and also uses variable $x$.

$$P_{tot} = \sum_{\{i\ in\ K\}} x_i P_i \tag{10}$$

For purely nonlinear problem formulations without integer variables (NLP) the variable $x$ may be transformed into a parameter. Enumeration is a possible approach to find the optimal number of active pumps. In this case only the rotational speeds for the different pump types are the remaining degrees of freedom. For booster stations with only one pump type the optimization problem becomes a feasibility problem.

(ii) The booster station has to promote the total volume flow $Q_{tot}$, which is given as input parameter, as the sum of the single volume flows in each pump:

$$Q_{tot} = \sum_{\{i\ in\ K\}} Q_i \tag{11}$$

Further input parameters for the point of duty are the input pressure and the required pressure head for the booster station. The pressure head (equation (9)) for the booster station accounts for any active pump. To deactivate a specific pump, we introduce a binary activity variable $y$ and use a Big-M-Formulation to modify equation (9) and decouple the pressure at the inlet and outlet of the pump.

$$\begin{aligned} H_{i,input} + \Delta H_i &\leq H_{i,output} + M(1-y) \quad \forall i \in K \\ H_{i,input} + \Delta H_i &\geq H_{i,output} - M(1-y) \quad \forall i \in K \end{aligned} \tag{12}$$

A deactivated pump ($y_i = 0$) has no volume flow or power input. This is realized by forcing all weights $\lambda$ of the linearization (compare [4]) to be zero. Thus the total energy input reads

$$P_{tot} = \sum_{\{i\ in\ K\}} P_i. \tag{13}$$

Due to the use of the binary variable, the application of a Big-M-Formulation is only possible in an approach which allows discontinuous variables (MINLP or MILP).

Please note: In the MINLP case, the deactivation model for single pumps is also possible in a manner similar to the approach described in (i): Instead of the integer variable $x$ the binary activity variable $y_M$ is used to modify the flow and the power equation of each pump. Pressure rise and rotational speed will be calculated also for the deactivation in this case, but do not influence the optimization program and result. The rotational speed and the power input for all pumps follow from model for the pump characteristics.

If binary variables must be avoided to use the NLP approach, the remaining degrees of freedom for the optimization are the continuous variables for the rotational speed. To find the optimal number of active pumps for this case, an enumeration is a possible approach.

### 3.3 Objective

The objective for the optimization program is to minimize the total power input of all pumps for the given point of duty.

$$min\ P_{tot} \tag{14}$$

### 3.4 Optimization Framework

In order to quickly gain solutions for optimization problems of booster related tasks, we created an optimization framework within the AIMMS software system. The parameters for the optimization, such as different pump types as well as their characteristics or maximum number of pumps, are given as user input, but remain the same for any optimization program. From the user input the flow graph is generated. Furthermore, the framework includes the constraints and variables for all different modelling options. The user chooses which type of optimization program he would like to use and the computer compiles only the relevant part of the model for the user's choice. Afterwards the resulting optimization program is passed to a matching solver. This framework enables us to compare the results for the different modelling approaches and compare them to each other.

The described optimization program finds an optimal control solution for one point of duty. In order to find control guidelines for the whole field of operation, we have to run the optimization model many times with varying load cases. We discretize the pressure and the volume flow and run the program for every point of duty.

## 4 RESULTS

### 4.1 Control Strategy for Booster Stations with one Pump Type

We compare three different approaches to optimize the control strategy of a booster station consisting of three pumps of type A. The first approach is a MILP: We model the pump characteristics by an ACC linearization with 20 intervals for the rotational speed and create one edge for each pump. The second model is an NLP approach, which uses a fitted pump model and one edge for each pump type. The discrete decision for the optimal number of pumps follows from an enumeration and comparison of the results. The third model is a MINLP with the same model as the NLP. Instead of the enumeration the Solver additionally makes a discrete decision for the number of active pumps.

The result for the discrete decision for the number of pumps is the same with all three approaches. Figure 3 shows the optimal control strategy for the booster station. We can identify two reasons for the switch of a pump: (1) For high constant pressure and increasing volume flow the controller increases the rotational speed of the active pumps. Once the maximum is reached, an additional pump is switched on. The decreasing gray lines show the maximum volume flow for fixed pressure head for one, two or three pumps. (2) For low constant pressure and increasing volume flow we find an efficiency argument: An additional pump is switched on, because it reduces the total power input, even though more pumps are working.

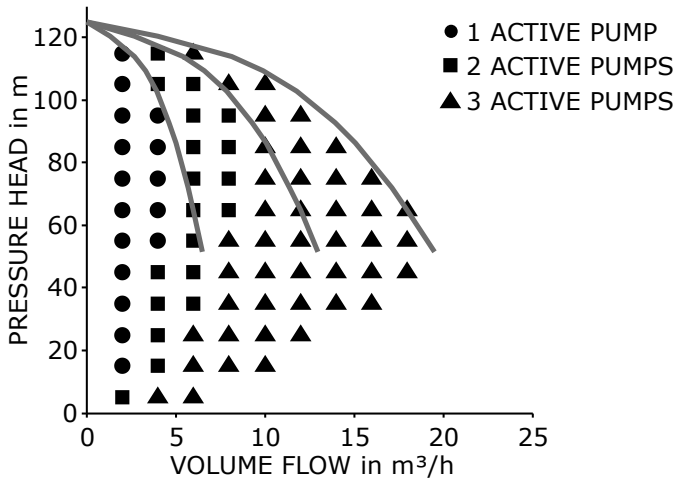



Figure 3: Optimization result for a booster station of three pumps of type A.

All calculations were made on an AMD Phenom™ II X6 1100T Processor with 3.3 GHz and 16 GB RAM. The applied solvers were CPLEX 12.6.2 for the linear model and BARON 15 for the nonlinear model. The sum of the run times for the different approaches is:

- MILP: 13.5 s
- MINLP: 25.2 s
- NLP + Enumeration: 51.6 s

The MILP is clearly the fastest, while, as expected, the NLP takes the most time. Even though the discrete decision of all the approaches remains the same, the solutions are slightly different: In the MILP results the rotational speed of the active pumps is not always exactly the same. For $\Delta H = 75\,mWC$ and $Q = 12\,m^3/h$ the rotational speeds for the three active pumps are $n_1 = 2607\,min^{-1}, n_2 = 2521\,min^{-1}, n_3 = 2607\,min^{-1}$. Even though it's proven, that all these pumps should run at the same rotational speed [6], the optimality gap is small enough to stop the solver. Highest settings for the numerical precision of the solver can't change this. Additionally, the linearization error can only be calculated and reduced [7] but not avoided for the optimization. A nonlinear correction would be necessary to use the results from this approach. The necessary correction is a feasibility problem (see 3.2).

### 4.2 Control Strategy for Booster Stations with two Pump Types

The next considered optimization problem is a booster station with three pumps of type A and one pump of different type. Table 3 shows the measurement points for pump type B. Once again we use the MILP and the MINLP approach as described in 4.1. The NLP approach is not tested anymore, due to the bad performance for the first test case and the increased number of decisions.

| Flow Rate in $m^3\ h^{-1}$ | Pressure head in mWC | Power Input in kW |
|---|---|---|
| 0.0000 | 123.86 | 0.4597 |
| 0.6000 | 120.80 | 0.6289 |
| 1.2215 | 114.09 | 0.7930 |
| 1.6223 | 107.30 | 0.9069 |
| 1.8152 | 103.55 | 0.9673 |
| 1.9219 | 101.30 | 0.9887 |
| 2.0529 | 86.11 | 1.0242 |
| 2.4428 | 72.11 | 1.1013 |
| 2.6902 | 75.95 | 1.1334 |
| 3.0266 | 59.32 | 1.1545 |

Table 3: Measurement points for pump type B.

The computer and the solvers stay the same. The summed solver times for the solution of the new problem are:

- MILP: 20.0 s
- MINLP: 654.0 s

One can see that the MILP is much faster than the MINLP. Figure 4 shows the results for the number of active pumps as given by both solvers. The solutions of the solvers differ in six of the 90 feasible cases. The objectives in these points are nearly equal. The reasons for theses slight differences are related to our given reasons for the different rotational speeds in Chapter 4.1. Due to the different model types we gain different solutions.

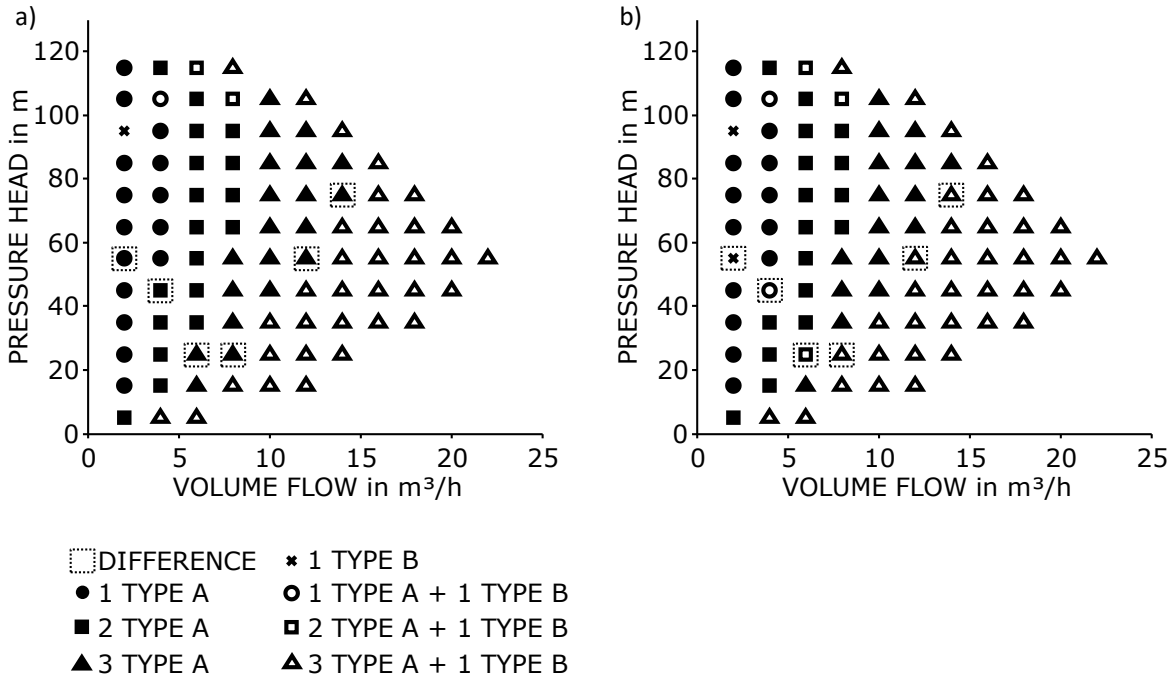



Figure 4: Optimization result for a booster station of three pumps of type A and one pump of type B. a) MILP, b) MINLP

Also in this case, we notice different rotational speeds for the active pumps of type A. The inevitable correction of the rotational speeds is not necessarily a feasibility problem anymore. Pumps of different types may but do not have to run at the same rotational speed to minimize the energy input. Thus this might become an optimization problem if pumps of both types are active.

As a next step we combine both methods: We use the MILP to make discrete decision for the active pumps. Afterwards, we use the discrete result as input for the NLP-Solver to gain an optimal and non-linear feasible solution for the original problem. The summed computation time for this approach is 223.6 s, so this approach is much quicker than the MINLP due to the excellent performance of the MILP solver.

## 5 CONCLUSIONS

- Pumping systems and booster stations use a significant amount of the produced electrical energy and thus need to be considered for optimization.
- We presented the technical requirements for the optimization of booster stations.
- Different approaches are suitable for the optimization of the control and layout of a booster station.
- We implemented a framework, which allows us to compare the different approaches and calculate appropriate solutions quickly.
- The effort of building a MILP is the highest, but the solvers work very quickly and deliver good results.
- Finding a nonlinear feasible solution corresponding to a MILP solution is a good alternative for the use of a MINLP solver.

## REFERENCES

[1] Pelz, P. F.: 250 Jahre Energienutzung: Algorithmen übernehmen Synthese, Planung und Betrieb von Energiesystemen, Fachvortrag anlässlich der Ehrenpromotion von Hans-Ulrich Banzhaf, 23.01.2014, Universität Siegen, 2014.

[2] Bidstrup, N.; Teepe, M.; Berge, G.; Ludwig, G.: Extended Product Approach for Pumps, Draft Version, Europump, 2014.

[3] Pelz, P.; Lorenz, U.: As good as can be! TOR plans the energetically optimal fluid system. ErP Special Issue, Delta p, 2013.

[4] Vielma, J.P.; Nemhauser, G.L.: Modeling disjunctive constraints with a logarithmic number of binary variables and constraints; Math. Program. Ser., 2011.

[5] Pöttgen, P.; Ederer, T.; Altherr, L. C.; Pelz, P. F.: Developing a Control Strategy for Booster Stations under Uncertain Load. Applied Mechanics and Materials, 807 pp. 241-246. ISSN 1660-9336, 2015.

[6] Groß, T; Pöttgen, P.; Pelz, P. F.: An Analytical Approach for the Optimal Operation of Pumps in Booster Systems, not yet published.

[7] Geißler, B: Towards Globally Optimal Solutions for MINLPs by Discretization Techniques with Applications in Gas Network Optimization, Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2011.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MANIFESTATION OF ASCENDANCY OF EXTINCTION-REIGNITION ON SOUNDING HYBRID ROCKET USING DESIGN INFORMATICS

**Kazuhisa Chiba**[1], **Masahiro Kanazaki**[2], **and Toru Shimada**[3]

[1]The University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 182-8585, Japan
e-mail: kazchiba@uec.ac.jp

[2] Tokyo Metropolitan University
6-6, Asahigaoka, Hino, Tokyo 191-0065, Japan
e-mail: kana@tmu.ac.jp

[3] Japan Aerospace Exploration Agency
3-1-1, Yoshinodai, Chuo, Sagamihara 252-5210, Japan
e-mail: shimada.toru@jaxa.jp

**Keywords:** Extinction-Reignition, Sounding Launch Vehicle, Hybrid Rocket Engine, Design Informatics, Multidisciplinary Design Optimization, Data Mining.

**Abstract.** *Information visualization have been performed for researching and developing a single-stage sounding launch vehicle with hybrid rocket engine by using design informatics; it has three points of view: problem definition, optimization, and data mining. The primary objective of this study is to reveal extinction-reignition ascendancy, which is one of the beneficial points of hybrid rocket, for expanding downrange and duration in the lower thermosphere. Swirling-flow oxidizer is furnished with solid fuel; we adopt polypropylene for solid fuel and liquid oxygen for oxidizer. A multidisciplinary design optimization was implemented by using a hybrid evolutionary computation; data mining was carried out by using a scatter plot matrix to efficiently perceive the entire design space. It is consequently revealed that extinction-reignition extends duration although it does not provide any effect on expanding downrange. Scatter plot matrix results express physical mechanisms of design-variable behaviors for the objective functions and also the roles of the design variables via bird's-eye visualization of the entire design-space constitution.*

## 1 INTRODUCTION

Single-stage rockets have been researched and developed for scientific observations and experiments of high-altitude zero-gravity condition, whereas multi-stage rockets have been for orbit injections of payloads. Institute of Space and Astronautical Science (ISAS)/Japan Aerospace Exploration Agency (JAXA) has been operating Kappa, Lambda, and Mu series rockets as the representatives of solid rocket to contribute to space science researches. Although Epsilon series began to be operated from September 2013, a next-generation rocket is necessary to fulfill higher-frequent and lower-cost space transportations.

Hybrid rocket engines (HREs) using different phases between fuel and oxidizer (solid fuel and liquid/gas oxidizer is generally used) have been researched and developed as an innovative technology in mainly the E.U.[1] and the U.S.[2] Each country has plans to adopt an HRE to a main engine of space transports because of several advantages: lower cost, higher safety[1], and pollution free flight due to no gunpowder use. In contrast, disadvantages of HREs proceed from their combustion. Since HREs have low regression rate of solid fuel due to turbulent boundary layer combustion, the thrust of HREs is less than that of pure-solid/liquid engines which premixed combustion[3] is implemented. Moreover, since the mixture ratio between solid fuel and liquid/gas oxidizer ($O/F$) is temporally fluctuated, thrust changes with time. Research topics of HREs are improving those performances via experiments.

Now in Japan, ISAS/JAXA recently researches HREs to develop a next-generation space transportation. Research topics are viporization of liquid oxygen, advancement of exhaust velocity $c^*$ efficiency, progress of regression rate, stable ignition, and numerical simulations of turbulent boundary layer combustion. In contrast, we will investigate hybrid-rocket ascendancy via conceptual design as a part of ISAS/JAXA's hybrid rocket project. The objective of this conceptual design is to quantitatively indicate hybrid-rocket advantage compared with the current rockets via conceptual design; multidisciplinary design requirements: chemical equilibrium, thrust, structural, aerodynamics, and trajectory analyses are driven. Furthermore, exhaustive design information will be obtained to additionally consider manufacturing, productive, and market factors for practical problems[2].

Design informatics (DI) has essential for not only an operating system itself but also its applications to practical problems so that science contributes toward the real world. Results themselves do not possess versatility in application problems due to their particularity; system versatility is indeed critical in application problems because it is revealed that application range is expanded. Furthermore, the application results indicate the guidance for system improvements. In this study, we conceptually explore a conceptual design of a single-stage sounding hybrid rocket using DI approach. The objective is that extinction-reignition advantage in the science mission for aurora observation on hybrid rocket will be quantitatively revealed. Since HREs are comparatively easy to perform multi-time ignition[4], extinction-reignition supremacy is especially predicted by using design informatics approach.

We researched step by step in the previous studies. As a first step, an optimization problem on single-time ignition, which is the identical condition of the current solid rocket, was defined to obtain the design information[5]. As a second step, the implication of solid fuels in the performance of hybrid rocket was revealed because the regression rate is one of the key elements for hybrid-rocket performance[6]. Finally, this study investigates an extinction-reignition sequence to reveal a hybrid-rocket ascendancy as multi-time ignition.

[1] this is especially important for manned mission.

[2] optimization is difficult to deal with them due to the difficulty of quantitative definition.

Table 1: Upper/lower limits of each design variable.

| *serial number* | *design variable* | *unit* | *design space* | | |
|---|---|---|---|---|---|
| dv1 | initial mass flow of oxidizer | [kg/sec] | $1.0 \leq$ | $\dot{m}_{\mathrm{ox}}(0)$ | $\leq 30.0$ |
| dv2 | fuel length | [m] | $1.0 \leq$ | $L_{\mathrm{fuel}}$ | $\leq 10.0$ |
| dv3 | initial radius of port | [m] | $0.01 \leq$ | $r_{\mathrm{port}}(0)$ | $\leq 0.30$ |
| dv4 | total combustion time | [sec] | $20.0 \leq$ | $t_{\mathrm{burn}}^{(\mathrm{total})}$ | $\leq 60.0$ |
| dv5 | 1st combustion time | [sec] | $10.0 \leq$ | $t_{\mathrm{burn}}^{(\mathrm{1st})}$ | $\leq 40.0$ |
| dv6 | extinction time | [sec] | $1.0 \leq$ | $t_{\mathrm{ext}}$ | $\leq 300.0$ |
| dv7 | initial pressure in combustion chamber | [MPa] | $3.0 \leq$ | $P_{\mathrm{cc}}(0)$ | $\leq 6.0$ |
| dv8 | aperture ratio of nozzle | [-] | $5.0 \leq$ | $\epsilon$ | $\leq 8.0$ |
| dv9 | elevation at launch time | [deg] | $60.0 \leq$ | $\phi(0)$ | $\leq 90.0$ |

The constitution of this paper is as follows. The optimization and data-mining techniques used in DI are explained in Chapter II. The problem definition for designing a hybrid rocket are shown in Chapter III. Otimization and data-mining results are revealed; the knowledge is also discussed in Chapter IV.

## 2 PROBLEM DEFINITION

We consider a conceptual design for a single-stage sounding hybrid rocket, simply composed of a payload chamber, an oxidizer tank, a combustion chamber, and a nozzle[7] shown in Fig. 1. A launch vehicle for aurora scientific observation will be focused because more efficient sounding rockets are desired due to successful obtaining new scientific knowledge on the aurora observation by ISAS/JAXA in 2009. In addition, a single-stage hybrid rocket problem fits for resolving fundamental physics regarding HREs and for improving the problem definition.

### 2.1 Objective functions

Three objective functions are defined. First objective is maximizing the downrange in the lower thermosphere (altitude of 90 to 150 [km]) $R_d$ [km] (obj1). Second is maximizing the duration in the lower thermosphere $T_d$ [sec] (obj2). It recently turns out that atmosphere has furious and intricate motion in the lower thermosphere due to energy injection, from which derives aurora, from high altitude. The view of these objective functions is to secure the hor-

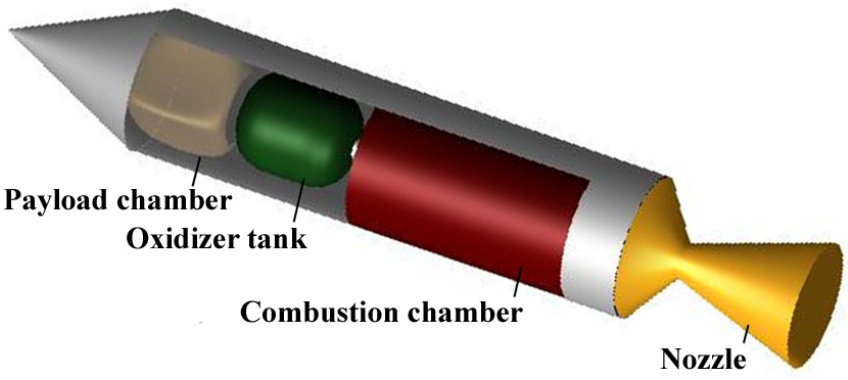



Figure 1: Conceptual illustrations of hybrid rocket.

izontal distance and time for competently observing atmospheric temperature and the wind so that thermal energy balance is elucidated on atmospheric dynamics. Third objective is minimizing the initial gross weight of launch vehicle $M_{\text{tot}}(0)$ [kg] (obj3), which is generally the primary proposition for space transportations.

## 2.2 Design variables

We use nine design variables: initial mass flow of oxidizer $\dot{m}_{\text{ox}}(0)$ [kg/sec] (dv1), fuel length $L_{\text{fuel}}$ [m] (dv2), initial radius of port $r_{\text{port}}(0)$ [m] (dv3), total combustion time $t_{\text{burn}}^{(\text{total})}$ [sec] (dv4), first combustion time $t_{\text{burn}}^{(\text{1st})}$ [sec] (dv5), extinction time from the end of first combustion to the beginning of second combustion $t_{\text{ext}}$ [sec] (dv6), initial pressure in combustion chamber $P_{\text{cc}}(0)$ [MPa] (dv7), aperture ratio of nozzle $\epsilon$ [-] (dv8), and elevation at launch time $\phi(0)$ [deg] (dv9). The design variables regarding rocket geometry are visualized in Fig. 2.

Note that since this problem assumes

$$\dot{m}_{\text{ox}}(t) = \dot{m}_{\text{ox}}(0) = \text{const.}, \tag{1}$$

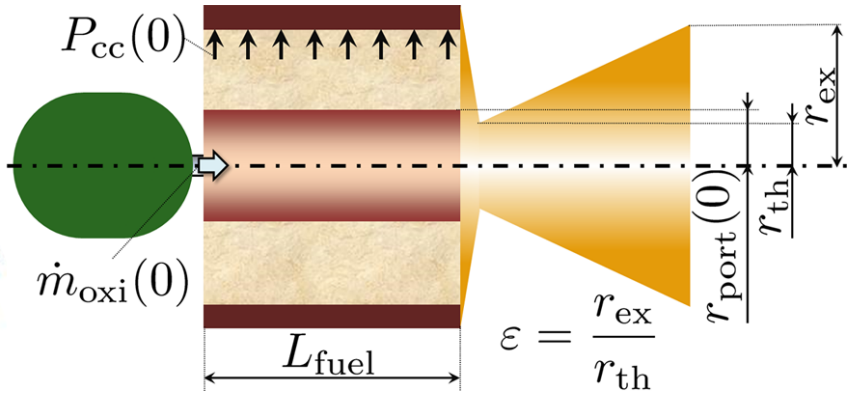



Figure 2: Design variables regarding rocket geometry. $\epsilon$ is described by using the radius at nozzle exit $r_{\text{ex}}$ and the radius at nozzle throat $r_{\text{th}}$.

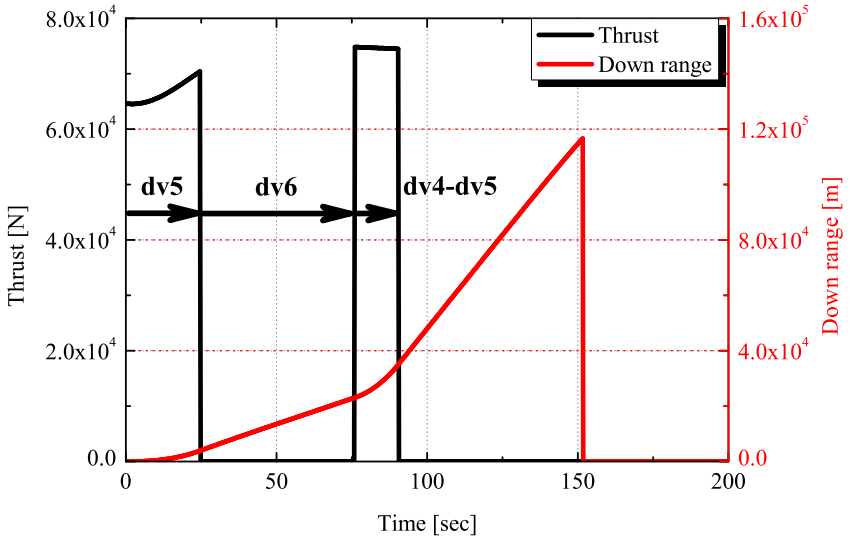



Figure 3: Conceptual graph of extinction-reignition. Two-time combustions generate two thrust pulses. Second thrust pulse increases $\mathrm{d}R_d/\mathrm{d}t$.

mass flow of oxidizer is equal between 1st and 2nd combustions. The relationship among dv4, dv5, and dv6 regarding extimction-reignition operation is conceptually shown in Fig. 3. We set two combustion times as follows:

$$
\begin{aligned}
t_{\text{burn}}^{(\text{1st})} &= \begin{cases} t_{\text{burn}}^{(\text{total})} & (t_{\text{burn}}^{(\text{total})} < t_{\text{burn}}^{(\text{1st})}) \\ t_{\text{burn}}^{(\text{1st})} & (t_{\text{burn}}^{(\text{total})} \geq t_{\text{burn}}^{(\text{1st})}) \end{cases}, \\
t_{\text{burn}}^{(\text{2nd})} &= \begin{cases} 0 & (t_{\text{burn}}^{(\text{total})} < t_{\text{burn}}^{(\text{1st})}) \\ t_{\text{burn}}^{(\text{total})} - t_{\text{burn}}^{(\text{1st})} & (t_{\text{burn}}^{(\text{total})} \geq t_{\text{burn}}^{(\text{1st})}) \end{cases}.
\end{aligned}
\tag{2}
$$

Under $t_{\text{burn}}^{(\text{total})} < t_{\text{burn}}^{(\text{1st})}$ condition, it is defined that $t_{\text{burn}}^{(\text{1st})}$ is set to be $t_{\text{burn}}^{(\text{total})}$ and second-time combustion is not performed. Note that there is no constraint except upper/lower limits of each design variable summarized in Table 1. These upper/lower values are exhaustively covering the region of the design space which is physically admitted. When there is a sweet spot[3] in the objective-function space, the exploration space would intentionally become narrow due to the operation of range adaptation on the evolutionary computation.

### 2.3 Evaluation method of hybrid rocket

First of all, $O/F(t)$ is computed by the following equation.

$$
O/F(t) = \frac{\dot{m}_{\text{ox}}(t)}{\dot{m}_{\text{fuel}}(t)}, \tag{3}
$$

where, $\dot{m}_{\text{ox}}(t)$ is set from eq. (1) and

$$
\begin{aligned}
\dot{m}_{\text{fuel}}(t) &= 2\pi r_{\text{port}}(t) L_{\text{fuel}} \rho_{\text{fuel}} \bar{\dot{r}}_{\text{port}}(t), \\
r_{\text{port}}(t) &= r_{\text{port}}(0) + \int \dot{r}_{\text{port}}(t) \mathrm{d}t.
\end{aligned}
\tag{4}
$$

$\dot{m}_{\text{ox}}(t)$ and $\dot{m}_{\text{fuel}}(t)$ are the mass flow of oxidizer/fuel [kg/sec] at time $t$, respectively. $r_{\text{port}}(t)$ is port radius [m] at $t$, $L_{\text{fuel}}$ describes fuel length, and $\rho_{\text{fuel}}$ is fuel density [kg/m$^3$]. $\dot{r}_{\text{port}}(t)$ describes the regression rate. Note that since HREs perform no premixed combustion which conventional rocket engine implements but turbulent boundary layer combustion, $O/F(t)$ is not constant but timely fluctuated. After that, an analysis of chemical equilibrium is performed by using NASA-CEA (chemical equilibrium with applications) [8], then trajectory, thrust, aerodynamic, and structural analyses are respectively implemented. A body is assumed as rigidness. As the time step is set to be 0.5 [sec], it takes roughly 10 [sec] for an individual evaluation on a general desktop computer.

A combustion chamber is filled with solid fuel with a single port at the center to supply oxidizer. As the regression rate to the radial direction $\dot{r}_{\text{port}}(t)$ [m/sec] generally governs thrust power of HREs, it is a significant parameter. This study uses the following experimental model[9, 10];

$$
\begin{aligned}
\dot{r}_{\text{port}}(t) &= a_{\text{fuel}} \times G_{\text{ox}}^{n_{\text{fuel}}}(t) \\
&= a_{\text{fuel}} \times \left( \frac{\dot{m}_{\text{ox}}(t)}{\pi r_{\text{port}}^2(t)} \right)^{n_{\text{fuel}}},
\end{aligned}
\tag{5}
$$

[3] the region that all objective functions proceed optimum directions.

where, $G_{\mathrm{ox}}(t)$ is oxidizer mass flux [kg/m$^2$/sec]. $a_{\mathrm{fuel}}$ [m/sec] and $n_{\mathrm{fuel}}$ [-] are the constant values experimentally determined by fuels. Swirling oxidizer flow is carried out; we adopt liquid oxygen for oxidizer and polypropylene as thermoplastic resin for solid fuel. Experiments[9, 10] respectively indicate $a_{\mathrm{fuel}} = 8.26 \times 10^{-5}$ [m/sec] and $n_{\mathrm{fuel}} = 0.5500$ [-] for polypropylene.

We assume a body to be rigid[11]. Figure 4 shows a three-degree-of-freedom (3DoF) flight simulation kinematic model. $T(t)$ is estimated; a flight path is computed by using the following 3DoF equations of motion:

$$\begin{cases} \ddot{x} = \dfrac{T(t)\cos\theta_1 - D(t)\cos\theta_2}{M_{\mathrm{tot}}(t)}, \\ \ddot{y} = \dfrac{T(t)\sin\theta_1 - D(t)\sin\theta_2}{M_{\mathrm{tot}}(t)} - g, \\ \ddot{\theta}_1 = \dfrac{N(t)|X_{\mathrm{c.p.}} - X_{\mathrm{c.g.}}|}{I(t)}, \end{cases} \tag{6}$$

where, $N(t)$ is the normal component of aerodynamic force. $N(t)$ is approximately evaluated as follows:

$$N(t) = \frac{1}{2}\rho(t)V^2(t)S_{\mathrm{ref}}\sin(|\theta_1 - \theta_2|). \tag{7}$$

$\rho(t)$ and $V(t)$ respectively denote the air density and rocket velocity at elapsed time $t$. $\theta_1$ and $\theta_2$ respectively describe vehicle attitude angle and flight path angle. Eq. 6 assumes that thrust vector corresponds to attitude direction, i.e., body axis. Therefore, thrust angle does not generally correspond to flight direction. $I(t)$ denotes the moment of inertia, which is estimated by using the coordinate of the center of gravity for an entire body $X_{\mathrm{c.g.}}(t)$ and that for the

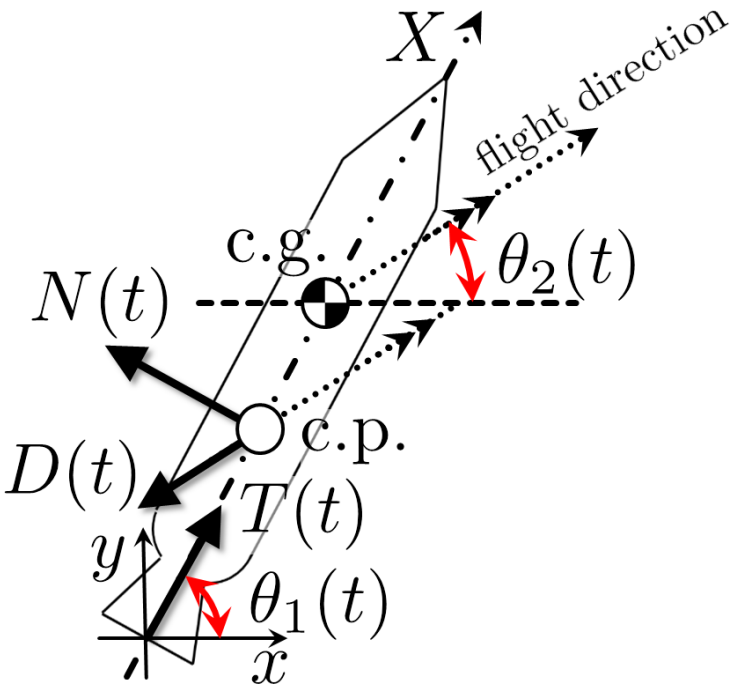



Figure 4: Illustration of a kinematic model of single-stage rocket for 3DoF flight simulation. Note that c.g. and c.p. respectively denote the center of gravity and the center of pressure.

components of body $X_{\text{c.g.}}^{(n_\text{comp})}(t)$ shown in Fig. 5; $I(t)$ is described as follows:

$$I(t) = \sum_{n_\text{comp}=1}^{5} m_{n_\text{comp}}(t) \times (X_{\text{c.g.}}^{(n_\text{comp})}(t) - X_{\text{c.g.}}(t))^2. \tag{8}$$

$X_{\text{c.g.}}$ is computed by the following equation, which considers the variance of fuel and oxidizer masses.

$$X_{\text{c.g.}}(t) = \sum_{n_\text{comp}=1}^{5} \frac{m_{n_\text{comp}}(t) \times X_{\text{c.g.}}^{(n_\text{comp})}(t)}{M_{\text{tot}}(t)}. \tag{9}$$

$X_{\text{c.g.}}^{(n_\text{comp})}(t)$ describes the distance between nozzle exit and the center of gravity of each component. $n_$comp denotes the serial number of each component; we consider five components. $n_$comp of 1, 2, 3, 4, and 5 respectively correspond to nozzle, combustion chamber, oxidizer tank, payload bay, and nose cone. Note that $m_{n_\text{comp}}(t)$ for nozzle, payload bay, and nose cone merely depends on the structures of these components; $m_{n_\text{comp}}(t)$ can be assumed to be constant. In contrast, $m_{n_\text{comp}}(t)$ for combustion chamber and oxidizer tank depends on not only the structure but also change in the amount of fuel and oxidizer over time. That is, $X_{\text{c.g.}}^{(2)}(t)$ and $X_{\text{c.g.}}^{(3)}(t)$ are time-dependent functions. $M_{\text{tot}}(t)$ denotes the total mass for a whole rocket body.

## 3 DESIGN INFORMATICS

DI is essential for practical design problems. Although solving design optimization problems is important under many-discipline consideration on engineering[12], the most significant part of the process is the extraction of useful knowledge of the design space from results of optimization runs[13, 14]. The results produced by multiobjective optimization (MOO) are not an individual optimal solution but rather an entire set of optimal solutions due to tradeoffs. That is, MOO results are not sufficient from the practical point of view as designers need a conclusive shape and not the entire selection of possible optimal shapes. But optimal-solution set produced by an MOO can be considered a hypothetical design database for design space. Thereupon, data mining techniques can be applied to a hypothetical database to acquire not only useful design knowledge but also structurizing and visualizing design space for conception support. This approach was suggested as DI[15].

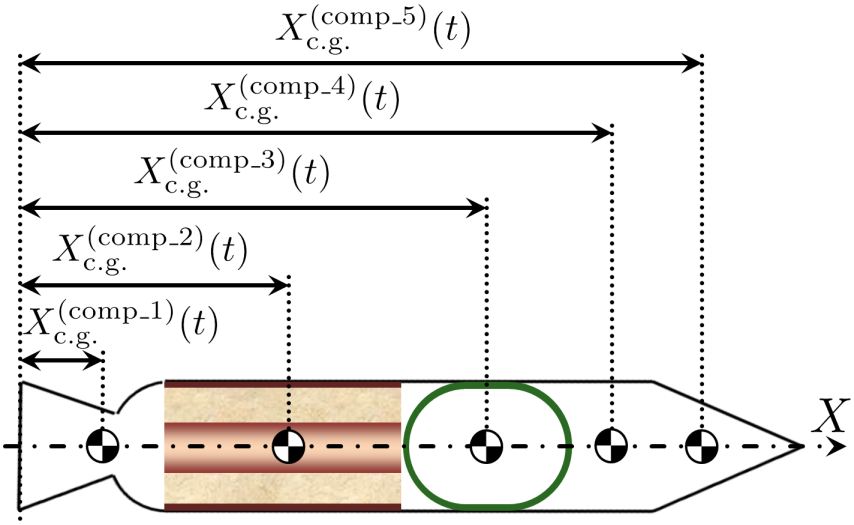



Figure 5: Definition of positions for the five-components' center of gravity.

The goal of this approach is the conception support for designers to materialize innovation. This methodology is constructed by three essences: problem definition, efficient optimization, and data mining for structurization and visualization of design space. A design problem including objective functions, design variables, and constraints, is strictly defined in view of the background physics for several months Note that problem definitions are the most important process because it directly gives effects on design space qualities. If we garrulously define a problem, unnecessary evolutionary exploration should be performed; needless mining will be also carried out because it is conceived to be low-quality design space. Then, optimization is implemented to acquire nondominated solutions[4] to become hypothetical database. Finally, data mining is implemented for this database to obtain design information. Mining has a postprocess role for optimization. Mining results might include significant knowledge for next design phase and also becomes the material to redefine a design problem.

### 3.1 Optimization method

DI second phase is optimization; we use evolutionary computations (ECs). Although we can employ a surrogate models[16]: the radial basis function and the Kriging model[17], which is a response surface model developed in the field of spatial statistics and geostatistics, we will not select them because they are generally difficult to deal with a large number of design variables. In addition, we would like to generate a hypothesis database using exact solutions. We also employ ECs so that plural individuals are parallel conducted. We use a hybrid EC between the differential evolution (DE) and the genetic algorithm (GA)[18].

First, multiple individuals are generated randomly as an initial population. Then, objective functions are evaluated for each individual. The population size is equally divided into subpopulations between DE and GA[5]. New individual candidates generated via DE and GA are combined. The nondominated solutions in the combined population are archived in common. Note that only the archive data is in common between DE and GA. The respective optimization methods are independently performed in the hybrid EC.

The hybrid EC is a real-coded optimizer[19]. Although GA is based on the real-coded NSGA-II (the elitist nondominated sorting genetic algorithm)[20], it is made several improvements on to be progressed with the diversity of solutions. Fonseca's Pareto ranking[21] and the crowding distance[20] are used for fitness values. The stochastic universal sampling[22] is employed for parents selection. Crossover rate is 100%; the principal component analysis blended crossover-$\alpha$ (PCABLX)[23] and the confidence interval based crossover using $L_2$ norm (CIX)[24] are used because of the high performance for convergence and diversity as well as the strength for noise[18]. The subpopulation size served by GA is equally divided for these two crossovers. Mutation rate is set to be constant as the reciprocal of the number of design variables. For alternation of generations, we employ a cross-generational elitist selection model, which uses the crowding distance for clustering; it selects $N$ solutions from all parents and offspring. Thereupon, candidates of the next generation are $2N$. The fitness of the region that individuals congregate because of falling into local optimum and s on is estimated to be low for diversity maintaining. DE is used as the revised scheme[25] for multiobjective optimization from DE/rand/1/bin scheme. The scaling factor $F$ is set to be 0.5. The hybrid EC has a range-adaptation function[26], which changes the search region according to the statistics of better

[4]quasi-Pareto solutions.

[5]although sub-population size can be changed at every generations on the optimizer, the determined initial sub-populations are fixed at all generations.

solutions, for all design variables. Range adaptations are implemented at every 20 generations.

### 3.2 Data mining

DI third phase is data mining. Scatterplot matrix (SPM)[27] or be simply named scatterplots remains one of the general visual descriptions for multidimensional data due to its simplicity. SPM is available to simultaneously visualize multidimensional data constructed by all of objective functions and design variables like as a bird's-eye view. Other data-mining techniques which have flexibility and effective visual expressiveness exist. However, we merely select SPM so that we will obtain first-step design information via observing design-space overview.

## 4 RESULTS

### 4.1 Optimization results

The population size is set to be 18 in one generation; the EC is performed until 4,000 generations. Population size of 18 is determined as a small number of the order of $10^1$ because the generation number will be earned as much in evolution. The generation number is decided by evolution convergence. Plots of acquired nondominated solutions are shown in Fig. 6(a); two discontinuous connecting and convex nondominated surfaces except several isolated individuals are generated.

The tradeoffs are identical to the previous study [6]; the difference whether extinction-reignition is considered will be explained here. The essential difference between the previous and the present results is to break the upper limit of $T_d$ and to expand it approximately 14%. This result indicates that extinction-reignition fulfills the hovering in the lower thermosphere. Extension of $T_d$ accompanies the increase of $M_{\text{tot}}(0)$. $R_d$ is not, however, improved due to extinction-reignition. This fact suggests that the significant design variable might not be $t_{\text{burn}}^{(\text{2nd})}$ but $t_{\text{ext}}$. When $t_{\text{burn}}^{(\text{2nd})}$ will serve functions, vehicles move away from the lower thermosphere; both of $R_d$ and $T_d$ are not increased.

If thrust direction can be horizontally controlled, $t_{\text{burn}}^{(\text{2nd})}$ is expected to give an effect on raising $R_d$ and $T_d$ simultaneously. Since thrust direction corresponds to body axis shown in Fig. 4 due to considering 3DoF equations of motion , $t_{\text{burn}}^{(\text{2nd})}$ is not performed functions; only $t_{\text{ext}}$ gives an effect on $T_d$. However, since detailed functions of design variables are not indicated from the optimization, data mining has essential role in DI.

### 4.2 Data mining results

Figure 7 shows the generated SPM so that data mining is implemented. The objective is to reveal design-variable roles in the design space via observing their behaviors. SPM does not indicate the tradeoff relations among more than the optimization results shown in Fig. 6 due to identical information.

#### 4.2.1 Knowledge for $R_d$

First, we will investigate design-variable roles for $R_d$. The optimization results shown in Fig. 6 indicate that $R_d$ is not improved despite extinction-reignition implementation; we will trace the physical mechanism of it. SPM shown in Fig. 7 indicates effective design variables: $\dot{m}_{\text{ox}}(0)$, $L_{\text{fuel}}$, and $\phi(0)$.

$R_d$-$\dot{m}_{\text{ox}}(0)$ plot in Fig. 7 reveals a nonlinear and positive correlation between them; $\dot{m}_{\text{ox}}(0)$ gives an effect on $R_d$ increase. $\dot{m}_{\text{ox}}(0)$ rise gives an effect on $\dot{r}_{\text{port}}(t)$ growth and thrust gain;

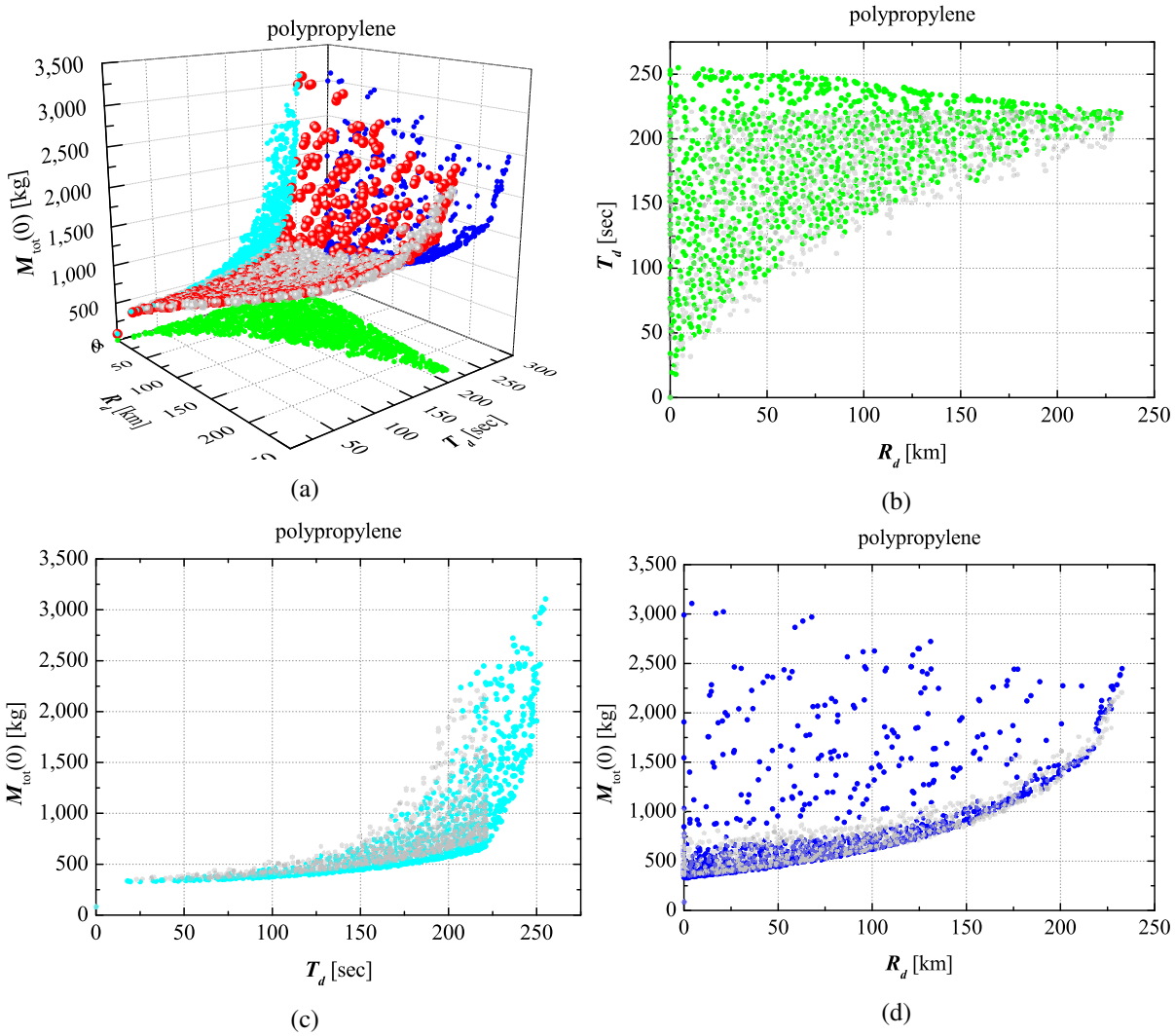



Figure 6: Plots of nondominated solutions derived by optimization, (a) plotted in the three-dimensional objective-function space (red) and their plots projected onto two dimensions, (b) plots projected onto two dimension between $R_d$ (obj1) and $T_d$ (obj2) (light green), (c) plots projected onto two dimension between $T_d$ (obj2) and $M_{\rm tot}(0)$ (obj3) (light blue), and (d) plots projected onto two dimension between $R_d$ (obj1) and $M_{\rm tot}(0)$ (obj3) (blue). Note that gray-colored plots represent the results of previous work[6] in view of no extinction-reignition.

this correlation is reasonable. However, this plot also indicates that $\dot{m}_{\rm ox}(0)$ reaches the upper limit shown in Table 1. Since this upper value descends from technical limit, $\dot{m}_{\rm ox}(0)$ is the bottleneck to realize the strategy to grow $R_d$ as horizontal flight path in the lower thermosphere using low $\phi(0)$.

$R_d$-$L_{\rm fuel}$ plot in Fig. 7 similarly shows a nonlinear and positive correlation between them; $L_{\rm fuel}$ also gives an effect on $R_d$ swell. However, $L_{\rm fuel}$ does not attain the upper limit of the design-variable space; the region to be close to $L_{\rm fuel}$ upper limit (from roughly 8 to 10 [m]) has no nondominated solution. If $L_{\rm fuel}$ increases, fuel mass rises; long-time combustion can be performed. But

- $R_d$-$t_{\rm burn}^{(1{\rm st})}$ plot in Fig. 7 indicates that $t_{\rm burn}^{(1{\rm st})}$ is not set to be close to upper limit.
- $R_d$-$t_{\rm burn}^{({\rm total})}$ plot in Fig. 7 reveals that $t_{\rm burn}^{({\rm total})}$ merely uses the lower half region.

That is, if $t_{\rm burn}^{(1{\rm st})}$ is set over the maximum value in $t_{\rm burn}^{(1{\rm st})}$ plots in Fig. 7, long combustion does not give an effect on $R_d$ growth. Furthermore, if $t_{\rm burn}^{({\rm total})} < t_{\rm burn}^{(1{\rm st})}$, we define Eq. (2); 2nd combustion

is not implemented. Thereupon, the absolute behavior of

$$t_{\text{burn}}^{(\text{2nd})} = t_{\text{burn}}^{(\text{total})} - t_{\text{burn}}^{(\text{1st})} > 0 \tag{10}$$

signifies that the hybrid EC evolves to strictly perform 2nd combustion. Thrust direction corresponds to body axis shown in Fig. 4. Body axis is diagonally downward although it depends on elevation. Therefore, if $t_{\text{burn}}^{(\text{2nd})}$ is long, a vehicle might deviate from the lower thermosphere. We consider two ways to fulfill a horizontal slide because of expanding $R_d$ and $T_d$.

- We have a way to employ vectored thrust. But since there is an upper limit of deflection angle, we cannot drastically implement it. Moreover, we are difficult to control deflection angle.
- We can consider several manners not to use vectored thrust; thrust direction keeps corresponding to body axis.
    1. $\phi(0)$ is set to be as low as possible so that horizontal launch is realized, i.e., horizontal component of thrust augments. However, since $\dot{m}_{\text{ox}}(0)$ reach the upper limit, a vehicle might not attain the lower thermosphere shown in $R_d$-$\phi(0)$ plot in Fig. 7.
    2. We control thrust so that it balances with vertical component due to rocket weight. If we control $\dot{m}_{\text{ox}}(t)$ for 2nd combustion, this manner might be fulfilled because present $\dot{m}_{\text{ox}}(0)$ is defined as constant shown in Eq. (1).
    3. Reignitions are discontinuously and repeatedly implemented for short time. But many-time ignition might be technically difficult.

    Feasibility studies regarding above 2nd manner should be quickly investigated.

$R_d$-$\phi(0)$ plot in Fig. 7 indicates a linear and negative correlation between them; $\phi(0)$ decrease gives an effect on $R_d$ augmentation. If $\phi(0)$ is low, flight path becomes horizontal; $R_d$ increases. However, $\phi(0)$ does not reach lower limit. This is because a vehicle might not attain the lower thermosphere when $\phi(0)$ is set to be under the minimum value: roughly 63 [deg].

If we consider knowledge for setting design variables summarized hereinbefore, an effective strategy to expand $R_d$ is that nearly vertical $\phi(0)$ is set so that a vehicle injection is implemented to the lower thermosphere; appropriate thrust is generated for 2nd combustion to horizontally move a body. If strictly vertical elevation: $\phi(0)$=90 [deg] is set, horizontal component of body velocity becomes zero, a rocket cannot have horizontal slide because thrust direction corresponds to body axis.

### 4.2.2 Knowledge for $T_d$

Second, we will consider design-variable roles for $T_d$. Since the optimization results shown in Fig. 6 already indicates that we can extend $T_d$, physical mechanisms of it will be revealed via SPM. Figure 7 indicates the effective design variables: $\dot{m}_{\text{ox}}(0)$, $L_{\text{fuel}}$, $t_{\text{burn}}^{(\text{total})}$, $t_{\text{burn}}^{(\text{1st})}$, $t_{\text{ext}}$, and $\phi(0)$. Since $\dot{m}_{\text{ox}}(0)$ and $L_{\text{fuel}}$ have similar effects for $R_d$, we will consider the roles of $t_{\text{burn}}^{(\text{total})}$, $t_{\text{burn}}^{(\text{1st})}$, $t_{\text{ext}}$, and $\phi(0)$ hereinafter. We especially focus extended feasible region due to extinction-reignition.

$T_d$-$t_{\text{burn}}^{(\text{1st})}$ plot in Fig. 7 describes positive correlation between them. Summit altitude becomes near 150 [km] as $t_{\text{burn}}^{(\text{1st})}$ is larger. $\phi(0)$ is also close to be vertical to efficiently implement. But if $t_{\text{burn}}^{(\text{1st})}$ is set to be roughly full time, a hull might go over altitude of 150 [km]; we observe maximum $t_{\text{burn}}^{(\text{1st})}$ on this plot. If $\phi(0)$ is close to the lower limit: 60 [deg], summit altitude does not reach 150 [km]; $T_d$ is not longer than that under vertical launch condition. $R_d$ is not also

expanded. Thereupon, $\phi(0)$ is set to be vertically so that fuel mass is restrained; $M_{\rm tot}(0)$ (obj3) will be reduced on the hybrid EC.

Furthermore, if $t_{\rm burn}^{\rm (2nd)}$ is set to be long, a vehicle also goes over altitude of 150 [km] because $t_{\rm burn}^{\rm (2nd)}$ (denoted as dv* in Fig. 7) is merely set to be low value on $T_d$-$t_{\rm burn}^{\rm (2nd)}$ plot in Fig. 7. That is, $t_{\rm burn}^{\rm (1st)}$ is appropriately determined first to reach the lower thermosphere; $t_{\rm burn}^{\rm (total)}$ is decided not to go beyond 150 [km]. Indeed we obtain infeasible region in upper area where no individual exists on $t_{\rm burn}^{\rm (total)}$ plots in Fig. 7 although we observe strongly positive correlation between $t_{\rm burn}^{\rm (total)}$ and $t_{\rm burn}^{\rm (1st)}$ shown in Fig. 7.

Physical mechanisms regarding combustion-time behaviors mentioned above will be considered. The range of dv* is from 0 to 20 [sec], i.e., $T_d$-$t_{\rm burn}^{\rm (2nd)}$ plot in Fig. 7 indicates that $t_{\rm burn}^{\rm (2nd)}$ absolutely fulfills Eq. 10 condition to become a nondominated solution. This problem likes to actively employ 2nd combustion. However, $t_{\rm burn}^{\rm (2nd)}$ does not like to set to be long because a hull might go over altitude of 150 [km]. If $t_{\rm burn}^{\rm (2nd)}$ is increased, $T_d$ linearly extends. However, there is the maximum $t_{\rm burn}^{\rm (2nd)}$: roughly 5 [sec]. This result indicates that properties of 1st and 2nd combustions are absolutely different.

- 1st needs powerful combustion to launch a body to the lower thermosphere.
- 2nd requires feeble combustion to merely sustain vehicle gross weight.

We should define $\dot{m}_{\rm ox}$ for 2nd combustion to simulate different properties and to realize a feeble 2nd combustion.

$T_d$-$t_{\rm ext}$ plot in Fig. 7 has the roughly linear and positive correlation. Since a vehicle draws roughly parabolic flight path, it is an efficiently operation that reignition is implemented after passing summit altitude and at the timing immediately before going down altitude of 90 [km]; this correlation is reasonable.

$T_d$-$\phi(0)$ plot in Fig. 7 expresses that $T_d$ becomes large as $\phi(0)$ is close to 90 [deg]. If $\phi(0)$ becomes the vertical condition, a rocket can use entire region of the lower thermosphere from 90 to 150 [km] taking advantage of 2nd combustion; $T_d$ is inevitably extended.

### 4.2.3 Knowledge for $M_{\rm tot}(0)$

$\dot{m}_{\rm ox}(0)$ has a peculiarity for $M_{\rm tot}(0)$ on their plot in Fig. 7; $M_{\rm tot}(0)$-$\dot{m}_{\rm ox}(0)$ plot shows two belts with different inclination $\mathrm{d}\dot{m}_{\rm ox}(0)/\mathrm{d}M_{\rm tot}(0)$. SPM indicates that high $\mathrm{d}\dot{m}_{\rm ox}(0)/\mathrm{d}M_{\rm tot}(0)$ is the cluster to perform long $t_{\rm ext}$ and short $t_{\rm burn}^{\rm (2nd)}$ (condition I); low $\mathrm{d}\dot{m}_{\rm ox}(0)/\mathrm{d}M_{\rm tot}(0)$ is the cluster to implement short $t_{\rm ext}$ and long $t_{\rm burn}^{\rm (2nd)}$ (condition II). Note that entire nondominated solutions in both clusters do not exist in the novel feasible region where extinction-reignition generates. If $M_{\rm tot}(0)$ is identical, we need more $\dot{m}_{\rm ox}(0)$ for a long $t_{\rm ext}$ sequence than for a short $t_{\rm ext}$ one. Because we require extra-thrust for extinction to convey fuel and oxidizer masses which burn down if we operate under no extinction-reignition condition. Thereupon, the difference of sensitivity

$$\left.\frac{\mathrm{d}\dot{m}_{\rm ox}(0)}{\mathrm{d}M_{\rm tot}(0)}\right|_{\rm condition\ I} > \left.\frac{\mathrm{d}\dot{m}_{\rm ox}(0)}{\mathrm{d}M_{\rm tot}(0)}\right|_{\rm condition\ II} \tag{11}$$

is strictly emerged. The fact indicates that vehicle gross weight with extinction-reignition swells rather than that with single-time ignition so that we actively adopt extinction-reignition to increase $R_d$ and $T_d$.

Although $M_{\rm tot}(0)$-$L_{\rm fuel}$ plot in Fig. 7 shows curved surface on the plots, these are divided into two clusters that we mentioned the above; each cluster has linear correlation between $M_{\rm tot}(0)$

and $L_{\mathrm{fuel}}$. If $L_{\mathrm{fuel}}$ shortens, fuel mass decreases; since weight for reduced fuel mass lightens, this is physically reasonable result.

### 4.2.4 Other knowledge

Figure 7 also indicates design variable behaviors. Especially note that there is a strong positive correlation between $\dot{m}_{\mathrm{ox}}(0)$ and $L_{\mathrm{fuel}}$. This result suggests that since $\dot{m}_{\mathrm{ox}}(0)$ arrives the upper limit 30 [kg/sec] shown in Table 1, $L_{\mathrm{fuel}}$ cannot reach the upper limit 10 [m] due to the strong correlation; accordingly this is one of the reason not to increase $t_{\mathrm{burn}}^{(\mathrm{total})}$. The shortcoming of HREs is low $\dot{r}_{\mathrm{port}}(t)$. $\dot{r}_{\mathrm{port}}(t)$ is the function described by $\dot{m}_{\mathrm{ox}}(t)$. The improvement of $\dot{m}_{\mathrm{ox}}(0)$ is essential work to effectively use extinction-reignition and to develop impressive hybrid rocket system.

Lowest value of $L_{\mathrm{fuel}}$ is observed to set to be 1.0 [m] on, e.g., $M_{\mathrm{tot}}(0)$-$L_{\mathrm{fuel}}$ plot in Fig. 7. If $L_{\mathrm{fuel}} = 1.0$ [m], $M_{\mathrm{tot}}(0))$ minimization is fulfilled whereas $R_d = 0$ and $T_d = 0$; an individual with $L_{\mathrm{fuel}} = 1.0$ [m] belongs to a nondominated solution. The lower limit of $L_{\mathrm{fuel}}$ to realize $R_d \neq 0$ and $T_d \neq 0$ is confirmed to be roughly 2.8 [m], which is the absolute minimum length to reach the lower thermosphere.

Figure 7 indicates that $r_{\mathrm{port}}(0)$ is not flexible; it has roughly constant. If we set a small $r_{\mathrm{port}}(0)$, we cannot implement sufficient oxidizer mass flow to fulfill appropriate thrust. In contrast, if we set a large $r_{\mathrm{port}}(0)$, fuel mass becomes insufficient. Being proportionate to a large $r_{\mathrm{port}}(0)$ cannot basically achieve $\dot{m}_{\mathrm{ox}}(0)$.

$P_{\mathrm{cc}}(0)$ is observed as an independent design variable on Fig. 7 because it is used to decide tanks' thicknesses. In addition, Fig. 7 shows the maximum $P_{\mathrm{cc}}(0)$ is approximately 4.9 [MPa]; upper design space for $P_{\mathrm{cc}}(0)$ is not completely used. Since tanks' thickness is huge as $P_{\mathrm{cc}}(0)$ is large, vehicles might not be existable for physical reasons: heavy $M_{\mathrm{tot}}(0)$ and large drag due to small aspect ratio of body.

There is a tendency that $\epsilon$ becomes large to gain thrust. Although $\epsilon$ likes to set to be the upper limit 8.0 [-] shown in Table 1, $\epsilon$ does not directly give effects on all objective functions: $R_d$, $T_d$, and $M_{\mathrm{tot}}(0)$.

## 5 CONCLUSIONS

Visualizing design space has been implemented for research and development of a single-stage sounding launch vehicle with hybrid rocket engine via design informatics, which manages a hybrid evolutionary computation for multidisciplinary design optimization and a scatter plot matrix for data mining. Especial objective of this study is to reveal extinction-reignition ascendancy in a hybrid rocket engine. The optimization result consequently indicates that extinction-reignition expands duration despite no improvement of downrange in the lower thermosphere. Furthermore, data-mining result reveals the intimate knowledge regarding physical mechanisms of the objective-function behavior and of design-variable roles in design space.

## ACKNOWLEDGEMENTS

This work is progressed via the discussion in the hybrid rocket research working group in ISAS/JAXA. We employed the interactive Scatter Plot Matrix; it was developed by ISAS/JAXA under Industrial Innovation R&D Topic 4 on High Performance Computing Initiative Field 4.

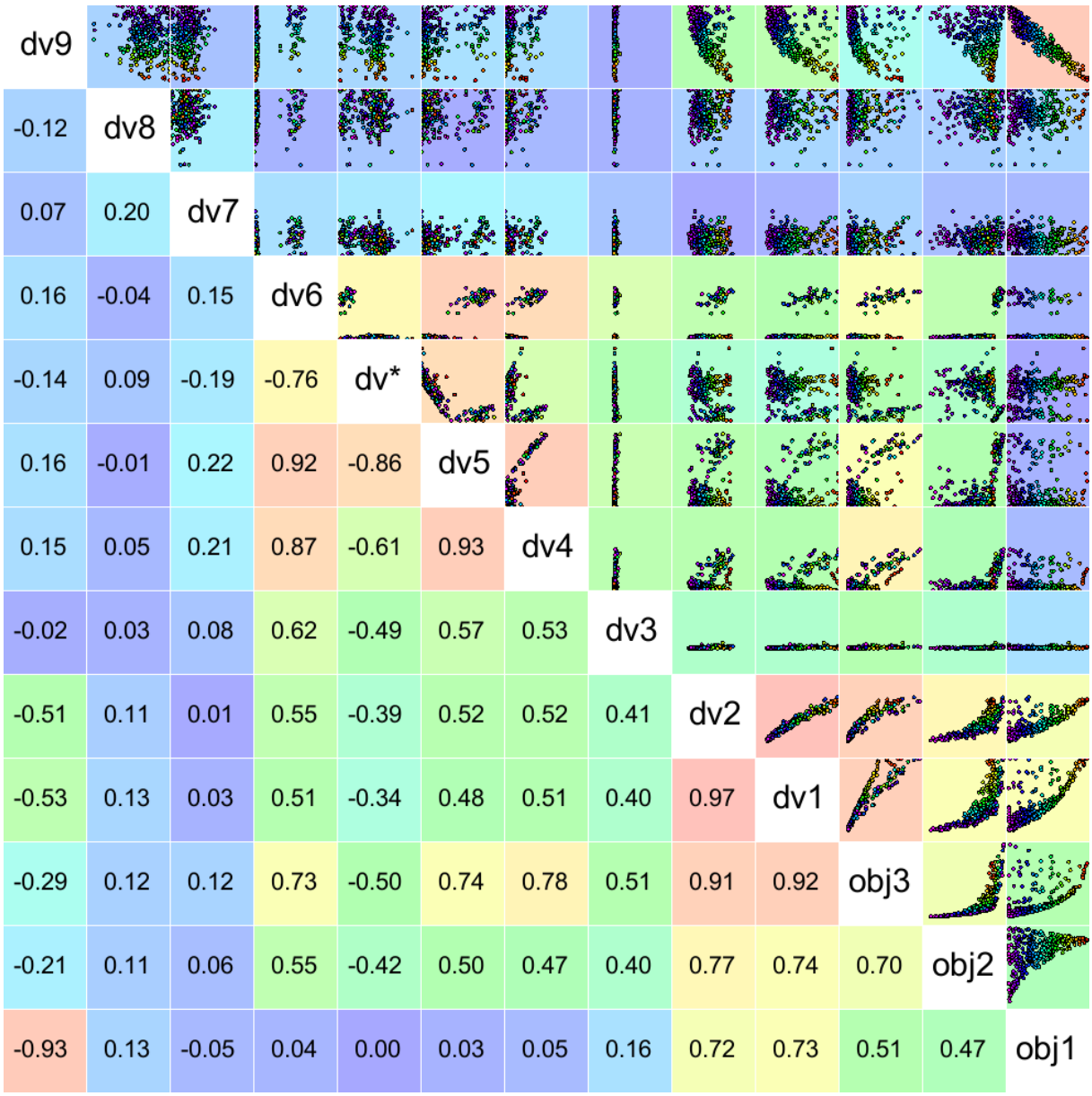



Figure 7: Scatterplots of the optimization results and their correlation coefficients. The upper/lower values of the axes of the design variables on graphs are set to be the upper/lower limits of each design variables shown in Table 1. Plots are colored by $R_d$ as obj1; each block is colored by the absolute value of their correlation coefficients. "dv*" denotes dv4−dv5, i.e., $t_{\mathrm{burn}}^{(\mathrm{2nd})}$; dv* range is from 0 to 20 [sec].

## REFERENCES


[1] M. A. Karabeyoglu, Advanced hybrid rockets for future space launch, in: Proceedings on 5th European Conference for Aeronautics and Space Sciences, EUCASS, 2013.

[2] L. Simurda, G. Zilliac, C. Zaseck, High performance hybrid propulsion system for small satellites, in: AIAA Paper 2013-3635, AIAA, 2013.

[3] M. A. Karabeyoglu, D. Altman, B. J. Cantwell, Combustion of liquefying hybrid propellants: Part 1, general theory, Journal of Propulsion and Power 18 (3) (2002) 610–620.

[4] M. A. Saraniero, L. H. Caveny, M. Summerfield, Restart transients of hybrid rocket engines, Journal of Spacecraft and Rockets 10 (3) (1973) 215–217.

[5] K. Chiba, M. Kanazaki, M. Nakamiya, K. Kitagawa, T. Shimada, Conceptual design of single-stage launch vehicle with hybrid rocket engine for scientific observation using design informatics, Journal of Space Engineering 6 (1) (2013) 15–27.

[6] K. Chiba, M. Kanazaki, M. Nakamiya, K. Kitagawa, T. Shimada, Diversity of design knowledge for launch vehicle in view of fuels on hybrid rocket engine, Journal of Advanced Mechanical Design, Systems, and Manufacturing 8 (3) (2014) JAMDSM0023, 1–14.

[7] Y. Kosugi, A. Oyama, K. Fujii, M. Kanazaki, Multidisciplinary and multi-objective design exploration methodology for conceptual design of a hybrid rocket, in: AIAA Paper 2011-1634, AIAA, 2011.

[8] S. Gordon, B. J. McBride, Computer program for calculation of complex chemical equilibrium compositions and applications I. analysis, in: NASA Reference Publication RP-1311, NASA, 1994.

[9] K. Hirata, C. Sezaki, S. Yuasa, N. Shiraishi, T. Sakurai, Fuel regression rate behavior for various fuels in swirling-oxidizer-flow-type hybrid rocket engines, in: AIAA Paper 2011-5677, AIAA, 2011.

[10] S. Yuasa, N. Shiraishi, K. Hirata, Controlling parameters for fuel regression rate of swirling-oxidizer-flow-type hybrid rocket engine, in: AIAA Paper 2012-4106, AIAA, 2012.

[11] K. Chiba, M. Kanazaki, A. Ariyarit, H. Yoda, S. Ito, K. Kitagawa, T. Shimada, Multidisciplinary design exploration of sounding launch vehicle using hybrid rocket engine in view of ballistic performance, International Journal of Turbo and Jet Engines 32 (3) (2015) 299–304.

[12] A. Arias-Montano, C. A. C. Coello, E. Mezura-Montes, Multiobjective evolutionary algorithms in aeronautical and aerospace engineering, IEEE Transactions on Evolutionary Computation 16 (5) (2012) 662–694.

[13] K. Chiba, S. Obayashi, K. Nakahashi, Design exploration of aerodynamic wing shape for reusable launch vehicle flyback booster, Journal of Aircraft 43 (3) (2006) 832–836.

[14] K. Chiba, S. Obayashi, Knowledge discovery in aerodynamic design space for flyback-booster wing using data mining, Journal of Spacecraft and Rockets 45 (5) (2008) 975–987.

[15] K. Chiba, Y. Makino, T. Takatoya, Design-informatics approach for intimate configuration of silent supersonic technology demonstrator, Journal of Aircraft 49 (5) (2012) 1200–1211.

[16] A. J. Keane, Statistical improvement criteria for use in multiobjective design optimization, AIAA Journal 44 (4) (2006) 879–891.

[17] S. Jeong, M. Murayama, K. Yamamoto, Efficient optimization design method using kriging model, Journal of Aircraft 42 (2) (2005) 413–420.

[18] K. Chiba, Evolutionary hybrid computation in view of design information by data mining, in: Proceedings on IEEE Congress on Evolutionary Computation, IEEE, 2013, pp. 3387–3394.

[19] A. Oyama, S. Obayashi, T. Nakamura, Real-coded adaptive range genetic algorithm applied to transonic wing optimization, Applied Soft Computing 1 (3) (2001) 179–187.

[20] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Transactions on Evolutionary Computation 6 (2) (2002) 182–197.

[21] C. M. Fonseca, P. J. Fleming, Genetic algorithms for multiobjective optimization: Formulation, discussion and generalization, in: Proceedings of the Fifth International Conference on Genetic Algorithms, Morgan Kaufmann, 1993, pp. 416–423.

[22] J. E. Baker, Adaptive selection methods for genetic algorithms, in: Proceedings of the International Conference on Genetic Algorithms and their Applications, Lawrence Erlbaum Associates, 1985, pp. 101–111.

[23] M. Takahashi, H. Kita, A crossover operator using independent component analysis for real-coded genetic algorithms, in: Proceedings of IEEE Congress on Evolutionary Computation 2001, IEEE, 2001, pp. 643–649.

[24] C. Hervas-Martinez, D. Ortiz-Bayer, N. Garcia-Pedrajas, Theoretical analysis of the confidence interval based crossover for real-coded genetic algorithms, in: The 7th International Conference on Parallel Problem Solving from Nature, LNCS 2439, Springer-Verlag, Berlin Heidelberg, 2002, pp. 153–161.

[25] T. Robic, B. Filipic, DEMO: Differential evolution for multiobjective optimization, in: The 3rd International Conference on Evolutionary Multi-Criterion Optimization, LNCS 3410, Splinger-Verlag, Guanajuato, Mexico, 2005, pp. 520–533.

[26] D. Sasaki, S. Obayashi, Efficient search for trade-offs by adaptive range multi-objective genetic algorithms, Journal of Aerospace Computing, Information, and Communication 2 (1) (2005) 44–64.

[27] N. Elmqvist, P. Dragicevic, J. Fekete, Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation, IEEE Transactions on Visualization and Computer Graphics 14 (6) (2008) 1539–1548.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPUTING AN OPERATING STRATEGY FOR AN ACTIVE BODY CONTROL WITH DYNAMIC PROGRAMMING IN THE AUTOMOTIVE AREA

**Marlene Utz, Phillipp Hedrich, and Peter F. Pelz**

Technische Universität Darmstadt, Chair of Fluid Systems
Otto-Berndt-Str. 2, 64287 Darmstadt, Germany
e-mail: {marlene.utz, phillipp.hedrich, peter.pelz } @fst.tu-darmstadt.de

**Keywords:** Active Body Control, Dynamic Programming, Operating Strategy, Skyhook

**Abstract.** *Customers expect a modern automobile to be comfortable and safe. The spring damper system in the wheel suspension is the most important component to ensure both. During the design process of the spring damping a physically based conflict between these two main functions arises. For passive systems improving the driving comfort will always worsen the driving safety and vice versa. We developed an active air spring damper system (AASD) which enables us to increase the driving comfort while keeping driving safety constant. Since auxiliary power is necessary for the adaptive control a new challenge is to handle the limited energy supply. In this paper we present a primal heuristic for the determination of an operating strategy for the AASD for a given route and a given limited energy supply. We appraise feasible solutions and present runtime benchmarks.*

## 1 INTRODUCTION

Automobiles are equipped with spring damper systems to ensure the safety and to allow a comfortable ride. Currently, most spring damper systems work passively. When designing a passive spring damper system one is always confronted by the conflict between driving safety and driving comfort. At a certain point it is not possible to improve the safety without simultaneously worsening the comfort and vice versa [1]. To further increase the driving comfort for a constant driving safety, active spring damper systems have to be used. Within the Collaborative Research Centre 805 "Control of Uncertainty in Load-Carrying Structures in Mechanical Engineering" an active air spring damper (AASD) was developed at the Chair of Fluid Systems at Technische Universität Darmstadt. The working principle is based on the adaptive control of the axial force by changing the effective area of the AASD dynamically [2]. Since auxiliary power is necessary for the adaptive control of the axial force, another challenge is to handle limited energy supply. In this paper we present a primal heuristic for the determination of an operating strategy for the AASD for a given route and a given energy supply. The route consists of sections distinguished by the corresponding uncertain road condition. The uncertain road condition is represented by an uncertainty set with scenarios. The working principle and the input data of the optimization are described in Section 2 of this paper.

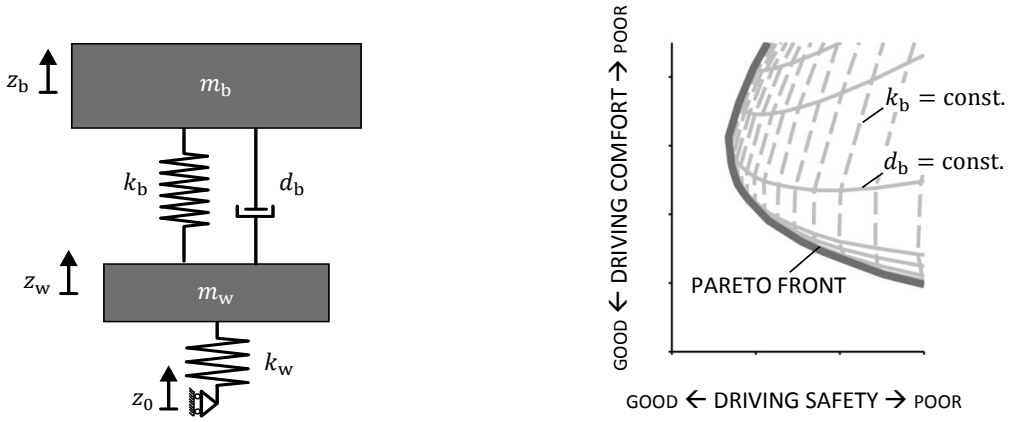
The goal is to find an operating strategy with maximal comfort, while keeping a level of safeness. Solving the problem in a reasonable timespan requires an efficient algorithm. The specific structure allows the usage of a dynamic program (DP), which is presented in the second part of the paper. The dynamic program, decides how much energy is allowed to be used on a single route section. The steps of the DP are the sections of the route and the considered states are the remaining energy supplies before entering the section. The DP finds the optimal solution for a given class of problems. Applying DP to solve this problem requires discretization of the energy supply. In order to relate an increase of demanded energy to an increase of comfort we modify the DP: Increasing the number of considered initial steps and reusing the calculated interim results enables us to compute operating strategies for a large number of energy supplies simultaneously.

Furthermore we give runtime benchmarks for the modified version of the DP. We increase the number of route sections and the number of considered states in every step of the DP. We present solution quality benchmarks and analyse a solution.

## 2 TECHNICAL DESCRIPTION

The requirements of an spring damper system are among others to guide the wheel safely to reach a good driving comfort and driving safety. When investigating a passive spring damper system in a quarter-car vehicle (cf. Fig.1a) we deal with a two degree-of-freedom-system, whose operating point is predefined by the spring stiffness and the damper constant. The operating point is fix if the spring damper system is not adjustable. When designing such systems a trade-off between driving comfort and driving safety has to be made. The standard deviation of the body-acceleration is used to rate the driving comfort and the standard deviation of wheel load fluctuation is equivalent to the driving safety [1]. A small value of the the standard deviation of the body-acceleration implies high driving comfort. The conflict between driving comfort and driving safety is illustrated in the conflict diagram in Figure 1b. The Pareto-front shows the limitation of a passive spring damper system.

In order to overstep the limitation of the passive system an active system which applies forces during usage is used. Within the Collaborative Research Centre 805 - "Control of Uncertainty




(a) (b)

Figure 1: Conventional topology of a vehicle suspension (a) and its limitation for vertical dynamics by a pareto front (b) [2].

in Load-Carrying Structures in Mechanical Engineering" an active air spring damper (AASD) was developed (cf. Fig 2) [2], which combines the advantages of an air spring damper system and an active system [3] . By changing the load carrying area of the air spring during the usage the axial force

$$F(t) = A(t)[p(t) - p_a)], \text{ with } A(t) = A_1(t) - A_2(t) \tag{1}$$

can be alerted.

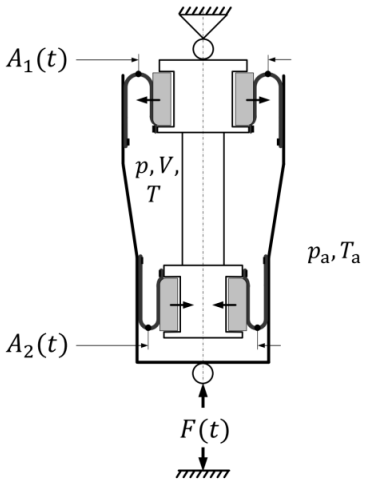



Figure 2: Schematic diagram of the double bellows active air spring. The axial force is $F(t) = A(t)[p(t) - p_a)]$, with the load-carrying area $A(t) = A_1(t) - A_2(t)$[2].

The AADS allows to improve the driving comfort at almost the same level of safety. For this reason the driving comfort is the objective of the optimization while keeping the safety level constant.

Without taking the random uneveness of the road into account the road excitation can be treated as a stochastic excitation [4]. The power spectral density (psd) of the road profile con-

tains informations about the unevenness of the road and is given by

$$\Phi_h(\Omega) = \lim_{X\to\infty} \frac{4\pi}{X} \left[\hat{h}(\Omega)\right]^2, \tag{2}$$

with $\hat{h}(\Omega)$ is the amplitude spectrum of the road profile and $X$ the period length. The psd is approximated by

$$\Phi_h(\Omega) = \Phi_h(\Omega_0) \left[\frac{\Omega_0}{\Omega}\right]^{-\omega}, \tag{3}$$

where $\Omega = 2\pi/L$ denotes the angular wavenumber and $\Phi_0 = \Phi_h(\Omega_0)$ describes the value of the psd at the reference angular wavenumber $\Omega_0 = 1$ m$^{-1}$. The road condition is characterized by $\Phi_0$ and $\omega = v\Omega$ where $v$ is the driving velocity. Often the condition of the roads and the driving velocity are uncertain. In order to take this fact into account sets Sc of all occurring triples $(\Phi_0, \omega, v)$ are used. The given route is divided into route sections $i = 1 \dots N$ with homogeneous road conditions. A road with the homogeneous condition can be divided into more than one route section. One calls these triple a section-scenario $\text{sc}_i$ and a sequence $\text{sc}_1, \dots, \text{sc}_N$ of section-scenarios corresponding with the sections of the consider route route-scenario $\text{sc}_r$. For every section-scenario a probability is given. We assume that the road condition get deterministic in the moment the vehicle enters the route section. One can generate information by a pre-scan of the road.

To use an active system it is necessary to implement a controller, which process signals provided by sensors of the vehicle to calculate the required active force. The optimization chooses from a set of controllers. A possible control strategy is a Skyhook control which is based on the assumption that the driving comfort is optimal when the oscillating body of the vehicle is damped with respect to the horizon. This strategy simulates the responses that would occur if a damper was installed between the vehicles body and the horizon [3]. The required active force between wheel and body is given by [5]

$$F_{\text{req}}(t) = d\dot{z}_b(t). \tag{4}$$

The optimization chooses from a set of damping constants $d \in D$.

Given a model of the AADS, a route section with a triple $\text{sc} = (\Phi_0, \omega, v)$ that characterizes the road condition and a Skyhook controller with damping constant $d$ one calculates the corresponding values of the standard deviation of the body-acceleration $\sigma(\ddot{z}_b(d, \text{sc}))$, the standard deviation of the wheel load fluctuation $\sigma(F_w(d, \text{sc}))$ and the mean power input $\bar{P}(d, \text{sc})$. In order to ensure a certain level of safety only those damping constants $d \in D$ are chosen for a section with certain road conditions whose $\sigma(F_w(d, \text{sc}))$ is lower than a given bound.

If the length $l_i$ of the route section is given one calculates the total energy consumption $e_i(d, \text{sc})$ for the section $i$ via

$$e_i(d, \text{sc}) = \bar{P}(d, \text{sc}) \frac{v}{l_i}. \tag{5}$$

## 3 DYNAMIC PROGRAMMING

The optimization problem is to find a partition of the available energy $e_r$ to the route section such that the resulting comfort is maximal and a minimum level of driving safety is guaranteed. Since the selection of the controller determines the energy usage for given section-scenarios one

has to find a sequence of controllers for every road-scenario which respects the energy partition. Due to the complexity of the problem and the vast amount of controller-sequences which need to be calculated, it is necessary to employ a efficient algorithm to solve the problem in reasonable time. The structure of the problem allows one to formulate a Dynamic Program (DP). DPs are used to solve problems which consist of a sequence of depending sub-problems. The DP starts with solving the smallest sub-problems and proceeds to solve the next bigger sub-problems by reusing the solution of the smaller sub-problems. This strategy leads to an optimal solution of the original problem [6]. It is possible to formulate a optimization problem containing a sequence of depending sub-problems as a DP if the optimal solution of a sub-problem contains the optimal solution of the next smaller sub-problem.

If a problem is represented by the following it can be solved with DP resp. algorithm 1 [6].

$$\begin{aligned} \min \quad F(x_1, \cdots, x_n) &= \sum_{k=1}^{n} f_k(z_{k-1}, x_k) \\ \text{s.t.} \quad z_k &= \text{tr}_k(z_{k-1}, z_k) \qquad \forall k = 1...n \\ z_0 &= a \\ z_n &= b \\ z_k &\in Z_k \qquad \forall k = 1...n \\ x_k &\in X_k(z_{k-1}) \qquad \forall k = 1...n \end{aligned} \tag{6}$$

whereby the terms are defined as follows:

| | |
|---|---|
| $n$ | number of *decision steps*, specifies how many decisions have to be taken. |
| $z_k$ | *state variable*, specifies the state of the system after decision step $k$. |
| $Z_k$ | *state set*: set of all possible states of the system after decision step $k$. |
| $z_0 = a$ | given *initial state* of the system. |
| $z_n = b$ | given *end state* of the system. |
| $x_k$ | *decision variable*, specifies the decision which is made in step $k$ |
| $X_k(z_{k-1})$ | *decision set*: set of all possible decisions which can be made in step $k$, if the system is in state $z_{k-1}$. |
| $\text{tr}_k(z_{k-1}, z_k)$ | *transfer function*, specifies the state $z_k$ of the system after taking decision $z_k$ starting from state $z_{k-1}$. |
| $f_k(z_{k-1}, x_k)$ | *evaluation function* of the step, which specifies the effect of decision $x_k$ of the objective function starting from state $z_{k-1}$. It is necessary that $f_k$ is only depending on $z_{k-1}$ and $x_k$. |

One calls a sequence of decisions $(x_h,\ x_{h+1}, ...,\ x_l)$ which transfer the system from state $z_{h-1}$ into $z_l$ a *policy*. A policy $(x_h^*,\ x_{h+1}^*, ...,\ x_l^*)$ which minimizes $\sum_{i=h}^{l} f_i(z_{i-1}, x_i)$ is called *optimal policy*. Let $P_k(z_k)$ be the problem of finding the optimal policy which transfers the system from state $z_k$ into $z_n$ and $F_k^*(z_k)$ the objective function of this problem. The fact that the evaluation function $f_k(z_{k-1}, x_k)$ only depends on $z_{k-1}$ and $x_k$ leads to the optimality principle of Bellman which says that if $(x_h^*, ...,\ x_l^*)$ is an optimal policy to transfer the system from $z_{h-1}$ into $z_l$ and the policy leads to a state $z_j^*$ $(h-1 \leq j \leq l)$, $(x_h^*, ...,\ x_{j+1}^*)$ and $(x_{j+1}^*, ...,\ x_l^*)$ are

optimal policies to transfer the system from $z_{h-1}$ into $z_j^*$ resp. $z_j^*$ into $z_l$. Using this principle leads to:

$$F_{k-1}^*(z_{k-1}) = \min_{x_k \in X_k(z_{k-1})} \Big\{ f_k(z_{k-1}, x_k) + F_k^*(z_k = t_k(z_{k-1}, x_k)) \Big\}. \tag{7}$$

**Algorithmus 1** DP via backward recursion

1: Solve $P_{n-1}(z_{n-1})$ for all $z_{n-1} \in Z_{n-1}$.
2: Set $x_n^*(z_{n-1})$ equals to the solution of $P_{n-1}(z_{n-1})$.
3: **for** $k = n-1$ **to** $1$ **do**
4: Solve $\mathbf{P_{k-1}(z_{k-1})}$ for all $z_{k-1} \in Z_{k-1}$ by solving the equations:
$F_{k-1}^*(z_{k-1}) = \min_{x_k \in X_k(z_{k-1})} \{\mathbf{f_k}(z_{k-1}, x_k) + F^*(z_k = \text{tr}_k(z_{k-1}, x_k))\}$
5: set $x_k^*(z_{k-1}) = \arg\min_{x_k \in X_k(z_{k-1})} \{\mathbf{f_k}(z_{k-1}, x_k) + F^*(z_k = \text{tr}_k(z_{k-1}, x_k))\}$
6: **end for**
7: Be $x_1^*$ the decision with minimize $F_0^*(a)$ and $z_1^* = \text{tr}_1(a, x_1^*)$
8: **for** $k = 2$ **to** $n$ **do**
9: set $x_k^* = x_t(z_{k-1}^*)$ and $z_k^* = \text{tr}_k(z_{k-1}^*, x_k^*)$.
10: **end for**
11: $x^* = (x_1^*, ..., x_n^*)$ is the solution of $P_0(a)$.

In the developed DP the sections $i = 1 \cdots N$ of the routes are the decision steps and the decision $x_i$ which is taken is how much energy is maximally consumed on the section. In case of a Skyhook controller one calculates the best damping constant $d$ for a given $x_i$ and occurring $\text{sc}_i$ by solving

$$d_i(x_i, \text{sc}_i) = \max \{d \in D : e_i(d, \text{sc}_i) \leq x_i\} \quad \forall i = 1 \dots N. \tag{8}$$

The states $z_i$ of the DP are the remaining energy supply after section $s_i$. Thus, the initial state $z_0 = e_r$ is the energy supply for the entire route. There is no requirement for the energy supply at the end of the route. As you can exchange energy for comfort it is advisable to set $z_N = 0$. There is always one controller which causes the highest power input on a route section. It is impossible to use more energy on this route section. Let $e_{i,max}$ be this energy and

$$\hat{e}_{i,max} = \sum_{l=i}^{N} e_{l,max} \quad \forall\, i = 1 \dots N \tag{9}$$

be the maximum amount of energy which can be consumed starting from section $i$ till the end of the route.

To get the state sets $Z_i$ we choose an energy-step $E$ and generate states $z_i = nE$ (starting with $n = 0$) by increasing $n$ by one as long as $nE \leq z_0$. We insert this states $z_i$ into $Z_i$. This leads to

$$Z_i = \{z_i = nE : nE \leq z_0, n \in \mathbf{N}\} \quad \forall i = 0 \ \dots N. \tag{10}$$

In order to relate an increase of demanded energy to an increase of comfort we modify our DP. The initial state $z_0$ is substituted by a set of $Z_0$ initial states. To get the state sets $Z_i$ we

choose an energy-step $E$ and generate states $z_i = nE$ (starting with $n = 0$) by increasing $n$ by one as long as $En \leq \hat{e}_{i,max}$. We insert this states $z_i$ into $Z_i$. This procedure leads to

$$Z_i = \{z_i = nE : nE \leq \hat{e}_{i,max}, n \in \mathbf{N}\} \quad \forall\, i = 0 \dots N. \tag{11}$$

The transfer function is given by

$$\mathrm{tr}_i(z_{i-1}, x_i) = z_{i-1} - x_i \quad \forall\, i = 1 \dots N. \tag{12}$$

To formulate a DP it is furthermore needed to define an evaluation function for every decision which can be made. Since the objective is the driving comfort and the comfort is measured by the standard deivation of the body acceleration $\sigma(\ddot{z}_b)$, this value is consulted. The body acceleration depends on the road conditions and the chosen controller. But it is independent of the energy supply before entering the section. Further we must remember that the sections of the route differs in their length $l_i$. This considerations lead to the following evaluation function for a route with deterministic road conditions:

$$f_i(x_i) = \sigma\Big(\ddot{z}_b\big(d_i(x_i, \mathrm{sc}_i), \mathrm{sc}_i\big)\Big) \frac{l_i}{\sum_{j=1}^{N} l_j} \quad \forall\, i = 1 \dots N. \tag{13}$$

We minimize the objective function because a small value of $f_i$ connotes a high driving comfort. Since the road conditions are uncertain we consider the three following evaluation functions – average-case, worst-case, best-case:

$$f_i^{AC}(x_i) = E\left[\sigma\Big(\ddot{z}_b\big(d_i(x_i, \mathrm{sc}_i), \mathrm{sc}_i\big)\Big)\right] \frac{l_i}{\sum_{j=1}^{N} l_j} \quad \forall\, i = 1 \dots N, \tag{14}$$

$$f_i^{BC}(x_i) = \min_{x_i \in X_i(z_{i-1})} \left\{\sigma\Big(\ddot{z}_b\big(d_i(x_i, \mathrm{sc}_i), \mathrm{sc}_i\big)\Big)\right\} \frac{l_i}{\sum_{j=1}^{N} l_j} \quad \forall\, i = 1 \dots N, \tag{15}$$

$$f_i^{WC}(x_i) = \max_{x_i \in X_i(z_{i-1})} \left\{\sigma\Big(\ddot{z}_b\big(d_i(x_i, \mathrm{sc}_i), \mathrm{sc}_i\big)\Big)\right\} \frac{l_i}{\sum_{j=1}^{N} l_j} \quad \forall\, i = 1 \dots N. \tag{16}$$

The runtime of the DP is mainly determined by the number of considered states and the evaluation of $d_i(x_i, \mathrm{sc}_i)$ and $f_i(x_i)$. Since $d_i$ and $f_i$ are independent of $z_{i-1}$ and $z_i$ they have only be evaluated ones for every possible $x_i$. In the modified version of the DP the number $M$ of evaluations is given by:

$$\begin{aligned} M &= \sum_{i=1}^{N} \frac{\frac{\hat{e}_{i-1,max}}{E}\left(\frac{\hat{e}_{i-1,max}}{E} + 1\right)}{2} + \frac{\hat{e}_{i,max}}{E}\left(\frac{\hat{e}_{i-1,max}}{E} - \frac{\hat{e}_{i,max}}{E}\right) \\ &= \sum_{i=1}^{N} \frac{\hat{e}_{i-1,max}^2}{2E^2} + \frac{\hat{e}_{i-1,max}}{2E} + \frac{\hat{e}_{i-1,max}\hat{e}_{i,max}}{E^2} - \frac{\hat{e}_{i,max}^2}{2E^2} \end{aligned} \tag{17}$$

The calculated solution is robust against all considered road-scenarios. In case a section-scenario $(\Phi_0, \omega, v)$ occurs which was not taken into account in the DP the energy consumption on the section will differ from the one in the solution. This leads to a different state respectively remaining energy supply after this section. Due to the optimality principle of Bellman one can change to the solution which is optimal based on the arising state.

## 4 BENCHMARKS AND EXAMPLES

The benchmarks where computed on a computer with 16 GB RAM and an AMD Phenon II X6 1100T processor. We used a linearised model of the AADS [7] with a Skyhook control. The used damping constants are $5000\ \mathrm{N\,s\,m^{-1}}$, $4500\ \mathrm{N\,s\,m^{-1}}$, $4000\ \mathrm{N\,s\,m^{-1}}$, $3500\ \mathrm{N\,s\,m^{-1}}$, $3000\ \mathrm{N\,s\,m^{-1}}$, $2500\ \mathrm{N\,s\,m^{-1}}$, $2000\ \mathrm{N\,s\,m^{-1}}$, $1500\ \mathrm{N\,s\,m^{-1}}$, $1000\ \mathrm{N\,s\,m^{-1}}$, $500\ \mathrm{N\,s\,m^{-1}}$ and $0\ \mathrm{N\,s\,m^{-1}}$. The routes are given by a sequence of roads with road conditions. For every road we use deviations $\Delta\Phi_0,\ \Delta\omega$ and $\Delta v$ to generate the section-scenarios so that one gets three possible value for $\Phi_0,\ \omega$ and $v$. For every section one has $27$ section-scenarios and for a route with $N$ sections one has $27^N$ route-scenarios. We consider 6 different routes. For detailed information about sections and road conditions see appendix.

- **route A** contains just autobahn sections and the velocity is $130\ \mathrm{km/h}$.
- **route B** contains just federal road sections and the velocity varies between 95 $\mathrm{km/h}$ and $110\ \mathrm{km/h}$.
- **route C** is a route from Darmstadt to Ulm (Germany).
- **route D** is a route with good road conditions and a high velocity.
- **route E** is a route from Frankfurt to Darmstadt (Germany).
- **route F** is the route of a commuter which contains only federal roads.

In order to investigate the runtime of the implemented DP we vary the number of road sections $N$ and the energy-step size $E$. This leads to a variation of the steps of the DP and the number of the states in every single step.

Figure 3 shows the runtime over the energy steps. We used the energy-step $5$ J, $10$ J, $50$ J, $100$ J, $500$ J, $1000$ J, $5000$ J, $10000$ J. As we expected the runtime behaves like $f(E) = 1/E^2 + 1/E$ (cf. Fig. 3) for small values of $E$. For small values of $E$ respectively a large number of sub-problems the runtime is asymptotically like $f(E)$.

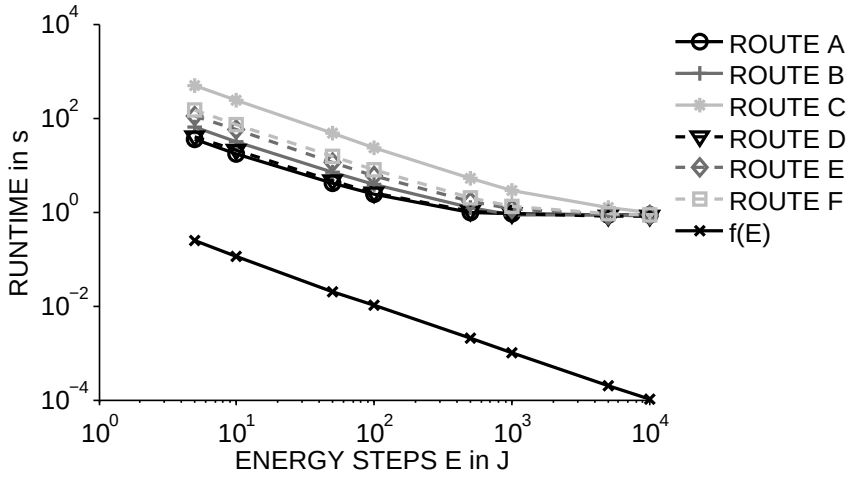



Figure 3: Runtime and function value over the energy steps for all considered routes with axes in a logarithmic scale.

Figure 4 shows the runtime over the number of route sections for $E = 10$ J. We generate routes with different numbers of sections, by dividing the section of the considered routes into subsections. As expected the runtime increases linearly with the number of route sections.

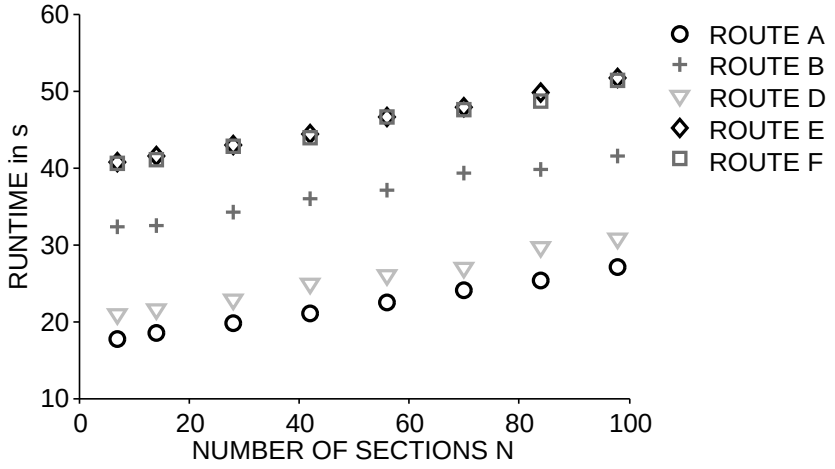



Figure 4: Runtime over the number of route section for all considered routes.

After investigating the runtime we want to have a look at the solution for the route E for $E = 10$ J. Figure 5 shows the objective values for all initial states and three evaluation functions. For the total energy supply for the route of $2.1$ kJ the objective values are BC $= 0.275$ m/s$^2$,

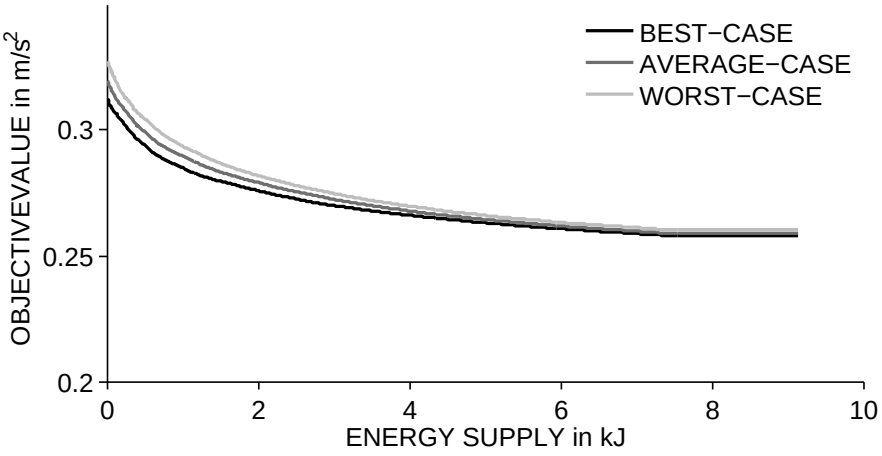



Figure 5: Objective values over the energy supply for route E and an energy step of $10$ J.

AC $= 0.278$ m/s$^2$ and WC $= 0.281$ m/s$^2$ . It is possible to halve the energy supply with a lose of 3.3%, 3.9 % and 4.2 % of comfort in the best-, average- and worst-case.

## 5 CONCLUSIONS

- To resolve the conflict between driving safety and driving comfort active spring damper systems are developed. Since auxiliary power is used for the adaptive control another

challenge is to handle a limited energy supply while the road conditions are uncertain.

- We compute a operating strategy for the AADS which is robust for all road-scenarios and leads to maximal comfort.
- Discretizeing the energy supply enables us to use DP to solve the problem in reasonable time.
- To relate an increase of demanded energy we use a set of initial energy supplies instead of one.
- The runtime of the DP mainly depends on the calculation of the evaluation function and the section-scenario depending controller. Since $d_i$ and $f_i$ are independent of $z_{i-1}$ and $z_i$ they only need to be evaluated once for every possible $x_i$.
- The employed damping constant depends on the occurring section-scenario and the energy supply before entering the section
- In a next step we wave the condition that the road condition get deterministic when entering the section. We will develop a DP which calculates one single damping constant for every section.

## 6 ACKNOWLEDGEMENT

The authors like to thank the German Research Foundation DFG for funding this research within the Collaborative Research Center (SFB) - "Control of Uncertainty in Load-Carrying Structures in Mechanical Engineering".

## REFERENCES

[1] M. Mitschke, H. Wallentowitz, *Dynamik der Kraftfahrzeuge*. Springer, 2004.

[2] P. Hedrich, F.J. Cloos, J. Würtenberger, P. F. Pelz, *Comparison of a New Passive and Active Technology for Vibration Reduction of a Vehicle under Uncertain Load*. Applied Mechanics and Materials, 807 pp 57-66, 2015.

[3] B. Heißing, M. Ersoy, *Chassis Handbook. Fundamentals, Driving dynamics, Components*. Vieweg Teubner, 2011.

[4] H. Braun, *Untersuchung von Fahrbahnunebenheiten und Anwendung der Ergebnisse*. Dissertation, TU Braunschweig, 1969.

[5] D. Karnopp, *Active damping in road vehicle suspension systems* Vehicle System Dynamics 12.6 pp 291-311, 1983.

[6] W. Domschke, A. Drexl, *Einführung in Operations Research, Vol. 5*. Springer, 2005.

[7] T. Bedarff, P. Hedrich, P. F. Pelz, *Design of an Active Air Spring Damper*. 9. IFK Proceedings Vol.3, 2014.

## APPENDIX

| $i$ | $l_i$ in $\mathrm{km}$ | $\Phi_0$ in $1$ | $\Delta\Phi_0$ in $1$ | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s^{-1}}$ | $\Delta v$ in $\mathrm{m\,s^{-1}}$ |
|---|---|---|---|---|---|---|---|
| 0 | 5.00 | 0.50 | 0.02 | 2.30 | 0.10 | 36.11 | 0.06 |
| 1 | 2.00 | 1.00 | 0.07 | 1.50 | 0.06 | 36.11 | 0.11 |
| 2 | 2.35 | 1.10 | 0.05 | 1.90 | 0.11 | 36.11 | 0.30 |
| 3 | 1.60 | 2.00 | 0.10 | 2.00 | 0.30 | 36.11 | 0.07 |
| 4 | 3.20 | 0.90 | 0.06 | 2.60 | 0.07 | 36.11 | 0.09 |
| 5 | 2.80 | 1.00 | 0.11 | 2.30 | 0.05 | 36.11 | 0.10 |
| 6 | 4.00 | 1.90 | 0.30 | 2.20 | 0.03 | 36.11 | 0.04 |

Table 2: route A contains just autobahn sections an the velocity is $\mathrm{km/h}$.

| $i$ | $l_i$ in $\mathrm{km}$ | $\Phi_0$ in $1$ | $\Delta\Phi_0$ in $1$ | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s^{-1}}$ | $\Delta v$ in $\mathrm{m\,s^{-1}}$ |
|---|---|---|---|---|---|---|---|
| 0 | 2.00 | 1.30 | 0.02 | 2.10 | 0.10 | 27.78 | 0.06 |
| 1 | 2.00 | 2.00 | 0.07 | 1.70 | 0.06 | 30.56 | 0.11 |
| 2 | 2.35 | 2.10 | 0.05 | 2.90 | 0.11 | 27.78 | 0.30 |
| 3 | 1.60 | 2.00 | 0.10 | 2.00 | 0.30 | 27.78 | 0.07 |
| 4 | 3.20 | 1.90 | 0.06 | 2.60 | 0.07 | 29.17 | 0.09 |
| 5 | 2.80 | 4.00 | 0.11 | 2.30 | 0.05 | 26.39 | 0.10 |
| 6 | 2.00 | 6.00 | 0.30 | 2.60 | 0.03 | 27.78 | 0.04 |

Table 3: route B contains just autobahn sections an the velocity varies between 95 $\mathrm{km/h}$ and 110 $\mathrm{km/h}$.

| $i$ | $l_i$ in $\mathrm{km}$ | $\Phi_0$ in $1$ | $\Delta\Phi_0$ in $1$ | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s^{-1}}$ | $\Delta v$ in $\mathrm{m\,s^{-1}}$ |
|---|---|---|---|---|---|---|---|
| 0 | 0.42 | 12.50 | 0.02 | 2.50 | 0.10 | 8.33 | 0.06 |
| 1 | 2.58 | 9.00 | 0.07 | 2.30 | 0.06 | 13.89 | 0.11 |
| 2 | 1.95 | 2.30 | 0.05 | 2.40 | 0.11 | 27.78 | 0.30 |
| 3 | 67.91 | 3.00 | 0.10 | 2.00 | 0.30 | 33.33 | 0.07 |
| 4 | 186.00 | 1.00 | 0.06 | 2.20 | 0.07 | 36.11 | 0.09 |
| 5 | 4.43 | 3.40 | 0.11 | 2.10 | 0.05 | 22.22 | 0.10 |
| 6 | 0.28 | 8.00 | 0.30 | 2.60 | 0.03 | 8.33 | 0.04 |

Table 4: route C is a route from Darmstad to Ulm (Germany).

| $i$ | $l_i$ in km | $\Phi_0$ in 1 | $\Delta\Phi_0$ in 1 | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s}^{-1}$ | $\Delta v$ in $\mathrm{m\,s}^{-1}$ |
|---|---|---|---|---|---|---|---|
| 0 | 2.00 | 3.00 | 0.02 | 2.60 | 0.10 | 33.33 | 0.06 |
| 1 | 2.00 | 1.00 | 0.07 | 1.50 | 0.06 | 36.11 | 0.11 |
| 2 | 2.35 | 1.10 | 0.05 | 1.90 | 0.11 | 41.67 | 0.30 |
| 3 | 1.60 | 2.00 | 0.10 | 2.00 | 0.30 | 36.11 | 0.07 |
| 4 | 3.20 | 0.90 | 0.06 | 2.60 | 0.07 | 41.67 | 0.09 |
| 5 | 2.80 | 1.00 | 0.11 | 2.30 | 0.05 | 36.11 | 0.10 |
| 6 | 2.00 | 12.00 | 0.30 | 2.40 | 0.03 | 22.22 | 0.04 |

Table 5: route D is a route with good road conditions and a high velocity.

| $i$ | $l_i$ in km | $\Phi_0$ in 1 | $\Delta\Phi_0$ in 1 | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s}^{-1}$ | $\Delta v$ in $\mathrm{m\,s}^{-1}$ |
|---|---|---|---|---|---|---|---|
| 0 | 0.28 | 12.20 | 0.02 | 2.40 | 0.10 | 8.33 | 0.06 |
| 1 | 5.90 | 8.00 | 0.07 | 2.30 | 0.06 | 13.89 | 0.11 |
| 2 | 1.35 | 2.10 | 0.05 | 2.30 | 0.11 | 27.78 | 0.30 |
| 3 | 20.60 | 1.00 | 0.10 | 2.00 | 0.30 | 36.11 | 0.07 |
| 4 | 3.20 | 2.30 | 0.06 | 2.20 | 0.07 | 27.78 | 0.09 |
| 5 | 1.75 | 9.00 | 0.11 | 2.40 | 0.05 | 13.89 | 0.10 |
| 6 | 0.70 | 10.00 | 0.30 | 2.30 | 0.03 | 8.33 | 0.04 |

Table 6: route E is a route from Frankfurt to Darmstadt (Germany).

| $i$ | $l_i$ in km | $\Phi_0$ in 1 | $\Delta\Phi_0$ in 1 | $\omega$ in $\mathrm{cm}^3$ | $\Delta\omega$ in $\mathrm{cm}^3$ | $v$ in $\mathrm{m\,s}^{-1}$ | $\Delta v$ in $\mathrm{m\,s}^{-1}$ |
|---|---|---|---|---|---|---|---|
| 0 | 1.00 | 1.70 | 0.02 | 2.10 | 0.10 | 8.33 | 0.06 |
| 1 | 2.00 | 2.00 | 0.07 | 1.80 | 0.06 | 13.89 | 0.11 |
| 2 | 12.35 | 4.90 | 0.05 | 2.30 | 0.11 | 22.22 | 0.30 |
| 3 | 5.80 | 5.30 | 0.10 | 2.40 | 0.30 | 25.00 | 0.07 |
| 4 | 4.20 | 2.90 | 0.06 | 2.60 | 0.07 | 13.89 | 0.09 |
| 5 | 0.75 | 4.00 | 0.11 | 2.30 | 0.05 | 8.33 | 0.10 |
| 6 | 2.00 | 2.00 | 0.30 | 2.60 | 0.03 | 8.33 | 0.04 |

Table 7: route F is the daily route of a commuter which contains only federal roads.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SEISMIC RELIABILITY-BASED DESIGN OF STRUCTURES ISOLATED BY FPS

**P. Castaldo[1], G. Amendola[2], B. Palazzo[3]**

[1] Department of Civil Engineering, University of Salerno.
Via Giovanni Paolo 132, 84084 Salerno, Italy.
e-mail: pcastaldo@unisa.it

[2] Department of Civil Engineering, University of Salerno.
Via Giovanni Paolo 132, 84084 Salerno, Italy.
e-mail: guglielmoamendola7@gmail.com

[3] Department of Civil Engineering, University of Salerno.
Via Giovanni Paolo 132, 84084 Salerno, Italy.
e-mail: palazzo@unisa.it

**Keywords:** Friction pendulum devices, Monte Carlo simulations, power spectral density method, medium soil condition, seismic fragility, seismic reliability.

**Abstract.** *The paper deals with the seismic reliability of base-isolated structural systems equipped with friction pendulum isolators (FPS) in order to provide useful design recommendations. A two-degree-of-freedom model is adopted by accounting for the superstructure flexibility, whereas the FPS isolator behaviour is described by adopting a widespread model which considers the variation of the friction coefficient with the velocity. The spectral displacement corresponding to the isolated period has been chosen as intensity measure (IM). The uncertainty in the seismic inputs as well as the friction coefficient at large velocity are considered as random variables modeled through appropriate probability density functions. Monte Carlo simulations are developed in order to evaluate the probabilities exceeding different limit states related to both superstructure and isolation level defining the seismic fragility curves. Finally, considering the seismic hazard curve related to an Italian site, closed-form expressions are derived with the aim to design the radius in plan of the friction pendulum isolators in function of the expected reliability level.*

## 1 INTRODUCTION

In the last decades, isolation systems have emerged as a very effective technique for the seismic protection of building frames [1], which, even if designed according to the most advanced codes, could suffer severe damages under strong earthquake events [2]. Among the base isolation devices currently employed for seismic isolation, friction pendulum system (FPS) isolators present some advantages, mainly related to their capability of providing an isolation period independent of the mass of the supported structure, their high dissipation and recentering capacity, and their longevity and durability characteristics [3]-[4].

Over the years, within the issue of the passive control, many works have developed new design strategies and methodologies [5]-[14], as well as other works have been focused on probabilistic analyses in structural dynamics, structural reliability methods, and reliability-based analysis [15]-[17]. Reliability evaluation of base-isolated systems has been presented by Chen et al. [18], as well as Monte Carlo simulations have been performed by Fan and Ahmadi [19] to analyze the stochastic response of sliding isolation systems under random earthquake excitations. In Barroso and Winterstein [20], the seismic performance of steel buildings isolated with FPS bearings was evaluated by taking into account the variability of both the seismic intensity and the record characteristics. Seismic reliability analyses of a 3D system isolated by FPS bearings have been carried out in [21]-[22] by accounting for the randomness of both the isolator properties (i.e., coefficient of friction) and of the earthquake main characteristics. Performance curves for the isolators and the superstructure have been estimated by considering both the vertical and horizontal components of each seismic excitation. This way, a reliability criterion has been defined to assist the design of the isolator dimensions in plan by considering the effects of the uncertainties relevant to the problem. In [23], the influence of FPS bearing properties and of the structural parameters on the seismic performance of base-isolated structures through the nondimentionalitation of the motion of equations is analyzed by providing useful results for seismic reliability analyses.

This paper deals with the seismic reliability of structural systems equipped with friction pendulum isolators (FPS) by presenting the fragility curves related to an extensive parametric study encompassing a wide range of building properties, seismic intensity levels and considering both the friction coefficient and soil characteristics as random variables. The isolated system is described by a two-degree-of-freedom (2dof) system in order to take account of the superstructure flexibility, and the FPS behavior is described by employing the model developed by Mokha et al. [4] for which the friction coefficient varies with the velocity. The uncertainty in the seismic inputs is taken into account by considering a set of artificial records [24], obtained through the power spectral density method [25], with different characteristics depending on soil dynamic parameters [26]-[27], and scaled to increasing intensity levels. Incremental dynamic analyses are developed in order to evaluate the probabilities exceeding different limit states related to both superstructure and isolation level through an extensive parametric study carried out for different structural properties. The estimates of the response statistics obtained are used for deriving seismic fragility curves of both the superstructure and isolation level assuming different values of the corresponding limit states. The seismic fragility curves are useful to evaluate the seismic reliability of base-isolated systems equipped with FPS, within the PEER-like modular approach [28]. In fact, in the final part of the work, considering the seismic hazard curve related to L'Aquila site (Italy), as provided by NTC08 [29], regarding a structural system isolated by FP bearings with a design life of 50 years, reliability-based abacuses are derived with the aim to design the radius in plan of the FP isolators in function of the structural properties and reliability level expected.

## 2 SYSTEM DESCRIPTION AND EQUATION OF MOTION

The equation of motion governing the response of a single-degree-of-freedom (SDOF) system on single concave FPS isolation devices to the seismic input $\ddot{u}_g(t)$ is:

$$\begin{aligned} &m_s \cdot \ddot{u}_s(t) + c_s \cdot \dot{u}_s(t) + k_s \cdot u_s(t) = -m_s \cdot \left[\ddot{u}_g(t) + \ddot{u}_b(t)\right] \\ &m_b \cdot \ddot{u}_b(t) + f_b(t) + c_b \cdot \dot{u}_b(t) + \\ &-c_s \cdot \dot{u}_s(t) - k_s \cdot u_s(t) = -m_b \cdot \ddot{u}_g(t) \end{aligned} \tag{1}$$

where $u_s$ denotes the displacement of the superstructure relative to isolation bearing, $u_b$ the isolator displacement relative to the ground, $m_s$ and $m_b$ respectively the mass of the superstructure and of the basement, $k_s$ and $c_s$ respectively the superstructure stiffness and inherent viscous damping constant, $c_b$ the bearing viscous damping constant, $\ddot{u}_g(t)$ the ground motion input, the dot differentiation over time, and where $f_b(t)$ denotes the FPS bearing resisting force. This latter can be expressed as:

$$f_b(t) = k_b \cdot u_b(t) + \mu(\dot{u}_b)(m + m_b) g Z(t) \tag{2}$$

where $k_b = (m_s + m_b) g / R$, $g$ is the gravity constant, $R$ is the radius of curvature of the FPS, $\mu(\dot{u}_b(t))$ the coefficient of sliding friction, which depends on the bearing slip velocity $\dot{u}_b(t)$, and $Z(t) = \mathrm{sgn}(\dot{u}_b)$, where $sgn(\cdot)$ is the sign function.

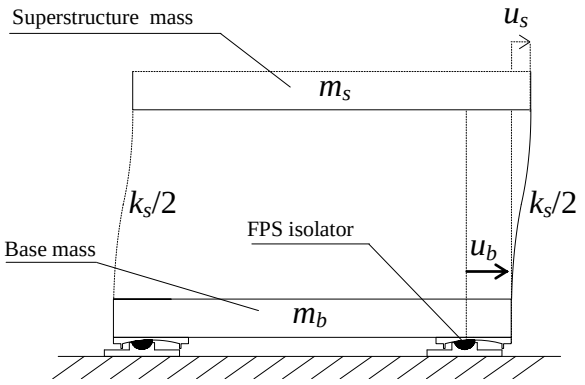



Figure 1: 2dof model of building isolated with FPS.

Experimental results [4],[30]-[31], suggest that the coefficient of sliding friction of Teflon-steel interfaces obeys to the following equation:

$$\mu(\dot{u}_b) = f_{\max} - Df \cdot \exp\left(-\alpha \left|\dot{u}_b\right|\right) \tag{3}$$

in which $f_{\max}$ represents the maximum value of friction coefficient attained at large velocities of sliding, $f_{\min} = f_{\max} - Df$ represents the value at zero velocity.

In order to generalize the problem and unveil the characteristic parameters controlling the seismic behaviour of the system, the equation of motion can be reduced to a non-dimensional form. By dividing Eqn.(1a) by $m_s$, and Eqn.(1b) by $m_b$, Eqn.(1) can be rewritten as:

$$\begin{aligned}
&\ddot{u}_s(t)+2\xi_s\omega_s\dot{u}_s(t)+\omega_s^2u_s(t)=-\left[\ddot{u}_g(t)+\ddot{u}_b(t)\right]\\
&\ddot{u}_b(t)+\frac{(m_s+m_b)}{m_b}\left[2\xi_b\omega_b\cdot\dot{u}_b(t)+\omega_b^2\cdot u_b(t)+\mu(\dot{u}_b)\,g\,\mathrm{sgn}(\dot{u}_b)\right]+\\
&-2\xi_s\omega_s\cdot\frac{m_s}{m_b}\cdot\dot{u}(t)-\omega_s^2\cdot\frac{m_s}{m_b}\cdot u(t)=-\ddot{u}_g(t)
\end{aligned}\tag{4}$$

where the parameters $\omega_s$ and $\xi_s$ denote the superstructure circular frequency and damping factor, whereas the parameters $\omega_b=\sqrt{\frac{k_b}{(m_s+m_b)}}=\sqrt{\frac{g}{R}}$ and $\xi_b$ denote the fundamental circular frequency and damping factor for a rigid mass $(m_s+m_b)$ on a linear frictionless isolator of stiffness $k_b$ and viscous damping constant $c_b$. The fundamental period of vibration of the base-isolated system, $T_b=2\pi/\omega_b$, corresponding to the pendulum component, results to be independent of the superstructure mass and related only to the radius of curvature of the spherical surface *R*. After introducing the mass ratio $\gamma=\frac{m_s}{(m_s+m_b)}$ [32], Eqn.(4) can be re-written as:

$$\begin{aligned}
&\ddot{u}_s(t)+2\xi_s\omega_s\dot{u}_s(t)+\omega_s^2u_s(t)=\\
&=-\left[a_g(t)+\ddot{u}_b(t)\right]\\
&\ddot{u}_b(t)+\frac{1}{1-\gamma}\left[2\xi_b\omega_b\cdot\dot{u}_b(t)+\omega_b^2\cdot u_b(t)+\mu(\dot{u}_b)\,g\,\mathrm{sgn}(\dot{u}_b)\right]+\\
&-2\xi_s\omega_s\frac{\gamma}{1-\gamma}\cdot\dot{u}(t)-\omega_s^2\frac{\gamma}{1-\gamma}\cdot u(t)=-\ddot{u}_g(t)
\end{aligned}\tag{5}$$

## 3 SEISMIC RELIABILITY OF STRUCTURES WITH FPS: RANDOM VARIABLES

Seismic reliability assessment of a building structure, according to the structural performance (SP) evaluation method [21],[33], is based on the coupling between structural performance levels [34] and associated exceeding probabilities during its design life [35]. Coherently with the PEER-like modular approach [28] and performance-based earthquake engineering (PBEE) approach [36], the uncertainties related to the seismic input intensity are separated from those related to the characteristics of the record (record-to-record variability) by introducing a scale factor, i.e., an intensity measure (*IM*). The approach is based on calculating the probabilities of exceeding different limit state thresholds, properly defined, given different values of the intensity measure with the aim to define the fragility curves of the systems. Afterward, the abovementioned fragility curves integrated with the seismic hazard curve, expressed in terms of the same *IM*, related to a reference site, lead to the mean annual rates of exceeding the limit states. Using a Poisson distribution, it is possible to transform the mean annual rates of exceeding the limit states into probabilities of exceedance in the time frame of interest (e.g., 50 years).

The aim of this work consists of evaluating the seismic reliability of structural systems equipped with friction pendulum isolators (FPS) through an extensive parametric study encompassing a wide range of building properties, different seismic intensity levels and consi-

dering both the friction coefficient and earthquake characteristics as random variables. With reference to the uncertainty in the seismic inputs, it is taken into account by considering a set of artificial records, obtained through the power spectral density (PSD) method [25]. Regarding the uncertainty of the friction coefficient at large velocity of the FP devices, another appropriate uniform probability density functions (PDF) is employed.

### 3.1 Non-Stationary stochastic processes: Power Spectral Density (PSD) method

In order to evaluate the record-to-record variability of the structural system response, several artificial earthquake excitations have been considered. In particular, if the evolution of the frequency with the time can be neglected, each earthquake excitation can be modeled as a Gaussian stationary process with mean value equal to zero and two-sided power spectral density (PSD) function $S_{ff}(\omega)$. It follows that the stochastic process $f(t)$ can be simulated by the following series as $N \to \infty$:

$$f(t) = \sqrt{2}\sum_{n=0}^{N-1} A_n \cos(\omega_n t + \Phi_n) \tag{6}$$

where $A_n = (2S_{ff}(\omega_n)\Delta\omega)^{1/2}$, $\omega_n = n\Delta\omega$ for $n = 0 \ldots N-1$, $\Delta\omega = \omega_u / N$, having assumed $T_0 = 2\pi / \Delta\omega$=31.25s (NTC08) and $\omega_u = 50 rad/s$, $\Phi_0, \Phi_1, \Phi_2, ....., \Phi_{N-1}$ are independent random phase angles distributed uniformly over the interval [0,2π]. A sample function $f^{(i)}(t)$ of the simulated stochastic process $f(t)$ can be obtained by replacing the sequence of random phase angles $\Phi_0, \Phi_1, \Phi_2, ....., \Phi_{N-1}$ with their respective *i*-th realizations $\Phi_0^i, \Phi_1^i, \Phi_2^i, ....., \Phi_{N-1}^i$, sampled through Monte Carlo simulations:

$$f^i(t) = \sqrt{2}\sum_{n=0}^{N-1} A_n \cos(\omega_n t + \Phi_n^i) \tag{7}$$

In this study, 100 sequence of random phase angles are sampled through Monte Carlo simulations in order to generate 100 input accelerometric signals. The power spectral density function (PSD) of the embedded stationary process is described by the widely-used Kanai and Tajimi [37]-[38], modified according to Clough and Penzien [39], which applies:

$$S_f(\omega) = \frac{\omega_g^4 + 4\xi_g^2\omega_g^2\omega^2}{(\omega_g^2 - \omega^2) + 4\xi_g^2\omega_g^2\omega^2} \cdot \frac{\omega^4}{(\omega_f^2 - \omega^2) + 4\xi_f^2\omega_f^2\omega^2} S_0 \tag{8}$$

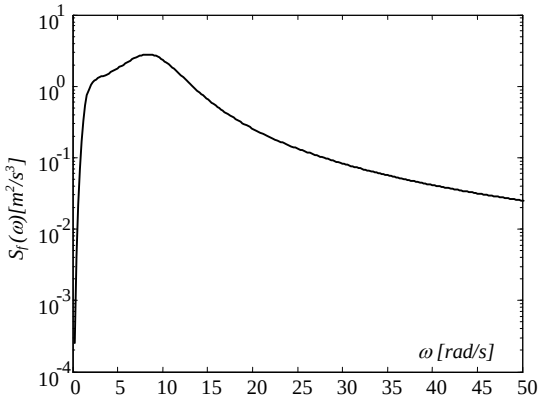



Figure 2: PSD function corresponding to Medium Soil condition.

In the following parametric study, with the aim to assume the uncertainty related to earthquake characteristics in terms of soil dynamics parameters corresponding to Medium Soil condition according to EC8 [40], $\omega_g$ and $\xi_g$ are modeled as random variables uniformly distributed, respectively, in the intervals [3π,5π] (rad/sec) and [40%,60%] [26]-[27], and sampled through Monte Carlo simulations. In Figure 2, the PSD function related to medium Soil with the sampled values of $\omega_g$ and $\xi_g$ equal respectively to 3π and 40% is represented. In order to obtain non-stationary stochastic processes, a time-modulating function proposed by [41], as shown in Figure 3, is adopted.

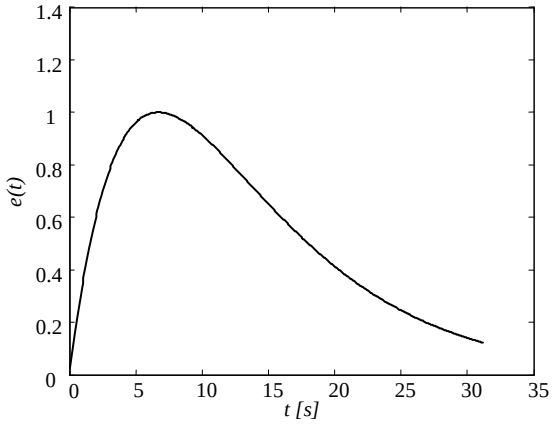



Figure 3: Time-modulating function [41].

### 3.2 Uncertainty of the friction coefficient

As regard to the friction coefficient, the experimental data, developed by [4],[30]-[31] on sheet type Teflon bearings, have pointed out that friction is a complex phenomenon, not complying with the Coulomb friction law and that several mechanisms contribute to its variability. In this study, a uniform density probability function (PDF), ranging from 3% to 12%, has been assumed to model the sliding friction at large velocity as random variable $f_{max}$. For the generation of the sampled values of the friction coefficient $f_{max}$, within the stratified sampling techniques to develop Monte Carlo simulations, the Latin Hypercube Sampling (LHS) method [42]-[43] has been used. In particular, in the following parametric study, 20 sampled values *(j = 20)* of the random variable $f_{\max}$ are employed and assuming a ratio $f_{\max}/f_{\min}$ equal to 3, based on regression of experimental results, whereas the exponent α of Eqn.(3) equal to 30 [4],[30]-[31].

## 4 PARAMETRIC STUDY:INCREMENTAL DINAMIC ANALYSIS RESULTS

Seismic reliability assessment of the equivalent base-isolated systems is based on developing incremental dynamic analyses (IDA) [44].

### 4.1 Intensity measure (IM): spectral displacement

In general, the *IM*'s choice should be driven by criteria of efficiency, sufficiency, and hazard computability [45]. In this study, the spectral displacement, $S_D(T_b,\xi_b)$, at the isolated period of the system, $T_b = 2\pi/\omega_b$, and for the damping ratio $\xi_b$, is assumed as intensity measure. In the analyses carried out in this study, the damping ratio $\xi_b$ is taken equal to zero,

consistently with [46]. The corresponding *IM*, hereinafter denoted as $S_D\left(T_b\right)$, in the IDA is assumed ranging from 0m to 0.5m.

### 4.2 Structural parameters and incremental dynamic analysis (IDA) results

The incremental dynamic analysis (IDA) is developed through an extensive parametric study encompassing a wide range of base-isolated building properties according to Eqn.(5). The parameters $\xi_b$ and $\xi_s$ are assumed respectively equal to 0% and 2%, $\gamma$ equal to 0.7, the radius *R* of the FPS equal to 1m and 4m, the fixed-base system period $T_s$ is considered varying between 0.3s and 1.5s. It follows that the isolation degree [47], ranges from 1.3 (flexible superstructure) to 13.3 (rigid superstructure).

The response parameters $u_s$ and $u_b$ are adopted as the engineering demand parameters (EDP). It follows that a set of samples is obtained for each output variable (EDP) representing the response variability. In this paper, the response parameters are assumed to follow a lognormal distribution according to [21]-[46]. A lognormal distribution can be fitted to the both response parameters (i.e., the extreme values of the EDPs), by estimating the sample lognormal mean, $\mu_{\ln}\left(EDP\right)$, and the sample lognormal standard deviation $\sigma_{\ln}\left(EDP\right)$, through the maximum likelihood estimation method.

In the hypothesis of regular buildings, the Eqn. (9), according to [29], is assumed as relationship between the fixed-base building period and its height *H*, and is employed to estimate the height *H* as the integer multiple of the inter-storey height assumed equal to *h*=3m and, so, the corresponding total number of floors $N_f$.

$$T_s = 0.075H^{\frac{3}{4}} \tag{9}$$

For each $T_s = 2\pi / \omega_s$, assuming the building floor mass equal to $m_{s,i}$ = 1000 kNs$^2$/m, for $i=1...N_f$, it is possible to determinate the floor stiffness and vector $\mathbf{\Phi}_1$ containing the floor displacements of the first mode of the fixed-base structure normalized to the top floor displacement. The base mass $m_b$ is assigned in order to respect the mass ratio $\gamma$ [32]:

$$\gamma = \frac{\Gamma_1^{\,2} M_{s1}}{\sum_{i=1}^{N_f} m_{s,i} + m_b} \tag{10}$$

where $\Gamma_1$ and $M_{s1}$ represent respectively the participation factor and modal mass of the fundamental mode of the fixed-base structure. It follows that the maximum absolute inter-story drift of the 1$^{st}$ floor can be evaluated as $u_{s,1,\max} = \Gamma_1 \phi_{11} u_{s,\max}$, and this response parameter, divided by the inter-storey height assumed equal to *h*=3m, corresponds to the overall maximum interstorey drift index (IDI) experienced over the different stories that controls the performance of the superstructure and can be assumed as EDP.

Fig.s 4-5 illustrate the IDA results regarding both the superstructure response in terms of IDI and the isolation level response $u_b$ obtained for different values of the system parameters varying in the range of interest. Each figure contains several surface plots, corresponding to different values of percentile (50$^{th}$, 84$^{th}$ and 16$^{th}$). Fig. 4 shows the IDA results regarding the superstructure response. The lognormal mean and dispersion decrease for higher values of $T_b$ and lower values of $T_s$ (high value of the isolation degree). Fig. 5 shows the IDA results re-

garding the isolation level response $u_b$. The lognormal mean and dispersion also decrease for higher values of $T_b$ (high value of the isolation degree) and for lower values of $T_s$.

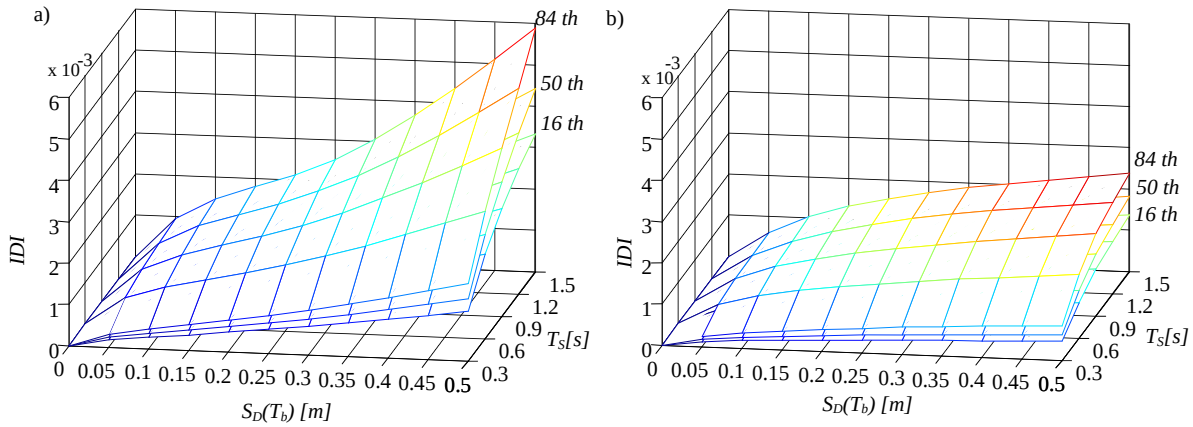



Figure 4: IDA curves of the superstructure 1st floor with $\gamma$=0.7, for $R$=1m (a) and $R$=4m(b).

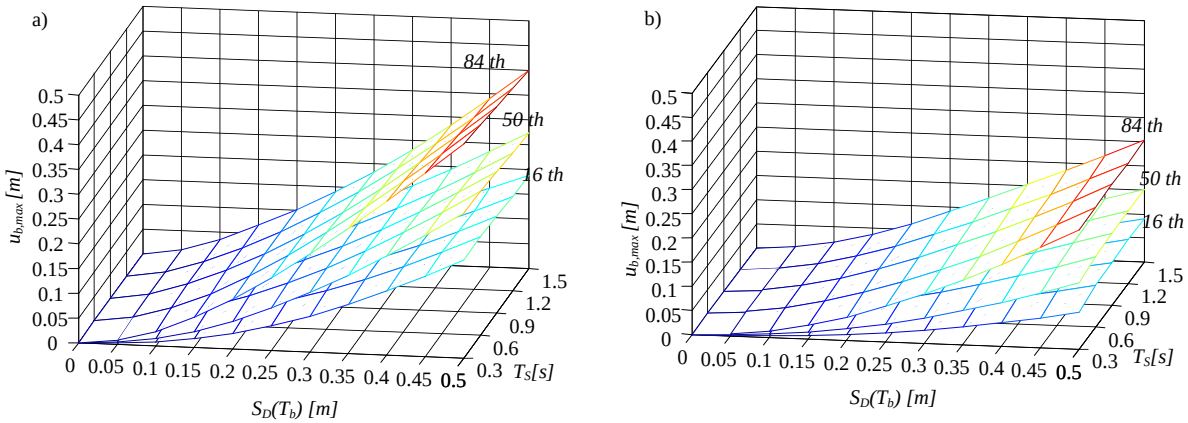



Figure 5: IDA curves of the isolation level with $\gamma$=0.7, for $R$=1m (a) and $R$=4m (b).

## 5 SEISMIC FRAGILITY OF STRUCTURES WITH FP DEVICES

This section describes the evaluation of the probabilities $p_f$ exceeding different limit states related to both the superstructure and the isolation level at each value of the *IM* defining the corresponding seismic fragility curves.

With reference to performance levels of the superstructure, four discrete performance levels or limit states (*LS1*,*LS2*,*LS3*,*LS4*), corresponding respectively to "fully operational", "operational", "life safety" and "collapse prevention" are provided from [34]. The performance limit states for base-isolated buildings, in accordance to provisions [48], have been defined by limiting the response of the lateral-load-resisting superstructure system, IDI limits, to a fraction of the limits provided for designing comparable fixed-base buildings [33].

| | *LS1 fully operational* | *LS2 operational* |
|---|---|---|
| *Inter-story drift (ISD) index* | 0.1% | 0.2% |
| $p_f$*(50 years)* | $5.0 \cdot 10^{-1}$ | $1.6 \cdot 10^{-1}$ |

Table 1: Limit state thresholds for the superstructure [33]-[48].

In Table 1, the *LS1* and *LS2* thresholds assumed for the seismic fragility of the superstructure as well as the corresponding failure probabilities in a design life of 50 years are reported depending on the limit state. At each value of the intensity measure *IM*, the probabilities $p_f$ exceeding different limit states related to the superstructure have been numerically computed for each considered combination of the superstructure/isolation level properties, as shown in Fig.s 6-7.

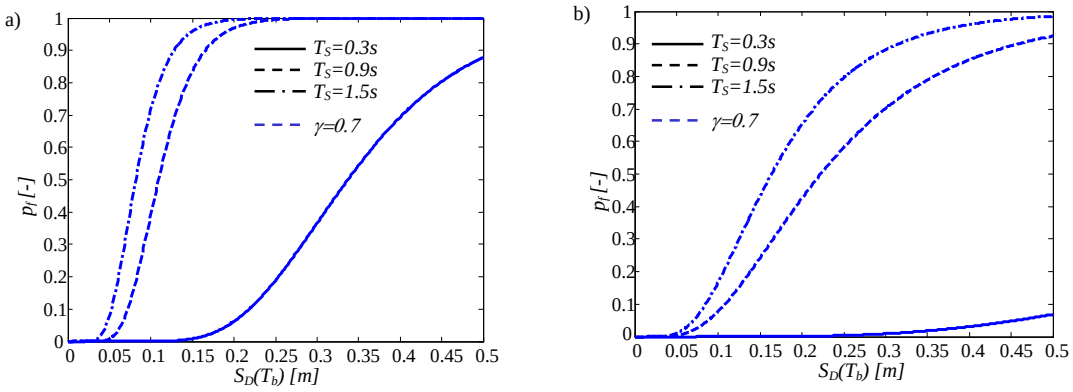



Figure 6: Seismic fragility curves of the superstructure 1st floor related to *LS1*, for *R*=1m (a), *R*=4m (b).

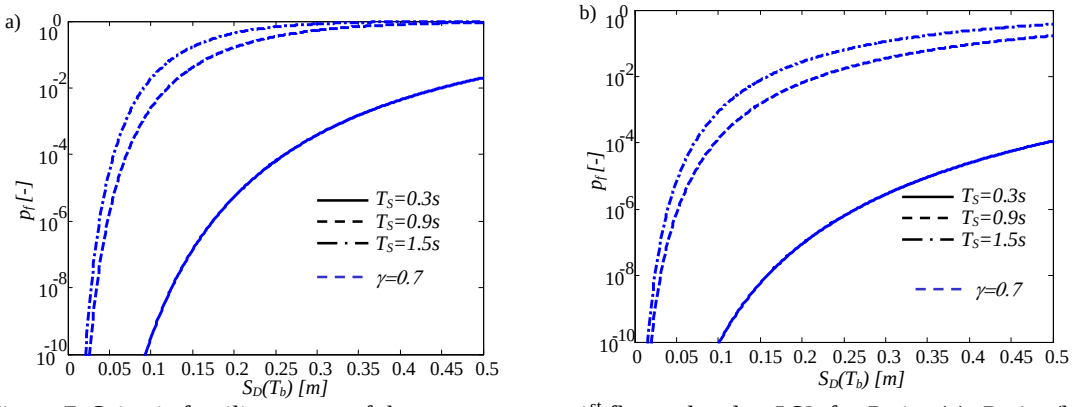



Figure 7: Seismic fragility curves of the superstructure 1st floor related to *LS2*, for *R*=1m (a), *R*=4m (b).

With reference to the performance levels of the isolation system, several different values for the plan dimension of the isolator (i.e. radius in plan of the concave surface), are considered. In Table 2, the limit state thresholds assumed for the seismic fragility of the FPS isolation level are reported.

| | *LS1* | *LS2* | *LS3* | *LS4* | *LS5* | *LS6* | *LS7* | *LS8* | *LS9* |
|---|---|---|---|---|---|---|---|---|---|
| *Maximum relative displacement [m]* | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 | 0.35 | 0.4 | 0.45 | 0.5 |

Table 2: Limit state thresholds for the isolation level.

Similarly, at each value of the *IM*, the probabilities $p_f$ exceeding different limit states related to the isolation level have been numerically computed for each combination of the structural properties. Afterward, the abovementioned exceeding probabilities $p_f$ have been fitted by a lognormal distribution. Fig.s 8-9 show the fragility curves regarding the isolation level for three values of $T_s$ (0.3s, 0.9s and 1.5s) and two different values of the limit state thresholds: *LS5*-*LS9*. The seismic fragility of the isolation level increases for higher values of $T_s$.

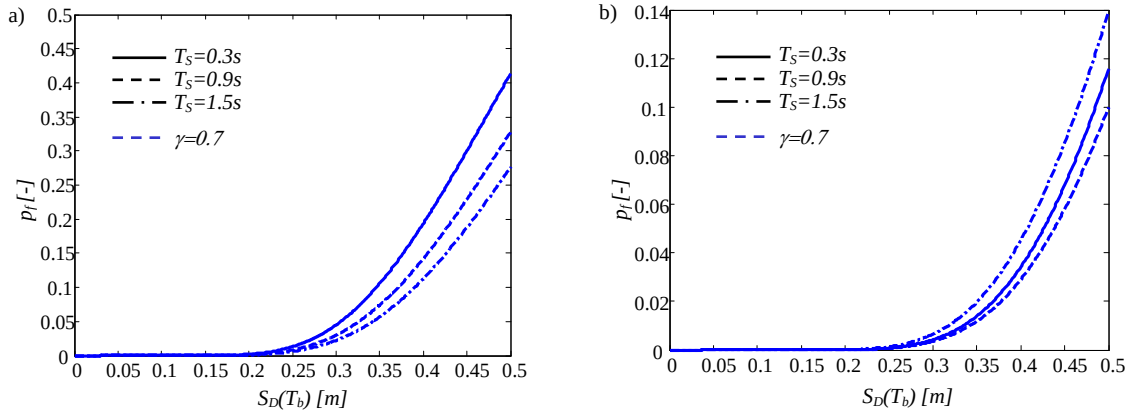



Figure 8: Seismic fragility curves of the isolation level related to *LS5*, for *R*=1m (a), *R*=4m (b).

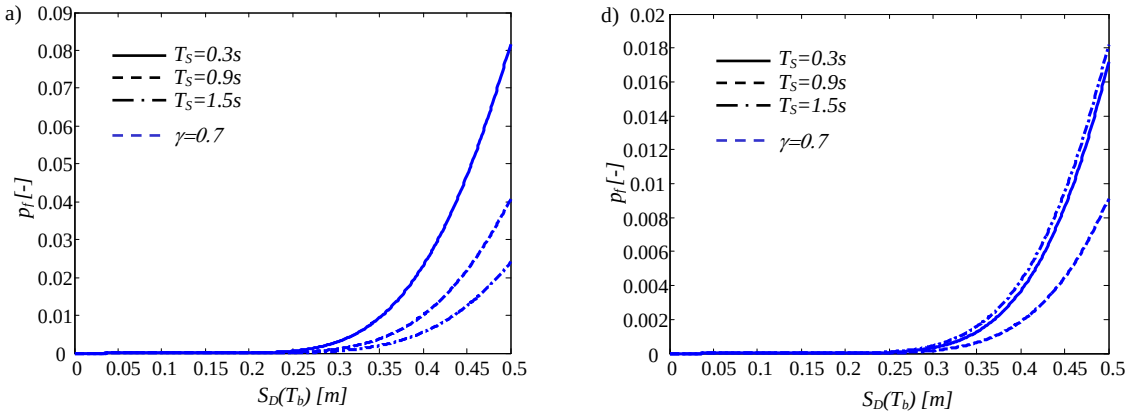



Figure 9: Seismic fragility curves of the isolation level related to *LS9*, for *R*=1m (a), *R*=4m (b).

## 6 SEISMIC RELIABILITY OF STRUCTURES WITH FP DEVICES

Considering L'Aquila site as the reference site, in Figure 10 the seismic hazard curves, expressed in terms of the same $IM = S_D(T_b)$, related to the different isolated periods analyzed in the parametric study are plotted according to NTC08. Each curve represents the average values of the annual rate λ of exceeding the $IM = S_D(T_b)$ level.

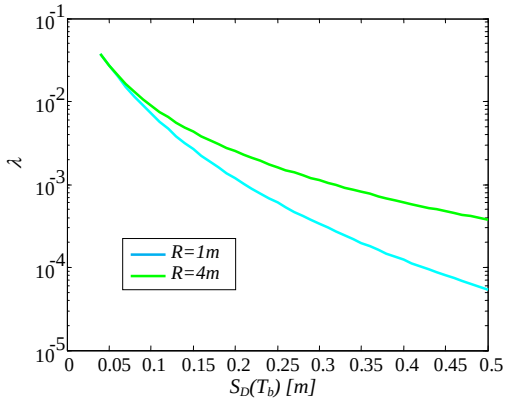



Figure 10: Seismic hazard curves related to the different isolated periods $T_b$ for a site near L'Aquila (Italy).

Integrating the fragility curves related to the superstructure with the seismic hazard curves and using a Poisson distribution, it is possible to evaluate the seismic reliability of the superstructure in the time frame of interest (50 years) for different values of the superstructure properties and having assumed the friction coefficient and soil dynamic parameters as random

variables. The seismic reliability of the superstructure increases for low values of $T_s$ (high values of the isolation degree), as shown in Figure 11. The results are consistent with those discussed by [21].

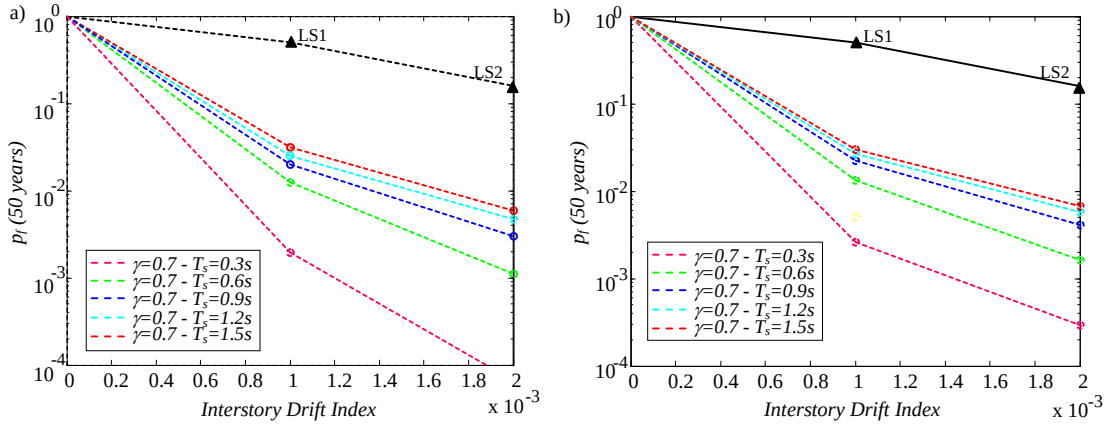



Figure 11: Seismic reliability curves of the superstructure 1st floor for $R$=1m (a), $R$=4m (b).

The seismic reliability of the isolation level decreases as $R$ increases and slightly depends on the values of $T_s$ (Fig. 12). Since the isolation level is not strongly influenced by the higher modes of the superstructure, the derived reliability-based abacuses are useful to design FP bearing devices depending on the properties of the superstructure and the expected reliability level in an area with a seismic hazard similar to that considered. In fact, an exceeding probability of $p_f$=1.5·10$^{-3}$ (related to collapse limit state, $\beta$=3 in 50 years) is achieved through a radius in plan $r$ ranging from about 0.2 m to about 0.4 m depending on system properties. The results are consistent with the monovariate structural performance curves in [21].

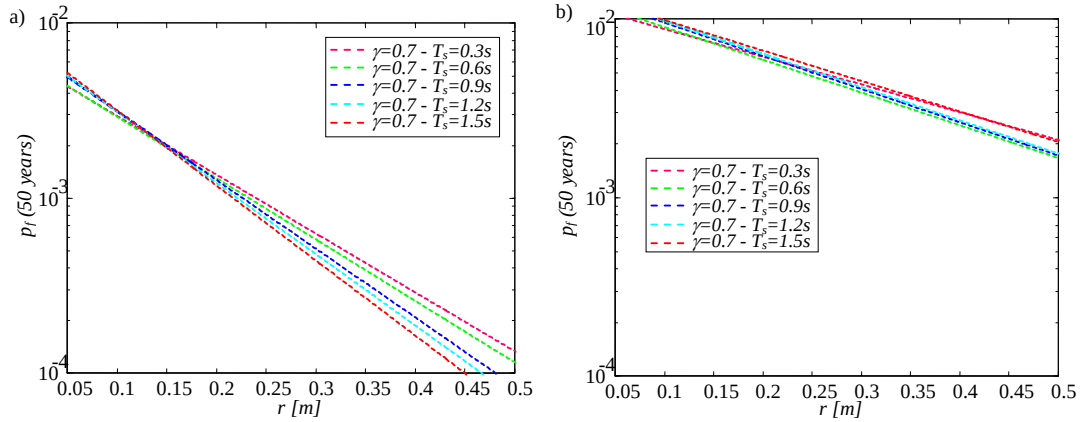



Figure 12: Seismic reliability curves of the isolation level for $R$=1m (a), $R$=4m (b).

## 7 CONCLUSIONS

This paper deals with the seismic reliability of structural systems equipped with friction pendulum isolators (FPS) by presenting the fragility curves related to an extensive parametric study encompassing a wide range of building properties, different seismic intensity levels and considering both the friction coefficient and earthquake characteristics as random variables. The uncertainty in the seismic inputs is taken into account by considering a set of artificial records, obtained through the power spectral density method, with different characteristics depending on soil dynamic parameters. IDA are developed to evaluate the probabilities exceeding different limit states related to both superstructure and isolation level for different structural system properties. The estimates of the response statistics are used for deriving

seismic fragility curves for the superstructure and the isolation level assuming different values of the corresponding limit states. In the final part, considering the seismic hazard curve related to a site near to L'Aquila (Italy), according to NTC08, and regarding a structure isolated by FPS with a design life of 50 years, reliability-based abacuses are derived with the aim to design the radius in plan of the friction pendulum isolators.

## REFERENCES


[1] C. Christopoulos, A. Filiatrault, *Principles of Passive Supplemental Damping and Seismic Isolation*. IUSS Press: Pavia, Italy, 2006.

[2] M. De Iuliis, P. Castaldo, B. Palazzo, Analisi della domanda sismica inelastica del terremoto de L'Aquila su sistemi dimensionati secondo le NTC2008. *Ingegneria Sismica*, Patron Editore XXVII(3), 55–68, 2010.

[3] V.A. Zayas, S.S. Low, S.A. Mahin, A simple pendulum technique for achieving seismic isolation. *Earthquake Spectra*, 6, 317-333, 1990.

[4] A. Mokha, M.C. Constantinou, A.M. Reinhorn, Teflon Bearings in Base Isolation. I: Testing. *J. Struct. Eng*, **116**(2), 438-454, 1990.

[5] P. Castaldo, *Integrated Seismic Design of Structure and Control Systems*, Springer International Publishing: New York, 2014.

[6] P. Castaldo, M. De Iuliis, Optimal integrated seismic design of structural and viscoelastic bracing-damper systems, *Earthquake Engineering and Structural Dynamics*, **43**(12), 1809–1827, 2014.

[7] M. De Iuliis, P. Castaldo, An energy-based approach to the seismic control of one-way asymmetrical structural systems using semi-active devices, *Ingegneria Sismica - International Journal of Earthquake Engineering* XXIX(4), 31-42, 2012.

[8] Giugliano M.T., Longo A., Montuori R., Piluso V., Plastic design of CB-frames with reduced section solution for bracing members. *Journal of Constructional Steel Research.* Vol. 66 pp 611-621. 2010.

[9] Montuori R., Nastri E., Piluso V. Theory of plastic mechanism control for the seismic design of braced frames equipped with friction dampers - *Mechanics Research Communications* – Vol. 58. pp. 112-123 2014.

[10] Montuori R., Piluso V., Troisi M. Innovative structural details in MR-frames for free from damage structures *Mechanics Research Communications* –58, 146-156, 2014.

[11] Longo A., Montuori R., Piluso V., Theory of Plastic Mechanism Control of Dissipative Truss Moment Frames, *Engineering Structures*, Vol. 37, pp. 63-75, 2012.

[12] A. Longo, R. Montuori, V. Piluso, Failure mode control and seismic response of dissipative truss moment frames, *Journal of Structural Eng.*, **138**(11), 1388-1397, 2012.

[13] B. Palazzo, P. Castaldo, I. Marino, The steel column damper: a new hysteretic device providing additional stiffness and damping, *EUROSTEEL 2014*, Naples, Italy, September 10-12, 2014.

[14] B. Palazzo, P. Castaldo, I. Marino, The Dissipative Column: A New Hysteretic Damper, *Buildings* **5**(1), 163-178, 2015.

[15] M.T. Giugliano, A. Longo, R. Montuori, V.Piluso, Influence of homoschedasticity hypothesis of structural response parameters on seismic reliability of CB-frames, *Georisk*, **5**(2), 120-131, 2011.

[16] M.T. Giugliano, A. Longo, R. Montuori, V. Piluso, Seismic reliability of traditional and innovative concentrically braced frames, *Earthquake Engineering and Structural Dynamics*, **40**(13), 1455–1474, 2011.

[17] A. Longo, R. Montuori, V. Piluso, Seismic reliability of V-braced frames: Influence of design methodologies, *Earthquake Engineering and Structural Dynamics*, **38**(14), 1587–1608, 2009.

[18] J. Chen, W.Liu, Y. Peng, J. Li, Stochastic seismic response and reliability analysis of base-isolated structures. *J Earthquake Eng*, **11**, 903-924, 2007.

[19] F.G. Fan, G. Ahmadi, Random response analysis of frictional base isolation system. *J Eng Mech*,**116**(9), 1881 ‑ 1901, 1990.

[20] L.R. Barroso, S. Winterstein, Probabilistic seismic demand analysis of controlled steel momentresisting frame structures. *Earthquake Engineering and Structural Dynamics*, **31**(12), 2049–2066, 2002.

[21] P. Castaldo, B. Palazzo, P. Della Vecchia, Seismic reliability of base-isolated structures with friction pendulum bearings. *Engineering Structures*, **95**, 80-93, 2015.

[22] B. Palazzo, P. Castaldo, P. Della Vecchia, Seismic reliability analysis of base-isolated structures with friction pendulum system, *IEEE Workshop on Environmental, Energy and Structural Monitoring Systems Proceedings*, Napoli, September 17-18, 2014.

[23] P. Castaldo, E. Tubaldi, Influence of FPS bearing properties on the seismic performance of base-isolated structures, *Earthquake Engineering and Structural Dynamics* , **44**(15), 2817–2836, 2015.

[24] R. Jangid, Stochastic Response of Bridges seismically isolated by friction pendulum system. *J. Bridge Eng*.**13**(4), 319-330, 2008.

[25] M. Shinozuka, G. Deodatis, Simulation of stochastic processes by spectral representation. *Applied Mechanics Reviews*, **44**(4), 191-203, 1991.

[26] P. Pinto, R. Giannini, P. Franchin, *Seismic Reliability Analysis of Structures*. Iuss Press, 2004.

[27] D.G. Talaslidis, G.D. Manolis, E.A. Paraskevopoulos, C.G. Panagiotopoulos, Risk analysis of industrial structures with hazardous materials under seismic input, *13th World Conference on Earthquake Eng.*, Vancouver, B.C., Canada, August 1-6, 2004.

[28] C.A. Cornell, H. Krawinkler, Progress and challenges in seismic performance assessment. *PEER Center News*, **4**(1), 1-3, 2000.

[29] NTC08. Norme tecniche per le costruzioni. Gazzetta Ufficiale del 04.02.08, DM 14.01.08, Ministero delle Infrastrutture.

[30] M.C. Constantinou, A. Mokha, A.M. Reinhorn, Teflon Bearings in Base Isolation. II: Modeling. *J. Struct. Eng*. **116**(2), 455-474, 1990.

[31] M.C. Constantinou, A.S. Whittaker, Y. Kalpakidis, D.M. Fenz, G.P. Warn, Performance of Seismic Isolation Hardware Under Service and Seismic Loading, Technical Report MCEER-07-0012, 2007.

[32] J.M. Kelly, *Earthquake-Resistant Design with Rubber, 2nd Edition*. Springer-Verlag (1997).

[33] R.D. Bertero, V.V. Bertero, Performance-based seismic engineering: the need for a reliable conceptual comprehensive approach. *Earthquake Engineering and Structural Dynamics*, **31**, 627–652, 2002.

[34] SEAOC Vision 2000 Committee. Performance-based seismic engineering. Report prepared by Structural Engineers Association of California, Sacramento, CA., 1995.

[35] T. Saito, J. Kanda, N. Kani, Seismic reliability estimate of building structures designed according to the current Japanese design code. *Proceedings of the Structural Engineers World Congress*, 1998.

[36] K.A. Porter, An overview of PEER's performance-based earthquake engineering methodology. *Proceedings, Proceedings of the 9th International Conference on Application of Statistics and Probability in Civil Engineering (ICASP9)*, San Francisco, California, 2003; 973-980.

[37] K. Kanai, Semiempirical formula for the seismic characteristics of the ground. *Bulletin of earthquake research institute* Vol. 35, 309-325, 1957.

[38] H. Tajimi, A statistical method of determining the maximum response of a building structure during an earthquake. *Proc., 2nd World Conf. on Earthquake Eng*, Vol.II, 781-798, 1960.

[39] R.W.Clough, J. Penzien, *Dynamics of Structures*, *2nd Edition*. McGraw-Hill,1993.

[40] European Committee for Standardization. Eurocode 8-Design of Structures for Earthquake Resistance. Part 1: General Rules, Seismic Actions and Rules for Buildings, Brussels, 2004.

[41] M. Shinozuka, Y.Sato, Simulation of nonstationary random process. *J. Engrg.Mech.Div*. **93**(1), 11-40, 1967.

[42] M.D. Mckey, W.J. Conover, R.J. Beckman, A comparison of three methods for selecting values of input variables in the analysis from a computer code. *Technometrics* , **21**, 239-45, 1979.

[43] D. Celarec, M. Dolšek, The impact of modelling uncertainties on the seismic performance assessment of reinforced concrete frame buildings. *Engineering Structures*, **52**, 340 – 354, 2013.

[44] D. Vamvatsikos, C.A. Cornell, Incremental dynamic analysis. *Earthquake Engineering and Structural Dynamics*, **31**(3), 491–514, 2002.

[45] N. Luco, C.A. Cornell, Structure-specific scalar intensity measures for near-source and ordinary earthquake ground motions. *Earthquake Spectra*, **23**(2), 357-92, 2007.

[46] K.L. Ryan, A.K. Chopra, Estimation of Seismic Demands on Isolators Based on Nonlinear Analysis. *J. Struct. Eng.*, **130**(3), 392–402, 2004.

[47] B. Palazzo, Seismic Behavior of base-isolated Buildings. *Proc. International Meeting on earthquake Protection of Buildings*, Ancona, 1991.

[48] Building Seismic Safety Council. NEHRP commentary on the guidelines for the seismic rehabilitation of buildings. Provisions (FEMA 274). Washington, DC, 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# SEISMIC RETROFIT OF EXISTING BUILDINGS THROUGH THE DISSIPATIVE COLUMNS

**Paolo Castaldo [(1)], Bruno Palazzo [(1)], Francesco Perri [(1)], Ivana Marino [(1)], Marco Maria Faraco [(2)]**

[1] Department of Civil Engineering, University of Salerno.
via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy
e-mail: {pcastaldo,palazzo,fperri,imarino}@unisa.it

[2] Department of Civil Engineering, University of Salerno.
via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy
e-mail: mfaraco@hotmail.it

**Keywords:** Hysteretic Damper, Seismic Damage Reduction, Passive Energy Dissipation

**Abstract.** *A new replaceable hysteretic damper to better control seismic building damage, consisting of two or more adjacent steel vertical elements (columns) connected to each other with continuous mild/low strength steel shear links, is investigated in this study. New Dampers, called Dissipative Columns (DC), are continuously linked with X-shaped steel plates and provide additional stiffness and damping to a lateral system. The Dissipative Column has been conceived as a device installed within a frame or as an external damper to provide macro-dissipation. In fact, considering different configurations, a parametric analysis is developed, through non-linear pushover and cyclic analyses carried out in ABAQUS, in order both to evaluate the effect of the main geometrical and structural parameters and provide the design capacity curves of this new damper. Moreover, non linear dynamic analyses of an existing building without and with the Dissipative Columns have been performed in SAP2000 in order to evaluate the supplemental damping provided by the new damper. The DC can be considered a new damping device, easy to install in new as well as existing buildings in order to protect them from seismic damage.*

## 1 INTRODUCTION

Strong earthquakes have shown that a large percentage of buildings in the affected areas, even if properly built and designed according to the most advanced codes, suffer such severe damages [1] that they need to be demolished after the quake, since they would be expensive to repair. As is known, the acceptance of such level of damage due to severe earthquakes is related to the ductility-based design criteria that assume design seismic actions decrease by reduction factors. This approach may lead to high social and economic costs to the affected communities, and to a long recovery time for essential services and production activities. Inspired by new performance criteria, there is a growing belief that code design criteria are not sustainable for the high level of accepted damage, are impossible to repair, and that common buildings should be designed with a higher performance level. At the beginning of this century, Performance Based Engineering [2] introduced new principles with the scope to select more articulated targets better corresponding to different building roles and use, defining a variety and complex subdivision of performance objectives for seismic events with different intensities and frequencies of occurrence. The "Direct Displacement Based Design" philosophy [3] relates the specified performance level to the strain or drift limits for a specified seismic intensity. With the scope of minimizing structural damage, several frictional isolation devices [4]-[6] and dampers [7]-[8], new replaceable hybrid composite or steel devices [9]-[14] have been recently proposed as well as integrated design approaches [15]-[16] and new strategies, i.e., based on the collapse mechanism control [17]-[30] or energy balance [31], have been developed in order to dissipate seismic input energy outside of the primary structure. Dampers should absorb a significant portion of the input energy reducing the hysteretic energy demand to the primary structural elements. Another damper employed to dissipate energy dynamic energy through stable hysteretic behavior [32] consists in the buckling-restrained braces (BRBs). In [33], Low Yielding Strength (LYS) steel is investigated for improving the ductility capacity of box-shaped steel bridge piers and experimental work is carried out for four specimens having different thickness and sectional configurations for cyclic loading test. The test results reveal that the LYS steel portion with longitudinal stiffeners greatly improves the strength and ductility capacity of box columns and it is observed that LYS steel has a great cyclic strain-hardening characteristic. The advantage of use of LYS steel is that it can effectively use large plastic deformation in component plates and the failure of column is concentrated at the LYS steel segment and the energy dissipation occurs far beyond the yield point.

The aim of this paper is to investigate a new replaceable hysteretic damper having a basic form of the art of building, minimally architecturally invasive, consisting of two or more dissipative steel columns directly connected to two floors linked to each other with X-shaped low/mild steel plates. It will be shown that the new element is able to add significant stiffness and damping to the structural system to reduce seismic response and damage in primary structural members under severe earthquakes. A similar concept is used when Reduced Beam Sections (RBS) are realized at beam ends of a steel moment resisting frames. In fact, in this case the beam section resistance reduction is obtained by trimming the flanges, realizing the so called "dog-bone" [34]-[35]. The Dissipative Columns (DC) element will be investigated through non-linear pushover and cyclic analyses in ABAQUS [36]-[37] in order to characterize the local and global behaviour of the device considering different steel grades. These analyses are useful to design the yielding properties of the proposed damper depending both on the characteristics of the primary structure and the expected performance as discussed by [38]. The efficiency of the proposed damper applied on the benchmark system is investigated through time history analyses in SAP2000 [39].

## 2 MECHANICAL PRINCIPLES OF THE DISSIPATIVE COLUMN

The Dissipative Column model, as shown in Figure 1, can be considered as a sort of framed bi-pendulum with height equal to $H$, connected in parallel to the primary structure, able to react to the story drift $\Delta$ with a lateral force $Q_D$ adding stiffness, strength and damping [12]-[13]. The design concept of the DC element aims to obtain a lever mechanism by which a small inter-story drift provides an amplified vertical drift between the X-plate ends (Figure 1) reacting with shear forces. The X-shaped steel plates made of mild or Low Yielding Strength steel, having length $a$, thickness $t$, width at the ends $b$ and vertical distance $i$, are also used as shear links between coupled elements.

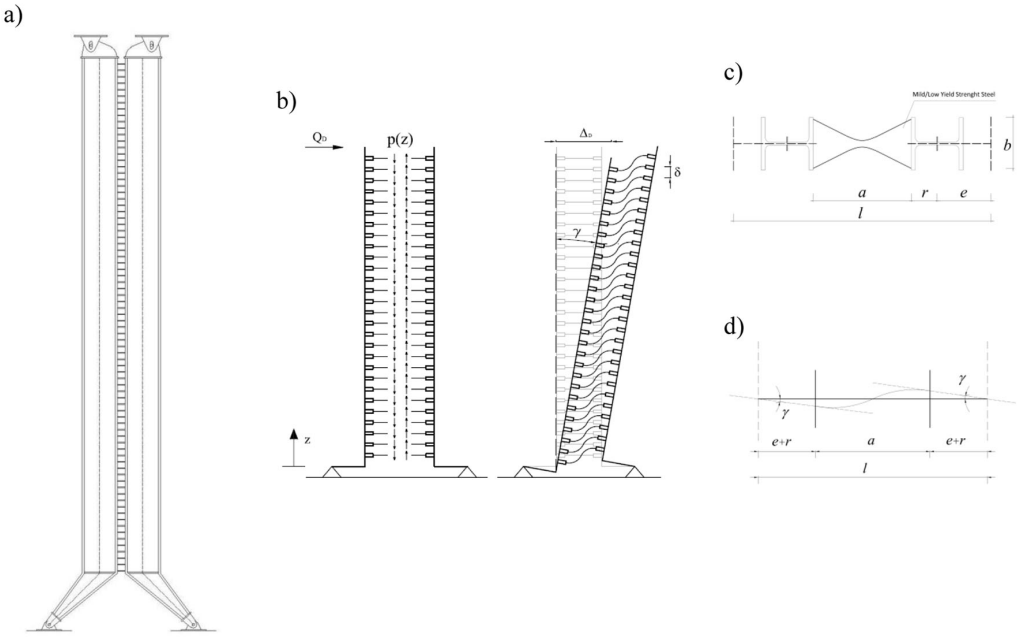



Figure 1: Eccentric Dissipative Column (a), structural model (b) and dissipative amplified mechanism (c)-(d).

Each lever arm is characterized by an eccentricity $e$, while, $r$ is the rigid element representing half section of each column (Figure 1). The top ends of the model are linked to the upper floor through slotted bolted connections to allow large vertical displacements. By yielding a large volume of steel, the shear devices are able to dissipate substantial input energy during earthquakes, while also increasing damping in the entire system with the aim to reduce the damage in the primary system. With reference to reinforced concrete structures, the limit values of Inter-Story Drift Angle (ISDA) corresponding to different structural performance levels are suggested by Ghobarah [40]. The great advantages of the DC, if compared with classical steel dissipative braces [32], are the reduced architectural invasiveness, so that it is able to be integrated in any building, the ease of installation everywhere, replacement after earthquakes and the stable behavior in cyclic reversal deformation. To the scope to this study, two steel are considered to model the ADAS mechanical properties. In particular, S235 and LYP100 (Low Yield Point) are implemented, while the DC are totally conceived in S355.

## 3 SIMPLIFIED MECHANICAL BEHAVIOUR OF THE DC

As extensively discussed in [13], the vertical drift $\delta$ between the ends of a generic shear link in the elastic range, being the curvature $\chi$ constant along each half plate, is related to shear force $V_d$ developed by each X-shaped plate, as:

$$\delta = 2\int_0^{a/2}\int_0^{a/2} \chi(x)dxdx = \frac{3}{2}\frac{a^3}{Ebt^3}V_d \tag{1}$$

where $a$, $t$, $b$ are, respectively, the X-plate length, thickness, width at the ends and $E$ is Young's modulus of the steel plates. The axis $x$ represents the barycentric axis of a generic steel plate. Hence, the single X-plate vertical stiffness is equal to:

$$K_d = \frac{V_d}{\delta} = \frac{2}{3}\frac{Ebt^3}{a^3} \tag{2}$$

In the case of small eccentricity, a simplified analysis of the DC behavior subjected to relative displacements can be easily carried out assuming that the column flexural deformation is negligible respect to the case of flexible inextensible links. An inter-story drift produces a shear drift angle $\gamma$ and vertical drifts $\delta$ along the X-shaped steel plate. Under such simplified assumptions, for the equilibrium, the top-base relative displacement $\Delta$ of a DC element with height $H$ is related to the drift angle $\gamma$ as:

$$\Delta = \gamma \cdot H \tag{3}$$

Therefore, each X-plate undergoes a vertical drift equal to:

$$\delta = (l-a)\Delta/H \tag{4}$$

where $l$ represents pin axes distance (Figure 1). Therefore, the uniform distributed vertical load, along the vertical axis having origin in the hinge (Figure 1), due to the shear drift angle $\gamma$ applies:

$$p = \frac{2}{3}\frac{Ebt^3}{a^3 i}(l-a)\Delta/H \tag{5}$$

where $i$ represents the plates vertical distance. The term $(l-a)/2 = r$ represents a small lever arm that can be amplified using eccentricity $e$ between the vertical axis and the supports as will be shown in the following. The axial force at the base of each column is equal to:

$$N_c = \frac{2}{3}\frac{Ebt^3}{a^3 i}(l-a)H\Delta/H \tag{6}$$

According to the experimental tests [41]-[42], the load-deformation curve of the X-shaped mild steel plates can be idealized as a bilinear curve with a ratio of post yielding stiffness to the initial one equal to 0.03 and available displacement ductility ratio $\mu=\delta/\delta_y$ varying in the range between 3 and 5 [43]. Since yielding strength $f_y$ is reached almost uniformly along the device, the yielding vertical load $p_y$ can also be expressed as:

$$p_y = \frac{2M_{d,y}}{ia} = f_y\frac{bt^2}{2ia} \tag{9}$$

while, the relative yielding displacement of each link can be written as:

$$\delta_y = \frac{V_{d,y}}{K_d} = f_y\frac{3a^2}{4Et} \tag{10}$$

Therefore, the lateral yield strength of the doubly hinged DC element can be evaluated as:

$$Q_{Dy} = f_y \frac{blt^2}{2ia} \tag{11}$$

and the yielding displacement applies:

$$\Delta_y = \frac{3}{4}\frac{f_y}{E}\frac{a^2 H}{t(l-a)} \tag{12}$$

Moreover, in presence of a significant eccentricity the column flexural deformation should not be neglected, therefore, as extensively described in [13], the top-base relative displacement Δ and lateral force of the DC element can respectively be expressed as:

$$\Delta = \gamma \cdot H + Q_D\left(\frac{eH^3}{3EI_c l} + \frac{e^2 H^2}{2EI_e l}\right) \tag{13}$$

$$Q_D = \frac{N_c l}{H} = \frac{\frac{2}{3}\frac{Ebt^3}{a^3 i}\gamma(l-a)H\frac{l}{H}}{1-\left[\frac{2}{3}\frac{Ebt^3}{a^3 i}\frac{l}{H}\left(2\frac{H^2}{l}\frac{e^3}{3EI_e}+\frac{H^2}{l}\frac{e^2}{EI_e}r+\frac{e}{l}\frac{H^3}{3EI_c}r\right)\right]} \tag{14}$$

where $K_D$ represents the lateral stiffness of the Dissipative Column is given by:

$$K_D = \frac{2}{3}\frac{Ebt^3}{a^3 i}\frac{l}{H}(l-a) \tag{15}$$

Therefore, the stiffness ratio between the lateral stiffness of the Dissipative Column and the single X-plate vertical stiffness is equal to:

$$\frac{K_D}{K_d} = 2\frac{(a+2(e+r))}{H}(e+r) \tag{16}$$

## 4 NON LINEAR ANALYSIS OF THE X-SHAPE STEEL PLATES

Non-linear analyses in ABAQUS [36]-[37] are performed in order to characterize the non-linear behavior of the X-shape steel plates. In Table 1, the geometric and mechanical properties are reported with reference to two different steel grades: S235 [44] and LYP100 [33],[45].

| Plate Length $a$ [mm] | Plate Thickness $t$ [mm] | Plate Width $b$ [mm] | Plate Yield Stress $f_y$ [N/mm$^2$] |
|---|---|---|---|
| 150 | 15 | 120 | 132 |
| 150 | 15 | 120 | 235 |

Table 1: Geometrical and structural properties of the X-shaped steel plates.

Figure 2 shows the geometric shape as well as the deformed shape of the steel plates. A FE model has been defined in ABAQUS [36]-[37] with means of 3D elements with nonlinear stress-strain behaviour. The geometric model proposed by [46] has been adopted for the X-shape steel plates.

The displacement-controlled pushover analyses of the X plates have been performed until to reach 50 mm in terms of relative displacement and the corresponding curves related to the two different steel grades are plotted in Figure 3.

The results clearly show that in the case of LYP100 steel grade the yielding point has been firstly reached against to the case of S235 steel grade, but it is also evident that, in this case, the stiffness is larger with the effect that larger is, in the global model, as detailing described in the following, the additional shear forces on the steel vertical elements of the DC.

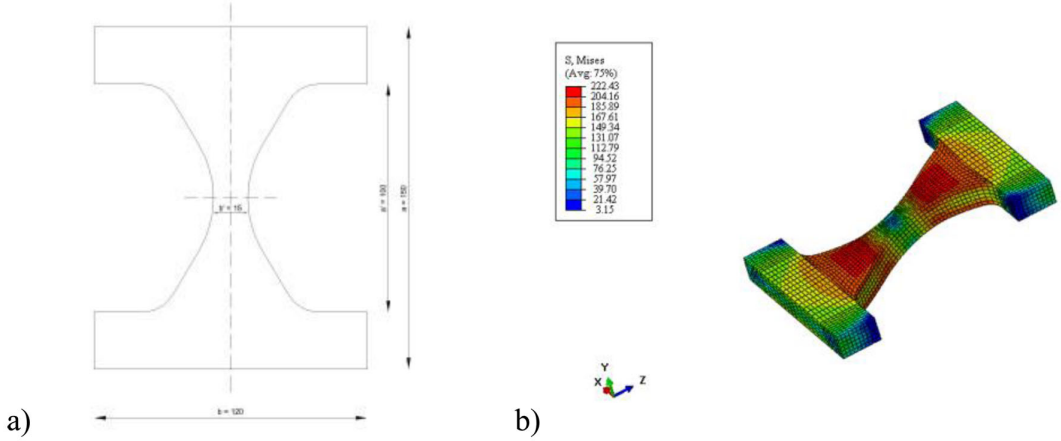



Figure 2: Geometric shape (a) and deformed shape (b) of a X-shaped steel plate.

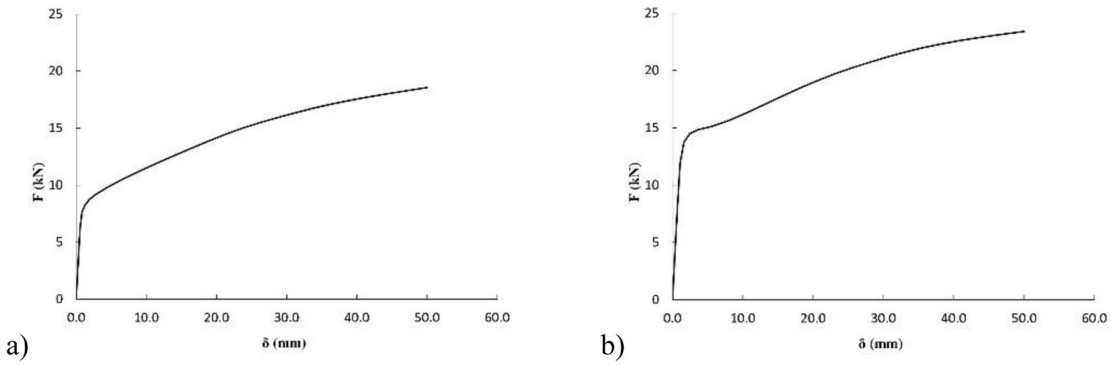



Figure 3: Pushover curves of the X plates: (a) LYP100 steel grade [33],[45]; (b) S235 steel grade [44].

## 5 NON LINEAR ANALYSIS OF THE DISSIPATIVE COLUMN

Defined the non-linear behavior of the X-shape steel plates, non-linear static and dynamic analyses of the Dissipative Column have been performed in ABAQUS [36]-[37]. In Table 2, the main geometric and mechanical properties of the DC are reported.

| Height H | Column Distance *l* | Plate Length *a* | Plate Thickness *t* | Plate Distance *i* | Plate Width *b* | LYP 100 X-shape Yield Stress $f_y$ | S235 X-shape Yield Stress $f_y$ | Eccentricity *e* | Column Profile |
|---|---|---|---|---|---|---|---|---|---|
| [mm] | [mm] | [mm] | [mm] | [mm] | [mm] | [N/mm$^2$] | [N/mm$^2$] | [mm] | |
| 10500 | 1650 | 150 | 15 | 125 | 120 | 132 | 235 | 400 | ISE700 |

Table 2: Geometrical and structural properties of the Dissipative Column (DC).

A FE model has been defined in ABAQUS with means of 3D elements with nonlinear stress-strain behaviour [36]-[37]. As shown in Figure 4, in ABAQUS [36]-[37] it has been possible to carefully model the restrains at the base and the top of DC as well as the contact between the pin and the slotted holes. In particular, the hinge at the top of the DC (Figure 4a) has been modeled implementing an appropriate slotted hole in the superior plate in order to

simulate the sliding due to the applied displacement to the pin causing translation of the base system. In Figure 4(b) it is also shown the connection at the base of the DC. The FE points belonging to the internal surface of the inferior hinges have rigidly been connected to the reference point located in the center of the pin with the aim to simulate the free rotation in the plane.

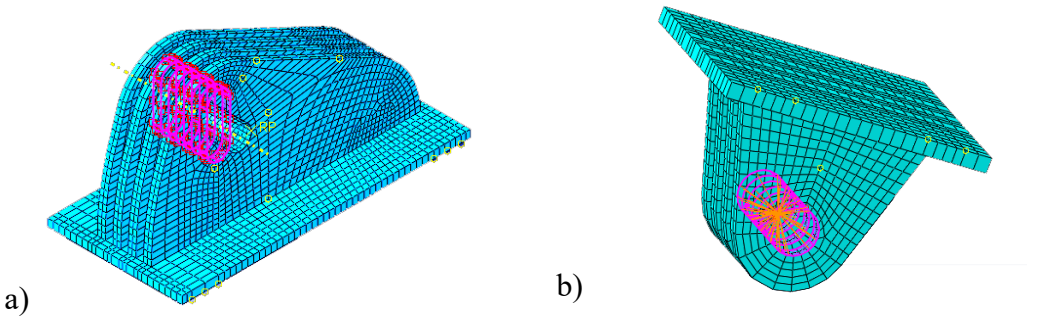



Figure 4: (a) Connection pin-hinge at the top of the DC; (b) Fix hinge at the base of the DC modeled in ABAQUS [36]-[37].

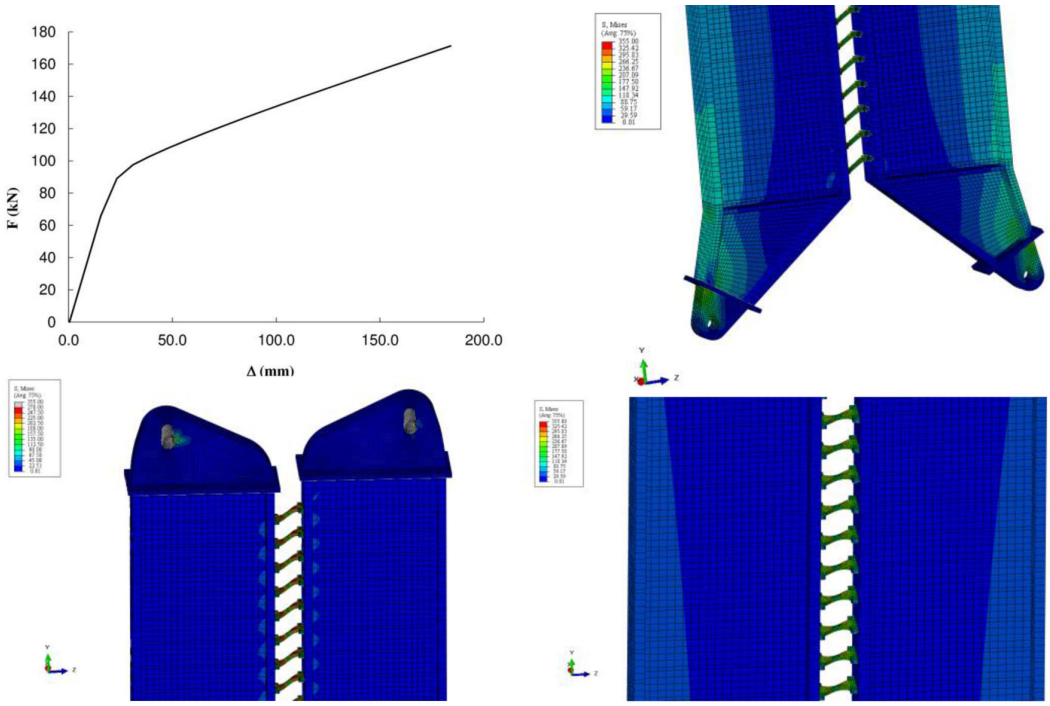



Figure 5: Pushover curve of the DC, using LYP100 X-plate, and deformed shape in ABAQUS [36]-[37].

Within nonlinear static analyses, in order to simulate the mechanical properties of the steel elements, a quadrilinear constitutive law [47] for S235 steel (X-shape plates) and S355 steel (vertical elements DC), while, a trilinear constitutive law regarding the LYP100 steel (X-shape plates) have been implemented in ABAQUS [36]-[37]. Figure 5 illustrates the geometric shape as well as the deformed shape of the DC. The displacement-controlled pushover analyses of the DC using LYP100 X-shape has been carried out until to reach 180 mm in terms of relative displacement as plotted in Figure 5, showing a yielding point at about 20 mm (0.17%). Figure 6 shows the comparison between the push-over curves of DC using LYP100 and S235 X-shape plates. The results demonstrate that in the case of LYP100 X-shape plates

the yielding point has been firstly reached against to the case of S235 X-shape plates, and the shear forces are lower although the lateral global stiffness remains the same.

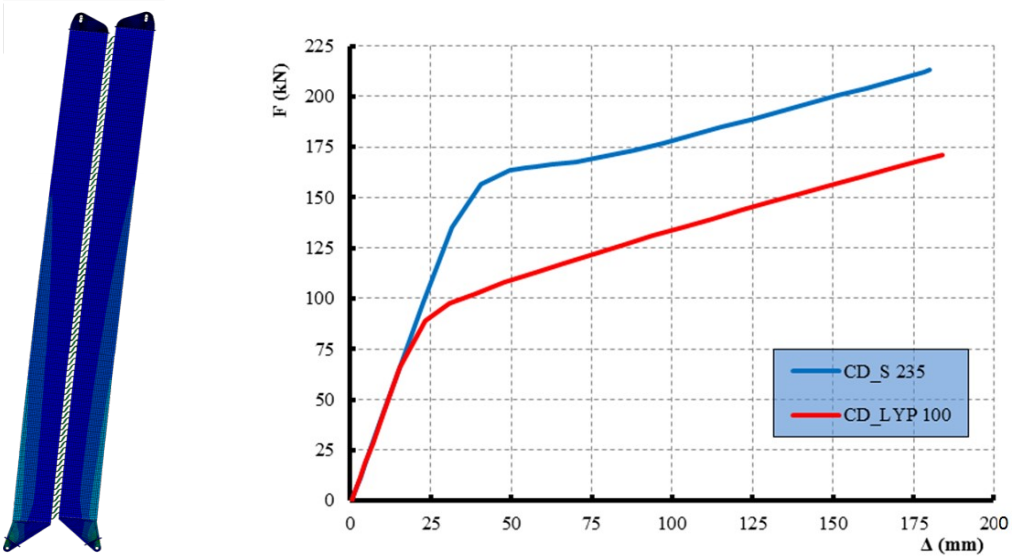
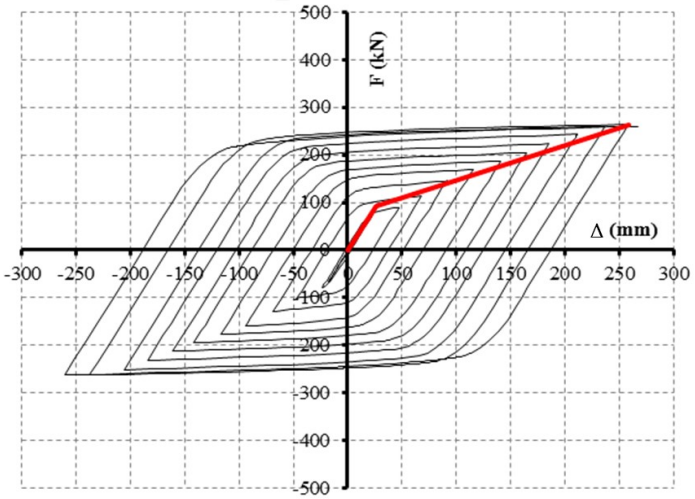



Figure 6: Pushover curve of the DC, using LYP100 and S235 X-shape plates, and deformed shape in ABAQUS [36]-[37].

In order to also simulate the behavior under seismic loads, the nonlinear behaviour of the X-shape plates has been described by a nonlinear kinematic hardening model in the case of nonlinear cyclic analyses. In particular, the nonlinear cyclic analyses has been performed until to reach a value of ductility equal to 10 in plastic range. As previously discussed, all the analyses of the DC are performed considering LYP100 and S235 X-shape plates. Figure 7 shows the results of displacement-controlled nonlinear cyclic analyses of the DC equipped with LYP100 X-shape plates. In particular, the hysteretic behaviour of the DC has been modeled considering the Bouc&Wen model [48]-[49] with an elastic stiffness $K_{el}$ and a ratio between the elastic and plastic stiffness $k = K_{el} / K_{pl}$. In Figure 7, the parameters of the Bouc&Wen model [48]-[49] are also reported for the DC with both LYP100 and S235 X-shape plates.




(a)

| | | |
|---|---|---|
| $K_{el}$ | 3560 | N/mm |
| $K_{pl}$ | 738 | N/mm |
| $k = K_{pl}/K_{el}$ | 0.207 | |
| *exp* | 10 | |

(b)

| | | |
|---|---|---|
| $K_{el}$ | 3630 | N/mm |
| $K_{pl}$ | 609 | N/mm |
| $k = K_{pl}/K_{el}$ | 0.168 | |
| *exp* | 10 | |

(c)

Figure 7: (a) Hysteretic behaviour of the DC using LYP100; Parameters of the Bouc&Wen [48]-[49] model for LYP100 (b) and S235 (c).

## 6 BUILDING MODEL EQUIPPED WITH THE DISSIPATIVE COLUMNS

As mentioned before, hysteretic dampers should be generally designed to yield before the primary structure, therefore their mechanical properties change along the building height. Assuming to have already established the shear strength of the primary system $Q^i_{Fy}$, mechanical properties of the i-th story may be summarized as follows [38],[50]:

$$Q^i{}_s = Q^i{}_{Fy} + Q^i{}_{Dy} \tag{17}$$

where $Q^i_S$, $Q^i_{Fy}$, and $Q^i_{Dy}$ represent the yield shear strength of the entire system, primary structure and damping system respectively at the i-th story. $Q^i_S$ and $Q^i_{Fy}$ correspond to the relative displacement $\Delta^i_{Fy}$, as well as $Q^i_{Dy}$ corresponds to $\Delta^i_{Dy}$. Defining $\Delta_{max}$ as the maximum story drift, $\mu_D$ and $\mu_F$ are, respectively, the story-drift ductility of the main frame and the story-drift ductility of the damper system. The lateral stiffness both of the primary structure and Dissipative Column system at the i-th story result being:

$$k^i{}_{Fy} = \frac{Q^i{}_{Fy}}{\Delta^i{}_{Fy}} \tag{18}$$

$$k^i{}_{Dy} = \frac{Q^i{}_{Dy}}{\Delta^i{}_{Dy}} \tag{19}$$

Defining the strength ratio β as the ratio of the yield strength of the damper system $Q^i_{Dy}$ to that of the entire system $Q^i_{Fy}$, $Q^i_{Dy}$ can be expressed as:

$$Q^i{}_{Dy} = \beta Q^i{}_s \tag{20}$$

$$Q^i{}_{Fy} = Q^i{}_s (1-\beta) \tag{21}$$

Moreover, the yield drift ratio ν can be defined as a ratio of damper yield drift to the one of the main structure, so, that the stiffness ratio $\kappa$ is defined as:

$$\kappa = \frac{K^i{}_{Dy}}{K^i{}_{Fy}} = \frac{Q^i{}_{Dy}}{\Delta^i{}_{Dy}} \frac{\Delta^i{}_{Fy}}{Q^i_{Fy}} = \frac{\beta}{(1-\beta)\nu} \tag{22}$$

The damper strength ratio β and the yield story drift ratio ν are the main parameters useful to design the damping system. Passive control is achieved by yielding the dampers prior to the yield in the R/C frame, in other words, by setting the value of ν smaller than unity. The value of the drift ratio ν is defined from 0 to 1.0, and it is intended to be constant for all stories and strength ratios, according to "Constant yield story-drift ratio" scheme [38],[50]. The value ν=1.0 means that both the damper system and the main frame yield at the same story-drift level, and it can be assumed as the lowest protection to the main frame since the damper system will have the smallest stiffness and require larger story drifts to start to dissipate energy. To the scope to observe the influence of the DC on the seismic performance of an entire system, the nonlinear static and time history analyses, in SAP2000 [39], have been performed on the benchmark plane 10-story model [38],[50]. The benchmark plane model (Frame C) is represented in Figure 9 and the geometric and mechanical properties are described in [38],[50]. By means of the building regularity, the vertical distribution of lateral loads used for the eval-

uation of the pushover curve useful to design the DCs has been assumed proportional to the first mode shape [3].

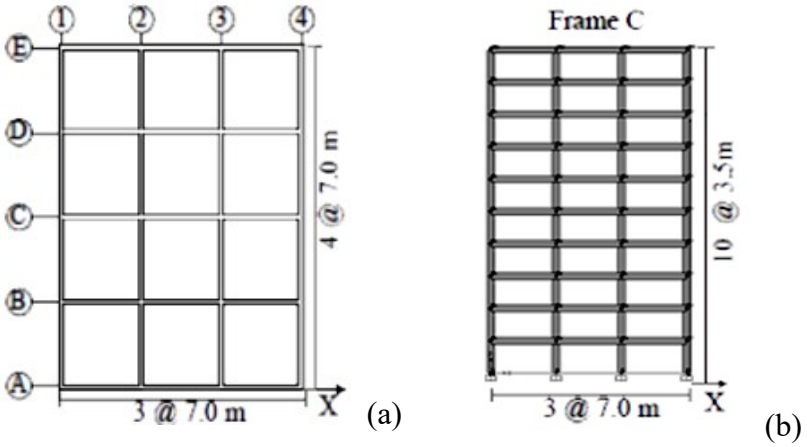



Figure 9: Typical plane (a) and elevation (b) benchmark model (modified from [38],[50]).

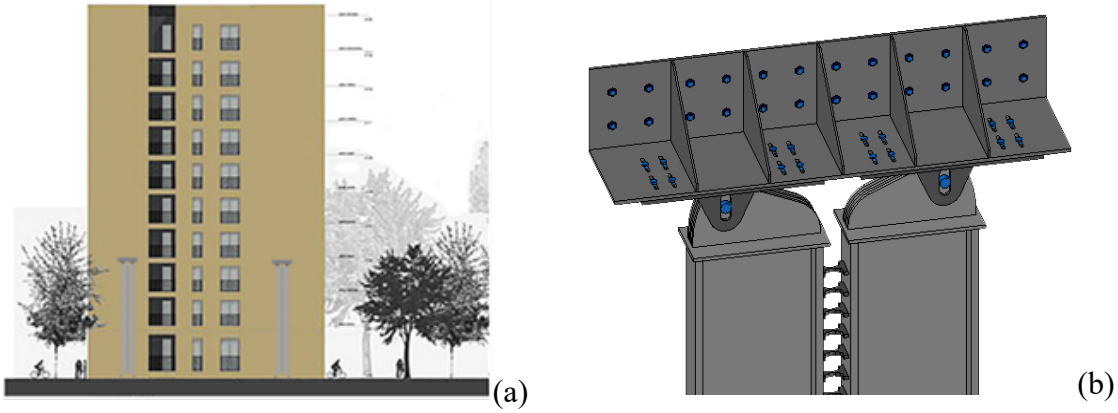
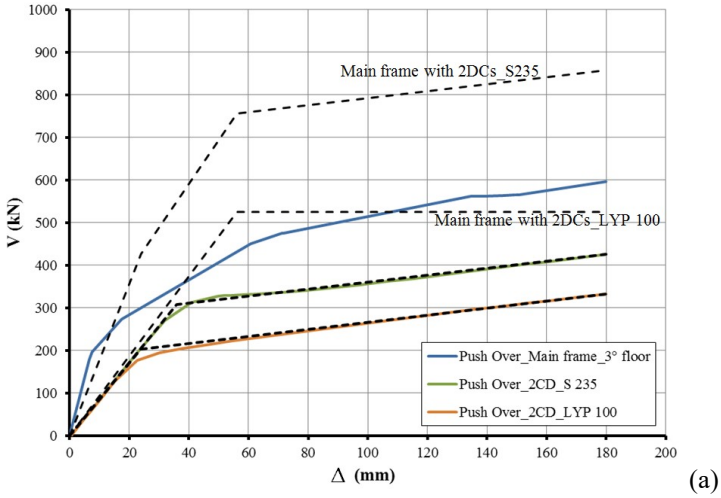



Figure 10: (a) Schematic installation of two DCs linked to the building; (b) Structural detail of the connection.

In the present study, two DCs have been applied between the third level and the ground. Figure 10 shows a schematic example of the connection between two DCs and the building and the structural detail of the link between the primary structure and damping system realized with two DCs so that both ones present the same deformation. Nonlinear analyses have been performed on the benchmark model with and without the DCs. The DCs have been modeled through nonlinear link element ruled by Bouc&Wen model, as described above, considering both LYP100 and S235 X-shape plates. The comparison between the bilinear pushover curve representative of the seismic response of the main frame in correspondence of the third level and the bilinear pushover curve descriptive of the seismic performance of the DCs with LYP100 and S235 X shape plates have allowed to determine the β and ν values for the considered system. As shown in Figure 11, the damper strength ratio β and the yield story drift ratio ν values are in accordance with the limit established by [38],[50] to obtain a good efficiency of the damper system. Considering the benchmark building located in L'Aquila site, within the "Direct Displacement Based Design" philosophy [3], the both DCs have been designed to increase the equivalent damping until to about 10% in addition to the inherent damping (5%) of the primary structure in order to achieve a target displacement between the roof and the ground equal about to 0.18 m. In particular, seven spectrum compatible input ground motions have been selected from the European database compatible in average with

the elastic spectrum (with viscous damping $\xi$ = 5 %) relative to the ultimate limit state ULS, life safe, provided by the new Italian seismic code NTC08 [51] and considering an ordinary structure with a nominal life of 50 years located in L'Aquila (Italy), soil type C and topography class T2, to perform nonlinear time history analyses.




| 2 CD_LYP 100 | 2 CD_S 235 |
|---|---|
| $\Delta'_{Dy} = 24\ mm$ | $\Delta''_{Dy} = 36\ mm$ |

(b)

2 CD_LYP 100: $\nu = 0.28$, $\beta = 0.42$

2 CD_S 235: $\nu = 0.38$, $\beta = 0.63$

(c)

Figure 11: (a) Comparison between main frame pushover curve and DC pushover curves; Yielding displacements (b) and parameters [38],[50] (c) of the two DCs with LYP110 and S235 X-shape plates.

| Waveform ID | Earthquake ID | Earthquake Name | $M_w$ | Fault Mechnism | Epicentral Ditance [km] | PGA [m/s$^2$] |
|---|---|---|---|---|---|---|
| 74 | 43 | Gazli | 6.7 | thrust | 11 | 6.0382 |
| 197 | 93 | Montenegro | 6.9 | thrust | 24 | 2.8797 |
| 290 | 146 | Campano Lucano | 6.9 | normal | 32 | 3.1662 |
| 535 | 250 | Erzincan | 6.6 | strike slip | 13 | 5.0275 |
| 1560 | 497 | Duzce 1 | 7.2 | oblique | 39 | 7.3108 |
| 4673 | 1635 | South Iceland | 6.5 | strike slip | 15 | 4.6775 |
| 7329 | 2343 | Faial | 6.1 | strike slip | 11 | 4.1204 |

Table 3. Input ground motions.

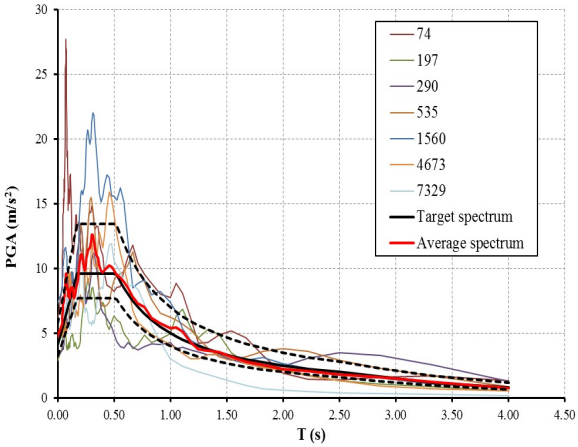



Figure 12: Response spectra of the selected ground motions and target spectrum ($\boldsymbol{\xi}$ **= 5 %**).

The values of the magnitude ranging from 6 to 8, and of the epicentral distance between 10 and 40 km are reported in Table 3. In Figure 12, the target spectrum as well as corresponding

elastic pseudo acceleration response spectra are shown. The time history analyses with and without the two DCs with LYP100 and S235 X shape plates linked to the main frame have been performed to evaluate the seismic performance in terms of relative displacements at each story. The effect of the damper system provided by the two DCs is significant until to 5-story and it is evident that the installation of two DCs to the 3rd story successfully affects the global seismic performance avoiding any soft stories and plastic hinges in the primary elements of the main frame.

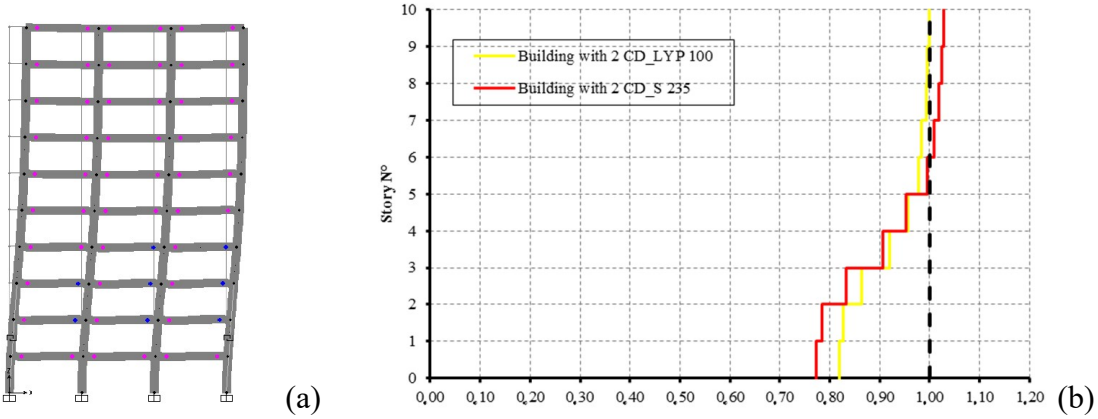



Figure 13: (a) Deformed Frame at the ultimate step (b) Normalized average maximum interstory drif ratio.

## 7 CONCLUDING REMARKS

The aim of this paper is to propose and investigate a new replaceable hysteretic damper consisting of two or more dissipative steel columns directly connected to two floors linked to each other with X-shaped low/mild steel plates. This element is able to add significant stiffness and damping to the structural system as well as to reduce seismic response and damage in primary structural members under severe earthquakes. The Dissipative Columns (DC) element equipped with both LYP100 and S235 X shape plates has been investigated through non-linear pushover and cyclic analyses. The efficiency of the proposed damper applied on a benchmark system has been investigated through time history analyses demonstrating successfully the benefits on the global seismic performance avoiding any soft stories and plastic hinges in the primary elements of the main frame.

## REFERENCES

[1] M. De Iuliis, P. Castaldo, B. Palazzo, Analisi della domanda sismica inelastica del terremoto de L'Aquila su sistemi dimensionati secondo le NTC2008. *Ingegneria Sismica*, Patron Editore **XXVII**(3), 55–68, 2010.

[2] California Office of Emergency Services (OES). Vision 2000: Performance Based Seismic Engineering of Buildings; Structural Engineers Association of California: Sacramento, CA, USA, 1995.

[3] M.J.N. Priestley, Myths and fallacies in earthquake engineering-conflicts between design and reality. *Bull. NZNSEE*, **26**, 329–341,1993**.**

[4] P. Castaldo, B. Palazzo, P. Della Vecchia, Seismic reliability of base-isolated structures with friction pendulum bearings. *Engineering Structures,* **95**, 80-93, 2015.

[5] P. Castaldo, E. Tubaldi, Influence of FPS bearing properties on the seismic performance of base-isolated structures, *Earthquake Engineering and Structural Dynamics*, **44**(15), 2817–2836, 2015.

[6] B. Palazzo, P. Castaldo, P. Della Vecchia, Seismic reliability analysis of base-isolated structures with friction pendulum system, *2014 IEEE Workshop on Environmental, Energy and Struct. Monitoring Systems Proceedings*, Napoli, September 17-18, 2014.

[7] M.D. Symans, F.A. Charney, A.S. Whittaker, M.C. Constantinou, C.A. Kicher, M.W. Johnson, R.J. McNamara, Energy dissipation system for seismic applications: Current practise and recent developments. *J. Struct. Eng.*, **134**, 3–21, 2008.

[8] A. Wada, A. Watanabe, J. Connor, H. Kawai, M. Iwata, Damage Tolerant Structures. In *Proceeding of the Fifth US-Japan Workshop on the Improvement of Building Structural Design and Construction Procedures*, ATC-15, San Diego, CA, USA, 8–10 September 1992.

[9] R. Montuori, The influence of gravity loads on the seismic design of RBS connections, *Open Construction and Building Technology Journal*, **8**, 248-261, 2014.

[10] S. Pampanin, Emerging solutions for high seismic performance of precast-prestressed concrete buildings. *J. Adv. Concr. Technol.*, 3, 202–222, 2005.

[11] S. Pampanin, Reality-check and renewed challenges in earthquake engineering: Implementing low-damage structural systems-from theory to practice. In Proceeding of the *15WCEE*, Lisbon, Portugal, 24–28 September 2013.

[12] B. Palazzo, P. Castaldo, I. Marino, The steel column damper: a new hysteretic device providing additional stiffness and damping, *EUROSTEEL 2014*, September 10-12, Naples, Italy, 2014.

[13] B. Palazzo, P. Castaldo, I. Marino, The Dissipative Column: A New Hysteretic Damper," *Buildings* **5**(1), 163-178, 2015, doi:10.3390/buildings5010163.

[14] V. Piluso, R. Montuori, M. Troisi, Innovative structural details in MR-frames for free from damage structures, *Mechanics Research Communications*, **58**:146-156, 2014.

[15] P. Castaldo, M. De Iuliis, Optimal integrated seismic design of structural and viscoelastic bracing-damper systems, *Earthquake Engineering and Structural Dynamics*, **43**(12), 1809–1827, 2014, DOI: 10.1002/eqe.2425.

[16] P. Castaldo, *Integrated Seismic Design of Structure and Control Systems*. Springer International Publishing: New York, 2014 . DOI 10.1007/978-3-319-02615-2.

[17] A. Longo, R. Montuori, E. Nastri, V. Piluso, On the use of HSS in seismic-resistant structures, *Journal of Constructional Steel Research*, **103**:1-12, 2014.

[18] R. Montuori, R. Muscati, Plastic design of seismic resistant reinforced concrete frame," *Earthquake and Structures,* **8**(1), 205-224, 2015.

[19] Giugliano M.T., Longo A., Montuori R., Piluso V., Plastic design of CB-frames with reduced section solution for bracing members. *Journal of Constructional Steel Research.* Vol. 66 pp 611-621. 2010.

[20] Montuori R., Piluso V., Troisi M. Innovative structural details in MR-frames for free from damage structures *Mechanics Research Communications* – Vol. 58 pp. 146-156. 2014.

[21] Longo A., Montuori R., Piluso V., Theory of Plastic Mechanism Control of Dissipative Truss Moment Frames, *Engineering Structures*, Vol. 37, pp.

[22] Montuori, R. , Muscati, R. A General Procedure for Failure Mechanism Control of Reinforced Concrete Frames. Accepted for publication to Engineering Structures.

[23] Montuori, R., Nastri, E., Piluso, V., Theory of Plastic Mechanism Control for Eccentrically Braced Frames with inverted Y-scheme, *Journal of Constructional Steel Research*, Volume 92, pp. 122-135, 2014.

[24] Montuori, R., Nastri, E., Piluso, V., "Seismic Design of MRF-EBF Dual Systems with Vertical Links: Ec8 Vs Plastic Design", *Journal of Earthquake Engineering*, (2015)

[25] Montuori, R., Nastri, E., Piluso, V., "Seismic Response of EB-Frames with Inverted Yscheme: TPMC Versus Eurocode Provisions", *Earthquakes and Structures*, (2015)

[26] Longo A., Montuori R., Piluso V. Seismic Reliability of Chevron Braced Frames with Innovative Conception of Bracing Members, *Advanced Steel Construction, an International Journal* - Vol. 5, No. 4, December 2009.

[27] Longo, A. , Montuori, R., Piluso, V Theory of plastic mechanism control for MRF–CBF dual systems and its validation

[28] Bulletin of Earthquake Engineering. Volume 12, Issue 6, 6 November 2014, Pages 2745-2775

[29] R. Montuori, E. Nastri, V. Piluso, Theory of plastic mechanism control for the seismic design of braced frames equipped with friction dampers, *Mechanics Research Communications* **58**:112-123, 2014.

[30] R. Montuori, E. Nastri, V. Piluso, Advances in theory of plastic mechanism control: Closed form solution for MR-Frames, *Earthquake Engineering and Structural Dynamics,* **44**(7):1035-1054, 2015.

[31] M. De Iuliis, P. Castaldo, An energy-based approach to the seismic control of one-way asymmetrical structural systems using semi-active devices, *Ingegneria Sismica - International Journal of Earthquake Engineering* **XXIX**(4):31-42, 2012.

[32] J. Kim, Y. Seo, Seismic design of low-rise steel frames with buckling-restrained Braces. *Eng. Struct.*, **26**, 543–551, 2004.

[33] K.A.S. Susantha, T. Aoki, T. Kumano, K. Yamamoto, Application of Low-Yield Strength Steel in Steel Bridge Piers, In Proceeding of the *13th WCEE*, Vancouver, BC, Canada, 1–6 August 2004.

[34] Montuori R. The Influence of Gravity Loads on the Seismic Design of RBS Connections. *The Open Construction and Building Technology Journal*, 2014, 8, (Suppl 1: M6) 248-261.

[35] Montuori, R. Design of "Dog-bone" connection: The role of vertical loads. COMPDYN 2015 - 5th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering 2015, Pages 3368-3387. Crete; Greece; 25 May 2015 through 27 May 2015.

[36] ABAQUS, User's Manual: Volume III: Materials, 2010.

[37] ABAQUS, User's Manual: Volume IV: Elements, 2010.

[38] A.J.A. Oviedo, M. Midorikawa, T. Asari, Earthquake response of ten-story story-drift-controlled reinforced concrete frames with hysteretic dampers. *Eng. Struct.*, **32**, 1735–1746, 2010.

[39] SAP2000 vers; Computers Structures Inc.: Berkeley, CA, USA, 2010.

[40] A. Ghobarah, On Drift Limits Associated with Different Damage Levels. In Proceedings of the International *Workshop on Performance-Based Seismic Design, Department of Civil Engineering*, McMaster University, Bled, Slovenia, 28 June–1 July 2004.

[41] I. Aiken, D. Nims, A. Whittaker, J. Kelly, Testing of passive energy dissipation systems. *Earthq. Spectra Earthq. Eng. Res. Inst. Calif.*, **9**, 335–370, 1993.

[42] A. Whittaker, V. Bertero, J. Alonso, Thompson, C. *Earthquake Simulator Testing of Steel Plate Added Damping and Stiffness Elements*. Report No. UCB/EERC-89/02; Earthquake Eng. Research Center, University of California, Berkeley, CA, USA, 1989.

[43] A.S. Whittaker, V.V. Bertero, Thompson, C.L., Alonso, L.J. Seismic testing of steel plate energy dissipation devices. *Earthq. Spectra*, **7**, 563–604, 1991.

[44] CEN-European Committee for Standardization *Eurocode 3 Part 1: General Rules and Rules for Buildings*, Brussels, Belgium., 2007.

[45] E. Saeki, M. Sugisawa, T. Yamaguchi, A. Wada, Mechanical properties of low yield point steels," *J. Mater. Civ. Eng.* **10**, 143-152, 1998.

[46] A. Tena-Colunga, Mathematical modelling of the ADAS energy dissipation device, *Engineering Structures* **19**(10), 811-821, 1997.

[47] V. Piluso, C. Faella, G. Rizzano Ultimate behavior of bolted T-stubs. Part I: theoretical model. *J Struct Eng* ASCE, **127**(6), 686–93, 2001.

[48] R. Bouc, Forced vibration of mechanical systems with hysteresis. Proceedings of the 4th Conference on Nonlinear Oscillations, Prague, Czechoslovakia, 315-315, 1967.

[49] YK. Wen, Method for random vibration of hysteretic systems. *Journal of the Engineering Mechanics Division*, **102**(2), 249-263, 1976.

[50] A.J.A. Oviedo, M. Midorikawa, T. Asari, Performance of Story-Drift-Controlled R/C Frames Equipped with Hysteretic Dampers Subjected to Earthquake Motions. 15 WCEE, Lisboa 2012.

[51] NTC08 - Norme tecniche per le costruzioni, Gazzetta Ufficiale del 04.02.08, DM 14.01.08, Ministero delle Infrastrutture, 2008.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EXAMPLE OF ENERGY DISSIPATION OPTIMIZATION FOR STEEL MRFs WITH PIN-JOINTED COLUMN BASES

**Elide Nastri**[1], **Annabella Paciello**[1]

[1] University of Salerno
Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
e-mail: enastri@unisa.it, annabella.paciello@gmail.com

**Keywords:** MRFs, Dissipation optimization, steel, dissipative devices, free from damage structures, Theory of Plastic Mechanism Control.

**Abstract.** *The proposed work has the aim to show a practical application of the Theory of Plastic Mechanism Control applied to Moment Resisting Frames with pin-jointed base connections whose traditional dissipative zones, i.e. beam ends, in the framework of supplementary energy dissipation strategies, are substituted by frictional devices. The control of devices activation and, as a consequence, the optimization of energy dissipation capacity is obtained by a rigorous design procedure assuring that only the dissipative zones are involved in plastic range while non dissipative ones remain elastic. It means that a mechanism of global type has to be assured by applying the Theory of Plastic Mechanism Control. In addition, being traditional dissipative zones substituted by frictional devices, column bases pin jointed and column sections designed in order to remain in elastic range by a properly design procedure, the structure can be considered as free from damage. In fact after a destructive seismic event all the damaged devices can be replaced by new ones. The accuracy of the design approach has been investigated by means push-over and non-linear dynamic analyses by applying a properly chosen set of earthquake ground motions.*

## 1 INTRODUCTION

The optimization of seismic energy dissipation has ever been one of the main purposes that has to be fulfilled in order to design structures able to withstand severe earthquake event. For this reason modern seismic codes have introduced simplified rules, such as the beam-column hierarchy criterion having the scope to promote the developing of plastic hinges at the beam ends constituting the dissipative zone, while non dissipative ones, i.e. columns, remain in elastic range [1]. However, this criterion that is reported also in the European seismic code (Eurocode 8 (EC8)) [2] is only able to assure that structures do not collapse with a soft-storey mechanism but it is not able to assure a determinate collapse mode and also a dissipation optimization of the seismic response of the structure. In fact, the seismic response of the structure, and in this case of steel Moment Resisting Frames (MRFs), happens when all the beam ends are engaged in yielding as well as the column bases at first storey. This mode of collapse is called global type.

The development of a collapse mechanism of global type is assured only by a revolutionary design procedure based on the Theory of Plastic Mechanism Control (TPMC) [3] that, differently by the rules proposed by EC8 allows designing structure always showing at the collapse a mechanism of global type. This result is assured thanks both to the strong background of the theory, that is based on the kinematic theorem of plastic collapse, than to the successive improvements of the design procedure that have led to a recent version that is easy to be carried out by means of hand calculation. This last form of the design procedure is the so called "closed form solution of TPMC" that has been satisfactorily applied to design both MRFs [4] than EBF-MRFs dual systems [5]. However, this procedure, in the previous but as the same efficient form as the improved one, has been applied to design all the structural steel typology [6] - [21] and also reinforced concrete MRFs [22]-[23].

Even though, this procedure is able to assure a collapse mechanism of global type, the seismic optimization has to be pursued by assuring that all the dissipative zones engage in yielding as possible as the same time. For this reason, in this structural example, with the scope to have a supplementary and substitutive dissipation [24]-[28], traditional dissipative zones, such as the beam ends, have been substituted by innovative frictional devices [29] developed in the framework of the European Research Project "FREEDAM". As a consequence, the design procedure based on TPMC has improved to promote the contemporary activation of all dissipative devices by additionally considering the column bases at the first storey pinned. In this way, being only the device activated, column bases pin jointed and column sections designed in order to remain in elastic range by a properly design procedure, the structure can be considered as free from damage [30]-[31]. In fact after a destructive seismic event all the damaged devices can be replaced by new ones. The accuracy of the design approach has been investigated by means push-over and non-linear dynamic analyses by applying a properly chosen set of earthquake ground motions.

## 2 DESIGN APPROACH

The "Theory of Plastic Mechanism Control" design approach has the aim to provide structures able to fail with a collapse mechanism of global type. Global mechanism represents the optimum in term of structural dissipation capacity being all the dissipative zones involved in the pattern of yielding while the non dissipative ones remain in elastic range. Dissipative zones of traditional MR-Frame are the beam ends and the bases of the column at the first storey. However, in this particular case, beam ends are substituted by special dissipative devices and column bases at the bottom of the first storey by means of pin connections. In this way MRFs plastic hinges cannot develop at the bottom of the first storey columns, so that, only the

beam ends are involved in plastic range. In order to design members to fulfil this design goal both devices at the beam ends and column sections have to be designed in order to assure on one hand the optimization of the seismic response and from the other hand a collapse mechanism of global type. Both dissipative zone and non-dissipative ones have been designed according to TPMC. The first design unknown having to be provided are the devices at the beam-to-column connections for which, the whole sum of design slip moment has to be provided by assuring that, according with the kinematic theorem of plastic collapse extended to the concept of mechanism equilibrium curve, the collapse multiplier $\alpha$ of the seismic storey forces $F_k$ has to be at least equal to 1 at a design displacement $\delta_u$ compatible with the ductility supply of dissipative zone:

$$\alpha = \alpha_0^{(g)} - \gamma^{(g)}\delta_u \geq 1 \;\Rightarrow\; \frac{M_{fb.Ed.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \geq 1 \tag{1}$$

where $\alpha_0^{(g)}$ is the first order collapse multiplier:

$$\alpha_0^{(g)} = \frac{M_{fb.Ed.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} \tag{2}$$

and $\gamma^{(g)}$ is the slope of mechanism equilibrium curve accounting for second order effects:

$$\gamma^{(g)} = \frac{\frac{1}{h_{ns}}\sum_{k=1}^{n_s}V_k h_k}{\sum_{k=1}^{n_s}F_k h_k} \tag{3}$$

In particular, $V_k$ and $h_k$ are the sum of all the vertical loads acting at each storey and the storey height, respectively. By designing dissipative zones proportionally to the storey shear they are assured to be engaged in yielding as contemporary as possible, benefit optimization of the seismic response. In addition, Eq. (1) is quite conservative because it assures that also at the collapse the structure is able to support the storey forces distribution. In addition, the design ultimate displacement has to be dependent on the device stroke, i.e. the ultimate distance the dissipative device is able to tread by slippage. By solving as an equality Eq. (1) it is possible to provide the sum of design slip moments at the last storey , $M_{fb.Ed.n_s}$:

$$M_{fb.Ed.n_s}\sum_{k=1}^{n_s}\beta_k = \left(\gamma^{(g)}\delta_u + 1\right)\sum_{k=1}^{n_s}F_k h_k \tag{4}$$

that shared by the term $\beta_k = \sum_{i_m=k}^{n_s}F_k\,/F_{n_s}$, [32]-[33] provides the sum of design slip moment at each storey optimized to promote the contemporary engagement in slippage, $M_{fb.Ed.k}$:

$$M_{fb.Ed.k} = \beta_k M_{fb.Ed.n_s} \tag{5}$$

Finally, $M_{fb.Ed.k}$ has to be shared between all the bay in function of the loads $q_j$ and the bay length $L_j$:

$$M_{fb.Ed.jk} = M_{fb.Ed.k}\frac{q_j L_j^2}{2\sum_{j=1}^{n_b}q_j L_j^2} \tag{6}$$

The design slip moment $M_{fb.Ed.jk}$ is used to design frictional devices whose resistance is expressed by means of the following term $M_{fb.Cd.jk}$.

Beams, in this case, have to be considered as non dissipative zones and to be designed by means as the maximum among the design moment belonging to the vertical load distribution and, according to the second capacity design principle, those provided by a local hierarchy criterion where $\gamma_{Rd}$ is an overstrength factor.

$$M_{b.Cd.jk} = \max\left(\frac{q_{v.j}L_j^2}{8}; \gamma_{Rd}M_{fb.Cd.jk}\right) \tag{7}$$

As regards the column design, all the possible mechanism typology which can affect a seismic resistant frame should be considered. In particular, it is possible to observe that MRFs under seismic horizontal forces should fail according to three main collapse typologies called type-1, type-2 and type-3, as reported in
Figure 1 for the specific case of pin-jointed column bases where the rectangles represents the dissipative devices involved while the solid circles the plastic hinges developing at the top or bottom end of columns for undesired mechanisms. Every one of these can involve each storey and, for this reason, being the number of storey equal to $n_s$ the number of possible mechanisms involving a structure can be easily computed as $3n_s$. Only one of these mechanisms is the desired one, i.e. the global one, and it is a particular case of type-2 mechanism extended to all the $n_s$ storeys. It means that all the other $3n_s - 1$ mechanisms are undesired and must be avoid by applying TPMC.

The design conditions also found their theoretical bases in the kinematic theorem of plastic collapse extended to the concept of collapse mechanism equilibrium curve which is representative of a straight line whose intercept with the vertical axis in a Cartesian diagram is the collapse mechanism multiplier of the first order, $\alpha_{i_m}^{(t)}$, while $\gamma_{i_m}^{(t)}$ is the slope where $i_m$ and $t$ are the mechanism index and the mechanism typology code respectively.

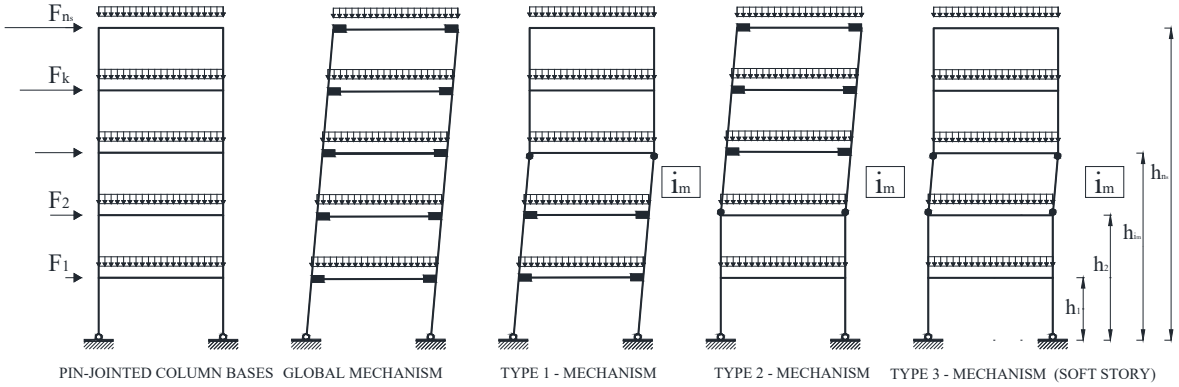



Figure 1: Collapse mechanism of pin-jointed column bases resistant frame

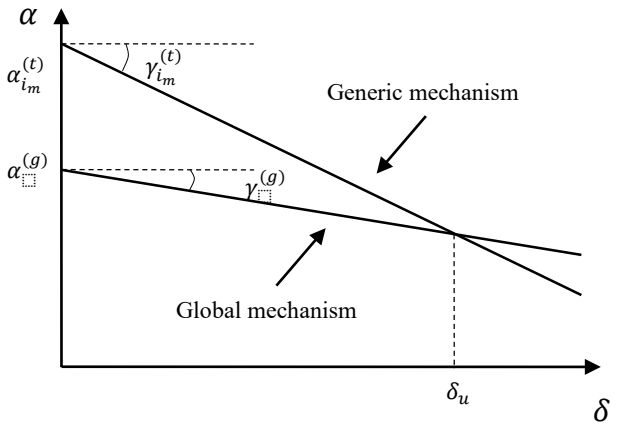



Figure 2: Design condition for the failure mode control

TPMC states that the global mechanism equilibrium curve has to be located below those corresponding to all the other mechanisms within a displacement range compatible with the plastic deformation capacity of members, $\delta_u$ (Figure 2). Consequently, the design conditions are provided by the following equations:

$$\alpha^{(g)} - \gamma^{(g)}\delta_u \leq \alpha_{i_m}^{(t)} - \gamma_{i_m}^{(t)}\delta_u \qquad i_m = 1,2,3,\dots,n_s \qquad t = 1,2,3 \tag{8}$$

The computation of the quantities $\alpha_{i_m}^{(t)}$ and $\gamma_{i_m}^{(t)}$ exploits the virtual work principle and it is amply detailed and discussed in a previous work [3]-[4]. In particular, the expressions of $\alpha_{i_m}^{(t)}$ and $\gamma_{i_m}^{(t)}$ are reported in Table 1. In particular, according to the second capacity design principle, non dissipative zones have to be designed according to the maximum actions that dissipative zones, i.e. dissipative devices, are able to transmit in their ultimate conditions. For this reason, frictional device overstrength have to be considered by introducing the overstrength factor $\gamma_{Rd}$.

| ***Type-1*** | ***Type-2*** | ***Type-3*** |
|---|---|---|
| $\alpha_{0.i_m}^{(1)} = \gamma_{Rd}\dfrac{M_{fb.Cd.n_s}\sum_{k=1}^{i_m-1}\beta_k + \sum_{i=1}^{n_c}M_{c.ii_m}^{(1)}}{\sum_{k=1}^{i_m}F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}F_k}$ | $\alpha_{0.i_m}^{(2)} = \dfrac{\sum_{i=1}^{n_c}M_{c.ii_m}^{(2)} + \gamma_{Rd}M_{fb.Cd.n_s}\sum_{k=i_m}^{n_s}\beta_k}{\sum_{k=i_m}^{n_s}F_k(h_k - h_{i_m-1})}$ | $\alpha_{0.i_m}^{(3)} = \dfrac{2\sum_{i=1}^{n_c}M_{c.ii_m}^{(3)}}{(h_{i_m} - h_{i_m-1})\sum_{k=i_m}^{n_s}F_k}$ |
| $\gamma_{i_m}^{(1)} = \dfrac{\sum_{k=1}^{i_m}V_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}V_k}{h_{i_m}\left(\sum_{k=1}^{i_m}F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}F_k\right)}$ | $\gamma_{i_m}^{(2)} = \dfrac{\sum_{k=i_m}^{n_s}V_k(h_k - h_{i_m-1})}{(h_{n_s} - h_{i_m-1})\sum_{k=i_m}^{n_s}F_k(h_k - h_{i_m-1})}$ | $\gamma_{i_m}^{(3)} = \dfrac{\sum_{k=i_m}^{n_s}V_k}{(h_{i_m} - h_{i_m-1})\sum_{k=i_m}^{n_s}F_k}$ |

Table 1. First order and second order collapse mechanism multiplier at the generic $i_m - th$ storey

Therfore, by substituting terms reported in Table 1 in Eq. **(8)**, the following design relationship are obtained for type-1 mechanism:

$$\gamma_{Rd}\frac{M_{fb.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{M_{fb.Cd.n_s}\sum_{k=1}^{i_m-1}\beta_k + \sum_{i=1}^{n_c}M_{c.ii_m}^{(1)}}{\sum_{k=1}^{i_m}F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}F_k} - \gamma_{i_m}^{(1)}\delta_u \tag{9}$$

for type-2 mechanism:

$$\gamma_{Rd}\frac{M_{fb.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{\sum_{i=1}^{n_c}M_{c.ii_m}^{(2)} + \gamma_{Rd}M_{fb.Cd.n_s}\sum_{k=i_m}^{n_s}\beta_k}{\sum_{k=i_m}^{n_s}F_k(h_k - h_{i_m-1})} - \gamma_{i_m}^{(2)}\delta_u \tag{10}$$

and for type-3 mechanism :

$$\begin{aligned} for\ i_m = 1 \qquad & \gamma_{Rd}\frac{M_{fb.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{\sum_{i=1}^{n_c}M_{c.i1}^{(3)}}{h_1\sum_{k=1}^{n_s}F_k} - \gamma_1^{(3)}\delta_u \\ for\ i_m > 1 \qquad & \gamma_{Rd}\frac{M_{fb.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{2\sum_{i=1}^{n_c}M_{c.ii_m}^{(3)}}{(h_{i_m} - h_{i_m-1})\sum_{k=i_m}^{n_s}F_k} - \gamma_{i_m}^{(3)}\delta_u \end{aligned} \tag{11}$$

That rearranged to provide the unknown of the design procedure, i.e. the sum of column plastic moment at each storey $M_{c.ii_m}^{(t)}$ reported in the following.

In particular, at the first storey the condition reported in Eq. (8) provide only one design relationship being easily demonstrable that mechanism type-1 and type-3 at the first storey are coincident and mechanism type-2 is coincident with the global one:

$$\sum_{i=1}^{n_c} M_{c.i1}^{(1or3)} \geq \left[\gamma_{Rd} \frac{M_{fb.Ed.n_s} \sum_{k=1}^{n_s} \beta_k}{\sum_{k=1}^{n_s} F_k h_k} + \left(\gamma_1^{(3)} - \gamma^{(g)}\right) \delta_u\right] \left(h_{i_m} - h_{i_m-1}\right) \sum_{k=i_m}^{n_s} F_k \tag{12}$$

At the $i_m > 1$ storeys, design conditions have to be make explicit for all the undesired mechanism typologies:

$$\sum_{i=1}^{n_c} M_{c.ii_m}^{(1)} \geq \left[\gamma_{Rd} M_{fb.Cd.n_s} \left(\frac{\sum_{k=1}^{n_s} \beta_k}{\sum_{k=1}^{n_s} F_k h_k} - \frac{\sum_{k=1}^{i_m-1} \beta_k}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m} \sum_{k=i_m+1}^{n_s} F_k}\right) + \left(\gamma_{i_m}^{(1)} - \gamma^{(g)}\right) \delta_u\right] \left(\sum_{k=1}^{i_m} F_k h_k + h_{i_m} \sum_{k=i_m+1}^{n_s} F_k\right) \tag{13}$$

for type-1 mechanism;

$$\sum_{i=1}^{n_c} M_{c.ii_m}^{(2.1)} \geq \left[\gamma_{Rd} M_{fb.Cd.n_s} \left(\frac{\sum_{k=1}^{n_s} \beta_k}{\sum_{k=1}^{n_s} F_k h_k} - \frac{\sum_{k=i_m}^{n_s} \beta_k}{\sum_{k=i_m}^{n_s} F_k (h_k - h_{i_m-1})}\right) + \left(\gamma_{i_m}^{(2)} - \gamma^{(g)}\right) \delta_u\right] \sum_{k=1}^{i_m} F_k (h_k - h_{i_m-1}) \tag{14}$$

for type-2 mechanism;

$$\sum_{i=1}^{n_c} M_{c.ii_m}^{(3)} \geq \left[\gamma_{Rd} M_{fb.Cd.n_s} \frac{\sum_{k=1}^{n_s} \beta_k}{\sum_{k=1}^{n_s} F_k h_k} + \left(\gamma_1^{(3)} - \gamma^{(g)}\right) \delta_u\right] \frac{\left(h_{i_m} - h_{i_m-1}\right)}{2} \sum_{k=i_m}^{n_s} F_k \tag{15}$$

for type-3 mechanism.

While for the $i_m = 1$ storey the design condition is only one Eq. (13) for the $i_m$>1 storeys, the sum of the reduced plastic moments of columns has to be calculated as the maximum among the three mechanism typologies, according to the following relation:

$$\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}} = \boldsymbol{max} \left\{ \sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(1)}}, \sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(2)}}, \sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(3)}} \right\} \tag{16}$$

The knowledge of these plastic moments, coupled with the estimated axial force at the collapse state, allows the evaluation of the required sections of columns.

The sum provided by Eq. (16) has to be distributed between all the columns at each storey according to two possible strategies [4]:

1) the sum are subdivided equally between each column
2) the sum are subdivided according to the axial load acting on the columns in the collapse condition.

However, in both the cases column sections need to verify the combination of the competent axial load at collapse and bending moment.

The axial load acting on the columns at collapse state can be easily provided in agreement with the global mechanism where axial forces in the columns at collapse state depend both on distributed loads acting on the beams and on the shear action coming from the development of plastic hinges at the beam ends, as depicted in Figure 3. So that, the total load transmitted by the beams to the columns is the sum of two contributions. The first one, $N_q$, is related to the vertical loads acting in the seismic load combination (i.e. the sum of $qL_j/2$ type contributions). The second one, $N_f$, is related to the shear actions due to the engagement in plastic range of the dissipative devices located at the beam ends (i.e. the sum of $2\gamma_{Rd}M_{fb.Cd.n_s}/L_j$ contributions) (Figure 3). Also in this case, devices overstrength has to be taken in account according to the second capacity design principle. However, seismic actions can be acting either in the

positive direction or in the negative direction, so that the maximum axial forces has to be considered.

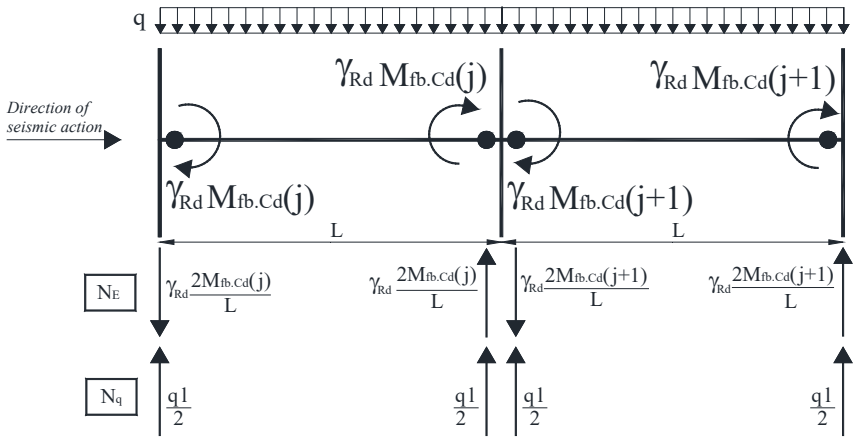



Figure 3: Loads transmitted by the beams to the columns at collapse state including the overstrength of the device

## 3 WORKED EXAMPLE

Aiming to practically show how to apply the proposed design procedure a 3 bays 4 storeys pin-jointed column bases structure whose structural scheme is depicted in Figure 4 has been designed. The bay span is 5 m while the interstorey height is equal to 3.0 m. The characteristic values of the vertical loads acting on beams are equal to 6,68 $kN/m^2$ and 5,01 $kN/m^2$ for permanent and live loads, respectively, while on columns concentrated loads belonging to the secondary warping of floor are applied (Figure 4). With reference to the seismic load combination ($G_k + \psi_2 Q_k + E_d$), the vertical loads acting on the beams of the analysed structure are 8,183 $kN/m$. The structural material adopted for all the structural members is an S275. The design horizontal forces have been determined according to EC8, assuming a peak ground acceleration equal to 0.20g, a seismic response factor equal to 2.5, a behaviour factor equal to 6. The design horizontal forces distribution is in accord to the first vibration mode as reported in Figure 4.

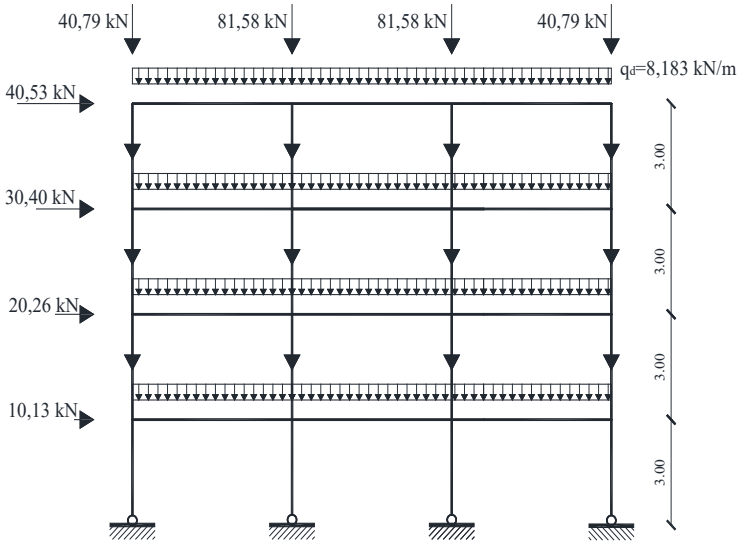



Figure 4: Worked example structural scheme

Beam-to-column devices (Figure 5) are bolted double split tee beam-to-column connections equipped with friction dampers designed to slip before the yielding of the beam [34] where the energy is dissipated through the slippage between the stem of bolted tee stubs and the beam flange with an interposed friction pad. Adopted bolts are high resistance class 8.8 while the friction material adopted (FREEDAM project) is sprayed aluminum that is able to exhibit a friction coefficient of about 0.5 [35].

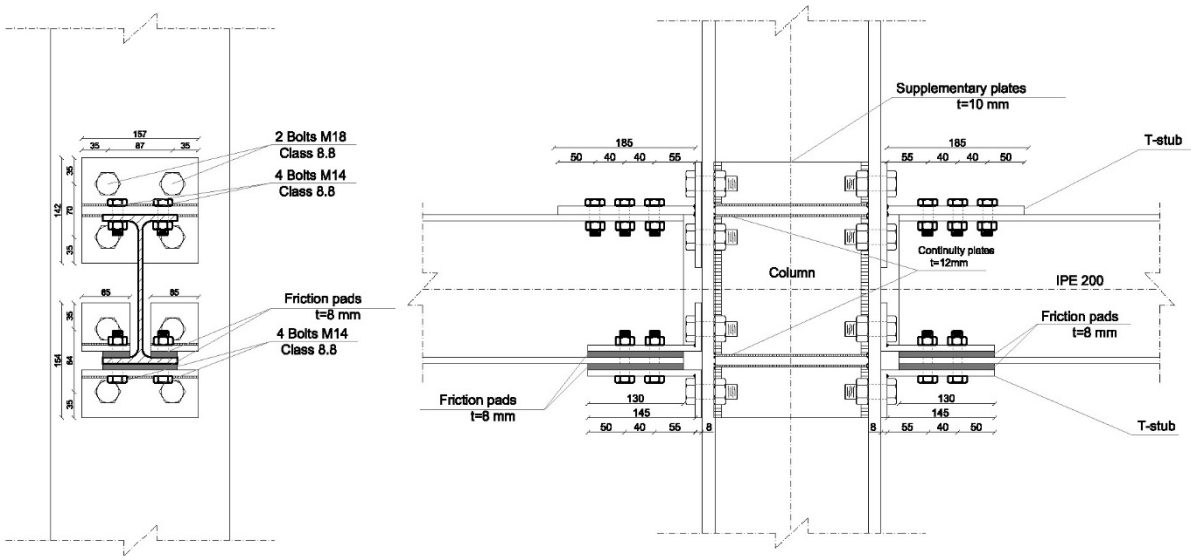



Figure 5: Beam to column connections equipped with friction dampers

### 3.1 Design of dissipative zones

The first quantity needed for the design of the structure is the design slip moment of devices at the last storey (storey $n_s$) provided by Eq. (4), being preliminarily compute the term $\sum_{k=1}^{n_s} \beta_k = 7{,}5$. Successively, the sum of the design slip moment $M_{fb.Ed.n_s}$ has to be shared between all the storeys by means of the term $\beta_k$ to design the dissipative device preliminarily described, according to the following relation:

$$M_{fb.Ed.n_s} = n_s \, n_b \, N_b \, \mu \, h_b \rightarrow M_{fb.Cd.n_s} \tag{17}$$

where, $n_s$ is the number of frictional surfaces that in this case is equal to 2, $n_b$ is number of bolts, $N_b$ is the preloading force equal to:

$$N_b = 0.7 \, f_{ub} A_s \tag{18}$$

$f_{ub}$ is the design bolt resistance, $A_s$ is the bolt area, $\mu$ is the dynamic frictional coefficient assumed in this case equal to 0.5 and $h_b$ is the beam height preliminarily assumed equal to the height of the beam designed by means of vertical load combination. In this case, beams are preliminarily selected from the standard shapes to withstand the vertical load combination $(1.3G_k + 1.5Q_k)$ according to a design moment computed as in Eq. (7) and are IPE200. In addition, in Table 2 the sum of the design slip moment $M_{fb.Ed.k}$ shared at all the storey by means of the term $\beta_k$, the slip resistant moment $M_{fb.Cd.k}$, belonging to the device design, the number of bolts $n_b$, the beam height, the preloading force $N_b$ are reported.

Finally, beam sections designed to withstand vertical loads have to fulfil also the local hierarchy criterion reported in the second term of Eq. (7).

| Storey (-) | $\beta_k$ (-) | $M_{fb.Ed.k}$ $(kN\ m)$ | $n_b$ (-) | $h_b$ (-) | Bolts (-) | $N_b$ $kN$ | $M_{fb.Cd.k}$ $(kN\ m)$ | $\gamma_{Rd}M_{fb.Cd.k}$ $(kN\ m)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 2.5 | 451.00 | 4 | 200 | ϕ18 | 120.00 | 460.80 | 552.96 |
| 2 | 2.25 | 405.90 | 4 | 200 | ϕ18 | 110.00 | 422.40 | 506.88 |
| 3 | 1.75 | 315.70 | 4 | 200 | ϕ18 | 850.00 | 326.40 | 391.68 |
| 4 | 1 | 180.40 | 4 | 200 | ϕ18 | 500.00 | 192.00 | 230.40 |

Table 2: Parameters for the design of beam-to-column connections

### 3.2 Design of non-dissipative zones

Being designed dissipative zones, also non-dissipative one have to be provided by means of TPMC. The ultimate design displacement governing the procedure has been computed in the following way:

$$\delta_u = 0.04\, h_{n_s} = 0.04 \cdot 12 = 0.48\, m \tag{19}$$

being assumed an ultimate device rotation equal to 0.04. In addition, in Table 3 the slopes of mechanism equilibrium curves are reported being computed according to the relationships reported in Table 1.

| STOREY $\boldsymbol{i_m}$ | $\boldsymbol{\gamma_{i_m}^{(1)}}$ (1/cm) | $\boldsymbol{\gamma_{i_m}^{(2)}}$ (1/cm) | $\boldsymbol{\gamma_{i_m}^{(3)}}$ (1/cm) |
|---|---|---|---|
| 1 | 0.0484 | <u>0.0101</u> | 0.0484 |
| 2 | 0.0223 | 0.0121 | 0.0403 |
| 3 | 0.0139 | 0.0165 | 0.0345 |
| 4 | 0.0101 | 0.0302 | 0.0302 |

Table 3. Slopes of the mechanism equilibrium curves

It is possible to observe that the slope of the global mechanism equilibrium curve (undersigned in Table 3) is the minimum among all the slope of mechanism equilibrium curves corresponding to the undesired mechanism. In addition, in Table 4 the design bending moment of columns at each storey are reported.

| STOREY $\boldsymbol{i_m}$ | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(1)}}$ (kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(2)}}$(kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(3)}}$(kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}}$(kN m) |
|---|---|---|---|---|
| 1 | 1119.24 | [-] | 1119.24 | 1119.24 |
| 2 | 850.36 | 51.12 | 450.74 | 850.36 |
| 3 | 544.82 | 97.52 | 321.17 | 544.82 |
| 4 | 230.40 | 111.46 | 170.93 | 230.40 |

Table 4. Value of the paramters $\sum_{\mathrm{i=1}}^{\mathrm{n_c}} \mathrm{M}_{\mathrm{c.ii_m}}^{(\mathrm{t})}$

In this worked example the approach that have been used to subdivide the reduced sum of plastic moments between the columns exploits the axial load acting on the columns at collapse state. In fact, the maximum value of $M_{c,i_m}$is subdivided between columns on the basis of axial force acting on each column at each storey.

The reduced sum of plastic moments of columns at each storey,$\sum_{i=1}^{n_c} M_{c.ii_m}$, has been subdivided proportionally to the axial load at the collapse state according to the following equation:

$$M_{c.ik} = \frac{N_{tot.ii_m} \sum_{i=1}^{n_c} M_{c.ii_m}}{\sum_{i=1}^{n_c} N_{tot.ii_m}} \tag{20}$$

where $M_{c.ik}$ is the design moment of column sections.

Finally, in Table 5 the axial load acting at the collapse state, the design moment of column sections and the column sections selected from standard shapes are reported with reference to external and internal columns. Finally the collapse mechanism multiplier of global type $\alpha = 1{,}84$).

| STOREY $i_m$ | External columns | | | Internal columns | | |
|---|---|---|---|---|---|---|
| | $N_{tot.ii_m}$ $(kN)$ | $M_{c.ik}$ (kN m) | PROFILES | $N_{tot.ii_m}$ $(kN)$ | $M_{c.ik}$ (kN m) | PROFILES |
| 1 | 357.12 | 292.20 | HE 240 B | 490.00 | 278.64 | HE 260 B |
| 2 | 259.01 | 289.30 | HE 200 B | 367.50 | 275.88 | HE 240 B |
| 3 | 163.97 | 249.63 | HE 180 B | 245.00 | 238.05 | HE 200 B |
| 4 | 76.61 | 154.81 | HE 140 B | 122.50 | 147.63 | HE 160 B |

Table 5. Axial force at collapse state, design moment of column sections and column sections selected from standard shapes at each storey

## 4 VALIDATION OF PROCEDURE

In order to validate the design procedure static non linear analysis (push-over) has been carried out for the designed frame by means of SAP 2000 computer program [36]. This analysis has the primary aim to predict the collapse mechanism typology, testing the accuracy of the proposed design methodology.

In SAP2000 ambiance all members have been modeled by means of beam-column elements, whose non-linearities have been concentrated in plastic hinges at their ends. In particular, plastic hinges accounting for the interaction between axial force and bending moment have been defined for columns while dissipative devices at beam ends have been modelled in pure bending. All hinges have been represented with a rigid plastic curve. The elastic behaviour is not considered in plastic hinge definition because it is directly taken into account by the beam-column element. The analysis has been led under displacement control taking into account both geometrical and mechanical non-linearities. In addition out of plan stability check of compressed members have been performed at each step of the non-linear analysis. The results of the push-over are mainly constituted by the frame capacity curve. In Figure 6 the push-over curve is provided with reference to two different structural models, the first one corresponding to the use of a Young modulus (E) equal to 210000 MPa the second one (dotted line) corresponding to a structural model with a ten times increased Young modulus value. This second one is proposed to more accurately point out that the softening branch of push-over curve is perfectly parallel to the global mechanism equilibrium curve. This condition shows this evidence once the structural model is more close to the hypothesis of rigid perfect plastic behavior that is herein obtained by increasing the Young Moduls.

In addition, the designed structure has been checked also for serviceability requirements under seismic action evaluated with a first period of vibration equal to 2,53 s. Beam to column connections have also been checked to be inactive under vertical load combination at the ultimate limit state.

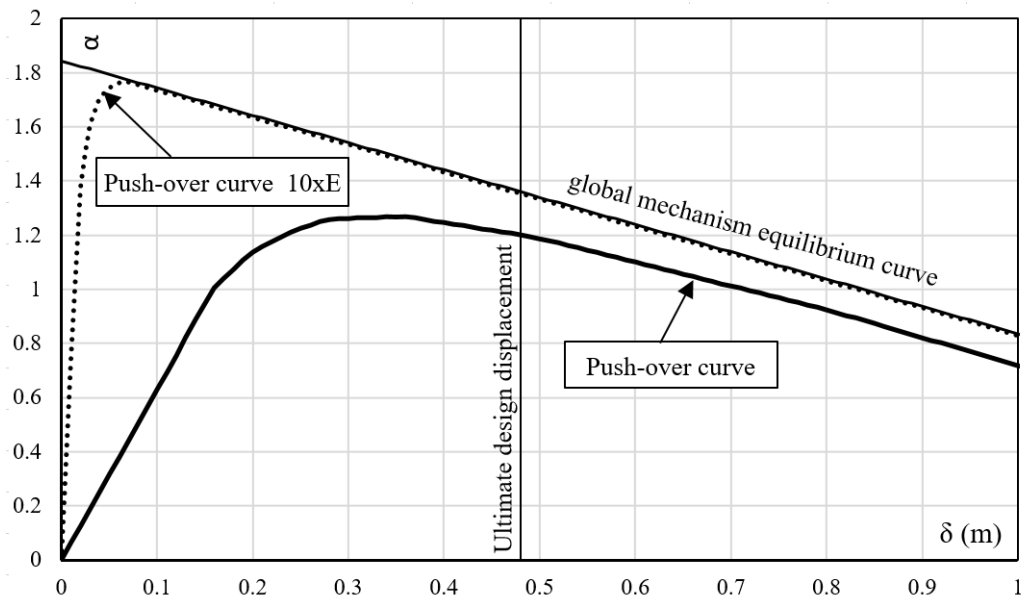



Figure 6: Push-over curves for E=210000 and 10xE

In addition, in order to provide a more robust validation of the design methodology, non-linear Incremental Dynamic Analyses (IDA) have been carried out with reference to the same structural model used for push-over analyses. In addition, 5% damping according to Rayleigh modelling has been assumed.

Record-to-record variability has been accounted for considering 10 recorded accelerograms selected from PEER [37] data base whose main characteristics (name, date, magnitude, ratio between PGA and gravity acceleration, length and step recording) are reported in Table 6. These earthquake records have been selected to approximately match the linear elastic design response spectrum of Eurocode 8 [2], for type A soil. Moreover, in order to perform IDA analyses, each ground motion has been scaled to obtain the same value of the spectral acceleration Sa ($T_1$) corresponding to the fundamental period of vibration $T_1$ of the structure ($T_1$=2,53 s). This is the seismic intensity measure (IM) adopted for IDA analyses where Sa($T_1$) values have been progressively increased until the occurrence of structural collapse, corresponding to anyone of the following ultimate limit states: column buckling, complete development of a collapse mechanism, attainment of the limit value of plastic rotation of frictional devices or columns.

| Earhquake (record) | Component | Date | PGA/g | Length (s) | Step recording (s) |
|---|---|---|---|---|---|
| Coalinga (Slack Canyon) | H-SCN045 | 1985/05/02 | 0.166 | 29.99 | 0.01 |
| Friuli, Italy (Buia) | B-BUI000 | 1976/09/15 | 0.110 | 26.385 | 0.005 |
| Imperial Valley (Agrarias) | H-AGR003 | 1979/10/15 | 0.370 | 28.35 | 0.01 |
| Kobe (Kakogawa) | KAK000 | 1995/01/16 | 0.251 | 40.95 | 0.01 |
| Northridge (Stone Canyon) | SCR000 | 1994/01/17 | 0.252 | 39.99 | 0.01 |
| Santa Barbara (Courthouse) | SBA132 | 1978/08/13 | 0.102 | 12.57 | 0.01 |
| Spitak Armenia (Gaukasian) | GKS000 | 1998/12/07 | 0.199 | 19.89 | 0.01 |
| Friuli, Italy (Tolmezzo) | TMZ000 | 1976/05/06 | 0.351 | 36.345 | 0.005 |
| Irpinia (Calitri) | A-CTR000 | 1980/11/23 | 0.132 | 35.79 | 0.0024 |
| Victoria, Mexico (Chihuahua) | CHI102 | 1980/06/09 | 0.150 | 26.91 | 0.01 |

Table 6: Accelerogram characteristics

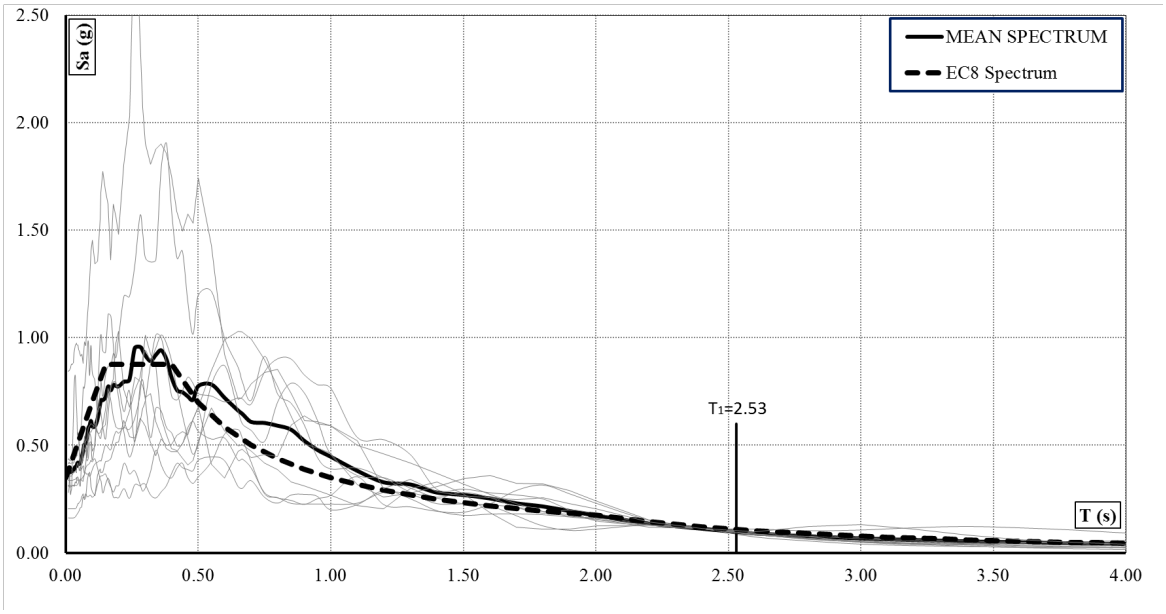



Figure 7: Response spectra (soil type A, ζ=5%) scale at the same value of Sa for T1=2,53

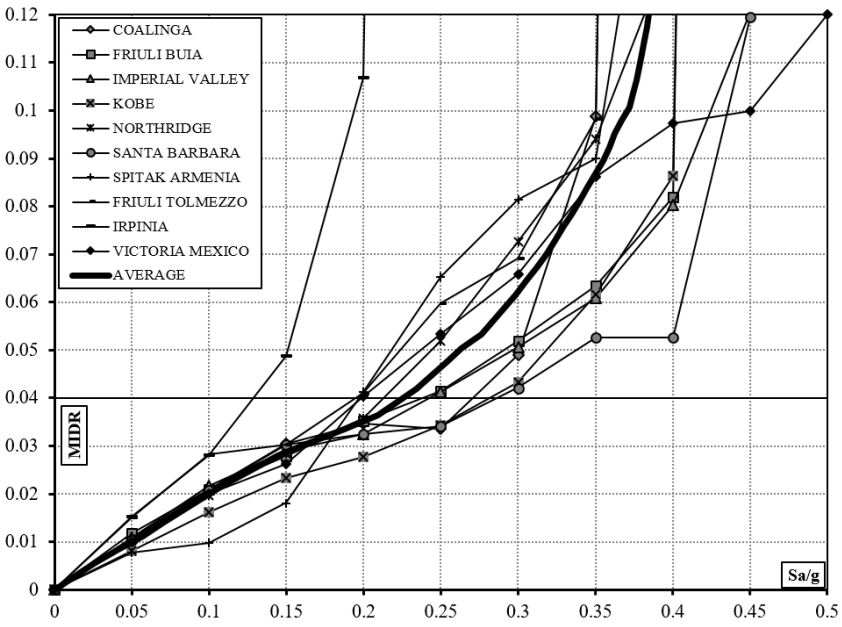



Figure 8: Maximum Interstorey Drift Ratio (MIDR) vs Spectral Acceleration (Sa/g)

In Figure 8 the Maximum Interstorey Drift Ratio (MIDR) vs Spectral Acceleration is reported. MIDR curves appears regular and always increasing reaching on average a spectral acceleration of 0.23 g corresponding to the achievement of the target drift equal to 0.04 rad. In addition, for each accelerogram, the obtained pattern of yielding has been monitored as far as the spectral acceleration increases confirming the development of a global mechanism.

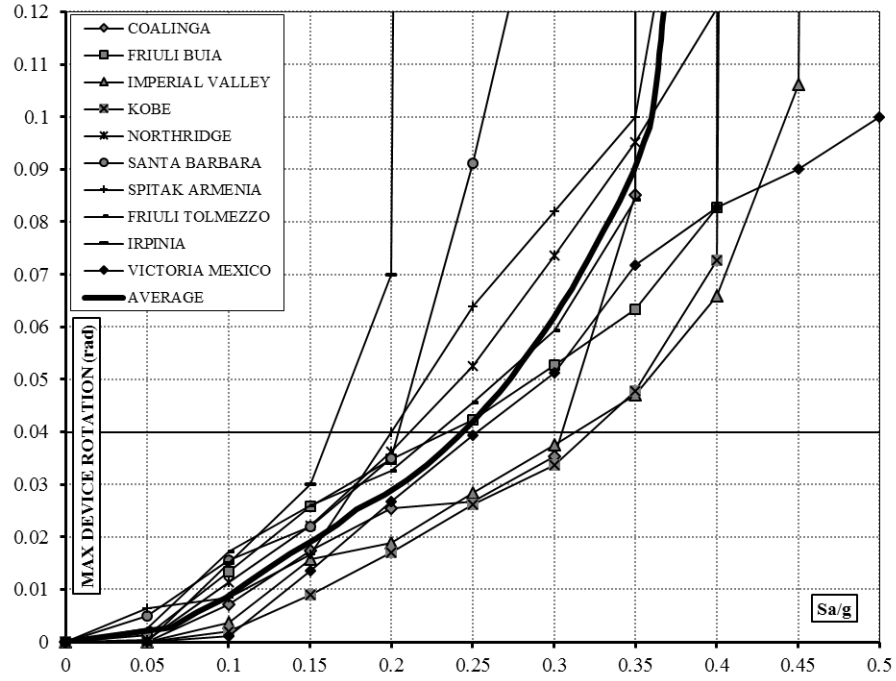



Figure 9: Maximum Device Rotation vs Spectral Acceleration (Sa/g)

In Figure 9 the Maximum Device Rotation vs Spectral Acceleration is reported. In addition, in this case, curves appear regular and increasing.

| | **MIDR=0.04 rad** | | **Collapse** | |
|---|---|---|---|---|
| **Earhquake (record)** | **Sa/g** | **$PGA_c$/g** | **Sa/g** | **$PGA_c$/g** |
| Coalinga (Slack Canyon) | 0.30 | 0.832 | 0.32 | 0.888 |
| Friuli, Italy (Buia) | 0.25 | 0.796 | 0.24 | 0.860 |
| Imperial Valley (Agrarias) | 0.25 | 0.617 | 0.32 | 0.790 |
| Kobe (Kakogawa) | 0.30 | 0.804 | 0.32 | 0.991 |
| Northridge (Stone Canyon) | 0.23 | 1.036 | 0.23 | 1.036 |
| Santa Barbara (Courthouse) | 0.30 | 0.900 | 0.20 | 0.600 |
| Spitak Armenia (Gaukasian) | 0.20 | 0.736 | 0.20 | 0.736 |
| Friuli, Italy (Tolmezzo) | 0.20 | 1.500 | 0.23 | 1.725 |
| Irpinia (Calitri) | 0.125 | 0.228 | 0.17 | 0.307 |
| Victoria, Mexico (Chihuahua) | 0.20 | 0.288 | 0.25 | 0.360 |
| **Mean value** | **0.23** | **0.774** | **0.248** | **0.829** |

Table 7: $S_a(T_1)$ and PGA values corresponding to the attainment of the structural collapse

Finally, for reason of clarity, the spectral acceleration values leading to collapse are given in Table 7 both in term of spectral acceleration ($S_a$) than PGA. In particular, the average value of $S_a(T_1)$ leading to collapse is about 0.248 g while the average PGA is about 0.859 g. However, target drift (0.04 rad) is achieved for lower values of spectral acceleration and PGA because of the high elastic displacement exhibited by the structure due to the pin-joint connection at the bottom of the first storey columns. Anyway, the collapse PGA is very high if compared with the design one (0,2 g) and it is considered always achieved when dissipative devices reach the target rotation equal to 0.04 rad being not observed unsatisfactory failure mechanism such as out of plane buckling of columns or the development of undesired mechanism of collapse. Finally, it important to observe that, although the structure exhibit low val-

ues of the spectral acceleration at the collapse, dissipative devices have only achieved the device stroke and are even able to resort to other ductility resources such as the yielding of bolt in shear.

## 5 CONCLUSIONS

- In this work reference is made to pin-jointed MR-Frames with beam-to-column connections equipped with frictional devices.
- The control of devices activation and, as a consequence, the optimization of energy dissipation capacity is obtained by a rigorous design procedure assuring that only the dissipative zones are involved in plastic range while non dissipative ones remain elastic called Theory of Plastic Mechanism Control (TPMC)
- Thanks to the TPMC and being traditional dissipative zones substituted by frictional devices, column bases pin jointed and column sections designed in order to remain in elastic range, the structure can be considered as free from damage.
- By designing dissipative zones proportionally to the storey shear, they are assured to be engaged in yielding as contemporary as possible, benefit optimization of the seismic response.
- The accuracy of the design approach has been investigated by means push-over and nonlinear dynamic analyses by applying a properly chosen set of earthquake ground motions.
- The global mechanism is confirmed to be achieved by both push-over than IDA analyses.
- The designed structure show quite high performances that are, anyhow, compromised by the high structural displacement due to the pin-jointed column bases.

## REFERENCES

[1] Mazzolani, F.M., Piluso, V., "Theory and Design of Seismic Resistant Steel Frames", E&FN Spon, an imprint of Chapman & Hall, 1996.

[2] EN 1998-1: "Eurocode 8: Design of Structures for Earthquake Resistance – Part 1: general Rules, Seismic Actions and Rules for Buildings", CEN, 2004

[3] Mazzolani, F.M., Piluso, V., "Plastic Design of Seismic Resistant Steel Frames", Earthquake Engineering and Structural Dynamics, Vol. 26, pp. 167-191, 1997.

[4] Montuori, R., Nastri, E., Piluso, V., "Advances in Theory of Plastic Mechanism Control: Closed Form Solution for Mr-Frames", *Earthquake Engineering and Structural Dynamics*, 2015.

[5] Montuori, R., Nastri, E., Piluso, V., "Theory of plastic mechanism control for MRF-EBF dual systems: Closed Form Solution", Submitted for publication to Engineering Structures, 2016.

[6] Longo, A.; Montuori, R.; Piluso, V. Theory of Plastic Mechanism Control for MRF-CBF dual system and its validation. *Bulletin of Earthquake Engineering* (BEEE). 2014.

[7] Giugliano M.T., Longo A., Montuori R., Piluso V. Failure mode and drift control of MRF-CBF dual systems *The Open Construction and Building Technology Journal*, 4, 121-133, 2010.

[8] Montuori R., Piluso V., Troisi M. Theory of plastic mechanism control of seismic-resistant MR-frames with set-backs. *Open Construction and Building Technology Journal*, vol. 6, pp. 404-413 , 2012.

[9] Longo A., Montuori R., Piluso V., Theory of Plastic Mechanism Control of Dissipative Truss Moment Frames, *Engineering Structures*, Vol. 37, pp. 63-75, 2012.

[10] Longo A., Montuori R., Piluso V., Failure Mode Control and Seismic Response of Dissipative Truss Moment Frames, *Journal of Structural Engineering*, Vol. 138, pp.1388-1397, 2012.

[11] Longo A., Montuori, R., Nastri, E., Piluso, V. On the use of HSS in seismic-resistant structures *Journal of Constructional Steel Research*, Vol. 103, p. 1-12, 2014.

[12] Longo A., Montuori, R., Piluso, V., Moment Frames - Concentrically Braced Frames Dual Systems: Analysis of Different Design Criteria, *Structure and infrastructure engineering*, Volume 12, Issue 1, 2 January 2016, Pages 122-14.

[13] Longo A., Nastri E., Piluso V. Theory of plastic mechanism control: State-of-the-art. Open Construction and Building Technology Journal. Volume 8, 2014, Pages 262-278

[14] Longo A., Montuori R., Piluso V. Failure mode control of X-braced frames under seismic actions. *Journal of Earthquake Engineering*, 12: 728-759, 2008

[15] Longo A, Montuori R, Piluso V. Plastic design of seismic resistant V-Braced frames. *Journal of Earthquake Engineering,* vol. 12: 1246-1266, 2008.

[16] Longo A, Montuori R, Piluso V. "Influence of design Criteria on the seismic Reliability of X-Braced Frames", *Journal of Earthquake Engineering*, Vol. 12, Issue 3– p.406-431. 2008

[17] Montuori, R., Nastri, E., Piluso, V.,"Rigid-plastic analysis and moment–shear interaction for hierarchy criteria of inverted Y EB-Frames" *Journal of Constructional Steel Research* vol. 95 71-80. 2014.

[18] Montuori, R., Nastri, E., Piluso, V., Theory of Plastic Mechanism Control for Eccentrically Braced Frames with inverted Y-scheme, *Journal of Constructional Steel Research*, Volume 92, pp. 122-135, 2014.

[19] Giugliano M.T., Longo A., Montuori R., Piluso V., Plastic design of CB-frames with reduced section solution for bracing members. *Journal of Constructional Steel Research.* Vol. 66 pp 611-621. 2010.

[20] Longo A., Montuori R., Piluso V. Seismic reliability of V-braced frames: Influence of design methodologies. *Earthquake Engineering and Structural Dynamics*, vol. 38 p. 1587-1608, 2009.

[21] Longo A., Montuori R., Piluso V. Seismic Reliability of Chevron Braced Frames with Innovative Conception of Bracing Members, *Advanced Steel Construction, an International Journal* - Vol. 5, No. 4, December 2009.

[22] Montuori, R. , Muscati, R. Plastic design of seismic resistant reinforced concrete frame. Earthquake and Structures Volume 8, Issue 1, 2015, Pages 205-224.

[23] Montuori, R. , Muscati, R. A General Procedure for Failure Mechanism Control of Reinforced Concrete Frames. Submitted for publication to Engineering Structures.

[24] Castaldo P, Palazzo B, Della Vecchia P., "Seismic reliability of base-isolated structures with friction pendulum bearings", *Engineering Structures*, Vol. 95, 80-93, 2015

[25] Castaldo P, Tubaldi E., "Influence of FPS bearing properties on the seismic performance of base-isolated structures", *Earthquake Engineering and Structural Dynamics* 2015; Vol. 44, 2817–2836, 2015.

[26] Castaldo P., "Integrated Seismic Design of Structure and Control Systems", *Springer International Publishing: New York*, 2014. DOI 10.1007/978-3-319-02615-2, 2014.

[27] Castaldo P., De Iuliis M., "Optimal integrated seismic design of structural and viscoelastic bracing-damper systems", *Earthquake Engineering and Structural Dynamics*, Vol. 43(12),1809–1827, 2014.

[28] De Iuliis M., Castaldo P., "An energy-based approach to the seismic control of one-way asymmetrical structural systems using semi-active devices", *Ingegneria Sismica - International Journal of Earthquake Engineering 2012*; XXIX(4):31–42, 2012.

[29] Latour, M., Piluso, V., Rizzano, G. Free from damage beam-to-column joints: Testing and design of DST connections with friction pads. Engineering Structures. Volume 85, February 05, Pages 219-233, 2015.

[30] Montuori R., Nastri E., Piluso V. Theory of plastic mechanism control for the seismic design of braced frames equipped with friction dampers - *Mechanics Research Communications* – Vol. 58. pp. 112-123 2014.

[31] Montuori R., Piluso V., Troisi M. Innovative structural details in MR-frames for free from damage structures Mechanics Research Communications – Vol. 58 pp. 146-156. 2014.

[32] Lee S-S., Goel S.C., "Performance-Based Design of Steel Moment Frames using Target Drift and Yield Mechanism", Research Report UMCEE 01-17, University of Michigan, December 2011.

[33] Goel S.C., Chao S-H., "Performance-Based Plastic Design: Earthquake Resistant steel Structures", International Code Council: Washington, DC, 2008.

[34] Latour M., Rizzano G., Piluso V. (2012): "Experimental Analysis of Innovative Dissipative Bolted Double Split Tee Beam-to-Column Connections", *Steel Construction*, 4 (2011), N.2, 53-64.

[35] Latour M., Piluso V., Rizzano G. (2014), "Experimental analysis on friction materials for supplemental damping devices", Construction and Building Materials, Volume 65, 29 August 2014, 159-176.

[36] CSI 2007. SAP 2000: Integrated Finite Element Analysis and Design of Structures. Analysis Reference. Computer and Structure Inc. University of California, Berkeley.

[37] Pacific Earthquake Engineering Research Center, PEER Strong Motion Database, http://peer.berkley.edu.smcat.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ECCENTRICALLY BRACED FRAMES DESIGNED FOR THE ENERGY DISSIPATION OPTIMIZATION

**Elide Nastri**[1]

[1] University of Salerno
Via Giovanni Paolo II, 132, 84084, Fisciano (SA), Italy
e-mail: enastri@unisa.it

**Keywords:** Eccentrically Braced Frames, dissipation optimization, steel, dissipative devices, free from damage structures, Theory of Plastic Mechanism Control.

**Abstract.** *The work herein presented has the main purpose to show a practical application of a proper design procedure able to dimension Eccentrically Braced Frames with Inverted Y-scheme equipped with removable dissipative devices substituting the link member. Column bases are conceived with pin-jointed base connections in order to assure that the global mechanism is achieved only when link members results activated in plastic range. The control of devices activation and the optimization of energy dissipation capacity is obtained by applying a rigorous design procedure assuring that only the dissipative zones are involved in plastic range while non dissipative ones remain in elastic range. Dissipative devices are designed in order to bear the whole storey seismic shear without a strength and stiffness degradation. In addition, being the link members substituted by dissipative devices, column bases pin jointed and column sections designed in order to remain in elastic range by a properly design procedure, the structure can be considered as free from damage. In fact after a destructive seismic event all the damaged devices can be replaced by new ones. The accuracy of the design approach has been investigated by means push-over and non-linear dynamic analyses by applying a properly chosen set of earthquake ground motions.*

## 1 INTRODUCTION

In the framework of seismic resistant structure, Eccentrically Braced Frames (EBFs) constitute a quite recent structural typology. They gained prominence thanks to the study of Popov and Kasai [1]-[3]. This structural typology is well suited for tall buildings located in areas of high seismic intensity. For this reason, EBFs are especially widespread in USA and New Zealand where the recent Christchurch earthquake of the February 11th 2011 put to the test a great number of structures. In particular, this unfortunate event allowed testing on real scale the damage a high intensity earthquake is able to bring on EBF steel frames. As regards their working under seismic actions, EBFs constitute a suitable compromise between seismic resistant MR-frames and concentrically braced frames because they exhibit both adequate lateral stiffness, due to the high contribution coming from the diagonal braces, and ductile behaviour, due to the ability of the links, constituting the dissipative zones of this structural typology, in developing wide and stable hysteresis loops. Therefore, the optimization of dissipative behaviour of this structural typology is obtained when all the links are involved in plastic range. In this paper, reference is made to Inverted Y-scheme EBFs, that are a particular configuration of EBFs reported in Eurocode 8 [4] still not sufficiently investigated and not largely widespread despite having many advantages both in term of performance and construction. Its main characteristic is that the link, i.e. the dissipative zone, does not belong to the beam member. In fact, one of the primary benefits in using such structural typology regards the chance to substitute easily the damaged link after a destructive seismic event, and, in addition, the possibility to conceive the scheme within the framework of supplementary energy dissipation [5]-[10], by substituting the vertical link member with a dissipative device, such as a friction damper [11] or hysteretic damper, which is able to exhibit a highly dissipative behaviour if compared with traditional link members. As damaged links can be easily removed and substituted after earthquake, such structural scheme exhibits the greatest advantages provided that the other structural members as beams, diagonals and columns, have been not damaged during the seismic event, i.e. have remained in elastic range. This is precisely why a proper design is of paramount importance. In fact, only with collapse mechanism of global type it is possible to assure that damage is concentrated only in dissipative zones while the other non-dissipative ones remain in elastic range. This important scope is, out of doubts, the strength of the Theory of Plastic Mechanism Control.

The Theory of Plastic Mechanism Control (TPMC) [12], differently from the rules proposed by EC8 allows designing structure always showing at the collapse a mechanism of global type. This result is assured thanks both to the strong background of the theory, that is based on the kinematic theorem of plastic collapse, than to the successive improvements of the design procedure that have led to a recent version that is easy to be carried out by means of hand calculation. This last form of the design procedure is the so called “closed form solution of TPMC” that has been satisfactorily applied to design both MRFs [13] than EBF-MRFs dual systems [14]. However, this procedure in the previous but as the same efficient form as the improved one has been applied to design all the structural steel typology [15] - [30] and also reinforced concrete MRFs [31]-[32].

Even though, this procedure is able to assure a collapse mechanism of global type, the seismic optimization has to be pursued by assuring that all the dissipative zones engage in yielding as possible as the same time [33]. For this reason, in this structural example, link members, have been substituted by innovative devices [34] inspired to frictional devices investigated in the framework of the European Research Project “FREEDAM”. As a consequence the design procedure based on TPMC has improved to promote the contemporary activation of all dissipative devices by additionally considering the column bases at the first

storey, beam and diagonal ends has simply pinned. In this way, being only the device activated, beams, diagonals and column bases pin jointed and column sections designed in order to remain in elastic range by a properly design procedure, the structure can be considered as free from damage [11], [35]. The accuracy of the design approach has been investigated by means push-over and non-linear dynamic analyses by applying a properly chosen set of earthquake ground motions.

## 2 DESIGN APPROACH

The "Theory of Plastic Mechanism Control" design approach has the aim to provide structures able to fail with a collapse mechanism of global type. Global mechanism represents the optimum in term of structural dissipation capacity being all the dissipative zones involved in the pattern of yielding while the non dissipative ones remain in elastic range. Dissipative zones of traditional EBFs are the links and the bases of the column at the first storey. However, in this particular case, links are substituted by special dissipative devices and column bases at the bottom of the first storey by means of pin connections. In this way plastic hinges cannot develop at the bottom of the first storey columns, so that, only the link ends are involved in plastic range being beam equally hinged at their ends. In order to design members to fulfil this design goal devices and column sections have to be designed in order to assure on one hand the optimization of the seismic response and from the other hand a collapse mechanism of global type. The first design unknown having to be provided are the devices substituting link members for which, the whole sum of design slip moment has to be provided by assuring that, according with the kinematic theorem of plastic collapse extended to the concept of mechanism equilibrium curve, the collapse multiplier $\alpha$ of the seismic storey forces $F_k$ has to be at least equal to 1 at a design displacement $\delta_u$ compatible with the ductility supply of dissipative zone:

$$\alpha = \alpha_0^{(g)} - \gamma^{(g)}\delta_u \geq 1 \;\Rightarrow\; \frac{2M_{fl.Ed.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \geq 1 \tag{1}$$

where $\alpha_0^{(g)}$ is the first order collapse multiplier:

$$\alpha_0^{(g)} = \frac{2M_{fl.Ed.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} \tag{2}$$

and $\gamma^{(g)}$ is the slope of mechanism equilibrium curve accounting for second order effects:

$$\gamma^{(g)} = \frac{\frac{1}{h_{ns}}\sum_{k=1}^{n_s}V_k h_k}{\sum_{k=1}^{n_s}F_k h_k} \tag{3}$$

In particular, $V_k$ and $h_k$ are the sum of all the vertical loads acting at each storey and the storey height, respectively. By designing dissipative zones proportionally to the storey shear they are assured to be engaged in yielding as contemporary as possible, benefit optimization of the seismic response. In addition, the design ultimate displacement has to be dependent on the device stroke, i.e. the ultimate distance the dissipative device is able to tread by slippage. By solving as an equality Eq. (1) it is possible to provide the sum of design slip moments of devices at the last storey , $M_{fl.Ed.n_s}$:

$$M_{fl.Ed.n_s}\sum_{k=1}^{n_s}\beta_k = \frac{\left(\gamma^{(g)}\delta_u + 1\right)}{2}\sum_{k=1}^{n_s}F_k h_k \tag{4}$$

that shared by the term $\beta_k = \sum_{i_m=k}^{n_s}F_k / F_{n_s}$, [36]-[37] provides the sum of design slip moment at each storey optimized to promote the contemporary engagement in slippage, $M_{fl.Ed.k}$:

$$M_{fl.Ed.k} = \beta_k M_{fl.Ed.n_s} \tag{5}$$

In addition, if the same frame has more braced bays the quantity provided by Eq. (5) need to be shared between all the devices belonging to the braced bay.

The design slip moment $M_{fl.Ed.jk}$ is used to design frictional devices whose resistance is expressed by means of the following term $M_{fl.Cd.jk}$.

Beams are non dissipative zones and to be designed by means as the maximum among the design moment belonging to the vertical load distribution and, according to the second capacity design principle, those provided by a local hierarchy criterion where $\gamma_{Rd}$ is an overstrength factor [26].

$$M_{b.Cd.jk} = \max\left(\frac{q_{v.j}L_j^2}{8}; \gamma_{Rd}\frac{M_{fl.Cd.jk}}{2}\right) \tag{6}$$

As regards the column design, all the possible mechanism typology which can affect a seismic resistant frame should be considered. In particular, it is possible to observe that EBFs under seismic horizontal forces should fail according to three main collapse typologies called type-1, type-2 and type-3, as reported in Figure 1 for the specific case of pin-jointed column bases where the rectangles represents the dissipative devices involved while the solid circles the plastic hinges developing at the top or bottom end of columns for undesired mechanisms. Every one of these can involve each storey and, for this reason, being the number of storey equal to $n_s$ the number of possible mechanisms involving a structure can be easily computed as $3n_s$. Only one of these mechanisms is the desired one, i.e. the global one, and it is a particular case of type-2 mechanism extended to all the $n_s$ storeys. It means that all the other $3n_s - 1$ mechanisms are undesired and must be avoid by applying TPMC.

The design conditions also found their theoretical bases in the kinematic theorem of plastic collapse extended to the concept of collapse mechanism equilibrium curve which is representative of a straight line whose intercept with the vertical axis in a Cartesian diagram is the collapse mechanism multiplier of the first order, $\alpha_{i_m}^{(t)}$, while $\gamma_{i_m}^{(t)}$ is the slope where $i_m$ and $t$ are the mechanism index and the mechanism typology code respectively.

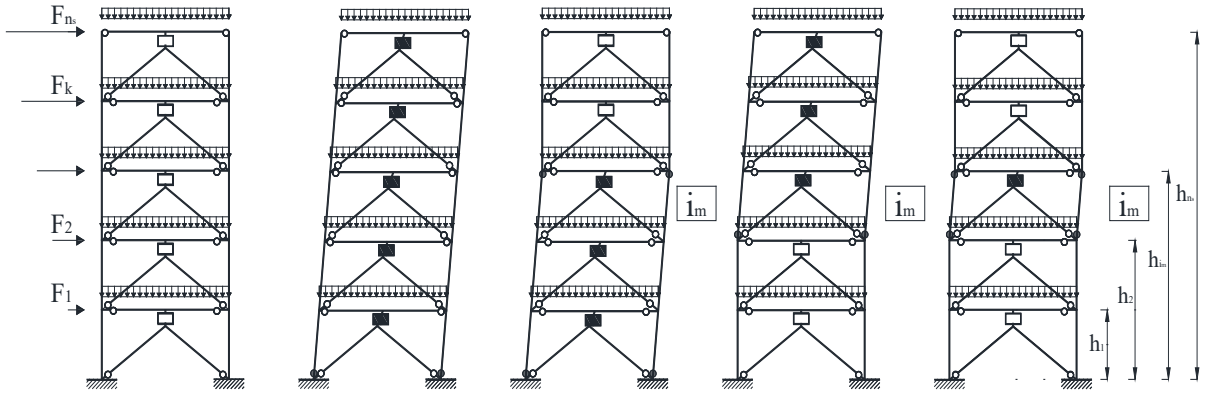



Figure 1: Collapse mechanism of Inverted Y EBFs equipped with dissipative devices

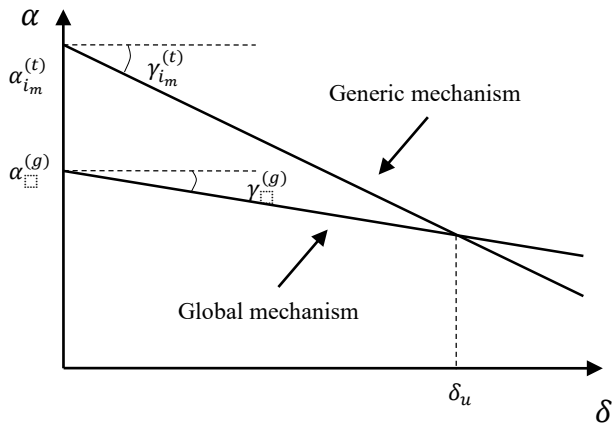



Figure 2: Design condition for the failure mode control

TPMC states that the global mechanism equilibrium curve has to be located below those corresponding to all the other mechanisms within a displacement range compatible with the plastic deformation capacity of dissipative zones, $\delta_u$ (Figure 2). Consequently, the design conditions are provided by the following equations:

$$\alpha^{(g)} - \gamma^{(g)}\delta_u \leq \alpha_{i_m}^{(t)} - \gamma_{i_m}^{(t)}\delta_u \qquad i_m = 1,2,3,\dots,n_s \qquad t = 1,2,3 \tag{7}$$

The computation of the quantities $\alpha_{i_m}^{(t)}$ and $\gamma_{i_m}^{(t)}$ exploits the virtual work principle and it is amply detailed and discussed in a previous work [5]-[13]. In particular, the expressions of $\alpha_{i_m}^{(t)}$ and $\gamma_{i_m}^{(t)}$ are reported in Table 1. In particular, according to the second capacity design principle, non dissipative zones have to be designed according to the maximum actions that dissipative zones, i.e. dissipative devices, are able to transmit in their ultimate conditions. For this reason, frictional device overstrength have to be considered by introducing the over-strength factor $\gamma_{Rd}$.

| ***Type-1*** | ***Type-2*** | ***Type-3*** |
|---|---|---|
| $\alpha_{0.i_m}^{(1)} = \frac{\gamma_{Rd} 2M_{fl.Cd.n_s}\sum_{k=1}^{i_m-1}\beta_k + \sum_{i=1}^{n_c} M_{c.ii_m}^{(1)}}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k}$ | $\alpha_{0.i_m}^{(2)} = \frac{\sum_{i=1}^{n_c} M_{c.ii_m}^{(2)} + \gamma_{Rd} 2M_{fl.Cd.n_s}\sum_{k=i_m}^{n_s}\beta_k}{\sum_{k=i_m}^{n_s} F_k(h_k - h_{i_m-1})}$ | $\alpha_{0.i_m}^{(3)} = \frac{2\sum_{i=1}^{n_c} M_{c.ii_m}^{(3)}}{(h_{i_m} - h_{i_m-1})\sum_{k=i_m}^{n_s} F_k}$ |
| $\gamma_{i_m}^{(1)} = \frac{\sum_{k=1}^{i_m} V_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} V_k}{h_{i_m}\left(\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k\right)}$ | $\gamma_{i_m}^{(2)} = \frac{\sum_{k=i_m}^{n_s} V_k(h_k - h_{i_m-1})}{(h_{n_s} - h_{i_m-1})\sum_{k=i_m}^{n_s} F_k(h_k - h_{i_m-1})}$ | $\gamma_{i_m}^{(3)} = \frac{\sum_{k=i_m}^{n_s} V_k}{(h_{i_m} - h_{i_m-1})\sum_{k=i_m}^{n_s} F_k}$ |

Table 1. First order and second order collapse mechanism multiplier at the generic $i_m - th$ storey

Therfore, by substituting terms reported in Table 1 in Eq. (7), the following design relationship are obtained for type-1 mechanism:

$$\gamma_{Rd}\frac{2M_{fl.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s} F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{2M_{fl.Cd.n_s}\sum_{k=1}^{i_m-1}\beta_k + \sum_{i=1}^{n_c} M_{c.ii_m}^{(1)}}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k} - \gamma_{i_m}^{(1)}\delta_u \tag{8}$$

for type-2 mechanism:

$$\gamma_{Rd}\frac{2M_{fl.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s} F_k h_k} - \gamma^{(g)}\delta_u \leq \frac{\sum_{i=1}^{n_c} M_{c.ii_m}^{(2)} + \gamma_{Rd} 2M_{fl.Cd.n_s}\sum_{k=i_m}^{n_s}\beta_k}{\sum_{k=i_m}^{n_s} F_k(h_k - h_{i_m-1})} - \gamma_{i_m}^{(2)}\delta_u \tag{9}$$

and for type-3 mechanism :

$$\begin{aligned} for\ i_m = 1 \qquad & \gamma_{Rd}\frac{2M_{fl.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \le \frac{\sum_{i=1}^{n_c}M_{c.i1}^{(3)}}{h_1\sum_{k=1}^{n_s}F_k} - \gamma_1^{(3)}\delta_u \\ for\ i_m > 1 \qquad & \gamma_{Rd}\frac{2M_{fl.Cd.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \gamma^{(g)}\delta_u \le \frac{2\sum_{i=1}^{n_c}M_{c.ii_m}^{(3)}}{\left(h_{i_m} - h_{i_m-1}\right)\sum_{k=i_m}^{n_s}F_k} - \gamma_{i_m}^{(3)}\delta_u \end{aligned} \tag{10}$$

That rearranged to provide the unknown of the design procedure, i.e. the sum of column plastic moment at each storey $M_{c.ii_m}^{(t)}$ reported in the following.

In particular, at the first storey the condition reported in Eq. (8) provide only one design relationship being easily demonstrable that mechanism type-1 and type-3 at the first storey are coincident and mechanism type-2 is coincident with the global one:

$$\sum_{i=1}^{n_c}M_{c.i1}^{(1or3)} \ge \left[\gamma_{Rd}\frac{2M_{fl.Ed.n_s}\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} + \left(\gamma_1^{(3)} - \gamma^{(g)}\right)\delta_u\right]\left(h_{i_m} - h_{i_m-1}\right)\sum_{k=i_m}^{n_s}F_k \tag{11}$$

At the $i_m > 1$ storeys, design conditions have to be make explicit for all the undesired mechanism typologies:

$$\begin{aligned} \sum_{i=1}^{n_c}M_{c.ii_m}^{(1)} \ge \Bigg[ & \gamma_{Rd}2M_{fl.Cd.n_s}\left(\frac{\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \frac{\sum_{k=1}^{i_m-1}\beta_k}{\sum_{k=1}^{i_m}F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}F_k}\right) \\ & + \left(\gamma_{i_m}^{(1)} - \gamma^{(g)}\right)\delta_u\Bigg]\left(\sum_{k=1}^{i_m}F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s}F_k\right) \end{aligned} \tag{12}$$

for type-1 mechanism;

$$\sum_{i=1}^{n_c}M_{c.ii_m}^{(2.1)} \ge \left[\gamma_{Rd}2M_{fl.Cd.n_s}\left(\frac{\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} - \frac{\sum_{k=i_m}^{n_s}\beta_k}{\sum_{k=i_m}^{n_s}F_k\left(h_k - h_{i_m-1}\right)}\right) + \left(\gamma_{i_m}^{(2)} - \gamma^{(g)}\right)\delta_u\right]\sum_{k=1}^{i_m}F_k\left(h_k - h_{i_m-1}\right) \tag{13}$$

for type-2 mechanism;

$$\sum_{i=1}^{n_c}M_{c.ii_m}^{(3)} \ge \left[\gamma_{Rd}2M_{fl.Cd.n_s}\frac{\sum_{k=1}^{n_s}\beta_k}{\sum_{k=1}^{n_s}F_k h_k} + \left(\gamma_1^{(3)} - \gamma^{(g)}\right)\delta_u\right]\frac{\left(h_{i_m} - h_{i_m-1}\right)}{2}\sum_{k=i_m}^{n_s}F_k \tag{14}$$

for type-3 mechanism.

While for the $i_m = 1$ storey the design condition is only one Eq. (12) for the $i_m$>1 storeys, the sum of the reduced plastic moments of columns has to be calculated as the maximum among the three mechanism typologies, according to the following relation:

$$\sum_{i=1}^{n_c}M_{c.ii_m} = max\left\{\sum_{i=1}^{n_c}M_{c.ii_m}^{(1)}, \sum_{i=1}^{n_c}M_{c.ii_m}^{(2)}, \sum_{i=1}^{n_c}M_{c.ii_m}^{(3)}\right\} \tag{15}$$

The knowledge of these plastic moments, coupled with the estimated axial force at the collapse state, allows the evaluation of the required sections of columns.

The sum provided by Eq. (15) has to be distributed between all the columns at each storey according to two possible strategies [13]:

1) the sum are subdivided equally between each column

2) the sum are subdivided according to the axial load acting on the columns in the collapse condition.

However, in both the cases column sections need to verify the combination of the competent axial load at collapse and bending moment.

The axial load acting on the columns at collapse state can be easily provided in agreement with the global mechanism where axial forces in the columns at collapse state depend both on distributed loads acting on the beams and on the maximum axial load belonging to diagonal members [14], [27]. Also in this case, devices overstrength has to be taken in account according to the second capacity design principle. However, seismic actions can be acting either in the positive direction or in the negative direction, so that the maximum axial forces has to be considered.

## 3 WORKED EXAMPLE

Aiming to practically show how to apply the proposed design procedure a 3 bays 4 storeys pin-jointed column bases EBF whose structural scheme is depicted in Figure 3 has been designed. The bay span is 5 m while the interstorey height is equal to 3.0 m. The characteristic values of the vertical loads acting on beams are equal to $3{,}34 kN/m^2$ and $2{,}50\ kN/m^2$ for permanent and live loads, respectively, while on columns concentrated loads belonging to the secondary warping of floor are applied (Figure 3). With reference to the seismic load combination ($G_k + \psi_2 Q_k + E_d$), the vertical loads acting on the beams of the analysed structure are $4{,}067\ kN/m$. The structural material adopted for all the structural members is an S275. The design horizontal forces have been determined according to EC8, assuming a peak ground acceleration equal to 0.35g, a seismic response factor equal to 2.5, a behaviour factor equal to 6. The design horizontal forces distribution is in accord to the first vibration mode as reported in Figure 3. In addition, vertical loads acting on the inner pendular structure are accounted for by means of a leaning column.

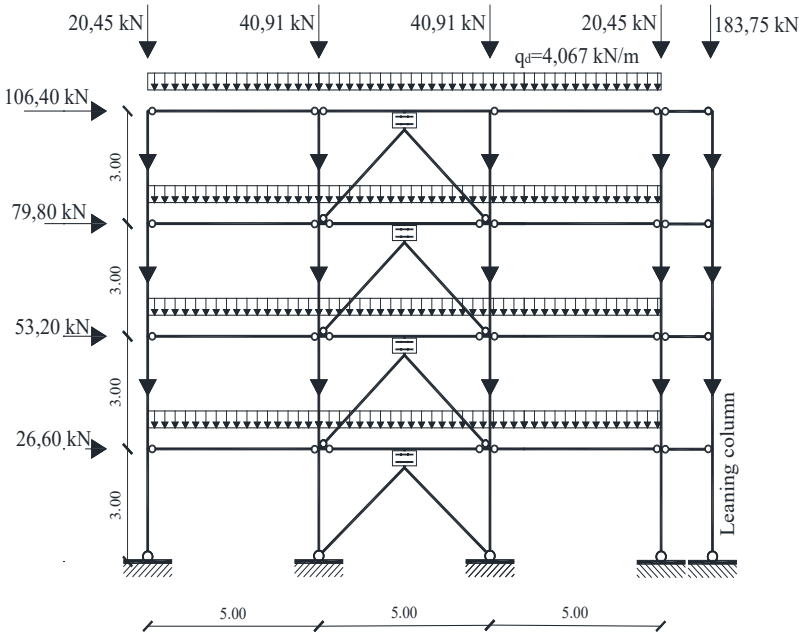



Figure 3: Worked example structural scheme

Frictional devices are made of a layer of friction material, placed between two bolted steel plates which accommodate an inner plate containing slotted holes that allow the slippage of

the device. In addition, external plates have slotted holes also in vertical direction to allow vertical displacement due to vertical loads (Figure 4). Adopted bolts are high resistance class 8.8 while the friction material adopted (FREEDAM project) is sprayed aluminum that is able to exhibit a friction coefficient of about 0.5 [38]-[39].

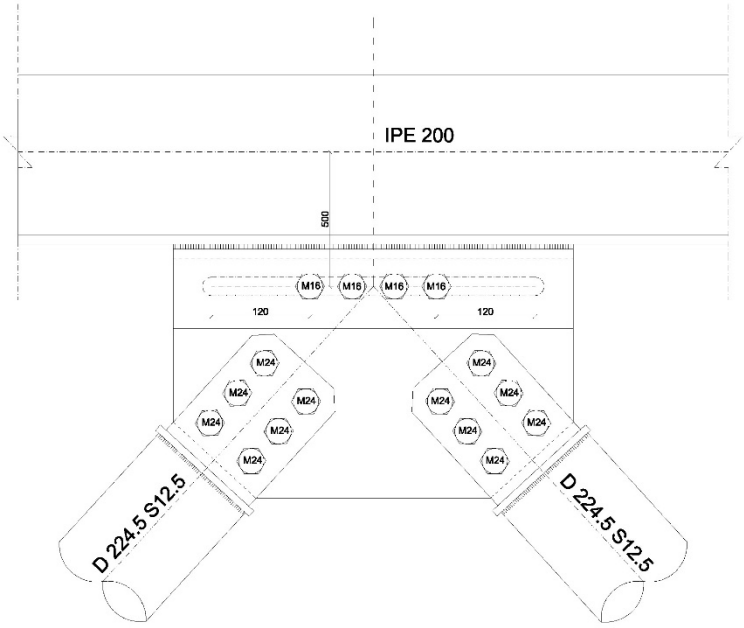



Figure 4: Scheme of dissipative device

### 3.1 Design of dissipative zones

The first quantity needed for the design of the structure is the design slip moment of devices at the last storey (storey $n_s$) provided by Eq. (4), being preliminarily compute the term $\sum_{k=1}^{n_s} \beta_k = 7{,}5$. Successively, the sum of the design slip moment $M_{fb.Ed.n_s}$ has to be shared between all the storeys by means of the term $\beta_k$ to design the dissipative device preliminarily described, according to the following relation:

$$M_{fl.Ed.n_s} = n_s \, n_b \, N_b \, \mu \, e/2 \rightarrow M_{fl.Cd.n_s} \tag{16}$$

where, $n_s$ is the number of frictional surfaces that in this case is equal to 2, $n_b$ is number of bolts, $N_b$ is the preloading force equal to:

$$N_b = 0.7 \, f_{ub} A_s \tag{17}$$

$f_{ub}$ is the design bolt resistance, $A_s$ is the bolt area, $\mu$ is the dynamic frictional coefficient assumed in this case equal to 0.5 and $e$ is the distance between the median axis of the beam and the intersection of brace axes assumed, in this case, equal to 0,5 m. In addition, beams are preliminarily selected from the standard shapes to withstand the vertical load combination $(1.3\mathrm{G_k} + 1.5Q_k)$ according to a design moment computed as in Eq. (6) and are IPE200. In Table 2 the sum of the design slip moment $M_{fl.Ed.k}$ shared at all the storey by means of the term $\beta_k$, the slip resistant moment $M_{fl.Cd.k}$, belonging to the device design, the number of bolts $n_b$, the distance $e$, the preloading force $N_b$ are reported.

| Storey (-) | $\beta_k$ (-) | $M_{fb.Ed.k}$ $(kN\ m)$ | $n_b$ (-) | $e$ (m) | Bolts (-) | $N_b$ $kN$ | $M_{fl.Cd.k}$ $(kN\ m)$ | $\gamma_{Rd} M_{fl.Cd.k}$ $(kN\ m)$ |
|---|---|---|---|---|---|---|---|---|
| | 2.5 | 66.50 | 4 | 0.5 | ϕ16 | 84.00 | 67.20 | 80.64 |
| 2 | 2.25 | 59.85 | 4 | 0.5 | ϕ16 | 75.00 | 60.00 | 72.00 |
| 3 | 1.75 | 46.55 | 4 | 0.5 | ϕ16 | 59.00 | 47.20 | 56.64 |
| 4 | 1 | 26.60 | 4 | 0.5 | ϕ16 | 35.00 | 28.00 | 33.60 |

Table 2: Parameters for the design of beam-to-column connections

Finally, beam sections designed to withstand vertical loads have also to fulfil also the local hierarchy criterion reported in the second term of Eq. (6). Diagonal braces, designed to support in and out of plane buckling are D244,5 s12,5 tubes.

### 3.2 Design of non-dissipative zones

Being designed dissipative zones, also non-dissipative one have to be provided by means of TPMC. The ultimate design displacement governing the procedure has been computed in the following way:

$$\delta_u = 0.04\, h_{n_s} = 0.04 \cdot 12 = 0.48\, m \tag{18}$$

being assumed an ultimate device rotation equal to 0.04 and as a consequence a maximum device stroke of 0,12 m. In addition, in Table 3 the slopes of mechanism equilibrium curves are reported being computed according to the relationships reported in Table 1.

| STOREY $\boldsymbol{i_m}$ | $\boldsymbol{\gamma_{i_m}^{(1)}}$ (1/m) | $\boldsymbol{\gamma_{i_m}^{(2)}}$ (1/m) | $\boldsymbol{\gamma_{i_m}^{(3)}}$ (1/m) |
|---|---|---|---|
| 1 | 0.308 | <u>0.064</u> | 0.308 |
| 2 | 0.142 | 0.077 | 0.256 |
| 3 | 0.089 | 0.105 | 0.220 |
| 4 | 0.064 | 0.192 | 0.192 |

Table 3. Slopes of the mechanism equilibrium curves

It is possible to observe that the slope of the global mechanism equilibrium curve (undersigned in Table 3) is the minimum among all the slope of mechanism equilibrium curves corresponding to the undesired mechanism. In addition, in Table 4 the design bending moment of columns at each storey are reported.

| STOREY $\boldsymbol{i_m}$ | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(1)}}$ (kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(2)}}$(kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}^{(3)}}$(kN m) | $\sum_{i=1}^{n_c} \boldsymbol{M_{c.ii_m}}$(kN m) |
|---|---|---|---|---|
| 1 | 841.741 | [-] | 841.741 | 841.741 |
| 2 | 521.3593 | 84.3591 | 302.8592 | 521.3593 |
| 3 | 235.0923 | 140.1408 | 187.6165 | 235.0923 |
| 4 | [-] | 161.8078 | 80.9039 | 161.8078 |

Table 4. Value of the paramters $\sum_{i=1}^{n_c} M_{c.ii_m}^{(t)}$

In this worked example the approach that have been used to subdivide the reduced sum of plastic moments between the columns exploits the axial load acting on the columns at col-

lapse state. In fact, the maximum value of $M_{c,i_m}$ is subdivided between columns on the basis of axial force acting on each column at each storey.

The reduced sum of plastic moments of columns at each storey, $\sum_{i=1}^{n_c} M_{c.ii_m}$, has been subdivided proportionally to the axial load at the collapse state according to the following equation:

$$M_{c.ik} = \frac{N_{tot.ii_m} \sum_{i=1}^{n_c} M_{c.ii_m}}{\sum_{i=1}^{n_c} N_{tot.ii_m}} \tag{19}$$

where $M_{c.ik}$ is the design moment of column sections.

Finally, in Table 5 the axial load acting at the collapse state, the design moment of column sections and the column sections selected from standard shapes are reported with reference to external and internal columns. The collapse mechanism multiplier of global mechanism $\alpha$ is equal to 1.2175.

| STOREY $i_m$ | External columns | | | Internal columns | | |
|---|---|---|---|---|---|---|
| | $N_{tot.ii_m}$ $(kN)$ | $M_{c.ik}$ (kN m) | PROFILES | $N_{tot.ii_m}$ $(kN)$ | $M_{c.ik}$ (kN m) | PROFILES |
| 1 | 122.50 | 74.50 | HE 160 B | 569.48 | 346.36 | HE 280 B |
| 2 | 91.87 | 52.51 | HE 140 B | 364.23 | 208.17 | HE 220 B |
| 3 | 61.25 | 28.70 | HE 120 B | 189.70 | 88.86 | HE 160 B |
| 4 | 30.62 | 26.97 | HE 100 B | 61.25 | 53.93 | HE 140 B |

Table 5. Axial force at collapse state, design moment of column sections and column sections selected from standard shapes at each storey

## 4 VALIDATION OF PROCEDURE

In order to validate the design procedure static non linear analysis (push-over) has been carried out for the designed EBF frame by means of SAP 2000 computer program [44]. This analysis has the primary aim to predict the collapse mechanism typology, testing the accuracy of the proposed design methodology.

In SAP2000 ambiance all members have been modeled by means of beam-column elements, whose non-linearities have been concentrated in plastic hinges at their ends. In particular, plastic hinges accounting for the interaction between axial force and bending moment have been defined for, columns while dissipative devices have been modelled in shear-elongation such as short links [40]-[43]. Beams and diagonals have been modelled with pure bending plastic hinges and force bending, respectively. The elastic behaviour is not considered in plastic hinge definition because it is directly taken into account by the beam-column element. The analysis has been led under displacement control taking into account both geometrical and mechanical non-linearities. In addition out of plan stability check of compressed members have been performed at each step of the non-linear analysis. The results of the push-over are mainly constituted by the frame capacity curve (Figure 4) whose softening branch is perfectly coincident to the global mechanism equilibrium curve. In addition, push-over curve show a very pronounced elbow that is symptomatic of a contemporary activation of all the dissipative devices and of a consequent seismic energy optimization in terms of dissipative capacity of the structure.

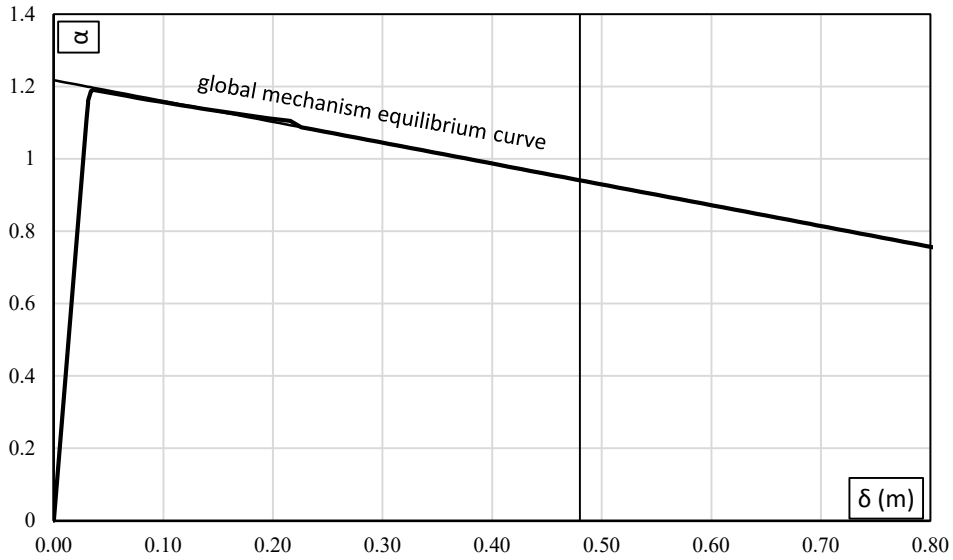



Figure 5: Push-over curve of the analyzed structure

In addition, in order to provide a more robust validation of the design methodology, non-linear incremental dynamic analyses have been carried out with reference to the same structural model used for push-over analyses. In addition, 5% damping according to Rayleigh modelling has been assumed.

Record-to-record variability has been accounted for considering 10 recorded accelerograms selected from PEER [45] data base whose main characteristics (name, date, magnitude, ratio between PGA and gravity acceleration, length and step recording) are reported in Table 6. These earthquake records have been selected to approximately match the linear elastic design response spectrum of Eurocode 8 [4], for type A soil (Figure 6). Moreover, in order to perform IDA analyses, each ground motion has been scaled to obtain the same value of the spectral acceleration Sa ($T_1$) corresponding to the fundamental period of vibration $T_1$ of the structure ($T_1$=0,785 s). This is the seismic intensity measure (IM) adopted for IDA analyses where Sa($T_1$) values have been progressively increased until the occurrence of structural collapse, corresponding to anyone of the following ultimate limit states: column buckling, complete development of a collapse mechanism, attainment of the limit value of the devices stroke.

| Earhquake (record) | Component | Date | PGA/g | Length (s) | Step recording (s) |
|---|---|---|---|---|---|
| Coalinga (Slack Canyon) | H-SCN045 | 1985/05/02 | 0.166 | 29.99 | 0.01 |
| Helena (Carroll College) | A-HMC180 | 1976/09/15 | 0.150 | 39.99 | 0.01 |
| Imperial Valley (Agrarias) | H-AGR003 | 1979/10/15 | 0.370 | 28.35 | 0.01 |
| Kobe (Kakogawa) | KAK000 | 1995/01/16 | 0.251 | 40.95 | 0.01 |
| Northridge (Stone Canyon) | SCR000 | 1994/01/17 | 0.252 | 39.99 | 0.01 |
| Santa Barbara (Courthouse) | SBA132 | 1978/08/13 | 0.102 | 12.57 | 0.01 |
| Spitak Armenia (Gaukasian) | GKS000 | 1998/12/07 | 0.199 | 19.89 | 0.01 |
| Friuli, Italy (Tolmezzo) | TMZ000 | 1976/05/06 | 0.351 | 36.345 | 0.005 |
| Irpinia (Calitri) | A-CTR000 | 1980/11/23 | 0.132 | 35.79 | 0.0024 |
| Victoria, Mexico (Chihuahua) | CHI102 | 1980/06/09 | 0.150 | 26.91 | 0.01 |

Table 6: Accelerogram characteristics

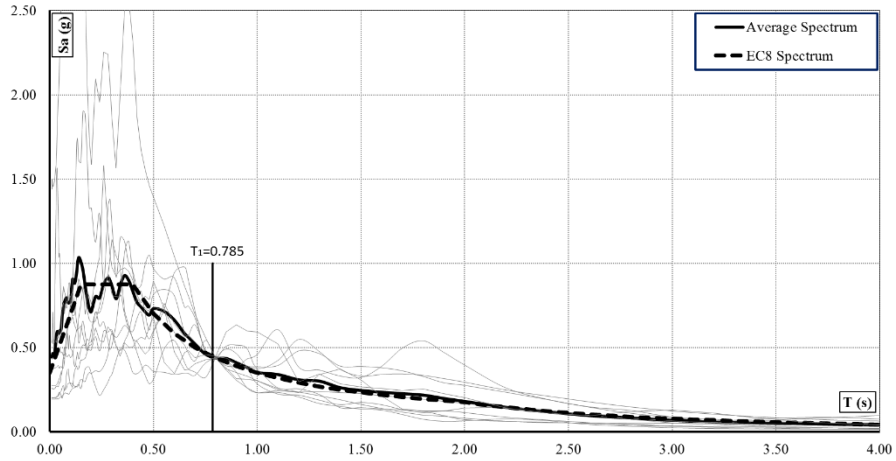



Figure 6: Response spectra (soil type A, ζ=5%) scale at the same value of Sa for T1=0,785

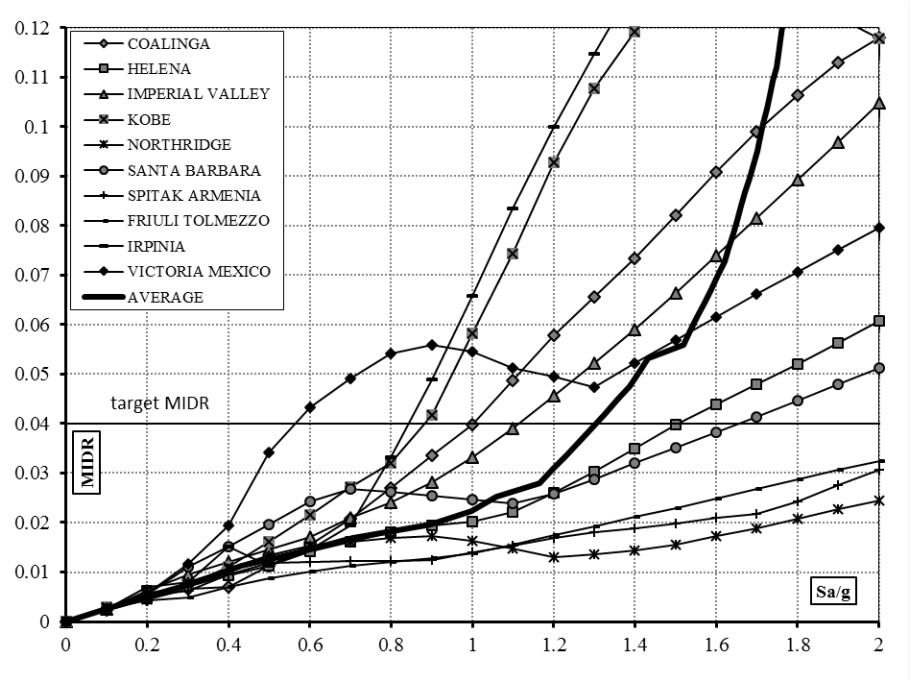



Figure 7: Maximum Interstorey Drift Ratio (MIDR) vs Spectral Acceleration (Sa/g)

In Figure 6 the Maximum Interstorey Drift Ratio vs Spectral Acceleration is reported. MIDR curves appears regular and quite always increasing reaching on average a spectral acceleration of 1,30 g corresponding to the achievement of the target drift equal to 0.04 rad. In addition, for each accelerogram, the obtained pattern of yielding has been monitored as far as the spectral acceleration increases confirming the development of a global mechanism.

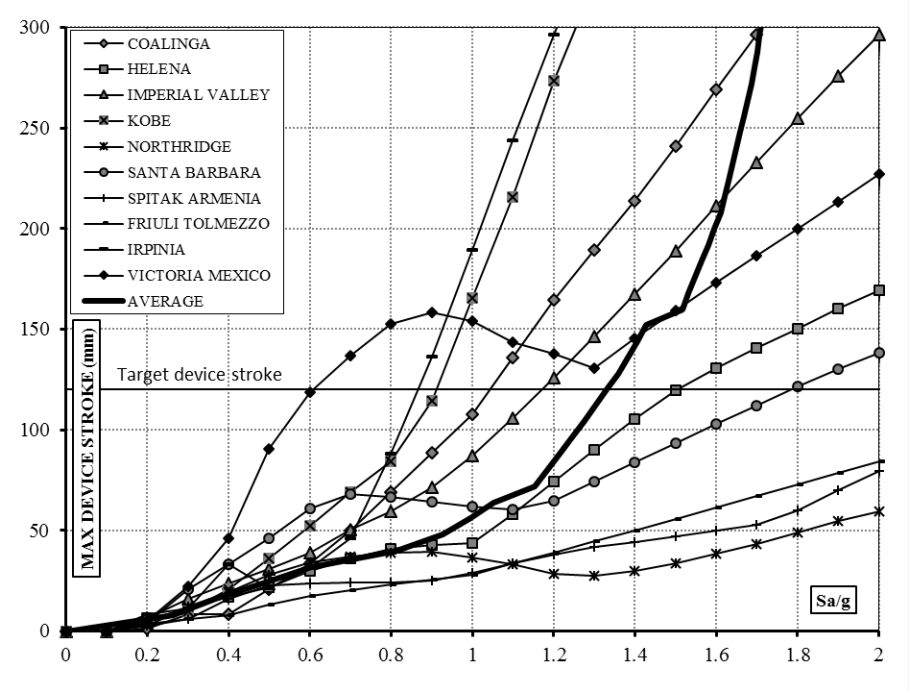



Figure 8: Maximum Device Stroke vs Spectral Acceleration (Sa/g)

Finally, in Figure 7 the maximum device stroke versus the spectral acceleration is reported. In particular, a spectral acceleration of 1.23g is achieved on average also for a target stroke of 0,12 m. This result is in perfect agreement with those deduced by the MIDR graph. This high values of spectral acceleration are very satisfactory and confirms the high dissipative performances of EBFs especially if equipped with dissipative friction devices.

## 5 CONCLUSIONS

- In this work EBFs equipped with frictional devices have been analyzed.
- The control of devices activation and, as a consequence, the optimization of energy dissipation capacity is obtained by a rigorous design procedure assuring that only the dissipative zones are involved in plastic range while non dissipative ones remain elastic called theory of Plastic Mechanism Control.
- Being all the non dissipative zones in elastic range and only frictional devices activated in plastic range the structure can be considered as free from damage. In fact, after a destructive seismic event all the damaged devices can be replaced by new ones.
- In addition, by designing dissipative zones proportionally to the storey shear they are assured to be engaged in yielding as contemporary as possible, benefit optimization of the seismic response.
- The accuracy of the design approach has been investigated by means push-over and nonlinear dynamic analyses by applying a properly chosen set of earthquake ground motions.
- The global mechanism is confirmed to be achieved by both push-over than IDA Analyses.
- The designed structure show very high performances due both to the accuracy of the design procedure than to the exploitation of dissipative capacity of devices.

## REFERENCES


[1] Kasai K, Popov E.P.: "General behavior of WF steel shear link beams", ASCE Journal of Structural Engineering, Vol. 112, Issue 2, pp. 362-382, 1986.

[2] Kasai K., Han X.: "New EBF design method and applications: Redesign and analysis of US-Japan EBF. Proceedings of Stessa 97 – 2nd international conference on behaviour of steel structures in seismi areas, 1997.

[3] Kasai K., Han X.: "Refined Design and Analysis of Eccentrically Braced Frames", Journal of Structural Engineering, ASCE, 1997.

[4] EN 1998-1: "Eurocode 8: Design of Structures for Earthquake Resistance – Part 1: general Rules, Seismic Actions and Rules for Buildings", CEN, 2004

[5] Castaldo P, Palazzo B, Della Vecchia P., "Seismic reliability of base-isolated structures with friction pendulum bearings", *Engineering Structures*, Vol. 95, 80-93, 2015

[6] Castaldo P, Tubaldi E., "Influence of FPS bearing properties on the seismic performance of base-isolated structures", *Earthquake Engineering and Structural Dynamics* 2015; Vol. 44, 2817–2836, 2015.

[7] Castaldo P., "Integrated Seismic Design of Structure and Control Systems", *Springer International Publishing: New York*, 2014. DOI 10.1007/978-3-319-02615-2, 2014.

[8] Castaldo P., De Iuliis M., "Optimal integrated seismic design of structural and viscoelastic bracing-damper systems", *Earthquake Engineering and Structural Dynamics*, Vol. 43(12),1809–1827, 2014.

[9] De Iuliis M., Castaldo P., "An energy-based approach to the seismic control of one-way asymmetrical structural systems using semi-active devices", *Ingegneria Sismica - International Journal of Earthquake Engineering 2012*; XXIX(4):31–42, 2012.

[10] Palazzo B, Castaldo P, Marino I. The Dissipative Column: A New Hysteretic Damper. Buildings 2015;5(1):163-178;

[11] Montuori R., Nastri E., Piluso V. Theory of plastic mechanism control for the seismic design of braced frames equipped with friction dampers - *Research Communications* – Vol. 58. pp. 112-123 2014.

[12] Mazzolani, F.M., Piluso, V., "Plastic Design of Seismic Resistant Steel Frames", Earthquake Engineering and Structural Dynamics, Vol. 26, pp. 167-191, 1997.

[13] Montuori, R., Nastri, E., Piluso, V., "Advances in Theory of Plastic Mechanism Control: Closed Form Solution for Mr-Frames", *Earthquake Engineering and Structural Dynamics*, 2015.

[14] Montuori, R., Nastri, E., Piluso, V., "Theory of plastic mechanism control for MRF-EBF dual systems: Closed Form Solution", Submitted for publication to Engineering Structures, 2016.

[15] Longo, A.; Montuori, R.; Piluso, V. Theory of Plastic Mechanism Control for MRF-CBF dual system and its validation. *Bulletin of Earthquake Engineering* (BEEE). 2014.

[16] Giugliano M.T., Longo A., Montuori R., Piluso V. Failure mode and drift control of MRF-CBF dual systems *The Open Construction and Building Technology Journal*, 4, 121-133, 2010.

[17] Montuori R., Piluso V., Troisi M. Theory of plastic mechanism control of seismic-resistant MR-frames with set-backs. *Open Construction and Building Technology Journal*, vol. 6, pp. 404-413 , 2012.

[18] Longo A., Montuori R., Piluso V., Theory of Plastic Mechanism Control of Dissipative Truss Moment Frames, *Engineering Structures*, Vol. 37, pp. 63-75, 2012.

[19] Longo A., Montuori R., Piluso V., Failure Mode Control and Seismic Response of Dissipative Truss Moment Frames, *Journal of Structural Engineering*, Vol. 138, pp.1388-1397, 2012.

[20] Longo A., Montuori, R., Nastri, E., Piluso, V. On the use of HSS in seismic-resistant structures *Journal of Constructional Steel Research*, Vol. 103, p. 1-12, 2014.

[21] Longo A., Montuori, R., Piluso, V., Moment Frames - Concentrically Braced Frames Dual Systems: Analysis of Different Design Criteria, *Structure and infrastructure engineering*, Volume 12, Issue 1, 2 January 2016, Pages 122-14.

[22] Longo A., Nastri E., Piluso V. Theory of plastic mechanism control: State-of-the-art. Open Construction and Building Technology Journal. Volume 8, 2014, Pages 262-278

[23] Longo A., Montuori R., Piluso V. Failure mode control of X-braced frames under seismic actions. *Journal of Earthquake Engineering*, 12: 728-759, 2008

[24] Longo A, Montuori R, Piluso V. Plastic design of seismic resistant V-Braced frames. *Journal of Earthquake Engineering,* vol. 12: 1246-1266, 2008.

[25] Longo A, Montuori R, Piluso V. "Influence of design Criteria on the seismic Reliability of X-Braced Frames", *Journal of Earthquake Engineering*, Vol. 12, Issue 3– p.406-431. 2008

[26] Montuori, R., Nastri, E., Piluso, V.,"Rigid-plastic analysis and moment–shear interaction for hierarchy criteria of inverted Y EB-Frames" *Journal of Constructional Steel Research* vol. 95 71-80. 2014.

[27] Montuori, R., Nastri, E., Piluso, V., Theory of Plastic Mechanism Control for Eccentrically Braced Frames with inverted Y-scheme, *Journal of Constructional Steel Research*, Volume 92, pp. 122-135, 2014.

[28] Giugliano M.T., Longo A., Montuori R., Piluso V., Plastic design of CB-frames with reduced section solution for bracing members. *Journal of Constructional Steel Research.* Vol. 66 pp 611-621. 2010.

[29] Longo A., Montuori R., Piluso V. Seismic reliability of V-braced frames: Influence of design methodologies. *Earthquake Engineering and Structural Dynamics*, vol. 38 p. 1587-1608, 2009.

[30] Longo A., Montuori R., Piluso V. Seismic Reliability of Chevron Braced Frames with Innovative Conception of Bracing Members, *Advanced Steel Construction, an International Journal* - Vol. 5, No. 4, December 2009.

[31] Montuori, R. , Muscati, R. Plastic design of seismic resistant reinforced concrete frame. Earthquake and Structures Volume 8, Issue 1, 2015, Pages 205-224.

[32] Montuori, R. , Muscati, R. A General Procedure for Failure Mechanism Control of Reinforced Concrete Frames. Submitted for publication to Engineering Structures.

[33] Ohsaki M., Nakajima T.: "Optimization of Link Member of Eccentrically Braced Frames for Maximum Energy Dissipation", Journal of Constructional Steel Research, Vol. 75, 2012, pp. 38-44, 2012.

[34] Latour, M., Piluso, V., Rizzano, G. Free from damage beam-to-column joints: Testing and design of DST connections with friction pads. Engineering Structures. Volume 85, February 05, Pages 219-233, 2015.

[35] Montuori R., Piluso V., Troisi M. Innovative structural details in MR-frames for free from damage structures Mechanics Research Communications – Vol. 58 pp. 146-156. 2014.

[36] Lee S-S., Goel S.C., "Performance-Based Design of Steel Moment Frames using Target Drift and Yield Mechanism", Research Report UMCEE 01-17, University of Michigan, December 2011.

[37] Goel S.C., Chao S-H., "Performance-Based Plastic Design: Earthquake Resistant steel Structures", International Code Council: Washington, DC, 2008.

[38] Latour M., Rizzano G., Piluso V. (2012): "Experimental Analysis of Innovative Dissipative Bolted Double Split Tee Beam-to-Column Connections", *Steel Construction*, 4 (2011), N.2, 53-64.

[39] Latour M., Piluso V., Rizzano G. (2014), "Experimental analysis on friction materials for supplemental damping devices", Construction and Building Materials, Volume 65, 29 August 2014, 159-176.

[40] Della Corte G., D'Aniello M., Landolfo R., (2013). Analytical and numerical study of plastic overstrength of shear links. Journal of Constructional Steel Research, 82, 19–32

[41] Della Corte G., D'Aniello M., Mazzolani F.M., "Inelastic response of shear links with axial restraints: numerical vs. analytical results" 5th International Conference on Advances in Steel Structures, Singapore, 5 – 7 December, 2007.

[42] Mastandrea L., Nastri E., Piluso, V., "Validation of a Design Procedure for Failure Mode Control of EB-Frames: Push-over and IDA analyses", The Open Construction and Building Technology Journal, 7, 193-207, 2013.

[43] Bosco, M., Marino, E.M., Rossi, P.P., "Modelling of steel link beams of short, intermediate or long length", Engineering Structures, Volume 84, February 01, Pages 406-418, 2015.

[44] CSI 2007. SAP 2000: Integrated Finite Element Analysis and Design of Structures. Analysis Reference. Computer and Structure Inc. University of California, Berkeley.

[45] Pacific Earthquake Engineering Research Center, PEER Strong Motion Database, http://peer.berkley.edu.smcat.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EXAMPLE OF COLUMN DESIGN OPTIMIZATION FOR FAILURE MODE CONTROL OF REINFORCED CONCRETE FRAMES

**Roberta Muscati**

Ph. D. student, Department of Civil Engineering
University of Salerno – Italy
e-mail: rmuscati@unisa.it

**Keywords:** Global Mechanism, Concrete Moment Resisting Frames, Plastic Collapse Theory.

**Abstract.** *In this paper an example of column design optimization for the control of collapse mechanism for RC-frames is presented. The applied theory has already been developed, by the authors, in previous work.*
*The procedure is based on the application of the kinematic theorem of the plastic collapse (TPMC: Theory of Plastic Mechanism Control). In particular, the outcome of the theory is the evaluation of the sum of the plastic moments of the columns required, at each storey, to prevent undesired failure modes such as soft-storey mechanism. In the proposed method the second-order effects, due to vertical loads, can play an important role in the seismic design of reinforced concrete frames; they can be taken into account by mean the mechanism equilibrium curve of the considered collapse mechanism.*
*In this work the authors show how the classical design methodology based on the beam-column hierarchy criterion does not allow to obtain a global mechanism.*
*Beam-column hierarchy criterion, commonly suggested by seismic codes, appears only as a very rough approximation when compared to TPMC and its theoretical background.*
*By applying this classic methodology, if a global mechanism is obtained, than the design of the column section and its reinforcement are not optimized. In fact, only with the theory already recalled we can obtain the minimum of section columns able to provide the development of a global mechanism.*
*In addition, in the presented design approach, also the role played by joists can be accounted for. In fact, if their contribution is neglected the collapse mechanism can be very worse that the desired one.*

## 1 INTRODUCTION

One of the most important requirement in the seismic design consists in avoiding collapse mechanism characterized by the development of the minimum energy dissipation capacity of the structure, such as “soft-storey” mechanism. In fact the development of a collapse mechanism of global type, characterized by the location of plastic hinges at all the beam ends and at the base sections of first storey columns is always preferred.

Relatively to the moment resisting frames, the maximum number of plastic hinges is obtained when two plastic hinges develops in each bay and they are usually located at beam ends. At the aim of avoiding undesired collapse mechanisms, hierarchy criterion reported in all the modern seismic codes, suggests that at any joint, the sum of the flexural strength of the columns is greater than the sum of the flexural strength of the beams converging in the same joint [1, 2]. Notwithstanding, the beam-column hierarchy criterion, being based on simple joint equilibrium, is generally able to prevent “soft-storey” mechanisms, but it does not assure the development of a collapse mechanism of global type; in fact it is a non-rigorous application of capacity design principles. In addition, several research are devote in recent years to the understand of seismic collapse mode of reinforced concrete frames, the induced loss and the retrofitting technics to be adopted in order to obtain a better and more dissipative collapse mechanism both in case of existing structures [3-9] and in case of new structures [10-16]. In order to overcame this problem, a more sophisticated design procedure, based on the kinematic theorem of plastic collapse and on second order plastic analysis (i.e. the concept of mechanism equilibrium curve) has been presented. “Theory of Plastic Mechanism Control” (TPMC) has been obtained as a powerful tool for the seismic design. In particular, it consists on the extension of the kinematic theorem of plastic collapse to the concept of mechanism equilibrium curve. In fact, for any given structural typology, the design conditions to be applied in order to prevent undesired collapse mechanisms can be derived by imposing that the mechanism equilibrium curve, corresponding to the global mechanism, has to be located below those corresponding to all the other undesired mechanisms up to a displacement level compatible with the local ductility supply of dissipative zones. This design approach has been applied to different steel structural typologies such as MRFs with RBS connections, EB-Frames, dissipative truss-moment frames, MRF-CBF dual systems and MRFs equipped with friction dampers [17-39]. Starting from the above background, the TPMC is developed also with reference to the reinforced concrete frames. In the present paper, a worked example and a validation of the proposed design procedure are presented in [40].

## 2 THEORY OF PLASTIC MECHANISM CONTROL

In general, three main collapse mechanism typologies that the structure is able to exhibit can be recognized: these mechanisms, depicted in Figure 1, are to be considered undesired because they do not involve all the dissipative zones. In order to apply the TPMC it is of paramount importance the introduction of the concept of linearized mechanism equilibrium curve for each considered mechanism. The mathematical expression of this curve can be written as:

$$\alpha = \alpha_0 - \gamma\delta \tag{1}$$

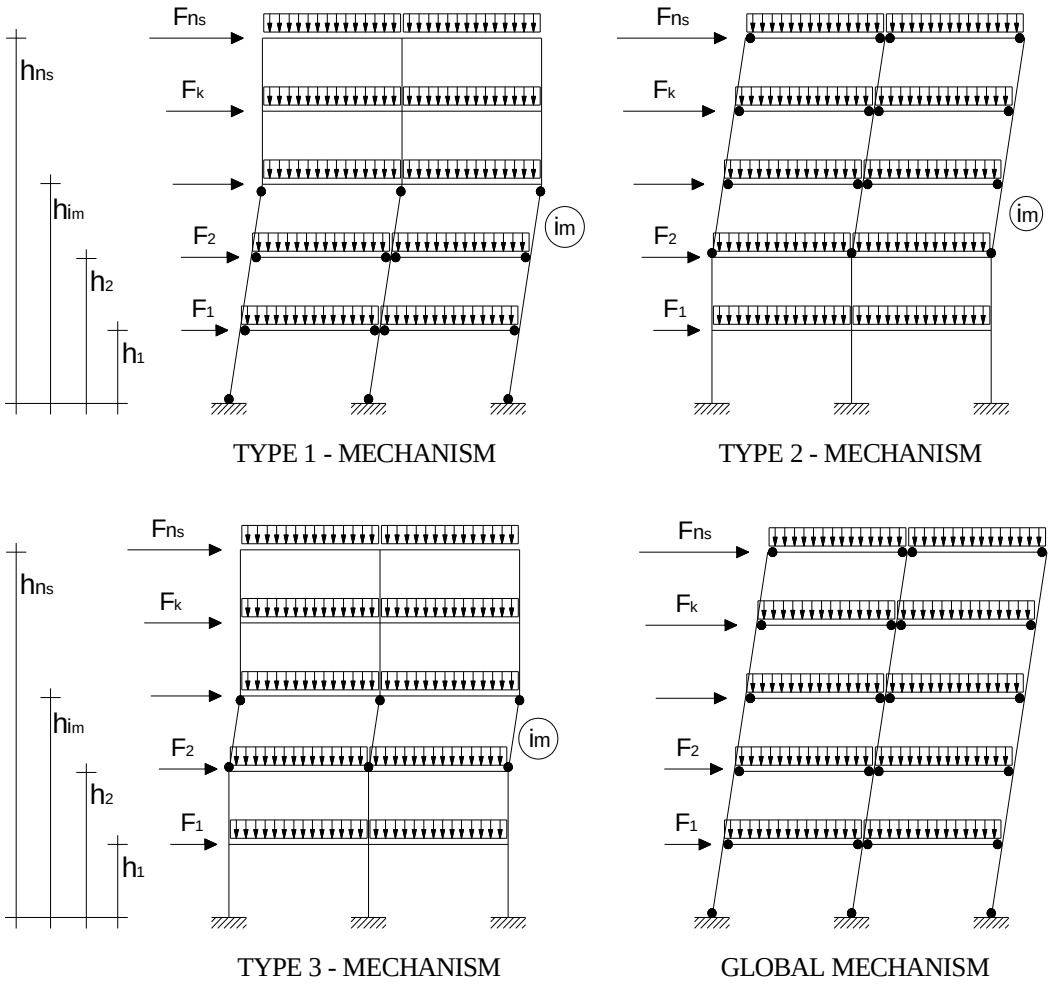



Figure 1: Collapse mechanism typologies

where $\alpha_0$ is the kinematically admissible multiplier of horizontal forces, $\gamma$ is the slope of the mechanism equilibrium curve and $\delta$ is the top-sway displacement. Both parameters can be derived, according to rigid-plastic theory, using the principle of virtual work. Within the framework of a kinematic approach, for any given collapse mechanism, the mechanism equilibrium curve can be easily derived by equalling the external work to the internal work. In addition, in order to account for second-order effects, the external second-order work due to vertical load is also evaluated. For what concerns the theory herein considered and applied, reference is to be made to [40] where the design algorithm and the presentation of the developed theory is reported. In is useful to remember that in the proposed method the beam section properties are assumed to be known quantities because they are designed to resist vertical loads. As a consequence, the unknowns of the design problem are the column sections. They could be determined by means of design conditions expressing that the kinematically admissible multiplier corresponding to the global mechanism is the minimum among all kinematically admissible multipliers corresponding to all other mechanisms (Figure 1). Obviously, this design condition is able to assure the desired collapse mechanism only in case of rigid-plastic behaviour, while actual structures are characterized by elastic displacements before the development of a plastic mechanism. Due to these elastic displacements, second-order effects of vertical loads cannot be neglected. For the better comprehension of the following, the adopted notation is reported to Table 1.

| | | | |
|---|---|---|---|
| $n_c$ | number of columns | $e$ | index of beam ends (e=L=left end, e=R=right end) |
| $n_b$ | number of bays | $M_{c,i,im}$ | plastic moment of the i-th column at $i_m$-th storey |
| $n_s$ | number of storeys | $M_{c,im} = \sum_{i=1}^{n_c} M_{c,i,im}$ | sum of column plastic moments at $i_m$-th storey |
| $i_m$ | index of mechanism | $M_V = \sum_{k=1}^{n_s} V_k h_k$ | second-order work due to vertical loads in global mechanism |
| $H_o$ | sum of the interstorey heights of the storeys involved by the generic mechanism | $M_F = \sum_{k=1}^{n_s} F_k h_k$ | external work due to horizontal forces in the global mechanism |
| $h_k$ | height of the k-th storey (with k=1, 2, .., $n_s$) | $F = \sum_{k=1}^{n_s} F_k$ | sum of the horizontal forces |
| $F_k$ | horizontal force applied to the k-th storey | $M_{b.jk}$ | plastic design resistance of beam at *j*-th bay of the *k*-th storey |
| $V_k$ | sum of all the vertical loads acting at k-th storey | $M_{b,Rd,e} = \sum_{k=1}^{n_s}\sum_{j=1}^{n_b} M_{b.jk}$ | sum of the plastic design resistances of beam ends (for $e$ end) in the global mechanism |

Table 1: Notation

These effects can be taken into account by imposing that the mechanism equilibrium curve corresponding to the global mechanism has to lie below those corresponding to all other mechanisms i.e. the upper bound theorem of plastic design is to be satisfied for each value of the displacements $\delta$ (Figure 2). However, the fulfilment of this requirement is necessary only up to a selected ultimate displacement $\delta_u$, which has to be compatible with the ductility supply of structural members.

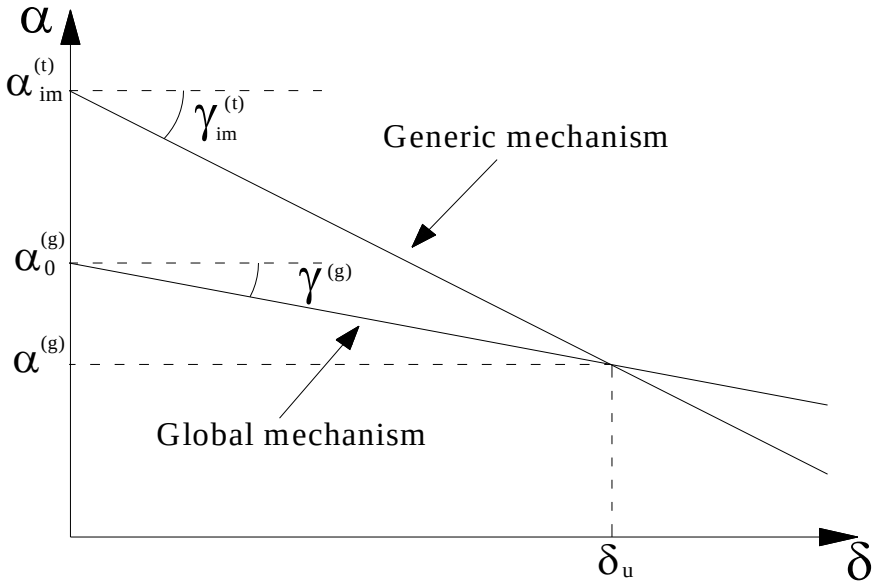



Figure 2: Design condition

This corresponds to impose the following conditions, for LR earthquake:

$$\alpha_{LR}^{(g)} = \alpha_{0,LR}^{(g)} - \gamma^{(g)}\,\delta_u \le \alpha_{0,im,LR}^{(t)} - \gamma_{im}^{(t)}\,\delta_u = \alpha_{im,LR}^{(t)} \tag{2}$$

and for RL earthquake:

$$\alpha_{RL}^{(g)} = \alpha_{0,RL}^{(g)} - \gamma^{(g)}\,\delta_u \le \alpha_{0,im,RL}^{(t)} - \gamma_{im}^{(t)}\,\delta_u = \alpha_{im,RL}^{(t)} \tag{3}$$

for $i_m = 1, 2\,3, \dots, n_s$ and $t = 1, 2, 3$.

Therefore, there are $6n_s$ design conditions to be satisfied for a structural scheme having $n_s$ storeys. With reference to $i_m$-th mechanism of type-1, the kinematically admissible multiplier of seismic horizontal forces is given, for LR earthquake, by:

$$\alpha_{0,i_{m,LR}}^{(1)} = \frac{M_{c,1} + \sum_{k=1}^{i_m-1}\sum_{j=1}^{n_b} M_{b.jk,L}^{+} + \sum_{k=1}^{i_m-1}\sum_{j=1}^{n_b} M_{b.jk,R}^{-} + M_{c,im}}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k} \tag{4}$$

and for RL earthquake by:

$$\alpha_{0,i_{m,RL}}^{(1)} = \frac{M_{c,1} + \sum_{k=1}^{i_m-1}\sum_{j=1}^{n_b} M_{b.jk,L}^{-} + \sum_{k=1}^{i_m-1}\sum_{j=1}^{n_b} M_{b.jk,R}^{+} + M_{c,im}}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k} \tag{5}$$

while the slope of the mechanism equilibrium curve, which is the same for both directions, is:

$$\gamma_{i_m}^{(1)} = \frac{1}{h_{i_m}} \frac{\sum_{k=1}^{i_m} V_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} V_k}{\sum_{k=1}^{i_m} F_k h_k + h_{i_m}\sum_{k=i_m+1}^{n_s} F_k} \tag{6}$$

With reference to $i_m$-th mechanism of type-2 the kinematically admissible multiplier of seismic horizontal forces is given, for LR earthquake, by:

$$\alpha_{0,i_{m,LR}}^{(2)} = \frac{M_{c,im} + \sum_{k=i_m}^{n_s}\sum_{j=1}^{n_b} M_{b.jk,L}^{+} + \sum_{k=i_m}^{n_s}\sum_{j=1}^{n_b} M_{b.jk,R}^{-}}{\sum_{k=i_m}^{n_s} F_k (h_k - h_{i_m-1})} \tag{7}$$

and for RL earthquake by:

$$\alpha_{0,i_{m,RL}}^{(2)} = \frac{M_{c,im} + \sum_{k=i_m}^{n_s}\sum_{j=1}^{n_b} M_{b.jk,L}^{-} + \sum_{k=i_m}^{n_s}\sum_{j=1}^{n_b} M_{b.jk,R}^{+}}{\sum_{k=i_m}^{n_s} F_k (h_k - h_{i_m-1})} \tag{8}$$

while the slope of the mechanism equilibrium curve is:

$$\gamma_{i_m}^{(2)} = \frac{1}{h_{n_s} - h_{i_m-1}} \frac{\sum_{k=i_m}^{n_s} V_k \left(h_k - h_{i_m-1}\right)}{\sum_{k=i_m}^{n_s} F_k (h_k - h_{i_m-1})} \tag{9}$$

Finally, with reference to $i_m$-th mechanism of type-3, the kinematically admissible multiplier of horizontal forces, is given by:

$$\alpha_{0,i_m}^{(3)} = \frac{2M_{c,im}}{\left(h_{i_m} - h_{i_m-1}\right)\sum_{k=i_m}^{n_s} F_k} \tag{10}$$

In this case the expression is the same for both directions of earthquake because the beams are not involved in this collapse mechanism. In addition, the corresponding slope of the mechanism equilibrium curve is given by:

$$\gamma_{\mathrm{i_m}}^{(3)} = \frac{\sum_{\mathrm{k=i_m}}^{\mathrm{n_s}} \mathrm{V_k}}{\left(\mathrm{h_{i_m}} - \mathrm{h_{i_m-1}}\right)\sum_{\mathrm{k=i_m}}^{\mathrm{n_s}} \mathrm{F_k}} \tag{11}$$

It is important to underline that, for any given geometry of the structural system, the slope of mechanism equilibrium curve attains its minimum value when the global type mechanism

is developed. In fact, it is easy to check that $\gamma^{(g)}$, which is equal to $\gamma_1^{(2)}$, is always the minimum value among all the $\gamma_{i_m}^{(t)}$. This issue assumes a paramount importance in TPMC allowing the extension of the kinematic theorem of plastic collapse to the concept of mechanism equilibrium curve by simply checking relations (2) and (3) for the value $\delta = \delta_u$, as depicted in Figure 2.

## 3 STUDY CASE

In order to show the practical application of the proposed design procedure, the seismic design of a four-bay six-storey moment resisting frame is presented in this section.

The inelastic behaviour of the designed structure is successively examined by means of a push-over static, confirming the fulfilment of the design goal, i.e. the location of the yielding zones at the beam ends with the only exception of the base section of first-storey columns. The structural scheme of the frame to be designed is shown in Figure 3. The interstorey height is equal to 3m. The characteristic values of the vertical loads acting on the beams are equal to 21.89 kN/m and 8 kN/m for permanent ($G_k$) and live ($Q_k$) actions, respectively. The structural materials adopted are concrete C25/30 and reinforcement of steel grade B450C.

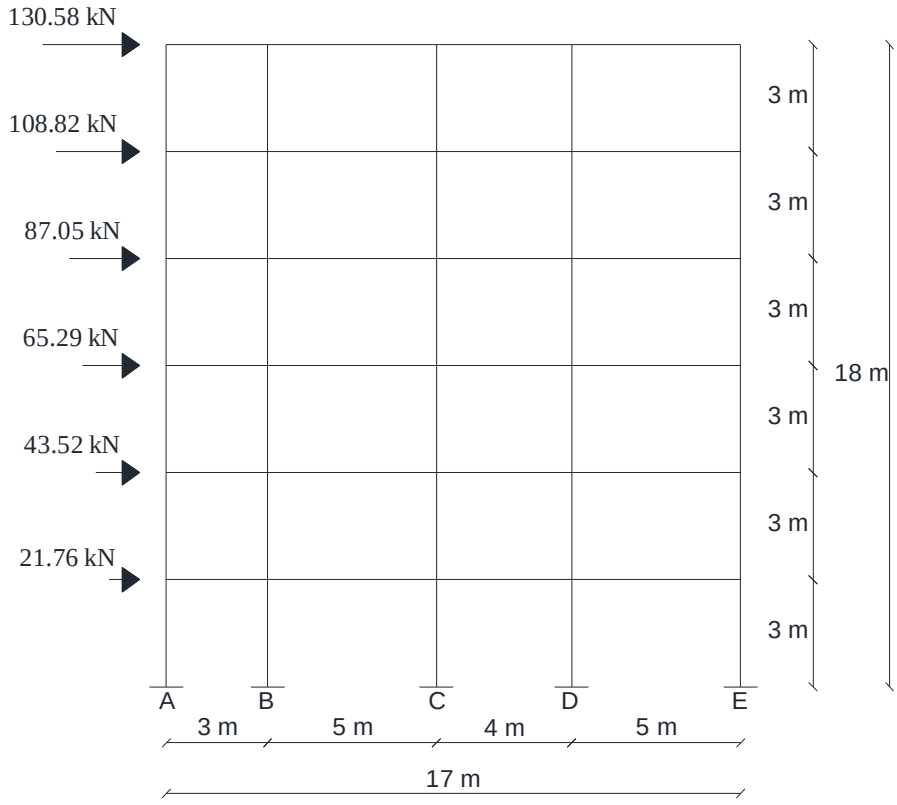



Figure 3: Structural scheme of the designed frame

According to Eurocode 8, the value of the period of vibration to be used for preliminary design is:

$$T = 0.075\, H^{3/4} = 0.075 \cdot 18^{3/4} \approx 0.65\, s \tag{12}$$

where $H$ is the total height of the frame. With reference to the design spectrum for stiff soil conditions (soil class A of Eurocode 8) and by assuming a behaviour factor q equal to 3.9, the horizontal seismic forces are those depicted in Figure 3. In the following, the numerical development of the design steps for the structural scheme described above is provided.

*a) Selection of the design top sway displacement*

The selection of the maximum top sway displacement up to which the global mechanism has to be assured is a very important design issue, because the value of this displacement governs the magnitude of second order effects accounted for in the design procedure. A good criterion to choose the design ultimate displacement $\delta_u$ is to relate it to the plastic rotation supply of beams or beam-to-column connections by assuming $\delta_u = \theta_u \cdot h_{ns}$ (where $\theta_u$ can be assumed equal to 0.01 rad). As a consequence, the design value of the top sway displacement has been assumed equal to:

$$\delta_u = 0.01 \cdot h_{ns} = 0.01 \cdot 18 = 0.18\, m \tag{13}$$

*b) Design of beam sections to withstand vertical loads.*

The load acting on the frame in the vertical load combination is:

$$Q_{SLU} = 1.3\, G_k + 1.5\, Q_k = 40.46\, kN/m \tag{14}$$

For the design of the beams has been considered a bending moment equal to:

$$M_{max} = \frac{Q_{SLU} \cdot L^2}{8} \tag{15}$$

Therefore, by imposing the base of the section equal to b=30 cm, is possible to calculate the height of the beam through the following design relation:

$$d = r \sqrt{\frac{M_{Sd}}{b}} \tag{16}$$

Assuming $\xi = 0.25$ and $\rho = 0.25$ a value of $r = 0.19$ is obtained. As a consequence the amount of reinforcement is given by:

$$A_s = \frac{M_{Sd}}{0.85 \cdot h \cdot f_{sd}} \tag{17}$$

Obviously the number of steel bars in the beam is such that:

$$M_{Rd} > M_{Sd} \tag{18}$$

where $M_{Rd}$ is the design value of the resistant moment and $M_{Sd}$ is the design value of the applied internal bending moment. The reinforcement at the beam ends are reported in Table 2.

| *PART OF SECTION* | $L_{AB}$ = 3m | | $L_{BC}$ = 5m | | $L_{CD}$ = 4m | | $L_{DE}$ = 5m | |
|---|---|---|---|---|---|---|---|---|
| | *e = L* | *e = R* | *e = L* | *e = R* | *e = L* | *e = R* | *e = L* | *e = R* |
| *Top* | *3F20* | *3F20* | *4F20* | *4F20* | *4F20* | *4F20* | *4F20* | *4F20* |
| *Bottom* | *3F20* | *3F20* | *4F20* | *4F20* | *3F20* | *3F20* | *4F20* | *4F20* |

Table 2: Reinforcement at the beam ends (L = left and R = right) for the first storey

*c) Computation of the slopes of mechanism equilibrium curve* $\gamma_{i_m}^{(t)}$.

By means of Eqs. (6), (9) and (11) the slopes of mechanism equilibrium curves are computed. These values are reported in Table 3 and they are the same for both directions of seismic input.

| STOREY $i_m$ | $\gamma_{i_m}^{(1)}$ | $\gamma_{i_m}^{(2)}$ | $\gamma_{i_m}^{(3)}$ |
|---|---|---|---|
| 1 | 0.0181 | 0.0024 | 0.0181 |
| 2 | 0.0085 | 0.0027 | 0.0158 |
| 3 | 0.0054 | 0.0032 | 0.0141 |
| 4 | 0.0038 | 0.0040 | 0.0127 |
| 5 | 0.0030 | 0.0056 | 0.0116 |
| 6 | 0.0024 | 0.0150 | 0.0105 |

Table 3: Slopes of mechanism equilibrium curves ($cm^{-1}$)

In particular it is important to underline that the slope value corresponding to the global mechanism $\gamma^{(g)} = \gamma_1^{(2)}$, is the minimum among all the $\gamma_{i_m}^{(t)}$ values:

$$\gamma^{(g)} = 0.002433\ cm^{-1} \tag{19}$$

*d) Computation of the required sum of plastic moments of columns at first storey $M_{c,1}$.*

The required sums of plastic moments of columns at first storey for LR and RL earthquake are equal to $M_{c,1,LR} = 1177.99$ kNm and $M_{c,1,RL} = 1177.99$ kNm, respectively.

*e) Distribution among the columns proportionally to their number.*

According to the global mechanism, axial forces in the columns at collapse state depend both from the distributed loads acting on the beams and from the shear action due to the development of plastic hinges at the beam ends, as depicted in Figure 4 (with reference to the earthquake from Left to Right). So that, the total load transmitted by the beams to the columns is the sum of two contributions. The first one, $N_q$, is related to the vertical loads acting in the seismic load combination (i.e. the sum of $ql/2$ type contributions).

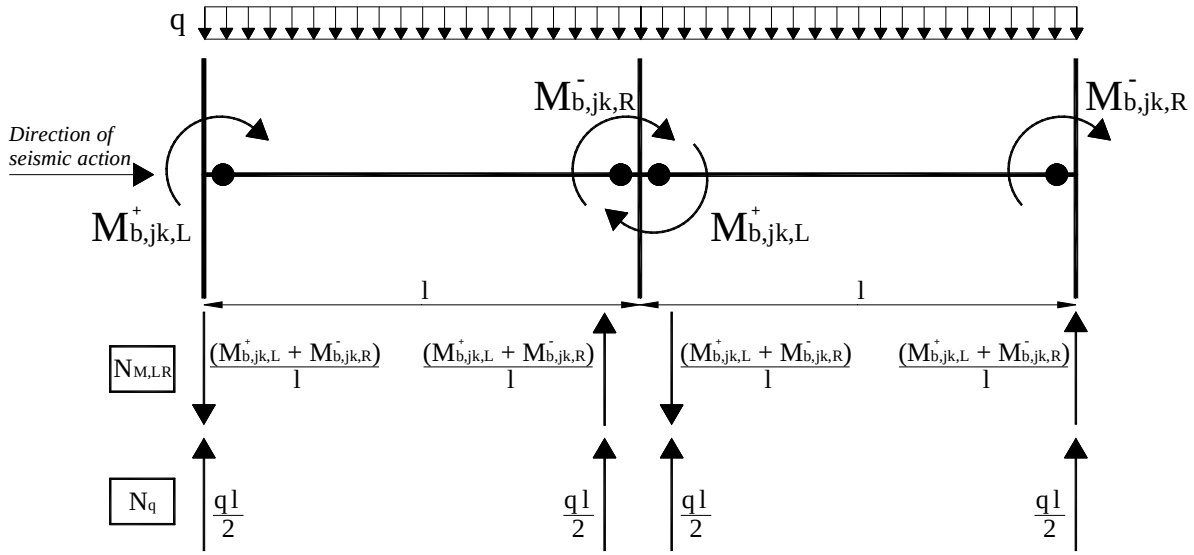



Figure 4: Loads transmitted by the beams to the columns at collapse state for LR earthquake

In Table 4 the axial forces due to vertical loads, for both directions of earthquake, are reported for each storey and for each column.

| Storey | Column A | Column B | Column C | Column D | Column E |
|---|---|---|---|---|---|
| $i_m$ | $N_q$ (kN) | $N_q$ (kN) | $N_q$ (kN) | $N_q$ (kN) | $N_q$ (kN) |
| 1 | 218.70 | 583.20 | 656.10 | 656.10 | 364.50 |
| 2 | 182.25 | 486.00 | 546.75 | 546.75 | 303.75 |
| 3 | 145.80 | 388.80 | 437.40 | 437.40 | 243.00 |
| 4 | 109.35 | 291.60 | 328.05 | 328.05 | 182.25 |
| 5 | 72.90 | 194.40 | 218.70 | 218.70 | 121.50 |
| 6 | 36.45 | 97.20 | 109.35 | 109.35 | 60.75 |

Table 4: Axial forces acting in the columns related to the vertical loads for both directions of earthquake

The second one, $N_{M,LR}$ (or $N_{M,RL}$), is related to the shear actions due to the plastic hinges developed at the beam ends i.e. the sum of $(M_{b.jk,L}^{+} + M_{b.jk,R}^{-})/l$ for earthquake from left to right (or the sum of $(M_{b.jk,L}^{-} + M_{b.jk,R}^{+})/l$ for earthquake from right to left). In Table 5 these contributions are reported.

| Storey | Column A | | Column B | | Column C | | Column D | | Column E | |
|---|---|---|---|---|---|---|---|---|---|---|
| $i_m$ | $N_{M,LR}$ (kN) | $N_{M,RL}$ (kN) | $N_{M,LR}$ (kN) | $N_{M,RL}$ (kN) | $N_{M,LR}$ (kN) | $N_{M,RL}$ (kN) | $N_{M,LR}$ (kN) | $N_{M,RL}$ (kN) | $N_{M,LR}$ (kN) | $N_{M,RL}$ (kN) |
| 1 | -508.96 | 508.96 | 103.42 | -103.42 | -38.47 | 38.47 | 38.47 | -38.47 | 405.53 | -405.53 |
| 2 | -424.13 | 424.13 | 86.18 | -86.18 | -32.06 | 32.06 | 32.06 | -32.06 | 337.94 | -337.94 |
| 3 | -339.30 | 339.30 | 68.95 | -68.95 | -25.65 | 25.65 | 25.65 | -25.65 | 270.36 | -270.36 |
| 4 | -254.48 | 254.48 | 51.71 | -51.71 | -19.24 | 19.24 | 19.24 | -19.24 | 202.77 | -202.77 |
| 5 | -169.65 | 169.65 | 34.47 | -34.47 | -12.82 | 12.82 | 12.82 | -12.82 | 135.18 | -135.18 |
| 6 | -84.83 | 84.83 | 17.24 | -17.24 | -6.41 | 6.41 | 6.41 | -6.41 | 67.59 | -67.59 |

Table 5: Axial forces acting in the columns related to the shear actions for both directions of earthquake

Therefore the required bending moment for each column $M_{c,i,1}$, the section, the upper and lower reinforcement and the axial force, for both directions of the earthquake are reported in Table 6.

| STOREY | Column | $M_{c,i,1,LR}$ [kNm] | $M_{c,i,1,RL}$ [kNm] | b x h | $A_s = A'_s$ | $N_{LR}$ [kN] | $N_{RL}$ [kN] |
|---|---|---|---|---|---|---|---|
| 1° | A | 235.59 | 235.59 | 30x40 | 7 Φ 20 | -290.26 | 727.66 |
| | B | | | 30x50 | 5 Φ 16 | 686.62 | 479.78 |
| | C | | | 30x60 | 5 Φ 16 | 617.63 | 694.57 |
| | D | | | 30x60 | 5 Φ 16 | 694.57 | 617.63 |
| | E | | | 30x40 | 6 Φ 20 | 770.03 | -41.03 |

Table 6: Final design of the column sections at first storey

The sums of obtained column plastic moments at first storey are: $M_{c,Rd,1,LR} = 1585.19\ kNm$ for LR earthquake and $M_{c,Rd,1,RL} = 1603.24\ kNm$ for RL earthquake which are greater than the required one. As a consequence the value of $\alpha_0^{(g)}$ for LR and RL earthquake are equal to $\alpha_{0,LR}^{(g)} = 1.5179$ and $\alpha_{0,RL}^{(g)} = 1.5227$, respectively.

*f) Computation of the required sum of plastic moments of columns $M_{c,im}^{(t)}$ at any storey to avoid undesired mechanism.*

The sum of the plastic moments of columns governing the column design at each storey is given in Table 7 and Table 8 by the underlined values. It can be recognized that, in the examined case, the need to avoid type-1 mechanism always governs the design of columns.
So for earthquake from Left to Right:

| STOREY $i_m$ | $M_{c,im;LR}^{(1)}$ [kNm] | $M_{c,im,LR}^{(2)}$ [kNm] | $M_{c,im,LR}^{(3)}$ [kNm] |
|---|---|---|---|
| 1 | <u>1585.19</u> | - | 1585.19 |
| 2 | <u>1509.60</u> | 770.57 | 1140.09 |
| 3 | <u>1963.60</u> | 51.38 | 1007.49 |
| 4 | <u>2131.34</u> | -476.97 | 827.18 |
| 5 | <u>1917.40</u> | -719.07 | 599.16 |
| 6 | <u>1226.37</u> | -579.49 | 323.43 |

Table 7: Sum of plastic moments of column required at each storey to avoid undesired mechanism

And for earthquake from Right to Left:

| STOREY $i_m$ | $M_{c,im,RL}^{(1)}$ [kNm] | $M_{c,im,RL}^{(2)}$ [kNm] | $M_{c,im,RL}^{(3)}$ [kNm] |
|---|---|---|---|
| 1 | <u>1603.24</u> | - | 1603.24 |
| 2 | <u>1499.68</u> | 784.46 | 1142.07 |
| 3 | <u>1957.25</u> | 61.29 | 1009.27 |
| 4 | <u>2127.97</u> | -470.62 | 828.67 |
| 5 | <u>1916.21</u> | -715.70 | 600.25 |
| 6 | <u>1226.37</u> | -578.30 | 324.03 |

Table 8: Sum of plastic moments of column required at each storey to avoid undesired mechanism

*g) Design of column sections at each storey.*

The required sum of column plastic moments $M_{c,i,i_m}$, the section, the upper and lower reinforcement, the axial force for both directions of the earthquake are reported in Table 9.

| STOREY | Column | $M_{c,im,LR}$ [kNm] | $M_{c,im,RL}$ [kNm] | b x h | $A_s = A'_s$ | $N_{LR}$ [kN] | $N_{RL}$ [kN] |
|---|---|---|---|---|---|---|---|
| 2° | A | 301.92 | 299.93 | 30x50 | 7 Φ 20 | -241.88 | 606.38 |
| | B | | | 30x40 | 6 Φ 20 | 572.18 | 399.82 |
| | C | | | 30x50 | 6 Φ 16 | 514.69 | 578.81 |
| | D | | | 30x50 | 6 Φ 16 | 578.81 | 514.69 |
| | E | | | 30x50 | 6 Φ 20 | 641.69 | -34.19 |
| 3° | A | 392.72 | 391.45 | 30x50 | 6 Φ 24 | -193.50 | 485.10 |
| | B | | | 30x50 | 6 Φ 20 | 457.75 | 319.85 |
| | C | | | 30x50 | 6 Φ 20 | 411.75 | 463.05 |
| | D | | | 30x50 | 6 Φ 20 | 463.05 | 411.75 |
| | E | | | 30x50 | 6 Φ 24 | 513.36 | -27.36 |
| 4° | A | 426.26 | 425.59 | 30x50 | 6 Φ 24 | -145.13 | 363.83 |
| | B | | | 30x50 | 7 Φ 20 | 343.31 | 239.89 |
| | C | | | 30x50 | 7 Φ 20 | 308.81 | 347.29 |
| | D | | | 30x50 | 7 Φ 20 | 347.29 | 308.81 |
| | E | | | 30x50 | 6 Φ 24 | 385.02 | -20.52 |
| 5° | A | 383.48 | 383.24 | 30x50 | 6 Φ 24 | -96.75 | 242.55 |
| | B | | | 30x50 | 7 Φ 20 | 228.87 | 159.93 |
| | C | | | 30x50 | 7 Φ 20 | 205.88 | 231.52 |
| | D | | | 30x50 | 7 Φ 20 | 231.52 | 205.88 |
| | E | | | 30x50 | 5 Φ 24 | 256.68 | -13.68 |
| 6° | A | 245.27 | 245.27 | 30x40 | 7 Φ 20 | -48.38 | 121.28 |
| | B | | | 30x40 | 6 Φ 20 | 114.44 | 79.96 |
| | C | | | 30x40 | 6 Φ 20 | 102.94 | 115.76 |
| | D | | | 30x40 | 6 Φ 20 | 115.76 | 102.94 |
| | E | | | 30x40 | 6 Φ 20 | 128.34 | -6.84 |

Table 9: Design of column sections at each storey

*h) Checking of technological condition*

By observing Table 6 and Table 9 it can be noted that there are some column sections at the first storey which are smaller than the corresponding ones required at the second storey, therefore, a technological condition at the first storey is not satisfied. As a consequence, the values of $M_{c,Rd,1,LR}$ and $M_{c,Rd,1,RL}$ need to be updated and the procedure needs to be repeated from the step e). In Table 10 and Table 11 the new value of required sum of plastic moments of columns $M_{c,im}^{(t)}$ at any storey are reported for both directions of earthquake.

## 4 VALIDATION OF THE DESIGN PROCEDURE

In order to validate the design procedure, a static non-linear analysis (push-over) has been carried out to investigate the actual seismic response of the designed frame by means SAP2000 computer program [41]. This analysis has the primary aim to confirm the development of the desired collapse mechanism typology and to evaluate the obtained energy dissipation capacity, testing the accuracy of the proposed design methodology. Regarding the structural modelling, the mechanical non-linearities, have been concentrated at beam and column ends by means of plastic hinge elements.

| STOREY $i_m$ | $M_{c,im}^{(1)}$ [kNm] | $M_{c,im}^{(2)}$ [kNm] | $M_{c,im}^{(3)}$ [kNm] |
|---|---|---|---|
| 1 | 1660.80 | - | 1660.80 |
| 2 | 1468.05 | 828.74 | 1148.39 |
| 3 | 1937.01 | 92.92 | 1014.96 |
| 4 | 2117.21 | -450.38 | 833.41 |
| 5 | 1912.42 | -704.95 | 603.73 |
| 6 | 1226.37 | -574.51 | 325.93 |

Table 10: Sum of plastic moments of column required at each storey for LR earthquake

and for the other direction of earthquake:

| STOREY $i_m$ | $M_{c,im}^{(1)}$ [kNm] | $M_{c,im}^{(2)}$ [kNm] | $M_{c,im}^{(3)}$ [kNm] |
|---|---|---|---|
| 1 | 1689.21 | - | 1689.21 |
| 2 | 1452.45 | 850.59 | 1151.52 |
| 3 | 1927.02 | 108.53 | 1017.77 |
| 4 | 2111.91 | -440.39 | 835.75 |
| 5 | 1910.54 | -699.64 | 605.45 |
| 6 | 1226.37 | -572.63 | 326.86 |

Table 11: Sum of plastic moments of column required at each storey for RL earthquake

In Table 12 are reported the final value of the columns.

| STOREY | Column | $M_{c,im,LR}$ [kNm] | $M_{c,im,RL}$ [kNm] | b x h | $A_s = A'_s$ | $N_{LR}$ [kN] | $N_{RL}$ [kN] |
|---|---|---|---|---|---|---|---|
| 1° | A | 235.59 | 235.59 | 30x50 | 6 Φ 20 | -290.26 | 727.66 |
| | B | | | 30x50 | 5 Φ 16 | 686.62 | 479.78 |
| | C | | | 30x60 | 5 Φ 16 | 617.63 | 694.57 |
| | D | | | 30x60 | 5 Φ 16 | 694.57 | 617.63 |
| | E | | | 30x50 | 5 Φ 20 | 770.03 | -41.03 |
| 2° | A | 293.61 | 290.49 | 30x50 | 7 Φ 20 | -241.88 | 606.38 |
| | B | | | 30x50 | 7 Φ 16 | 572.18 | 399.82 |
| | C | | | 30x50 | 6 Φ 16 | 514.69 | 578.81 |
| | D | | | 30x50 | 6 Φ 16 | 578.81 | 514.69 |
| | E | | | 30x50 | 6 Φ 20 | 641.69 | -34.19 |
| 3° | A | 387.40 | 385.40 | 30x50 | 6 Φ 24 | -193.50 | 485.10 |
| | B | | | 30x50 | 6 Φ 20 | 457.75 | 319.85 |
| | C | | | 30x50 | 6 Φ 20 | 411.75 | 463.05 |
| | D | | | 30x50 | 6 Φ 20 | 463.05 | 411.75 |
| | E | | | 30x50 | 6 Φ 24 | 513.36 | -27.36 |
| 4° | A | 423.44 | 422.38 | 30x50 | 6 Φ 24 | -145.13 | 363.83 |
| | B | | | 30x50 | 7 Φ 20 | 343.31 | 239.89 |
| | C | | | 30x50 | 7 Φ 20 | 308.81 | 347.29 |
| | D | | | 30x50 | 7 Φ 20 | 347.29 | 308.81 |
| | E | | | 30x50 | 6 Φ 24 | 385.02 | -20.52 |
| 5° | A | 382.48 | 382.10 | 30x50 | 6 Φ 24 | -96.75 | 242.55 |
| | B | | | 30x50 | 7 Φ 20 | 228.87 | 159.93 |
| | C | | | 30x50 | 7 Φ 20 | 205.88 | 231.52 |
| | D | | | 30x50 | 7 Φ 20 | 231.52 | 205.88 |
| | E | | | 30x50 | 5 Φ 24 | 256.68 | -13.68 |
| 6° | A | 245.27 | 245.27 | 30x40 | 7 Φ 20 | -48.38 | 121.28 |
| | B | | | 30x40 | 6 Φ 20 | 114.44 | 79.96 |
| | C | | | 30x40 | 6 Φ 20 | 102.94 | 115.76 |
| | D | | | 30x40 | 6 Φ 20 | 115.76 | 102.94 |
| | E | | | 30x40 | 6 Φ 20 | 128.34 | -6.84 |

Table 12: Design of column sections at each storey for earthquake

The constitutive law of such plastic hinge elements is provided by a rigid plastic moment-rotation curve. The type of hinge depends on the element considered i.e. by its internal action. In fact, for the beams and the columns M3 and P-M3 hinge type have been considered, respectively. In case of P-M3 hinge type, the interaction domain P-M has been evaluated for each column and used in SAP2000 computer program. The results of the push-over analysis are mainly constituted by base shear – top sway displacement curve which is depicted in Figure 5. In the same figure also a straight line is given, i.e. the one corresponding to the linearized mechanism equilibrium curve of global mechanism whose expression, for the designed frame and for earthquake from Left to Right, is:

$$\alpha_{LR}^{(g)} = 1.5179 - 0.002433\,\delta \tag{20}$$

For earthquake to Right to Left, the expression is:

$$\alpha_{RL}^{(g)} = 1.5227 - 0.002433\,\delta \tag{21}$$

As already underlined there is a mechanism equilibrium curve for both direction of earthquake. The two mechanism equilibrium curves are different but only for what concern the $\alpha_0$ value while the slope is the same as reported in Figure 5.

As it was expected, also the LR push-over curve is different from RL one. This difference can be easily understood if we consider that the axial forces in the columns are different in the two considered push-over and, as a consequence, also the plastic moment is different. Notwithstanding, in this case, the two curves are very close one each other but there is no proof of the fact that this represent a general result. So that both curves should be always considered when a non-symmetric moment-resisting frame is analized. Obviously, the base shear is, in this case, obtained by multiplying the value of $\alpha$, given by Eq. (20) and Eq. (21), for the design base shear corresponding to $\alpha = 1$. The comparison between the push-over curve and the global mechanism equilibrium curve provides a first confirmation of the accuracy of the proposed design procedure. In fact, the last branches of push-over curves parallel to global mechanism equilibrium curves as showed in Figure 5.

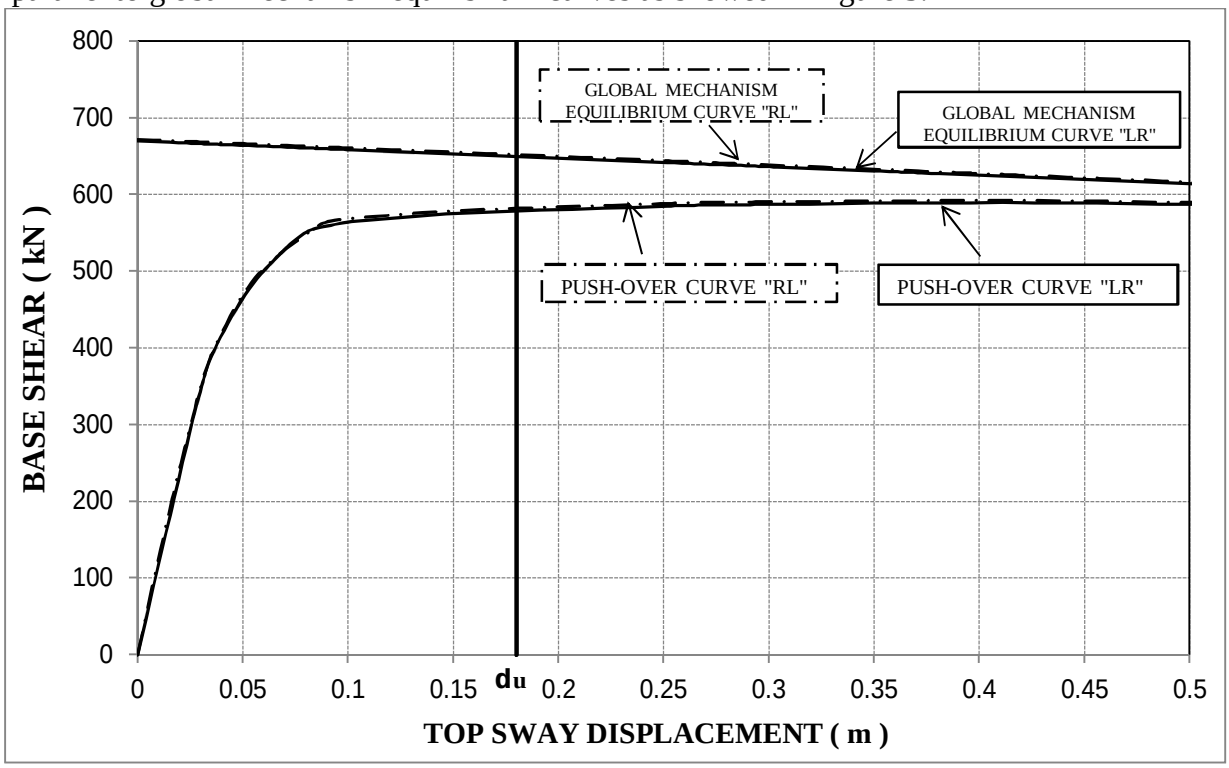



Figure 5: Overlap of the push-over curve with the global mechanism equilibrium curve

A further confirmation, even the most important, of the fulfilment of the design objective is represented by the pattern of yielding developed at the occurrence of the design ultimate displacement. In fact, developed plastic hinges are shown in Figure 6 and their pattern is in perfect agreement with the global mechanism. In order to fulfill the serviceability requirements the interstorey drift have been checked with reference to the limit reported in the Eurocode 8.

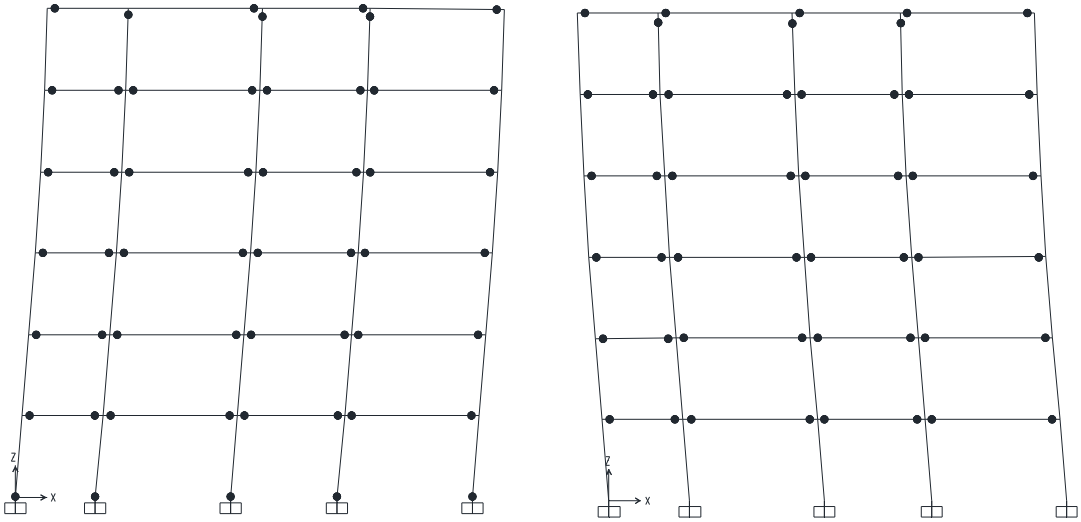


Figure 6: Pattern of yielding of the designed frame at $\delta = \delta_u$ for LR and RL earthquake direction

In particular the considered limit refers to buildings having non structural elements of brittle materials attached to the structure:

$$d_r \, \nu \leq 0.005 \, \mathrm{h} \tag{22}$$

where $d_r$ is the the maximum relative displacement between two consecutive storeys and $h$ is the correspondinbg stprey height. If this serviceability requirement is not verified the structural stiffness can be improved by increasing the beam sections or the ultimate design displacement. In fact, in both cases the final results will be a more rigid structure with respect to the one obtained in the worked example herein presented. In Table 13 the final results are reported.

| STOREY | $d_s$ [cm] | $d_r$ [cm] | $\nu$ | $d_r \, \nu$ | 0.005 h |
|---|---|---|---|---|---|
| 6° | 1.7687 | 0.7173 | 0.5 | 0.3586 | 1.5 |
| 5° | 1.5848 | 1.0963 | | 0.5481 | 1.5 |
| 4° | 1.3037 | 1.4030 | | 0.7015 | 1.5 |
| 3° | 0.9439 | 1.5538 | | 0.7769 | 1.5 |
| 2° | 0.5455 | 1.4072 | | 0.7036 | 1.5 |
| 1° | 0.1847 | 0.7203 | | 0.3601 | 1.5 |

Table 13: Limitation of interstorey drift

## 5 CONCLUSIONS

In this work a methodology able to control the failure mechanism of reinforced concrete moment resisting frames has been applied. On the base of the extension of the kinematic theorem of plastic collapse to the concept of mechanism equilibrium curve, the Theory of Plastic Mechanism Control allows to evaluate the sum of plastic moments of the columns required at each storey in order to develope a collapse mechanism of global type.

The closed form solution of the design conditions makes the design procedure very easy to be applied. In fact, in the author opinion it could also be suggested for code purpose by definitely solving the problem of collapse mechanism control whose importance in seismic design is universally recognised. Beam-column hierarchy criterion, commonly suggested by seismic codes, appears only as a very rough approximation when compared to TPMC and its theoretical background. The reliability of the proposed design procedure has been also demonstrated through its application to a four-bays, six-storeys frame, leading to the fulfilment of the design objective, i.e. the development of a collapse mechanism of global type, as it has been confirmed by the results of both push-over analysis.
The proposed methodology can be considered as belonging to the Performance Based Seismic Design philosophy [42]. In fact, in order to satisfy the limit states of "Life Safe" or "Near Collapse" the designer has to promote a dissipative collapse mechanism avoiding the so called "soft storey mechanism". In addition, it is useful to underline that the proposed procedure constitutes a rigorous application of the capacity design principles. In fact, beams are designed in order to bear external loads, while columns are designed according to the maximum internal actions transmitted by the dissipative zones. It is important to underline that the proposed procedure can be applied for MRFs characterized by a non-symmetric scheme and a non-symmetric reinforcement in each beam section. In the future development, the procedure should account also for the joists influence. In fact resistance of the joists is always neglected, but, as shown in [43] it can be really significative.

## REFERENCES

[1] EN 1998-1, *Eurocode 8: Design of Structures for Earthquake Resistance – Part 1: general Rules, Seismic Actions and Rules for Buildings*, CEN, 2004.

[2] New Zealand *Standard Code of Practice for the Design of Concrete Structures*, NZS 3101: Part 1; Commentary NZS 3101: Part 2; Standard Association of New Zealand, Wellington, New Zealand, 1982.

[3] R. Montuori, V. Piluso , A. Tisi, *Comparative analysis and criticalities of the main constitutive laws of concrete elements confined with FRP*. Compos Part B. 2012; 43: 3219–3230. Doi: 10.1016/j.compositesb.2012.04.001.

[4] R. Montuori, V. Piluso, A. Tisi, *Ultimate behaviour of FRP wrapped sections under axial force and bending: Influence of stress–strain confinement model*. Compos Part B 2013; 54: 85–96. Doi: http://dx.doi.org/10.1016/j.compositesb.2013.04.059.

[5] R. Montuori, V. Piluso, *Reinforced concrete columns strengthened with angles and battens subjected to eccentric load*. Eng Struct 2009; 31:539–50. Doi: 10.1016/j.engstruct.2008.10.005.

[6] M. D'Aniello, G. Della Corte, F. M. Mazzolani, *Seismic upgrading of RC buildings by steel eccentric braces: Experimental results vs numerical modeling* - Proceedings of the 5th International Conference on Behaviour of Steel Structures in Seismic Areas - Stessa 200, Pages 809-814 Yokohama; Japan; 14- 17, August 2006.

[7] M. D'Aniello, G. Della Corte, F. M. Mazzolani, *Seismic upgrading of RC buildings by buckling restrained braces: Experimental results vs numerical modeling* - Proceedings of the 5th International Conference on Behaviour of Steel Structures in Seismic Areas - Stessa 200, Pages 809-814 Yokohama; Japan; 14- 17, August 2006.

[8] G. Della Corte, M. D'Aniello, R. Landolfo, *Field testing of all-steel buckling-restrained braces applied to a damaged reinforced concrete building* - Journal of Structural Engineering (United States), Volume 141, Issue 1, 1 January 2015.

[9] E. S. Mistakidis, G. De Matteis, A. Formisano, *Low yield metal shear panels as an alternative for the seismic upgrading of concrete structures* - Advances in Engineering Software, Volume 38, Issue 8-9, Pages 626-636, August 2007.

[10] P. Castaldo, *Integrated Seismic Design of Structure and Control Systems.* Springer International Publishing: New York, 2014. Doi: 10.1007/978-3-319-02615-2.

[11] P. Castaldo, E. Tubaldi, *Influence of FPS bearing properties on the seismic performance of base-isolated structures.* Earthquake Engineering and Structural Dynamics 2015; 44 (15): 2817–2836.

[12] P. Castaldo, B. Palazzo, P. Della Vecchia, *Seismic reliability of base-isolated structures with friction pendulum bearings.* Engineering Structures 2015; 95:80-93.

[13] B. Palazzo, P. Castaldo, P. Della Vecchia, *Seismic reliability analysis of base-isolated structures with friction pendulum system,* 2014 IEEE Workshop on Environmental, Energy and Structural Monitoring Systems Proceedings Napoli September 17-18, 2014.

[14] M. Latour , V. Piluso, G. Rizzano, *Free from damage beam-to-column joints: Testing and design of DST connections with friction pads.* Engineering Structures,

[15] P. Castaldo, M. De Iuliis, *Optimal integrated seismic design of structural and viscoelastic bracing-damper systems.* Earthquake Engineering and Structural Dynamics, Volume 43, Issue 12, Pages 1809-1827, 10 October 2014.

[16] M. De Iuliis, P. Castaldo, *An energy-based approach to the seismic control of one-way asymmetrical structural systems using semi-active devices.* Ingegneria Sismica - International Journal of Earthquake Engineering 2012; XXIX(4): 31–42.

[17] A. Longo, R. Montuori, V. Piluso, *Theory of Plastic Mechanism Control for MRF-CBF dual system and its validation.* Bulletin of Earthquake Engineering (BEEE). 2014.

[18] M. T. Giugliano, A. Longo, R. Montuori, V. Piluso, *Failure mode and drift control of MRF-CBF dual systems.* The Open Construction and Building Technology Journal, 4, 121-133, 2010.

[19] R. Montuori, V. Piluso, M. Troisi, *Theory of plastic mechanism control of seismic-resistant MR-frames with set-backs.* Open Construction and Building Technology Journal, Volume 6, Pages 404-413, 2012.

[20] A. Longo, R. Montuori, V. Piluso, *Theory of Plastic Mechanism Control of Dissipative Truss Moment Frames.* Engineering Structures, Volume 37, Pages 63-75, 2012.

[21] R. Montuori, R. Muscati, *Plastic design of seismic resistant reinforced concrete frame.* Earthquake and Structures, Volume 8, Issue 1, 2015, Pages 205-224.

[22] A. Longo, R. Montuori, V. Piluso, *Failure Mode Control and Seismic Response of Dissipative Truss Moment Frames.* Journal of Structural Engineering, Volume 138, Pages 1388-1397, 2012.

[23] R. Montuori, V. Piluso, M. Troisi, *Innovative structural details in MR-frames for free from damage structures.* Mechanics Research Communications, Volume 58, Pages 146-156. 2014.

[24] R. Montuori, E. Nastri, V. Piluso, *Advances in Theory of Plastic Mechanism Control: Closed Form Solution for Mr-Frames*. Earthquake Engineering and Structural Dynamics, 2015.

[25] A. Longo, R. Montuori, E. Nastri, V. Piluso, *On the use of HSS in seismic-resistant structures.* Journal of Constructional Steel Research, Volume 103, Pages 1-12, 2014.

[26] A. Longo, R. Montuori, V. Piluso, *Moment Frames - Concentrically Braced Frames Dual Systems: Analysis of Different Design Criteria.* Structure and infrastructure engineering, Volume 12, Issue 1, Pages 122-141, 2 January 2016.

[27] A. Longo, E. Nastri, V. Piluso, *Theory of plastic mechanism control: State-of-the-art*. Open Construction and Building Technology Journal, Volume 8, Pages 262-278, 2014.

[28] A. Longo, R. Montuori, V. Piluso, *Failure mode control of X-braced frames under seismic actions*. Journal of Earthquake Engineering, 12: 728-759, 2008.

[29] A. Longo, R. Montuori, V. Piluso, *Plastic design of seismic resistant V-Braced frames*. Journal of Earthquake Engineering, Volume 12: 1246-1266, 2008.

[30] A. Longo, R. Montuori, V. Piluso, *Influence of design Criteria on the seismic Reliability of X-Braced Frames*, Journal of Earthquake Engineering, Volume 12, Issue 3, Pages 406-431, 2008.

[31] R. Montuori, E. Nastri, V. Piluso, *Rigid-plastic analysis and moment–shear interaction for hierarchy criteria of inverted Y EB-Frames*. Journal of Constructional Steel Research, Volume 95 71-80, 2014.

[32] R. Montuori, E. Nastri, V. Piluso, *Theory of Plastic Mechanism Control for Eccentrically Braced Frames with inverted Y-scheme*. Journal of Constructional Steel Research, Volume 92, Pages 122-135, 2014.

[33] M. T. Giugliano, A. Longo, R. Montuori, V. Piluso, *Plastic design of CB-frames with reduced section solution for bracing members*. Journal of Constructional Steel Research, Volume 66, Pages 611-621, 2010.

[34] R. Montuori, E. Nastri, V. Piluso, *Theory of plastic mechanism control for the seismic design of braced frames equipped with friction dampers*. Mechanics Research Communications, Volume 58, Pages 112-123, 2014.

[35] A. Longo, R. Montuori, V. Piluso, *Seismic reliability of V-braced frames: Influence of design methodologies*. Earthquake Engineering and Structural Dynamics, Volume 38, Pages 1587-1608, 2009.

[36] A. Longo, R. Montuori, V. Piluso, *Seismic Reliability of Chevron Braced Frames with Innovative Conception of Bracing Members*, Advanced Steel Construction, an International Journal, Volume 5, No. 4, December 2009.

[37] L. Mastrandrea, E. Nastri, V. Piluso, *Validation of a design procedure for failure mode control of EB-Frames: Push-over and IDA analyses.* Open Construction and Building Technology Journal, Volume 7, Pages 193-207, 2013.

[38] R. Montuori, E. Nastri, V. Piluso, *Seismic Design of MRF-EBF Dual Systems with Vertical Links: Ec8 Vs Plastic Design*. Journal of Earthquake Engineering, 2015.

[39] R. Montuori, E. Nastri, V. Piluso, *Seismic Response of EB-Frames with Inverted Yscheme: TPMC Versus Eurocode Provisions.* Earthquakes and Structures, 2015.

[40] R. Montuori, R. Muscati, *A General Procedure for Failure Mechanism Control of Reinforced Concrete Frames*. Accepted for publication to Engineering Structures (2016).

[41] CSI 2007. *SAP 2000: Integrated Finite Element Analysis and Design of Structures. Analysis Reference*. Computer and Structure Inc. University of California, Berkeley.

[42] SEAOC. *Vision 2000 - A Framework for Performance Based Design, Volumes I, II, III. Structural Engineers Association of California, Vision 2000 Committee*. Sacramento, California, 1995.

[43] R. Montuori, E. Nastri, V. Piluso. *Modelling of Floor Joists Contribution to the Lateral Stiffness of RC Buildings Designed for Gravity Loads*. Accepted for publication in Engineering Structures (2016).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A COMPARISON OF MILP AND MINLP SOLVER PERFORMANCE ON THE EXAMPLE OF A DRINKING WATER SUPPLY SYSTEM DESIGN PROBLEM

**Lea Rausch, Philipp Leise, Thorsten Ederer, Lena C. Altherr and Peter F. Pelz**

Technische Universität Darmstadt, Chair of Fluid Systems
Otto-Berndt-Str. 2, 64287 Darmstadt, Germany
e-mail: {lea.rausch, thorsten.ederer, lena.altherr, peter.pelz}@fst.tu-darmstadt.de, philipp.leise@live.de

**Keywords:** Technical Operations Research, Mixed-Integer Nonlinear Optimisation, Solver Performance, Drinking Water Supply, System Design Problem

**Abstract.** *Finding a good system topology with more than a handful of components is a highly non-trivial task. The system needs to be able to fulfil all expected load cases, but at the same time the components should interact in an energy-efficient way. An example for a system design problem is the layout of the drinking water supply of a residential building. It may be reasonable to choose a design of spatially distributed pumps which are connected by pipes in at least two dimensions. This leads to a large variety of possible system topologies. To solve such problems in a reasonable time frame, the nonlinear technical characteristics must be modelled as simple as possible, while still achieving a sufficiently good representation of reality. The aim of this paper is to compare the speed and reliability of a selection of leading mathematical programming solvers on a set of varying model formulations. This gives us empirical evidence on what combinations of model formulations and solver packages are the means of choice with the current state of the art.*

# 1 INTRODUCTION

This paper will focus on the run-time comparison of different mathematical optimisation solvers, complementarity constraint methods and approximation methods applied to an exemplary model of the drinking water supply system in a residential building. For a more comprehensive coverage of this comparison, we vary the number of floors and the inlay pressure, whereas the overall method of modelling the pump system will remain the same. Depending on the chosen solver, complementarity constraint and approximation method, the problem will be modelled as a mixed-integer linear problem (MILP) or a mixed-integer nonlinear problem (MINLP).

Key decision for the topology problem is to find the optimal placement of pumps and pipes. Pumps can be positioned between the floors and pipes have to be laid out to connect the main water connection in the basement with the pumps and the consumers. In addition, the operational parameters for pumps are calculated, e.g. flow rate, power input and rotational speed.

To analyse the solver behaviour for different model sizes and formulations, different variations of the number of floors and the inlay pressure were created. Each combination of a number of floors and inlay pressure specifies a load setting of the pump network model. In this paper $40$ different load settings were solved for all feasible combinations of solvers, constraint methods and approximation methods. This led to a total of $400$ solved problems. Following solvers and methods were used and will be introduced in more detail in Section 3:

- *Solvers*: MINLP: Couenne, Scip MILP: Gurobi, GLPK
- *Complementarity Constraint Methods:* bigM (linear) and product (nonlinear) constraint
- *Approximation Methods for Characteristic Curves:*

  Linear: piecewise linear approximation by aggregated convex combination (ACC),

  Nonlinear: cubic, blending-function (Coons) and quadratic approximation

## 1.1 Estimation of Input Parameters

The following values, which are required to model the pump system, are calculated upfront.

### 1.1.1 Expected Load for Individual Floors

The minimal pressure head $H_{min}$, flow rate $Q$, and maximal flow velocity $v_{\text{max}}$ for an individual floor are estimated by assuming a standard consumption for a flat following the European industrial standard guideline (DIN EN) [1]. The DIN EN guideline defines a standard flat with a dishwasher, kitchen sink, washbasin, shower, washing machine, and toilette. Information on the flow rates, minimal pressure, resistance coefficients, and maximal flow velocity are given. By applying the foundations of hydraulic systems [2] and assuming the hydraulic case with the highest volume flow and pressure head, the total pressure loss due to friction can be calculated. Including the minimal flow pressure from the DIN EN guideline leads to the estimated load of an individual floor given in Table1.

| Parameter | Symbol | Value |
|---|---|---|
| minimal pressure head | $H_{\text{min}}$ | $13\,\text{m}$ |
| flow rate | $Q$ | $1.50\,\text{m}^3\,\text{h}^{-1}$ |
| maximal flow velocity | $v_{\text{max}}$ | $2.0\,\text{m}\,\text{s}^{-1}$ |

Table 1: Summary of parameter estimation for individual floor.

### 1.1.2 Selection of Pipe Diameters

A selection rule for the pipe diameters depending on the flow rate $Q$ is determined. The pipe diameters given in the DIN EN guideline [1] are listed in Table 2. The guideline also specifies the flow velocity $v$ to to be within specific bounds in order to avoid microbial load and noise. We consider $0\,\text{m}^3\,\text{h}^{-1} \leq Q \leq 60\,\text{m}^3\,\text{h}^{-1}$ and $1\,\text{m}\,\text{s}^{-1} \leq v \leq 2\,\text{m}\,\text{s}^{-1}$.

| Pipe diameter in $\text{mm}$ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 13 | 16 | 19.6 | 25.6 | 32 | 39 | 51 | 60 | 72.1 | 84.9 | 104 |

Table 2: Possible diameters for pipes.

This leads to the selcetion of pipe diameters depending on the flow rate which is shown in Figure 1.

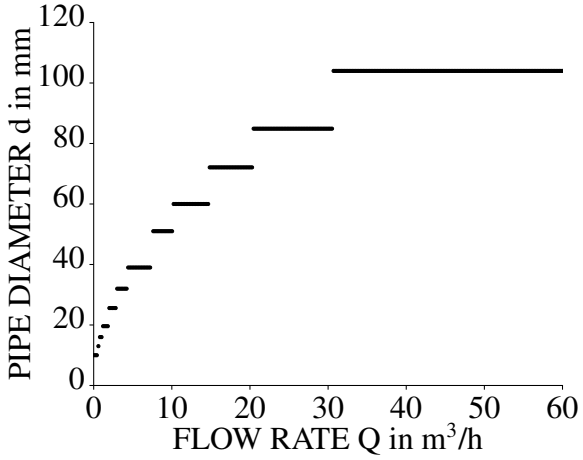



Figure 1: Pipe diameter as function depending on the flow rate.

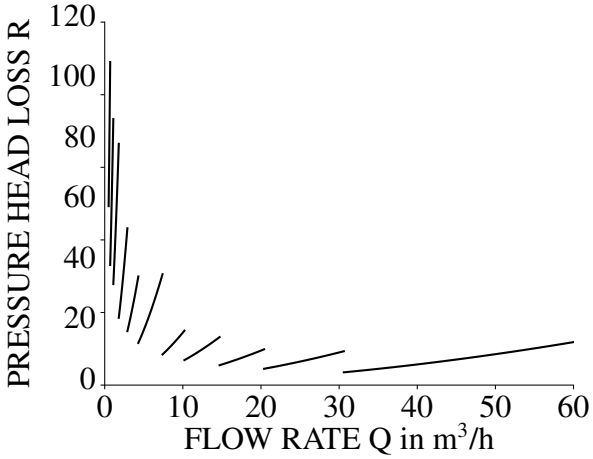



Figure 2: Linear approximation of the specific pressure head loss.

### 1.1.3 Approximation of Pipe Characteristics

Pressure loss due to pipe friction can be represented by the *Darcy-Weisbach* equation [2]

$$\Delta H_R = \frac{\Delta p_R}{\rho\, g} = R \cdot L = \lambda \frac{1}{2\, d_i\, g} v^2 L\,, \tag{1}$$

with pipe diameter $d_i$, pipe length $L$, gravitational constant $g$, average flow velocity $v$ in pipe, friction factor $\lambda$ and specific pressure loss $R$. Using the flow rate representation [2]

$$Q = vA = v\frac{1}{4}\pi d_i^2 \tag{2}$$

the specific pressure loss can be transformed to

$$R(Q, d_i) = \lambda \frac{1}{d_i^5}\frac{8}{\pi^2}\frac{Q^2}{g}\,. \tag{3}$$

The friction factor $\lambda$ can be calculated iteratively for a fixed pipe diameter $d_i$ and fixed flow rate $Q$ with Newton's method. Figure 2 shows that the specific pressure loss $R$ depends strongly on the pipe diameter and less on the volume flow. For simplification, we use the maximal specific pressure loss $R_{\text{max}}$ for each pipe diameter as an approximation.

### 1.2 Pump models

In this paper, we consider three exemplary pumps from the KSB pump series Hya-Solo: EV 1/0206B, EV 1/0406B, and EV 1/0605B. The characteristics of these pumps can be found in [3]. The pump curves describe the pressure head increase and the power consumption as functions of the flow rate and rotational speed. These pump curves will be approximated with different methods which are introduced in Section 3.3.

## 2 MODEL

In this section, the optimisation model is presented. We do not yet focus on constraints that can be modelled in different ways: For complementarity constraints, only the formulation as disjunction is given – the explicit formulations are introduced in Section 3.2. The approximation methods of he pump characteristics are discussed in Section 3.3. The main variables and parameteres which are used in this model can be found in Table 4.

### 2.1 Objective Function

The model aims to minimise the sum of investment costs $C$ of the components and the energy costs for operating the pumps:

$$\min \quad C_{\text{pipes}} + C_{\text{pumps}} + c_{\text{energy}} \cdot T \cdot P_{\text{total}}\,.$$

The operational costs $c_{\text{energy}}$ are estimated with $0.2951\,€/\text{kWh}$ for an operating time of $T = 5$ years. The costs for the pipe purchase are estimated with $c_{\text{pipe}} = 50\,€/\text{m}$. For the pumps the investment costs are stated in the Table 3.

| Pump Model | Price in € |
|---|---|
| 10206/B | 2344.55 |
| 10406/B | 2409.35 |
| 10605/B | 2484.75 |

Table 3: Investment costs for the different pump models.

| Type | Name | Description | Value Range | Dimension |
|---|---|---|---|---|
| *parameters* | $n_{\text{floors}}$ | number of floors | | |
| | $H_0$ | inlet pressure head | | $L$ |
| | $n_{\text{pipes}}$ | number of possible pipes | {1,2,3,...,$\infty$} | |
| | $n_{\text{pumps}}$ | number of pumps | {3} | |
| | $T$ | expected operating time of the system | [0,$\infty$] | $T$ |
| | $D$ | ingoing and outgoing flow on each floor | | $L^3T^{-1}$ |
| | $h_{S\hat{S}}$ | height offset between floor $S$ and $\hat{S}$ | | $L$ |
| | $Q_{\text{max}}$ | maximal flow for pipe diameter | [0, 100] | $L^3T^{-1}$ |
| *variables* | $l_{\text{pipe}}$ | pipe length | [0,$\infty$] | $L$ |
| | $P$ | power consumption of each pump | [0, 1800] | $ML^2T^{-3}$ |
| | $Q$ | flow rate in each pipe | [0, 100] | $L^3T^{-1}$ |
| | $Q_{\text{square}}$ | quadratic flow rate in each pipe | [0, 3600] | $L^6T^{-2}$ |
| | $\hat{n}$ | normalised pump rotational speed | [0.6,1] | |
| | $\Delta H$ | pressure head increase for each pump | [0,100] | $L$ |
| | $H$ | pressure head in each floor | [0,100] | $L$ |
| | $k_{\text{pipe}}$ | indicator for pipe selection | {0,1} | |
| | $k_{\text{pump}}$ | indicator for pump placement on a pipe | {0,1} | |
| | $R_{\text{pipe}}$ | specific pressure loss due to pipe friction | [0,$\infty$] | |
| | $d_{\text{pipe}}$ | selected pipe diameter | [9,105] | $L$ |
| | $k_{\text{Q}}$ | non-zero flow indicator | {0,1} | |

Table 4: Main input parameters and decision variables of the optimisation model.

Expressed in decision variables, the objective is

$$\begin{aligned} \min \quad & \sum_{i=1}^{n_{\text{pipes}}} \sum_{j=1}^{n_{\text{pumps}}} k_{\text{pump}}(i,j) \cdot c_{\text{pump}}(j) \\ & + \sum_{i=1}^{n_{\text{pipes}}} k_{\text{pipe}}(i) \cdot c_{\text{pipe}} \cdot l_{\text{pipe}}(i) \\ & + c_{\text{energy}} \cdot T \cdot \sum_{i=1}^{n_{\text{pipes}}} \sum_{j=1}^{n_{\text{pumps}}} P(i,j)\,. \end{aligned} \tag{4}$$

## 2.2 Constraints

Pipe indicators model the interconnection of floors. The correspondence of pipes and floors can be encoded by a matrix $A$ where $a_{ij}$ is equal to the index $k \in \{1, ..., n_{\text{pipes}}\}$ of the pipe which connects floor $i$ with floor $j$. Pipes can only be placed from a lower floor to one above, i.e. $a_{ij} = 0\ \forall j \leq i$. Hence, the number of pipes in the model is given by $n_{\text{pipes}} = \frac{1}{2} n_{\text{floors}}(n_{\text{floors}} - 1)$. Only one pipe can be chosen to reach one floor:

$$\sum_{i=1}^{n_{\text{floors}}} k_{\text{pipe}}(a_{ij}) \leq 1, \qquad \forall j \in \{2, ..., n_{\text{floors}}\}\,. \tag{5}$$

In addition, a bigM or product constraint for the following complementarity expression (c.f. Section 3.2) is required to allow a flow rate only if the pipe is selected:

$$expr = Q(k,j) \text{ depending on } k_{\text{pipe}}(k) \qquad \forall\, k \in \{1,2,...,n_{\text{pipes}}\}\,. \tag{6}$$

The pipe diameters need to be selected as described in Section 1.1.2. Let the parameters $d_{\text{values}}(i)$ be the possible pipe diameters and $Q_{jd}(i)$ the flow rates at the jump discontinuities in Figure 1.

$$\begin{aligned} Q(j) &\leq Q_{\text{max}} && \forall j \in \{1,2,...,n_{\text{pipes}}\}\,, \\ Q(j) &\geq Q_{jd}(i) \cdot k_{\text{Q}}(j,i) && \forall j \in \{1,2,...,n_{\text{pipes}}\}\,, \\ & && \forall i \in \{1,2,...,12\} \\ Q(j) &\leq Q_{jd}(i+1) \cdot k_{\text{Q}}(j,i) + Q_{\text{max}} \cdot (1 - k_{\text{Q}}(j,i)) && \forall j \in \{1,2,...,n_{\text{pipes}}\}\,, \\ & && \forall i \in \{1,2,...,12\}\,, \\ d_{\text{pipe}} &= \sum_{i=1}^{12} k_{\text{Q}}(j,i) \cdot d_{\text{values}}(i) && \forall j \in \{1,2,...,n_{\text{pipes}}\}\,. \end{aligned} \tag{7}$$

To model the flow requirements, let $I$ be the $(n_{\text{pipes}} \times m)$ incidence matrix and $D$ the input (positive value) and output (negative value) flows of the floors. The continuity equation [4] on networks can be implemented by the following constraints:

$$\sum_{k=1}^{m} I(i,k)\, Q(k) = D(i) \qquad \forall i \in \{1,...,n_{\text{pipes}}\}\,. \tag{8}$$

The pipe friction is included based on the estimation in Section 1.1.3 and can be modelled by a bigM or product constraint for the following complementarity expression which is active in a pipe if the flow rate is higher than 0, i.e., if $k_{\text{Q}}(i,k) = 1$:

$$\begin{aligned} expr = R(i) - R_{\text{max}}(k) \quad &\text{depending on } k_{\text{Q}}(i,k) \quad \forall\, i \in \{1,2,...,n_{\text{pipes}}\}, \\ & \qquad\qquad\qquad\quad \forall\, k \in \{2,...,12\}\,. \end{aligned} \tag{9}$$

The pressure requirement for each floor from Table 1 gives the minimal pressure head $H_{\text{min}}$:

$$\begin{aligned} H(1) &= H_0 && , \\ H(k) &\geq H_{\text{min}} && \forall\, k \in \{2,...,n_{\text{floors}}\}\,. \end{aligned} \tag{10}$$

In addition, a bigM or product constraint for the following complementarity expression is required to model the differences of the pressure head between the floors $S$ and $\hat{S}$ for each selected pipe:

$$\begin{aligned} expr = H(\hat{S}) - \Delta H(i,1) - \Delta H(i,2) - \Delta H(i,3) + h_{S\hat{S}} + R(i) \cdot h_{S\hat{S}} - H(S) \\ \text{depending on } k_{\text{pipe}}(k) \qquad \forall\, k \in \{1,2,...,n_{\text{pipes}}\}\,. \end{aligned} \tag{11}$$

The network dependencies are given by:

$$\begin{aligned} k_{\text{pump}}(k,i) \leq k_{\text{pipe}}(k) \qquad & \forall\, i \in \{1,2,3\}\,, \\ & \forall\, k \in \{1,2,...,n_{\text{pipes}}\}\,. \end{aligned} \tag{12}$$

To include the relationship of flow rate $Q$, rotational speed $\hat{n}$, power $P$ and pressure head $\Delta H$ as given by the pump curves, the general constraints are shown here. The exact formulas for $\tilde{P}$

and $\tilde{H}$ depend on the selected approximation method of the pump curves which are described in Section 3.3:

$$P(i,j) = k_{\text{pump}}(i,j) \cdot \tilde{P}(Q(i), \hat{n}(i,j)) \qquad \begin{array}{l} \forall\, i \in \{1,..,n_{\text{pipes}}\}\,, \\ \forall\, j \in \{1,..,n_{\text{pumps}}\}\,, \end{array} \tag{13}$$

$$\Delta H(i,j) = k_{\text{pump}}(i,j) \cdot \tilde{H}(Q(i), \hat{n}(i,j)) \qquad \begin{array}{l} \forall\, i \in \{1,..,n_{\text{pipes}}\}\,, \\ \forall\, j \in \{1,..,n_{\text{pumps}}\}\,. \end{array} \tag{14}$$

In addition, the variables $Q$ and $\hat{n}$ need to be restricted to the value range which is included in the pump curves.

## 3 SELECTED SOLVERS AND METHODS

This section gives a short introduction of all compared solvers, constraint methods and approximation methods including a quality analysis of the approximation methods.

### 3.1 Solvers

The presented comparison includes four different solvers. Two solvers, Gurobi and GLPK, for linear problems and two for nonlinear problems, Couenne and Scip. Except of Gurobi all solvers are open source. A short general description of the used solver can be found below. *Gurobi* is a commercial solver for mixed-integer linear problems (MILP), as well as linear problems (LP). A new version of Gurobi can also include quadratic aspects but is not used in this paper. The solver is provided by the company Gurobi [5]. *GLPK* (GNU Linear Programming Kit) includes an open source solver for mixed-integer linear problems (MILP), as well as linear problems (LP). GLPK works with the GNU MathProg modelling language, which is a supported by the Free Software Foundation [6]. *Couenne* (Convex Over and Under ENvelopes for Nonlinear Estimation) is an open source solver for mixed-integer nonlinear problems (MINLP). The solver is provided by the Computational Infrastructure For Operations Research Foundation (COIN-OR) [7]. *Scip* (Solving Constraint Integer Programs) is a commercial open source solver for mixed-integer nonlinear problems (MINLP) as well as for mixed-integer linear problems (MILP). Scip is directed towards meeting the needs of Mathematical Programming experts who are interested in a high level of control and detailed information of the solution process. The solver is provided by the Zuse Institute Berlin (ZIB) [8].

### 3.2 Methods to Model Complementarity Constraints

In general, complementarity constraints describe the following relation:

$$expr_1 = 0 \ \vee \ expr_2 = 0\,. \tag{15}$$

We will compare two different methods to model complementarity constraints, which are used to activate and deactivate equality constraints depending on the value of an auxiliary decision variable. The linear bigM method and the nonlinear product method. As the product complementarity constraint is nonlinear, it can not be used with the linear solvers Gurobi and GLPK.

For both methods assume $expr = 0$ is a linear constraint for a number of decision variables which only needs to be met if the auxiliary binary decision variable $k \in \{0,1\}$ is active, i.e. $k = 0$. This requirement can be modelled with the following constraints.

### 3.2.1 bigM Method

Let $bigM \in \mathbb{R}^+$ be a sufficiently big positive constant. Then the above described requirement can be modelled with the constraint (16).

$$\begin{aligned} expr &\leq +\, k \text{ bigM}\,, \\ expr &\geq -\, k \text{ bigM}\,. \end{aligned} \tag{16}$$

The choice of $bigM$ needs to guarantee $-\text{bigM} \leq expr \leq \text{bigM}$ for any feasible combinations of the decision variables in $expr$. Nevertheless, $bigM$ should be estimated as small as possible, as a big value for $bigM$ increases the risk of numerical errors.

### 3.2.2 Product Method

To reduce numerical errors caused by the bigM method, one can use the standard complementarity constraint in product form (17) for nonlinear solvers instead.

$$(expr)\,(1-k) = 0\,. \tag{17}$$

## 3.3 Approximation Method

Pump curves are used to represent the key pump characteristics and show the pressure head increase $\Delta H(\hat{n}, Q)$ as a function over the rotational speed and flow rate as well as the required power input $P(\hat{n}, Q)$. These pump curves need to be approximated for the MI(N)LPs. For this comparison one linear and three nonlinear approximation methods are used and the quality of all four methods is analysed by validating the affinity laws, c.f. Section 3.3.5.

### 3.3.1 Piecewise Linearization by Aggregated Convex Combination

An aggregated convex combination formulation based on a Freudenthal triangulation [9] is used as piecewise linear approximation method. The surfaces which describe the pump curves $\Delta H(\hat{n}, Q)$ and $P(\hat{n}, Q)$ are cut into triangles. In each triangle the curve is approximated by a linear triangle spanned by the grid points of the pump curve. Hence, in each grid point the approximation equals the actual pump curve. In every other point, the approximation is a linear combination of the three surrounding grid points. In this paper a 7x7 triangulation is used which leads to 49 grid points.

### 3.3.2 Cubic Approximation

For the cubic method the pump curves are approximated with the cubic functions (18) where the coefficients $a, b, c, \alpha, \beta, \gamma$ and $\delta$ are determined by a regression analysis. In the regression analysis the rotational speed is scaled to the maximal pump rotational speed $n_{\text{ref}}$, i.e. $\hat{n} = \frac{n}{n_{\text{ref}}}$.

$$\begin{aligned} \Delta H(\hat{n}, Q) &= aQ^2 + bQ\hat{n} + c\hat{n}^2\,, \\ P(\hat{n}, Q) &= \alpha Q^3 + \beta Q^2\hat{n} + \gamma Q\hat{n}^2 + \delta\hat{n}^3\,. \end{aligned} \tag{18}$$

The results of the regression analysis for the three different pump models pumps Hya-Solo EV 1/0206B, Hya-Solo EV 1/0406B and Hya-Solo EV 1/0605B are shown in Table 5 for the pressure head increase $\Delta H$ and for the power consumption $P$. In this table one can also find the

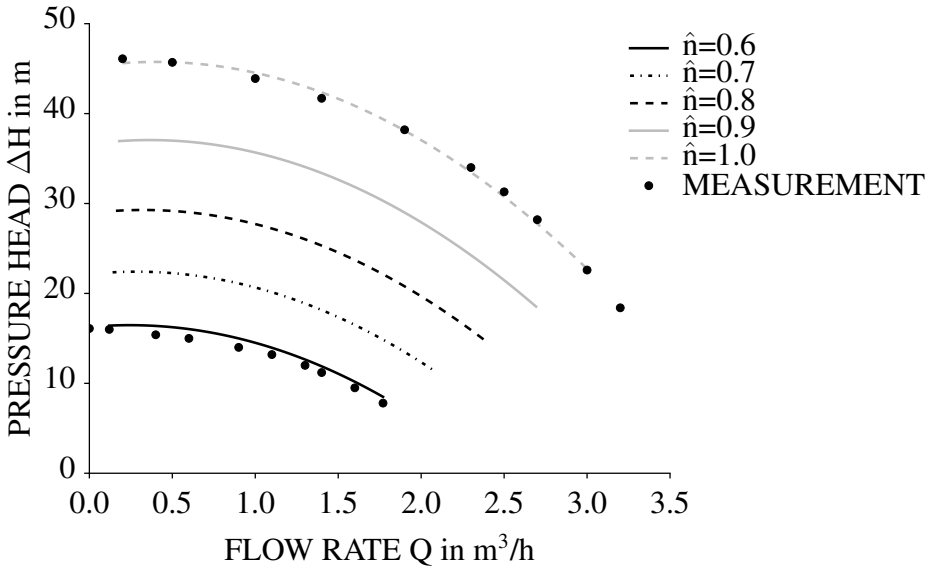



Figure 3: Cubic approximation of pressure head increase in pump curve for the pump model Hya-Solo EV 1/0206B.

standard error $\epsilon_\beta(j)$ for each coefficient of the approximation which is calculated with the MSE (*mean squares of errors*) $\sigma_r^2 = \frac{\sum_{i=1}^{m} \epsilon^2}{m-p-1}$ leading to the standard error $\epsilon_\beta(j) = \sigma_r \sqrt{(X^T X)^{-1}_{jj}}$, c.f. [10].

Figure 3 shows the cubic approximation of the pump curve for the pressure increase $\Delta H(\hat{n}, Q)$.

### 3.3.3 Bending-Function (Coons) Approximation

The Coons approximation method [11] is not physically motivated but based on graphical data processing. It interpolates the pump curve with an auxiliary variable $s$ with $0 \leq s \leq 1$. The pump curve is approximated as a linear function of $s$ and a quadratic function of the second variable $Q$:

$$\begin{aligned} \Delta H(s, Q) &= (a_1 Q^2 + b_1 Q + c_1)s + (a_2 Q^2 + b_2 Q + c_2)(1 - s)\,, \\ P(s, Q) &= (\alpha_1 Q^2 + \beta_1 Q + \gamma_1)s + (\alpha_2 Q^2 + \beta_2 Q + \gamma_2)(1 - s)\,. \end{aligned} \tag{19}$$

For all pump models the results of the regression analysis can be found in the appendix in Table 7 and 8.

### 3.3.4 Quadratic Approximation

A further nonlinear approximation method is a complete quadratic approximation [12] for $0.6 \leq \hat{n} \leq 1$. This approximation is *not* physically motivated and does not adhere to the affinity laws. However, due to its lower polynomial degree, it might speed up the computation without too much of a deviation from the cubic approximation.

$$\begin{aligned} \Delta H(\hat{n}, Q) &= aQ^2 + bQ\hat{n} + c\hat{n}^2\,, \\ P(\hat{n}, Q) &= \alpha Q^2 + \beta\hat{n}^2 + \gamma Q\hat{n} + \delta Q + \epsilon\hat{n} + \zeta\,. \end{aligned} \tag{20}$$

The approximation for the pressure head increase is the same as for the cubic approximation 3.3.2. Therefore, the coefficients $a, b$ and $c$ can be taken from the cubic regression analysis in

| Coefficient | Value | Error |
|---|---|---|
| $a_{206}$ | $-3.415$ | 0.25 |
| $b_{206}$ | 2.760 | 0.87 |
| $c_{206}$ | 45.193 | 0.63 |
| $a_{406}$ | $-0.921$ | 0.085 |
| $b_{406}$ | 1.177 | 0.55 |
| $c_{406}$ | 52.296 | 0.75 |
| $a_{605}$ | $-0.345$ | 0.04 |
| $b_{605}$ | 0.373 | 0.35 |
| $c_{605}$ | 47.973 | 0.65 |

| Coefficient | Value | Error |
|---|---|---|
| $\alpha_{206}$ | $-11.142$ | 2.52 |
| $\beta_{206}$ | 41.522 | 13.29 |
| $\gamma_{206}$ | 54.318 | 19.61 |
| $\delta_{206}$ | 191.211 | 7.75 |
| $\alpha_{406}$ | $-1.774$ | 0.79 |
| $\beta_{406}$ | 5.902 | 7.59 |
| $\gamma_{406}$ | 135.815 | 19.57 |
| $\delta_{406}$ | 245.276 | 12.86 |
| $\alpha_{605}$ | $-0.664$ | 0.12 |
| $\beta_{605}$ | 1.154 | 1.59 |
| $\gamma_{605}$ | 125.728 | 5.84 |
| $\delta_{605}$ | 276.781 | 5.81 |

Table 5: Coefficients for the cubic approximation of the pressure head $\Delta H$ (right) and the power consumption $P$ (left) for each of the three pump models 10**206**/B, 10**406**/B, 10**605**/B. The coefficients of the other approximation methods can be found in the appendix.

Table 5. The coefficients $\alpha$, $\beta$, $\gamma$, $\delta$, $\epsilon$ and $\zeta$ need to be determined in an additional regression analysis. The results for the power consumption $P$ for the three different pump models can be found in the appendix in Table 9.

### 3.3.5 Quality Analysis of Approximation Methods

For comparison, four approximation methods are applied. We want to analyse how well they match reality. The functional correlation of pump rotational speed $\hat{n}$, flow rate $Q$ and pressure head $\Delta H$ are described by affinity laws [13]. They imply the following proportionality:

$$Q \sim \hat{n}, \quad \Delta H \sim \hat{n}^2, \quad \text{and} \quad P \sim \hat{n}^3 \,. \tag{21}$$

These proportionality relations can be transformed to the following quality measure (22) for the approximation methods. The error value $E$ indicates the adherence to the affinity law and therefore the fit quality of the approximation method:

$$E_Q = \frac{Q_1}{Q_2} - \frac{\hat{n}_1}{\hat{n}_2} \approx 0, \quad E_H = \frac{\Delta H_1}{\Delta H_2} - \left(\frac{\hat{n}_1}{\hat{n}_2}\right)^2 \approx 0 \text{ and } E_P = \frac{P_1}{P_2} - \left(\frac{\hat{n}_1}{\hat{n}_2}\right)^3 \approx 0 \,. \tag{22}$$

To measure the quality of the approximation method, a set of 1000 grid points were inserted in the approximation of the pump curve and the results used to calculate the errors in (22). The results are listed in Table 6 in the appendix. The error analysis shows that all used approximation methods have a similar and low error level. The only exception is the bending-function (Coons) approximation method which causes a higher average error for the pressure head $E_H$ for all tested pump models.

## 4 OPTIMISATION RESULTS FOR A SELECTED EXAMPLE

In most cases, the topology results of different solver and method combinations for one load setting are the same or vary only slightly, c.f. Section 5.4.

We chose an example with a residential building with three floors. In the optimised topology in Figure 4 one pump of type Hya-Solo EV 1/0206B is used between the second and third floor.

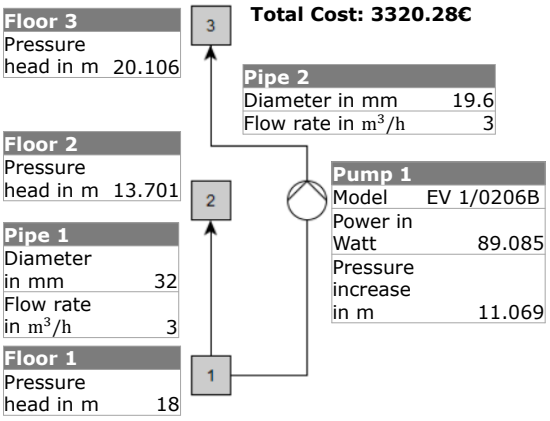



Figure 4: Example of optimal topology with 16% savings compared to standard approach.

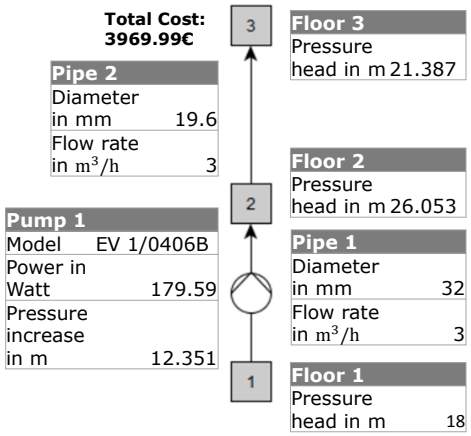



Figure 5: Example topology with pumps only on the first floor.

Compared to the basic approach to place all pumps in the first floor, the optimal solution for this example reduces the sum of investment costs of the components and energy costs for operating the pumps by 16%. Following the basic approach it is necessary to use the more expensive pump Hya-Solo EV 1/0406B to gain the relevant flow rate and pressure head from the pump in the first floor to the third floor, see Figure 5.

## 5 EMPIRICAL RESULTS

All methods show the expected result of the commercial solver Gurobi excelling clearly in run-time performance, especially for a growing model size.

When comparing the geometric mean of run-times over all problems for the best method per solver, it is necessary to focus on lower number of floors, e.g. 7 or 8 floors. The reason is that the problems with a high number of floors could not be solved to optimality with all solvers and all methods within the time limit of six hours (details see Section 5.2).

For 7 floors the best method combination for Counne and Scip is far closer to Gurobi than for the geometric mean over all methods. This shows the relevance of the right method choice: Scip takes about $9$ times the geometric mean time of Gurobi over all problems. When comparing the best suitable method combination for Scip (bigM constraints and quadratic pump curve approximation), it takes only $4$ times the geometric mean time of Gurobi. For Counne the difference of the geometric mean run times to Gurobi drops from factor $135$ to $28$ for the best suitable combination (product complementarity constraints with quadratic approximation).

### 5.1 Number of Constraints and Variables

For all pump characteristic approximations and complementarity constraint methods, the number of constraints as well as the number of variables grow quadratically in the number of floors. The linear approximation has by far the highest numbers with around 4,500 constraints and 8,300 variables for 7 floors compared to the nonlinear approximation methods which all have about 1,500 constraints and 1,000 variables. For 10 floors the linear approach leads to 9,700 constraints and 18,000 variables, whereas the nonlinear methods have 2,500 constraints and 1,600 variables.

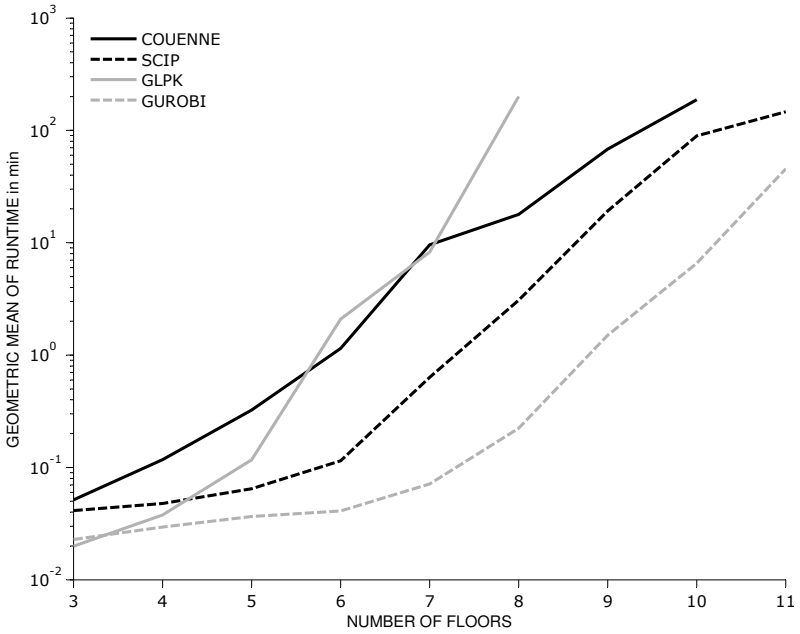



Figure 6: All solvers with geometric mean of run-times over all approximation and constraint methods as well as all inlet pressures for each number of floors.

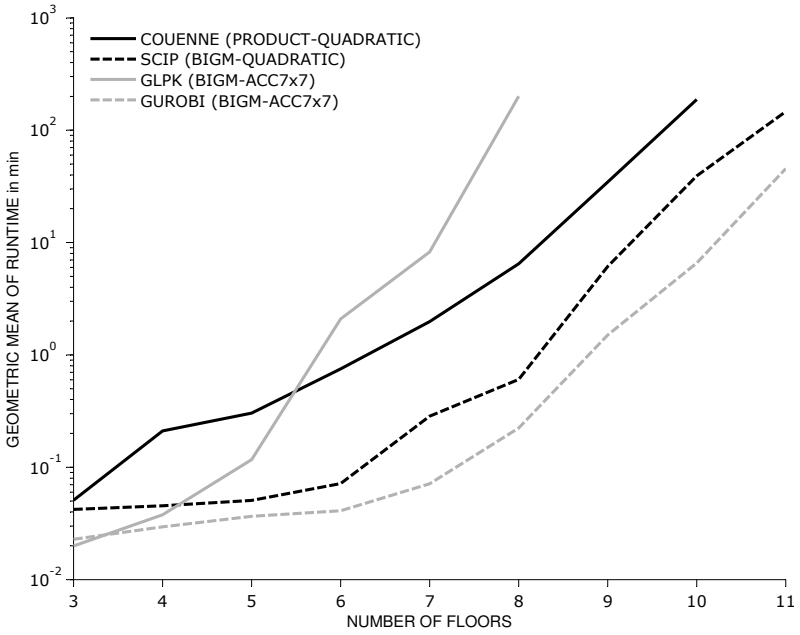



Figure 7: All solvers with geometric mean over all problems solved with the best combination of methods for teach solver over all inlet pressures for each number of floors.

## 5.2 Method Analysis for Individual Solvers

For the nonlinear solvers, we used a total of six different combinations of pump characteristic approximations and complementarity constraint methods. In this section, we show how they influence the run-time of the individual solvers.

### 5.2.1 Couenne

The highest influence on the run-time of the Couennne solver is caused by the method to model the complementarity constraints, as each approximation method combined with the bigM method has a higher run-time than any approximation method combined with the product method for complementarity constraints. There is also a strong increase of the differences in run-time between the complementary methods starting with seven floors.

Comparing the approximation methods one can see that for bigM complementarity constraints the cubic approximation methods outperforms the quadratic approximation, whereas for complementarity constraints it is the other way around and the quadratic approximation outperforms the cubic approximation, especially for a high number of floors. Even though Coons is a simplification of the cubic equations, it has the highest run time for both complementary methods for a high number of floors; for the product method starting with 7 floors, for bigM starting with 5 floors.

### 5.2.2 Scip

Contrary to Couenne, for the Scip solver the approximation method for the pump curves has the highest influence on run time. This is shown as every method to model the complementarity constraints combined with the Coons approximation is worse than every complementarity constraints method combined with the other two approximation methods.

Comparing the two other approximation methods, one can see that for bigM complementarity constraints the quadratic approximation method outperforms the cubic approximation. Whereas, for product complementarity constraints it is the other way around, especially for a high number of floors. BigM is the better complementarity constraint method for Scip since for each approximation method the bigM method has a shorter geometric mean run-time than the same approximation method combined with product complementarity constraints.

## 5.3 Comparison of Scip and Couenne

In Figure 5 one can see that Scip with its best combination (bigM complementarity constraint and quadratic approximation) outperforms compared to Couenne's best combination (product complementarity constraint and quadratic approximation), as well as in the overall comparison with the geometric mean run-time of all combinations.

As already discussed in the individual solver analysis, the product complementarity constraint methods works better for Couenne as for Scip. In Figure 5.4 one can see that for 10 floors this difference is so severe that the in average slower solver Couenne outperforms Scip when using the product complementarity constraint method.

## 5.4 Objective Values and Boundaries of the Presented Methods

In most cases, all solvers, approximation methods and constraint types result in similar topology decisions for the same load setting with only a small deviation in the objective function value due to numerical differences. Only for $4\%$ of all solved problems the objective value deviates more than $5\%$ (but no more than $26\%$) from the average objective value of all problems belonging to this load setting. In these cases, a small deviation of the topology is possible.

An exception with a different topology decision and objective value deviation of $88\%$ occurs when using the solver Scip with cubic approximation method for one load setting with a high inlet pressure ($H_0 = 17\,\mathrm{m}$) with a low number of floors ($n_{\text{floors}} = 3$). For this load setting,

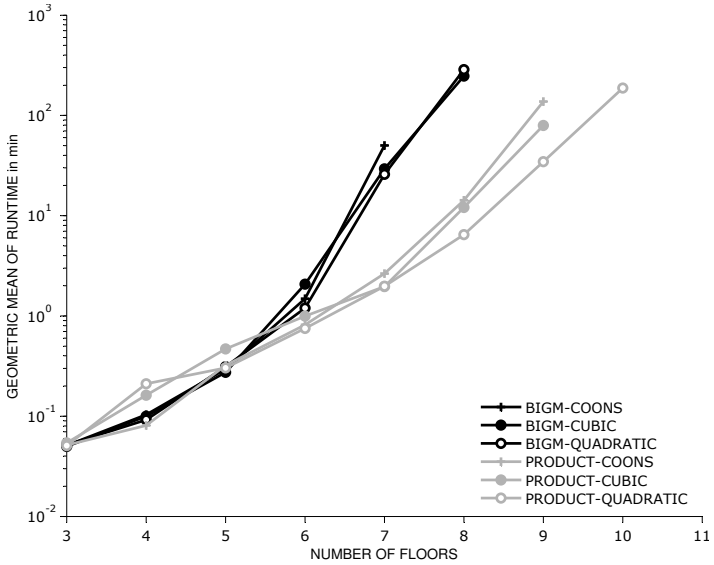



Figure 8: Couenne solver with geometric mean of run-times over all inlet pressures for each number of floors for each combination of approximation and complementary methods.

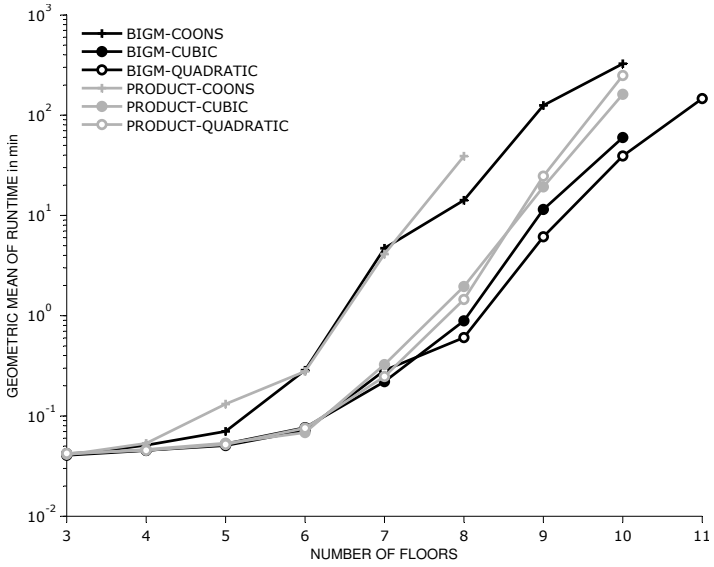



Figure 9: Solver Scip with geometric mean of run-times over all inlet pressures for each number of floors for each combination of approximation and complementary methods.

the topological solution using Scip with cubic approximation method with bigM as well as with product complementarity constraint method differs strongly from all other solutions of the same load setting. With the cubic approximation method no pump is selected in the whole system whereas one pump is selected in all other solutions of this load setting. We assume it is caused by the high numerical complexity of the cubic approximation method, as the inlet pressure is relatively high for the number of floors. This hypothesis is supported by the solutions for a different load setting with an even higher inlet pressure of ($H_0 = 23\,\mathrm{m}$) with the same number of floors ($n_{\text{floors}} = 3$). For this second load setting, all solutions (with all combinations of solvers, constraint and approximation method) contain the topology decision to use no pump, as the inlet pressure is sufficient to supply all floors.

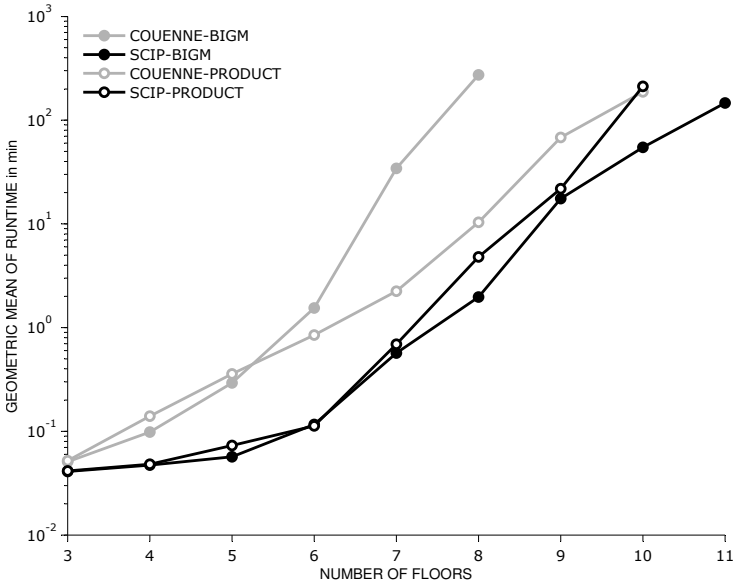



Figure 10: Geometric mean of run time over all inlet pressures and all approximation methods for each number of floors for each complementary method for Couenne and Scip.

## 6 CONCLUSIONS

- A model to design spacially distributed fluid systems has been developed and verified.
- Savings of energy and investment costs by applying the developed model could be detected for an example load setting.
- Solver runtime was compared depending on various combinations of solvers, complementarity constraint methods and approximation methods for pump characteristics.
- Optimal solutions from different solvers and methods vary only slightly, but the run-times deviate severely.
- Approximation methods with simpler equations can lead to higher run-times.

## 7 ACKNOWLEDGEMENT

This work is partially supported by the German Federal Ministry for Economic Affairs and Energy funded project ‘Entwicklung hocheffizienter Pumpensysteme’ and by the German Research Foundation (DFG) funded Collaborative Research Centre SFB 805.

# 8 APPENDIX

| Pump Model | Approximation | Error Type | Average | Standard Deviation |
|---|---|---|---|---|
| EV 1/0206B | cubic | $E_Q$ | $-7.11 \times 10^{-19}$ | $3.74 \times 10^{-33}$ |
| | | $E_H$ | $-4.27 \times 10^{-18}$ | $1.36 \times 10^{-16}$ |
| | | $E_P$ | $-3.69 \times 10^{-19}$ | $1.52 \times 10^{-32}$ |
| | coons | $E_Q$ | $5.55 \times 10^{-19}$ | $3.79 \times 10^{-33}$ |
| | | $E_H$ | $-3.50 \times 10^{-2}$ | $6.94 \times 10^{-3}$ |
| | | $E_P$ | $-1.83 \times 10^{-2}$ | $3.76 \times 10^{-3}$ |
| | quadratic | $E_Q$ | $-9.44 \times 10^{-19}$ | $3.61 \times 10^{-33}$ |
| | | $E_H$ | $-4.35 \times 10^{-18}$ | $1.40 \times 10^{-16}$ |
| | | $E_P$ | $6.26 \times 10^{-4}$ | $4.35 \times 10^{-5}$ |
| | linear | $E_Q$ | $4.00 \times 10^{-19}$ | $3.70 \times 10^{-33}$ |
| | | $E_H$ | $-1.16 \times 10^{-3}$ | $3.12 \times 10^{-3}$ |
| | | $E_P$ | $-3.54 \times 10^{-4}$ | $1.29 \times 10^{-5}$ |
| EV 1/0406B | cubic | $E_Q$ | $-3.33 \times 10^{-20}$ | $3.70 \times 10^{-33}$ |
| | | $E_H$ | $-8.30 \times 10^{-18}$ | $1.40 \times 10^{-16}$ |
| | | $E_P$ | $1.72 \times 10^{-19}$ | $1.56 \times 10^{-32}$ |
| | coons | $E_Q$ | $-8.55 \times 10^{-19}$ | $3.52 \times 10^{-33}$ |
| | | $E_H$ | $2.60 \times 10^{-1}$ | $1.95 \times 10^{-1}$ |
| | | $E_P$ | $2.24 \times 10^{-1}$ | $1.22 \times 10^{-1}$ |
| | quadratic | $E_Q$ | $6.66 \times 10^{-19}$ | $3.78 \times 10^{-33}$ |
| | | $E_H$ | $-9.92 \times 10^{-18}$ | $1.38 \times 10^{-16}$ |
| | | $E_P$ | $-2.22 \times 10^{-4}$ | $2.66 \times 10^{-6}$ |
| | linear | $E_Q$ | $-4.33 \times 10^{-19}$ | $3.70 \times 10^{-33}$ |
| | | $E_H$ | $-1.45 \times 10^{-3}$ | $3.96 \times 10^{-3}$ |
| | | $E_P$ | $-1.42 \times 10^{-3}$ | $2.78 \times 10^{-5}$ |
| EV 1/0605B | cubic | $E_Q$ | $5.55 \times 10^{-19}$ | $3.91 \times 10^{-33}$ |
| | | $E_H$ | $-5.42 \times 10^{-18}$ | $1.34 \times 10^{-16}$ |
| | | $E_P$ | $2.03 \times 10^{-18}$ | $1.66 \times 10^{-32}$ |
| | coons | $E_Q$ | $-5.33 \times 10^{-19}$ | $3.49 \times 10^{-33}$ |
| | | $E_H$ | $2.61 \times 10^{-1}$ | $1.78 \times 10^{-1}$ |
| | | $E_P$ | $2.38 \times 10^{-1}$ | $1.40 \times 10^{-1}$ |
| | quadratic | $E_Q$ | $-3.00 \times 10^{-19}$ | $3.86 \times 10^{-33}$ |
| | | $E_H$ | $-5.91 \times 10^{-18}$ | $1.34 \times 10^{-16}$ |
| | | $E_P$ | $-1.75 \times 10^{-4}$ | $6.67 \times 10^{-5}$ |
| | linear | $E_Q$ | $6.88 \times 10^{-19}$ | $4.05 \times 10^{-33}$ |
| | | $E_H$ | $9.33 \times 10^{-4}$ | $2.57 \times 10^{-3}$ |
| | | $E_P$ | $-1.51 \times 10^{-3}$ | $2.42 \times 10^{-5}$ |

Table 6: Error in affinity laws for the pump characteristic approximation methods.

| Coefficient | Value | Error |
|---|---|---|
| $a_{206}(1)$ | $-3.415$ | $0.25$ |
| $a_{206}(2)$ | $-2.761$ | $0.23$ |
| $b_{206}(1)$ | $2.760$ | $0.87$ |
| $b_{206}(2)$ | $0.397$ | $0.42$ |
| $c_{206}(1)$ | $45.193$ | $0.63$ |
| $c_{206}(2)$ | $15.940$ | $0.15$ |
| $a_{406}(1)$ | $-0.921$ | $0.08$ |
| $a_{406}(2)$ | $-0.902$ | $0.08$ |
| $b_{406}(1)$ | $1.177$ | $0.55$ |
| $b_{406}(2)$ | $0.601$ | $0.32$ |
| $c_{406}(1)$ | $52.296$ | $0.75$ |
| $c_{406}(2)$ | $17.991$ | $0.25$ |
| $a_{605}(1)$ | $-0.345$ | $0.04$ |
| $a_{605}(2)$ | $-0.329$ | $0.03$ |
| $b_{605}(1)$ | $0.373$ | $0.35$ |
| $b_{605}(2)$ | $0.129$ | $0.14$ |
| $c_{605}(1)$ | $47.973$ | $0.65$ |
| $c_{605}(2)$ | $16.257$ | $0.15$ |

Table 7: Coefficients for the bending-function (Coons) approximation of the pressure head characteristic for each of the three pump models 10**206**/B, 10**406**/B, 10**605**/B.

| Coefficient | Value | Error |
|---|---|---|
| $\alpha_{206}(1)$ | $-16.441$ | $3.97$ |
| $\alpha_{206}(2)$ | $-14.684$ | $3.23$ |
| $\beta_{206}(1)$ | $134.804$ | $14.42$ |
| $\beta_{206}(2)$ | $51.325$ | $6.34$ |
| $\gamma_{206}(1)$ | $165.890$ | $10.51$ |
| $\gamma_{206}(2)$ | $39.560$ | $2.38$ |
| $\alpha_{406}(1)$ | $-10.958$ | $1.68$ |
| $\alpha_{406}(2)$ | $-3.328$ | $2.42$ |
| $\beta_{406}(1)$ | $175.426$ | $11.20$ |
| $\beta_{406}(2)$ | $44.981$ | $10.05$ |
| $\gamma_{406}(1)$ | $231.090$ | $14.55$ |
| $\gamma_{406}(2)$ | $72.840$ | $9.05$ |
| $\alpha_{606}(1)$ | $-7.667$ | $0.78$ |
| $\alpha_{605}(2)$ | $-3.125$ | $1.34$ |
| $\beta_{605}(1)$ | $154.842$ | $7.27$ |
| $\beta_{605}(2)$ | $42.054$ | $6.96$ |
| $\gamma_{605}(1)$ | $260.489$ | $13.48$ |
| $\gamma_{605}(2)$ | $72.679$ | $7.40$ |

Table 8: Coefficients for the bending-function (Coons) approximation of the power characteristic for each of the three pump models 10**206**/B, 10**406**/B, 10**605**/B.

| Coefficient | Value | Error |
|---|---|---|
| $\alpha_{206}$ | $-10.509$ | $0.09$ |
| $\beta_{206}$ | $554.656$ | $0.34$ |
| $\gamma_{206}$ | $177.974$ | $0.62$ |
| $\delta_{206}$ | $-57.797$ | $0.46$ |
| $\epsilon_{206}$ | $-552.849$ | $5.40$ |
| $\zeta_{206}$ | $169.975$ | $2.13$ |
| $\alpha_{406}$ | $-9.462$ | $0.03$ |
| $\beta_{406}$ | $888.804$ | $4.60$ |
| $\gamma_{406}$ | $241.806$ | $0.45$ |
| $\delta_{406}$ | $-72.719$ | $0.33$ |
| $\epsilon_{406}$ | $-963.466$ | $7.29$ |
| $\zeta_{406}$ | $304.907$ | $2.89$ |
| $\alpha_{605}$ | $-6.729$ | $0.02$ |
| $\beta_{605}$ | $1034.160$ | $5.54$ |
| $\gamma_{605}$ | $208.700$ | $0.38$ |
| $\delta_{605}$ | $-60.132$ | $0.28$ |
| $\epsilon_{605}$ | $-1125.230$ | $8.73$ |
| $\zeta_{605}$ | $354.673$ | $3.43$ |

Table 9: Coefficients for quadratic approximation of power characteristic for each of the three pump models 10**206**/B, 10**406**/B, 10**605**/B. The coefficients for the pressure head characteristic are given by the cubic coefficients in Table 5 (left).

## REFERENCES

[1] DIN EN 806-3:2006-07 (and underlying DIN 1988-300):2012-05, *Technische Regeln f¨r Trinkwasser-Installationen - Teil 3: Berechnung der Rohrinnendruchmesser - Vereinfachte Verfahren.* Beuth Verlag, 2006.

[2] H. Martin, R. Pohl, *Technische Hydromechanik 4: Hydraulische und numerische Modelle.* Beuth Verlag, 2015.

[3] KSB Aktiengesellschaft, *Baureihenheft hya-solo*, https://shop.ksb.com/ims_docs/00/005056B602221ED4A9CD0BF2DE4AAB88.pdf, 04.09.2015 17:55.

[4] D. Aigner, D. Carstensen. *Technische Hydromechanik 2: Spezialfälle.* Volume 2, Beuth Verlag, 2013.

[5] Gurobi *Website Gurobi Optimiser.* http://www.gurobi.com/products/gurobi-optimizer 10.02.2016 09:10

[6] GNU *Website GLPK* https://www.gnu.org/software/glpk/ 10.02.2016 09:10

[7] Coin OR *Website.* http://www.coin-or.org/Couenne/ 10.02.2016 08:50

[8] Zuse Institue Berlin *Website SCIP* http://scip.zib.de/ 10.02.2016 09:10

[9] K. Weiss, L. De Floriani, *Simplex and diamond hierarchies: Models and applications.* Computer Graphics Forum, volume 30, pages 21272155, Wiley Online Library, 2011.

[10] S. Chatterjee, J. S. Simonoff. *Handbook of Regression Analysis* John Wiley & Sons, Inc, 2013.

[11] R. Forrest *On coons and other methods for the representation of curved surfaces.* Computer Graphics and Image Processing, 1(4):341359, 1972.

[12] I.N. Bronstein, K.A. Semendjajew, G. Musiol, H Mhlig. *Taschenbuch der Mathematik* 7. vollstndig berarbeitete und ergnzte Auflage. Verlag Harri Deutsch, 2008.

[13] G.K. M. Pedersen, Z. Yang. *Efficiency optimization of a multi-pump booster system.* Proceedings of the 10th annual conference on Genetic and evolutionary computation, pages 16111618. ACM, 2008.

Equation Chapter 1 Section 1ECCOMAS Congress 2016
Equation Chapter 1 Section 1 VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds. )
Crete Island, Greece, 5–10 June 2016

# RELIABILITY TEST DEMONSTRATION METHOD FOR EXPONENTIAL LIFE SYSTEM WITH RELIABILITY GROWTH UNDER THE CONDITION OF IN-TIME CORRECTIVE STRATEGY

**Yunyan Xing[1], Ping Jiang[1], Feng Yao[1], Zhiwei Yang[1,2], Michael Emmerich[1] and Thomas Bäck[2]**

[1] National Univeristy of Defence Techonology
Deya road, No.109, Kaifu district, Changsha, Hunan, China
e-mail: xingyy2005@sina.com,jiangping@nudt.edu.cn,yf_nudt@126.com

[2] Leiden Institute of Advanced Computer Science,
Niels Bohrweg 1, Leiden, 2333CA, the Netherlands
{z.yang,m.t.m.emmerich,t.h.w.baeck}@liacs.leidenuniv.nl

**Keywords:** Bayesian methods, reliability growth, exponential life system, In-Time corrective strategy, hypothesis testing.

**Abstract.** *The development process of weapon system usually composes of multiple test stages. Especially for expensive weapon system which test is destructive, engineers often apply In-Time corrective strategy to analyze and modify failures exposed during development. That is to say once failure is observed at each stage, test is then stopped and a design modification is introduced to remove the cause of the observed failure, and thus release improved system to new test stage until the system reliability level satisfy the predefined development requirement. In this paper, we present a Bayesian statistical decision method to judge whether system reliability level meet development requirement or not, based on multiple stages test data during development of exponential life system and taking In-Time corrective strategy into account. Firstly, the reliability growth model, derived from non-homogeneous Poisson process, under the condition of In-Time corrective strategy is presented. Then prior distribution of system reliability index is introduced by applying reliability growth model to incorporate various test data from development process, and the Bayesian hypothesis testing method is given. Finally, the numerical example validates effectivity of the method compared with traditional methods.*

## INTRODUCTION

For most complex repairable system, the first prototype usually contains design and/or engineering flaws, thus the system has to experience an iterative process of TAAF to eliminate the observed failure modes and then achieve the reliability growth. Reliability growth model is an effective means to analyze the growth trend of product reliability, and many scholars has made lots of research. Fries [1], [2] presented a discrete reliability growth model, derived from the learning-curve property, to describe reliability growth. Xing [3] presented a discrete reliability growth model base on In-Time correction strategy. Hall & Mosleh [4] introduced a framework for the evaluation of reliability growth for one-shot systems. Pulcini [5], [6], [7] proposed an exponential reliability growth model where several identical copies of equipment are put on test, once a failure is observed, design modification for all copies of equipment is introduced. In addition, Bayesian reliability growth models are very popular in describing the reliability growth trend of product [8], [9], [10], [11]. However, those reliability growth models are usually used to evaluate or predict product reliability index, little effort focused on hypothesis testing for whether the product reliability index has reached the predefined requirement during development.

As for weapon equipment, development process generally consist of multiple test stages. Especially for expensive system which test is destructive, test is stopped and design modifications are made immediately once failure is observed at each stage, such a corrective strategy for failure mode is often referred to as In-Time corrective strategy. While under the condition of In-Time corrective strategy, how to describe the reliability growth trend of system and how to judge whether the system has achieved the predefined reliability goal are prominent problems that reliability researchers concern about. So we make following four assumptions:

(i) System development process composes of multiple test stages, test data from various stages are independent.

(ii) In-Time corrective strategy is adopted to fix observed system failure at each test stage.

(iii) Test data at the ith $(i=1,2,\cdots,m)$ stage is $(1,\tau_i)$. Where 1 is the number of failure at the ith stage. $\tau_i$ denotes the time on test, and it is r. v. .

(iv) System reliability level at the ith stage is changeless, and system life $t$ conform to exponential distribution with parameter $\lambda_i$, i. e.

$$f(t|\lambda_i)=\lambda_i e^{-\lambda_i t},\quad \lambda_i,t>0,\quad i=1,2,\cdots \tag{1}$$

Where $\lambda_i$ is the system failure rate at the ith stage.

The paper is organized as follows. A reliability growth model under the condition of In-Time corrective strategy is presented In Section 1. The Bayesian hypothesis testing method is introduced in Section 2. A numerical example is given in Section3. Section 4 summarizes the paper.

## 1 RELIABILITY GROWTH MODEL UNDER THE CONDITION OF IN-TIME CORRECTIVE STRATEGY

Crow proposed AMSAA model which thought the cumulative number of observed failures $K(t)$ of a repairable system in the development duration $(0,t]$ follows the non-homogeneous Poisson process (NHPP) with the expectation of $K(t)$ as $E[K(t)]=ct^d$ and the probability is

$$P\{K(t)=n\}=\frac{(ct^d)^n}{n!}e^{-ct^d},\quad n=0,1,2,\cdots \tag{2}$$

Where parameters $c,d$ denote shape parameter and scale parameter respectively. When $0<d<1$, system reliability is increased. $1-d$ denotes growth rate.

In practice, people are more concerned with analysis on time on test $T(n)$ when $n$ failures happen, especially for analysis on the expectation of time on test $E[T(n)]$. It is obvious that $n$ failures happen before moment $t$ equals if and only if the number of failures until $t$ is more than $n$, namely,

$$K(t)\geq n \Leftrightarrow T(n)\leq t \tag{3}$$

So the distribution function $F_{T(n)}(t)$ of $T(n)$ is

$$F_{T(n)}(t)=P\{T(n)\leq t\}=P\{K(t)\geq n\}=\sum_{j=n}^{\infty}e^{-ct^d}\frac{(ct^d)^j}{j!} \tag{4}$$

Thus the probability density function $f_{T(n)}(t)$ of $T(n)$ is obtained by computing the derivative of $t$ from (4) as

$$f_{T(n)}(t)=e^{-ct^d}\frac{(ct^d)^{n-1}}{(n-1)!}cdt^{d-1} \tag{5}$$

So the expectation of time on test $E[T(n)]$ when $n$ failures happen is

$$E[T(n)]=\int_0^{\infty}tf_{T(n)}(t)\,dt=\frac{d}{(n-1)!}\int_0^{\infty}(ct^d)^n e^{-ct^d}dt \tag{6}$$

Let $y=ct^d$, we have $t=\left(\frac{y}{c}\right)^{\frac{1}{d}}$, hence $dt=\frac{1}{c^{\frac{1}{d}}b}y^{\frac{1}{d}-1}dy$, (6) can be transformed into

$$E[T(n)]=\frac{\Gamma\left(n+\frac{1}{d}\right)}{c^{\frac{1}{d}}\Gamma(n)} \tag{7}$$

Where $\Gamma(n)=\int_0^{\infty}x^{n-1}e^{-x}dx=(n-1)!$。

That is to say the expectation of time on test when $n$ failures happen in the NHPP is

$$E[T(n)]=\frac{\Gamma\left(n+\frac{1}{d}\right)}{c^{\frac{1}{d}}\Gamma(n)} \tag{8}$$

As above mentioned that development process of system consists of several stages with respective tests, and modification of design is introduced once failure is observed at each stage, so the reliability growth model should reflect the growth character of step-wise jump.

From (8), the relationship between the cumulative time on test $T(n_i)$ and the cumulative number of failures $n_i$ until the ith stage can be expressed as

$$E\left[T\left(n_i\right)\right]=\frac{\Gamma\left(n_i+\frac{1}{d}\right)}{c^{\frac{1}{d}}\Gamma\left(n_i\right)} \tag{9}$$

According to the assumption (iii), the cumulative number of observed failure $n_i$ till the ith stage is

$$n_i=i \tag{10}$$

From (9) and (10), the expectation of time on test till the ith stage is

$$E\left[T\left(i\right)\right]=\frac{\Gamma\left(i+\frac{1}{d}\right)}{c^{\frac{1}{d}}\Gamma\left(i\right)} \tag{11}$$

Let $a=c^{\frac{1}{d}}$, $b=\frac{1}{d}$, then (11) can be rewritten as

$$E\left[T\left(i\right)\right]=\frac{\Gamma\left(i+b\right)}{a\Gamma\left(i\right)} \tag{12}$$

Similarly, we have

$$E\left[T\left(i-1\right)\right]=\frac{\Gamma\left(i-1+b\right)}{a\Gamma\left(i-1\right)} \tag{13}$$

By referring to (12) and (13), the expectation of time on test at the ith stage can be obtained by

$$\tau_i=E\left[T\left(i\right)\right]-E\left[T\left(i-1\right)\right]=a^{-1}\left[\frac{\Gamma\left(i+b\right)}{\Gamma\left(i\right)}-\frac{\Gamma\left(i-1+b\right)}{\Gamma\left(i-1\right)}\right] \tag{14}$$

From (14), the reliability growth model under the condition of In-Time corrective strategy is formulated as

$$\lambda_i=\frac{a}{\frac{\Gamma\left(i+b\right)}{\Gamma\left(i\right)}-\frac{\Gamma\left(i-1+b\right)}{\Gamma\left(i-1\right)}} \tag{15}$$

Where $\lambda_i$ denotes system failure rate at the ith stage. $b>1$ is required to guarantee the reliability growth, and $a>0$.

The point estimation $\hat{a}$ and $\hat{b}$ are obtained by maximum likelihood (ML) procedures, the ML function is

$$L\left(a,b\right)=\prod_{i=1}^{m}\lambda_i e^{-\lambda_i\tau_i} \tag{16}$$

Where $m$ denotes the number of test stages, $\tau_i$ denotes the cumulative time on test at the ith stage, and $\lambda_i\left(i=1,2,\cdots,m\right)$ is given in model (15).

## 2 BAYESIAN HYPOTHESIS TESTING METHOD

The objective of this paper is to construct a Bayesian hypothesis testing method to judge whether system reliability index has achieved the predefined requirement by applying reliability growth model introduced in Section 1 to incorporate test data from diverse development stages. The prior distribution is the premise to perform Bayesian statistical analysis.

### 2.1 The prior distribution of $\lambda_i$

Let $(1,\tau_1),(1,\tau_2),\cdots,(1,\tau_{i-1})$ denote test data from $i-1$ test stages respectively, according to reliability growth model (15), we can obtained point estimation $\hat{a},\hat{b}$ by calculating (16). Then the prediction estimation of failure rate $\lambda_i$ is

$$\hat{\lambda}_{i,0}=\frac{\hat{a}}{\dfrac{\Gamma\left(i+\hat{b}\right)}{\Gamma(i)}-\dfrac{\Gamma\left(i-1+\hat{b}\right)}{\Gamma(i-1)}} \tag{17}$$

Let $\hat{\lambda}_{i,0}$ be the prior mean of the failure rate $\lambda_i$ at ith stage, i. e.

$$E(\lambda_i)=\hat{\lambda}_{i,0} \tag{18}$$

According to the maximum entropy method, the prior distribution $\pi(\lambda_i)$ can de expressed as

$$\pi(\lambda_i)=\frac{e^{\mu\lambda_i}}{\int_0^{+\infty}e^{\mu\lambda_i}d\lambda_i} \tag{19}$$

Where $\mu$ is a parameter to be determined.

By referring to (18) and (19), the value of $\mu$ can be acquired by solving the following equation

$$E(\lambda_i)=\frac{\int_0^{+\infty}\lambda_i e^{\mu\lambda_i}d\lambda_i}{\int_0^{+\infty}e^{\mu\lambda_i}d\lambda_i}=\hat{\lambda}_{i,0} \tag{20}$$

After the unique parameter $\mu$ is determined, we can obtain the prior distribution $\pi(\lambda_i)$ of $\lambda_i$. Hence the prior second moment of $\lambda_i$ can be obtained by

$$E(\lambda_i^2)=\frac{\int_0^{+\infty}\lambda_i^2 e^{\mu\lambda_i}d\lambda_i}{\int_0^{+\infty}e^{\mu\lambda_i}d\lambda_i} \tag{21}$$

As conjugate prior distribution has the advantages of tractability in mathematical computation and practical meanings for its parameters, we use Gamma distribution $Ga(a_{i,0},b_{i,0})$ as the approximate distribution to fit the prior distribution $\pi(\lambda_i)$. Thus both $Ga(a_{i,0},b_{i,0})$ and $\pi(\lambda_i)$ must have the same prior first and second moments. Hence

$$\begin{cases}\dfrac{a_{i,0}}{b_{i,0}}=E(\lambda_i)\\[2ex]\dfrac{a_{i,0}(a_{i,0}+1)}{b_{i,0}^2}=E(\lambda_i^2)\end{cases} \tag{22}$$

By solving above equations, we get

$$\begin{cases} a_{i,0} = \dfrac{\left[E(\lambda_i)\right]^2}{E(\lambda_i^2) - \left[E(\lambda_i)\right]^2} \\ b_{i,0} = \dfrac{E(\lambda_i)}{E(\lambda_i^2) - \left[E(\lambda_i)\right]^2} \end{cases} \tag{23}$$

Thus, the prior distribution $Ga(a_{i,0}, b_{i,0})$ of $\lambda_i$ at ith stage is obtained by applying the reliability growth model (15) to incorporate test data from $i-1$ development stages, where $a_{i,0}, b_{i,0}$ are given by (23).

## 2.2 Bayesian hypothesis testing method

Let $\theta$ denotes system MTBF. $\theta_U$ is the upper limit (if $\theta$ surrounds upper limit closely, products will be accepted by high probability), $\theta_L$ is the lower limit (if $\theta$ surrounds lower limit closely, products will be rejected by high probability), $\theta_U$ and $\theta_L$ are given in development contract, and $\theta_U > \theta_L$. $\alpha_{\pi_0}$ denotes producer risk, $\beta_{\pi_1}$ denotes consumer risk. The hypothesis testing problem need to be researched is to judge that $\theta_i$ satisfies $\theta_U$ or $\theta_L$ when development process has come to the ith stage and takes In-Time corrective strategy into account, by utilizing test data from $i$ stages under the condition of given $\alpha_{\pi_0}, \beta_{\pi_1}, \theta_U, \theta_L$. So the hypotheses can be formulated as

$$H_0 : \theta_i = \theta_U, \quad H_1 : \theta_i = \theta_L \tag{24}$$

Where $\theta_i$ denotes system MTBF at the ith stage.

To convenient calculation, the hypothesis testing about MTBF can be transformed into hypothesis testing about failure rate. So (24) is equivalent to

$$H_0 : \lambda_i = \lambda_L, \quad H_1 : \lambda_i = \lambda_U \tag{25}$$

Where $\lambda_i$ denotes system failure rate at the ith stage, $\lambda_L = 1/\theta_U$, $\lambda_U = 1/\theta_L$, $\lambda_L < \lambda_U$.

According to the Bayesian principle, the prior probability of null hypothesis $H_0$ from (25) is

$$P(H_0) \triangleq \pi_0 = P\left\{\lambda_L \left| (a_{i,0}, b_{i,0}) \right.\right\} = \frac{P(\lambda_L) P\left\{(a_{i,0}, b_{i,0}) \left| \lambda_L \right.\right\}}{P(\lambda_L) P\left\{(a_{i,0}, b_{i,0}) \left| \lambda_L \right.\right\} + P(\lambda_U) P\left\{(a_{i,0}, b_{i,0}) \left| \lambda_U \right.\right\}} \tag{26}$$

Where $P(\lambda_L), P(\lambda_U)$ donote prior probability of $\lambda_L, \lambda_U$ respectively.

Here let $P(\lambda_L) = P(\lambda_U) = 0.5$, thus the above prior probability of $H_0$ is

$$\begin{aligned}\pi_0 &= \frac{P\left\{\left(a_{i,0},b_{i,0}\right)\middle|\lambda_L\right\}}{P\left\{\left(a_{i,0},b_{i,0}\right)\middle|\lambda_L\right\}+P\left\{\left(a_{i,0},b_{i,0}\right)\middle|\lambda_U\right\}} \\ &= \frac{\frac{1}{a_{i,0}!}\left(\lambda_L a_{,0}\right)^{a_{i,0}}\mathrm{e}^{-\lambda_L b_{i,0}}}{\frac{1}{a_{i,0}!}\left(\lambda_L b_{i,0}\right)^{a_{i,0}}\mathrm{e}^{-\lambda_L b_{i,0}}+\frac{1}{a_{i,0}!}\left(\lambda_U b_{i,0}\right)^{a_{i,0}}\mathrm{e}^{-\lambda_U b_{i,0}}} \\ &= \frac{1}{1+\left(\frac{\lambda_U}{\lambda_L}\right)^{a_{i,0}}\mathrm{e}^{(\lambda_L-\lambda_U)b_{i,0}}}\end{aligned} \tag{27}$$

So the prior probability of alternative hypothesis $H_1$ is

$$P\left(H_1\right) \triangleq \pi_1 = 1-\pi_0 \tag{28}$$

From the assumption (iii), the likelihood function of test data from $i$ stages is

$$L\left\{\left(1,\tau_i\right)\middle|\lambda_i\right\} = \lambda_i e^{-\lambda_i\tau_i} \tag{29}$$

So the posterior probability ratio of hypothesis testing (25) is

$$\begin{aligned}O_i &= \frac{P\left\{H_1\middle|\left(1_i,\tau_i\right)\right\}}{P\left\{H_0\middle|\left(1_i,\tau_i\right)\right\}} = \frac{P\left(H_1\right)L\left\{\left(1,\tau_i\right)\middle|\lambda_U\right\}}{P\left(H_0\right)L\left\{\left(1,\tau_i\right)\middle|\lambda_L\right\}} \\ &= \frac{\pi_1}{\pi_0}\frac{\lambda_U e^{-\lambda_U\tau_i}}{\lambda_L e^{-\lambda_L\tau_i}} = \frac{\pi_1}{\pi_0}\frac{\lambda_U}{\lambda_L}\exp\left[\tau_i\left(\lambda_L-\lambda_U\right)\right]\end{aligned} \tag{30}$$

To perform the hypothesis testing, a decision threshold $A_i$ is introduced here, and the decision criteria are as follows:

(i) If $O_i \le A_i$ accept null hypothesis $H_0$;

(ii) If $O_i > A_i$, accept alternative hypothesis $H_1$.

The successive question we have to solve is to calculate $A_i$. It is appropriate to consider computation of the two types of risks for statistical testing according to the above decision criteria before determining $A_i$. The type I risk, namely producer risk $\alpha_{\pi_0}$ is

$$\alpha_{\pi_0} = P\left(O_i > A_i\middle|\lambda_i=\lambda_L\right)P\left(H_0\right) \tag{31}$$

As for $P\left(O_i > A_i\middle|\lambda_i=\lambda_L\right)$, we have

$$\begin{aligned}&P\left(O_i > A_i\middle|\lambda_i=\lambda_L\right) \\ &= P\left\{\frac{\pi_1}{\pi_0}\frac{\lambda_U}{\lambda_L}\exp\left[\tau_i\left(\lambda_L-\lambda_U\right)\right] > A_i\middle|\lambda_i=\lambda_L\right\} \\ &= P\left\{\exp\left[\tau_i\left(\lambda_L-\lambda_U\right)\right] > A_i\frac{\pi_0}{\pi_1}\frac{\lambda_L}{\lambda_U}\middle|\lambda_i=\lambda_L\right\} \\ &= P\left\{\tau_i\left(\lambda_L-\lambda_U\right) > \ln A_i + \ln\frac{\pi_0}{\pi_1} + \ln\frac{\lambda_L}{\lambda_U}\middle|\lambda_i=\lambda_L\right\}\end{aligned} \tag{32}$$

Because $\lambda_L < \lambda_U$, so (32) can be transformed into

$$P\left(O_i > A_i \middle| \lambda_i = \lambda_L\right) = P\left\{\tau_i < \frac{1}{\lambda_L - \lambda_U}\left(\ln A_i + \ln\frac{\pi_0}{\pi_1} + \ln\frac{\lambda_L}{\lambda_U}\right) \middle| \lambda_i = \lambda_L\right\} \tag{33}$$

According to the assumption (iv), we know time on test $\tau_i$ conforms to Gamma distribution $Ga\left(1_i, \lambda_i\right)$, so we have

$$P\left(O_i > A_i \middle| \lambda_i = \lambda_L\right) = \lambda_L \int_0^{\frac{1}{\lambda_L - \lambda_U}\left(\ln A_i + \ln\frac{\pi_0}{\pi_1} + \ln\frac{\lambda_L}{\lambda_U}\right)} e^{-\lambda_L x} dx \tag{34}$$

From (31) and (34), the producer risk can be obtained as

$$\alpha_{\pi_0} = \pi_0 \lambda_L \int_0^{\frac{1}{\lambda_L - \lambda_U}\left(\ln A_i + \ln\frac{\pi_0}{\pi_1} + \ln\frac{\lambda_L}{\lambda_U}\right)} e^{-\lambda_L x} dx \tag{35}$$

Similarly, type II risk, namely consumer risk is

$$\begin{aligned} \beta_{\pi_1} &= P\left(O_i \le A_i \middle| \lambda_i = \lambda_U\right) P\left(H_1\right) \\ &= \pi_1\left[1 - P\left(O_i > A_i \middle| \lambda_i = \lambda_U\right)\right] \\ &= \pi_1 - \pi_1 \lambda_U \int_0^{\frac{1}{\lambda_L - \lambda_U}\left(\ln A_i + \ln\frac{\pi_0}{\pi_1} + \ln\frac{\lambda_L}{\lambda_U}\right)} e^{-\lambda_U x} dx \end{aligned} \tag{36}$$

From (35) and (36), we can find that $\alpha_{\pi_0}, \beta_{\pi_1}$ are both functions on $A_i$, and $\alpha_{\pi_0}$ is the non-increasing function on $A_i$, while $\beta_{\pi_1}$ is the non-decreasing function on $A_i$. Therefore, given $\alpha_{\pi_0}$, there exists $A_{i,0}$ that satisfies $\alpha_{\pi_0}\left(A_{i,0}\right) = \alpha_{\pi_0}$. At the same time, for any $A_i$ within the interval $\left[A_{i,0}, \infty\right), \alpha_{\pi_0}\left(A_i\right) \le \alpha_{\pi_0}$ holds, where $\alpha_{\pi_0}\left(A_i\right)$ denotes the probability of type I risk when the decision threshold is $A_i$. Similarly, there is $A_i$ which satisfies $\beta_{\pi_1}\left(A_{i,1}\right) = \beta_{\pi_1}$ when $\beta_{\pi_1}$ is given, and for any $A_i$ within the interval $\left(0, A_{i,1}\right]$, $\beta_{\pi_1}\left(A_i\right) \le \beta_{\pi_1}$ holds. Where $\beta_{\pi_1}\left(A_i\right)$ denotes the probability of type II risk when the decision threshold is $A_i$. Given the probabilities of the two types of risks，the decision threshold $A_i$ will be determined by the following procedure:

(i) if $A_{i,0} \le A_{i,1}$, any $A_i \in \left[A_{i,0}, A_{i,1}\right]$ can be chosen as the decision threshold. The real probabilities of the two types of risks are $\alpha'_{\pi_0}\left(A_i\right), \beta'_{\pi_1}\left(A_i\right)$ respectively.

(ii) if $A_{i,0} > A_{i,1}$, $A_i$ can be determined by consulting the producer and consumer. When we choose $A_i$ that meets the requirement of (35), it should also minimize the value of (36), so the decision threshold can be chosen as $A_i = A_{i,0}$. In the same manner, when we choose $A_i$ which meets the requirement of (36), it should also minimize the value of (35), so the decision threshold can be chosen as $A_i = A_{i,1}$.

To sum up, the Bayesian hypothesis testing procedure for system failure rate $\lambda_i$ at the ith stage when system development process takes In-Time corrective strategy into account is shown as the following steps:

Step1 Obtain prior distribution of $\lambda_i$ by applying reliability growth model (15) to incorporate test data from previous $i-1$ test stages.

Step2 Calculate the posterior probability ratio $O_i$ and decision threshold $A_i$.

Step3 Make a decision according to the criteria:

(i) If $O_i \le A_i$, terminate the development process and accept null hypothesis $H_0$.

(ii) If $O_i > A_i$, accept alternative hypothesis $H_1$, improve system design and release improved system into new test stage until system failure rate has reached upper limit.

## 3 NUMERICAL EXAMPLE

A system experiences a series of test stages during development and its lifetime follows exponential distribution. The upper limit of system MTBF is 40 hours, the lower limit of system MTBF is 32 hours, and the producer risk and consumer risk are both set to be 0. 2. The problem is to judge which requirement the system MTBF has reached.

The test data is in Table 1 which lists the simulated test data derived by sampling from the NHPP when the truth values of the MTBF are known.

Table 1 test data

| stage | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Truth value of MTBF | 8. 9551 | 16. 2524 | 22. 8743 | 29. 0875 | 35. 0133 |
| Time on test(h) | 8. 3278 | 15. 7918 | 21. 4445 | 34. 8337 | 35. 9414 |

To simplify the computation, the question of above hypothesis testing for the MTBF can be transformed into the testing for the failure rate, namely, we will accept system by high probability when failure rate approaches to $\lambda_L = 0.025$; otherwise we will reject system by high probability when failure rate approaches to $\lambda_U = 0.0313$. According to test data from previous 4 stages, by referring to (16), we can obtain the ML estimate of reliability growth model parameters $\hat{a} = 0.1892, \hat{b} = 1.8125$, and the prediction estimate of the failure rate $\lambda_5$ is $\hat{\lambda}_{5,0} = 0.0287$ through (17). So the prior probability of null hypothesis is $\pi_0 = 0.4456$ by referring to (18)-(27), and the posterior probability ratio is $O_5 = 1.2420$ by referring to (30). The corresponding decision threshold are $A_{5,0} = 1.1461, A_{5,1} = 1.0765$ respectively when two types of risks are both 0. 2. As $A_{5,0} > A_{5,1}$, we choose the decision threshold as $A_5 = 1.1461$ after consulting between consumer and producer, and the truth probabilities of the two types of risks are $\alpha'_{\pi_0} = 0.2, \beta'_{\pi_1} = 0.2814$ ,respectively. Since $O_5 > A_5$, we reject system and demand design modification before starting the next test stage. After improving the system design and performing the sixth stage test, the observed test data is shown in Table2. Now we continue to determine which requirment the improved system MTBF has achieved.

Table2 test data at the 6th stage

| stage | Truth value of MTBF | Time on test(h) | The number of failure |
|---|---|---|---|
| 6 | 40. 7196 | 42. 0064 | 1 |

Same with above calculation procedure, the ML estimate of reliability growth model parameters are $\hat{a} = 0.2311, \hat{b} = 1.8868$ according to test data from previous 5 stages, and the prediction estimate of the failure rate $\lambda_6$ is $\hat{\lambda}_{6,0} = 0.0252$. So the prior probability of null hypothesis is $\pi_0 = 0.5063$, and the posterior probability ratio is $O_6 = 0.9370$. The corresponding decision threshold are $A_{6,0} = 1.1549, A_{6,1} = 0.9468$ respectively when the probabilities of the

two types of risks are both 0. 2. As $A_{6,0} > A_{6,1}$ , we choose the decision threshold as $A_6 = 1.1549$, and the truth probabilities of the two types of risks are $\alpha'_{\pi_0} = 0.2, \beta'_{\pi_1} = 0.2814$, respectively. Since $O_6 < A_6$, the system can be accepted and the development process can be terminated. Actually, the truth value of MTBF (40. 7196) at the 6th test stage indeed more approached to upper limit, so the decision results concluded from our methods coincide with the real condition.

## 4 CONCLUSION

- A reliability growth model to describe reliability growth trend of system under the condition of In-Time corrective strategy is proposed, which well describes the features of stepwise trend of system reliability growth.
- A reliability test demonstration method is presented, which especially adapts to expensive system with destructive test when development process takes In-Time corrective strategy to modify failure. The method makes use of test data from development stages, offsets the absence of information, and effectively enhances the credibility of statistical decision.

## REFERENCES

[1] A. Fries. Discrete reliability-growth models based on learning-curve property. IEEE Trans. on Reliability, 1993, 42(21): 303-306.

[2] A. Fries, A. Sen. A survey of discrete reliability-growth models. IEEE Trans. on Reliability, 1996, 45(4): 582-604.

[3] Y. Y. Xing, X. Y. Wu, P. Jiang, Q. Liu. Dynamic Bayesian evaluation method for system reliability growth based on In-Time correction. IEEE Trans. on Reliability, 2010, 59(2): 309-312.

[4] J. B. HaII, A. Mosleh. An analytical framework for reliability growth of one-shot systems. Reliability Engineering and System Safety, 2008, 93: 1751-1760.

[5] G. Pulcini. An exponential reliability-growth model in multicopy testing program. IEEE Trans. on Reliability, 2001, 50(4): 365-373.

[6] M. Giorgio, M. Guida, G. Pulcini. Reliability-growth analysis of locomotive electrical equipment. Journal of quality technology, 2006, 38(1):14-29.

[7] G. Pulcini. Correction to “how to model reliability-growth when times of design modifications are known”. IEEE Trans. on Reliability, 2002, 51(2): 252-253.

[8] R. Calabfia, M. Guida, G. Pulcini. A Bayes procedure for estimation of current system reliability. IEEE Trans. on Reliability, 1992, 41(4): 616-620.

[9] R. Calabria, M. Guida, G. Pulcini. A reliability-growth mode in a Bayes-decision framework. IEEE Trans. on Reliability, 1996, 45(3): 505-510.

[10] Z. Q. Yan, X. X. Li, H. W. Xie, et al. Bayesian synthetic evaluation of multistage reliability growth with instant and delay fix modes. Journal of systems engineering and electronics, 2008, 19(6):1287-1294.

[11] Y. Y. Xing, X. Y. Wu. Bayesian sequential testing for exponential life system with reliability growth. Journal of systems engineering and electronics, 2011, 22(6):1023-1029.

This work was supported by the National Science Foundation of China (71401170, 71371182) and the Research Fund for the Doctoral Program of Higher Education of China (20134307120026 ).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EVALUATION OF LINE-FITTING METHOD OF MODEL ORDER REDUCTION

**V. Makhavikou[1], R. Kasper[1], and D. Vlasenko[2]**

[1] Otto von Guericke University
Universitaetsplatz, 2, 39106 Magdeburg, Germany
{valery.makhavikou, roland.kasper}@ovgu.de

[2] Schaeffler Technologies AG & Co. KG
Industriestrasse 1-3, 91074 Herzogenaurach, Germany
dmitry.vlasenko@schaeffler.com

**Keywords:** Elastic multibody systems, Floating frame of reference, Model order reduction.

**Abstract.** *Computer aided simulation is an important part of development of modern technical products. Many modern technical systems operate at high speeds and include lightweight components. These conditions usually lead to appearance of deformation effects that must be taken into account in processes of system modeling and simulation.*

*In this contribution the dynamics of elastic multibody systems is formulated using the floating frame approach and the finite element method. In many applications a large number of elastic coordinates have to be employed to properly describe body deformations. This results in large simulation costs and memory deficiency problems. The problem can be solved using model order reduction techniques, which approximate a large number of equations by a much smaller number of equations that keep important dynamic properties of the original system.*

*This work is focused on the evaluation of the line-fitting method that is a linear structure preserving method of model order reduction for elastic multibody simulations. The method is based on the approximation of transfer functions and enables tuning of a reduced order model for certain transfer functions and certain frequency ranges. Line-fitting reduced order models possess such desirable properties as high accuracy, preservation of stability, low order and stiffness.*

*We apply the line-fitting method to a model of elastic bar, analyze properties of a reduced order model, and compare the results with results of the powerful SEREP reduction approach.*

# 1 INTRODUCTION

## 1.1 Research field

The subject of this research relates to simulation of elastic multibody systems (EMBS). The term EMBS denotes a system of rigid and elastic bodies interconnected by joints or coupling elements, where the bodies may undergo large rigid body motions and elastic deformations. Under elastic deformations a body returns to its initial state after applied loads are removed. In particular, in this article we focus on linear elastic deformations. It means that we assume small deformations and linear relationship between deformations and applied forces. EMBS appear in applications of many engineering fields: robotics, biomechanics, and vehicle and aircraft dynamics.

Modeling of EMBS is based on methods of multibody system dynamics and the theory of elasticity. The most efficient way to describe the dynamics of EMBS undergoing linear deformations is a floating frame formulation [1]. According to this method the total motion of elastic body is divided into two parts: rigid body motion represented by the motion of body reference frame and deformations with respect to this frame.

Linear elastic deformations are usually modeled by the finite element method which results in a system of ODEs. The problem is that in many cases, the system of equations is extremely high dimensional, and its solution on standard computers is either not feasible or it suffers from vast computational time and memory deficiency problems. The problem can be solved by model order reduction methods. These methods approximate the large set of equations by a small set of equations that keep important dynamic properties of the original system.

## 1.2 Past studies

Over the last decades a variety of reduction techniques have been developed. The set of classical reduction approaches applied in elastic multibody dynamics consists of the methods based on modal truncation, condensation [2, 3], and component mode synthesis [4, 5, 6]. The traditional reduction methods also include a System Equivalent Reduction Expansion Process (SEREP), which was introduced in [7] and extensively used in structural and multibody mechanics. The classical approaches are implemented in many simulation software and remain state of the art techniques for model order reduction.

More recently, a few alternative reduction methods have come from the field of control theory, namely, techniques based on the singular value decomposition (SVD) using Gramian matrices and moment matching via Krylov subspaces [8, 9, 10, 11]. The methods are aimed at the approximation of input-output behavior of dynamical systems. Nowadays, the application of these techniques in the context of EMBS is on the focus of intensive research.

The comprehensive description of basic model order reduction methods can be found in the books [8, 12]. Relative advantages and disadvantages of widespread reduction approaches are presented in [13, 14].

## 1.3 Problem statement

In this article we examine a recently developed line-fitting method of model order reduction. The approach was originally proposed in [15]. In addition, the method was tested on academic and industrial application examples and compared with the widely used Craig-Bampton method in [16, 17]. The line-fitting approach is based on an approximation of transfer functions of original finite element model and it enables tuning of a reduced order model for certain transfer functions and certain frequency ranges. In [16, 17] line-fitting reduced or-

der models showed better results in comparison with Craig-Bampton models in the accuracy, order, and stiffness.

The limitation of past studies of the line-fitting method consists in the fact that they focus primarily on the comparison with the Craig-Bampton method. The aim of this contribution is to evaluate the line-fitting method relative to the powerful SEREP approach. Despite the SEREP method is not a typical approach for the approximation of input-output behavior of a system, it can be applied for such applications and provide high quality reduced order models.

The relative comparison of promising reduction approaches is an important task because this contributes to further improvements of methods and to an appropriate choice of methods for certain applications examples.

### 1.4 Main steps and outline

In order to evaluate reduction approaches, it is necessary to define evaluation parameters, methods, and criteria. Application examples often make the following demands on the reduction approaches:

- high accuracy of reduced model,
- fast simulation of reduced model,
- simple tuning of reduced model for certain transfer functions and certain frequency ranges,
- preservation of model stability,
- preservation of second order structure of equations of motion,
- computational efficiency of reduction method.

We evaluate accuracy of reduced order models using a relative error of natural frequencies, correlation of eigenvectors, and a relative error for system transfer functions. In order to assess simulation speed, we compare an order of reduced models and the largest stiffness. The difficulty of tuning of reduced order model can be assessed using the time needed for the achievement of predefined accuracy demands. Preservation of stability and preservation of second order structure of equations of motion follow from theoretical descriptions of the methods. Finally, we determine computational efficiency of the reduction approaches by analyzing their computational complexity.

Theoretical description of methods allows partial evaluating of their properties. Complete assessment requires application of methods to numerical examples with subsequent analysis of reduced order models.

The structure of this article is as follows. We begin with a peace of theoretical information about equations of motion for EMBS and about the model order reduction. After that we describe SEREP and line-fitting approaches and identify their strength and weaknesses from the theoretical point of view. Next, we apply both methods to a model of elastic bar, validate reduced order models, and compare the results. Finally, we summarize and generalize the obtained results and discuss further work on the subject.

## 2 FUNDAMENTALS OF ELASTIC MULTIBODY DYNAMICS

### 2.1 Modeling of multibody dynamics

The equations of motion of a single unconstrained body can be derived using Jordain's principle. This leads to the following representation:

$$\begin{bmatrix} \mathbf{M}_r^i & \mathbf{M}_{re}^i \\ \mathbf{M}_{er}^i & \mathbf{M}_e^i \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{q}}_r^i \\ \ddot{\mathbf{q}}_e^i \end{Bmatrix} + \begin{Bmatrix} \mathbf{0} \\ \mathbf{K}_e^i \cdot \mathbf{q}_e^i + \mathbf{D}_e^i \cdot \dot{\mathbf{q}}_e^i \end{Bmatrix} = \begin{Bmatrix} \mathbf{h}_r^i \\ \mathbf{h}_e^i \end{Bmatrix}. \tag{1}$$

Here the superscript $i$ refers to a number of body. The vector of coordinates consists of the rigid body coordinates $\mathbf{q}_r$ and the elastic coordinates $\mathbf{q}_e$. The vector $\mathbf{q}_r$ includes 3 translational and 3 rotational coordinates; the term $\mathbf{q}_e$ has a length $N$ equal to a number of degrees of freedom of the finite element model. The mass matrix includes a rigid body part $\mathbf{M}_r \in \mathbb{R}^{6\times 6}$, an elastic part $\mathbf{M}_e \in \mathbb{R}^{N\times N}$, and coupling parts $\mathbf{M}_{re} = \mathbf{M}_{er}^T$. The matrices $\mathbf{K}_e$ and $\mathbf{D}_e$ represent stiffness and damping matrices of the elastic body. The Coriolis, centrifugal, as well as external forces are summarized in the force vector $\mathbf{h}$.

The equation of motion of the whole EMBS can be written as

$$\mathbf{M}(\mathbf{q})\cdot\ddot{\mathbf{q}} + \mathbf{D}\cdot\dot{\mathbf{q}} + \mathbf{K}\cdot\mathbf{q} = \mathbf{g}(\dot{\mathbf{q}},\mathbf{q},t), \tag{2}$$

where $\mathbf{q}$ is a vector of generalized coordinates of the whole system, $\mathbf{g}$ includes the vector $\mathbf{h}$ and the vector of constraint forces. More detailed description of EMBS dynamics can be found in [18, 19].

## 2.2 Modeling of elastic bodies

In this subsection we focus on the concept of model order reduction. Let $\mathbf{u}(t)$ and $\mathbf{y}(t)$ be an input distribution and an output measurement arrays, respectively. The elastic part of the equation of motion (1) is transformed into a linear time invariant second order multi-input and multi-output (MIMO) system

$$\begin{aligned} &\mathbf{M}_e\cdot\ddot{\mathbf{q}}_e(t) + \mathbf{D}_e\cdot\dot{\mathbf{q}}_e(t) + \mathbf{K}_e\cdot\mathbf{q}_e(t) = \mathbf{B}_e\cdot\mathbf{u}(t) \\ &\mathbf{y}(t) = \mathbf{C}_e\cdot\mathbf{q}_e(t) \end{aligned} \tag{3}$$

with the input matrix $\mathbf{B}_e \in \mathbb{R}^{N\times p}$ and the output matrix $\mathbf{C}_e \in \mathbb{R}^{r\times N}$, where $p$ and $r$ are numbers of input and output coordinates, respectively.

Due to the representation of elastic multibody systems using the floating frame approach, the elastic degrees of freedom can be reduced by means of methods of linear model order reduction. The basic idea of model reduction is to approximate the original set of ODEs by a low-dimensional set of equations that preserves significant characteristics of the full model. The large vector of elastic coordinates $\mathbf{q}_e \in \mathbb{R}^N$ is approximated using a vector $\tilde{\mathbf{q}}_e \in \mathbb{R}^n$ of a smaller dimension $n \ll N$ as

$$\mathbf{q}_e(t) \approx \mathbf{V}\cdot\tilde{\mathbf{q}}_e(t), \tag{4}$$

with $\mathbf{V} \in \mathbb{R}^{N\times n}$ being a projection matrix.

The transformation (4) leads to the following equations of motion:

$$\begin{aligned} &\tilde{\mathbf{M}}_e\cdot\ddot{\tilde{\mathbf{q}}}_e(t) + \tilde{\mathbf{D}}_e\cdot\dot{\tilde{\mathbf{q}}}_e(t) + \tilde{\mathbf{K}}_e\cdot\tilde{\mathbf{q}}_e(t) = \tilde{\mathbf{B}}_e\cdot\mathbf{u}(t) \\ &\mathbf{y}(t) = \tilde{\mathbf{C}}_e\cdot\tilde{\mathbf{q}}_e(t) \end{aligned} \tag{5}$$

with the reduced matrices $\tilde{\mathbf{M}}_e = \mathbf{V}^T\cdot\mathbf{M}_e\cdot\mathbf{V}$, $\tilde{\mathbf{D}}_e = \mathbf{V}^T\cdot\mathbf{D}_e\cdot\mathbf{V}$, $\tilde{\mathbf{K}}_e = \mathbf{V}^T\cdot\mathbf{K}_e\cdot\mathbf{V} \in \mathbb{R}^{n\times n}$, and $\tilde{\mathbf{B}}_e = \mathbf{V}^T\cdot\mathbf{B}_e \in \mathbb{R}^{n\times p}$, $\tilde{\mathbf{C}}_e = \mathbf{C}_e\cdot\mathbf{V} \in \mathbb{R}^{r\times n}$.

The difference of model order reduction methods lies in the way they generate the transformation matrix $\mathbf{V}$.

## 3 MODEL ORDER REDUCTION METHODS

The task of model reduction methods is to find a coordinate transformation matrix **V** in the relation (4). This section shows the constuction of **V** according to the line-fitting method and the SEREP approach. We also point out advantages and disadvantages of both methods relative to the requiremets stated in the introduction section.

### 3.1 SEREP

The idea of SEREP method is to include to a reduced order model only $c$ modes and $r$ degrees of freedom that are necessary to the model. Here $c$ is defined using a bandwidth of excitation loads and $r$ is defined based on the number of degrees of freedom where external loads are applied or model output coordinates are defined.

According to the SEREP method, all degrees of freedom of finite element model are divided into two sets: retained and truncated. These degrees of freedom are denoted with the indexes $r$ and $t$, respectively. The elastic part of equations of motion from (1) with no damping can be written as

$$\mathbf{M}_e \cdot \ddot{\mathbf{q}}_e + \mathbf{K}_e \cdot \mathbf{q}_e = \mathbf{h}_e \tag{6}$$

For the sake of simplicity we omit the index $e$ for the subsequent equations in this section. Partition of (6) into retained and truncated coordinates yields:

$$\begin{bmatrix} \mathbf{M}_{rr} & \mathbf{M}_{rt} \\ \mathbf{M}_{rt} & \mathbf{M}_{tt} \end{bmatrix} \cdot \begin{bmatrix} \ddot{\mathbf{q}}_r \\ \ddot{\mathbf{q}}_t \end{bmatrix} + \begin{bmatrix} \mathbf{K}_{rr} & \mathbf{K}_{rt} \\ \mathbf{K}_{rt} & \mathbf{K}_{tt} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{q}_r \\ \mathbf{q}_t \end{bmatrix} = \begin{bmatrix} \mathbf{h}_r \\ \mathbf{h}_t \end{bmatrix} \tag{7}$$

The coordinate transformation from physical coordinates to modal coordinates is defined as

$$\mathbf{q} = \mathbf{U} \cdot \mathbf{x} \tag{8}$$

where $\mathbf{U}$ and $\mathbf{x}$ are a modal matrix and modal coordinates. According to the SEREP method, the physical coordinates are approximated by a subset of modal vectors and corresponding modal coordinates as follows

$$\mathbf{q} = \mathbf{U}_c \cdot \mathbf{x}_c \tag{9}$$

Next, the approximation of physical coordinates is partitioned as

$$\begin{bmatrix} \mathbf{q}_r \\ \mathbf{q}_t \end{bmatrix} = \begin{bmatrix} \mathbf{U}_{cr} \\ \mathbf{U}_{ct} \end{bmatrix} \cdot \mathbf{x}_c \tag{10}$$

Solving of $\mathbf{q}_r = \mathbf{U}_{cr} \cdot \mathbf{x}_c$ for $\mathbf{x}_c$ gives

$$\mathbf{x}_c = \mathbf{U}_{cr}^{+} \cdot \mathbf{q}_r \tag{11}$$

where $\mathbf{U}_{cr}^{+} = (\mathbf{U}_{cr}^T \cdot \mathbf{U}_{cr})^{-1} \cdot \mathbf{U}_{cr}^T$ defines a pseudo inverse matrix of $\mathbf{U}_{cr}$. Substituting (11) into (9) yields

$$\mathbf{q} = \mathbf{U}_c \cdot \mathbf{U}_{cr}^{+} \cdot \mathbf{q}_r \tag{12}$$

It follows that the coordinate transformation matrix takes the form

$$\mathbf{V} = \mathbf{U}_c \cdot (\mathbf{U}_{cr}^T \cdot \mathbf{U}_{cr})^{-1} \cdot \mathbf{U}_{cr}^T. \tag{13}$$

The accuracy of eigenfrequencies and eigenvectors retained in SEREP reduced order model is exact [13]. The modes may be arbitrarily selected from the modes in the original model. In addition, the accuracy of the eigensolution of the reduced model does not depend on the location of interface degrees of freedom. The quality of predefined transfer functions in the reduced model is determined by the eigenvectors chosen in the SEREP procedure. The opened questions are what and how many modes to include into the model to meet accuracy demands on the transfer functions of interest.

The order and stiffness of reduced model are defined by the number of chosen interface coordinates and the highest eigenfrequency of retained modes.

The tuning of reduced model for certain transfer functions and certain frequency ranges can be a not trivial task because of the difficult choice of important eingevectors.

The non-proper definition of number of interface coordinates and retained modes can result in the generation of rank-deficient mass and stiffness matrices. In this case, stability of full model is not preserved in the reduced model.

The transformation (13) preserves the second order structure of equations of motion for the reduced system.

Computation of eigensolution takes in general $O(N^3)$ arithmetical operations. In some applications the eigenvalues can be known in advance, e.g., from experimental tests. The computational cost of pseudo inverse using SVD is $O(N^3)$ operations.

### 3.2 Line-fitting method

The line-fitting method approximates the dynamical behavior of mechanical systems by fitting transfer functions

$$\mathbf{H}(s) = \mathbf{C}_e \cdot (s^2\mathbf{M}_e + s\mathbf{D}_e + \mathbf{K}_e)^{-1} \cdot \mathbf{B}_e \tag{14}$$

that are determined from the system of equations (3) by using the Laplace transformation with a complex frequency $s = i\omega$.

In order to obtain the projection matrix $\mathbf{V}$ without rigid body modes, we exclude the contribution of rigid body motion form (14). The translational and rotational rigid modes can be written as follows:

$$\mathbf{V}_0 = \begin{bmatrix} \mathbf{V}_{0t} & \mathbf{V}_{0r} \end{bmatrix} = \left[ \begin{bmatrix} \vdots \\ [\mathbf{I}]^k \\ \vdots \end{bmatrix} \quad \begin{bmatrix} \vdots \\ [-\tilde{\mathbf{r}}]^k \\ \vdots \end{bmatrix} \right], k = 1,..,N. \tag{15}$$

Here the matrix $\mathbf{I}$ is a 3 × 3 identity matrix, the term $[\mathbf{r}]^k$ gives an undeformed position of k-th node of the FE model and consequently $[\tilde{\mathbf{r}}]^k$ annotates an associated skew symmetric matrix. The given calculation of $\mathbf{V}_0$ is valid for finite element nodes that exhibit three translational degrees of freedom.

The part of transfer matrix that corresponds to rigid body motion can be found as

$$\mathbf{H}_0(s) = \mathbf{C}_0 \cdot (s^2\mathbf{M}_0 + s\mathbf{D}_0 + \mathbf{K}_0)^{-1} \cdot \mathbf{B}_0 \tag{16}$$

with $\mathbf{M}_0 = \mathbf{V}_0^T \cdot \mathbf{M}_e \cdot \mathbf{V}_0$, $\mathbf{D}_0 = \mathbf{V}_0^T \cdot \mathbf{D}_e \cdot \mathbf{V}_0$, $\mathbf{K}_0 = \mathbf{V}_0^T \cdot \mathbf{K}_e \cdot \mathbf{V}_0 \in \mathbb{R}^{6\times 6}$, and transformed input and output matrices $\mathbf{B}_0 = \mathbf{V}^T \cdot \mathbf{B}_e \in \mathbb{R}^{6\times p}$, $\mathbf{C}_0 = \mathbf{C}_e \cdot \mathbf{V}_0 \in \mathbb{R}^{r\times 6}$. The component of transfer matrix (14) corresponding to elastic deformations is defined as

$$\mathbf{H}_1(s) = \mathbf{H}(s) - \mathbf{H}_0(s) \tag{17}$$

Further, we are looking for the transformation matrix **V**. We want to retain the output coordinates in the reduced order model, therefore the upper part of the coordinate transformation matrix **V** is an r × r identity matrix:

$$\mathbf{V} = \begin{bmatrix} \mathbf{I} \\ \mathbf{W} \end{bmatrix} \tag{18}$$

The matrix **W** can be found from the relation of non-output coordinates and output coordinates in terms of their transfer functions as

$$\mathbf{H}_1^{\perp}(s) \approx \mathbf{W} \cdot \mathbf{H}_1(s) \tag{19}$$

We construct the matrix **W** as a trade-off between frequency points of interest $s_k = i \cdot 2\pi f_k, k = 1..z$. The equation (19) is turned into

$$\begin{bmatrix} \mathbf{H}_1^{\perp}(s_1) & \dots & \mathbf{H}_1^{\perp}(s_z) \end{bmatrix} = \mathbf{W} \cdot \begin{bmatrix} \mathbf{H}_1(s_1) & \dots & \mathbf{H}_1(s_z) \end{bmatrix} \tag{20}$$

The decomposition of (20) into real and imaginary parts yields:

$$\begin{aligned} &\begin{bmatrix} \Re(\begin{bmatrix} \mathbf{H}_1^{\perp}(s_1) & \dots & \mathbf{H}_1^{\perp}(s_z) \end{bmatrix}) & \Im(\begin{bmatrix} \mathbf{H}_1^{\perp}(s_1) & \dots & \mathbf{H}_1^{\perp}(s_z) \end{bmatrix}) \end{bmatrix} = \\ &\mathbf{W} \cdot \begin{bmatrix} \Re(\begin{bmatrix} \mathbf{H}_1(s_1) & \dots & \mathbf{H}_1(s_z) \end{bmatrix}) & \Im(\begin{bmatrix} \mathbf{H}_1(s_1) & \dots & \mathbf{H}_1(s_z) \end{bmatrix}) \end{bmatrix} \end{aligned} \tag{21}$$

that can be rewritten as

$$\begin{aligned} \mathbf{T}_n &= \mathbf{W} \cdot \mathbf{T}_o \\ \mathbf{T}_o^T \cdot \mathbf{W} &= \mathbf{T}_n^T \end{aligned} \tag{22}$$

The matrices $\mathbf{T}_o^T \in \mathbb{R}^{2pz \times r}$ and $\mathbf{T}_n^T \in \mathbb{R}^{2pz \times (N-r)}$, where $p, r, z$ are the number of inputs, outputs, and reference frequency points, respectively. The order of reduced model $n$ is equal to the number of outputs $r$. The matrix **W** can be found as a least-squares solution of (22). After the projection matrix **V** is obtained, the reduced order model is constructed using (5).

The main positions of reference frequency points have to be points of resonances and anti-resonances. Besides, additional reference frequency points can be used to enhance quality of reduced model in special frequency ranges or points.

The error of predefined transfer functions in the line-fitting method depends on a residuum in equation (22). The error can be influenced by a number of interface coordinates, position of interface coordinates, and a choice of reference frequency points. As for the majority of other reduction methods, the error in this approach is unknown before the reduction, but it can be calculated after the reduction. Recommendations on the achievement of highly precise results are given in [15].

Simulation speed depends in addition to other factors on the number of coordinates of reduced order model and stiffness of the system. According to the results presented in [16, 17], line-fitting models require relative low number of coordinates to achieve desired accuracy demands. The stiffness is defined by the highest eigenfrequency of the system. The line-fitting method can influence the accuracy of eigenfrequencies contained in the predefined frequency range, but the spectrum of model beyond this range is uncontrolled. For this reason, no statements about the limits of the spectrum can be made. However, numeric examples in [16, 17]

show that the stiffness of line-fitting models is significantly lower than the stiffness of Craig-Bampton models.

The tuning of reduced model for certain transfer functions and certain frequency ranges is a straight forward process. The local accuracy of transfer functions can be improved by additional reference frequency points, whereas the accuracy over a whole frequency range of interest is enhanced by increasing the order of the reduced model.

The preservation of stability is an important property of reduction approaches because this ensures that a reduced model does not cause any type of failure to elastic multibody systems. It was proven in [15] that the line-fitting technique meets this essential requirement.

The systems reduced by the line-fitting method preserve the second order structure of equations of motion. This enables integration of reduced order models into a multibody system of second order type.

The main computational burdens of this reduction approach consist in the calculation of transfer function matrices for several frequency points and the solution of large least-squares problem. In applications, where the input-output behavior of a system is under consideration, the most of reduction approaches require calculation of transfer function matrices for all points in a predefined frequency range to evaluate accuracy of reduced order model. Therefore the values of transfer function matrices at several frequency points are considered as given information. In the case where the values of transfer function matrices are unknown, the complexity of their computation has a cubic dependency of $N$. The computational cost of least squares solution of equation (22) is $O(zn^2N)$. As before, $z$ is the number of reference frequency points, $n$ is the number of interface coordinates, and $N$ is the number of coordinates for the original finite element model. The total computational cost of the line-fitting method is larger in comparison with the Craig-Bampton approach, but it remains acceptable for moderately dimensional finite element models.

## 4 NUMERICAL EXAMPLE

In this section we reduce a finite element model of unconstrained elastic bar illustrated in Figure 1. The SEREP and line-fitting methods are applied and results are compared.

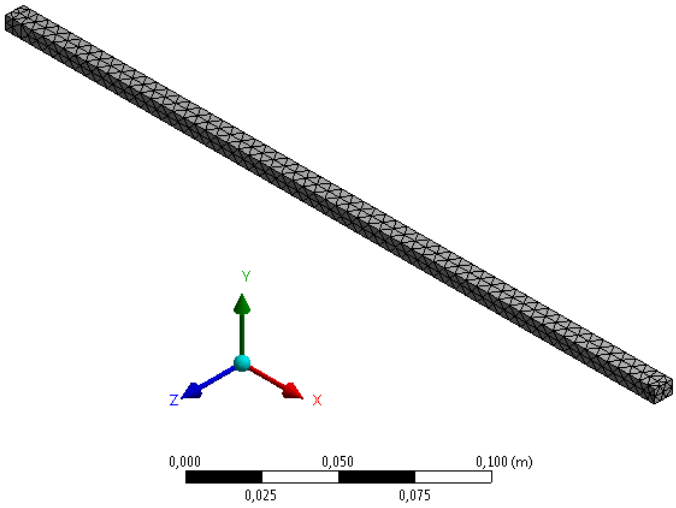



Figure 1: Finite element model of elastic bar.

### 4.1 Model description

The model of the bar has the following characteristics: size 6×8×300 mm, mass 0,1 kg,

Young's modulus $2 \cdot 10^{10}$ Pa, damping factor beta $10^{-5}$, number of degrees of freedom 1656. As a frequency range of interest we define an interval [0, 1000] Hz. The interval contains 6 zero eigenfrequencies and 6 deformation eigenfrequencies of the model. The set of input and output degrees of freedom for the bar is demonstrated in Figure 2. Here input and output coordinates coincide with each other.



Figure 2: Input and output coordinates of the model.

### 4.2 Initialization of reduction methods

For the line-fitting method the frequency range of interest, input and output coordinates are considered as given information. The task of the engineer is to choose an appropriate order of reduced model, to define interface degrees of freedom, and to allocate reference frequency points. The order of reduced model depends on a number of transfer functions that need to be tuned and a number of eigenfrequencies contained in a frequency range of interest. According to [15], the order equal to a double number of eigenfrequencies in the frequency range of interest is usually enough to achieve proper accuracy of reduced model. In the case of the bar model, we define the order of reduced model to be 14. After that the same number of interface degrees of freedom must be chosen. Nine interface degrees of freedom are assigned to nine user defined input and output coordinates, while the remaining coordinates are uniformly distributed throughout the model. The interface coordinates are located as it is shown in Figure 3.

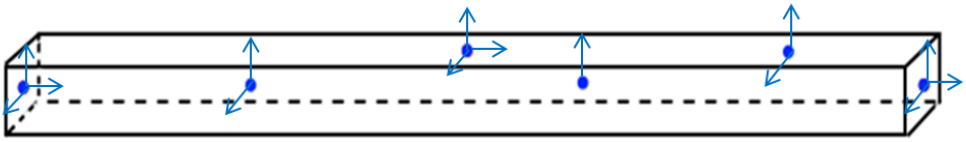


Figure 3: Interface coordinates for the reduced order model.

In order to choose reference frequency points, we compute transfer functions specified by the interface DoFs. The set of reference frequency points is common for all interface transfer functions; therefore it is possible to obtain a complete set of frequency points analyzing only a few of them. The reference points are defined at resonance, antiresonance, and some intermediate frequencies of the interval [0,1000] Hz, see Figure 4. For this model 20 frequency points of interest are assigned.

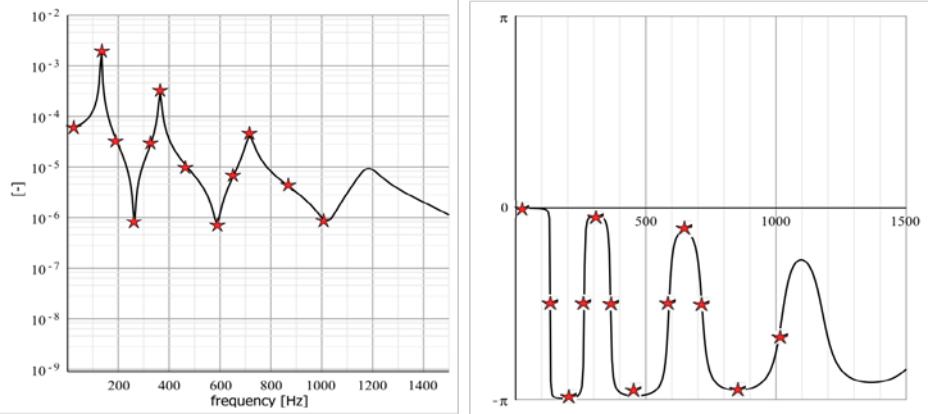



Figure 4: Choice of reference frequency points.

In order to initialize the SEREP method, it is necessary to choose interface coordinates and modes retained in the reduced model. We defined nine interface coordinates shown in Figure 2 as degrees of freedom of reduced order model. In order to avoid rank-deficiency problems in the SEREP method, we assigned the same amount of modes as the number of retained coordinates. Thus, the set of chosen modes consists of nine deformation modes with lowest frequencies. It also follows that the number of retained modes is larger than the amount of modes in the frequency range of interest.

### 4.3 Evaluation of reduced order models

For the comparison of the reduction methods we employ three criteria: normalized relative eigenfrequencies difference, modal assurance criterion, and a relative error of transfer functions. Figure 5 represents the result of the former test for the eigenfrequencies in the range of interest [0, 1000] Hz.

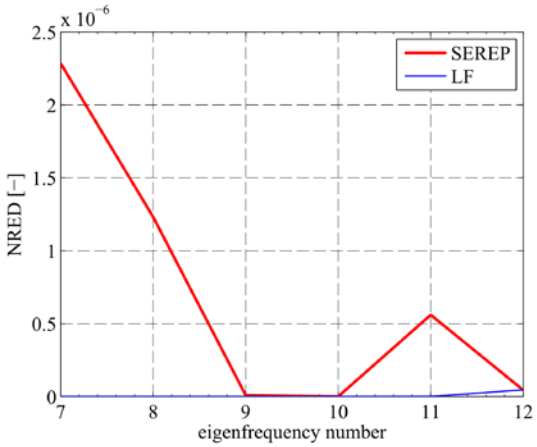



Figure 5: Relative error of non-zero eigenfrequencies for reduced models.

The diagram shows that the eigenfrequencies in the predefined frequency range are accurately approximated by both methods, but the precision is considerably better for the line-fitting model.

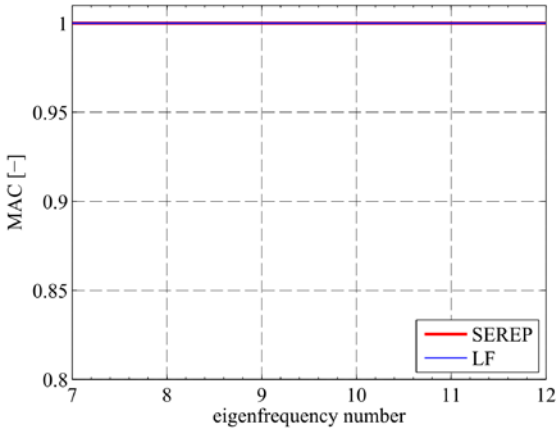



Figure 5: Correlation of deformation modes.

The results of MAC test are illustrated in Figure 6. The diagram shows that MAC values for all modes in the predefined frequency range are higher than 0.8. This implies a high correlation of deformation forms of the reduced and full models and, as a result, a qualitatively successful approximation.

Further we evaluate dynamics properties of reduced order models using transfer functions. For multi-input multi-output systems it is problematically to analyze errors of transfer functions separately, therefore the error of reduction is usually evaluated using the formula:

$$\varepsilon(\omega) = \frac{\left\| \mathbf{H}(i\omega) - \tilde{\mathbf{H}}(i\omega) \right\|_F}{\left\| \mathbf{H}(i\omega) \right\|_F} \tag{23}$$

where $\mathbf{H}$ and $\tilde{\mathbf{H}}$ are transfer matrices of original and reduced systems, and $\|\cdot\|_F$ denotes the Frobenius norm. The function $\varepsilon(\omega)$ shows a total error introduced by a model reduction approach in a certain frequency range. Figure 6 represents a relative error of transfer functions for both the reduced order models and interface degrees of freedom shown in Figure 2.

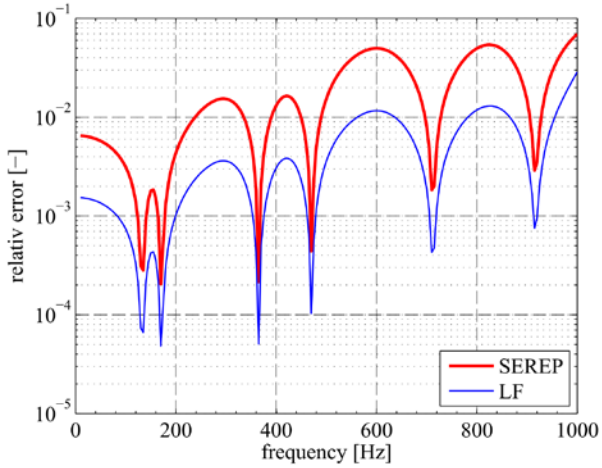



Figure 6: Relative error of frequency response.

The plot shows that the line-fitting method approximated the relevant transfer functions over the whole interval considerably more accurately than the SEREP approach. The error of line-fitting model does not exceed 3%. The SEREP method provides results with errors up to 7%. Increasing of order of SEREP model and the amount of modes to 14 improves the accuracy of transfer functions, but the line-fitting model is still superior, see Figure 7.

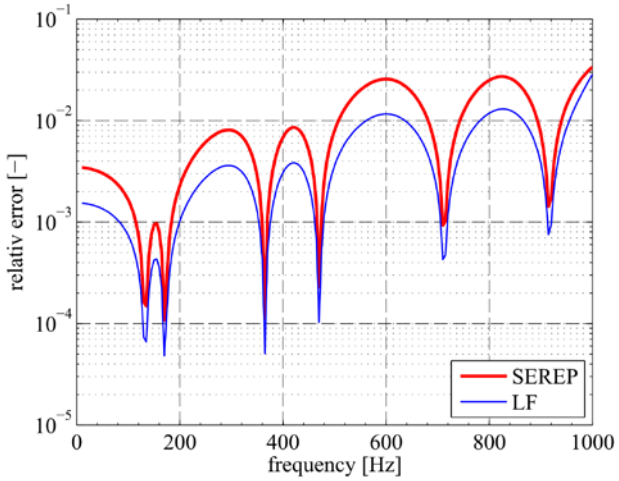



Figure 7: Relative error of frequency response with the enhanced SEREP model.

## 5 SUMMARY AND FURTHER RESEARCH

In this paper we evaluated the line-fitting method relative to the demands defined by application examples. The demands included a high accuracy of reduced order model, fast simulation, simple tuning of reduced model for certain transfer functions and certain frequency ranges, preservation of model stability, preservation of second order structure of equations of motion, and computational efficiency of reduction method. The evaluation was made based on the theoretical description and results of the numerical example. The analysis showed that the line-fitting method satisfies all the requirements for moderately dimensional finite element models.

In addition, the line-fitting method was compared with the widely used SEREP approach. We found that the line-fitting approach outperforms the SEREP method in accuracy, possibility to tune reduced order models for certain transfer functions and certain frequency ranges, and preservation of stability. Both methods preserve the second order structure of equations of motion. The relative drawback of the line-fitting method is the higher computational cost of a coordinate transformation matrix. Besides, the stiffness of the line-fitting model was higher relative to the stiffness of SEREP model.

Further we intend to improve computational efficiency of the line-fitting method and to compare the approach with reduction techniques from the control theory.

## REFERENCES

[1] T. Wasfy, A. Noor, Computational strategies for flexible multibody systems. *Applied Mechanics Reviews*, Bd. 56, Nr. 6, 553-613, 2003.

[2] J. Guyan, Reduction of stiffness and mass matrices. *AIAA Journal*, Bd. 3, 380-380, 1965.

[3] A. Leung, An accurate method of dynamic condensation in structural analysis. *International Journal for Numerical Methods in Engineering*, Bd. 12, Nr. 11, 1705-1715, 1978.

[4] M. C. C. Bampton, R. R. Craig, Coupling of substructures for dynamic analyses. *AIAA*, Bd. 6, 1313-1319, 1968.

[5] R. Craig, Coupling of substructures for dynamic analyses: an overview. *Proceedings of the AIAA Dynamics Specialists Conference*, Atlanta, USA, 2000.

[6] S. Dietz, *Vibration and Fatigue Analysis of Vehicle Systems Using Component Modes*. VDI-Verlag, 1999.

[7] J. O'Callahan, P. Avitabile, R. Riemer, System Equivalent Reduction Expansion Process (SEREP). In *7. International Modal Analysis Conference*, Las Vegas, 1989.

[8] A. C. Antoulas, *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.

[9] Z. Bai, Y. Su, Dimension reduction of second-order dynamical systems via a second-order Arnoldi method, *SIAM J. Sci. Comp.*, Vol.25, No.5, 1692-1709, 2005.

[10] B. Salimbahrami, B. Lohmann, Order reduction of large scale second-order systems using krylov subspace methods. *Linear Algebra and its Applications*, 415, 385-405, 2006.

[11] P. Benner, J. Saak, Efficient balancing-based MOR for large-scale second-order systems. *Mathematical and Computer Modelling of Dynamical Systems*, Bd. 17, Nr. 2, 123-143, 2011.

[12] Z.-Q. Qu, *Model Order Reduction Techniques: With Applications in Finite Element Analysis*. Springer London, London, 2004.

[13] P. Koutsovasilis, *Model order reduction in structural mechanics: Coupling the rigid and elastic multi body dynamics*, VDI-Verl., Duesseldorf, 2009.

[14] B. Besselink, U. Tabak, A. Lutowska, van de Wouw, N., H. Nijmeijer, D. J. Rixen, M. E. Hochstenbach, W.H.A. Schilders. A comparison of model reduction techniques from structural dynamics, numerical mathematics and systems and control. *Journal of Sound and Vibration*, 332(19), 4403-4422, 2013.

[15] V. Makhavikou, *Line-fitting method of model order reduction for elastic multibody systems*, PhD thesis, Otto von Guericke University, Magdeburg, 2015.

[16] V. Makhavikou, R. Kasper, D. Vlasenko, Method of model reduction for elastic multibody systems, *Proceedings of 11th World Congress on Computational Mechanics, 5th European Conference on Computational Mechanics*, 475-486, Barcelona, July 20-25, 2014.

[17] V. Makhavikou, R. Kasper, D. Vlasenko, Line-fitting method of model order reduction in a context of elastic multibody simulation, *Proceedings of 12th Magdeburger Maschinenbau-tage,* Otto von Guericke University, Magdeburg, September, 2015.

[18] R. Schwertassek, O. Wallrapp, *Dynamik flexibler Mehrkoerpersysteme: Methoden der Mechanik zum rechnergestützten Entwurf und zur Analyse mechatronischer Systeme*, Braunschweig, Vieweg, 1999.

[19] A. A. Shabana, *Dynamics of multibody systems*, Cambridge and U.K. and New York, Cambridge University Press, 1998.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EFFICIENT ORDER REDUCTION STRATEGY IN EARTHQUAKE NONLINEAR RESPONSE ANALYSIS OF STRUCTURES

**F. Bamer[1], A. Kazemi Amiri[2] and C. Bucher[3]**

[1] RWTH-Aachen University, Institute of General Mechanics,
Templergraben 64, 52062 Aachen, Germany
e-mail: bamer@iam.rwth-aachen.de

[2] Vienna Doctoral Programme on Water Resource Systems,
TU Vienna, Karlsplatz 13/222, 1040 Vienna, Austria
e-mail: mehrad_hoo@yahoo.com

[3] Center for Mechanics and Structural Dynamics,
TU Vienna, Karlsplatz 13/206, 1040 Vienna, Austria
e-mail: Christian.Bucher@tuwien.ac.at

**Keywords:** Model order reduction, Proper orthogonal decomposition, Friction pendulum system

**Abstract.** *Since earthquake dynamic response analysis of large and complex structures are computationally time demanding, efficient methods that can reduce the system order are of high interest. In this sense, there are different methods available, which try to provide a proper equivalent model. However, in the presence of nonlinearities in the structural elements, most of those methods are ruled out due to their linear assumptions. Therefore, this contribution aims at providing an efficient strategy, which can reduce the order of the nonlinear structural model while retaining important structural characteristics for further earthquake dynamic response analysis. The model order reduction (MOR) strategy is developed based on the proper orthogonal decomposition (POD) method to derive a set of nonlinear deterministic POD modes according to the information of the response history (snapshots) of the full order structure under one or a set of representative earthquake excitations. Subsequently, the POD modes are utilized to create the reduced-order models of the structure subjected to different earthquake excitations. Then, the reduced order models need substantially less amount of computational time in comparison to the full order models. This study presents the application results of the introduced new strategy to a realistic building structure, which is base-isolated by means of frictional bearing elements for better seismic performance. The results demonstrate accurate approximations of the physical (full) responses by means of this new MOR strategy if the probable behavior of the structure has already been captured in the POD snapshots.*

## 1 INTRODUCTION

The evaluation of the response history of a structure in the time domain is one of the main topics in earthquake engineering and structural dynamics. Often it is common practice to create simple structural models, e.g. multistory shear frames, which should be able to describe the structural behavior and peculiarities of the real structure. This approach leads to useful results for the investigation of rather simple and uniform structures in order to draw meaningful engineering decisions regarding structural resistance. On the contrary, the analysis of complicated systems can require the application of nonlinear high-order systems, as a characterization by a low dimensional structural model could lead to an oversimplification, i.e. important motion patterns could be ignored. Therefore, an effective strategy is to obtain a set of a low number of "important" equations of motion that approximates the high-dimensional nonlinear dynamical system as accurately as possible, that is, model order reduction (MOR).

The solution of the nonlinear set of equations of motion in the time domain is realized by numerical algorithms, which require computational effort if the number of DOF is high. Even the response calculation of linear systems can be expensive, as a factorization of the stiffness matrix is necessary to solve the eigenvalue problem and calculate the natural modes of vibration.

An alternative is to replace a high-dimensional nonlinear set of equations of motion by a reduced set, providing the main dynamic behavior of the system to reach the required level of accuracy. MOR methods are used in many fields of research, where high-dimensional systems are dealt. Some review papers of MOR, especially for structural dynamic applications, are presented by Rega and Troga [3] and Koutsovasilis and Beitelschmidt [4] as well as the books of Qu [2] and Schilders et al. [5]. The classical but also effective method of modal truncation is well-known in the field of earthquake engineering, which is however mainly applicable to linear systems.

This paper concentrates on a new MOR strategy based on the proper orthogonal decomposition (POD) method. The POD provides a low dimensional uncorrelated description (basis vectors), by which a high-dimensional correlated process, e.g. structural response, is spanned. Firstly, it was used as a statistical formulation in the papers of Kosambi [6], Karhunen [7] and Loeve [8]. Following this mathematical basis of the POD, which is also known as Karhunen-Loeve Decomposition and Principal Component Analysis, was applied in many fields of research, including turbulence and coherent structures, wind engineering, image processing and structural dynamics.

The first paper regarding the field of structural dynamics was written by Cusumano et al. [9] in the early 1990's, who presented an experimental study of dimensionality in an elastic impact oscillator. In the papers [10] and [11] of Feeny and Kappagantu a relation of the proper orthogonal modes to normal modes of vibration is investigated. Then they used the POD as they so call optimal modal reduction and exploit the benefits of the application of these modes in comparison to the linear natural modes. Furthermore, Kappagantu and Feeny investigated in [12] and [13] the dynamics of an experimental frictionally excited beam and they verified that the proper orthogonal modes are efficient in capturing the dynamics of the system. Liang et. al. [14] discuss the realizations of the POD, i.e. Karhunen-Loeve Decomposition, principal component analysis and singular value decomposition and compare these three methods. Kerschen and Golivani [15] analyze the physical interpretation of the POD modes and its relation to the singular value decomposition and in [16] they investigated POD based on auto-associative neural networks.

The necessity to describe a high dimensional set by a small set of equations of motion, i.e.

MOR, has aroused interests mainly in the last two decades in the field of earthquake engineering, some papers related to this issue are e.g. [27] and [28]. More specified to the related issue Tubino et. al. [17] investigated the seismic ground motion of the support points of a structure and denominate the POD as a very efficient tool to simulate multi-variate processes. Bucher [32] examined the stabilization of explicit time integration methods for analysis of nonlinear structural dynamics by modal reduction. Gutierrez and Zaldivar investigated in [30] how to handle the stability problem of explicit time integration by modal truncation methods more related to problems in earthquake engineering and structural dynamics and following in [29], they applied the Karhunen-Loeve Decomposition, which is formally identical to the POD analysis, to capture the essential characteristics of nonlinear systems and provide experimental examples conducted on a shaker table. Bamer and Bucher [18] developed a MOR strategy applying the POD method for transient excited structures resting on one-dimensional friction elements. This study presented a powerful combination of the POD and explicit time integration schemes.

The current work investigates the extension of the POD-based MOR strategy, which is applicable to nonlinear systems in contrast to the method of modal truncation. The new strategy pursues the following objective: a low number of deterministic nonlinear modes (i.e. set of POD modes) is determined that defines a representative characterization of the structural behavior. Therefore, due to the information content of the full or a part of the time response of the structure to one representative excitation a set of deterministic modes, i.e. POD modes, is evaluated. Subsequently, this set of modes is utilized to project the equations of motion of a structure under different earthquake excitations onto POD coordinates and following an order truncation is performed in a similar manner as the application of modal truncation to linear systems.

It is attached importance to demonstrate this new strategy and its advantages on a practical application. The method is applied to the dynamic model of a realistic building structure. The building is erected on friction pendulum bearings for the sake of seismic isolation to minimize the transferred acceleration to the building during an earthquake. A three-dimensional dynamic model of the base isolated structure is derived by implementing the finite element model of the structure and the bi-directional friction pendulum systems. The subsequent section deals with the nonlinear dynamic model of the base-isolated structure. Following, the new POD-based MOR strategy and the example of its practical application is provided. Finally, the discussion on the results and conclusions are given.

## 2 Nonlinear model order reduction

The $n$-dimensional set of equations of motion of a structure with nonlinear material behavior excited by horizontal components of ground acceleration is expressed as (cf. Chopra [20])

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{R}(\mathbf{x}) = -\mathbf{M}\left(\mathbf{f}_{\mathbf{x}}\ddot{x}_g + \mathbf{f}_{\mathbf{y}}\ddot{y}_g\right) \ , \tag{1}$$

where $\mathbf{M}$ and $\mathbf{C}$ are mass- and damping square matrices of order $n$ and $\mathbf{R}(\mathbf{x})$ is the nonlinear internal restoring force vector dependent on the displacement $\mathbf{x}$ with the dimension $n \times 1$. The right hand side of the set of equations of motion describes the earthquake excitation term, while $\ddot{x}_g$ and $\ddot{y}_g$ denote the ground acceleration in $x$- and $y$-direction and $\mathbf{f}_j, (j = x, y)$ are the influence vectors in the corresponding direction. It is

$$\mathbf{f}_{\mathbf{x}}(x_i) = 1 \ , \quad \mathbf{f}_{\mathbf{y}}(y_i) = 1 \ , \quad i = 1...n \ , \tag{2}$$

at the global $x$ and $y$ degrees of freedom of all nodes, whereas the other components of $\mathbf{f}_{\mathbf{x}}$ and $\mathbf{f}_{\mathbf{y}}$ are zero. Thus, $i$ describes the number of nodes of the FE discretized structure. This general

approach indicates that in this paper the ground acceleration in the corresponding direction, i.e. $x$-, $y$- or $z$-direction, is equal in all structural support points. In the following equations the term on the right hand side of the set of equations of motion (1) is denominated by $\mathbf{F}(t)$, which has the unit of a force.

Nonlinear systems, as they are depicted in Equation (1), have generally to be solved by the application of a numerical algorithm, that is, a step by step procedure in the time domain in order to obtain the response of the structure as a function of time. The necessity of application of a numerical method inevitably leads to the existence of computational effort if $n$ is a large number. Therefore, the approximation by a low, dimensional description of the system seems to be useful, i.e. the application of MOR.

The main goal of MOR techniques is primarily to define a transformation matrix $\mathbf{T} \in \mathbb{R}^{n\times m},\ m \ll n$ to approximate the displacement vector $\mathbf{x} \in \mathbb{R}^n$ through a reduced coordinate vector $\mathbf{q_r} \in \mathbb{R}^m$ by the relation (cf. Koutsovasilis and Beitelschmidt [19])

$$\mathbf{x} = \mathbf{T}\mathbf{q_r}\ , \tag{3}$$

such that the dynamic properties of the system are preserved and the error is small. The notation of the variable $m \in \mathbb{N}$ is the dimension of the reduced subspace.

The projection of the nonlinear system defined by Equation (1) onto that subspace leads to another second-order ordinary differential equation (cf. Koutsovasilis and Beitelschmidt [19])

$$\mathbf{m_r}\ddot{\mathbf{q}}_\mathbf{r} + \mathbf{c_r}\dot{\mathbf{q}}_\mathbf{r} + \mathbf{r} = \mathbf{f_r}\ , \tag{4}$$

where $\mathbf{m_r} = \mathbf{T^T M T},\ \mathbf{c_r} = \mathbf{T^T C T} \in \mathbb{R}^{m\times m}$ are mass- and damping matrix and $\mathbf{f_r} = \mathbf{T^T F}(t) \in \mathbb{R}^{m\times 1}$ is the force vector in the reduced subspace. It should be noted that the reduced system matrices $\mathbf{m_r}$ and $\mathbf{c_r}$ are generally not diagonal. The vector of the restoring forces in the reduced subspace is

$$\mathbf{r} = \mathbf{T^T R}(\mathbf{x}) = \mathbf{T^T R}(\mathbf{T}\mathbf{q_r})\ . \tag{5}$$

Consequently, one necessity of nonlinear MOR is the evaluation of the vector of the restoring forces in the physical (full) coordinate at every time step.

## 3 The proper orthogonal decomposition and nonlinear modes

Modal truncation is a widely-used tool and an effective method for order reduction of linear systems in the field of earthquake engineering. An accurate approximation of the response history is achieved by applying a small number of lower structural modes proportional to the number of degrees of freedom. In this work, the objective is to find a new strategy that is applicable to nonlinear systems in a similar manner to modal truncation. The approach is to define a set of deterministic modes that can be evaluated from the information of an existing response history of the structure. Consequently, this set of modes contains nonlinear motion patterns if the structure shows nonlinear response behavior to the excitation.

The proposed strategy is established based on the theory of the POD method. Generally, the POD ([2], , [14], [21], [22], [23], [24]) is a straightforward approach to obtain a low-dimensional uncorrelated process from a correlated high dimensional or even infinite-dimensional process. Holmes et al. [24] examined the theoretical background of the POD and its properties profoundly. In the following, the mathematical basics of the POD are discussed shortly, but as the paper is more targeted to the strategic approach in earthquake engineering, the dealing of the mathematical background and the numerical problems are limited to an essential minimum.

The aim of the POD is to find a set of ordered orthonormal basis vectors in a subspace so that samples in a sample space are expanded in terms of $l$ basis vectors in an optimal form. This means that the POD is able to find an orthonormal basis, which describes an observation vector in a subspace better than any other orthonormal basis can do. A measure for this problem is the mean square error (cf. Qu [2])

$$E\left\{\|\mathbf{x}-\mathbf{x}(l)\|^2\right\} \leq E\left\{\|\mathbf{x}-\hat{\mathbf{x}}(l)\|^2\right\} \ , \tag{6}$$

where $\mathbf{x} \in \mathbb{R}^{n\times 1}$ is the random vector, $\mathbf{x}(l)$ is the approximation of this random vector in an $l$-dimensional POD subspace and $\hat{\mathbf{x}}(l)$ is the approximation of the random vector by any other possible orthonormal basis. Therefore, the random vector can be expressed as (cf. Qu [2])

$$\mathbf{x} = \mathbf{\Phi_p q_p} \ , \quad \mathbf{\Phi_p} = [\varphi_{\mathbf{p,1}}, \varphi_{\mathbf{p,2}}, ..., \varphi_{\mathbf{p,s}}] \quad \text{and} \quad \mathbf{q_p} = [q_{p,1}, q_{p,2}, ..., q_{p,s}] \ , \tag{7}$$

where $\varphi_{\mathbf{p,i}}$ are the POD modes and $q_{p,i}$ denote the coordinates in the POD subspace and $s$ is the number of realizations of the random vector (also called snapshots). This leads to an optimization problem with the following objective function (cf. Qu [2])

$$\epsilon^2(l,t) = E\left\{\|\mathbf{x}-\mathbf{x}(l)\|^2\right\} \rightarrow min \tag{8}$$

subject to the orthonormality condition (cf. Qu [2])

$$\varphi_{\mathbf{p,i}}^{\mathbf{T}} \varphi_{\mathbf{p,j}} = \delta_{ij} \ (i,j = 1,2,...,s) \ . \tag{9}$$

The transformation into the $l$-dimensional POD subspace is a truncation of the first $l$ lower POD modes (cf. Qu [2])

$$\mathbf{x}(l) \approx \mathbf{\Phi_p q_p} \ , \quad \mathbf{\Phi_p} = [\varphi_{\mathbf{p,1}}, \varphi_{\mathbf{p,2}}, ..., \varphi_{\mathbf{p,l}}] \ , \quad l < s \ll n \ . \tag{10}$$

In structural dynamics, systems are discretized in space and time and the random vector is realized by $s$ observations at different time instances (cf. Han and Feeny [25])

$$\mathbf{X_s} = [\mathbf{x_{t_1}}, \mathbf{x_{t_2}}, ..., \mathbf{x_{t_s}}] = \begin{pmatrix} x_1(t_1) & \cdots & x_1(t_s) \\ \cdots & \cdots & \cdots \\ x_n(t_1) & \cdots & x_n(t_s) \end{pmatrix} \ . \tag{11}$$

These observation vectors $\mathbf{x_{t_i}}$ are called snapshots and, therefore, in the literature often the observation matrix $\mathbf{X_s}$ is called snapshot matrix containing $s$ snapshots (observations). $\mathbf{x_{t_i}}$ can be measurements or solution vectors of a dynamical system at different time instances ([21]). If $\boldsymbol{\mu}$ is the expectation of all observations, then the sample covariance matrix $\mathbf{\Sigma_s}$ of this random vector, which is realized by the observation matrix, is defined by (cf. Kerschen et. al. [26])

$$\mathbf{\Sigma_s} = E\{(\mathbf{x}-\boldsymbol{\mu})^T(\mathbf{x}-\boldsymbol{\mu})\} \ . \tag{12}$$

The POD modes and the POD values are defined by the eigensolution of the sample covariance matrix. If the data have zero mean, the covariance matrix is (cf. Kerschen et al. [26])

$$\mathbf{\Sigma_s} = \frac{1}{s}\mathbf{X_s}^T\mathbf{X_s} \tag{13}$$

and the POD is realized by the singular value deStructurescomposition (SVD) of the observation matrix $\mathbf{X_s}$. The POD modes $\varphi_{\mathbf{p,i}}$ are equal to the left singular vectors and the POD values $\lambda_{p,i}$ to

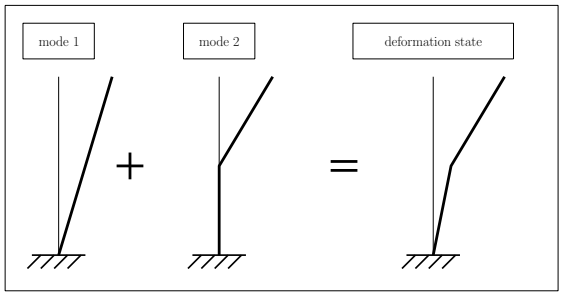



Figure 1: Appropriate mode superposition

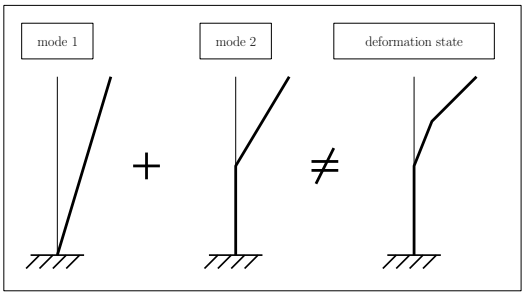



Figure 2: Inappropriate mode superposition

the singular values of $\mathbf{X_s}$, which are all real and positive and arranged in a rectangular diagonal matrix in descending order. The energy, which is contained by the snapshot matrix, is defined by the summation of the POD values, i.e. $V = \sum_{i=0}^{s} \lambda_{p,i}$. As a consequence, the energy ratio of the $i_{th}$ POD mode is (cf. Kerschen et al. [26])

$$V_i = \frac{\lambda_{p,i}}{\sum_{i=0}^{s} \lambda_{p,i}} \ . \tag{14}$$

In structural dynamics applications the sum of only a few POD values often captures $99.99$ percent of the total energy included in the observation matrix, which reflects the big advantage of the POD, i.e. the property of optimality with respect to energy in a least square sense.

Practically, nonlinear effects must be sufficiently captured in the snapshot matrix (11) derived from the representative earthquake exciStructurestation. This is in order to create the capability of providing the possible nonlinear responses that can appear in the response history of the structure excited by another earthquake event. Consequently, nonlinear effects that are not contained in the snapshot matrix, such as plastic hinges or nonlinear sliding of a friction isolator cannot be displayed in the system response. This issue is depicted qualitatively in Figures 1 and 2. Two fictitious deformation states of a simple surrogate model are shown. Obviously, the two deterministic modes in Figure 1 are sufficient in order to represent the deformation state of the model. On the contrary, the set of deterministic modes is insufficient to represent the deformation state in Figure 2 and, consequently, the existing set of modes must be expanded by an additional mode in order to ensure a representation of the correct deformation state. Related to this paper, the conclusion is to capture nonlinear structural reactions in the snapshot matrix in order to provide accurate approximations of the structural response excited by different earthquake events.

Structures

## 4 The new strategy

In this strategy, a snapshot matrix, which contains information about the response of the structure including nonlinear actions, must be found. Therefore, the set of full-order equations of motion of a structure is integrated numerically over the time history of the whole or a part of a representative earthquake excitation and the snapshots are selected from a chosen amount of response vectors, i.e. the displacement vectors at different time instances. Evaluation of the snapshot matrix comprises the time-consuming part of this MOR approach, therefore, the most customary and time-saving approach is to integrate over a small time window in the beginning of a representative earthquake excitation. Nevertheless, in order to increase the probability of recording sufficient linear and nonlinear response patterns, it seems to be reasonable to spread

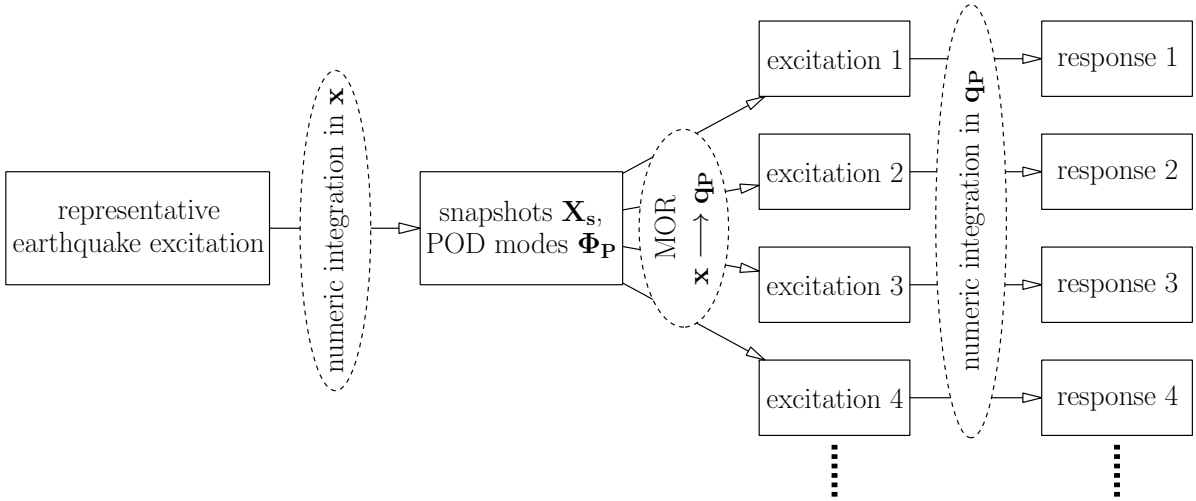



Figure 3: Approach of the new strategy

the snapshot records in equidistant time instances over the whole time history of the earthquake response. Additionally, it is of highest advantage to evaluate the full solutions to more than one earthquake event and spread the snapshots over all of the response histories. This procedure increases the probability of capturing all necessary motion patterns in order to provide an accurate representation by the reduced set of equations. Different earthquake excitations can affect nonlinearities, such as plastic hinges, at different parts of the structure. However, the goal is to define as much nonlinear effects as possible in order to be prepared to assemble an adequate reduced order model of Structuresthe structure excited by new earthquake excitations.

The transformation matrix $\mathbf{\Phi_p}$, containing the POD modes is calculated from the snapshot matrix by applying the SVD algorithm. Afterwards the transformation into the reduced subspace is performed in the same manner, if the classical method of modal truncation would be applied to a linear system. The low-order set of equations of motion is then

$$\widetilde{\mathbf{M}}\ddot{\mathbf{q}}_\mathbf{P} + \widetilde{\mathbf{C}}\dot{\mathbf{q}}_\mathbf{P} + \widetilde{\mathbf{R}} = \widetilde{\mathbf{F}} \,, \tag{15}$$

where $\widetilde{\mathbf{M}} = \mathbf{\Phi}_\mathbf{P}^\mathrm{T}\mathbf{M}\mathbf{\Phi}_\mathbf{P}$ and $\widetilde{\mathbf{C}} = \mathbf{\Phi}_\mathbf{P}^\mathrm{T}\mathbf{C}\mathbf{\Phi}_\mathbf{P}$ are mass- and stiffness matrices and $\widetilde{\mathbf{F}} = \mathbf{\Phi}_\mathbf{P}\mathbf{F}$ is the excitation vector in the POD reduced subspace. The reduced vector of the inner restoring forces $\widetilde{\mathbf{R}}$ is still dependent on the displacement in the physical coordinate $\mathbf{x}$,

$$\widetilde{\mathbf{R}} = \mathbf{\Phi}_\mathbf{P}^\mathrm{T}\mathbf{R}(\mathbf{\Phi}_\mathbf{P}\mathbf{q}_\mathbf{P}) = \mathbf{\Phi}_\mathbf{P}^\mathrm{T}\mathbf{R}(\mathbf{x}) \,. \tag{16}$$

Consequently, the vector of the inner restoring forces $\mathbf{R}(\mathbf{x})$ has to be evaluated from the physical model in the full-order coordinates in every calculation time step. Furthermore, the equations of motion in the reduced-order set are not decoupled and have to be solved numerically. Finally, after the time integration procedure is conducted, the solution vector $\mathbf{q_P}$ dependent on time is transformed back into the physical coordinate $\mathbf{x}$. A visualization of this approach is depicted in Figure 3.

The first remarkable advantage of this procedure is not only, as the matter of fact, the low-dimensional matrix operations, but more importantly is the substantial numerical benefit when applying an explicit time integration scheme such as the central difference method. The critical time step $\Delta t_{cr}$, which is $\frac{2\pi}{\omega_n}$, is inversely proportional to the highest eigenfrequency $\omega_n$. Consequently, for finely meshed structures a large number of loops has to be executed in order to

perform the full time integration procedure to solve the equation of motion (1). On the contrary, the critical time step of the POD reduced system described in equation (15) is much larger and, therefore, the number of calculation time steps is much smaller, which requires a fraction of computational effort compared to full-order system. These numerical issues are discussed in a similar manner by Gutierrez and Zaldivar [30] applied to modal truncation. For the numerical benefit of the combination of the POD with explicit numeric time integration, the reader is referred to Bamer and Bucher [18]. The second big advantage of the new strategy is that the actual time consuming process, which is the evaluation of the snapshot matrix, is only executed once at the beginning of the whole calculation procedure. This a priori assumption of nonlinear mode patterns makes sense if the excitations show physical "similarities", which is the case in earthquake analysis, where a considerably small number of lower modes is mainly affected.

## 5 Practical application

In addition to dealing with the development of the introduced POD-based MOR approach, it is within this section to represent the application of the new proposed MOR strategy on a realistic example. For this purpose, a dynamic structural model of a medical complex, according to its constructional plan, was derived. A schematic three-dimensional sketch of the building is depicted in Figure 4.

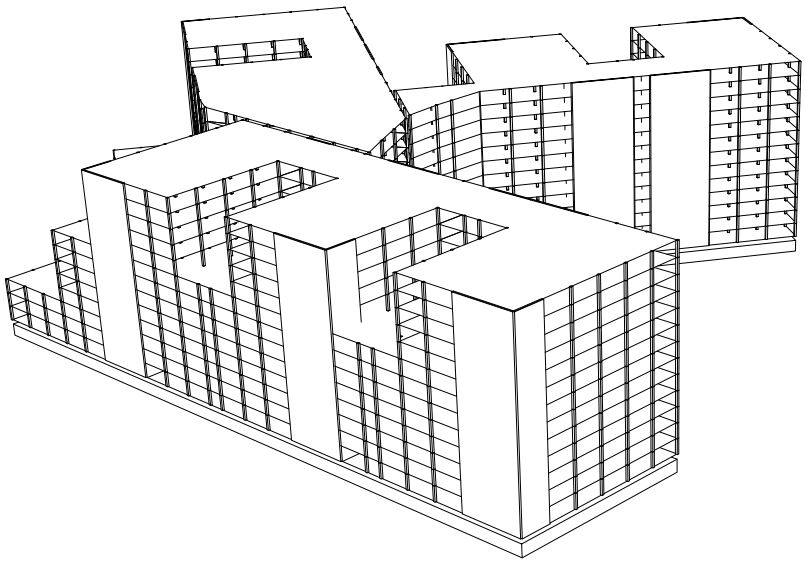


Figure 4: Three-dimensional visualization of the building construction

As shown in Figure 4, the building structure exhibits complex geometries. As a result, it seems to make sense to discretize the geometry by a finite element model in order to capture the main dynamic specifications.
If such a structure with medical function is located in an earthquake prone region, one way to improve its seismic performance can be realized through base isolation by means of frictional pendulum bearings. Consequently, the analytical simulations demand large computational time and storage due to the presence of nonlinearity imposed by those frictional isolators. In the following, firstly, the structural system specifications and implementation of frictional bearings

are presented. Then, the displacement responses to a set of six earthquake events are evaluated. The numerical evaluations compare the new introduced strategy, as an alternative means, with the iterative Newmark integration scheme, which is known as an efficient and exact method.

### 5.1 Structural system and model specifications

The building structure consists of three wings, referred to as wing I, II and III. Figure 5 shows a schematic sketch of the ground plan of the building containing the basic dimensions. The floor slabs of each wing are separate from the others except for the basement slab, which is indiscrete over all three wings. This means that all three wings are coupled through this slab and they work all together during earthquake excitations. However, the distance between the wings, which are connected by the basement slab, is about $1.5$ meters, consequently, contact problems induced by ground motion are not considered in the computations.
The grid indicates the location of the columns and the binding beams, and the red lines indicate the location of the shear walls, which are responsible for the lateral reinforcement. The regular distance between the columns is $6.5\ [m]$. The building structure has three stories below the ground level, while the highest parts of the building above ground level have $13$ stories and the remaining parts have eight stories including the basement levels. Therefore, the plan of the structure is irregular along its height along with the irregularities in the horizontal area. The dashed lines define the area, where the building is only located below the ground. The height of one story is three meters; this leads to a total construction height of $42$ meters.

Below the three stories at the basement level, there is the indiscrete slab on the top of the isolators at level of $-9.00\ [m]$. Below this slab, along each of the columns, a single friction pendulum (FP) bearing system is attached. Figure 6 depicts a part of the cross section A-A of the basement level shown in Figure 5. The horizontal diameter of the FP system is $2.00$ meter. Thus, the dimension of the quadratic cross section of the columns in the basement and FP story is $2.00 \times 2.00\ [m^2]$, while in the remaining stories the columns are modeled as quadratic cross sections with the dimensions $0.40 \times 0.40\ [m^2]$. All FP bearings have the same radius of the concave surface, which is equal to $3.00$ meters.

A representative full-scale finite element model of the building structure was created in the software package slangTNG [39]. The shear walls and slabs were modeled by shell/plate elements and the columns and beams by beam elements. A linear elastic material was considered for the modeling purpose (Young's modulus $E = 3.5 \cdot 10^{10}\ [\frac{N}{m^2}]$, Poisson's ratio $\nu = 0.3\ [-]$, density $\rho = 2500\ [\frac{kg}{m^3}]$). Nonlinear FP elements, whose implementation in slangTNG is presented in section 5.2, are assigned below the lowest basement plate of the structure. The total number of degrees of freedom is $33000$.

### 5.2 Dynamic model of the frictional pendulum element

This is to present how the frictional pendulum (FP) element in the finite element model of the structure behaves. The geometrical diagram of the FP element, which is realized as a spherical shell, is defined in Figure 7. As depicted, $R$ denotes the radius of the concave spherical surface and the origin of the local coordinate system is chosen to be in the center of the sphere. The position vector of the slider is described by $\mathbf{U} = [u, v, w]^T$. Since the desired behavior of the FP element is an in-plane elasto-plastic bidirectional action, the change of the vertical position $w$ can be neglected. Accordingly, the displacement of the FP element is reduced to an in-plane motion defined only by the components $u$ and $v$, i.e. $\mathbf{U} = [u, v]^T$. This simplification makes sense as the radius $R$ is much larger relative to the horizontal displacement

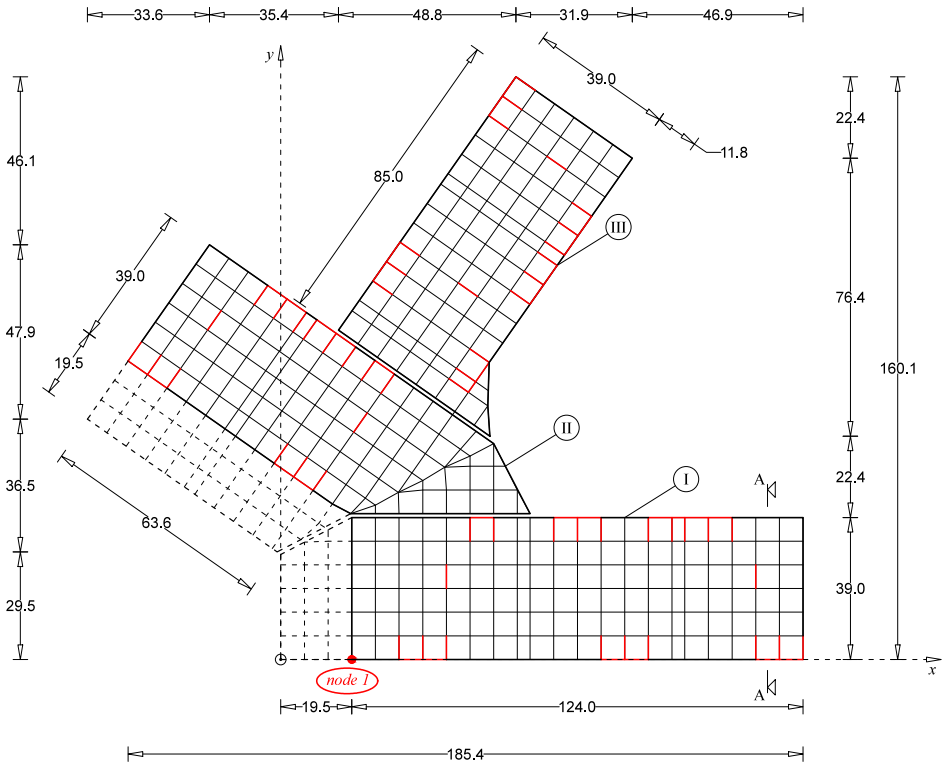



Figure 5: Schematic ground plan, building construction (units in meters), output node 1 $[19.5, 0.0, -9.5]^T$ $[m]$

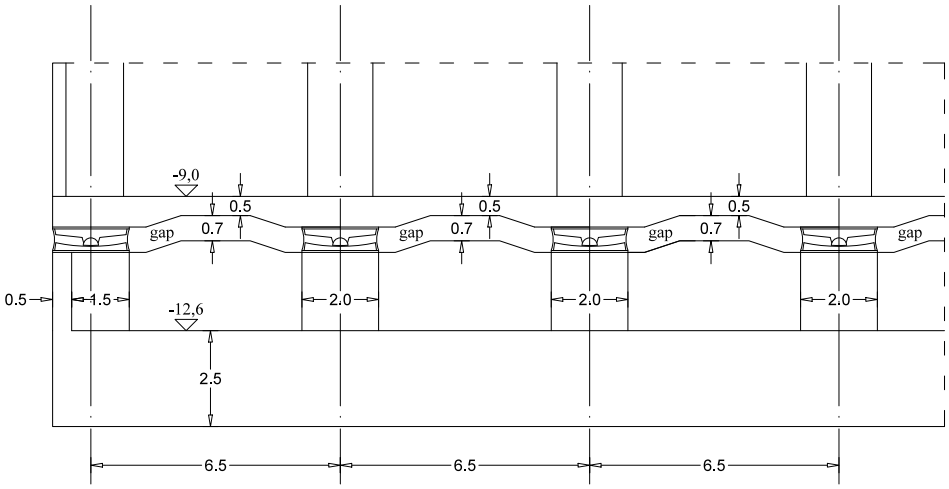



Figure 6: A schematic cutout of the vertical section A-A of the basement levels presented in Figure 5 (units in meters)

$|\mathbf{U}| \approx \sqrt{u^2+v^2}$. The equivalent representation of such an element together with the acting forces on it is represented in Figure 8.
The horizontal force equilibrium of the dynamical system is

$$\mathbf{F_{Fr}} + \mathbf{F_k} = \mathbf{F_{ex}} \ , \tag{17}$$

where $\mathbf{F_{Fr}}$ and $\mathbf{F_k}$ are the elasto-plastic frictional- and centring force and $\mathbf{F_{ex}} = [F_x, F_y]^T$ accounts for the interacting horizontal force, which couples the FP element to the super structure. Following, the force equilibrium is split into two parts as two dynamic situations can occur: situation stick and situation slide. The force equilibrium during the situation stick yields to

$$\mathbf{F_{ex}} = \underbrace{k_1 \left\{ \begin{array}{c} u \\ v \end{array} \right\}}_{\mathbf{F_k}} + \underbrace{k_2 \left\{ \begin{array}{c} \Delta u \\ \Delta v \end{array} \right\}}_{\mathbf{F_{Fr}}} \quad \text{if} \quad |\mathbf{F_{ex}} - \mathbf{F_k}| < \mu N \ . \tag{18}$$

This relation renders a linearly-elastic system, where the friction coefficient $\mu$ must be a value between $0$ and $1$ (about $0.04$ for a realistic implementation of the friction bearing) and the normal contact force $N$ acts orthogonal to the contact area of the slider and the concave surface. The vector $\mathbf{\Delta U} = [\Delta u, \Delta v]^T$ defines the radial distance with respect to the current sticking point of the slider if the sticking condition is true.
During the situation slide the FP element is described by the following horizontal force equilibrium

$$\mathbf{F_{ex}} = \underbrace{k_1 \left\{ \begin{array}{c} u \\ v \end{array} \right\}}_{\mathbf{F_k}} + \underbrace{\frac{\mu N}{|\dot{\mathbf{U}}|} \left\{ \begin{array}{c} \dot{u} \\ \dot{v} \end{array} \right\}}_{\mathbf{F_{Fr}}} \quad \text{if} \quad |\mathbf{F_{ex}} - \mathbf{F_k}| \geq \mu N \ , \tag{19}$$

where $\dot{\mathbf{U}} = [\dot{u}, \dot{v}]^T$ is the velocity vector.
In both relations, i.e. Eq. (18) and Eq. (19), the centring force $|\mathbf{F_k}| = k_1 r = k_1 \sqrt{u^2+v^2}$ acts linearly orthogonal to the vertical axis through the deepest point of the surface and the center of the sphere. The fact that the centring force is linear indicates that the spherical shell of the real system is approximated by the paraboloid, whose potential energy increases with $\frac{W}{R} r$ in radial distance from the deepest point, i.e. the stiffness is inversely proportional to the radius of the sphere $k_1 = \frac{W}{R}$.

The frictional force $\mathbf{F_{Fr}}$ is modeled either linearly elastic or elastic-perfectly plastic as presented in Eqs. (18) and (19), respectively. Note that the force corresponding to $\mathbf{\Delta U}$ accounts for the elastic behavior of the bearing coating material, in a small elastic range (situation stick, Eq. (18)) and acts towards the current location of the slider (not the center of the concave sphere). Generally, the implementation of a realistic model requires $k_2$ to be much larger than $k_1$, i.e. $k_2 \gg k_1$. During the situation slide, the frictional force acts in opposite direction to the velocity with the magnitude (perfectly plastic) $\mu N$. This is discussed in Eq. (19). Another point regarding Eq. (18) is that the reacting force $N$ is assumed to be constant throughout the calculation procedure. This is justified by the following reasons: Firstly, just $x$ and $y$ components of the exciting ground motion are taken into account for the computations. Secondly, the motion has already been simplified to be planar and therefore no additional force component due to vertical motion is generated. Finally, in our preliminary analysis, the uplift force on the isolator slap was observed to be extremely small in comparison with the downward force due

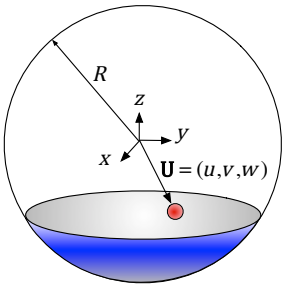



Figure 7: Geometric definitions of the FP element

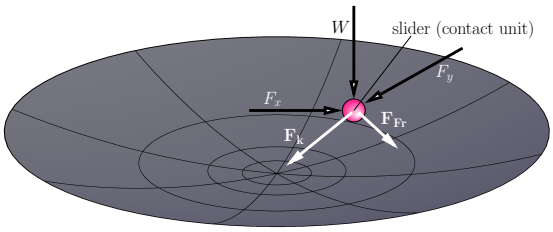



Figure 8: Internal specifications of the FP element; $F_x$, $F_y$, $N$, recentering force $\mathbf{F_k}$, friction force $F_{Fr}$

| **Event** | year | location | $n_t$ | $T$ | $d$ | $M$ | PGA |
|---|---|---|---|---|---|---|---|
| Bam | 2003 | Iran | 1995 | 19.95 | - | 6.6 | 7.16 |
| Imperial Valley | 1979 | California / Huston Road | 3905 | 39.05 | 10 | 6.5 | 4.79 |
| Landers | 1992 | California / Barstow | 4932 | 49.32 | 36 | 7.3 | 4.13 |
| Loma Prieta | 1989 | California / Gilroy | 2507 | 25.07 | 12 | 7.0 | 9.51 |
| North Palm Springs | 1986 | California / Palm Springs | 6009 | 60.09 | 6.7 | 6.0 | 9.99 |
| Northridge Rinaldi | 1994 | California / Newhall | 1200 | 12.00 | 6.7 | 6.7 | 5.23 |

Table 1: Earthquake excitation list; $n_t$ $[-]$ number of time steps, $T$ $[s]$ duration of the record, $d$ $[km]$ distance from epicenter, $M$ moment magnitude, PGA $[m/s^2]$ peak ground acceleration

to the weight of the structure. Considering the above-mentioned fact together with the force diagram given in in Figure 8, follows that the normal contact force $N$ is approximately constant and equal to the weight induced force of the super structure, $W$, i.e, $N = W$ in Eq. (19).

The FP bearing element governed by Eqs. (17) to (19) has been implemented in the software package slangTNG [39]. For a comparable study on this implemented friction pendulum system, the experimental work of Mosqueda et al. [40] is suggested. For additional information about friction pendulum systems the reader is referred to the relevant literature (e.g. [33], [34], [35], [36], [37] and [38]). More detailed examination of this topic would lead beyond the scope of this paper, which should focus more on the methodical extension of the new MOR strategy as well as the application on a complex realistic system.

### 5.3 Numerical evaluation

The evaluation of the introduced MOR strategy is dealt with displacement response calculations of the building structure to six different earthquake excitations. The earthquake records are applied in fault-parallel and fault-normal directions. The excitation set includes the Bam earthquake (2003) in Iran and the following five representative events in California, US: Northridge Rinaldi (1994), Imperial Valley (1979), Landers (1992), Loma Prieta (1989), North Palm Springs (1986). Table 1 presents a list of the events taken from the Pacific Earthquake Engineering Research Center (PEER) [31]. Fault-parallel is defined in $x$- and fault-normal in $y$-direction. Concerning the Bam event only a one-dimensional record was accessible, therefore, an excitation angle of $30$ degrees with respect to the $x$-axis was chosen. Calculation outputs are presented including the displacement translational response, namely the degree of freedom $x$ of the red marked node in Figure 5. This is the horizontal component of response in $x$-direction of the slider of the FP element. Additionally, the response regarding in-plane motion of the slider

in $x$- and $y$-direction are provided. The output node is called "node 1" and has the coordinates $[19.5, 0.0, -9.5]^T$ $[m]$, this is depicted in Figure 5. This node defines the location of a moving friction pendulum, it therefore shows directly the nonlinear response behavior of the system.

The method, according to Section 4, begins with the earthquake dynamic analysis of the structure under representative excitation. Therefore, the response $\mathbf{X}_{Bam} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_m}]$ of the full system to the Bam earthquake is evaluated. This full system response is evaluated applying the Newmark method. The response history in $x$-direction of node 1 (output node) is depicted in the left subplot of Figure 9. The motion of the slider is shown in the right subplot of Figure 9.

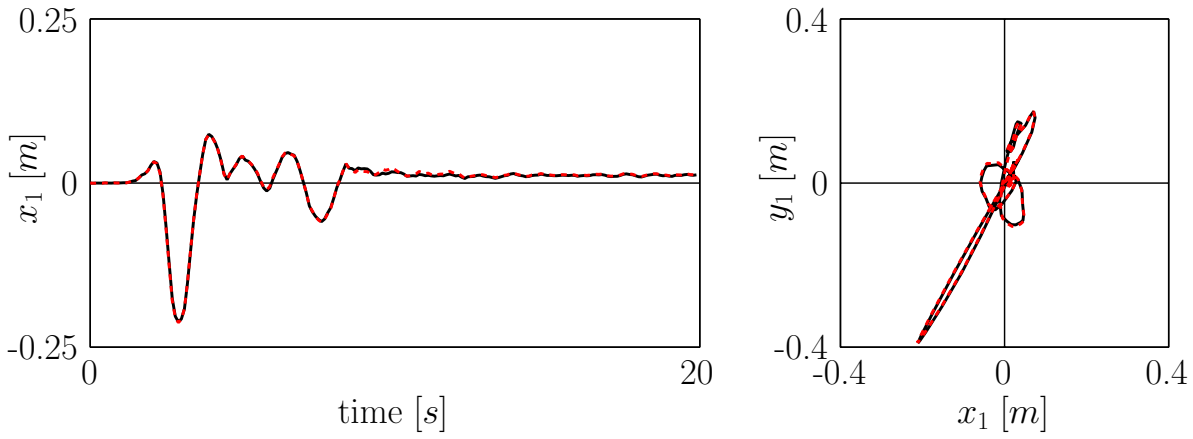



Figure 9: Response functions to the Bam earthquake excitation, full Newmark response and reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

The snapshot matrix $\mathbf{X_S}$ is assembled by taking into account 400 snapshots in equidistant time intervals spread over the whole displacement response history to the Bam earthquake excitation. It is essential to capture the main deformation behavior of the system in this stage, otherwise certain deformation states cannot be constituted by the set of linearly independent vectors (c.f. Figure 2), i.e. the sliding process, as can be seen in Figure 9, has to be recorded in the snapshot matrix. Afterwards, the evaluation of the left singular vectors of the snapshot matrix leads to the POD modes and its singular values to the POD values in descending order. The sum of all POD values define the total energy content of the snapshot matrix. Hence, the number of POD modes that have to be taken into account in order to capture $99,99$ percent of the total energy was evaluated to be 31. A plot of the first 31 singular values (POD values) dependent on the corresponding energy content is shown in Figure 10.

The physical equation of motion is now transformed into the 31-dimensional POD subspace and integrated over the whole time history by application of the explicit central difference scheme. Transformation back into the physical coordinate $\mathbf{x}$ produces the POD-based response of the structure. The integration time step for the reduced time integration is $5 \cdot 10^{-3}$ seconds, which is equivalent to the time step of the Newmark integration scheme, which is applied for the calculation of the exact response. In the response analysis of the reduced order system via central difference no iteration is needed and only a fractional amount of storage compared to the full Newmark method is required. Note that the critical time step of the central difference scheme for the full system is $10^{-7}$ seconds. Therefore, concerning the basic central difference method, an immense speed-up factor of $10^4$ is achieved. Within the presented examples generally a min-

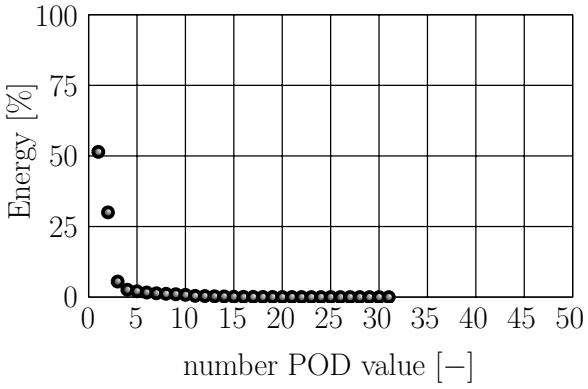



Figure 10: Number of POD values (singular values)

imum speed up factor of about 2 compared to the Newmark algorithm is achieved. However, it has to be added that for this type of nonlinearity, the Newmark integration scheme appears to have a relative slow convergence rate. As a consequence, the computational time by applying the Newmark method can exceed to a period of time, which is comparable if the full central difference integration scheme is applied or even to infinite period of time, if there is no convergence. Following, the new integration strategy gains the benefit of both of the two algorithms, i.e. central difference and Newmark, which is stability without requiring iteration algorithms. The red dashed line in Figure 9 (response to the Bam excitation) shows the response obtained by the new POD strategy, which approximates the full Newmark response accurately. This is not surprising as the snapshots are taken in equidistant time intervals spread over this whole response history. The derived POD transformation matrix, which is the set of POD modes, is now applied to reduce the order of the set of equations for the structure excited by the rest of the presented earthquake events presented in Table 1. It means that the time integration in the full (physical) space no longer has to be performed. As for the next transformations into the POD space only one transformation matrix is applied. This strategy is here called *universal* POD method. Now this approach unveils its similarities to the method of modal truncation, which is mainly applicable to linear systems. Figures 11, 12, 13, 14 and 15 compare the displacement responses applying the new MOR strategy and the Newmark scheme.

Accurate approximations are achieved by means of the proposed MOR strategy. The response to the Imperial Valley earthquake (Figure 11), the Landers earthquake (Figure 12) and the North Palm Springs earthquake (Figure 14) were highly accurate, when the proposed MOR strategy was applied. Concerning the response functions of the Loma Prieta (Figure 13) and the North Palm Springs excitation (Figure 15) small variations concerning the full and the reduced solutions can be observed. This is because the responses to those excitations include deformation states that are not captured in the snapshot matrix and, hence, not in the *universal* POD modes (POD transformation matrix) and thus cannot assemble the exact response history.

## 6 Conclusions

In this paper, a model order reduction (MOR) strategy, which is applicable to the dynamic response analysis of linear and nonlinear structural systems was presented. Usually, the analysis of building structures with complex geometries makes the engineer to create a finite element model with a large number of degrees of freedom, which is associated with computational effort in the response analysis. Therefore, the goal of this paper is to provide a new practical

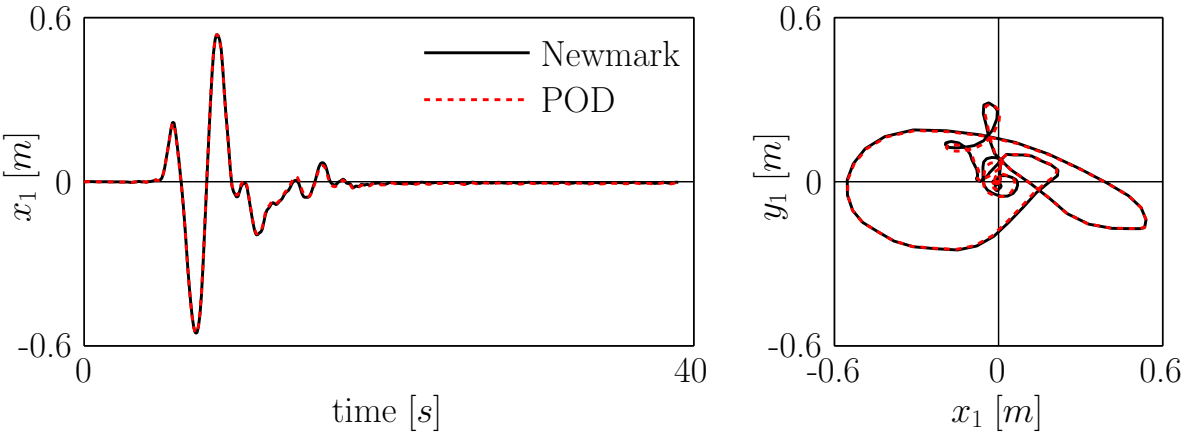



Figure 11: Response functions to the Imperial Valley excitation, full Newmark response and universal reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

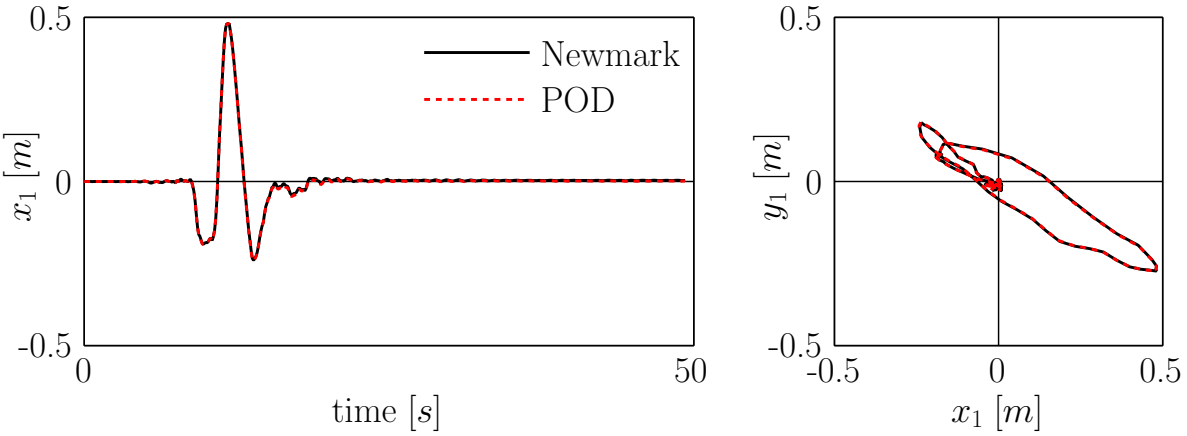



Figure 12: Response functions to the Landers excitation, full Newmark response and universal reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

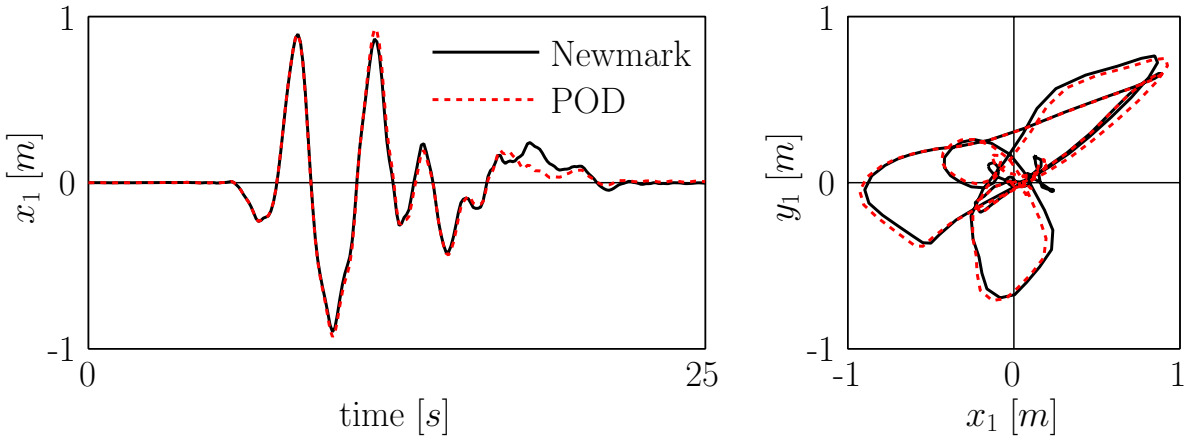



Figure 13: Response functions to the Loma Prieta excitation, full Newmark response and universal reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

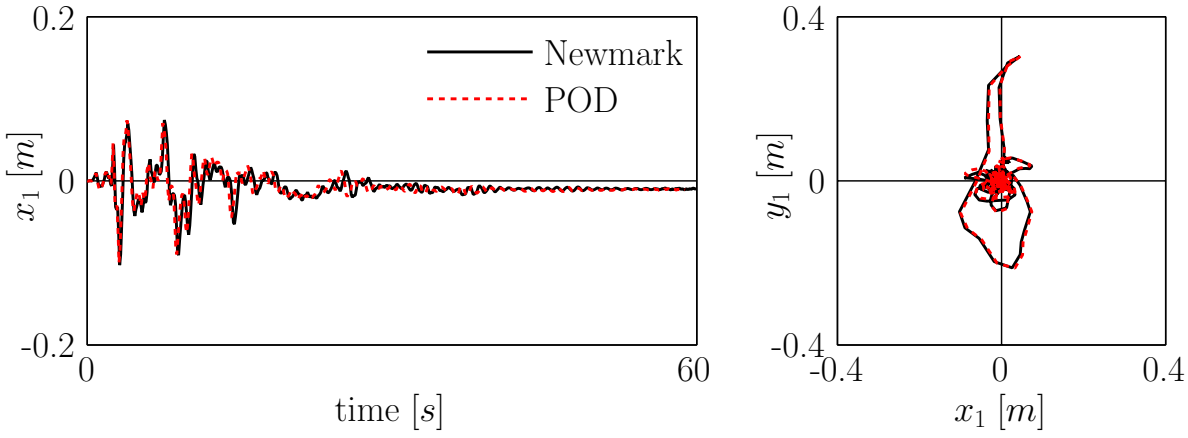



Figure 14: Response functions to the North Palm Springs excitation, full Newmark response and universal reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

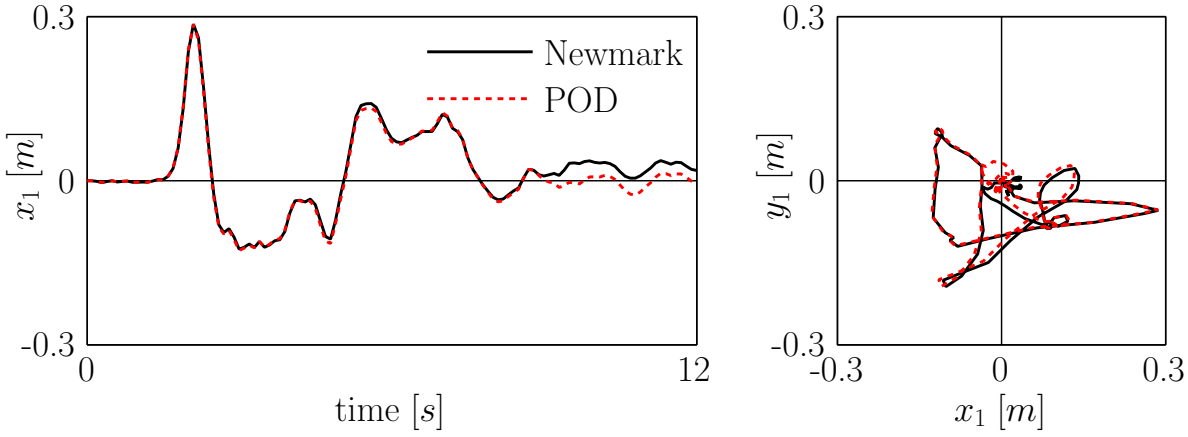



Figure 15: Response functions to the Northridge excitation, full Newmark response and universal reduced POD response; left subplot: displacement $x$ of node 1 dependent on time; right subplot: movement of the slider, displacement $x$ and $y$ of node 1

model order reduction strategy that is simple in application, but also very effective even in the presence of nonlinearities for problems in the field of earthquake engineering and structural dynamics. This strategy is extended based on the proper orthogonal decomposition (POD) method to derive a proper transformation matrix in order to transform the nonlinear systems into another low-dimensional subspace, which demands considerably less computational effort for the response calculation. Once the transformation matrix is derived, the approach of the strategy is similar to the method of modal truncation for linear systems.

In addition to the development of the MOR strategy, its application for the response calculation of a realistic numerical nonlinear example is demonstrated. The example is the displacement response calculation of a building structure serving as a medical complex, which is base-isolated by friction pendulum bearing systems excited by six earthquake excitations. In order to evaluate the accuracy of the introduced approach, the exact structural responses were also calculated by the iterative Newmark method in the full order (physical) coordinates. Numerical evaluations show that accurate approximations can be achieved if nonlinear response patterns of the structure are already captured in the POD snapshots to extract the transformation matrix. The advantage of this strategy is that obviously the transformation matrix is derived just once and it can be used for response calculation of the structure under different earthquake excitations.

Another substantial advantage of the introduced MOR concerns the speed of the response calculations. Firstly, compared to the basic central difference algorithm, the new introduced strategy has a much larger critical time step. Secondly, compared to the Newmark method, which allows usually larger time steps, no iteration procedure is required.

## 7 Acknowledgements

The authors would like to acknowledge financial support from the Austrian Science Funds (FWF) as part of the Vienna Doctoral Programme on Water Resource Systems (DK-plus W1219-N22). This research has also been supported by Stiftung Maurer Söhne (Forschungsförderung Technische Dynamik), which is gratefully acknowledged by the authors. Special thanks of the authors is dedicated to Professor Bernd Markert, head of the Institute of General Mechanics at the RWTH Aachen University, for his critical feedbacks and constant supports.

## REFERENCES


[1] K.J. Bathe, *Finite Element Procedures*. Prentice Hall, 1995.

[2] Z.Q. Qu, *Model order reduction techniques with application in finite element analysis*. Springer London Limited, 2004.

[3] G. Rega, H Troger, Dimension Reduction of Dynamical Systems: Methods, Models, Applications. *Nonlinear Dynamics*, **41**, 1–15, 2005.

[4] P. Koutsovasilis, M. Beitelschmidt, Comparison of model reduction techniques for large mechanical sytems, *Multibody System Dynamics*, **20**, 111–128, 2008.

[5] H.A. Schilders, H.A. Vorst, J. Rommes, *Model Order Reducion, Theory, Research Aspects and Application*. Springer, 2008.

[6] D. Kosambi, Statistics in function space, *Journal of the Indian Mathematical Society*, 1943.

[7] K. Karhunen, *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*, PhD thesis, University of Helsinki, 1947.

[8] M. Loeve, Fonctions al Eatoire de second ordre, *Revue Scientifique*, **48**, 195–206, 1946.

[9] J.P. Cusumano, M.T. Sharkady, B.W. Kimble, Spatial coherence measurements of a chaotic flexible-beam impact oscillator, *American Society of Mechanical Engineers, Aerospace Division (Publication) AD*, **33**, 13–22, 1993.

[10] B.F. Feeny, R. Kappagantu, On the physical interpretation of proper orthogonal modes in vibrations, *Journal of Sound and Vibration*, **211**, 607–616, 1998.

[11] R. Kappagantu, B.F. Feeny, An "optimal" modal reduction of a system with frictional excitation, *Journal of Sound and Vibration*, **224**, 863–877, 1999.

[12] R. Kappagantu, B.F. Feeny, Part 1: Dynamical Characterization of a frictionally excited beam, *Nonlinear Dynamics*, **22**, 317–333, 2000.

[13] R. Kappagantu, B.F. Feeny, Part 2: Proper Orthogonal Modal Modeling of a Frictionally Excited Beam, *Nonlinear Dynamics*, **23**, 1–11, 2000.

[14] Y.C. Liang, H.P. Lee, S.P. Lim, W.Z. Lin, K.H. Lee, C.G. Wu, Proper Orthogonal Decomposition and its Applications - Part 1: Theory, *Journal of Sound and Vibration*, **252**, 527–544, 2002.

[15] G. Kerschen, J.C. Golivani, Physical interpretation of the proper orthogonal modes using the singular value decomposition, *Journal of Sound and Vibration*, **249**, 849–865, 2002.

[16] G. Kerschen, J.C. Golivani, A model updating strategy of non-linear vibrating structures, *International Journal of Numerical Methods in Engineering*, **60**, 2147–2164, 2003.

[17] F. Tubino, L. Carassale, G. Solari, Seismic response of multi-supported structures by proper orthogonal decomposition, *Earthquake Engineering and Structural Dynamics*, *32*, 1639–1654, 2003.

[18] F. Bamer, C. Bucher, Application of the proper orthogonal decomposition for linear and nonlinear structures under transient excitation, *Acta Mechanica*, **223**, 2549–2563, 2012.

[19] P. Koutsovasilis, M. Beitelschmidt, Comparison of model reduction techniques for large mechanical sytems, *Multibody System Dynamics*, **20**, 111–128, 2008.

[20] A.K. Chopra, *Dynamics of Structures, Theory and Applications to Earthquake Engineering*. Prentice Hall, 2001.

[21] A. Chatterjee, An introduction to the proper orthogonal decomposition, *Current Science*, **78**, 808–817, 2000.

[22] S. Volkwein, *Model reduction using proper orthogonal decomposition*. Karl-Franzens-Universität Graz, Institut für Mathematik und Wissenschaftliches Rechnen, 2008.

[23] S. Volkwein, *Proper Orthogonal Decomposition and Singular Value Decomposition*. Karl-Franzens-Universität Graz, Institut für Mathematik und Wissenschaftliches Rechnen, 2008.

[24] P. Holmes, J.L. Lumley, G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press, 1996.

[25] S. Han, B. Feeny, Application of proper orthogonal decomposition to structural vibration analysis, *Mechanical Systems and Signal Processing*, **17**, 989-1001, 2003.

[26] G. Kerschen, J.C. Golinval, A.F. Vakakis, L.A. Bergman, The Method of Proper Orthogonal Decomposition for Dynamical Characterization and Order Reduction of Mechanical Systems: An Overview, *Nonlinear Dynamics*, **41**, 147–169, 2005.

[27] Z. Qu, Y. Shi, H. Hua, A reduced-order modeling technique for tall buildings with active tunded mass damper, *Earthquake Engineering and Structural Dynamics*, *30*, 349–362, 2001.

[28] A.G. Schemann, H.A. Smith. Vibration control of cable-stayed bridges - part 1: modelling issues, *Earthquake Engineering and Structural Dynamics*, **27**, 811–824, 1998.

[29] E. Gutierrez, J.M. Zaldivar, The application of Karhunen-Loeve, or principal component analysis method, to study the non-linear seismic response of structures, *Earthquake Engineering and Structural Dynamics*, **29**, 1261–1286, 2000.

[30] E. Gutierrez, J.J.L. Cela, Improving explicit time integration by modal truncation techniques, *Earthquake Engineering And Structural Dynamics*, **27**, 1541–1557, 1998.

[31] Pacific Earthquake Engineering Research Center (PEER), *Homepage: http://nisee.berkeley.edu*, 2015.

[32] C. Bucher, Stabilization of explicit time integration by modal reduction, *Trends in computational structural mechanics*, 2001.

[33] J. Almazan, J. Llera, J. Inaudi, Modelling aspects of structures isolated with the frictional pendulum system, *Earthquake Engineering and Structural Dynamics*, **27**, 845–867, 1998.

[34] J. Llera, J. Almazan, An experimental study of nominally symmetric and asymmetric structures isolated with the FPS, *Earthquake Engineering and Structural Dynamics*, **32**, 891–918, 2003.

[35] D. Ordonez, D. Foti, L. Bozzo, Comparative study of the inelastic response of base isolated buildings, *Earthquake Engineering and Structural Dynamics*, *32*, 151–164, 2003.

[36] K. Ryan, A.K. Chopra, Estimating the seismic displacement of friction pendulum isolators based on non-linear response history analysis, *Earthquake Engineering and Structural Dynamics*, **33**, 359–373, 2004.

[37] T. Ray, A. Sarlis, A. Reinhorn, M. Constantinou, Hysteretic models for sliding bearings with varying frictional force, *Earthquake Engineering and Structural Dynamics*, **42**, 2341–2360, 2013.

[38] C. Bucher, Probability-based optimal design of friction-based seismic isolation devices, *Structural Safety*, **31**, 500–507, 2009.

[39] C. Bucher, Open source structural dynamics and FE software (slangTNG), *Homepage: http://info.tuwien.ac.at/bucher/Private/Welcome.html*, 2015.

[40] G. Mosqueda, A. Whittaker, G. Fenves, Characterization and Modeling of Friction Pendulum Bearings Subjected to Multiple Components of Excitation, *Journal of Structural Engineering ASCE*, **130**, 433–442, 2004.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ACTIVE SUBSPACES FOR THE PRELIMINARY FLUID DYNAMIC DESIGN OF UNCONVENTIONAL TURBOMACHINERY

**Juan S. Bahamonde**[1], **Matteo Pini**[1] **and Piero Colonna**[1]

[1] Propulsion and Power
Aerospace Faculty, Delft University of Technology
Kluywerweg 1, 2629HS, Delft, The Netherlands
e-mail: {S.Bahamonde,M.Pini,P.Colonna}@tudelft.nl

**Keywords:** Organic Rankine Cycle, active subspaces, surrogate models, turbine design.

**Abstract.** *The fluid dynamic preliminary design of unconventional turbomachinery is customary done with meanline design procedures coupled with gradient-free optimizers. This method features various drawbacks, since it might become computationally expensive, and it does not provide design insights or guidelines to the designer. This work proposes a strategy to abate this disadvantages, namely, the construction of a reduced-order model by means of active subspaces, and the use of the surrogate combined with a gradient-based optimizer. The case study is the design optimization of a Organic Rankine Cycle radial inflow turbine. The results show that active subspaces exist for this application, and that it is possible to construct a surrogate with an approximate error of* $\pm 1\%$ *for the total-to-static efficiency. Additionally, the optimization using the surrogate leads to accurate results and a computational cost at least four times faster. Furthermore, the results reveal that the models for unconventional turbomachinery feature multiple regions containing constrained optima. Active subspace methods thus prove to be a promising alternative for optimization of unconventional turbomachinery.*

## 1 Introduction

The realization of high-efficiency unconventional turbomachinery (e.g., supercritical CO2 compressors, or Organic Rankine Cycle (ORC) expanders) demands for sound optimization strategies to tackle the preliminary design phase. Meanline methods based on empirical loss correlations (e.g., see Ref. [1]) are typically adopted to execute this stage. The mathematical nature of these correlations yields a highly non-linear system of equations, which is somehow discontinuous and noisy. As a result, in order to solve this challenging design problem, the meanline method is frequently coupled with a gradient-free optimizer (e.g., genetic algorithms, see Ref. [2, 3]).

Although the combination of a meanline code and a gradient-free optimizer proved to be succesful, it features several drawbacks: i) particularly for ORC machines, the computational cost might become an issue, for the optimization must consider different working fluids, turbine architectures, and a wide range of operating conditions [4]. ii) A genetic algorithm is a heuristic, hence it likely yields sufficiently good solutions, though it does not guarantee an optimum. iii) Valuable information regarding the physical response of the system is lost in gradient-free methods (e.g., sensitivity of the objective to individual design inputs).

This work aims to to investigate the capability of a new technique to construct reduced-order models to circumvent the aforementioned disadvantages. Active subspace methods (AS) reveal the dominant directions of the gradient of a scalar function. By using these directions, it is possible to transform a multidimensional input space to a lower-dimension version formed by the so-called active variables [5]. If applicable to turbomachinery design problems, this method holds great advantages: i) the computations with surrogates might require lower computational time. ii) Reduced-order models can be used to build response surfaces, thus permitting the designer to perform a visual inspection of the optimum solution and to switch to a gradient-based optimizer. iii) Finally, the response surface can be used to derive relevant physical insights.

This work presents the results of the application of the active subspaces to the design of a single-stage Organic Rankine Cycle (ORC) turbine. A meanline code for the fluid dynamic preliminary design of turbomachinery is used to demonstrate the existence of active subspaces, and to build surrogate models of the total-to-static efficiency and the machine geometry. Furthermore, a genetic algorithm optimization is compared against a gradient-based method using the surrogates. Ultimately, the surrogate model is exploited to gain physical insight on the design problem.

This document is structured as follows: Section 2 presents the methods used to construct the surrogate models and to perform the constrained optimization. Section 3 discusses the application of the methods to the design/optimization of a mini ORC radial inflow turbine. The document ends with some concluding remarks in Section 4.

## 2 Methods

This section describes the standard procedure conceived to provide the preliminary fluid dynamic design of non-conventional turbomachinery, i.e., a meanline code coupled with a genetic algorithm. It also reports the description of the process to construct the surrogate via AS, and the corresponding definition of the gradient-based optimization. The method has been implemented in a engineering programming environment [6], which is coupled with a computational library for the calculation of the fluid thermophysical properties [7].

### 2.1 Software for the turbine preliminary design

*zTurbo* is a meanline code for the fluid dynamic preliminary design of turbomachinery. It uses empirical loss correlations for the estimation of the turbine performance (see Ref. [1]), and it has been implemented in Fortran. More details about this tool can be found in the paper of Pini et al. [8]

Mathematically, *zTurbo* can be expressed as a vector function,

$$\mathbf{\Psi} = \mathbf{\Psi}(\mathbf{x}), \quad \mathbf{x} \in [-1, 1]^m, \tag{1}$$

where $\mathbf{x}$ is a normalized and centered input space of size $m$. The input space might include variables like (per stage) rotational speed, inlet diameter, blade height, etc. Function $\mathbf{\Psi}$ yields an array output, so that

$$\mathbf{\Psi} = [\eta_{\text{ts}}, \phi_i], \quad i \in \{1, ..., N\}, \tag{2}$$

where $\eta_{\text{ts}}$ is the turbine total-to-static efficiency, and $\phi_i$ are outputs that will be compared against $N$ number of constraints. *zTurbo* is imported to the programming environment as a function.

#### 2.1.1 Turbine design optimization by means of genetic algorithms

As a design tool, *zTurbo* delivers the full geometry of the turbine after the calculation is done, which precludes the possibility of constraining the optimization from the input space. As a consequence, the function is defined such that the solutions violating the constraints are discarded, i.e.,

$$\eta_{\text{ts}}(\mathbf{x}) = \begin{cases} -\infty, \text{ if } \phi_i(\mathbf{x}) < \psi_{\text{min},i}, \quad i \in \{1, ..., N\}, \\ \text{or,} \\ -\infty, \text{ if } \phi_i(\mathbf{x}) > \psi_{\text{max},i}, \\ \text{else,} \\ \eta_{\text{ts}}, \end{cases} \tag{3}$$

where $\psi_i$ is the $i^{\text{th}}$ constraint defined by a minimum and a maximum value. Due to the definition of the objective function from Eq. 3, the optimization problem reads

$$\text{Maximize} \quad \eta_{\text{ts}}(\mathbf{x}). \tag{4}$$

The built-in genetic algorithm provided by the host software is then used to solve the optimization problem [9].

### 2.2 Surrogate modeling by means of Active Subspace methods

This section discusses briefly the construction of the surrogate model based on the Active Subspace method. For an extense description, the reader is directed the work of Constantine [5].

Consider the scalar function

$$f = f(\mathbf{x}), \quad \mathbf{x} \in [-1, 1]^m, \tag{5}$$

where $\mathbf{x}$ is an array with $m$ number of continuous, centered and normalized inputs. Additionally, $f$ has has to be smooth and differentiable. The objective of the active subspace method is to approximate $f$ to a new function $\hat{f}$, which features a lower dimensional input space, i.e.,

$$f(\mathbf{x}) \approx \hat{f}(\mathbf{x}_{\text{ac}}), \quad \mathbf{x}_{\text{ac}} \in \mathbb{R}^n, \quad n < m. \tag{6}$$

Note the *hat* accent, which indicates $\hat{f}$ is an approximation of the exact function. The new input space ($\mathbf{x}_{\text{ac}}$) is constituted by the so-called "active variables", which are a linear combination of the original inputs. For instance, assuming that the input space is reduced to a single dimension ($n = 1$):

$$f(\mathbf{x}) \approx \hat{f}(x_{\text{ac}}), \quad x_{\text{ac}} = a_1 x_1 + a_2 x_2 + a_3 x_3 + ... + a_m x_m. \tag{7}$$

Infinite number of linear combinations are available. The active subspace method provides a strategy to discard the trivial linear combinations, and select the important ones. A short explanation on this procedure follows.

The gradient of $f$ is defined here as the column vector of partial derivatives,

$$\nabla f_{\text{x}} = \left[ \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_{\text{i}}}, ..., \frac{\partial f}{\partial x_{\text{m}}} \right]^T. \tag{8}$$

This method relies on the analysis of matrix $\boldsymbol{C}$,

$$\boldsymbol{C} = \mathbb{E}\left[ (\nabla f_{\text{x}})(\nabla f_{\text{x}})^T \right], \tag{9}$$

where $\mathbb{E}$ is the expectancy of the function inside the square brackets. $\boldsymbol{C}$ might be considered as the uncentered covariance of the gradient vector. Therefore, it contains the information of the directions on which $f$ changes the most (in average).

The exact evaluation of $\boldsymbol{C}$ requires multidimensional integration. Nonetheless, it is possible to approximate $\boldsymbol{C}$ by sampling throughout the input space, namely

$$\hat{\boldsymbol{C}} = \frac{1}{M} \sum_{j=1}^{M} (\nabla f_{\text{x}})(\nabla f_{\text{x}})^T, \tag{10}$$

where $M$ is the number of samples. Note again the *hat* accent, which indicates that $\hat{\boldsymbol{C}}$ is an approximation of the exact covariance matrix. It is important to mention that much of the active subspaces research aims to estimate the accuracy of $\hat{\boldsymbol{C}}$, and to study the corresponding strategies to improve this approximation [5].

$\hat{\boldsymbol{C}}$ is real and symmetric, and therefore a real eigenvalue decomposition is feasible, as in

$$\hat{\boldsymbol{C}} = \hat{\boldsymbol{W}} \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{W}}^T. \tag{11}$$

Large gaps between the eigenvalues indicate that there are directions where the function changes the most. $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{W}}$can be accordingly split in two families, so that

$$\hat{\boldsymbol{W}} = \begin{bmatrix} \hat{\boldsymbol{W}}_{\text{ac}} & \hat{\boldsymbol{W}}_{\text{ic}} \end{bmatrix}, \quad \hat{\boldsymbol{\Lambda}} = \begin{bmatrix} \hat{\boldsymbol{\Lambda}}_{\text{ac}} & \hat{\boldsymbol{\Lambda}}_{\text{ic}} \end{bmatrix}. \tag{12}$$

The subscripts ac and ic correspond to active and inactive, hence the "ac" matrices contain the largest eigenvalues (separated by a gap) and the corresponding eigenvectors. The input space is then rotated according the $\hat{\boldsymbol{W}}_{\text{ac}}$, in order to "hide" the directions where the function variability is lower, namely

$$\mathbf{x}_{\text{ac}} = \hat{\boldsymbol{W}}_{\text{ac}}^T \begin{bmatrix} x_1 \\ ... \\ x_m \end{bmatrix}, \tag{13}$$

Notice that the number of dimensions of $\mathbf{x}_{\text{ac}}$ equals the number of selected active eigenvectors.

Due to the sampling required to construct $\hat{\boldsymbol{C}}$, there are $M$ sets of transformations in the active subspace ($\hat{f} : \mathbf{x}_{\mathrm{ac}} \to f$). Ultimately, $f$ is approximated to a reduced-order function, i.e.,

$$f(\mathbf{x}) \approx \hat{f}\Big(\hat{\boldsymbol{W}}_{\mathrm{ac}}^T \mathbf{x}^T\Big), \tag{14}$$

The reduced order model can take any algebraic structure, e.g.,

$$\hat{f}(x_{\mathrm{ac}}) = a_1 x_{\mathrm{ac}}^2 + a_2 x_{\mathrm{ac}} + a_3. \tag{15}$$

Note that in this example that the original $f$ has been transformed to a new function with a one dimensional input or active variable.

### 2.2.1 Surrogate model of *zTurbo*

The nature of the turbomachinery loss correlations makes *zTurbo* inherently unstable, namely, the function outputs might present discontinuities or non-smooth intervals. As a consequence, the design space for the test case (see Fig.1) is selected such that it guarantees a well-behaved $f$.

The objective of this work is performing a constrained optimization, hence surrogate models of additional quantities of interest are also created. Ultimately, *zTurbo* is transformed in a system of equations, as in,

$$\text{Surrogate model} \begin{cases} \hat{\eta}_{\mathrm{ts}} & = \hat{\eta}_{\mathrm{ts}}(\mathbf{x}_\eta), \\ \hat{\phi}_i & = \hat{\phi}_i(\mathbf{x}_{\phi,i}), \quad i \in \{1, ..., N\}, \end{cases} \tag{16}$$

where $\hat{\eta}_{\mathrm{ts}}(\mathbf{x}_\eta)$ is the surrogate model of the objective function, and $\hat{\phi}_i(\mathbf{x}_{\phi,i})$ is the $i^{\mathrm{th}}$ surrogate model corresponding to a *zTurbo* output to be compared against a constraint.

### 2.2.2 Optimization based on gradient-search methods

Once the surrogates are constructed, it is possible to formulate a constrained optimization problem using gradient-based search methods, i.e.,

$$\begin{aligned} \text{Maximize} \quad & \hat{\eta}_{\mathrm{ts}}(\mathbf{x}_\eta), \\ \text{subject to} \quad & \psi_{\mathrm{min},i} \leq \hat{\phi}_i(\mathbf{x}_{\phi,i}) \leq \psi_{\mathrm{max},i}, \quad i \in \{1, ..., N\}. \end{aligned}$$

Since the active variables are a result of a matrix multiplication, it is possible to set the optimization using the original variables, so that the problem reads

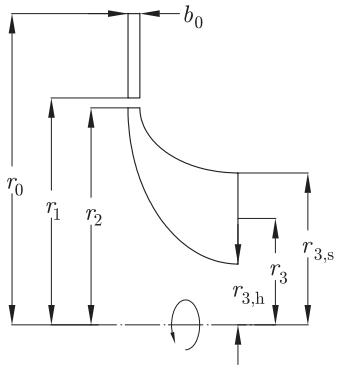
$$\begin{aligned} \text{Maximize} \quad & \hat{\eta}_{\mathrm{ts}}\Big(\hat{\boldsymbol{W}}_\eta^T \mathbf{x}^T\Big), \\ \text{subject to} \quad & \psi_{\mathrm{min},i} \leq \hat{\phi}_i\Big(\hat{\boldsymbol{W}}_{\phi,i}^T \mathbf{x}^T\Big) \leq \psi_{\mathrm{max},i}, \quad i \in \{1, ..., N\}, \end{aligned}$$

where $\hat{\boldsymbol{W}}_\eta^T$ and $\hat{\boldsymbol{W}}_{\phi,i}^T$ are matrices containing the approximated active eigenvectors for the objective function and the $i^{\mathrm{th}}$ constraint, respectively. As in the genetic algorithm case, the built-in optimization toolbox provided by the host software is used [9].

## 3 Application and results

The test case consists on the optimization of a single-stage mini (*m*)ORC expander. Figure 1 shows the specifications of the selected machine, namely, a radial inflow turbine with its corresponding constant parameters, design space, and constraints. This test case has been adapted from the work of Lang et.al. [10].

Notice that Figure 1 shows a non-normalized design space with $m = 7$. Moreover, as discussed in Section 2.2.1, this input space has been selected such that a continuous, smooth solution, is mostly guaranteed. The computations were run on a quad-core 3.60GHz processors machine. No parallelization is used in this study.




| **Constant parameters** | | | | | | |
|---|---|---|---|---|---|---|
| $P_3$ | 1.0 | bar | $r_1/r_2$ | 1.02 | - | |
| $b_0$ | 2.4 | mm | Fluid | $D_4$ | - | |
| $r_0/r_1$ | 1.3 | - | | | | |
| **Design space (model inputs, $m = 7$)** | | | | | | |
| | min | max | | min | max | |
| $r_2/r_3$ | 1.9 | 2.3 | - | $\beta_3$ | 50.1 | 61.3 | ° |
| $D_0$ | 181.4 | 221.7 | mm | $\Omega$ | 23.0 | 28.1 | krpm |
| $P_0$ | 3.5 | 4.3 | bar | $R$ | 0.3 | 0.4 | - |
| $\dot{m}$ | 0.24 | 0.30 | kg/s | | | | |
| **Constraints** ($N = 4$) | | | | | | |
| | min | max | | min | max | |
| $M_{2,\mathrm{rl}}$ | - | 0.8 | - | $\gamma_2$ | - | 75.0 | ° |
| $r_{3,\mathrm{s}}/r_2$ | - | 0.7 | - | $r_{3,\mathrm{h}}/r_{3,\mathrm{s}}$ | 0.4 | - | - |
| **Optimization objective** | | | | | | |
| $\eta_{\mathrm{ts}}$ | | | | | | |

Figure 1: Constant parameters, design variables and constraints used for the optimization of the radial inflow turbine. This test case has been adapted from to the work of Lang et.al. [10].

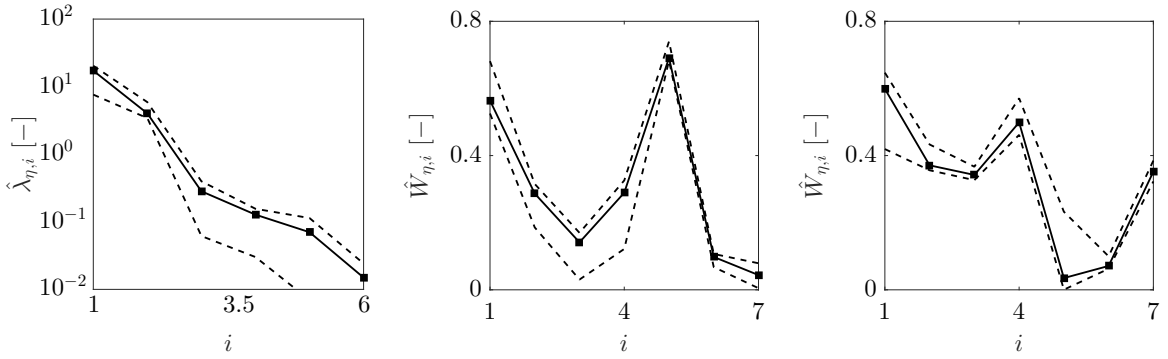
### 3.1 Construction of the surrogate model

In order to test the feasibility of the method, a primary attempt to reveal active subspaces is done with the current test case and an oversampling factor of $\alpha = 12$, so that

$$M = \alpha m. \tag{17}$$

Several sampling methods are available, and they are compared later in Section 3.2. For the moment, a sparse grid combined with a latin hypercube strategy suffices. Note also that the *zTurbo* gradients are obtained by forward finite differences.

Figure 2 presents the results of the eigenvalue decomposition of matrix $\hat{\boldsymbol{C}}$ computed for the turbine total-to-static efficiency. Figure 2a displays the eigenvalues in a logarithmic scale, in order to ease the detection of active subspaces. The dashed lines correspond to a boostrap confidence interval of $95\%$ (see Ref. [5] for information on bootstrapping). Notice that, independently of the boostrap interval, the largest gap is located between eigenvalues $2$ and $3$, thus suggesting the presence of two active variables. The smallest eigenvalue is not reported in Figure 2, for its value is insignificant (compared to the finite difference $\Delta x$) to draw any valid conclusion from it [5].



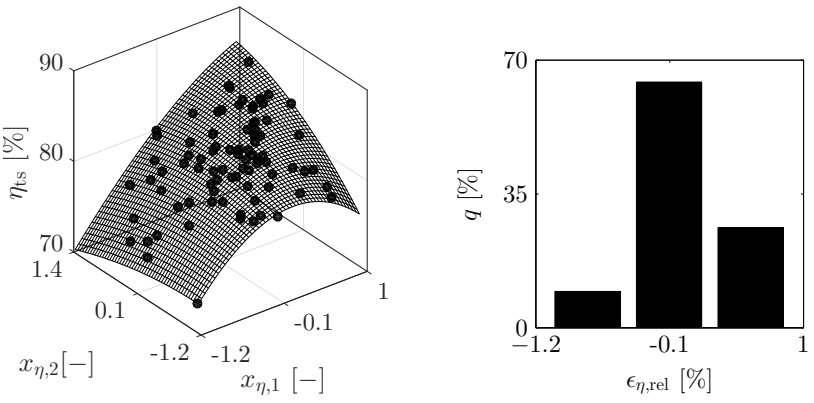

(a) (b) (c)

Figure 2: a) Eigenvalues corresponding to the decomposition of $\hat{\boldsymbol{C}}$. b) Eigenvector corresponding to the first eigenvalue. c) Eigenvector corresponding to the second eigenvalue. The dashed lines correspond to a bootstrap confidence interval of $95\%$.

Figures 2b and 2c present the coordinates of the eigenvectors corresponding to the $2$ largest eigenvalues from Figure 2a. Differently from the eigenvalue computation, the boostrap intervals are large for various coordinates. Nonetheless, the mean value is always close to a confidence bound, thus suggesting that only few samples feature considerable deviation from the mean.

Following the analysis done on $\hat{\boldsymbol{C}}$, it is possible to build a surface response for the turbine efficiency as a function of two active variables. No information exists regarding the potential structure of the surface response, yet polynomial functions lead to satisfactory results in this work. It is well-known that, due to Runge's phenomenon, high-order polynomials (third or higher) might lead to incorrect predictions at the edges of the input space. It is then important to inspect the surrogates, and even visualize them if possible.

Figure 3a presents the results of this process, namely, the surrogate-based surface and the samples for the turbine efficiency. Although being a third order polynomial, the figure presents a smooth, well-behaved, response. Additionally, Figure 3b presents the percentage distribution




(a) (b)

Figure 3: Results of the surrogate construction with an oversampling factor of $12$ for the turbine total-to-static efficiency. a) Surface response based on the surrogate, and *zTurbo* samples (●). b) Percentage distribution of the relative error between the polynomial fit and the samples.

of the related relative error between the polynomial fit and the samples. Note that the error

prominently surrounds $-0.1\%$, thus confirming a satisfactory approximation for the scope of this work.

A similar analysis for the other quantities of interest is performed, also leading to satisfactory results. It is possible now to perform a gradient-based optimization, by means of the procedure exposed in Section 2.2.2. Moreover, it is of interest to evaluate the optimization results as a function of several sampling techniques and number of samples. This is the matter of the following section.

### 3.2 Sensitivity analysis of the optimization

In order to analyze the robustness of the optimization problem described in Section 2.2.2, several sampling strategies are tested: i) random sampling with replacement, ii) random sampling with replacement in a sparse grid, iii) latin hypercube in a sparse grid. Various oversampling factors are also used. The sensitivity analysis follows the next steps:

**a.** for a specific sampling strategy, and oversampling factor, an optimization using the reduce-order model is performed following the procedure discussed in Section 2.2.2.

**b.** The solution drawn by the optimizer is recalculated with the original model (*zTurbo*).

**c.** The relative error of the objective and the constraints are calculated.

Figure 4a and Figure 4b present the results of these computations: the relative error of the optimization goal $(\epsilon_{\eta,\mathrm{rel}})$ and the constraints (averaged of absolute values, $|\bar{\epsilon}|_{\phi,\mathrm{rel}}$), as a function of the sampling technique and the oversampling factor $(\alpha)$. Although $\epsilon_{\eta,\mathrm{rel}}$ is small for the optimizations done with $\alpha = 4$, $|\bar{\epsilon}|_{\phi,\mathrm{rel}}$ is comparatively large, thus invalidating any conclusions drawn from these cases. On the other hand, independently of the sampling technique, $\epsilon_{\eta,\mathrm{rel}}$ somehow stabilizes for $\alpha \geq 8$, and it seems to reach convergence as the sampling population increases. However, notice that the estimation of the constraints is poor when using random and sparse grid-random sampling strategies. Conversely, the sue of a sparse grid-latin hypercube guarantees accurate estimation of the constraints, hence making this sampling strategy the most reliable.

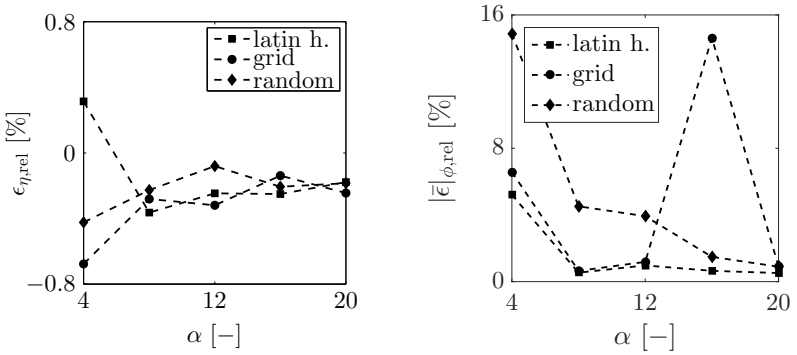



Figure 4: Sensitivity analysis of the optimization. a) Objective relative error as a function of the sampling strategy and the oversampling factor. b) Constraints average relative error as a function of the sampling strategy and the oversampling factor.

As previously discussed, the analysis on the eigendecomposition of $\hat{\boldsymbol{C}}$ indicated that the original model might be reduced to a surrogate function of two independent variables. This

means that, in average, perturbations in the selected two active variables change $\eta_{\text{ts}}$ more than perturbations in the five residuary (inactive) variables [5]. It follows that the surrogate model *inherently* features an error, for not all the directions where the function is varying are taken in account. Such characteristic manifests in Figures 4a and 4b, which illustrate that, independently of the number of samples, the surrogate will always present an offset respect to the original model value.

Worthy highlights can be extracted from Section 3.1 and the last paragraphs: i) this design problem can be transformed into a surrogate with two active variables, ii) sampling by means of a sparse grid-latin hypercube strategy yields the best results. These conclusions are used now to perform a comparison between a genetic algorithm optimization and a gradient based optimization using the active subspace models.

### 3.3 Comparison between optimization strategies

This section contains the results of the comparison between two optimization strategies: i) the original procedure conceived to perform the preliminary fluid dynamic design of turbomachinery, i.e., *zTurbo* combined with a genetic algorithm (see Sec. 2.1.1); ii) the new approach, which relies the surrogate model combined with a gradient-based search method (see Sec. 2.2.2). In order to make a fair comparison, the population number and sample size of both methods are expressed as a function of the oversampling factor, so that

$$P = M = \alpha m, \tag{18}$$

where $P$ is the genetic algorithm population size, $m$ is the number of design inputs, and $M$ is the number of samples used to build the active subspace model. Figure 5a presents the optimum as a function of the oversampling factor, and obtained by the two optimization methods.

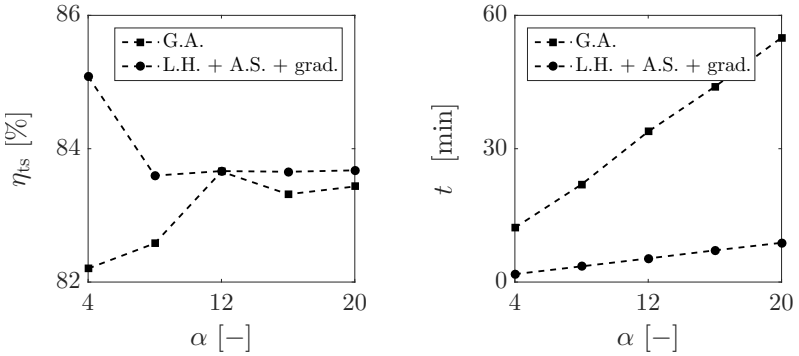



Figure 5: Comparison between optimization strategies: i) genetic algorithm (G.A.), and ii) latin hypercube combined with an active subspace transformation and a gradient based optimization (L.H. + A.S. + grad.) a) Objective $(\eta_{\text{ts}})$ as a function of the oversampling factor. b) Optimization time (including the construction of the surrogate) as a function of the oversampling factor.

In genetic algorithms, larger $P$ increases the likelihood of finding a global optimum at a cost of increasing the computational time. This response features an asymptotic response, hence, at a certain point, increasing $P$ will not improve the solution considerably, yet it will consume more resources. This is the response observed in Figure 5a, where it seems that $\alpha = 12$ is a compromise between solution optimality and population size. On the other hand, the optimum

obtained by the gradient base method combined with the surrogate converges after $\alpha = 8$, since, as presented in Figure 4a, $\epsilon_{\eta,\mathrm{rel}}$ stabilizes for $\alpha \geq 8$. Note that eventually both solution methods coincide, for the difference between their optima lays within the expected error of approximately $\pm 1\%$ (see Fig. 3b). Finally, although the surrogate corresponding to $\alpha = 4$ apparently yields a better optimum, recall that this solution presents a comparatively large deviation in the constraints (see Fig. 4b), hence disqualifying these results.

Figure 5b presents the computational time required by both optimization procedures as a function of the oversampling factor. The chart includes the time required to build the surrogate, in order to make a proper comparison. Note that the optimization by means of active subspaces presents a lower computational burden, almost independently from the population size (at least four times faster). That is, even comparing a genetic algorithm with $\alpha = 4$ with a surrogate optimization with $\alpha = 20$, it is observed that the latter is computationally lighter.

### 3.4 Physical insight given by the surrogate

Relevant knowledge is gained by comparing the optimum design solutions obtained by both methods. The genetic algorithm presents the best results when using $\alpha = 12$ (see Fig. 5a), hence this solution is selected. Although the optimum and the constraints given by the surrogate converge for $\alpha \geq 8$ (see Fig. 4), a closer look reveals that the error slightly decreases as $\alpha$ increases. For this reason, the comparison is done with the surrogate built with $\alpha = 20$. Figure 6a presents the error in the constraints corresponding to this solution. The deviation does not exceed $\pm 1\%$, hence demonstrating again the robustness of the optimization.

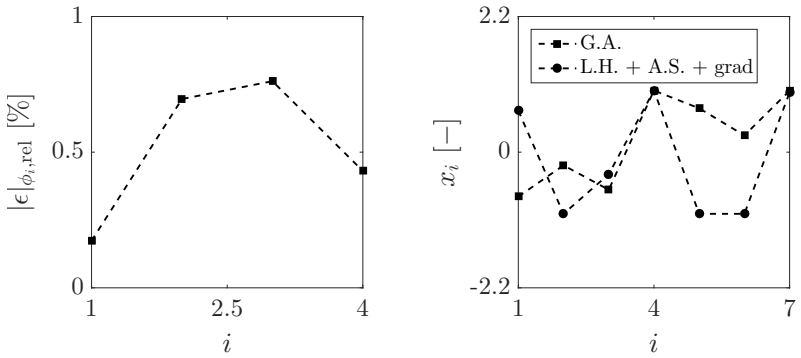



Figure 6: a) Constraint relative error for the optimum drawn by the surrogate model $(\alpha = 20)$. b) Optimal design inputs drawn by the genetic algorithm (G.A), and the surrogate model (L.H. + A.S. + grad.).

The optimum design inputs are presented in Figure 6b. Although the turbine efficiency practically coincides (see Fig. 5a), the values of the design variables are mostly different. This suggests that, for turbomachinery design problems, there might be *more* than one ($m$-dimension) region containing constrained optima. Nonetheless, by virtue of the active subspace transformation, it is possible to gather these optima in a single region in the reduced-order response surface.

To prove this point, the seven dimension input of both optima are transformed according to the active subspace computed with $\alpha = 20$. The results of this procedure are presented in Figure 7. To derive physical conclusions, it is necessary to recall some theory. The eigenvalue

decomposition of $\hat{C}$ reveals the directions where the function variability is the largest. This decomposition is sorted in order of influence, meaning that the first active variable (corresponding to the largest eigenvalue) affects more $\eta_{\text{ts}}$. Due to this fact, it is likely that both optimizers converge to constrained solutions in the region surrounding the first gradient dominant direction, i.e., the first optimum active variable should concur. This is confirmed in Figure 7a, which shows the optimum design inputs in the active subspace.

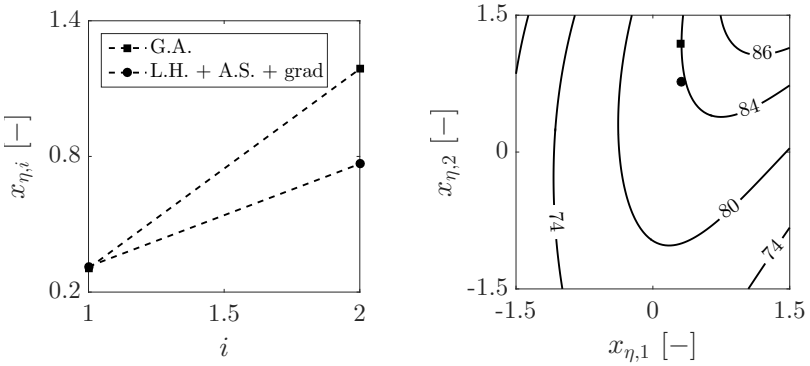



Figure 7: Coordinates of the genetic algorithm optimum (G.A) and the surrogate model optimum (L.H. + A.S. + grad.) in the active subspace. b) Contour plot of the surrogate response surface ($\alpha = 20$), and solutions drawn by the genetic algorithm (■) and the surrogate model (●).

Figure 7b depicts the contour plot of the $\eta_{\text{ts}}$ reduced-order surface response ($\alpha = 20$). It also presents the coordinates of the solutions drawn by both optimizers. Notice that there are a small regions where the effect of $x_{\eta,2}$ is negligible. As a matter of fact, the optima reside in one of them, thus featuring different $x_{\eta,2}$, and proving that this region contains multidimensional solutions that respect the constraints.

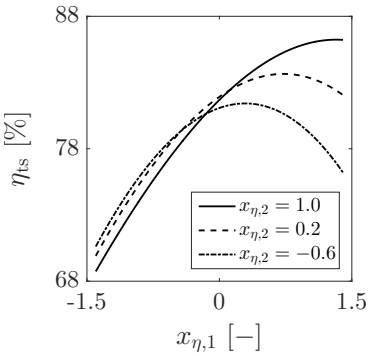



Figure 8: Reduced-order response surface (see Fig. 7b) for the turbine efficiency observed from an orthogonal point of view.

Figure 7b exposes the existence of a parabolic shape for $\eta_{\text{ts}}$, which converges to a peak as both active variables approach $1.5$. In an attempt to understand this response, Figure 8 presents the $\eta_{\text{ts}}$ surface from an orthogonal point of view. It is clear how the efficiency lines present a convex shape with an (non-constrained) optimum that is mainly a function of $x_{\eta,1}$. The second active variable $x_{\eta,2}$ scales the effects of $x_{\eta,1}$.

In order to determine the origin of the efficiency convex shape, the contributions of the loss mechanisms from various samples following the $\eta_{\text{ts}}$ line ($x_{\eta,2} = 0.2$) are presented in Figure 9. Observe in Figure 9a that the kinetic energy, the rotor profile, and secondary losses shape the efficiency curve, while the other contributions remain relatively constant. In particular, the convex shape observed in Figure 8 is induced by the kinetic energy loss.

The reasons for the losses variation can be deducted from Figure 9b, which presents the rotor profiles and the velocity triangles for the selected samples. Observe that, as $x_{\eta,1}$ increases, the flow deviation decreases, hence abating the rotor profile and secondary losses. Moreover, as $x_{\eta,1}$ increases, the component of the velocity normal to the outlet area decreases, thus requiring larger blade heights. Likewise, notice that larger $x_{\eta,1}$ decreases the absolute velocity at the outlet of the machine, hence increasing the efficiency until a maximum is reached ($x_{\eta,1} \approx 0.7$), after which the outlet kinetic energy again increases.

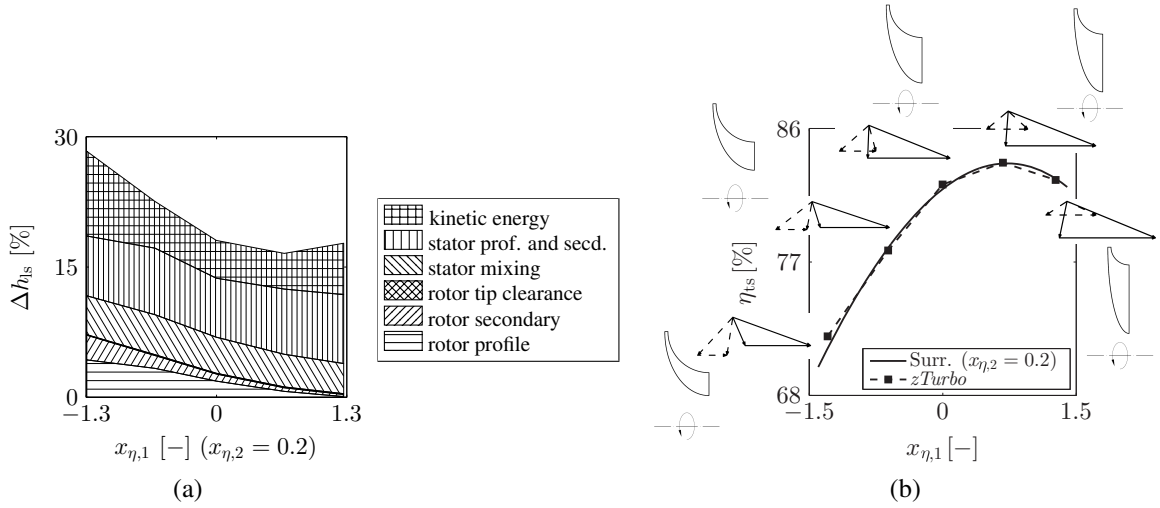



Figure 9: a) Loss contributions relative to the eulerian work for selected samples along $x_{\eta,2} = 0.2$. b) Velocity triangles and rotor meridional channel for the selected samples. The solid lines (-) in the velocity triangles correspond to the rotor inlet, the dashed lines to the rotor outlet (- -).

## 4 Concluding remarks

This work investigates the use of active subspaces to perform the fluid dynamic preliminary design of turbomachinery. The case study consists on the design of an ORC single-stage radial inflow turbine. Active subspaces are used to construct a reduced-order model of the machine total-to-static efficiency and the meridional channel geometry. The surrogate is then tested in a constrained optimization, and the results are compared against the original optimization procedure (genetic algorithms combined with the original model). Ultimately, the response surface is analyzed to gain physical knowledge on the design problem. The main results are summarized as follows:

**i)** active subspaces exist in a (limited) design space. It is therefore possible to construct reduced-order models for the turbine optimization objective, and the constraints. The surrogates present an accuracy considered satisfactory for the objectives of this work ($\pm 1\%$for the total-to-static efficency).

**ii)** By means of the surrogate models, it is possible to switch to gradient-based search methods. The optimization results show that, compared with the original method, using the surrogates is computationally less demanding (at least four times faster) and comparatively accurate.

**iii)** The design optimization of ORC turbomachinery likely feature multiple regions containing constrained optima. The active subspace transformation gathers this optima in a single region in the reduced-order response surface.

Although the work has been done in a limited design space, the active subspace methods have proved to be a valuable tool for turbomachinery design. Future work will deal with other expander configurations and a larger design space.

| **Nomenclature** | | **Subscripts** | |
|---|---|---|---|
| $\alpha$ | oversampling factor | 0, 1, 2, 3 | stage position |
| $\beta_3$ | rotor outlet blade angle | ac | active |
| $\hat{\boldsymbol{C}}$ | uncentered covariance matrix | ic | inactive |
| $D$ | mean diameter | min | minimum |
| $\eta_{\mathrm{ts}}$ | turbine total-to-static efficiency | max | maximum |
| $\gamma_2$ | stator outlet blade angle | h | hub |
| $\hat{\boldsymbol{\Lambda}}$ | eigenvalue matrix | s | shroud |
| $\hat{\lambda}$ | eigenvalue | $\eta$ | related to efficiency |
| $M$ | number of samples | $\phi$ | related to *zTurbo* output |
| $M_{2,\mathrm{rl}}$ | relative Mach number at rotor inlet | ls | loss |
| $m$ | dimension of design inputs vector | **Abbreviations** | |
| $\dot{m}$ | mass flow rate | A.S. | active subspace methods |
| $N$ | number of constraints | G.A. | genetic algorithm |
| $n$ | dim. of reduced-order inputs vector | L.H. | latin hypercube |
| $\Omega$ | rotational speed | grad. | gradient |
| $P$ | turbine inlet pressure | | |
| $\psi$ | constraint | | |
| $\phi$ | *zturbo* output | | |
| $R$ | degree of reaction | | |
| $r$ | mean radius | | |
| $\hat{\boldsymbol{W}}$ | eigenvector | | |
| $\mathbf{x}$ | design input vector | | |
| $x$ | design input value | | |

## REFERENCES


[1] N. Baines, *A meanline prediction method for radial turbine efficiency*. In 6th International Conference on Turbocharging and Air Management Systems, 1982.

[2] A. La Seta, J. Andreasen, L. Pierobon, G. Persico, and F. Haglind, *Design of organic Rankine cycle power systems accounting for expander performance*. Proceedings of the 3rd International Seminar on ORC Power Systems, 2015.

[3] S. Bahamonde, C. Di Servi, and P. Colonna, *Optimized preliminary design method for mini-ORC power plants: an integral approach*. Proceedings of the International Conference on Power Engineering (ICOPE-15), 2015.

[4] P. Colonna, E. Casati, C. Trapp, T. Mathijssen, J. Larjola, T. Turunen-Saaresti, A. Uusitalo, *Organic rankine cycle power systems: from the concept to current technology, applications and an outlook to the future*. Journal of Engineering for Gas Turbines and Power, 137:100801-119, 2015.

[5] P. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies*. arXiv, 2015.

[6] MATLAB Version 8.2.0.701 (R2013b), The MathWorks Inc., Natick, Massachussets, 2013a.

[7] ASIMPTOTE, *FluidProp: a software package for computation of thermodynamic properties*, `http://www.asimptote.nl/software/fluidprop`, 2015.

[8] M. Pini, G. Persico, E. Casati, V. Dossena, *Preliminary Design of a Centrifugal Turbine for Organic Rankine Cycle Applications*. Journal of Engineering for Gas Turbines and Power, 135:04231219, 2013.

[9] MATLAB Optimization Toolbox 6.4., The MathWorks Inc., Natick, Massachussets, 2013a.

[10] W. Lang, P. Colonna, and R. Almbauer, *Assessment of Waste Heat Recovery From a Heavy-Duty Truck Engine by Means of an ORC Turbogenerator*. Journal of Engineering for Gas Turbines and Power, 135:042313110, 2013.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# EXTENSION AND APPLICATION OF A NONLINEAR REDUCED ORDER MODEL TO GUST LOAD PREDICTION IN TIME DOMAIN

**C. Strobach**[1], **K. Lindhorst**[1], **M. C. Haupt**[1], **P. Horst**[1]

[1] TU Braunschweig, Institute of Aircraft Design and Leightweight Structures
Hermann-Blenk-Str. 35
38108 Braunschweig
e-mail: {c.strobach, k.lindhorst, m.haupt, p.horst}@tu-bs.de

**Keywords:** reduced order models, gust load prediction, disturbance velocity approach

**Abstract.** *This paper aims at extending a markov chain based reduced order model to discrete gust load prediction in an aeroelastic simulation. An method for the incorporation of the disturbance velocity approach is presented and evaluated for the AGARD445 wing based on different training strategies. The reduced order model trained under elastic and gust load conditions can successfully predict the gust response in a rigid and in an elastic setup. Thus the presented ROM approach can serve as one single CFD surrogate model to predict aerodynamic forces under multiple loading conditions.*

## 1 Introduction

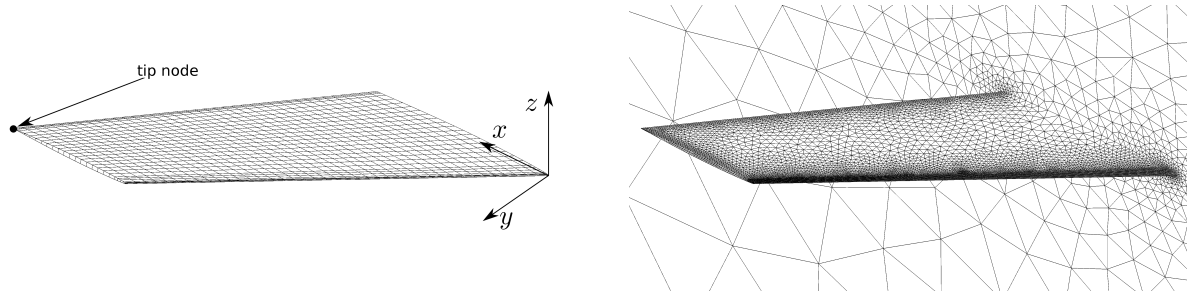
Efficient gust response analysis is crutial in aircraft design. Compared to linear aerodynamic models, URANS-solver can capture nonlinear effects in the transonic flow regime such as separation and shock movement at the cost of high computing times. Reduced order models (ROM) can reduce this computational cost significantly which is very important for coupled fluid-structure interaction simulations of large aircraft structures [16, 9].

ROMs for gust load caluations have been established using a linear aerodynamic model [4] or by a correction of doublet-lattice results [2]. [17] applied a disturbance velocities approach(DVA) in a linear aerodynamic state space formulation for the construction of a ROM that is capable of predicting generalized aerodynamic coefficients.

Recently [8] presented a reduced order model, that combines proper orthogonal decomposition (POD) and a discrete Markov chain with nonlinear mapping functions for the prediction of aerodynamic loads in the time domain. This paper presents an approach for the direct incorporation of the DVA [6] into this markov chain based ROM and an application to the 3D AGARD445.6 benchmark configuration under a discrete gust.

Using a predefined displacement sweep function to superimpose the first structural eigenmodes the ROM is trained with the computed nonlinear aerodynamic response of a forced motion analysis in TAU. Limit cycle oscillations (LCO) can be sufficiently predicted even for complex 3D configurations in the transsonic flow regime [9]. Extending the ansatz for the input displacements of the ROM with the discrete disturbance velocities, gust loads can be predicted.

**AGARD model** The well known benchmark case of the AGARD445 wing, experimentally studied by Yates [15], is choosen as a 3D example of the presented methods. The aerodynamic and structural model are depicted in Fig. 1. The corresponding structural eigenmodes are shown in Fig. 2 and listed in Table 1.




(a) (b)

Figure 1: AGARD445.6 model surface grids: (a) structural (b) aerodynamic

## 2 Numerical methods

The following paragraphs summarize the overall methodology including the solution of the fluid-structure problem, the training and creation process of the reduced order models, for further information consult [7].

### 2.1 Fluid-structure interaction

For the solution of the coupled *fluid-structure interaction*(FSI) problem a partitioned implicit coupling scheme is applied to solve the Dirichlet-Neumann iteration. The data transfer and in-

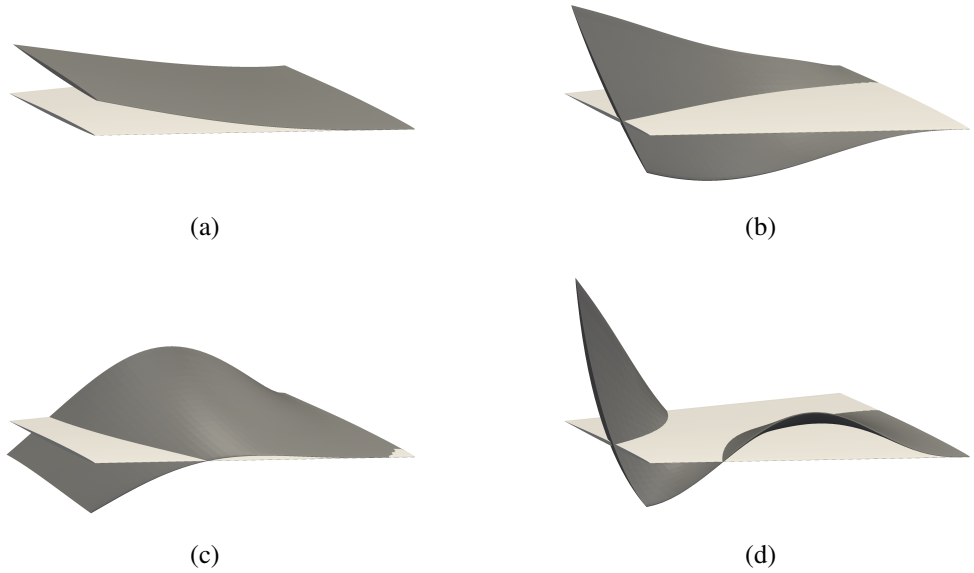



Figure 2: First four eigenmodes of the AGARD445.6-wing [7]: (a) first bending mode $\underline{\Phi}_1$; (b) first torsion mode $\underline{\Phi}_2$; (c) second bending mode $\underline{\Phi}_3$; (d) second torsion mode $\underline{\Phi}_4$

| Mode | $f_s/Hz$ | $f_{s,Unger}/Hz$ | $f_{s,Yates}/Hz$ |
|---|---|---|---|
| 1 | $9,813$ | $9,813$ | $9,60$ |
| 2 | $38,795$ | $38,800$ | $38,10$ |
| 3 | $50,514$ | - | $50,70$ |
| 4 | $94,119$ | - | $98,50$ |

Table 1: Comparison of the eigenfrequencies of the employed structural mesh using the reference values from $f_{s,Unger}$ Unger [13] and the experimental detemined values of Yates [15]

terpolation of the displacements and forces between the nonconforming grids is established with *ifls*, see [5]. Using the *computational fluid mechanics*(CFD) solver TAU [11] for the solution of the Euler equation (Dirichlet boundary conditions) the unknown aerodynamic forces of the fluid surface mesh can be computed. The structural problem with Neumann boundary conditions is solved for the unknown structural displacements using a second order implicit newmark algorithm. The mass **M** and stiffness matrix **K** are exported from an ANSYS model.

**CFD Gust calculation** In contrast to a stochastic nature of atmospheric turbulence, a gust is considered in an idealized form as a discrete gust profile of disturbance velocities, see [14]. The CFD solver TAU provides a *disturbance velocity approach* (DVA) for the generation of discrete gust profiles. The DVA assumes that the disturbance velocities can be superimposed to the flow field by altering the flux balance, as described in [6], neglecting the mutual influence between the aircraft and the gust. One common gust profile for design purposes is the *1-cos* gust (see FAR25.341 [3]) where the disturbance velocities can be expressed as a function of time and

space

$$w_g(x_g) = \begin{cases} \frac{w_{g0}}{2}\left(1 - \cos\frac{2\pi x_g}{\lambda_g}\right) & 0 \leq x_g \leq \lambda_g \\ 0 & x_g < 0 \;||\; x_g > \lambda_g \end{cases} \tag{1}$$

with the gust wave length $\lambda_g$, the relative gust moving velocity $U_g$ and the maximum disturbance velocity $w_{g0}$.

### 2.2 Training process

The training process consists of two steps and is summarized in the follwing paragraphs, for further information consult [7, 8, 9]:

1. Generation of training data by collecting forced displacements $u_f(t)$ and aerodynamic forces $F_f(t)$ using TAU

2. Model identification, construction of the ROM

In the first step the CFD data are calculated using an forced motion unsteady analysis in order to collect fluid displacements $\underline{u}_f(t)$ and corresponding force coefficients $\underline{c}_F(t) = \frac{\underline{F}_f(t)}{\frac{1}{2}\rho_\infty U_\infty^2 A_{ref}}$ values for each node of the aerodynamic surface mesh in every timestep $t$. The forced motion is defined by a superposition of the utilized modes in Eq. (2) and a sweep function in Eq. (3). The sweep function incorporates several features that can be tuned by parameters: The maximum amplitude of each mode is controlled by the maximum value of the generalised coordinate $q_{s,max}$, the exponential amplitude increase over time is controlled by $k_{Amp}$ and the frequency modulation can be tuned with $k_\omega$.

$$\underline{u}^f(t) = \sum_{s=1}^{3} q_s(t)\underline{\Phi}_s \tag{2}$$

$$q_s(t) = q_{s,max} \underbrace{\left(1 - e^{-k_{Amp}t}\right)}_{\text{exp. amplitude increase}} \sin(\omega_s t(\underbrace{1 - 0,1\sin\left(\frac{\omega_s}{k_\omega}t\right)}_{\text{frequency modulation}})) \tag{3}$$

In the second step the ROM is build based on the training data. The elements of the ROM are depicted in Fig. 3:

Using the truncated *singular value decomposition*(SVD) as a *proper orthogonal decomposition*(POD) technique both the $\underline{u}_f(t)$ and $\underline{c}_F(t)$ can be mapped into an subspace and can be described with their corresponding POD coefficients $\underline{\hat{u}}_f(t)$, $\underline{\hat{c}}_f(t)$ (***input/ouput POD***, cp. Fig. 3 ). The mode truncation is established using a user-defined threshold value $v_{red}$ and a partial sum criterion regarding the singular values $\sigma_i$ of each mode, see e.g [12].

$$v = \frac{\sum_i^M \sigma_i}{\sum_j^q \sigma_j} \geq v_{red} \;\text{ mit }\; 0.0 \leq v_{red} \leq 1.0 \tag{4}$$

Combining a NARMA(*Nonlinear Autoregressive Moving Average*) ***Markov chain*** model [10] and a single layer RBF neuronal network[1] as the ***nonlinear mapping*** function $G(\underline{x})$, a surrogate function $\underline{\hat{c}}_F(t)$ can be defined Eq. (5). The surrogate function only depends on the

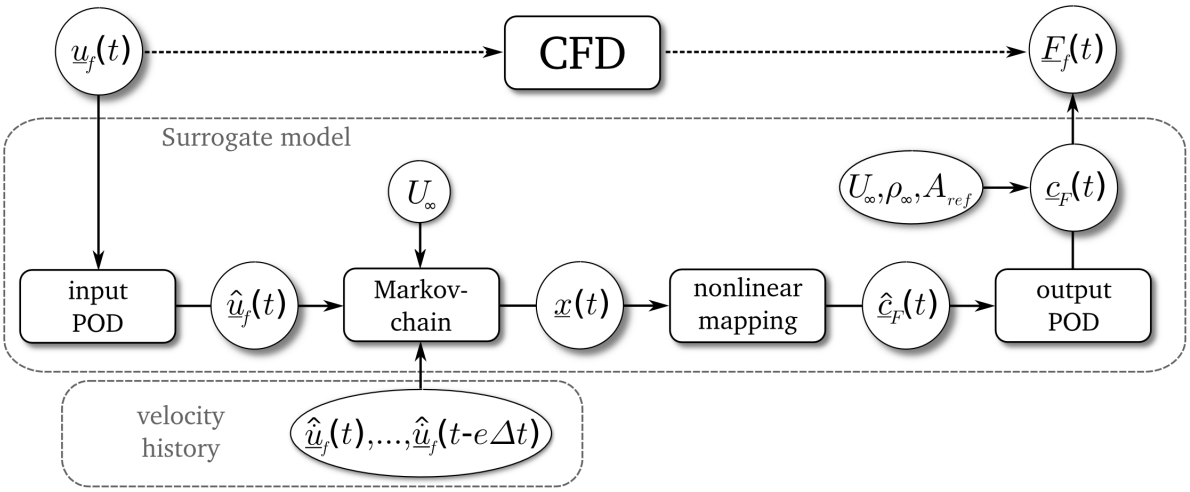



Figure 3: Modular process chain of the reduced order model

current static displacement input $\underline{\hat{u}}_f(t)$ and some discrete values for time history of unsteady velocities $\underline{\hat{\dot{u}}}_f$.

$$\hat{\underline{c}}_F(t) = G\left(\underbrace{\underline{\hat{u}}_f(t)}_{\text{steady part}}, \underbrace{\frac{\hat{\underline{\dot{u}}}_f(t)}{U_\infty}, \frac{\hat{\underline{\dot{u}}}_f(t-f_{sp}\Delta t)}{U_\infty}, \frac{\hat{\underline{\dot{u}}}_f(t-2f_{sp}\Delta t)}{U_\infty}, \ldots, \frac{\hat{\underline{\dot{u}}}_f(t-l_{\dot{u}}\Delta t)}{U_\infty}}_{\text{unsteady part}}\right) \tag{5}$$

where $f_{sp} \in \mathbb{N}$ is defined as a sparse factor that reduces the number of input parameters by considering only each $f_{sp}$th time step in the history of velocities.
For all simulations a time window number of $l_{\dot{u}} = 50$, an unsteady physical time step size of $\Delta t = 0.001$ and a sparse factor of $f_{sp} = 5$ are employed.

**Handling of gusts in ROM approach** The main idea for the incorporation gust disturbance velocities - as published in [7] - is to superimpose the disturbance velocities $\underline{w}_g(t)$ with the structural displacements velocities $\underline{\dot{u}}_f(t)$ to form a modified velocity distribution $\underline{\dot{u}}_{f+g}(t) = \underline{\dot{u}}_f(t) + \underline{w}_g(t)$. Consequently the surrogate function Eq. (5) is modified to

$$\hat{\underline{c}}_F(t) = G\left(\underline{\hat{u}}_f(t), \frac{\hat{\underline{\dot{u}}}_{f+g}(t)}{U_\infty}, \frac{\hat{\underline{\dot{u}}}_{f+g}(t-f_{sp}\Delta t)}{U_\infty}, \ldots, \frac{\hat{\underline{\dot{u}}}_{f+g}(t-l_{\dot{u}}\Delta t)}{U_\infty}\right) \tag{6}$$

### 2.2.1 Elastic modes trained ROM E

In the training data generation step only the first three elastic eigenmodes (Table 1) of the AGARD wing are choosen in the forced motion process using Eq. (2). Two analyses with the maximum values of the generalized coordinates as listed in Table 2 are performed and the resulting matrices $\underline{u}_{f,E}(t)$ and $\underline{c}_{F,E}(t)$ serve as input for the training process.

Under the flow conditions listed in Table 3, 1000 time steps are calculated using empirical parameters of $k_{Amp} = 5$ and $k_\omega = 25$, compare Eq. (3).

| Analyse | $q_{1,max}$ | $q_{2,max}$ | $q_{3,max}$ |
|---|---|---|---|
| I | 0.02 | 0.005 | 0.004 |
| II | 0.04 | 0.01 | 0.008 |

Table 2: Maximum values of the generalized coordinates of the forced motion analysis

| parameter | value | parameter | value |
|---|---|---|---|
| $Ma$ | $0,901$ | $\rho_\infty$ | $1,29251\ \frac{kg}{m^3}$ |
| $p_\infty$ | $101325,0\ Pa$ | $U_\infty$ | $298,49\ \frac{m}{s}$ |
| $T_\infty$ | $273,15\ K$ | $\alpha$ | $0°$ |

Table 3: Flow conditions

Using the truncation criterion of Eq. (4) with $v_{red} = 0.99$ results in 3 input POD modes (accoring to the three exited eigenmodes) and 254 output POD modes.

### 2.2.2 Gust trained ROM G

In order to generate a gust training set a gust analysis over the rigid wing is performed using the already mentioned flow conditions (Table 3) and the following gust parameters:

The relative gust moving velocity $u_g$ is set to $U_\infty$ and the gust wave length $\lambda_g$ is calculated as follows in order to exite the first structural eigenmode $f_{s,1}$.

$$\lambda_g = \frac{U_{\text{inf}}}{f_{s,1}} = 30.4179m \tag{7}$$

In every timestep the disturbance velocities $\underline{w}_g(t)$ as well as the force coefficients $\underline{c}_{F,G}(t)$ are collected from three analyses with varying maximum disturbance velocities, see Table 4.

| Gust analysis | $w_{g0}$ |
|---|---|
| I | $4\ \frac{m}{s}$ |
| II | $8\ \frac{m}{s}$ |
| III | $20\ \frac{m}{s}$ |

Table 4: maximum gust disturbance velocities for all three gust analyses for the rigid wing

For the ROM identification process virtual deformations $\underline{u}_{f,G}(t) = \underline{w}_g(t) \cdot \Delta t$ are calculated and serve as the only input for the ROM training process. The reduced order model G is thus trained with $\underline{u}_f(t) = \underline{u}_{f,G}(t)$ and $\underline{c}_F(t) = \underline{c}_{F,G}(t)$ of all three gust analyses. Consequently using Eq. (4) with a reduction threshold value of $v_{red} = 0.999$ overall 2 input modes and 21 output modes are identified.

### 2.2.3 Gust trained ROM GE

For the reduced order model GE that should predict the elastic as well as the gust CFD response no new data has to be generated. It is trained with forced motion data of 2.2.1 and 2.2.2 and thus with $u_f(t) = \underline{u}_{f,E}(t) \cup \underline{u}_{f,G}(t)$ and $c_f(t) = \underline{c}_{F,E}(t) \cup \underline{c}_{F,G}(t)$. Consequently using Eq. (4) with a reduction threshold value of $v_{red} = 0.999$ overall 4 input POD modes and 246

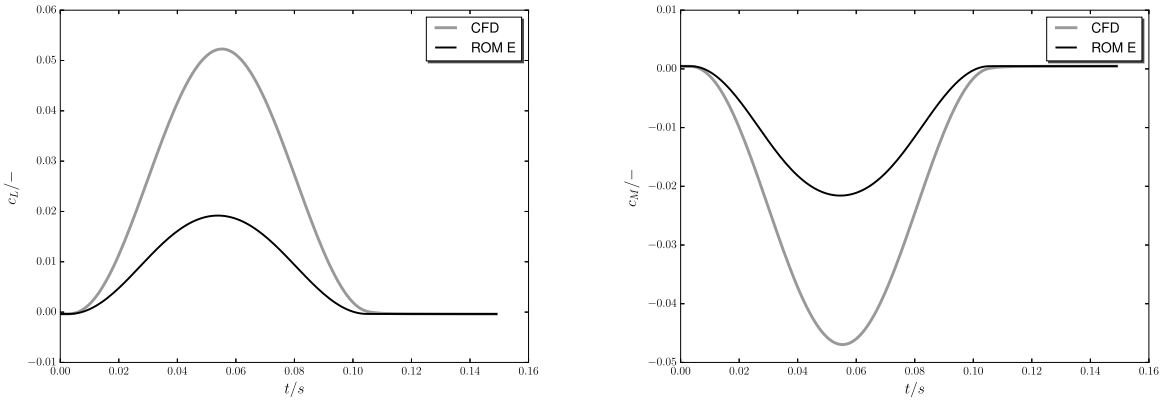
output POD modes are identified which means that one additional input POD mode is identified compared to ROM E.

## 3 Results

In the following section the results for the CFD and ROM calculation are compared. In the first sections the results for a gust over the rigid AGARD wing are compared for the flow conditions depicted in Table 3. Finally the gust trained ROMs are employed in an aerostructural coupling environment.

### 3.1 Rigid - ROM E

Fig. 4 shows the comparative results for the rigid setup, i.e. no deformations of the strucutural model are considered. The reduced order model E can predict the the time response accurately but the amplitude of the gust is underestimated, due to the improper training set. The exited strutural modes in the training set do not include displacements at the wing root and thus disturbance velocities in the root of the wing cannot be taken into account, as alreay argued in [7]. Thus an improved training set should at least include modes with displacements/velocities at the wing root.




(a) (b)

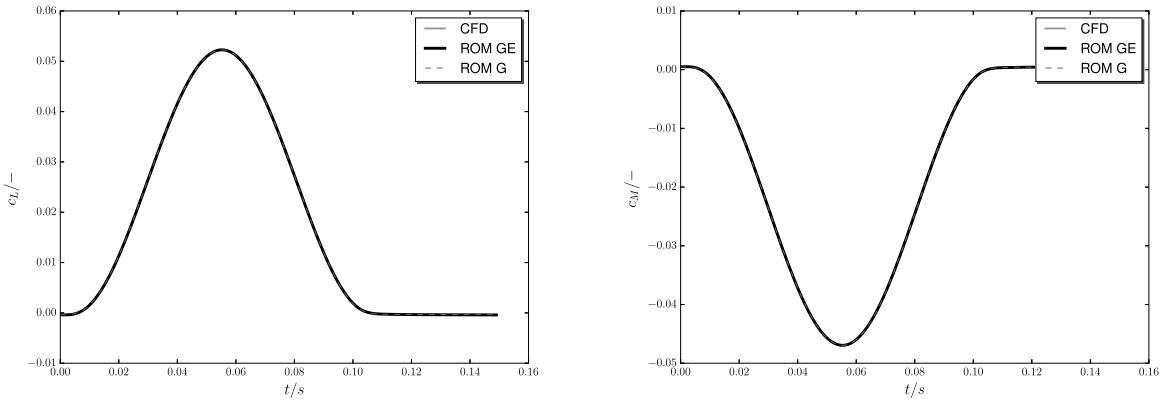
Figure 4: CFD versus ROM E results: (a) global lift coefficient $c_L$ (b) global moment coefficient $c_M$

### 3.2 Rigid - gust trained ROMs

Comparing the global lift $c_L$ and moment $c_M$ coefficient both the gust trained models can predict the gust response of the CFD analysis for the rigid setup, see Fig. 5.

### 3.3 CFD-CSM and ROM-CSM coupling results

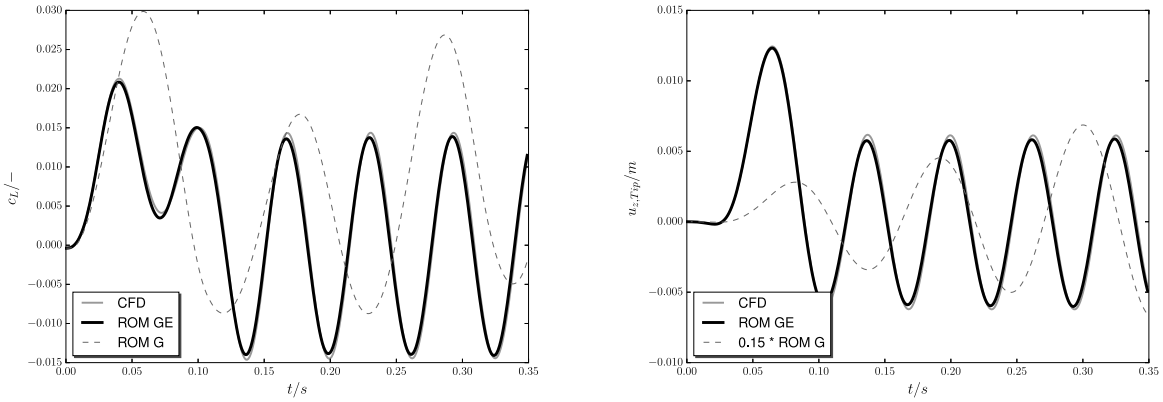
In Fig. 6 the results for the aeroelastic AGARD wing being exposed to a *1-cos* gust with a maximum disturbance velocity of $w_{g0} = 2\frac{m}{s}$ are depicted. Due to the zero angle of attack the wing is initially in rest and reacts to the gust induced change in angle of attack with an increase in lift and hence it deflects. After the gust has completly passed the wing at $t = 0.107$ the wing exibits flutter, as reported by [15] for these flow conditions (see Table 3).
As expected the ROM G - only trained with pseudo disturbance velocities - isn't capable of predicting the aerodynamic response of the elastically deforming wing.




(a) (b)

Figure 5: Results for ROM G and GE compared to CFD: (a) global lift coefficient $c_L$ (b) global moment coefficient $c_M$

The ROM GE - incorporation both gust and elastic modes - can predict the reaction of the wing quite well showing only slight deviations in $c_L$ and $u_{z,Tip}$. The maximum disturbance velocity is limited here to $w_{g0} = 2\frac{m}{s}$ because at higher values the maximum displacement would exceed the maximum trained value, see Table 2. In order to capture higher disturbance velocities the training with elastic modes should be set up with increased maximum generalised coordinates $q_{i,max}$.




(a) (b)

Figure 6: FSI results for model GE with gust amplitude $w_{g0} = 2\frac{m}{s}$: (a) global lift coefficient $c_L$ (b) structural displacement of tip node $u_{z,Tip}$

## 4 CONCLUSIONS

Three different reduced order models have been compared for a gust over a rigid AGARD model. The first ROM trained only with the structural elastic eigenmodes cannot sufficiently capture the gust response, due to missing training modes including displacements and velocities at the wing root.
Both gust trained models without elastic modes (ROM G) and including elastic modes (ROM GE) can accurately predict the gust response in the rigid setup.

In the FSI simulation only the gust *and* elastic modes trained ROM GE can successfully predict the aerostructural behaviour of the AGARD wing in the transonic flight regime. In addition the enhanced ROM GE can still be used to predict the flutter boundary and thus can serves as a surrogate model for multiple flow conditions.

Future studies should include parameter variations for different gust shapes and amplitudes. In addition the presented method could be applied to different configurations, e.g. more realistic 3D configurations including viscous effects. It could also be coupled with flight mechanical simulations in order to resolve the dynamic response of the aircraft.

## REFERENCES

[1] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[2] D. Dimitrov and R. Thormann. DLM-correction method for aerodynamic gust response prediction. *Proceedings" IFASD 2013"*, 2013.

[3] F.A.A. FAR 25.341 Online Document, Feb. 2016.

[4] H. Haddad Khodaparast, G. Georgiou, J. Cooper, L. Riccobene, S. Ricci, G. Vio, and P. Denner. Efficent worst case"1-cosine" gust loads prediction. *Journal of Aeroelasticity and Structural Dynamics*, 2(3), 2012.

[5] M. Haupt, R. Niesner, R. Unger, and P. Horst. Computational Aero-Structural Coupling for Hypersonic Applications. In *9th AIAA/ASME Joint Thermophysics and Heat Transfer Conference*, 2006. San Francisco, California, USA.

[6] R. Heinrich and L. Reimer. Comparison of different approaches for gust modeling in the CFD Code TAU. In *International Forum on Aeroelasticity and Structure Dynamics, Bristol*, 2013.

[7] K. Lindhorst. *Nichtlineare Ersatzmodellierung in transsonischer instationärer Aeroelastik*. PhD thesis, 2015. Forschungsbericht 2015-01.

[8] K. Lindhorst, M. Haupt, and P. Horst. Efficient surrogate modelling of nonlinear aerodynamics in aerostructural coupling schemes. *AIAA Journal*, 52(9):1952–1966, 2014.

[9] K. Lindhorst, M. Haupt, and P. Horst. Aeroelastic Analyses of the High-Reynolds-Number-Aerostructural-Dynamics Configuration Using a Nonlinear Surrogate Model Approach. *AIAA Journal*, 53(9):2784–2796, 2015.

[10] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer London Ltd., London, 1993.

[11] D. Schwamborn, T. Gerhold, and V. Hannemann. On the Validation of the DLR-TAU Code. In *New Results in Numerical and Experimental Fluid Mechanics II*, pages 426–433. Springer, 1999.

[12] H. Spiess. *Reduction methods in finite element analysis of nonlinear structural dynamics*. PhD thesis, 2006.

[13] R. Unger, M. C. Haupt, and P. Horst. Coupling techniques for computational non-linear transient aeroelasticity. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 222(4):435–447, 2008.

[14] J. R. Wright and J. E. Cooper. *Introduction to aircraft aeroelasticity and loads*. John Wiley & Sons, Ltd., 2007.

[15] E. C. Yates. AGARD standard aeroelastic configuration for dynamic response, candidate configuration I. *Wing 445.6, NASA TM-100462, Langley Research Center, Hampton, VA*, 1987.

[16] W. Zhang, B. Wang, Z. Ye, and J. Quan. Efficient method for limit cycle flutter analysis based on nonlinear aerodynamic reduced-order models. *AIAA journal*, 50(5):1019–1028, 2012.

[17] W. Zhang, Z. Ye, Q. Yang, and A. Shi. Gust response analysis using CFD-based reduced order models. In *47th AIAA Aerospace Sciences Meeting Including The New Horizons Forum and Aerospace Exposition, AIAA Paper*, volume 895, 2009.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# RESEARCH ON HYSTERETIC BEHAVIORS OF A SAPARATED SHOCK ABSORBER APPLIED IN RAILWAY BRIDGE

**Aili Li[1], Ri Gao[1], Mingde Sun[1] and Jilei Zhang[1]**

[1] School of Civil Engineering, Beijing Jiaotong University, Beijing 100044, China
E-mail:liaili2016@outlook.com, rigao@bjtu.edu.cn, mingdesun@163.com, lei1024747@sina.com

**Keywords:** separated shock absorber, hysteretic tests, mechanical model, energy dissipation, damping ratio.

**Abstract.** *Shock absorbers are installed between piers and superstructures in railway bridges. They can act as bridge members which bear the longitudinal force and displacement of the bridges in the serviceability state. When large earthquake occurs, the shock absorbers dissipate energies to reduce the seismic responses of bridges and protect the other bridge members. Based on the single cantilever shock absorber, a new structural type of shock absorber is proposed in this paper. To investigate the hysteretic behaviors of the improved shock absorber, quasi-static cyclic tests using displacement-controlled cyclic loading, including two separated shock absorbers, were carried out. Besides, the hysteresis behavior parameters of the separated shock absorbers, such as hysteretic curves, skeleton curves and damping ratio etc., were analyzed. The test results indicate that the hysteresis curve of this kind of separated shock absorbers is plump and stable. The maximum displacement and force could reach 160mm and 220kN, respectively. And bilinear model could be adopted to describe mechanical model for the separated shock absorber. The initial stiffness is high so that the shock absorbs can meet the demand of stiffness in train operation process. Besides, the equivalent viscous damping ratios can reach 0.46. In addition, this type of shock absorber exhibits sufficient deformation capacity and energy dissipation capacity. Therefore, this new separated shock absorber could be an ideal energy dissipator used in the high speed railway bridges.*

## 1 INTRODUCTION

The construction of high-speed railway is developing rapidly in China and the use of bridges takes a great proportion in the high-speed railway. In recent destructive earthquakes, many bridges sustained serious damage, leading to a great number of casualties and enormous economic loss [1-2]. So, improving the seismic behavior of high-speed railway bridges is one of the key problems that must be solved [3]. In the past 20 years, the passive energy dissipation technology has attracted high attention and it has been widely used in the practical structural engineering [4]. The passive energy dissipation technology used in the bridges mainly includes the seismic isolation technology, energy dissipation technology and tuning damping technology [5] .

Dampers which is belong to the energy dissipation technology are able to provide additional damping effects to reduce the seismic responses of the structures. According to the energy dissipation materials, commonly available dampers mainly include viscous dampers, viscous elastic dampers, steel dampers, etc [6]. Many steel dampers dissipate the energy by elastic-plastic deformation. Compared with other dampers, steel dampers offer significant advantages in terms of durability and economic efficiency [7]. However, most of steel dampers,

such as X-shaped ADAS, HADAS device [8-9], buckling restrained brace [10] and so on, were developed for building structures, which may not be able to satisfy the large deformation demands in bridge engineering [11].

Based on the idea of "separate bearing function", a new type of steel dampers named shock absorber was proposed by Professor Li Chenggen and Gao Ri in China [12]. So a new bearing system (Fig.1) is formed by combining the shock absorber and sliding bearing. The design concept is to change the fixed bearing into the sliding bearing so that the vertical and horizontal function can be separated in railway simply supported beam bridge [13]. Meng Xi and Gao Ri investigated the effects of the single cantilever shock absorber on a bridge, the test and numerical simulation results indicated that the single cantilever shock absorber could provided large energy consumption capacity, and protected the bridge structure effectively [14-15]. However, there are some disadvantages of the single cantilever shock absorber, for example, the high cost and inconvenient installation.

Therefore, a novel type of shock absorber called separated shock absorber is proposed in this paper to solve the problems above. The structure of the shock absorber is first presented. So the first part in this paper introduces the designed details of the shock absorber. In order to investigate the mechanical performance of the separated shock absorber, two specimens were tested by using quasi-static cyclic loading schemes in the second part. The third part presents and analyses the test results.

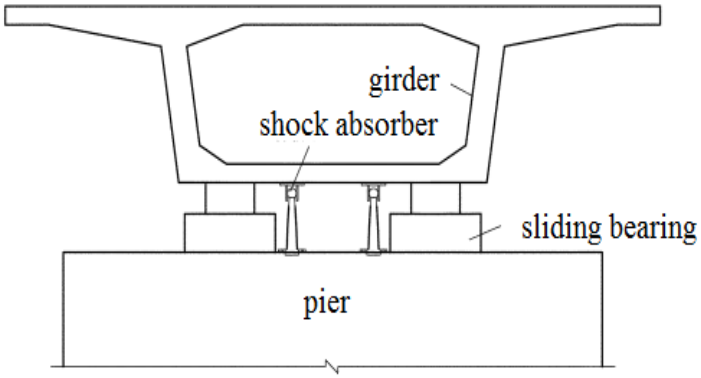



Figure 1: Shock absorber-sliding bearing system [15].

## 2 STRUCTURE OF THE SEPARATED SHOCK ABSORBER

The separated shock absorbers made of mild steel are commonly installed with the sliding bearing between the upper structures and lower piers. And The separated shock absorbers are distributed uniformly around the sliding bearing. The longitudinal stiffness and force of bridge determine the number of the separated shock absorbers.

As shown in Fig.2, the mild steel separated shock absorber mainly consists of two shock absorber bodies, sleeve and two connection plates. The shock absorber body and connection plate are made of the same steel and they process model together. The two connection plates, which are respectively designated as the upper connection plate and lower connection plate, are bolted to the upper girder and lower pier by using high strength bolts. The sleeve has a function that combines the upper shock absorber body and lower shock absorber body by using limit steel pin. Another function of the sleeve is to transfer horizontal force between the two shock absorber bodies. The transverse section shape of shock absorber body is round and the body can be divided into two parts, i.e. load transfer part and deformed part. When a large earthquake happens, the relative displacement will be produced in horizontal direction between the upper connection plate and lower connection plate. Then the shock absorber bodies

occur yielded plastic deformation to dissipate energies, thus reducing the responses of the bridge.

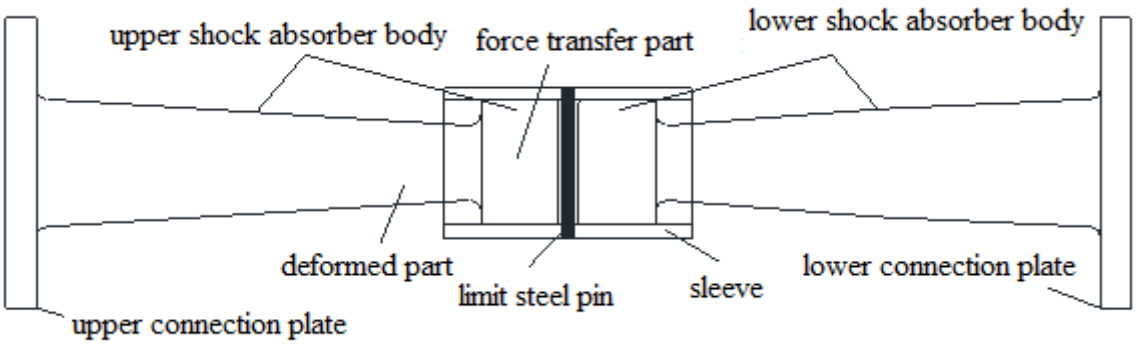



Figure 2: Schematic diagram of separated shock absorber.

At the stage of design, the shock absorber body can be seen as cantilever beam. According to the material mechanics, for more sections of the deformed part along the same shock absorber body to become plastic, the relationship between diameter d and distance x (Fig.3) is presented as follows:

$$d = \begin{cases} d_1 & (0 \le x \le L_1) \\ \partial(\Delta + \dfrac{L-\Delta}{L}x)^{1/3} & (L_1 \le x \le L) \end{cases} \tag{1}$$

Where x is the distance from the top of the shock absorber body to cross section and $d_1$ is the diameter of the force transfer part, which is fixed constant only in connection with the height of the separated shock absorber. $\partial$ represents the section coefficient and it is fixed value. $L_1$ and $L$ are the height of force transfer part and shock absorber body, respectively. $\Delta$ is the vertical deformation of the shock absorber body when the shock absorber occurs the designed maximal horizontal displacement.

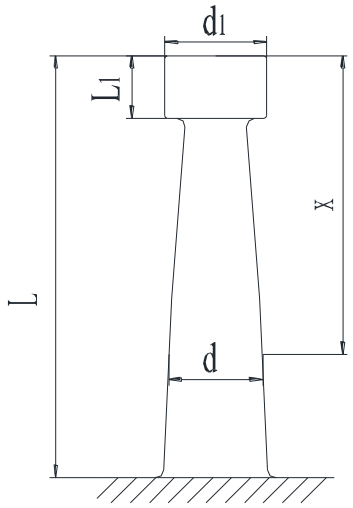



Figure 3: Theoretical calculation model.

The height of deformed part largely determines deformation capacity of the separated shock absorber. Hence the deformation capacity of the shock absorber can be enlarged by increasing the height of the shock absorber body. In addition, such design can significantly improve the low cycle fatigue behavior of the separated shock absorber because it avoids the concentration of plasticity at one cross section.

Compared with the single cantilever shock absorber, the structural form of separated shock absorber is simpler. And the shock absorber body can be more easily installed and replaced if damaged.

## 3 EXPERIMENTAL PROCESS

### 3.1 Specimen design

There were two specimens designed for this experiment and they were named A1 and A2. The dimensions of two specimens were the same. The major parameters of specimens are given in the table 1.

The designed maximal horizontal displacement was 160 mm in this experiment. As mentioned before, the value of $\Delta$ was 18.8 mm through finite element analysis with the maximal horizontal displacement. The length and inner diameter of the sleeve with a thickness of 40 mm were 250 mm and 120 mm, respectively. Specimen adopt the connection plate with a height of 30 mm and its diameter was 250 mm. The number and diameter of the high strength bolts were 6 and 32 mm, respectively.

The shock absorber body and connection plate for the two specimens are all made of one specially made mild steel in China. Uniaxial tensile test for the mild steel was carried out. The results indicated that the yield strength and ultimate tensile strength were 250 Mpa and 420 Mpa, respectively. In order to limit the vertical displacement of the shock absorber body, the strength of sleeve must be higher than the strength of the specially made mild steel. So we choose 45 steel (in Chinese rule) as the base metal of sleeve.

| Parameter | $L$ (mm) | $L_1$ (mm) | $d_1$ (mm) | $\partial$ |
|---|---|---|---|---|
| Value | 430 | 75 | 120 | 15.472 |

Table 1: The major parameters of specimens.

### 3.2 Test setup and loading scheme

The test was conducted in the structural laboratory of Beijing Jiaotong University. Fig.4 gives a photograph of the loading frame, which mainly includes one truss and an electro-hydraulic loading system. The loading system mainly consists of three subsystems: the controlling system, oil hydraulic and measurement system.

A1 and A2 are to investigate the hysteretic behaviors subjected to cyclic loading. And the experiment was all conducted using a displacement controlled cyclic loading. The loading scheme follows a triangular loading pattern with increasing amplitudes. The first cycle amplitude was 2 mm, and increased by 2 mm for the succeeding cycles until it reached 20 mm. After 20 mm, the amplitude was increased by 10 mm until it reached 80mm. Then the amplitude was increased by 20 mm until it reached 160 mm. When the amplitude was smaller than 20 mm, the loading speed was 2 mm/s. While the amplitude was larger than 20mm, the loading speed was 8 mm/s. And the loading test was repeated for 5 cycles at every amplitude.



Figure 4: Test setup.

## 4 EXPERIMENTAL RESULTS

### 4.1 Experimental phenomena and analysis

The experimental phenomena of A1 and A2 was turned out to be similar. When the amplitude was smaller than 30 mm, the center line of separated shock absorber was approximately a straight line which was perpendicular to the horizontal ground at the equilibrium position. When the amplitude was from 30 mm to 140 mm, there were no cracking on the mild steel separated shock absorber. However, we found that the deformed parts of the shock absorber bodies were buckled. At the amplitude of 140 mm, the first micro-crack was observed at the junction site of shock absorber body and connection plate. Finally, as the amplitude increased to 160 mm, the crack was not expanding. The photographs of two specimens' crack at the fifth cycle for amplitude of 160 mm were shown in Figs.5(a), 5(b). It is found that the two specimens sustained large deformation without fracture for the designed maximal horizontal displacement.

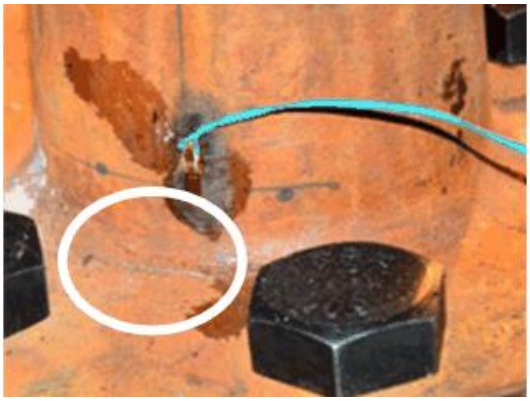


Fig.5(a): The crack state of A1

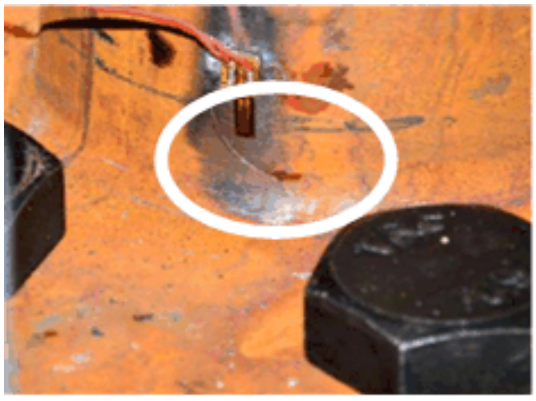


Fig.5(b): The crack state of A2

Figure 5: The crack states of A1 and A2 at the amplitude of 160mm.

After that, the loading test was still conducted at the amplitude of 180 mm. Finally, loading at the amplitude of 180mm for A1 was repeated for six cycles. while A2 got fractured at the

fifth cycle 180 mm loading amplitude. The conjunction crack of the shock absorber body and connection plate failed suddenly. This was because that the thickness of connection plate was too thin to resist pulling force of the high strength bolts. And another reason is that the conjunction is stress concentration area. It is recommended to increase the thickness of the connection plate to add the pulling force and enhance the conjunction of the shock absorber body and connection plate.

## 4.2 Hysteretic behavior

### 4.2.1 Hysteretic curve and skeleton curve

The loading-displacement hysteretic loops of two specimens at the third cycle for every amplitude were plotted in Figs.6(a), 6(b). From the Fig.6, we can see that the two hysteretic curves are plumper and the shape of the hysteretic loops are spindle. The elastic stiffness of two specimens is high in the condition of small displacement. When the displacement reaches about 20 mm, the separated shock absorbers sustain significant plasticity. Besides, degradations of stiffness and strength are not obvious. The low cycle fatigue behaviors of A1 and A2 are not obviously produced at the amplitude of 160 mm.

As presented in Fig.6(a) and Fig.6(b), the hysteresis loops of A1 and A2 are similar. It indicates that the properties of separated shock absorber are stable.

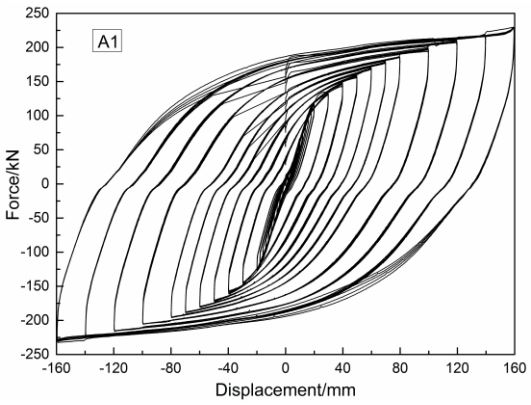



Fig.6(a): A1

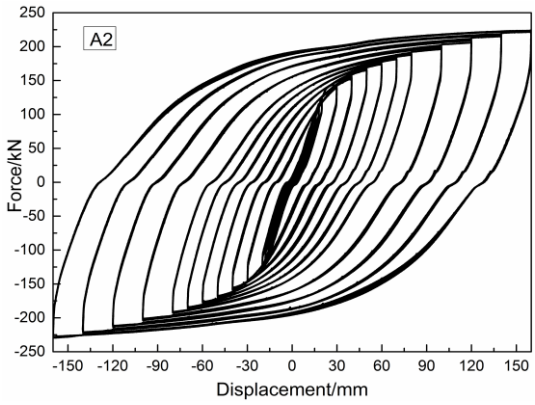



Fig.6(b): A2

Figure 6: Hysteresis curve for the specimens:(a) A1 and (b) A2.

### 4.2.2 Skeleton curve

Fig.7 shows the skeleton curves for A1 and A2 from envelope diagram of the hysteresis loops. It could be seen from the figure that the skeleton curves are approximately bilinear. Therefore, the mechanical model for the separated shock absorber could be described by bilinear mode. The bearing capacity of both A1 and A2 could reach 220 kN when loading amplitude reached 160 mm. And it is not decreased of the bearing capacity for A1 and A2 with increasing of displacement. So it exhibits sufficient deformation capacity of the mild steel separated shock absorber.

From the Fig.7, the skeleton curve for A1 was almost coincided with that of A2. The mechanical behaviors of the mild steel shock absorber are stable.

## 4.3 Energy dissipation capacity

As mentioned above, the primary function of the mild steel shock absorber is to dissipate energy and protect the bridge structure under large earthquake occurs. In this paper, the

equivalent viscous damping ratio $\xi$ is taken to evaluate the energy dissipation capacity of the mild steel separated shock absorber. The calculating formula is as follow [16].

$$\xi = \frac{E_{DS}}{2\pi E_S} \tag{2}$$

Where $E_{DS}$ is the damping energy dissipation and it equals to the area of hysteresis loop in one cycle. $E_S$ is the maximal strain energy.

The comparison of equivalent viscous damping ratio between A1 and A2 is shown in the Fig.8. It is found that the equivalent viscous damping ratio increases with the increase of displacement. The growth becomes slowly after the displacement is larger than 80 mm. And the maximum damping ration could be reach 0.46.

In addition, the equivalent viscous damping ratios of A1 and A2 are almost at the same. Hence the energy consumption of mild steel separated shock absorber is excellent, and the aseismic performance is stable.

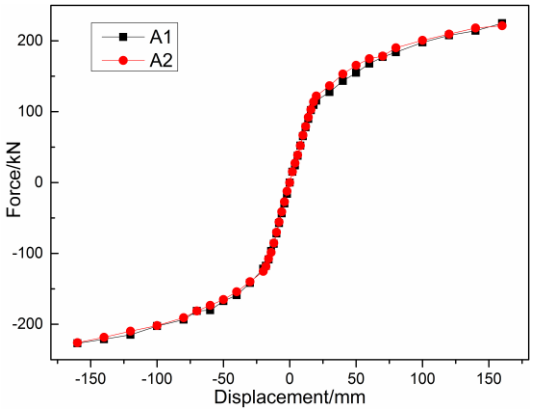



Figure 7: Skeleton cure for A1 and A2.

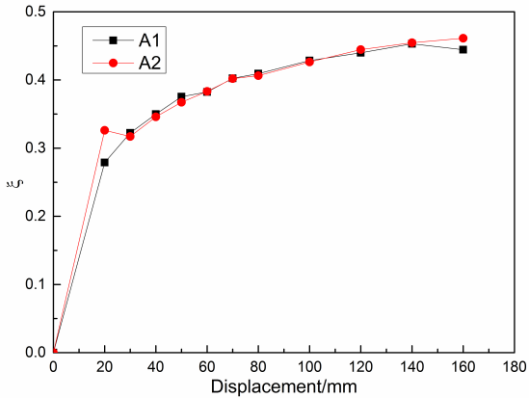



Figure 8: The damping ratio for A1 and A2.

## 5 CONCLUSIONS

New type of shock absorber for railway bridges has been proposed in this paper. The quasi-static tests of two specimens were conducted to investigate hysteretic performances. The major findings obtained are as follows:

- The separated shock absorber consists of two shock absorber bodies, sleeve and two connection plates. The shock absorber body is the main energy dissipation region through plastic deformation.
- Experimental results show that this shock absorber has good strength and deformation capacity. The hysteresis curve is stable and plump. The mechanical model for the separated shock absorber could be described by bilinear mode.
- The initial stiffness is high so that the shock absorbs can meet the demand of stiffness in train operation process. Besides, the energy consumption of mild steel separated shock absorber is excellent the aseismic performance is stable.
- The structural type of separated shock absorber is reasonable except that the thickness of the connection plate must increase. And it provides a theoretical basis for its future design.

## 6 ACKNOWLEDEMENT

Financial support from the Project of National High-tech Research and Development Program of China (Program863) under Grant Nos. 2008AA030707.

## REFERENCES

[1] Liu LB. Bridge design concepts and methods of seismic and wind resistant, *Ph.D Thesis*, Chang'an University, 2009.(in Chinese)

[2] Wang LP. Damping performance research in high-speed railway, *Master Thesis*, Central South University, 2011. (in Chinese)

[3] Wang Y. Research on seismic performance and fragility of isolated railway bridges, *Ph.D Thesis*, Zhejiang University, 2013. (in Chinese)

[4] G. W. Housner, L. A. Bergman, T. K. Caughey, et al. Structural control: past, present, and future. *Journal of Engineering Mechanics*, 1997, 123(9):897-971.

[5] Pei YL, Lin TK. Control of seismically isolation bridges by magneto rheological dampers and a rolling pendulum system, *Structural Control and Health Monitoring*, 2012,19(2):278-294.

[6] Zhou Y. Design of metallic energy dissipation damping structures. Wuhan: Press of Wuhan University of Technology, China, 2006. (in Chinese)

[7] Yasushi K, Mitsuo S and Toshikazu Y. Structural design of a tall building with elastic-plastic steel dampers for the attention of torsional and lateral motions. *Structure Design Tall Build*, 1998, 7:21-32.

[8] A.S Whittaker, V.V. Bertero, C. L. Thompson, et al. Seismic testing of steel plate energy dissipation devices. *Earthquake Spectra*, 1991,17(4):563-604.

[9] K.C. Tsai, H.W. Chen, C. P. Hong, et al. Design of steel triangular plate energy absorbers for seismic-resistant construction. *Earthquake Spectra*, 1993,19(3):505-528.

[10] Zhao J X, Wu B, Ou J P. Uniaxial quasi-static cyclic tests on the hysteretic behavior of a novel type of all-steel buckling restrained brace. *China Civil Engineering Journal*, 2011, 44(4):60-70. (in Chinese)

[11] Pan P, Yan H, Wang T, et al. Development of steel dampers for bridges to allow large displacement through a vertical free mechanism. *Earthquake Engineering and Engineering Vibration*,2014,13(3):375-388.

[12] Li CG, Gao R. Study on the shock absorbing technique of high-speed railway bridges. *Engineering Sciences*. 2009,11(1):81-86. (in Chinese)

[13] Li AL, Gao R, Sun MD, et al. Study on low-cycle fatigue life of base metal of the shock absorber. *2014 International Conference on Materials Science and Engineering Technology.* Shanghai. Trans Tech Publications Ltd, 2014, 936:1361-1365

[14] Meng X, Gao R, Li CG. Theoretical and experimental study of hysteretic behavior of shock absorber. *Bridge Construction*, 2015, 45(3):20-25. (in Chinese)

[15] Meng X, Gao R, Li CG. Design and seismic mitigation performance study of shock absorber for railway simply-supported beam bridge. *Bridge Construction*, 2014, 44(3):81-86. (in Chinese)

[16] A. K. Chopra. Dynamics of structures: theory and applications to earthquake engineering. Englewood Cliffs, NJ, USA: Prentice Hall,1995.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FINITE ELEMENT ANALYSIS OF TIMBER BEAMS WITH FLAWS

**Janka Kováčiková[1,2], Mats Ekevad[1], Oľga Ivánková[2],and Sven Berg[1]**

[1] Luleå University of Technology
Division of Wood Science and Engineering, SE-931 87 Skellefteå, Sweden
e-mail: janka.kovacikova@stuba.sk,{mats.ekevad,sven.berg}@ltu.se

[2] Slovak University of Technology in Bratislava
Department of Structural Mechanics, Radlinskeho 11, SK-810 05 Bratislava, Slovakia
olga.ivankova@stuba.sk

**Keywords:** FEA, Fracture Mechanics, Timber Beams, XFEM, Wood.

**Abstract.** *The work was focused on modelling of cracks and simulations of their propagation in timber beams. The aim was to find out the influence of flaws on load-carrying capacity. In beam design and beam inspection, it is necessary to determine the load-carrying capacity of a beam with flaws. There is not much information in literature about the influence of flaws on bending and shear strength of timber beams. Standards for fracture mechanics design approach for timber structures are not easily available.*

*The results from simulations of loaded timber beams with flaws are discussed in this paper. Two different types of timber beams were analysed. First type was a sawn timber beam and for that four different models were analysed: first model was a beam without flaw; second model was a beam with a straight central flaw; next was a beam with an oblique crack and fourth beam had a round hole in the middle of the span. Second type of beam was a glued laminated timber beam (glulam). For this type, five different models were analysed. First four variants were the same as for the solid timber beam and the fifth variant had an initial crack along a glue line. We analysed these types of flaws, since they are the most common flaws in wood.*

*Simulations were conducted in ABAQUS. Material properties of wood used in the models were retrieved from standards; C24 for solid timber and GL24c for glulam timber beams. Dimensions of beams were the same for all variants and beams were considered simply supported. For an analysis of crack propagation, linear elastic fracture mechanics was considered. Modelling fracture was conducted using the extended finite element method (XFEM). The energy approach was used for the analysis of crack propagation. Comparing results for solid and glulam beams with and without cracks gave us an overview how different flaws influence load-carrying capacity of the beams and under which loading failure occurs. Simulating timber beams is more complicated compering to steel or plastic. By understanding how to simulate flaws in wood material it is possible to obtain reliable results with finite element analysis.*

## 1 INTRODUCTION

Sawn timber and glued laminated beams (glulam) are used as bearing structure members of building constructions. Flaws which occur in beams due to different causes can have an effect on their load-carrying capacity. These imperfections can, in the worst case scenario, cause an ultimate failure of the constructions.

This study is focused on simulating different types of flaws in beams and analysing their effect on fracture behaviour and load- carrying capacity of beams. Energy criterion approach was assumed for fracture analysis. This approach states that a crack will grow when the energy required for the crack growth overcome the resistence of the material. $G$ is the energy release rate and the critical value of $G$ is the resistence of the material $G_c$. A crack occures at the moment when $G=G_c$ and will grow when $G>G_c$. This fracture parameter describes fracture behaviour of solids taking into account stress-strain state, and deflections in the area near the crack [1].

Fracture toughness is the ability to withstand flaws that initiate failure. This parameter can be defined for three ways of loading a crack (mode I, II, and III) (Fig.1). The crack or failure occurs when crack is loaded in one of these modes or by their combination.

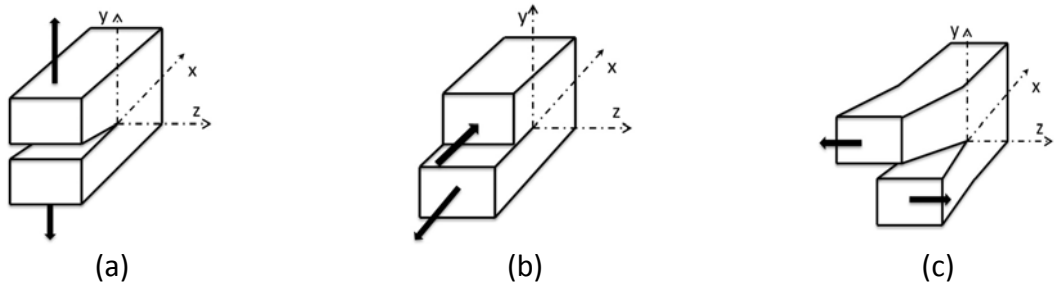



Figure 1: Modes of loading of a crack: (a) mode I, (b) mode II, (c) mode III

In an orthotropic material such as wood we need to consider not only the mode of loading but also in which plane the crack lies and in which direction it may propagate [2]. A crack can lie in six planes and propagate in one or two direction. Considering 6 planes and 3 modes of loading, 18 basic fracture situations can be identified (Fig.2) [3].

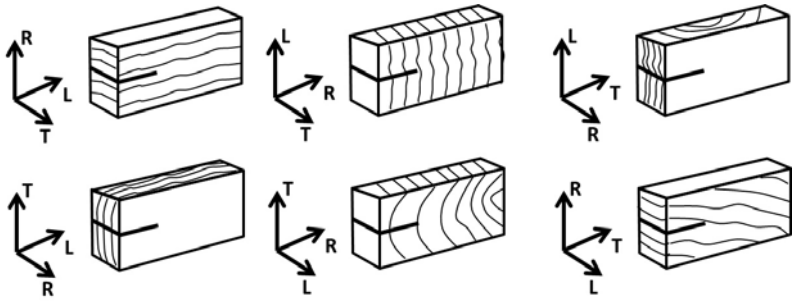



Figure 2: Possible crack propagation situation in wood

Fracture toughness is determined from values of stress intensity factors (SIFs) in modes I, II, III. Stress intensity factors have usually a subscript to denote the mode of loading i.e., $K_I$, $K_{II}$ and $K_{III}$. Some closed form solutions for stress intensity factors can be found in fracture mechanics handbooks. These solutions differs for case of loading, stress fields ahead of a crack tip, mode of loading, shape of a crack,… There is a relationship between fracture

toughness and critical values of energy release rates. When we consider this relationship, we obtain three different values which again depend on the mode of loading; i.e. $G_I$, $G_{II}$; $G_{III}$.

Simulating wood fracture behaviour is a challenging task. Especially, when we consider orthotropic wood models (sawn beams) or transverse isotropic wood models (glulam). Simulations are always three-dimensional (3D). However, investigating fracture behaviour of wood can help us to understand how wood will behave in certain circumstances and how flaws influence its characteristics.

This study is focused on modelling and simulating beams with flaws which can occur on timber beams in practical applications. The effect of the flaws on shear and bending capacity of the timber beams is studied.

## 2 MATERIAL AND METHODS

For this study, two types of beams: sawn timber, labelled N, and glued laminated beams, LN, and five different flaws models were considered. First four models were same for N beams and LN beams. Fifth model was created only for LN beam. First model was beam with no crack (Fig.3 (a)); second model had a central vertical crack (Fig.3 (b)); third beam had an incline crack (Fig.3 (c)); fourth variant was a model with a circular hole (Fig.3 (d)); and last model was a beam with an initial horizontal crack (Fig.3 (e)).

Simulations have been conducted in ABAQUS 6.14-4. Material for sawn timber beam was considered orthotropic linear elastic, which means we used 9 independent elastic constants for describing material response. Material property for glulam beams was considered transverse isotropic. Therefore, 4 independent constants were taken into account. Constants used for material modelling are shown in (Tab.1).

Global coordinate system LRT was used where L=1=X, R=2=Y, T=3=Z and also local coordinate systems were assigned for lamellas in glulam beams describing material orientation. In our simulations, we considered the crack in TR plane and loaded by mode I for first four models. For fifth model, the crack was in RL plane and was loaded by a mixed mode (mode I + mode II).

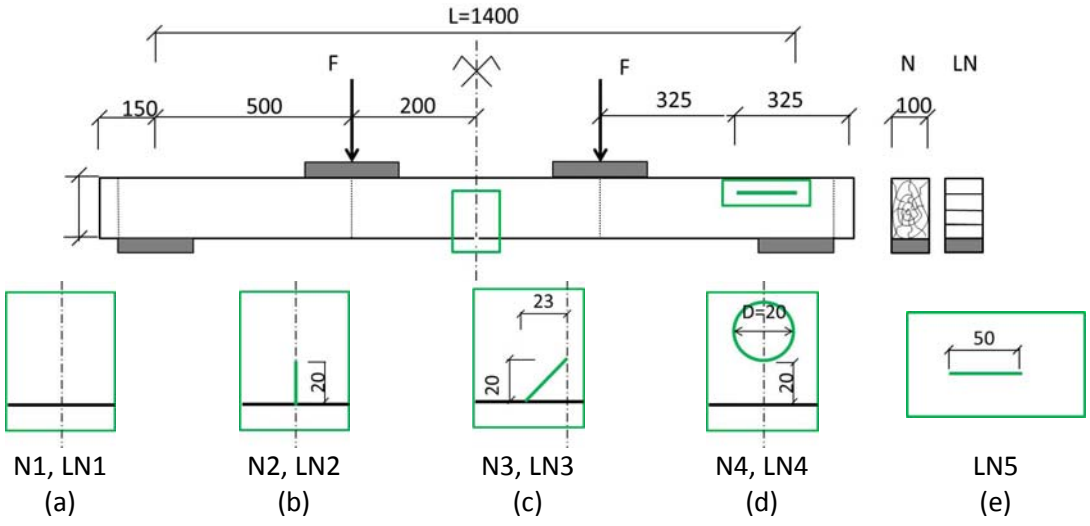



Figure 3: Variants of beams

Loading control was used for loading of N1, LN1 beams and deflection control for other types of beams. Therefore, N1 and LN1 beams were loaded by two equally concentrated forces (Fig. 3). Values of loading forces were the ultimate maximum bending and shear forces. Loading forces were: bending force was 20.48kN for N1 and LN1 (characteristic value of

maximum bending force), shear force 49.23kN and 44.8kN for N1 and LN1, respectively (design value of maximum shear force). Forces were analytically obtained considering maximal bending moment; bending strength and maximal shear force; shear strength of a timber simply supported beam with a rectangular cross section loaded by two equal concentrated forces. Beams N2 - N4 have been loaded by 13.32mm and LN2 - LN5 by 14.9mm which are deflections obtained from a conditional finite element analysis of N1 and LN1.

Resulting ultimate forces were obtained by analysing dissipation energy and load-deflections diagrams. Ultimate force was force where a considerable jump in dissipation energy comes into view. This is the force under which beams fail. Also we observed the crack initiation force. It is the force for which a crack starts to propagate; however, jump in dissipation energy is not so significant. In some cases ultimate force is equal to crack initiation force.

| Label | | Density | Modulus of Elasticity | | | Poisson's ratio | | | Shear modulus | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\rho$ | [MPa] | | | [-] | | | [MPa] | | |
| | | [kg/m$^3$] | $E_{LL}$ | $E_{RR}$ | $E_{TT}$ | $\nu_{LR}$ | $\nu_{LT}$ | $\nu_{RT}$ | $G_{LR}$ | $G_{LT}$ | $G_{RT}$ |
| N | C24 | 350 | 10700 | 710 | 430 | 0,38 | 0,51 | 0,31 | 500 | 620 | 23 |
| LN | GL24c | 385 | 11000 | 300 | 300 | 0* | 0* | 0* | 650 | 650 | 23 |
| *Poisson ratio may be zero value according EC5 [4] | | | | | | | | | | | |

Table 1: Engineering constants

The XFEM-based LEFM [5] approach was used for simulating crack propagation along an arbitrary, solution-dependent path. Two crack initial criterions were used: the maximum principal stress criterion (MAXPS), and the maximum nominal stress criterion (MAXS). The values for criterions were obtained analytically by considering fundamental laws of elasticity of materials. These criterions were user defined. Value of MAXSP criterion was 24MPa. For defining MAXS criterion it was necessary to define normal-only, first direction and second direction mode. Therefore, for normal-only and first direction mode 24MPa load was assumed for both types of beams (N, LN). For second direction mode, the criterion was 4.6MPa and 4.2MPa for N and LN beams respectively.

The virtual crack closure technique (VCCT) fracture criterion was defined. For computing the equivalent fracture energy release rate, the Benzeggagh and Kenane (BK) law was used. The BK law model was described by [6,7,8]. Critical values of the energy release which were needed for describing the fracture criterion were: $G_{IC}$=176N/m, $G_{IIC}$=734N/m [9]. Value of $G_{IIIC}$=10000N/m was considered in order to avoid mode III loading for a crack. C3D8 element type for enrichment areas and C3D8R element type for other parts were used in simulations. Enrichment areas were placed where the crack should occur and propagate. In these areas it was needed to consider enrichment functions for the purpose of fracture analysis [5]. Problem size is shown in Tab.3. Interaction between a glue line and lamella was created using surface to surface hard contact.

| Number | N1 | LN1 | N2 | LN2 | N3 | LN3 | N4 | LN4 | LN5 |
|---|---|---|---|---|---|---|---|---|---|
| Elements | 51676 | 63224 | 60692 | 51746 | 117260 | 51746 | 60884 | 54496 | 52768 |
| Variables | 265314 | 372087 | 142188 | 247021 | 428196 | 197988 | 316158 | 178665 | 271398 |

Table 3: Problem size

## 3 RESULTS

The values of maximum forces together with the values of crack initiation forces for all beams are shown in Tab. 4. These values were obtained by analysing damage dissipation energy and load-deflection curves. The load-deflection diagram for all beams is shown in Fig. 4. In these diagrams is shown relationship between deflections vs. maximal load. Deflection was 7.83mm for N1, 8.13mm for LN1, 5.18mm for N2, 7.28mm for LN2, 3.9mm for N3, 9.85mm for LN3, 4.42mm for N4, 6.99mm for LN4, and 15.86mm for LN5.

For N type beams we obtained following results: N1 (referential) beam failed under 21.35kN force. Comparing to the referential flaw-free beam, N1, the load-carrying capacity decreased about 54% for N2, 39% for N3 and 23% for N4.

| Force [kN] | Beam | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **N1** | **LN1** | N2 | LN2 | N3 | LN3 | N4 | LN4 | LN5 |
| Max. force | **21.35** | **23.21** | 9.73 | 10.66 | 13.02 | 21.37 | 16.41 | 19.40 | 45.55 |
| Initial crack force | **21.08** | **23.21** | 7.89 | 8.07 | 13.02 | 10.33 | 16.41 | 12.59 | no |

Table 4: Resulting values for beams obtained by XFEM analysis

LN1 beam failed under 23.21kN loading force. Load-carrying capacity decreased about 54%, 8%, and 16% for LN2, LN3, LN4 beams respectively. Different results were obtained for LN1 and LN5 beams. In this case, the load - carrying capacity increased about 96%.

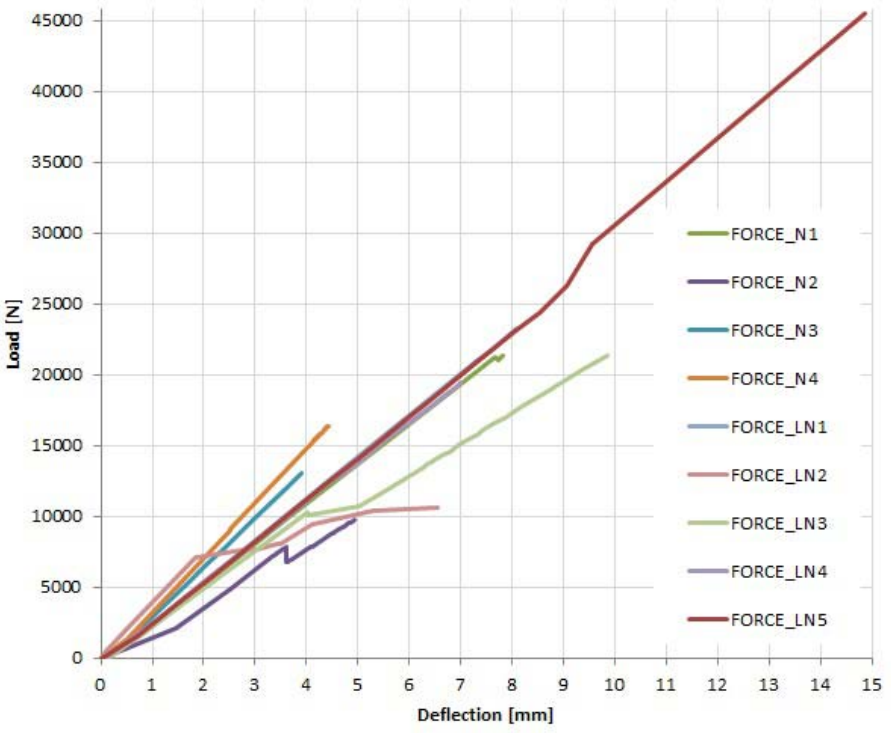



Figure 4: Load-deflection diagrams

## 4 DISSCUSION

Analysing results we can state that the worst influence on load-carrying capacity within the studied models has a vertical initial crack in the middle of a beam's span, for N2 and LN2

beams. In this case, beams will fail under much lower force compared to the flaw-free beam. N2 and LN2 beams can carry only 46% of the referential load. Load - carrying capacity decreased significantly also for the N3 type beam. This beam could carry only 61% of the reference beam loading. For the LN3 beams decreasing of load-carrying capacity is not so significant. Glulam beam with the oblique crack can carry 92% of the referential loading. Significant influence on load-carrying capacity had the circular hole in the middle of the beam's span. The N4 and LN4 beams can withstand 77% and 84% of the referential force, respectively. Interesting increase in load-carrying capacity was observed for the LN5 beam. In this case capacity of beam increased about 96% for 14.9 mm deflection. No jump in dissipation energy occurred in this case. This significant increase of capacity might be caused by stress concentration near this crack and the position of the initial crack. However, it might be also due to numerical errors in FEA analysis. It is needed to study more models and to look at the influence of user-defined parameters. In order to understand why this significant increase of load-carrying capacity occurred, it is necessary to compare the simulation results with experimental ones.

Generally, we can say that flaws have negative influence on load-carrying capacity of beams. Influence of flaws on load-carrying capacity is worse on sawn beams than on glulam beams. It validates our hypothesis that glulam beams are able to bear a higher loading. Positive effect on load-carrying capacity of beams had the horizontal crack on glulam beam. Future work will be focused on studying the effect of this crack type on load-carrying capacity.

## REFERENCES

[1] T.L. Anderson, *Fracture Mechanics - Fundamentals and Applications, Third edition*. Boca Raton : CRC Pres. Taylor & Francis Group, 2005. p. 611. 13: 978-1-4200-5821-5.

[2] Laboratory, Forest Products. *Wood Handbook. Wood as an Engineering Material.* Madison : United States Department of Agriculture Forest Service, 2010. p. 508.

[3] S. Aicher, P.J. Gustafsson, P. Haller, H. Petersson, *Fracture mechanics models for strength anylysis of timber beams with hole ore notch.* A Report of RILEM TC-133, Lund : Structural Mechanics, LTH, Sweden, 2002. ISSN 0281-6679

[4] EN 1995-1-1:2004, Euro code 5

[5] 6.14-4, Abaqus/CAE Users guide.

[6] D. Xie., S.B. Biggers Jr, *Progressive crack growth analysis using interface element based on the virtual crack closure technique.* 11, 2006, Finite Elements in Analysis and Design, Zv. 42, p. 977-984. 0168-874X.

[7] G.E. Mabson, L.R. Deobald, B. Dopker, D.M. Hoyt, J.S. Baylor, D.L. Graesser, *Fracture interface elements for static and fatigue analysis..* [ed.] 16th International conference on composite materials. Kyoto : s.n., 2007.

[8] A. Leski, *Implementation of the virtual crack closure technique in engineering FE calculations.* 3, 2007, Finite Elements in Analysis and Design, Zv. 43, p. 261-268. 0168-874X.

[9] P. Haller, R. Putzger, *Fracture Energy in Mode I and Mode II of Reinforced Wood.* Dresden : Technische Universität Dresden: Eigenverlag, 2003. Proceedings of the 2nd Colloquium on Textile Reinforced Structures CTRS2. s. 247-258.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
Crete Island, Greece, 5–10 June 2016

# NONLINEAR FORCED RESPONSE OF A STATOR VANE WITH MULTIPLE FRICTION CONTACTS USING A COUPLED STATIC/DYNAMIC APPROACH

**M. Lassalle[1], C. M. Firrone[2]**

[1]Politecnico di Torino, Department of Mechanical and Aerospace Engineering, Aer-Mec Laboratory
Corso Duca degli Abruzzi, 24, Torino, Italy
email: marco.lassalle@polito.it
www.aermec-dimec.polito.it

[2]Politecnico di Torino, Department of Mechanical and Aerospace Engineering, Aer-Mec Laboratory
Corso Duca degli Abruzzi, 24, Torino, Italy
email: christian.firrone@polito.it
www.aermec-dimec.polito.it

**Keywords**: Vane-segment, non-linear dynamics, friction joint, MHBM, coupled static/dynamic approach, coupled/uncoupled comparison

**Abstract.** *The purpose of this paper is the nonlinear calculation of the forced response of a vane-segment (a sector of a row of stator blades) for aeronautical applications in the presence of multiple joints using a coupled static/dynamic method to calculate the contact forces between assembled components. The solution of the static/dynamic equilibrium equations is calculated using the Multi Harmonic Balance Method (MHBM) that is a technique widely accepted as an effective approach to calculate the steady-state forced response in the frequency domain in spite of the Direct Time Integration (DTI).The MHBM technique allows, in fact, a strong reduction of the computation in particular when non-linearities are introduced as in this case for a more realistic modeling of the physical behavior of the joints and system response in general.*

*The non-linearity is introduced by means of contact elements that produce additional stiffening and damping in the system due the introduction of contact stiffness and friction forces based on Coulomb's law. Since the periodical nonlinear contact forces may depend on the history of the motion of the vibrating system, the MHBM steady-state solution may lead to a poor correlation with the DTI solution. To consider this dependence, in this paper, it is presented a coupled method which permits to solve together the static balance equations and the dynamic balance equations to have a better approximation of the real behavior of the system.*

*A sensitivity analysis is performed to investigate the effect of numerical, geometrical and contact parameters to optimize the design of the system. Calculated responses are finally compared with the uncoupled approach where the static equilibrium is calculated independently from the vibratory response calculation as usually done to investigate the difference of the two methods.*

## 1. INTRODUCTION

The forced vibrations in resonance conditions shorten the life and reliability of aircraft engines and turbo-machinery in general. Although the sources of dynamic excitation producing oscillations of the whole structure are many, the time-varying forces loading gas turbines of aeronautic engines are caused by the considerable non-uniformity of the flow of the hot gases that pass through the turbine blades [1]. In a regular operating condition, the aerodynamic excitation pattern loading the blade array can be represented as a combination of multiple excitations having a harmonic pattern in space along the circumferential direction that travel with respect to the bladed disk. The circumference is divided into an integer number (called Engine Order EO [2]) of wavelengths for each travelling harmonic force. The single blade is therefore subjected to a wide excitation spectrum and multiple excitations at resonance condition may occur for a given rotating speed of the rotor. The same issue occurs for stator vane segments as well where travelling forces are generated by the wakes of front and aft-rotating disks. For this reason the designers of aircraft engines are called to introduce damping systems to reduce peak stress values during the vibratory phenomenon. Friction dampers are often used to dissipate energy in the system through friction, i.e. the amplitudes of oscillation are reduced increasing the fatigue life of the blade arrays. The major sources of friction damping in the bladed disk turbine are attributable to the blade-disc interfaces (blade root joint) [3,4], the contact between adjacent blades connected by interference at the tip (shrouds) or mid-span airfoil (snubber, [5]) and the presence of underplatform dampers [6, 7]. Typical joints for vane segments are the so-called interlocking joints [8] located at the inner radius of the sector connecting adjacent sectors.

In this paper, a particular excitation is considered to study the forced response of a stator vane segment subjected to Engine Order 1 travelling force. It is verified, in fact, that a mechanical excitation due to the unbalance of the rotor shaft in the engine may excite the vane segment according to the 1 per rev excitation frequency at not negligible levels. For this reason it is interesting to verify at what extent the interlocking contacts are able to limit the vibration amplitudes. Moreover, the study is extended assuming that also the joint connecting the vane segment to the casing (hook contacts) may act as a friction damper that is added to the interlocking contact. In the first part of the paper the vane segment is described and the rationale of the study is specifically introduced; in the second part of the paper the modelling of the vane segment and of the contact model are illustrated; in the third part the methodology to solve the equation of motions are explained; in the fourth part the sensitivity analysis and the comparison with the uncoupled method is shown. The uncoupled method provides the normal loads acting on the contacts before the vibratory response is calculated. Usually a nonlinear static calculation of the equilibrium is performed to achieve the set of normal loads that are applied on the contact and these values are kept constant through the forced response calculation. In this sense the normal load are a sort of contact pre-load in the system that is acceptable to simulate the assembly phase when stator vanes are mounted but not the operative phase. It was proven that in the presence of vibrations, when slip occurs, the normal pre-load may vary for a single contact when the tangential and normal direction of the contact are both involved in the vibrations. Furthermore, in case of multiple contacts in the vibrating system as in this case, if one of the contact is in slip state during vibrations, it may change the normal pre-load of other contacts in the system. These are the main reasons to adopt a coupled approach instead of the uncoupled approach. The conclusions are drawn in the final part of the paper.

## 2. TEST ARTICLE AND EXPERIMENTAL EVIDENCE

The problem studied in this paper is to evaluate the forced response of a stator vane segment shown in **Figure 1** in the presence of a predominant radial excitation due to the unbalance of the rotor shaft. The excitation is transferred from the unbalanced rotating shaft to the bearing outer race and then to the

struts connecting the engine casing and finally to the vane segment that is connected to the casing by means of the hook joints (**Figure 2**). For this reason the load can be described as a radial force applied to the casing with unitary engine order (EO1, **Figure 3**) since the excitation frequency of a single vane segment can be directly connected to the 1 x rev excitation.

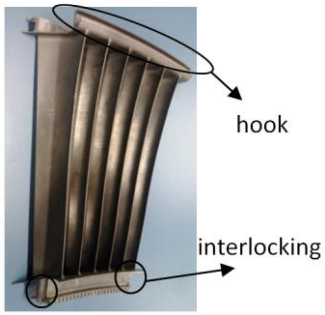



Figure 1: Vane-segment

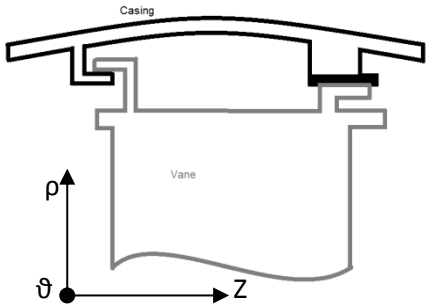



Figure 2: Hook joints

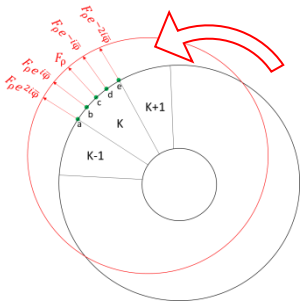



Figure 3: EO1 travelling excitation

It is important to study this phenomenon because it is predominant with respect to the aerodynamic forces due to the low engine order excitation.

For reason of confidentiality, in **Figure 4** is presented a sketch of an experimental Campbell diagram. Usually the values reported in the Campbell diagram are calculated after the measurement of strains obtained by means of a strain gauge placed on an airfoil of the vane segment. When vibrations are synchronous, i.e., when the vibration frequency is an integer number of times the rotation frequency of the rotor, the values of stress associated to the measured strain is produced by an engine order EO excitation, where EO is the integer number. A particular value of rotation speed exists where the natural frequency of a given mode shape matches the EO excitation. In this case, the second mode shape of the vane segment at Nodal Diameter ND = 1 is excited by an EO=1 excitation.

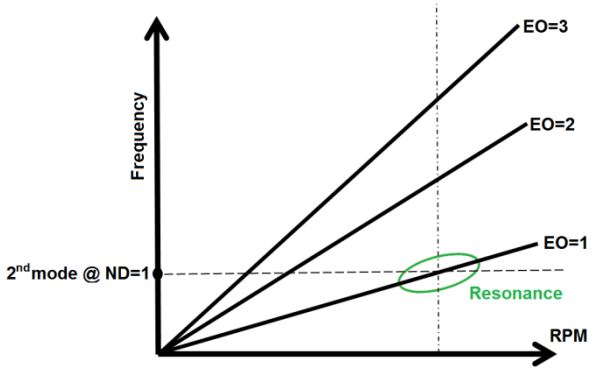



Figure 4: Campbell diagram

## 3. COMPONENT MODELING

The damping of vibrations through friction is a non-linear problem that is solved with iterative numerical methods. These methods require considerable computation time. Both the vane-segment and the casing have been modeled by FEM using a number of nodes so large that the resolution of the equations of motion in the time domain and in a direct form is not generally obtainable in an acceptable time. A reduction of the degrees of freedom (dofs) to solve is necessary in order to obtain a Reduced Order Model (ROM) without losing the critical information for the analysis. The reductions of the vane

segment and casing stiffness and mass matrices have been done using three methods: for the vane segment only the Craig-Bampton Component Mode Synthesis (CB-CMS) [9] was used, while for the casing the CB-CMS technique, the Cyclic Symmetry (CS) technique [9], and the Tran technique [10] were sequentially used in the aforementioned order. The ROM techniques are applied to only one vane-segment of the stator and to a fundamental sector of the casing associated with the vane segment by means of the hook joints.

### 3.1. Craig Bampton - Component Mode Synthesis (CB-CMS)

The CB-CMS reduction technique is implemented in most FE softwares (Ansys, Nastran) and it is based on the definition of a set of master $\{x_m\}$ and slave $\{x_s\}$ degrees of freedom (dofs) of the physical model plus an additional set of slave modal dofs $\{\eta_s\}$. The first set is the set that remains explicit in the ROM while the remaining part of dofs is omitted as a function of the first set according to the Guyan static reduction technique formulation [11] where the static deflection shapes of the slave dofs are written by imposing a unitary displacement of each master dofs one by one. With this process the slave nodes are forced to move according to the deformation of the master nodes. Since the Guyan reduction technique is not sufficiently representative of the dynamics of the system, a number of slave modal shapes is added to the static deflection base obtained from a modal analysis where all the master dofs are constrained simultaneously.

Both the casing fundamental sectors and the vane-segment are first reduced according to the following transformation for each component:

$$\{x_{tot}\} = \begin{Bmatrix} x_m \\ x_s \end{Bmatrix} = \begin{bmatrix} I & 0 \\ G & \Phi_s \end{bmatrix} \begin{Bmatrix} x_m \\ \eta_s \end{Bmatrix} = [\Psi_{CMS}] \begin{Bmatrix} x_m \\ \eta_s \end{Bmatrix}$$

where I is the identity matrix, G is the Guyan matrix, $\Phi_s$ is a subset of the slave modal matrix and $\eta_s$ is the vector of the modal slave coordinate.

For the vane-segment the master dofs are those belonging to the following nodes:

- 60 nodes on the hooks: 30 nodes on the front hook and 30 nodes on the rear hook arranged according to two bands of 15 nodes each;
- 20 nodes on the interlocking: 10 on the left interlocking and 10 on the right interlocking arranged according to two bands of 5 nodes each;
- 3 accessory nodes placed on one blade of the vane-segment.
- 10 Modal slave dofs

For the casing the master dofs are those belonging to the following nodes:

- 326 nodes for both the left and right interfaces. The interfaces are the cross sections obtained by selecting only the fundamental sector from the full casing;
- 60 nodes for the hooks mating the 60 nodes of the hooks of the vane;
- 16 nodes where the external force is applied along the circumferential direction of the casing simulating the EO1 unbalance mechanical force;
- 3 accessory nodes on the external face of the casing.
- 10 Modal slave dofs

**Figure 5** and **Errore. L'origine riferimento non è stata trovata.** show the position of the master dofs for both components.

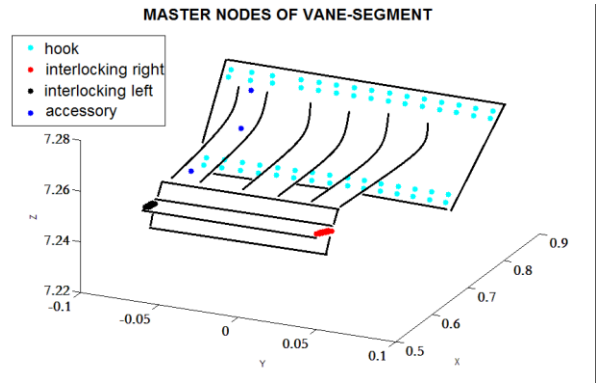



Figure 5: Master nodes on the vane

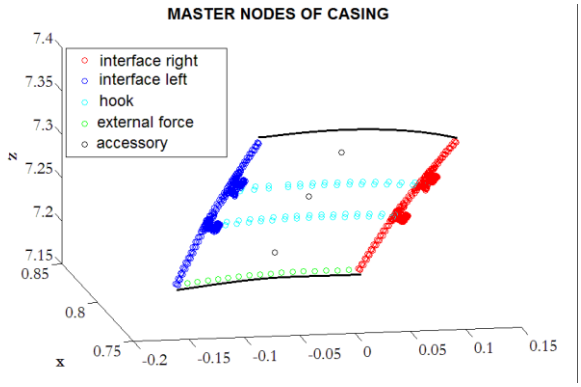



Figure 6: Master nodes on the casing

### 3.2. Cyclic symmetry

The casing is a cyclic symmetric component, therefore it shows the typical dynamic behavior and mode shapes of drums and rotors [9], in particular if a mode shape is excited by a travelling rotating force, each sector vibrates as the adjacent ones but with a phase lag given by the Inter Sector Phase Angle (ISPA) $\varphi = 2\pi ND/N$ with ND=Nodal Diameter and N=number of sector. The Nodal Diameter is a line passing from the center of the cyclic symmetric component where the modal displacements are null and its value is a characteristic of every mode shapes. A ND mode shape can be excited by an EO travelling force only if this relationship is verified:

$$EO = zN \pm ND, \quad z = 0,1,2,\dots$$

In this case we are interested in the response of the second resonance of the ND1 mode family excited by an EO1 travelling force (z=0).

Since the displacement of each sector is the same of the adjacent but with a phase lag, the full component can be reduced to a single sector by applying cyclic symmetry constraints to the left and right interfaces. The right interfaces dofs are written in function of the left interfaces dofs or vice-versa depending if the travelling force is clockwise or counterclockwise with respect to a defined cylindrical reference system. In this case for the casing:

$$\{x_l\} = \{x_r\}e^{i\varphi}$$

Where $x_l$ are the dofs of the left interface while $x_r$ are the dofs of the right interface of the casing. The set of dofs of the casing becomes then:

$$\begin{Bmatrix} x_m \\ \eta_s \end{Bmatrix} = \begin{Bmatrix} x_l \\ x_r \\ x_i \\ x_f \\ \eta_s \end{Bmatrix} = \begin{bmatrix} e^{i\varphi} & 0 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{Bmatrix} x_r \\ x_i \\ x_f \\ \eta_s \end{Bmatrix} = [\Psi_{CS}] \begin{Bmatrix} x_r \\ x_i \\ x_f \\ \eta_s \end{Bmatrix}.$$

where $x_f$ is the sub-vector of the dofs where the external force is applied and $x_i$ are the inner dofs (accessory and contact dofs).

### 3.3. Tran reduction technique

The Tran reduction technique [10] is applied on the dofs of the right interface because its number of dofs is still too large and consists in:

- Build an auxiliary ROM of the casing according to a Guyan static reduction technique where the master dofs are only the left and right interface dofs;
- Apply the cyclic symmetry reduction technique in order to write the left dofs with respect to the right dofs;
- Compute the right interface mode shape by means of a modal analysis.
- Of all the calculated mode shapes, substitute only few of them to the right dofs of the primary ROM model. in particular we considered 10 interface modal shape for the casing.

The array of the explicit dofs of the nodes becomes:

$$\begin{Bmatrix} x_r \\ x_i \\ x_f \\ \eta_s \end{Bmatrix} = \begin{bmatrix} \Phi_{cs} & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix} \begin{Bmatrix} \eta_{cs} \\ x_i \\ x_f \\ \eta_s \end{Bmatrix} = [\Psi_{TRAN}] \begin{Bmatrix} \eta_{cs} \\ x_i \\ x_f \\ \eta_s \end{Bmatrix}.$$

The whole transformation for the ROM of the casing is then:

$$\{x_{tot}\} = [\Psi_{CMS}][\Psi_{CS}][\Psi_{TRAN}] \begin{Bmatrix} \eta_{cs} \\ x_i \\ x_f \\ \eta_s \end{Bmatrix} = [\Psi_{TOT}] \begin{Bmatrix} \eta_{cs} \\ x_i \\ x_f \\ \eta_s \end{Bmatrix}$$

The stiffness and mass matrix of the two components can be obtained by pre-multiplication and post-multiplication of the transformation matrices:

$$[m] = [\Psi_{TOT}]^T [M][\Psi_{TOT}]$$
$$[k] = [\Psi_{TOT}]^T [K][\Psi_{TOT}]$$

The vane segment is kept in its position by means of a key/keyseat system close to the hooks that connect the vane segment to the casing. In order to model this kind of constraint, four contact nodes on the vane segment hooks close to the key are linked to the mating nodes of the casing hooks.

## 4. SOLUTION STRATEGY

In order to compute the forced response of a vane-segment with friction contacts the commercial finite element codes are not suitable since they are based on the time integration method of the non-linear differential balance equations which require very large calculations times. For this reason, ad hoc numerical codes must be developed in order to compute the forced response in the frequency domain. These codes are based on the Harmonic Balance Method (HBM) [12]. The periodic variables (displacements and forces) are written as a sum of harmonic terms by Fourier analysis and then the balance of each harmonic is imposed, turning the original nonlinear differential equations in a set of nonlinear algebraic equations.

The starting point in the forced response calculation of a mechanical system with friction contacts is writing the equations of motion in matrix form:

$$[M]\ddot{Q} + [C]\dot{Q} + [K]Q = F_e + F_{nl}, \tag{6}$$

where M, C and K are the mass, damping and stiffness matrices of the system, Q is the displacements array of degrees of freedom (dofs), $F_e$ is the periodical external forces array acting on the system and $F_{nl}$ is the non-linear array generated by friction contacts.

In order to reduce the calculation time of a numerical integration, the harmonic balance method (HBM) can be used to solve the equations of motion of the system [5,12,13, 14]. This is possible due to the periodicity of the external excitation that involves also the periodicity of the displacements *Q* and

the non-linear forces $F_{nl}$. This property of periodicity permits to express the displacements ($Q$) and the forces ($F_e$ and $F_{nl}$) in a truncated Fourier series of H orders, where H is the maximum number of harmonics considered. Consequently displacements and forces become:

$$Q = Q^{(0)} + \Re\left(\sum_{h=1}^{H} Q^{(h)} e^{ih\omega t}\right) \tag{2}$$

$$F_e = F_e^{(0)} + \Re\left(\sum_{h=1}^{H} F_e^{(h)} e^{ih\omega t}\right) \tag{3}$$

$$F_{nl} = F_{nl}^{(0)} + \Re\left(\sum_{h=1}^{H} F_{nl}^{(h)} e^{ih\omega t}\right) \tag{4}$$

where ω is the fundamental frequency of the excitation forces acting on the system, the 0th order represents the static components and the coefficients $Q^{(h)}$, $F_e^{(h)}$ e $F_{nl}^{(h)}$ are complex quantities. The Equations (2)-(4) are replaced into the equations of motion (1) obtaining two sets of algebraic equations, a set of static algebraic equations and a set of algebraic complex dynamic equations [14]:

$$[K_{stat}]Q^{(0)} = F_e^{(0)} + F_{nl}^{(0)} \tag{5}$$

$$\left[D^{(h)}\right]Q^{(h)} = F_e^{(h)} + F_{nl}^{(h)}, \qquad h = 1,2,\dots H \tag{6}$$

where $\left[D^{(h)}\right] = -(h\omega)^2[M] + ih\omega[C] + [K]$ is the hth dynamic stiffness matrix of the system. The equations (5) and (6) are non-linear because contact forces $F_{nl}^{(0)}$ and $F_{nl}^{(h)}$ depend on the relative static/dynamic displacements of the contact dofs that are a part of the total dofs $Q$.

In the classical approach, the static balance equations (5) are uncoupled from the dynamic balance equations, i.e. these equations are solved separately and then the 0th order coefficients of the displacements $Q^{(0)}$ are used as inputs for the solution of the dynamic equations. As a consequence, the value of the static contact forces, computed in the static balance equations is no more updated when the dynamic balance is imposed. In a more general approach, the static (5) and the dynamic (6) balance equations formulated by the HBM are solved together, i.e. the overall balance conditions (static and dynamic) are computed together.

This procedure will be referred as the '*coupled*' calculation of the forced response, where the term coupled means that the static and the dynamic quantities involved in the dynamic response of the vibrating system are updated at each iteration and may change for different excitation frequencies [14]. In order to separate the solution of the non-linear part of the system from its linear part it is convenient to rearrange the balance equations (5) and (6):

$$Q^{(h)} = \left[R^{(h)}\right]F_e^{(h)} + \left[R^{(h)}\right]F_{nl}^{(h)}, h = 0,\dots H \tag{7}$$

where $\left[R^{(h)}\right]$ is the receptance matrix (inverse of static or dynamic matrices: respectively $[K_{stat}]$ or $\left[D^{(h)}\right]$), $\left[R^{(h)}\right]F_e^{(h)} = Q_e^{(h)}$ is the linear response due to external excitation that does not change once the excitation frequency is fixed and $\left[R^{(h)}\right]F_{nl}^{(h)}$ is the non-linear response due to the contact forces that changes with the state of the contacts. Because the non-linear forces act only on the contact dofs, the vector of total displacements $Q^{(h)}$ can be split in two part: non-linear dofs $Q_{nl}^{(h)}$ and linear dofs $Q_l^{(h)}$; in this way the equation (7) becomes:

$$\begin{Bmatrix} Q_{nl}^{(h)} \\ Q_l^{(h)} \end{Bmatrix} = Q_e^{(h)} + \begin{bmatrix} R_{nl,nl}^{(h)} & R_{nl,l}^{(h)} \\ R_{l,nl}^{(h)} & R_{l,l}^{(h)} \end{bmatrix} \begin{Bmatrix} F_{nl}^{(h)} \\ 0 \end{Bmatrix}$$

where only the first equation $Q_{nl}^{(h)} = Q_e^{(h)} + \left[R_{nl,nl}^{(h)}\right] F_{nl}^{(h)}$ must be solved iteratively with a non-linear solver. Once the non-linear forces $F_{nl}^{(h)}$ are known it is possible to easily solve the second part of the equilibrium equations.

It must be observed that the *h* sets of equations $Q_{nl}^{(h)} = Q_{nl}^{(h)} + \left[R_{nl,nl}^{(h)}\right] F_{nl}^{(h)}$ are coupled to each other, because the number of harmonics that compose the nonlinear contact forces $F_{nl}^{(h)}$ depends on all the harmonics components of the displacements of the nonlinear dofs $Q_{nl}(t)$.

## 5. CONTACT MODEL

In order to solve the non-linear equations, described in the previous paragraph, it is necessary to introduce a contact model that computes the harmonic nodal contact force components $F_{nl}^{(h)}$ from a given set of harmonics contact displacements $Q_{nl}^{(0)}$ and $Q_{nl}^{(h)}$.

The contact model is based on the Coulomb friction law where the generic friction force $f_t(t)$ is equal to a normal preload $n(t)$ multiplied by the friction coefficient $\mu$ when sliding occurs:

$$f_t(t) = \mu n(t) \tag{8}$$

In this case the Coulomb friction law is implemented in a contact model [15] with a tangential $k_t$ and normal $k_n$ stiffness (**Figure 7**) in order to take into account a stick state of the contact that allows a relative motion between the two bodies in contact without sliding:

$$f_t(t) = f_0 + k_t(u(t) - u_0) \tag{9.a}$$

$$n(t) = k_n v(t) \tag{9.b}$$

(where $f_0$ and $u_0$ are respectively the value of the tangential force and tangential relative displacement at the end of the contact state preceeding the stick state) and the separation of the two parts in contact when $n(t) = k_n v(t)$ is less than zero, in that case:

$$\begin{cases} f_t(t) = 0 \\ n(t) = 0 \end{cases} \tag{10}$$

The couple static/dynamic approach, introduced to solve the balance equations of the system, involves that the tangential and normal relative displacements of the contact model include the correspondent static displacements, i.e.:

$$\begin{aligned} u(t) &= u_0 + \bar{u}(t) \\ v(t) &= v_0 + \bar{v}(t) \end{aligned}$$

In this way the normal pre-load $n_0 = k_n v_0$ of the contact model is no not more a design datum, but an information that comes by the iterative solution of the static and dynamic balance equation and it may change for different excitation frequencies. For this reason the normal relative static displacement $v_0$ (interference) is the necessary value to calculate the normal load of the contact model.

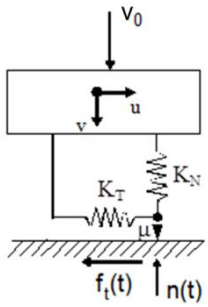



Figure 7: Contact model

The hysteresis loop can assume different shapes according to the relative phase lag of the relative tangential and normal displacements. In the **Figure** a, **Figure** b, **Figure** c three types of hysteresis loops are shown: the first when u(t) and v(t) are in-phase, the second when partial lift off occurs, the third when u(t) and v(t) do not vibrate in-phase.

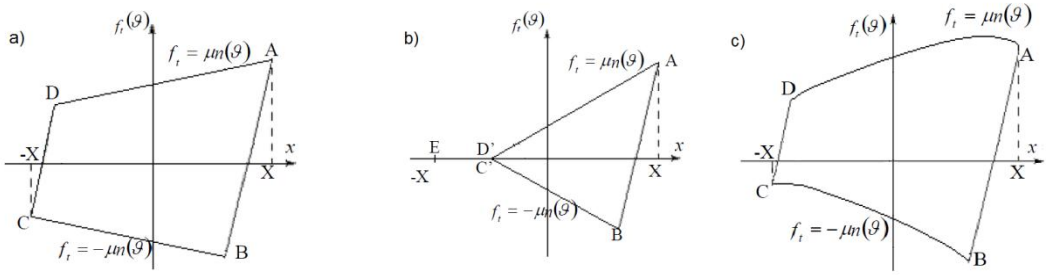



Figure 8: Hysteresis loops

The contact model is a 2-D model but it can be used to describe a 3-D behavior of the contact by using two orthogonal 2-D models that share the same normal relative static displacement $v_0$.

Since the equations of motion are formulated in the frequency domain the inputs of the contact model are the Fourier coefficients of the relative displacements $u^{(h)}$ and $v^{(h)}$ (with h=0,1,2,…,H) while the contact model operates in the time domain. For this reason for each iteration an Alternating Frequency Time (AFT) [15,16] also known as Hybrid Frequency Time (HFT) method [17] should be used in order to pass from the frequency domain to the time domain to calculate the hysteresis loop and then again to go back to the frequency domain by the Fourier series:

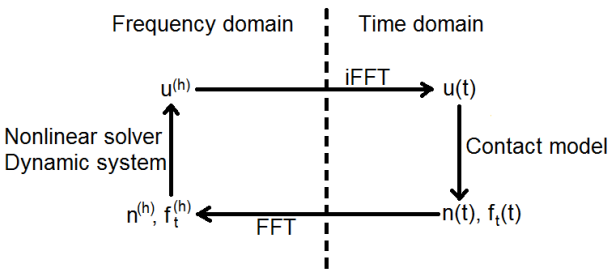



Figure 9: AFT/HFT method

The calculation procedure follows the following steps:

1. Periodical relative displacements *u(t)* and *v(t)* are computed from Fourier coefficients $u^{(h)}$ and $v^{(h)}$ (with h=0,1,2,…,H) by Inverse Fast Fourier Transform (IFFT);
2. Contact forces *n(t)* and $f_t(t)$ are computed by means of equations (9) and (10);
3. Fourier coefficients $n^{(h)}$ and $f_t^{(h)}$ (with h=0,1,2,…,H) are computed from *n(t)* and $f_t(t)$ respectively by Fast Fourier Transform (FFT).

With the Fourier coefficient of the contact forces it is possible to assemble the contact force vector $F_{nl}$ that will be used in the resolution of the non-linear equations $Q_{nl}^{(h)} = Q_{nl}^{(h)} + \left[R_{nl,nl}^{(h)}\right]F_{nl}^{(h)}$ (with h=0,1,2,…,H).

## 6. SENSITIVITY ANALYSIS, TEST SIMULATION AND COMPARISON WITH MEASUREMENTS

The ROM assembly is used to compute the nonlinear forced response of a vane-segment. A sensitivity analysis has been performed changing the values of some parameters in the calculation code, in details:

- Number of joints (presence and absence of the interlocking contact);
- Contact parameters (friction coefficient and interlocking interference);
- Number of harmonics retained in the MHBM calculation.

These analysis have been then compared with the corresponding analyses made with the uncoupled code to show the difference between the two methods.

### 6.1. Variation of number of joints

The first analyses are performed by varying the number of joints in contact: (i) only the hook joints working and (ii) both hook and interlocking joints in contact. For these two types of analyses, the joints have been first considered in fully stick condition because the analyses, in this way, are linear (the contact points are linked together through the contact stiffness) and the two obtained forced responses are the limit linear ‘free’ and ‘stick’ interlocking conditions that the system can encounter.

The **Figure 10** shows the forced response with the hook joints working in the fully stick condition while the interlocking joints is free. Curves are normalized according to the resonant frequency and peak amplitude of the second mode that is the mode of interest. The dynamic load given by the unbalance is a traveling force characterized by EO=1 and fixed to the value assigned by industry data recovery as well as the interference at the hook contacts. In this conditions the response of the system obtained from the non-linear analysis should be coincident with the free linear response obtained linking together vane and casing through the contact stiffness of the hook joints. This happens because the major contribution to the damping of the system is given by the interlocking joints, while the hook joints give a very low contribution (see [18]). In fact, the nonlinear calculation is actually linear because slip does not occur and this is the case where the static equilibrium of the system is independent from the dynamic equilibrium.

The forced response with both hook and interlocking joints working in the stick condition is presented in **Figure 11**. Curves are obtained by decreasing the value of the excitation amplitude at the casing in order to have no slip at the contacts. In this condition, it is possible to see the increase of the resonance frequency of interest due to the full sticking of all the joints (the structure is more rigid). On the same figure, it is shown the forced response obtained from the uncoupled method setting the same parameters used in the coupled code (value of contact stiffness, value of forcing and same frequency range) and the value of the pre-load on the interlocking contacts corresponding to the interlocking interference. The value of pre-load is obtained from a static non-linear analysis of the vane in which the equation (7) is solved at the zero-order to obtain the static displacements of the interlocking. These

values are then multiplied for the normal contact stiffness to get the pre-load on the interlocking. This process is necessary to compare the coupled method with the uncoupled method in order to verify differences in the results obtained from the coupled method. In **Figure 11** the two forced responses, couple and uncouple, are coincident and this is right because in the fully stick condition the static and dynamic calculations are not coupled.

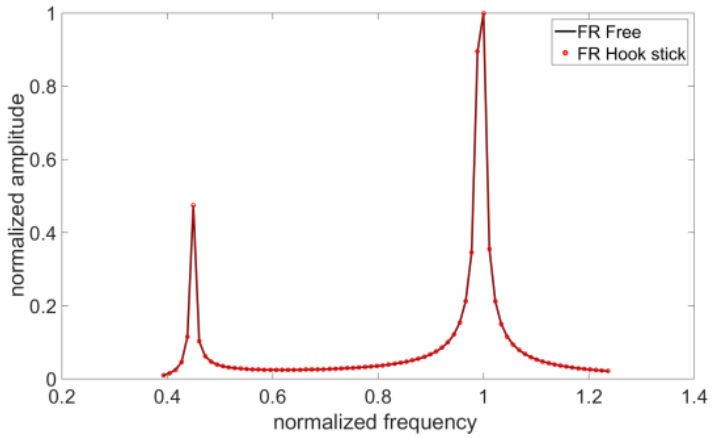



Figure 4: FR with hook joints in stick condition

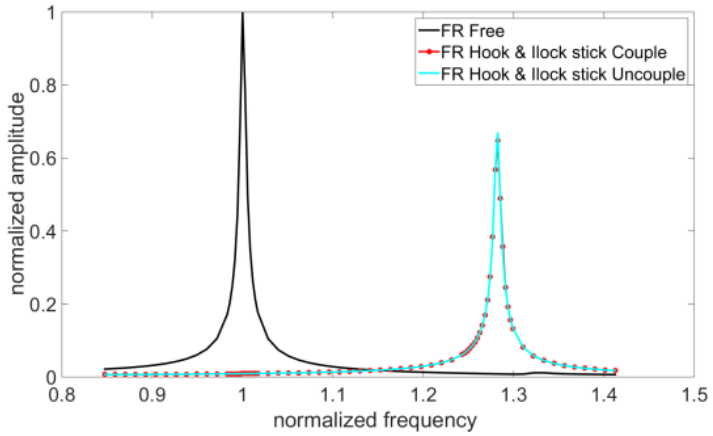



Figure 5: FR with hook & Ilock in stick condition

**Figure 12** shows the comparison between different nonlinear responses by varying the friction coefficient μ. These preliminary analyses have been made using only the first harmonic of the Fourier series fixing the design value of interference in the interlocking contact. It is possible to see that the nonlinear responses amplitudes are lower than the fully stick response of **Figure 11**, moreover the amplitude decreases if the friction coefficient increases. This effect is described in [8], basically when the friction coefficient value is low, friction forces are low as well and the dissipated energy is not effective to limit the vibration amplitudes because the contact tends to a 'frictionless' contact.

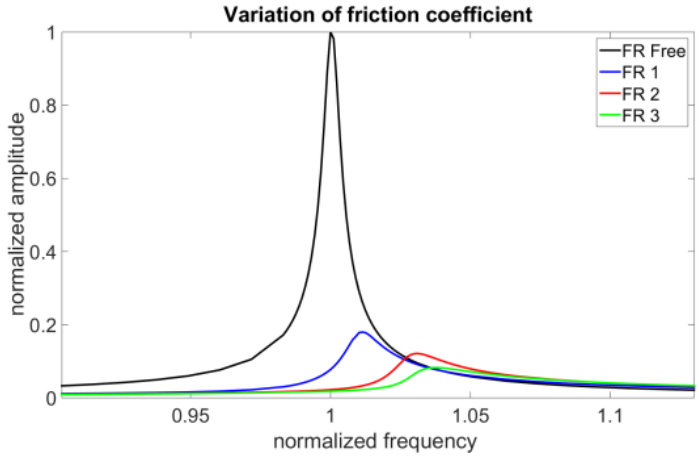



Figure 6: Variation of μ

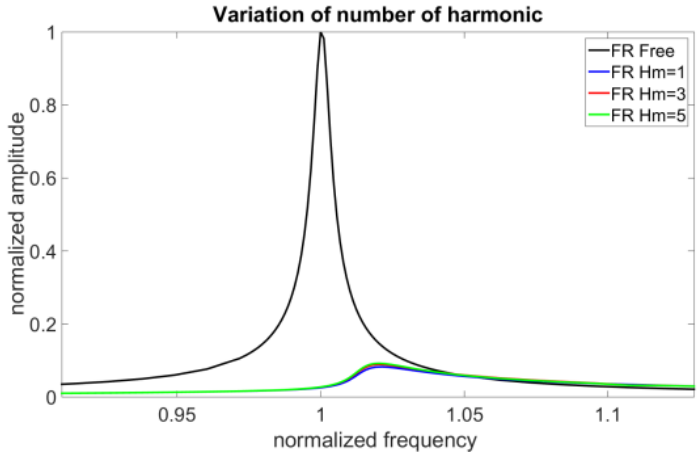



Figure 7: Variation of harmonics

We now take as a reference the value of friction coefficient of the response 3 (the value is not given for reasons of confidentiality). With this value, we have evaluated the effect of the variation of the number of harmonic taken into account during the resolution of the equations of motion. It is possible to see (**Figure 13**) a slight increase of the forced response when the number of harmonic is increased to three but no further increase is found if the number of harmonics is five. However, the increase of the number of harmonic do not produce a significant improvement of the approximation of the response. This means that is not necessary to add harmonics higher than one for a good approximation of the nonlinearities using the HBM produced by the friction contacts.

**Figure 14** shows the comparison between different nonlinear response by varying the value of the interlocking interference. Also these analyses have been made using only the first harmonic of the Fourier series (to contain the calculation time which increases fast with the number of harmonics taken into account) fixing the highest value of friction coefficient. Similarly to what happened in the case of variation of friction coefficient, the nonlinear responses amplitudes are lower than the fully stick response of **Figure 11,** moreover there is a decrease of the amplitude of resonance when the value of interference increases.

The variation of interlocking interference plays a similar role to the variation of friction coefficient, i.e. small interferences produce low friction forces and, as a consequence, the dissipated energy is not effective to limit the vibration amplitudes because the contact tends to a 'frictionless' contact.

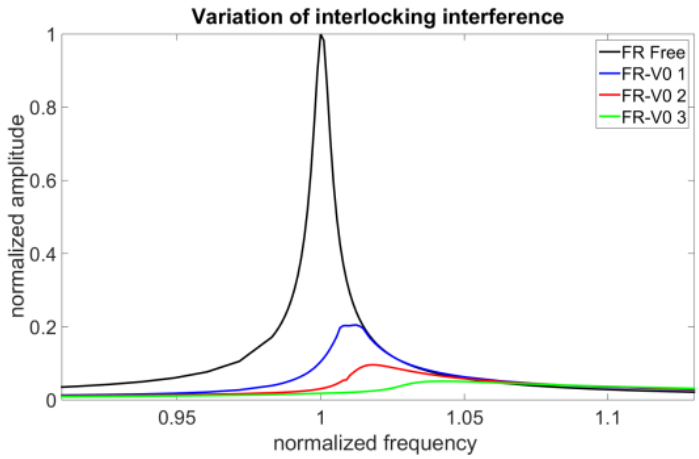



Figure 8: Variation of interference

### 6.2. Couple/uncouple comparison

In this section the difference between the two types of approaches are investigated. For this reason two conditions are analyzed: forced responses varying the friction coefficient (**Figure 15**) and forced responses varying the interlocking interference (**Figure 16**). The way to realize this comparison is explained in the section 6.1 where the response in the fully stick condition is shown for the two methods.

**Figure 15** shows the comparison of the forced response varying the friction coefficient between the coupled methods (continuous lines) and the uncoupled method (dashed lines). It is possible to see how the qualitative response of the two approaches is similar, i.e. the response has a similar form in the two methods varying the frequency of the exciting force. However, the uncoupled responses, fixed the contact parameters, show a shift of the resonance peak to higher frequency than the frequency found using the uncoupled method with a stronger decrease of the amplitude. This behavior is due to the fact that, in the uncouple method, the pre-load acting on the contact elements of the interlocking is set to a fixed value for the whole simulation and it is independent from the dynamic behavior of the system while in the coupled approach the static behavior is influenced by the dynamic one and vice-versa. **Figure 16** shows a similar result as the **Figure 15** and this is correct because the effect of the variation of the interference is the same of the variation of the friction coefficients (see **Figure 14**).

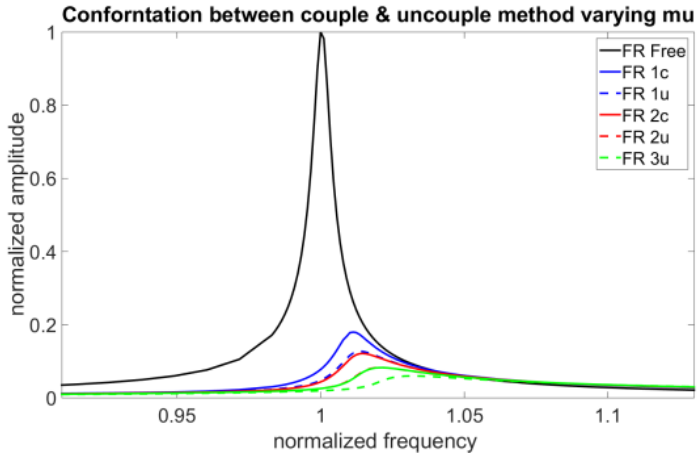



Figure 15: Couple/uncouple comparison varying μ

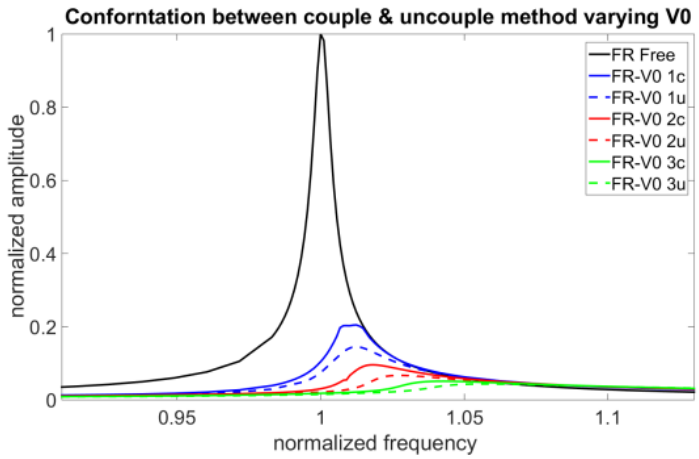



Figure 16: Couple/uncouple comparison varying interference

## 7. CONCLUSION

The calculation code for the study of the nonlinear dynamics of a vane-segment in the presence of friction damping is very complex because it introduces many numerical methods that try to reduce and simplify the system without reducing the level of quality of the approximation.

A new approach has been implemented to increase the level of approximation in the numerical simulations. To validate this new approach the simulated responses of the model have been compared with the responses obtained from an uncoupled method based on MHBM that is widely used to calculate the forced responses of the rotor and static blades in the aeronautic engines.

The comparison between the coupled and the uncoupled approaches shows that the coupled method results more conservative than the uncouple one, at the same time an increase of the calculation time has to be expected. However, if the velocity of calculation is not a priority, it would be better to use a more accurate model as the one that is realized with the coupled method.

## ACKNOWLEDGMENTS

The work described in this paper has been developed within the MIUR CLUSTER 'Greening the Propulsion' project funded by the Italian Government in collaboration with Avio Aero within the Politecnico di Torino-GE Avio Aero partnership 'Greatlab' (GReen Engine for the AirTransport LABoratory).

## REFERENCE

[1]Petrov E.P., Ewins D.J. (2004). "State-of-the-art dynamic analysis for non-linear gas turbine structures", *J Aerospace Eng. In: Proceedings of the IMechE*, Part G, vol 218, pp 199-211

[2]Ewins D.J. (2010). "Control of vibration and resonance in aero engines and rotating machinery – An overview", *International Journal of Pressure Vessels and Piping*, vol 87, pp 504-510

[3]Petrov E.P. & Ewins D.J. (2006). "Effects of damping and varying contact area at blade-disc joints in forced response analysis of bladed disk assemblies", *Journal of Turbomachinery*, Vol.128, No.2, (April 2006), pp. 403-410, ISSN 0889-504X

[4] Zucca, S. , Firrone, C.M. , Gola, M.M., "Numerical assessment of friction damping at turbine blade root joints by simultaneous calculation of the static and dynamic contact loads", Nonlinear Dynamics Volume 67, Issue 3, February 2012, Pages 1943-1955.

[5]Petrov E.P. & Ewins D.J. (2003). "Analytical formulation of friction interface elements for analysis of nonlinear multiharmonic vibrations of bladed discs", *Transactions of ASME Journal of Turbomachinery*, Vol.125, No.2, (April 2003), pp.364-371, ISSN 0889-504X
[6]Firrone C.M.; Zucca S. & Gola M.M. (2011). "The effect of underplatform dampers on the forced response of bladed disks by a coupled static/dynamic harmonic balance method", *International Journal of Non-Linear Mechanics*, Vol.46, No.2, (March 2011), pp.363-375, ISSN 0020-7462
[7] Firrone, C.M. "Measurement of the kinematics of two underplatform dampers with different geometry and comparison with numerical simulation" Journal of Sound and Vibration Volume 323, Issue 1-2, 5 June 2009, Pages 313-333
[8]Zucca S., Filippi S, Droetti F, Gola M.M, "Forced Response of Interlocked Vane Segments: Numerical Prediction and Experimental Results", *8th Biennal ASME Conference on Engineering Systems Design and Analysis*, Torino, Italy July 4-7, 2006, Vol. 3, pag 413- 420, ISBN: 0791842509, DOI: 10.1115/ESDA2006-95551
[9]Tran D.M.; "Reduced models of multi-stage cyclic structures using cyclic symmetry reduction and component mode synthesis", *Journal of Sound and Vibration*, Volume 333, Issue 21, 13 October 2014, Pages 5443–5463
[10]Tran D.M., "Component mode synthesis methods using interface modes. Application to structures with cyclic symmetry", *Computers and Structures*, Vol. 79, 2001, pp. 209-222
[11]Guyan R.J., "Reduction of stiffness and mass matrices", *AIAA Journal*, Vol. 3 (2), 1965, pp. 380.
[12]Cardona, A.; Lerusse, A. & Geradin, M. (1998). "Fast Fourier Nonlinear Vibration Analysis", *Computational Mechanics*, Vol.22, No.2, (August 1998), pp. 128-142, ISSN 0178-7675
[13]Griffin J.H. (1980). "Friction damping of resonant stresses in gas turbine engine airfoils", *Journal of Engineering for Power*, Vol.102, No.2, pp.329-333, ISSN 0022-0825
[14] Zucca S.; Firrone C.M. (2014). "Nonlinear dynamics of mechanical systems with friction contacts: Harmonic Balance Method and multiple solutions", *Journal of Sound and Vibration*, Volume 333, Issue 3, 3 February 2014, Pages 916-926
[15]Siewert C.; Panning L.; Wallaschek J. & Richter C. (2009). "Multi-harmonic forced response analysis of a turbine blading coupled by nonlinear contact forces", *Proceedings of ASME Turbo Expo 2009: Power for Land, Sea and Air*, June 8-12, 2009, Orlando, Florida, USA
[16]Cameron, T. M., Griffin, J. H. (1989). "An Alternating Frequency/Time Domain Method for Calculating the Steady-State Response of Nonlinear Systems", *Journal of Applied Mechanics*, Vol. 56, pp. 149-154, 1989
[17]Poudou, O., Pierre, C.: Hybrid (2003). "Frequency-Time Domain Methods for the Analysis of Complex Structural Systems with Friction Damping", *Collection of Technical Papers - AIAA / ASME / ASCE / AHS / ASC Structural Dynamics and Materials Conference*, Norfolk (USA), Vol. 1, pp. 111-124, 2003
[18]Lassalle M.; Firrone C.M. (2015). "Nonlinear forced response of stator vanes with multiple friction contacts", *AIAS Conference 2015*, Messina, Italy September 2-5, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPUTATIONAL METHOD OF DETERMINATION OF INTERNAL EFFORTS IN LINKS OF MECHANISMS AND ROBOT MANIPULATORS WITH STATICALLY DEFINABLE STRUCTURES CONSIDERING THE DISTRIBUTED DYNAMICALLY LOADINGS

**Zh.Zh. Baigunchekov[1], M.U. Utenov[2], N.M. Utenov[3], S.K. Zhilkibayeva[4]**

[1] Institute of Industrial Engineering, Kazakh National Research Technical University
Almaty, Republic of Kazakhstan
e-mail: bzh47@mail.ru

[2]Department of Mechanics, Kazakh National University
Almaty, Republic of Kazakhstan
e-mail: umu53@mail.ru

[3]Department of Mechanics, Kazakh National University
Almaty, Republic of Kazakhstan
e-mail: un.86@mail.ru

[4]Department of Mechanics, Kazakh National University
Almaty, Republic of Kazakhstan
e-mail: saltanatzhilkibayeva@gmail.com

**Keywords:** Manipulators, Mechanisms, Distributed Inertia Forces, Internal Loads, Kinematic Characteristic, Animation.

**Abstract.** *The technique of analytical determination of internal loads in links of planar rod mechanisms and manipulators with static definable structures taking into account the distributed dynamic stress, a self weight and the operating external loads is designed in this paper. The programs using the MAPLE are made on the given algorithm and animations of the motion of mechanisms with construction on links the intensity of cross and longitudinal distributed inertia loads, the bending moments, cross and longitudinal stress, depending on kinematic characteristics of links are obtained.*

## 1 INTRODUCTION

There are a variety of graph-analytical and numerical calculation methods on durability and rigidity of rod robotic systems and mechanisms, in which the distributed inertia forces of difficult character aren't considered [1-4]. The groups of Assur, that form the designed scheme of mechanism, can be statically definable, and also statically indefinable in concept of determination of internal stress. In this paper a new analytical approach of solution of problems of dynamic calculation on durability and rigidity taking into account the distributed dynamic stress in links of robotic systems and mechanisms with statically definable structures is proposed.

The distributed inertia forces of difficult character appear in links of rod mechanisms within the motion process. The intensity of distribution of inertia forces along the link depends on the mass distribution along the link and the kinematic characteristics of the mechanism changing rapidly. Rise of that sort of loads causes a set of problems, namely, breaking problems, which are specified by large-scale inertia forces; significance of elastic deformation of mechanism, that puts the mechanism out of action; because of deformation of links the mechanism can’t meet kinematic claims.

Therefore, relations between the intensity of distributed inertia forces and a self weight of links with geometrical, physical and kinematic characteristics are determined in our work. The laws of distribution of inertia forces and self weight allow to output laws of distribution of internal forces on the axis of link in each position of links, where there is a force attached to any point of a link. Their maximum values allow to optimize the design data of a link, which provides durability and rigidity of links and, entirely, of robotic systems and mechanisms.

As internal loads of each continual link are defined unambiguously by a set of internal loads in its separate cross-sections and by the matrixes of approximations, so the task is to calculate the internal loads in finite number of cross-sections of elements.

As a result, we refer to discrete model of elastic calculation of links of rod mechanisms. For elastic calculation of rod mechanisms based on Dalamber's principle, mechanisms are casted to structures which degree of freedom is equal to zero. For definition of internal loads in links of designed scheme of mechanism, the structure is divided into elements, both the hinged and rigid joints. The elements are divided into three types of beams for the first time. Discrete models of these three types of the beams with constant cross-sections which are under the action of cross and the longitudinal distributed loads of a trapezoidal view are constructed. The constructed discrete models for these three types of beams with constant cross sections along the axis allow to determine quantity of the independent dynamic equations of balance, components of a vector of forces in calculated cross-sections and to construct discrete model of all structure.

The dynamic equations of balance for discrete model of an element of the link with constant cross-sections which is under the influence of cross and longitudinal inertial loads of a trapezoidal look are also received in this work as well as the equations of balance of hinged and rigid knots expressed through required parameters of internal forces.

If we unite the equations of dynamic balance of elements and knots in one system, we will receive the equations of dynamic balance of all discrete model of system. A sort of systems of equations is sufficient for definition of internal forces in links of mechanisms, which structure is a static definable. The vector of forces and vector of loads in calculated cross-sections of discrete models of mechanisms are formed from vectors of forces and vectors of loads in calculated cross-sections of their separate elements. On the given algorithm the programs in the MAPLE system are made and animations of the motion of mechanisms with construction

on links the intensity of cross and longitudinal distributed inertia loads, the bending moments, cross and longitudinal forces, depending on kinematic characteristics of links are obtained.

## 2. INERTIA FORCES AND MATRIX OF APPROXIMATIONS

Considering the plane-parallel motion of an ***k***th link of mechanism with constant cross-sections comparatively fixed system of coordinates *OXY*, the following laws of distribution of the cross and longitudinal inertia forces along a link, that arise from self mass of a link are defined [5]:

$$\begin{aligned} q_k\left(x_k^{'}\right) &= a_{kq} + b_{kq}x_k^{'} \\ n_k\left(x_k^{'}\right) &= a_{kn} + b_{kn}x_k^{'} \end{aligned} \qquad (1)$$

where $a_{kq} = -\gamma_k A_k \cos\theta_k - \frac{\gamma_k A_k}{g} w_{kp}^{y_k^{'}}$, $b_{kq} = -\frac{\gamma_k A_k}{g}\varepsilon_k$, $a_{kn} = -\gamma_k A_k \sin\theta_k - \frac{\gamma_k A_k}{g} w_{kp}^{x_k^{'}}$, $b_{kn} = \frac{\gamma_k A_k}{g}\omega_k^2$, $\theta_k$ - an angle, which determines the position of the ***k***th link comparatively fixed system of coordinates *OXY*, respectively, $\omega_k$, $\varepsilon_k$ - angular velocity and angular acceleration of the ***k***th link, respectively, $w_{kp}^{x_k^{'}}$ и $w_{kp}^{y_k^{'}}$ - components of $P_k$ (pole) point acceleration of the ***k***th link put on the axis of link and perpendicular to it, respectively, $\gamma_k$ - specific weight of material of the ***k***th link, $A_k$ - square of cross-section of the ***k***th link, $g$ - acceleration of gravity.

The obtained expressions show that the distribution of cross and longitudinal inertia forces along the axis of link with constant cross-sections is characterized by trapezoidal law.

For the ***k***th link, which is under the influence of longitudinal trapezoidal distributed stress, Fig. 1, the bending moments along the length of element are distributed by the law of polynomial of third-degree.

$$M_k\left(x_k^{'}\right) = a_0 + a_1 x_k^{'} + a_2\left(x_k^{'}\right)^2 + a_3\left(x_k^{'}\right) \qquad (2)$$

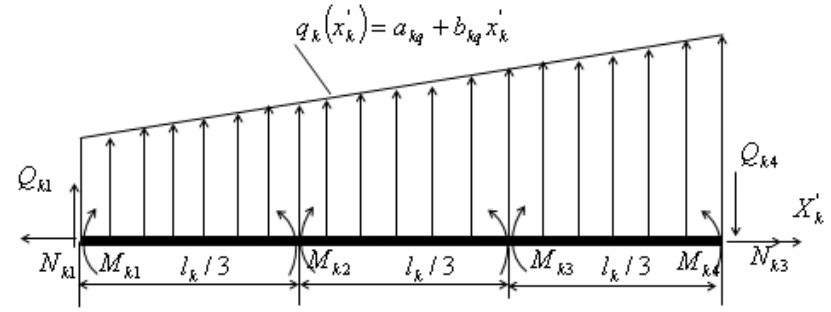



Figure 1: Longitudinal trapezoidal distributed load acting on the element

Now, let express the bending moments in $x_k^{'}$ cross-section through the sought bending moments $M_{k1}$, $M_{k2}$, $M_{k3}$, $M_{k4}$ in the cross-sections demonstrated in Fig. 1, respectively. For this purpose it is enough to express coefficients $a_0, a_1, a_2, a_3$ through $M_{k1}, M_{k2}, M_{k3}, M_{k4}$. As a result we have [6]:

$$M_k\left(x_k^{'}\right)=\left[1-\frac{11}{2l_k}x_k^{'}+\frac{9}{l_k^2}\left(x_k^{'}\right)^2-\frac{9}{2l_k^3}\left(x_k^{'}\right)^3\right]M_{k1}+\left[\frac{9}{l_k}x_k^{'}-\frac{45}{2l_k^2}\left(x_k^{'}\right)^2+\frac{27}{2l_k^3}\left(x_k^{'}\right)^3\right]M$$
$$+\left[-\frac{9}{2l_k}x_k^{'}+\frac{18}{l_k^2}\left(x_k^{'}\right)^2-\frac{27}{2l_k^3}\left(x_k^{'}\right)^3\right]M_{k3}+\left[\frac{1}{l_k}x_k^{'}-\frac{9}{2l_k^2}\left(x_k^{'}\right)^2+\frac{9}{2l_k^3}\left(x_k^{'}\right)^3\right]M_{k4k2} \quad (3)$$

Differentiating $M_k\left(x_k^{'}\right)$ to $x_k^{'}$ gives the equation of shear force:

$$Q_k\left(x_k^{'}\right)=\left[-\frac{11}{2l_k}+\frac{18}{l_k^2}x_k^{'}-\frac{27}{2l_k^3}\left(x_k^{'}\right)^2\right]M_{k1}+\left[\frac{9}{l_k}-\frac{45}{l_k^2}x_k^{'}+\frac{81}{2l_k^3}\left(x_k^{'}\right)^2\right]M_{k2}+$$
$$\left[-\frac{9}{2l_k}+\frac{36}{l_k^2}x_k^{'}-\frac{81}{2l_k^3}\left(x_k^{'}\right)^2\right]M_{k3}+\left[\frac{1}{l_k}-\frac{9}{l_k^2}x_k^{'}+\frac{27}{2l_k^3}\left(x_k^{'}\right)^2\right]M_{k4}. \quad (4)$$

Let the element be affected by the longitudinal trapezoidal distributed load, except the distributed shear force. In that case, the longitudinal force in arbitrary cross-section of an element can be expressed analogously to previous by means of longitudinal forces in calculated cross-sections as follows:

$$N_k\left(x_k^{'}\right)=\left[1-\frac{3}{l_k}x_k^{'}+\frac{2}{l_k^2}\left(x_k^{'}\right)^2\right]N_{k1}+\left[\frac{4}{l_k}x_k^{'}-\frac{4}{l_k^2}\left(x_k^{'}\right)^2\right]N_{k2}+\left[-\frac{1}{l_k}x_k^{'}+\frac{2}{l_k^2}\left(x_k^{'}\right)^2\right]N_{k3} \quad (5)$$

Thus, for the element which is acted by cross and the longitudinal trapezoidal distributed loads, the approximation matrix connecting internal loads in arbitrary cross-section of the element with values of internal loads in cross-sections has an appearance:

$$\left[H_k\left(x_k^{'}\right)\right]=\begin{bmatrix} h_{11}\left(x_k^{'}\right) & h_{12}\left(x_k^{'}\right) & h_{13}\left(x_k^{'}\right) & h_{14}\left(x_k^{'}\right) & 0 & 0 & 0 \\ h_{21}\left(x_k^{'}\right) & h_{22}\left(x_k^{'}\right) & h_{23}\left(x_k^{'}\right) & h_{24}\left(x_k^{'}\right) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & h_{35}\left(x_k^{'}\right) & h_{36}\left(x_k^{'}\right) & h_{37}\left(x_k^{'}\right) \end{bmatrix} \quad (6)$$

Elements of the first line of this matrix can be seen from the Eq. (3), elements of the second line can be seen from the Eq. (4), and elements of the third line can be seen from the Eq. (5), respectively.

The given expression of a matrix of approximation of loads defines dependence between a vector of forces $\left\{S_k\left(x_k^{'}\right)\right\}$ in arbitrary section of an $x_k^{'}$ element and a vector of forces $\left\{S_k\right\}$ in the appointed cross-sections. For an element of rod system the matrix of approximation is obtained accurately as it is solved on the basis of known laws of distribution of sought forces.

Note, the equations of the bending moment, the cross and longitudinal forces (3,4,5) respectively, which are expressed by the same values in calculated cross-sections, show that for definition of internal loads of each element of the mechanism it is enough to know values of these loads in final number of cross-sections of each of these elements. Number of sections in which it is necessary to know values of internal loads, are defined by polynomial degrees of external actions. Thus, internal loads of each continual link are determined unambiguously by a set of internal loads in its separate cross-sections and by the matrixes of approximations, therefore, the task is reduced to calculation of internal forces in final number of cross-sections of elements. Hence, we come to a discrete model of elastic calculation of links of rod mechanisms.

## 3. DISCRETE MODELS OF ELASTIC CALCULATION OF ELEMENTS AND MECHANISMS IN GENERAL

For elastic calculation of rod mechanisms based on Dalamber's principle, all inertial, external forces, gravity of links are attached and the unknown driving moments (forces) are applied to support assigned laws of their motion. If the ground hinges connecting drive link with rigid fixed-end are replaced, then the frames which joint is equal to zero are received.

For definition of internal stress in links (in elements) of calculated scheme of mechanism, the frame is divided into elements and joints. The link or its part can act as the elements, whereas the joints are ground hinges connecting links and cross-sections in the middle part, where concentrated external stress is occurred.

The process of frame sectioning consists of giving function and signs for element calculated section. While dividing the elements of calculated scheme of frame into calculated cross-sections and joints, it is necessary to set what internal relations between elements are remained or removed. If we reject any internal relations or their combinations in the element, so the element breaks up to two elements which can turn, move or be removed relatively each other. With the purpose to prevent it, internal forces-loads have to be applied at the joint rejecting places. Thereafter, these loads are regarded as primary unknowns.

Let's decompose an element of planar rod mechanisms on three types of beams, for convenience of working up the solving equations to determine the internal loads in the appointed cross-sections of elements of the mechanism [7].

Such beams can be the rods of basic linkage, if they are connected among themselves rigidly.

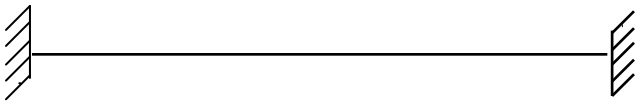


Figure 2: Beam's both ends are fixed rigidly (first type of a beam)

For determination of coefficients of expressions of the bending moment, it is necessary to know values of the bending moments in four cross-sections, and for determination of coefficients of expressions of longitudinal force, it is necessary to know values in three sections of an element. Therefore, we will choose four sections with unknown bending moments and three sections with unknown longitudinal forces in this beam. Then, by means of conditional schemes with the corresponding unknown, we will construct discrete model of the considered beam, Fig. 3.

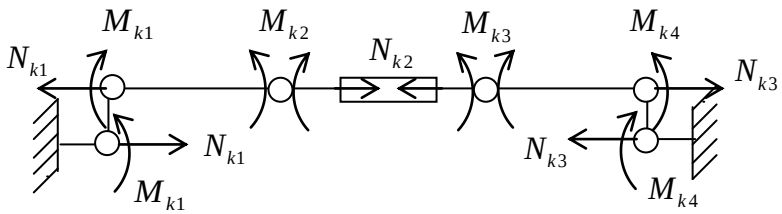



Figure 3: Discrete model of the first type beam under the action of the distributed trapezoidal load

Then the vector of forces in calculated cross-sections of the beam's discrete model is expressed by the following vector:

$$\{S_k\} = \{M_{k1}, M_{k2}, M_{k3}, M_{k4}, N_{k1}, N_{k2}, N_{k3}\}^T \quad (7)$$

There is dependence between degree of freedom of discrete model $m$, number of the attached external loads $n$ and degree of redundancy of calculated scheme $k$ [8]:

$$m = n - k \quad (8)$$

The matter is that total number of loads $n$ of calculated cross-sections is counted easily, and degree of redundancy of calculated scheme is obtained by formula $k = 3K - III$, where $K$ - number of the closed contours, $III$ - number of simple (single) joints, $k$ - degree of redundancy of calculated scheme of mechanism.

Degree of freedom of discrete model $m$ determines the quantity of necessary independent equations of statics.

Let's define the degree of freedom of discrete model of this beam. For this discrete model of beam the number of unknowns $n = 7$, the redundancy of beam $k = 3$, so the degree of freedom of discrete model $m = 4$. In other words, it is possible to work out four independent equilibrium equations for this discrete model of beam.

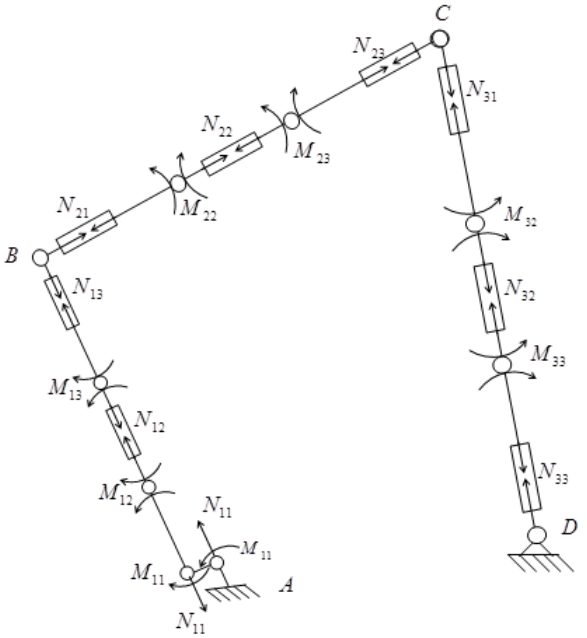



Figure 4: Discrete model of the four-link mechanism with constant cross-sections of links

The second type of an element is this beam, which one end is fixed rigidly and other end – joint-fixed. As an example, it can be drive links of planar rod mechanisms. The elements of the third type are beams of interlinks. They can be considered as the beams joint-fixed on the ends. The discrete models for beams of the second and third type are constructed similarly to the first type of beams.

The discrete model of the four-bar mechanism is constructed on Fig. 4; all sought values are shown here, these help to define all internal forces in any cross-section of rods of the mechanism.

For the first link (the second type of a beam) of this mechanism the vector of forces in cross-sections $\{S_1\}$ has the following components:

$$\{S_1\} = \{M_{11}, M_{12}, M_{13}, N_{11}, N_{12}, N_{13}\}^T \tag{9}$$

For the second and third links (the third beam type) regard to this mechanism, the vector of forces in calculated cross-sections have the following components, respectively:

$$\{S_2\} = \{M_{22}, M_{23}, N_{21}, N_{22}, N_{23}\}^T; \{S_3\} = \{M_{32}, M_{33}, N_{31}, N_{32}, N_{33}\}^T \tag{10}$$

For all discrete model of the mechanism the vector of forces in calculated cross-sections has an appearance:

$$\begin{aligned} &\{S\} = \{\{S_1\}, \{S_2\}, \{S_3\}\}^T = \\ &\{M_{11}, M_{12}, M_{13}, N_{11}, N_{12}, N_{13}, M_{22}, M_{23}, N_{21}, N_{22}, N_{23}, M_{32}, M_{33}, N_{31}, N_{32}, N_{33}\}^T \end{aligned} \tag{11}$$

## 4. DYNAMIC EQUATIONS OF EQUILIBRIUM OF DISCRETE MODELS OF ELEMENTS AND JOINTS

Let's remove the equations of dynamic equilibrium of an element. From the attached concentrated external loads $(Q_{k1}, M_{k1})$ and from the cross trapezoidal distributed loads on the axis of element, in arbitrary cross-section of $x_k^{'}$ element there is a bending moment determined by Eq. (2). On the other hand, the bending moment in cross-section of $x_k^{'}$ element, which is expressed through the sought moments in calculated cross-sections, is solved by Eq. (3).

If the Eq. (2) and Eq. (3) will be differentiated three times on $x_k^{'}$, then they will be equated and substituted to value $b_{kq}$, respectively, then the primary equation of dynamic equilibrium of element will be:

$$-\frac{27}{l_k^3} M_{k1} + \frac{81}{l_k^3} M_{k2} - \frac{81}{l_k^3} M_{k3} + \frac{27}{l_k^3} M_{k4} = -\frac{\gamma_k A_k}{g} \varepsilon \tag{12}$$

Relation between the values of the unknown quantities of bending moments in the calculated cross-sections and geometric, physical and kinematic characteristics of ***k***th element of mechanism is found. Thus, the second equation is expressed through relation of the sum of moments of all the acting forces on $k$ - element to center of gravity of $k4$ cross-section, Fig. 1:

$$Q_{k1} l_k + a_{kq} \frac{l_k^2}{2} + b_{kq} \frac{l_k^3}{6} + M_{k1} - M_{k4} = 0; \tag{13}$$

where $Q_{k1} = -\frac{11}{2l_k} M_{k1} + \frac{9}{l_k} M_{k2} - \frac{9}{2l_k} M_{k3} + \frac{1}{l_k} M_{k4}$, this equations is easy to get, if the value $x_k^{'} = 0$ is substituted into the Eq. (4).

Substituting the values $Q_{k1}$ and $a_{kq}$, $b_{kq}$ into Eq. (10), respectively, and summing the coefficients of unknowns of the same name, and also known quantities in the right end of the equation, the second equilibrium equation can be written as:

$$-\frac{9}{2} M_{k1} + 9 M_{k2} - \frac{9}{2} M_{k3} = \left( \gamma_k A_k \cos\theta_k + \frac{\gamma_k A_k}{g} w_{k1}^{y_k^{'}} \right) \frac{l_k^2}{2} + \frac{\gamma_k A_k}{g} \varepsilon_k \frac{l_k^3}{6} \tag{14}$$

From the longitudinal trapezoidal distributed loads acting on the element, as well as from the force $N_{k1}$ of the $k1$ cross-section, in the $x_k^{'}$ cross-section of element the longitudinal force is occurred, which can be solved by equation:

$$N_k\left(x_k^{'}\right)=N_{k1}-a_{kn}x_k^{'}-b_{kn}\frac{\left(x_k^{'}\right)^2}{2} \tag{15}$$

On the other hand, the longitudinal force in the $x_k^{'}$ cross-section of the element, expressed by means of longitudinal forces in the calculated cross-section, has the form (5).

Differentiating twice on $x_k^{'}$ the Eq. (12 and 5), respectively, equating them and substituting the value $b_{kn}$, the third equation of equilibrium can be expressed as:

$$\frac{4}{l_k^2}N_{k1}-\frac{8}{l_k^2}N_{k2}+\frac{4}{l_k^2}N_{k3}=-\frac{\gamma_k A_k}{g}\omega_k^2 \tag{16}$$

Projecting all forces acting on the ***k***th element on the $x_k^{'}$ axis and substituting the values $a_{kn}, b_{kn}$ the third equation of equilibrium is found. Thus

$$-N_{k1}+N_{k3}=\left(\gamma_k A_k\sin\theta_k+\frac{\gamma_k A_k}{g}w_{k1}^{x_k^{'}}\right)l_k-\frac{\gamma_k A_k}{g}\omega_k^2\frac{l_k^2}{2} \tag{17}$$

Obtained system of equation, which consist of Eq. (9), (11), (13) and (21) are assembled in a matrix form as:

$$\left[A_k\right]\left\{S_k\right\}=\left\{F_k\right\} \tag{18}$$

where

$$\left[A_k\right]=\begin{bmatrix} -\frac{27}{l_k^3} & \frac{81}{l_k^3} & -\frac{81}{l_k^3} & \frac{27}{l_k^3} & 0 & 0 & 0 \\ -\frac{9}{2} & 9 & -\frac{9}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{4}{l_k^2} & -\frac{8}{l_k^2} & \frac{4}{l_k^2} \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

$$\left\{S_k\right\}=\left\{M_{k1},M_{k2},M_{k3},M_{k4},N_{k1},N_{k2},N_{k3}\right\}^T$$

$$\left\{F_k\right\}=\left\{ b_{kq},\ -a_{kq}\frac{l_k^2}{2}-b_{kq}\frac{l_k^3}{6},\ -b_{kn},\ -a_{kn}l_k-b_{kn}\frac{l_k^2}{2}\right\}^T$$

Then, in the section of the element adjacent to the site (for kinematic pair) there are internal forces, as shown in Figure 5. For these units have two equilibrium conditions:

Let the two elements $j$ and $k$ of mechanism form a rotational kinematic pair, i.e. permit rotational motion relative to each other. Also let the length of these elements has a constant cross-section. Cut out of the mechanism a kinematic pair with surrounding cross-sections of the elements constituting this pair. Then, in the cross-section of the element adjacent to the

joint (to kinematic pair) there are internal loads, as shown in Fig. 5. There are two equilibrium conditions for these joints:

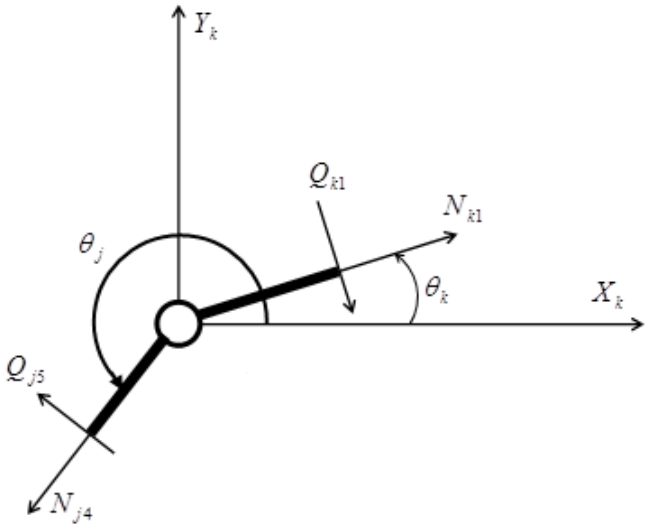



Figure 5: The hinge joint mechanism with constant cross-section elements

The equation of equilibrium for this joint can be described as:

$$\begin{cases} N_{k1}\cos\theta_k + Q_{k1}\sin\theta_k + N_{j3}\cos\theta_j + Q_{j4}\sin\theta_j = 0; \\ N_{k1}\sin\theta_k - Q_{k1}\cos\theta_k + N_{j3}\sin\theta_j - Q_{j4}\cos\theta_j = 0 \end{cases} \tag{19}$$

Further, the values will be expressed by means of sought moments in the calculated cross-sections of discrete model of the element, for this purpose we use the Eq. (4), substituting here the values $x_k^{'} = 0$ and $x_j^{'} = l_j$, respectively, hence:

$$Q_{k1} = -\frac{11}{2l_k}M_{k1} + \frac{9}{l_k}M_{k2} - \frac{9}{2l_k}M_{k3} + \frac{1}{l_k}M_{k4};$$
$$Q_{j4} = -\frac{1}{l_j}M_{j1} + \frac{9}{2l_j}M_{j2} - \frac{9}{l_j}M_{j3} + \frac{11}{2l_j}M_{j4}$$

Now, substituting the values $Q_{k1}$ and $Q_{j4}$ in the Eq. (19), the following equilibrium equations for joint have an appearance:

$$\begin{cases} -\frac{11\sin\theta_k}{2l_k}M_{k1} + \frac{9\sin\theta_k}{l_k}M_{k2} - \frac{9\sin\theta_k}{2l_k}M_{k3} + \frac{\sin\theta_k}{l_k}M_{k4} + \cos\theta_k N_{k1} - \\ -\frac{\sin\theta_j}{l_j}M_{j1} + \frac{9\sin\theta_j}{2l_j}M_{j2} - \frac{9\sin\theta_j}{l_j}M_{j3} + \frac{11\sin\theta_j}{2l_j}M_{j4} + \cos\theta_j N_{j3} = 0; \\ \frac{11\cos\theta_k}{2l_k}M_{k1} + \frac{9\cos\theta_k}{l_k}M_{k2} + \frac{9\cos\theta_k}{2l_k}M_{k3} - \frac{\cos\theta_k}{l_k}M_{k4} + \sin\theta_k N_{k1} + \\ +\frac{\cos\theta_j}{l_j}M_{j1} - \frac{9\cos\theta_j}{2l_j}M_{j2} + \frac{9\cos\theta_j}{l_j}M_{j3} - \frac{11\cos\theta_j}{2l_j}M_{j4} + \sin\theta_j N_{j3} = 0. \end{cases} \tag{20}$$

The linkage cross-sections can be as rigid joints, the concentrated external loads are attached here. For example, the concentrated forces $P_{kx_k'}$, $P_{ky_k'}$ and the concentrated moment $M_k$ are occurred in the $G$ cross-section of ***k***th link, Fig.5.

Then the $k$ link is divided into two elements, for $k$ th and $i$ th. If the cross-sections of elements are constant along the length of the link, then by means of cutting the $G$ joint out of mechanism, the scheme of $G$ joint with adjacent internal loads in cross-sections is displayed below. For this joint it is possible to write the three equilibrium equations that are expressed through sought parameters of elements.

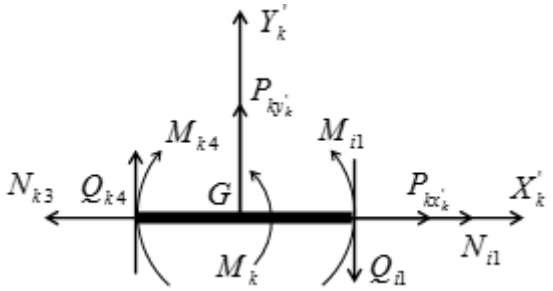



Figure 6: Rigid joints of a link with a constant cross-section of elements, where the external concentrated forces are attached

## 5. RESOLVING EQUATIONS OF DETERMINATION OF INTERNAL FORCES

By combining the equilibrium equations of elements and joints into a single system, the equilibrium equations of the discrete model of entire mechanism is obtained. They can be written in general form:

$$[A]\{S\} = \{F\} \tag{21}$$

Such systems of equations are sufficient to determine the internal forces in the links of the mechanism, which frame includes a statically definable group of Assur.

The matrix of equilibrium equations for the discrete model of mechanisms consists of matrices of equilibrium equations of their individual elements, as well as the equilibrium equations of their joints. The matrix of dynamic equilibrium equations of discrete models of mechanisms is as follows:

$$[A] = \begin{bmatrix} [A_1] & 0 & . & . & . & 0 \\ 0 & [A_2] & . & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & [A_n] \\ \textit{the equilburiu m equations} \end{bmatrix}$$

The vector of force and the vector of load in calculated cross-sections for discrete models of mechanisms is formed by vector of forces and loads in calculated cross-sections of their separate elements. These vectors in vector form, respectively, have the following species:

$$\{F\} = \{\{F_1\},\{F_2\},. . .,\{F_n\}\}^T; \{S\} = \{\{S_1\},\{S_2\},. . .,\{S_n\}\}^T$$

Now, for determination of internal loads in links, we give an example of six-bar second class mechanism with single drive linkage as shown in Fig. 7. The computer programs for

determination and construction of the inertia forces and internal loads on the links by means of using the MAPLE system are made. Therefore, the results of obtained inertia forces and internal loads for some positions of the mechanism are shown in Figs. 7-12.

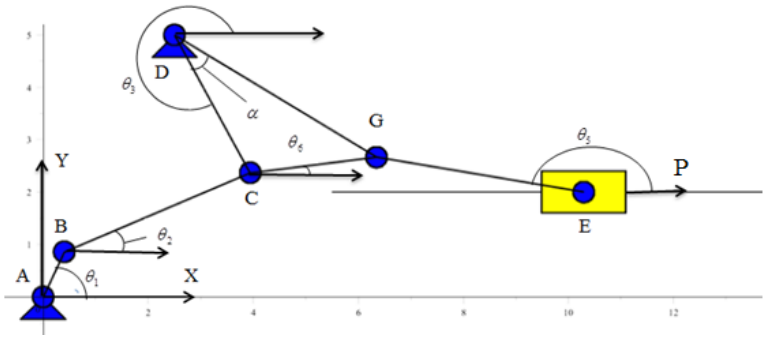



Figure 7: A six-bar second class mechanism with single drive linkage

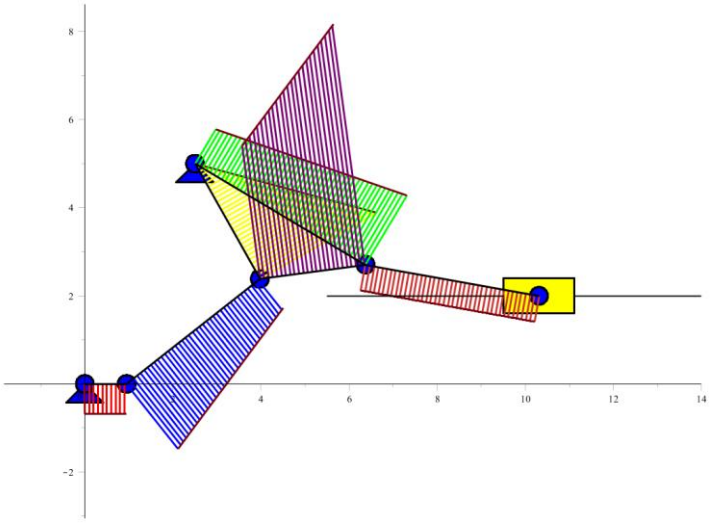


Figure 8: The investigating mechanism,
on which links the diagrams of cross inertia forces are constructed

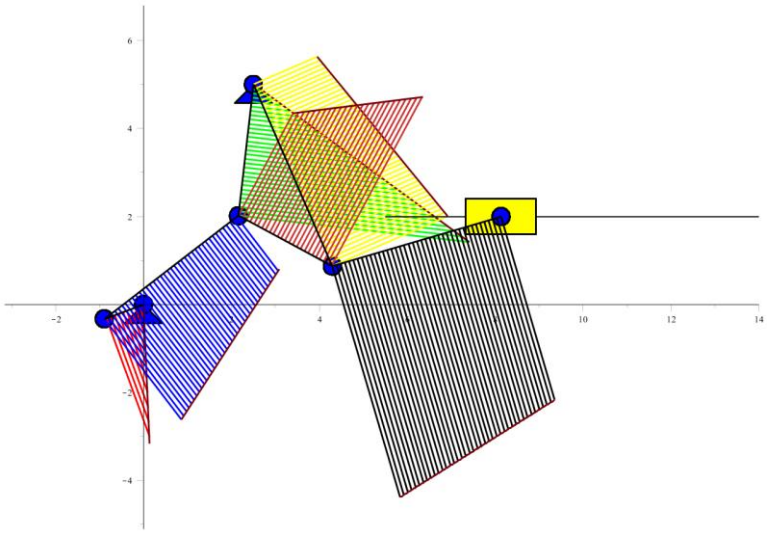


Figure 9: The investigating mechanism,
on which links the diagrams of longitudinal inertia forces are constructed

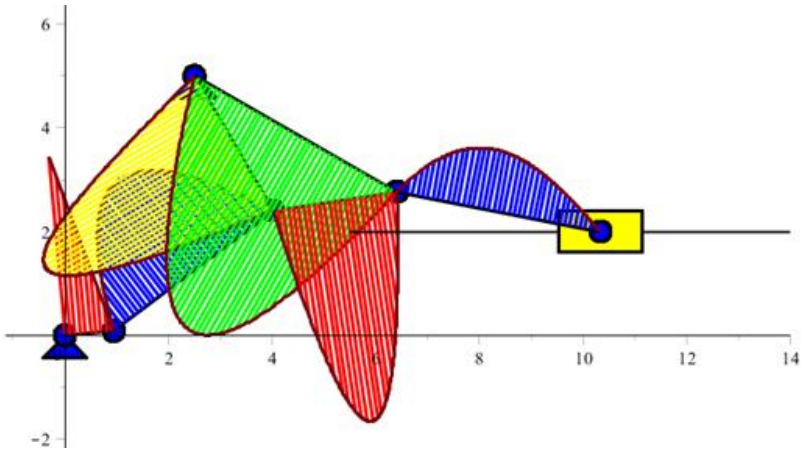



Figure 10: The investigating mechanism,
on which links the diagrams of bending moments are constructed

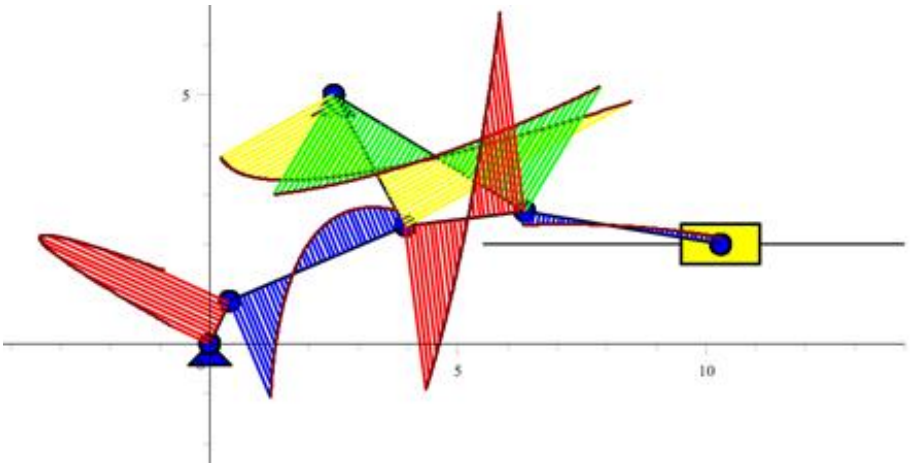



Figure 11: The investigating mechanism,
on which links the diagrams of shearing forces are constructed

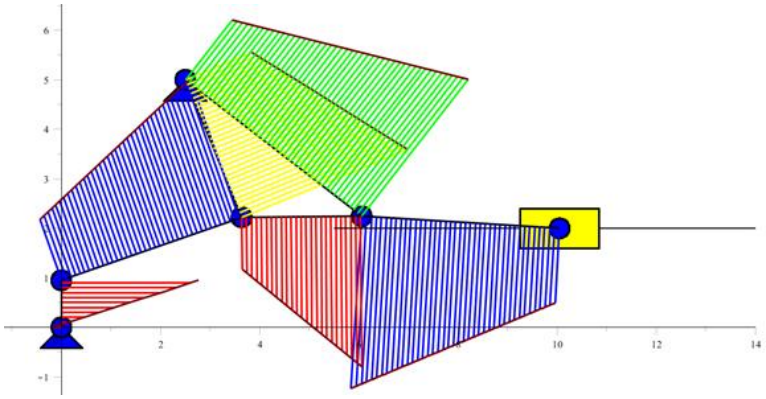



Figure 12: The investigating mechanism,
on which links the diagrams of longitudinal forces are constructed

## 6. CONCLUSIONS

The developed technique can be applied in the study of stress-strain state of the projected and existing mobile and fixed beam systems with statically definable structures (planar rod mechanisms, manipulators, frames, etc.).

## REFERENCES

[1] Du Zhaocai, Yu Yueqing, Yang Jianxin. *Analysis of the dynamic stress of planar flexible-links parallel robots*. Vol. 2, Issue 2, pp 152-158. Higher Education Press and Springer-Verlag, 2007.

[2] Santosha K.D., Peter E. *Dynamic analysis of flexible manipulators: A literature review*. Mechanism and Machine Theory, 2006, 41: 749-777.

[3] Gaultier P.E. *Modeling of flexible manipulator dynamics: A literature survey*. 1st nat. Applied Mechanics Conference, Cincinnati, 1989, 2C-3: 1-10.

[4] Kang B, Mills J.K. *Dynamic modeling of structurally-flexible planar parallel manipulator*. Robotica, 2002, v20: 329-339.

[5] Utenov M.U. *Investigations of forces, arising from the self mass of links with constant and variable cross-sections in their plane-parallel motion*. Proceedings of the First International Scientific and Practical Conference, Vol. 2, pp 30-34. Almaty, 2000.

[6] Utenov M.U. *The matrix of approximation of element force under the action of distributed load with parabolic intensity*. Proceedings of the First International Scientific and Practical Conference, Vol. 2, pp 55-58. Almaty, 2000.

[7] Utenov M.U. *Construction of discrete models of planar rod mechanisms in the elastic calculation.* Bulletin of the Kazakh Academy of Transport and Communications. №6 [12].-2001,- pp 61-64.

[8] Chiras A.A. Structural mechanics. – M.: Stroyizdat, 1989. – 255 p.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FE MODELLING OF SFRC BEAMS UNDER IMPACT LOADS

**Pegah Behinaein[1] , Ali A. Abbas[2], Demetris M. Cotsovos[3]**
[1] PhD Student, Institute of Infrastructure and Environment,
Heriot-Watt University, Edinburgh EH14 4AS, UK
e-mail: pb2@hw.ac.uk
[2] Senior Lecturer in Structural Engineering, School of Architecture,
Computing and Engineering, University of East London, London E16 2RD, UK
e-mail: abbas@uel.ac.uk
[3] Assistant Professor, Institute of Infrastructure and
Environment, School of the Built Environment, Heriot-Watt
University, Edinburgh EH14 4AS, UK
e-mail: d.cotsovos@hw.ac.uk

**Keywords:** Steel-fibre-reinforced concrete, finite elements, dynamic analysis, brittle cracking, impact.

**Abstract.** *The present work aims at investigating the structural behaviour of steel-fibre-reinforced concrete (SFRC) beams under high-rate loading conditions mainly associated with impact problems. A simple, yet practical non-linear finite-element analysis (NLFEA) model was used in the study. The model is mainly focused on realistically describing the fully brittle tensile behaviour of plain concrete as well as the contribution of steel fibres to the post-cracking response. The constitutive relations were incorporated into ABAQUS software brittle-cracking concrete model in order to adjust the latter to allow for the effects of fibres. Comparisons of the numerical predictions with their experimental counterparts demonstrated that the model employed herein, despite its simplicity, was capable of providing realistic predictions concerning the structural responses up to failure for different SFRC structural configurations. In the present study, the previous work is extended in order to numerically investigate the structural responses of simply-supported SFRC beams under impact loading. Data obtained from drop-weight tests on RC beams (without fibres) indicates that the response under impact loading differs significantly from that established during equivalent static testing. Essentially, there is (i) an increase in the maximum sustained load and (ii) a reduction in the portion of the beam span reacting to the impact load. However, there is considerable scatter making it difficult to ascertain the effect of loading rate on various aspects of RC structural response. To achieve this dynamic NLFEA is employed which is capable of realistically accounting for the characteristics of the problem at hand, a wave propagation problem within a highly non-linear medium. Following validation, a further study was conducted to assess the effect of steel fibers (provided at a dosage of Vf = 1%) on key aspects of structural response such as maximum sustained load, load-deflection curves, deformation profiles and ductility) under different rates and intensities of impact loading. The predictions reveal that steel fibers can potentially increase the maximum sustained load, ductility, toughness exhibited by SFRC members under impact loading compared to their RC counterparts.*

## 1 INTRODUCTION

The increasing demands for a variety of services and activities (i.e. faster transportation, higher energy production, safety, housing, etc.) has led to the increase in complex structural systems such as high-rise buildings, tunnels, bridges, slab-track for high-speed railways, off-shore and marine structures, storage and industrial facilities and nuclear power plants. Such structures are fully or partially constructed form Reinforced Concrete (RC) and their behaviour is required to be characterised by an elevated level of resilience in order to safely withstand loads imposed at higher rates and intensities than those considered in current design codes. This has also resulted in – and highlighted the need for – the development of new advanced construction materials, such as steel fibres.

It has been established numerically [1-4] and experimentally [5-6] that the dynamic structural responses of RC elements differ to those observed under corresponding static loading once certain thresholds of applied loading rate are surpassed. The shift in structural response with increasing loading rates is considered to be associated with (a) structural arrangements (such as geometry, reinforcement and boundary details), (b) the brittle nature and triaxiality characterising concrete material behavior, (c) the nature of the problem at hand (i.e. a wave propagation problem within a highly nonlinear material) and (d) the development of high strain rates within concrete and steel which is widely considered to affect material behaviour (i.e. strain-rate sensitivity) [7]. The current work aim was to examine the potential benefits of introducing steel fibres to the concrete mix in order to enhance the responses of (the otherwise RC) structural elements under high rate loading due to impact. The work is based on dynamic Nonlinear Finite Element Analysis (NLFEA), which was initially validated using exiting experimental data to ascertain its accuracy before the subsequent parametric studies were carried out. The work builds on previous NLFEA-based studies on various SFRC structural configurations subjected to both static monotonic and cyclic loading (the specimens covered a wide practical range from simply-supported beams to more complex structural systems characterised by a certain degree of static indeterminacy, such as continuous columns and beam-column sub-assemblages) [8-11]. Both previous and current studies utilize a material model for SFRC that is focused on realistically describing the fully brittle tensile behaviour of plain concrete as well as the contribution of steel fibres to the post-cracking response. The constitutive relations were incorporated into ABAQUS software brittle-cracking concrete model [12] in order to allow for the effects of fibres. The previous work (particularly under seismic loading) is extended herein in order to realistically describe the response – up to failure – of SFRC structural simply-supported beams under high-rate (i.e. impact) loading. The ultimate aim is to form the basis of a new practical assessment and design method for SFRC under impact.

The history of fibre-reinforced materials started in ancient Egypt over 2000 years ago with mud bricks reinforced with straw fibres. In more recent times, alternative fibres were introduced in the 1960s, such as steel, glass, and synthetic fibres such as polypropylene fibres etc. Early investigation into the application of steel fibre-reinforced concrete (SFRC) was carried out by Romualdi and Batson [13] resulted in its application in the pavement construction.

Since then, fibres have become widely adopted in different structural applications, for instance in concrete pipes, pavement slabs and more recently, it has been used in suspended slabs [14]. The introduction of steel fibres into the concrete mix can also lead to a reduction of transverse (i.e. shear) reinforcement such as beams, walls, joint regions [8-11, 15] without compromising design codes [16, 17] performance requirements for strength and ductility. The main benefit of such reduction is to alleviate reinforcement congestion due to dense arrangement and spacing of shear links in the critical regions of RC structures, as dictated in design codes such as the Eurrocode 8 for seismic-resistant design [17] in order to avoid brittle failure modes. The present research work focuses on the behaviour of steel fibres and other types are outside the scope of this study.

## 2 METHOD

### 2.1 Background of SFRC impact tests

To date, a number of drop weight tests have been conducted on SFRC specimens [18-20]. The majority of these studies employ drop weight testing, in which a steel striker is allowed to fall from a pre-defined height onto a specific (usually the mid-span) region of the structural element considered such as beams and slabs. A wide range of SFRC specimens have been tested to date under impact loading in order to assess the effect of various parameters associated with the fibres used (e.g. shape, aspect ratio, content) on the exhibited behaviour. The data obtained reveals that fibres capable of developing more adhesive and frictional bond as well as more effective anchorage at their ends can increase the maximum sustained load, ductility, toughness and crack controllability exhibited by SFRC members under impact loading compared to their counterparts exhibited by conventionally reinforced RC members [19]. However, it should be noted that the data is characterized by considerable scatter severely reducing its usefulness in accurately quantifying the above effects. The scatter predominantly reflects the difficulty in correlating the measured response to the actual physical state of the specimens. In fact, the measured maximum value of imposed load frequently corresponds to a specimen physical-state characterized by high concrete disintegration as well as low residual load-bearing capacity and stiffness [2, 21].This stage of structural response has little practical significance as it depends heavily on *post*-failure mechanisms for transferring the applied loads to the specimen supports. In view of the above, the available test data cannot provide detailed insight into the mechanisms underlying structural response. It can, however, provide a *qualitative* description of the effect of loading-rate on specimen behaviour.

### 2.2 Limitations of existing NLFEA software packages

NLFEA can be used in a safer and more cost-efficient manner for investigating a wider range of problems than those studied experimentally. It is capable of providing realistic predictions and more detailed insight into the mechanisms underlying SFRC structural responses under high-rate loading. However, as it usually employs dense three-dimensional (3D) finite element meshes combined with complex constitutive material laws implemented through the

use of iterative solution strategies, the required computational resources are high. As a result, its use is generally limited for the analysis of relatively simple structural forms. Moreover, its ability for providing realistic predictions of structural behaviour is, in most cases, linked with the use of case-study dependent constitutive models [22] often incorporating empirical amplification factors to account for the effect of strain-rate sensitivity on concrete *material* behaviour. This apparent lack of generality is the main drawback of NLFEA for RC structures, the results of which are often treated with some doubt regarding the validity of the predictions obtained.

### 2.3 General aspects of the FE model adopted

A finite-element (FE) model is generally considered capable of yielding realistic predictions of the nonlinear response of concrete structures when the deviation of the predicted values from their experimentally-measured counterparts (of structural characteristics considered) does not exceed a value in the order of 20% [22, 23]. Such structural characteristics usually include the load-bearing capacity, the relation between applied load and corresponding displacements, reactions and first-order deformation derivatives, i.e. rotations.

The present work is based on the use of the well-known commercial FE analysis software ABAQUS [12], which is capable of carrying out 3D static and dynamic NLFEA which incorporates a simple brittle model (termed "brittle cracking model") in order to describe concrete material behaviour. It is purpose-built for materials the behaviour of which is dominated by tensile cracking [12]. This is largely true in the case of RC flexural structural elements where cracks form due to the development of tensile strains within the concrete medium. Such cracks gradually extend with increasing levels of applied loading, ultimately leading to structural failure and collapse. This is particularly useful for the present study on SFRC performance as it allows for modelling the effect of steel fibers on the concrete behaviour in tension, especially after the *onset* of cracking. It is interesting to note that in the "brittle cracking model", the behaviour of concrete in compression is modelled as being linear elastic which safeguards the numerical stability of the solution process. By adopting this simplification, emphasis is focused on realistically describing the all-important tensile response. Furthermore, as the present NLFEA is carried out using 3D modelling, at least one principal stress is required to be tensile to trigger cracking (this is the salient feature of concrete behaviour which is predominantly brittle) and this is well captured by the brittle cracking model. So although the model assumes elastic behaviour in compression for efficiency purposes, this does not affect the accuracy as the predictions are in good agreement with experimental data (as discussed in the present study). So in summary, the brittle cracking model focuses on the important fundamental parameters affecting concrete behaviour (i.e. brittleness and cracking). This basic yet profound and focused approach allows one to develop a more fundamental understanding of aspects affecting the structural response of SFRC.

## 2.4 Material modelling

The introduction of fibres into the concrete mix can enhance both the concrete material behaviour (predominantly in tension) and RC structural response. More specifically fibres attribute ductile characteristics to concrete material behaviour resulting in an increase in strength, toughness and ductility. To date, many experimental investigations have been conducted on the effect of fibers on structural concrete material behaviour. The vast majority of these tests have been carried out on prisms and cylinders subjected to compression, direct and indirect tension and flexure. Such studies aim at assessing the effect of fibers on:

- important aspects of concrete material behaviour such as compressive strength $f_c$, tensile strength $f_t$, the elasticity modulus $E_c$, and the stress-strain curve under uniaxial compression or tension after cracking, and
- the cracking process that fibrous concrete undergoes, which is dependent on the fibre content, the strength and pull out behaviour of fibres bridging a crack as the crack becomes wider with increasing levels of applied loading.

The performance of SFRC is influenced by a number of parameters associated with the fibres such as shape, length, aspect ratio, and volume fraction [24, 25] and the concrete mix. The effect of fibres on concrete material behaviour in tension can be mainly observed after the onset of cracking as they attribute ductile characteristics to the post cracking behaviour of SFRC compared to the fully brittle behaviour exhibited by plain concrete [23, 26, 27]. Depending on the type and amount of fibres used, the post cracking behaviour may be described as strain-softening or strain-hardening. The residual strength of SFRC is the result of steel fibers bridging the cracks and the bond between the fiber and the surrounding concrete. The small use of fibers concrete mix results in softening post-cracking response and failure is caused by breaking fibers as they bridge the cracks. Using higher values of fiber-content results in a softening post-cracking behaviour, but exhibiting more ductile characteristic and in some cases with using higher fiber content, a strain-hardening response is observed as the fibers undertake the tensile forces which are acting normal to the plane of crack and will cause the increase in tensile strength.

## 2.5 Nonlinear solution strategy

During each time step the equation of motion governing the nonlinear dynamic impact problem considered is solved as a sequence of equivalent static problems through the use of the Newmark family of approximation methods. At the beginning of each iteration and based on the values of displacement, velocity and acceleration obtained from the previous iteration, the effective stiffness and load matrix are calculated and an equivalent static problem is formulated [27]. The equivalent static problem is solved through an iterative procedure based on a modified Newton-Raphson method [23]. During the solution process of the equivalent static problem every Gauss point is checked to determine whether loading or unloading takes place and to establish whether any cracks close or form. Depending on the results of the previous

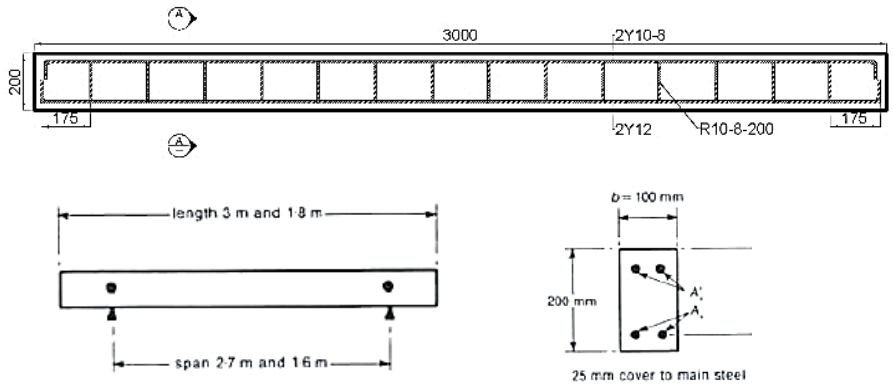
checks, changes are introduced to the stress-strain matrices of the individual FE's and to the stiffness matrix of the structure. Convergence is checked locally at each Gauss point; this involves the use of the constitutive relations for the calculation of the stresses increments which correspond to the estimated values of the strain increments. Once the values of the strain and the corresponding stress increments become less than a small predefined value (i.e. convergence criterion) then convergence is accomplished. When this is not achieved, the residual forces are calculated and are then re-imposed onto the FE model of the RC form investigated until convergence is finally achieved.

### 2.6 Modelling of cracking

The cracking process that concrete undergoes is modelled through the use of the smeared crack approach [23, 27, 28]. A crack is considered to form when the predicted stress in a given part of the structure corresponds to a point in the principal stress space that lies outside the surface defining the failure criterion for concrete, thus resulting in localised failure of the material. The plane of the crack is assumed normal to the direction in which the smallest principal stress acts (smallest compressive or largest tensile stress). A simple Rankine failure criterion is used to detect crack initiation (i.e. a crack forms when the maximum principal tensile stress exceeds the specified tensile strength of concrete). Constitutive calculations are performed independently at each integration point of the finite element model. The presence of cracks enters into these calculations by the way in which the cracks affect the stress and material stiffness associated with the integration point. After crack formation the residual shear stiffness along the plane of the crack is determined through the use of a *shear retention* factor. Its value is affected by the presence of the fibres bridging the two sides of the crack. The shear stiffness is considered to decrease as cracks widen. Therefore, in order to allow for degradation in shear stiffness due to crack propagation, the shear modulus is reduced linearly form full shear retention (i.e. no degradation) at the cracking strain to 50% of that at the ultimate tensile strain. It is worth noting that the shear retention does not diminish altogether due to the presence of the fibres which enhance dowel action as well as aggregate interlock by reducing crack opening. Crucially, the fibres contribute to shear resistance by providing tensile resistance (across the crack) to the shear induced *diagonal tension* stresses.

### 2.7 Structural forms investigated

The behaviour of the RC beam specimens considered herein has been experimentally investigated by Hughes and Spiers [29] under static and impact loading. The design details of the specimens (C2) investigated numerically in the present study are depicted in Figure1. The modulus of elasticity $E_S$, the yield stress $f_y$ and the ultimate strength $f_u$ of both the longitudinal and transverse reinforcement bars are 206 GPa, 460 MPa and 560 MPa, respectively. The uniaxial compressive strength $f_c$ of concrete is 45 MPa. The beams were subjected to drop weight testing at their mid-spans. Mild steel, rubber or ply pads were placed on the top face of the specimen in order to prevent or moderate local damage in the impact area and to some extent control the rate of loading.




| Beam type | Length | Tensile Steel $A_s$ | Compression Steel $A_{s'}$ | Stirrups |
|---|---|---|---|---|
| C2 | 3m | 2x12Φ | 2x6Φ | 14x6 |

Figure 1: Arrangements of RC beam investigated, adopted from Hughes and Spiers [29]

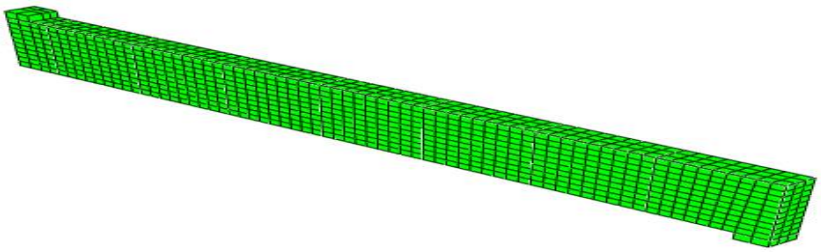


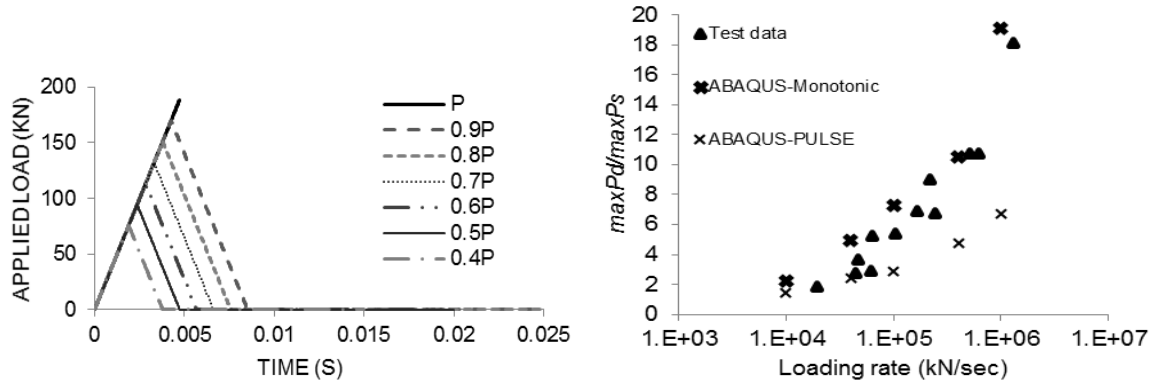
Figure 2: FE mesh adopted

## 3 RESULTS AND DISCUSSION

### 3.1 Numerical predictions

Initially, the case of RC beam specimen (C2) tested by Hughes and Spiers [29] was modelled using NLFEA under both static and dynamic (i.e. high-rate/impact) loading and subsequently a further study was carried out by adding fibres to this specimen in order to study SFRC under impact. Different loading rates were considered in the numerical study ranging from 10 to 1000 kN/msec, which were applied once monotonically and then in the form of a pulse load. The pulse load was applied in different stages, initially using the value of the maximum load sustained monotonically and consequently that load was reduced every time by 10% (of the maximum monotonic load sustained). The reduction continued in this manner to the point that the beam can bear the applied load, as shown in Figure 3(a). The aim was to establish the actual load-carrying capacity as the load applied during the experimental testing led to the complete destruction of the specimen and hence was not useful in determining the load-carrying capacity of the specimen. This is a common shortcoming of experimental testing under high rates of loading and limit its usefulness in terms of *design*, where a containment is needed up to a certain load threshold. To address this, the NLFEA-based study considered applying in the reduced pulse manner described in order to help pinpoint the actual failure point and establish the all-important load-carrying capacity value.

The variation of the dynamic increase factor (DIF) was considered as depicted in Figure 3(b), i.e. the maximum load sustained by the RC beams under high rate loading normalised with respect to its counterpart under static loading $maxP_d/maxP_s$, which shows a good correlation between the experimental and numerical results. Another indicator of the DIF was also considered, namely the ratio between the maximum displacement at mid-span under the dynamic load and the displacement under static load $Max\delta d/Max\delta s$. Both NLFEA-based $maxP_d/maxP_s$ and $Max\delta d/Max\delta s$ values at different loading rates as depicted in Figure 4(a,b), which shows that an increase in the loading rate leads to an increase in stiffness and load-carrying capacity combined with a decrease of the maximum deflection at mid-span. As regards the cracking and deformation profiles, Figures 5 and 6 indicate that, as the rate of loading increases, the length portion ($L_{eff}$) of the beam mostly affected by the applied load reduces. For relatively high rates of loading, this *effective* length is confined in the region of the beam mid-span extending on either side of the mid-span cross section to a distance marked by the formation of vertical *flexural* cracking initiating at the upper face and extending downwards, whereas the remainder of the beam (i.e. the portions extending between the supports and the aforementioned cracking) remains practically unaffected by the applied load. Therefore, under high rates of loading, the beam behaviour is essentially characterized by $L_{eff}$.




(a) (b)

Figure 3: (a) Applied load time history for monotonic and pulse load for RC and (b) comparison of experimental and numerical results for $maxP_d/maxP_s$ for different monotonic and pulse loading rates

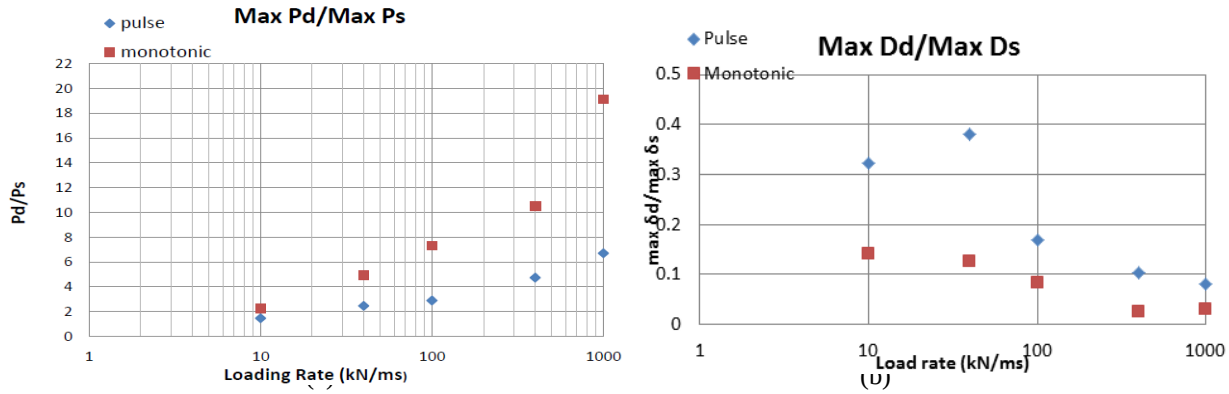



Figure 4: Comparison of NLFEA-based (a) $maxP_d/maxP_s$ and (b) $Max\delta d/Max\delta s$ for different monotonic and pulse loading rates for RC beam

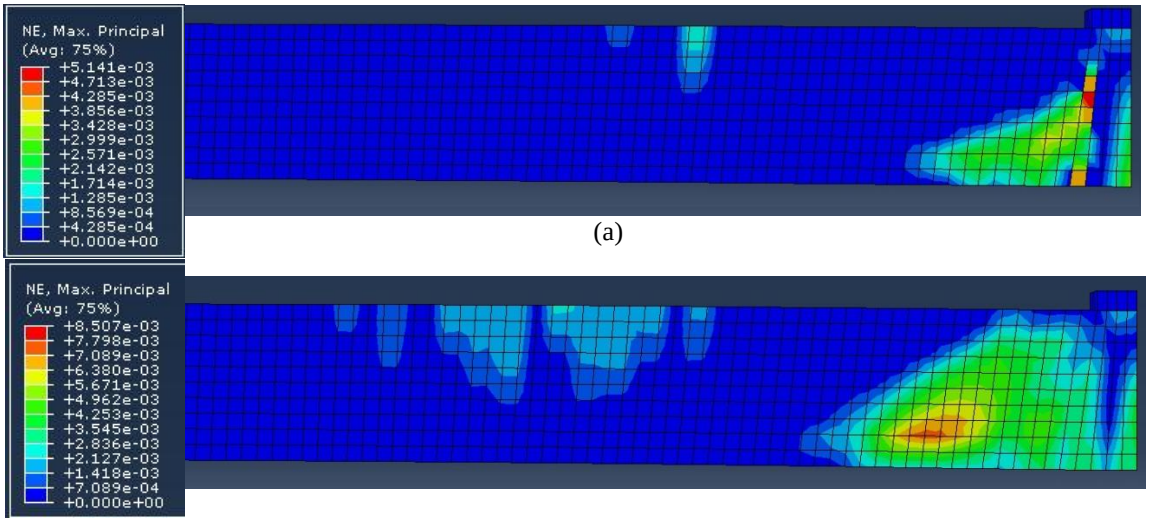



Figure 5: Strain distribution along the RC beam for loading rate 100kN/msec at (a) 75% and (b) 100% of load

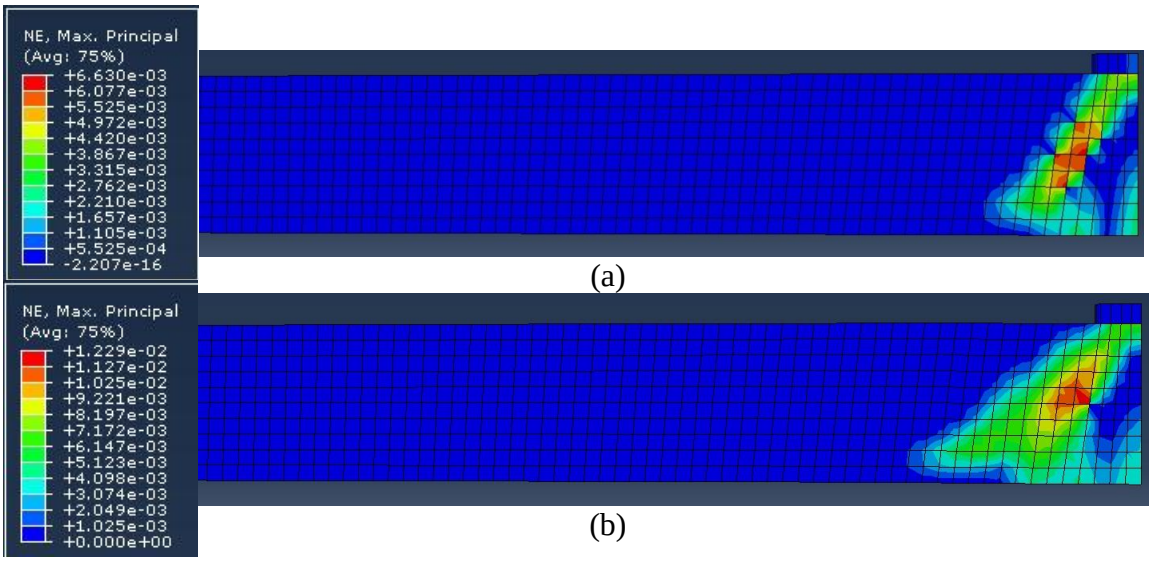



Figure 6: Strain distribution along the RC beam for loading rate 1000kN/msec at (a) 75% and (b) 100% of load

## 3.2 SFRC beam

Following the study of the RC beam (specimen C2), which also served to validate the FE model and strategy used, steel fibres were added with a volume fraction ($V_f$) value of 1%. The aim of this NLFEA-based study is to enhance understanding of the effect of steel fibers on structural behavior of RC elements under low and high rates of loading. The material constitutive model proposed by Lok and Xiao [30] was adopted in order to define the post-cracking tensile stress-strain relationship for SFRC. The load-deflection results under static loading are presented alongside the ones for RC beams (i.e. without fibres) in Figure 7. The curves are obtained based on incremental displacement acting on the mid-span of the beam. As it is shown, by adding 1% steel fibers to the concrete matrix, both the load carrying capacity and displacement in the mid-span of the beam increase. A comparison between the ratios of maximum dynamic to static load carrying capacity for different rates of loading ($maxP_d/maxP_s$) and maximum dynamic to static displacement ($MaxDd/MaxDs$), for both RC and SFRC are

shown in Figures 8(a,b), respectively. An increase in load carrying capacity and maximum deflection in the mid span of the beam is observed for SFRC compared to RC counterparts. Similar trends were found when the loading was applied as a pulse as can be seen in Figure 9(a,b) indicating the potential benefits of fibres in design terms as the load-bearing capacity increases. Figures 10(a,b) and 11(a,b) indicate that the fibres help widen the *localised* zone affected by the load ($L_{eff}$), albeit this still decreasing as the loading rate increases.

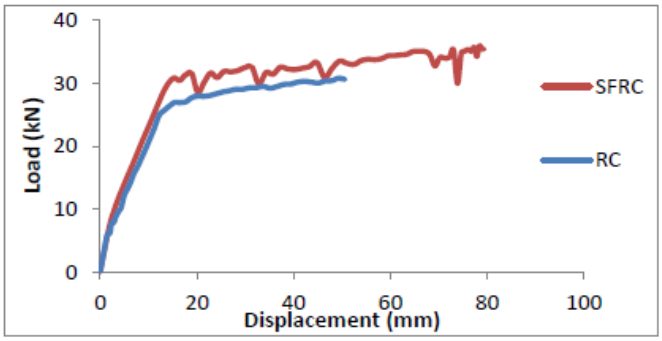
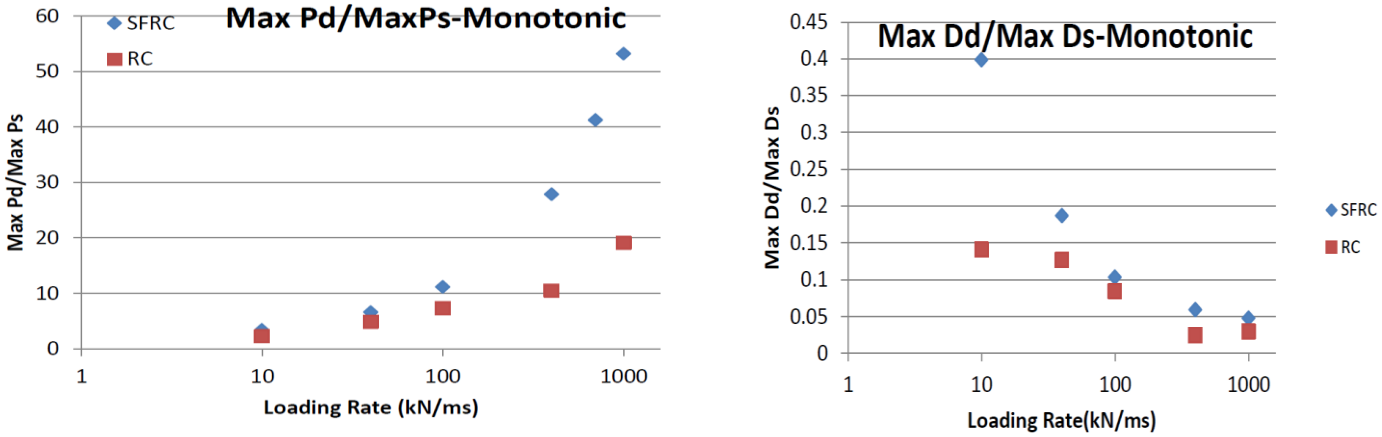
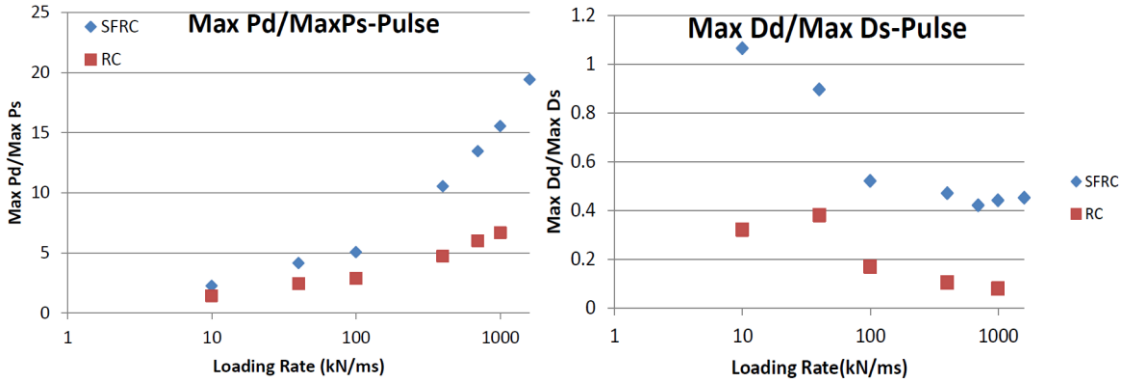



Figure 7: Comparison of static load-deflection in the middle of the beam for SFRC and RC



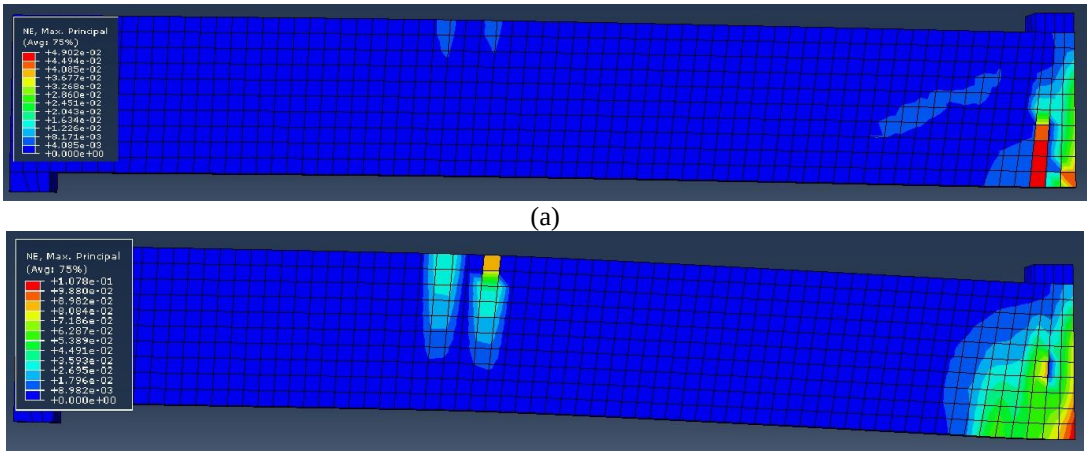

(a) (b)

Figure 8: (a) Comparison of MaxPd/MaxPs for RC and SFRC, and (b) Comparison of $Max\delta d/Max\delta s$ for RC and SFRC under monotonic load at different loading rates




(a) (b)

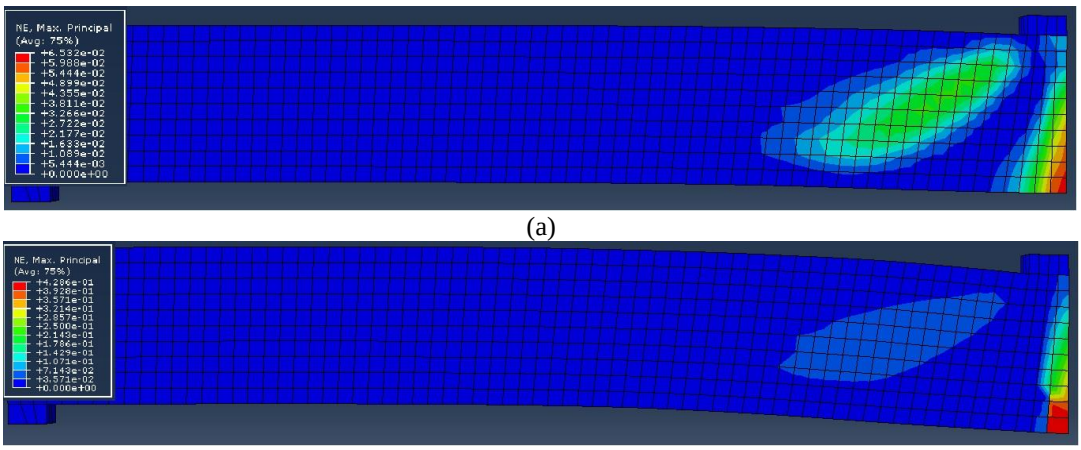
Figure 9: (a) Comparison of MaxPd/MaxP for RC and SFRC, and (b) Comparison of $Max\delta d/Max\delta s$ for RC and SFRC under pulse load at different loading rates




(a)

(b)

Figure 10: Strain distribution along SFRC beam for loading rate 100kN/msec at (a) 75% and (b) 100% of load




(a)

(b)

Figure 11: Strain distribution along SFRC beam for loading rate 1000kN/msec at (a) 75% and (b) 100% of load

## 4 CONCLUSIONS

A simplified, yet practical, brittle constitutive model was used by the NLFEA program adopted in the present study to analyse the behaviour of RC beams under impact loading applied at the mid span. For the beam case considered, the comparative study between numerical and experimental data revealed that such constitutive models are capable of providing realistic predictions of the experimentally established behaviour of RC beams under increasing rates of loading. The numerical results obtained using ABAQUS NLFEA software on RC beams under different loading rates show good agreement with previous numerical and experimental studies. A comparison of the load-deflection curves under static and impact loads shows that under high-rate loading conditions, the response of both RC and SFRC beams becomes stiffer and allows the beam to attain higher levels of loading. This change in structural response is essentially linked to the deformation profile exhibit by the beams with increasing loading rates which becomes more localised and confined to the mid-span loading region (i.e. the area at which the load is exerted) as the loading rate increases. This is likely to be due to the influence of inertia loads and the nature of stress waves invariably associated with impact which reduce the length of the beam that responds to the applied load. This leads to an in-

crease of the beam's stiffness and load-carrying capacity as well as to a reduction of its maximum vertical displacement under the load point as the loading rates increase. The study of RC and SFRC beam specimens shows that the use of fibers, at the dosage of $V_f$ = 1% studied, results in significant increase of load carrying capacity and energy absorption of the structural elements under impact. The deflection at the beam mid-span also increases compared to the RC case indicating the ductility provided by the steel fibres. The extent of the effective length around the point of application of the impact load seems to also improve (albeit still decreasing as the loading rate increases) due to the addition of fibres.

## REFERENCES


[1] D.M. Cotsovos, N.D. Stathopoulos, C. Zeris, Fundamental behaviour of RC beams subjected to high rates of concentrated loading. *ASCE Journal of Structural Engineering,* **134**, 1839-1851, 2008.

[2] D.M. Cotsovos, A simplified approach for assessing the load-carrying capacity of reinforced concrete beams under concentrated load applied at high rates. *International Journal of Impact Engineering*, **37**(8), 907–917, 2010.

[3] S. Saatci, F.J. Vecchio, Nonlinear finite element modelling of reinforced concrete structures under impact loads. *ACI Structural Journal*, **106**(5), 717-725, 2009.

[4] N. Kishi, S.G. Khasraghy, H. Kon-No, Numerical simulation of reinforced concrete beams under consecutive impact loading. *ACI Structural Journal*, **108**(4), 444-452, 2011.

[5] A.A. Abbas, A.D. Pullen, D.M. Cotsovos, Structural response of RC wide beams under low-rate and impact loading. *Magazine of Concrete Research*, **62**, 723-740, 2010.

[6] I.M. May, Y. Chen, D.R.J. Owen, Y.T. Feng, P.J. Thiele, Reinforced concrete beams under drop-weight impact loads. *Computers and Concrete*, **3**(2-3), 79-90, 2006.

[7] D.M. Cotsovos, M.N. Pavlovic, Modelling of RC Beams under Impact Loading. *ICE Proceedings, Structures & Buildings,***165**, 2, 77-94, 2012.

[8] A.A. Abbas, S. Syed Mohsin, D.M. Cotsovos, A.M. Ruiz-Teran, Seismic response of steel fibre reinforced concrete beam-column joints, *Engineering Structures*, **59**, 261-283, 2014.

[9] A.A. Abbas, S. Syed Mohsin, D.M. Cotsovos, A.M. Ruiz-Teran, Shear behaviour of SFRC simply-supported beams, *ICE Proc. Structures and Buildings*, **167(**SB9), 544-558, 2014.

[10] A.A. Abbas, S. Syed Mohsin, D.M. Cotsovos, A.M. Ruiz-Teran, Nonlinear analysis of statically-indeterminate SFRC columns, *Structural Concrete*, **15(**1), 94-105, 2014.

[11] A.A. Abbas, S. Syed Mohsin, D.M. Cotsovos, A.M. Ruiz-Teran, A M, Statically-indeterminate SFRC columns under cyclic loads, *Advances in Structural Engineering*, **17(**10), 1403-1417, 2014.

[12] ABAQUS Version 6.12-3 Documentation, Available online at http://www.3ds.com/products-services/simulia/products/abaqus (accessed 29/02/2016).

[13] J.P. Romualdi, G.B. Batson, Mechanics of crack arrest in concrete, *Journal of Engineering Mechanics*, **89**(EM3), 147-168, 1963.

[14] X. Destrée, Steel fibre reinforcement for suspended slabs. *Concrete*, **35**(8),58-59, 2001.

[15] M.P. Nadine, Healthy Doses of Steel Fiber 'Clear' Rebar Congestion in Concrete Coupling Beams. *ENR: Engineering News-Record*, **266**(2), 36, 2011.

[16] British Standard Institution, *Eurocode 2: Design of Concrete Structures - Part 1: General Rules and Rules for Buildings*. British Standard, DD ENV1992-1-1, 2004.

[17] British Standard Institution, *Eurocode 8: Design of Structures for Earthquake Resistance – Part 1: General Rules, Seismic Actions and Rules for Buildings*. Bristish Standard, BS EN 1998-1, 2004.

[18] A.E. Naaman, V.S. Gopalaratnam, Impact properties of steel fibre reinforced concrete in bending. *International journal of cement composites and lightweight concrete*, **5**(4), 225-233, 1983.

[19] Z. Xu, H. Hao, H.N. Li, Dynamic tensile behaviour of fibre reinforced concrete with spiral fibres. *Materials & Design*, **42**, 72-88, 2012.

[20] X.X. Zhang, A.M. Abd Elazim, G. Ruiz, R.C. Yu, Fracture behaviour of steel fibre-reinforced concrete at a wide range of loading rates. *International Journal of Impact Engineering*, **71**, 89-96, 2014.

[21] M.N. Pavlović, D.M. Cotsovos, Modelling of RC beams under impact loading. *Proceedings of the ICE - Structures and Buildings*, **165**(2), 77-94, 2012.

[22] M.D. Kotsovos, M.N. Pavlović, D.M. Cotsovos, Characteristic features of concrete behaviour: Implications for the development of an engineering finite-element tool, *Computers and Concrete*, **5**(3), 243-260, 2008.

[23] M.D. Kotsovos, M.N. Pavlović, *Structural concrete, Finite-element analysis for limit-state design*. Thomas Telford, London, UK, 1995.

[24] J.F. Trottier, N. Banthia, Toughness characterization of steel-fiber reinforced concrete. *Journal of materials in civil engineering*, **6**(2), 264-289, 1994.

[25] P.Robins, S.Austin, P. Jones, Pull-out behaviour of hooked steel fibres. *Materials and structures*, **35**(7), 434-442, 2002.

[26] P.M. Zisopoulos, M.D. Kotsovos, M.N. Pavlovic, Deformational behaviour of concrete specimens in uniaxial compression under different boundary conditions. *Cement and concrete research*, **30**(1), 153-159, 2000.

[27] M.D. Kotsovos, *Finite-element Modelling of Structural Concrete: Short-term Static and Dynamic Loading Conditions*. Boca Raton, CRC Press, 2015.

[28] O.C. Zienkiewicz, R.L., Taylor, *Vol. 2: The Finite Element Method for Solid and Structural Mechanics, 6th Edition*. Butterworth-Heinemann, Oxford, UK, 2005.

[29] G. Hughes, D.M. Spiers, An Investigation on the Beam Impact Problem. *Cement and Concrete Association,* Technical Report 546, 1982.

[30] T.S. Lok, J.R. Xiao, Flexural strength assessment of steel fiber reinforced concrete. *Journal of Materials in Civil engineering*, **11**(3), 188-196, 1999.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN APROXIMATE ANALYTICAL SOLUTION FOR NONLINEAR FGM SHELL STRUCTURE WITH VARIABLE IN TIME PARAMETERS

**V. Z. Gristchak[1], Yu. A. Fatieieva**[2]

Department of Applied Mathematics and Mechanics, Zaporizhzhe National University Zhukovskogo str., 66, Zaporizhzhe, 69063, Ukraine

grk@znu.edu.ua[1], fateevajulia@gmail.com[2]

**Keywords:** Functionally Gradient Material**,** Hybrid Asymptotic Approach, Approximate Analytical Solution, Time Dependent Parameters.

**Abstract.** *This paper deals with the problem of dynamic behavior of thin geometrically imperfect shell structures made of functionally gradient material (FGM) with time dependent parameters. Hybrid asymptotic approach is used to obtain an approximate analytical solution of the problem. The material properties are graded in the thickness direction according to the given law distribution and initial conditions. The non-linear strain-displacement relationships based upon the von Karman theory for moderately large normal deflections. Discussed problem leads to a singular non-linear second order non homogeneous differential equation with variable in time coefficients. An analytical solution by hybrid perturbation-WKB-Galerkin (P-WKB-G) method in some parameters of structure is compared with direct numerical integration results of initial equation of the problem.*

## 1 INTRODUCTION

Thin walled structures made of functionally graded materials (FGM) with metal inner surface and ceramic in outer surface widely used, for example, in modern air-space systems. In recent years important studies have been researched about vibration and stability FGM plates and shells with using mostly by numerical approaches [1-4, 12].

The present research devoted to an approximate analytical solution of nonlinear dynamic problem of FGM imperfect shallow shells with time dependent parameters (for example, with thickness depending on time) on the basis of hybrid (P-WKB-G) asymptotic method, which was successfully applied in some mechanical problems [5-11,15].

## 2 BASIC CONCEPT OF THE HYBRID APPROACH TO SOLUTION OF NONLINEAR PROBLEMS

To solve the non-linear differential equations with variable coefficients the approach is applied in three stages. On the first step the solution is determined using perturbation method by forming an expansion in parameter near the non-linear term of initial equation and we obtain the related system of linear non-homogeneous equations with variable coefficients. On the second step the solutions of linear system are determined using the (one or two-term) WKB-approximation by forming an expansion in parameter. On the third step the correction functions are obtained by classical Galerkin procedure.

In this paper an approximate analytical solution of nonlinear dynamic behavior of shallow shells with time-dependent thickness, which is described by singular nonlinear non homogeneous equation with variable coefficients was found.

## 3 THE NONLINEAR DYNAMIC BEHAVIOR OF IMPERFECT SHALLOW SHELLS WITH VARIABLE IN TIME THICKNESS

An approximate analytical non-linear analysis is given on the basic system of equations in terms of the stress and deflection following to the paper [1]. Suppose the FGM shallow shell is simply supported at its edges and subjected to a transverse load $q_0(t)$ and compressive edge loads $r_0(t)$, $p_0(t)$. We assume that modulus of elasticity and the mass density changes in the thickness direction, while the Poisson ratio is assumed to be constant and thickness of shell is function of time.

Appling Bubnov-Galerkin procedure with assumption that initial imperfection of middle surface of shell has the form:

$$w_0(x_1, x_2) = f_0 \sin\frac{m\pi x_1}{a} \sin\frac{n\pi x_2}{b} \tag{1}$$

where $f_0$ is given amplitude, the non-linear second – order ordinary differential equation for function $f(t)$ with deflection function $w = (x_1, x_2, t)$:

$$w(x_1, x_2, t) = f(t) \sin\frac{m\pi x_1}{a} \sin\frac{n\pi x_2}{b} \tag{2}$$

that are correspond to simply support boundary conditions, is given in the form [1]:

$$\varepsilon^2 \frac{d^2 f}{dt^2} + f\left(1 + 2 f_0 \bar{A}_2(t) - \bar{A}_3(t) f_0^2 - \bar{A}_1(t)\right) + f^2\left(-3\bar{A}_2(t)\right) + f^3 \bar{A}_3(t) = Q_0 - \bar{A}_0(t) + f_0 - \bar{A}_2(t) f_0^2 \quad (3)$$

where

$$\varepsilon^2 = \frac{1}{\omega_{mn}^2},\ \omega_{mn}^2 = \frac{1}{\rho_1(t)}\left[\frac{\left(E_1 E_3 - E_2^2\right)}{E_1\left(1-\nu^2\right)} \cdot \frac{\left(m^2 + n^2\lambda^2\right)\pi^2}{a^4} + \frac{E_1\left(k_1 n^2\lambda^2 + k_2 m^2\right)^2}{\left(m^2 + n^2\lambda^2\right)^2}\right],\ \bar{A}_i = \frac{A_i}{\omega_{mn}^2}$$

$$A_1(t) = \frac{\pi^2 h(t)}{a^2}\left(m^2 r_0 + n^2\lambda^2 p_0\right),\ A_2(t) = \frac{16 E_1(t) mn\lambda^2\left(k_1 n^2\lambda^2 + k_2 m^2\right)}{3a^2\left(m^2 + n^2\lambda^2\right)^2},\ A_3(t) = \frac{512 E_1(t) m^2 n^2\lambda^4}{9a^4\left(m^2 + n^2\lambda^2\right)^2} \quad (4)$$

$$E_1(t) = \left(E_m + \frac{E_c - E_m}{k+1}\right)h(t),\ \rho_1 = \left(\rho_m + \frac{\rho_c - \rho_m}{k+1}\right)h(t)$$

$k_1$, $k_2$ – curvatures of middle surface shell in $x_1$ and $x_2$ directions.

Initial differential equation (3) we rewrite in the form:

$$\varepsilon^2 \frac{d^2 f}{dt^2} + B_1(t) f + \mu\left(B_2(t) f^2 + B_3(t) f^3\right) = \bar{Q}_0(t) \quad (5)$$

where

$$B_1(t) = 1 + 2 f_0 \bar{A}_2(t) - \bar{A}_3(t) f_0^2 - \bar{A}_1(t)$$

$$B_2(t) = \frac{-3}{\mu}\bar{A}_2(t)$$

$$B_3(t) = \frac{1}{\mu}\bar{A}_3(t) \quad (6)$$

$$\bar{Q}_0(t) = Q_0 - \bar{A}_0(t) + f_0 - \bar{A}_2(t) f_0^2$$

$\varepsilon$, $\mu$ are parameters.

According to perturbation method with respect to parameter of nonlinearity $\mu$, solution of differential equation (5) we obtain in the form of two terms approximation:

$$f(t) = \varphi_0(t) + \mu\varphi_1(t) \quad (7)$$

Substituting (7) into equation (5) and acquainted the terms with the same order of nonlinear parameter we obtain the system equations for unknown functions $\varphi_0(t)$ and $\varphi_1(t)$:

$$\mu^0 : \varepsilon^2 \varphi_0''(t) + B_1(t)\varphi_0 = \bar{Q}_0 \tag{8}$$

$$\mu^1 : \varepsilon^2 \varphi_1''(t) + B_1(t)\varphi_1 = -B_2(t)\varphi_0^2 - B_3(t)\varphi_0^3 \tag{9}$$

Ordinary singular differential equation with variable in time coefficient $B_1$ is solved by two terms WKB-approximation [5]:

$$\varphi_0^0(t) = 1/B_1(t)^{0.25}\left(c_1 \sin K(t) + c_2 \cos K(t)\right) \tag{10}$$

where

$$K(t) = \int \varepsilon^{-1} B_1^{0.25}(t)\,dt \tag{11}$$

Particular solution of equation (9) can be present in the form:

$$\varphi_0^p(t) = 1/B_1(t)^{0.25}\left[\bar{c}_1(t)\sin K(t) + \bar{c}_2(t)\cos K(t)\right] \tag{12}$$

where

$$\bar{c}_1(t) = \varepsilon \int \frac{\bar{Q}_0(t)\cos K(t)}{B_1^{-0.25}(t)}dt$$

$$\bar{c}_2(t) = -\varepsilon \int \frac{\bar{Q}_0(t)\sin K(t)}{B_1^{-0.25}(t)}dt \tag{13}$$

The solution of equation (5) in the first approximation is:

$$\varphi_0(t) = \varphi_0^0(t) + \varphi_0^p(t) = B_1^{0.25}\left[\sin K(t)\left(c_1 + \bar{c}_1(t)\right) + \cos K(t)\left(c_2 + \bar{c}_2(t)\right)\right] \tag{14}$$

Second term in (9) for ordinary equation is obtained from:

$$\varepsilon^2 \varphi_1''(t) + B_1(t)\varphi_1 = 0 \tag{15}$$

where

$$K(t) = \int \varepsilon^{-1} B_1^{0.25}(t)\,dt \tag{16}$$

Particular solution of equation (5) can be present in the form:

$$\varphi_0^p(t) = 1/B_1(t)^{0.25}\left[\bar{c}_1(t)\sin K(t) + \bar{c}_2(t)\cos K(t)\right] \tag{17}$$

where

$$\begin{aligned}\bar{c}_1(t) &= \varepsilon\int \frac{\bar{Q}_0(t)\cos K(t)}{B_1^{-0.25}(t)}dt \\ \bar{c}_2(t) &= -\varepsilon\int \frac{\bar{Q}_0(t)\sin K(t)}{B_1^{-0.25}(t)}dt\end{aligned} \tag{18}$$

The solution of equation (5) in the first approximation is:

$$\varphi_0(t) = \varphi_0^0(t) + \varphi_0^p(t) = B_1^{0.25}\left[\sin K(t)\left(c_1 + \bar{c}_1(t)\right) + \cos K(t)\left(c_2 + \bar{c}_2(t)\right)\right] \tag{18}$$

Second term in (9) for ordinary equation is obtained from:

$$\varepsilon^2\varphi_1''(t) + B_1(t)\varphi_1 = 0 \tag{19}$$

in the form:

$$\varphi_1^0(t) = 1/B_1(t)^{0.25}\left(d_1\sin K(t) + d_2\cos K(t)\right) \tag{20}$$

Corresponding particular solution of equation (9) is

$$\varphi_1^p(t) = 1/B_1(t)^{0.25}\left[\bar{d}_1(t)\sin K(t) + \bar{d}_2(t)\cos K(t)\right] \tag{21}$$

where

$$\begin{aligned}\bar{d}_1(t) &= \varepsilon\int \frac{\left(-B_2(t)\varphi_0^2 - B_3(t)\varphi_0^3\right)\bar{Q}_0(t)\cos K(t)}{B(t)^{-0.25}}dt \\ \bar{d}_2(t) &= -\varepsilon\int \frac{\left(-B_2(t)\varphi_0^2 - B_3(t)\varphi_0^3\right)\bar{Q}_0(t)\sin K(t)}{B_1(t)^{-0.25}}dt\end{aligned} \tag{22}$$

Finally, we have obtained the solution of nonlinear problem on the basis of perturbation – (two-term) WKB method:

$$f(t)=\varphi_0(t)+\mu\varphi_1(t)=\sin K(t)\left(c_1+\bar{c}_1(t)\right)+\cos K(t)\left(c_2+\bar{c}_2(t)\right)+$$
$$+\mu\left(\sin K(t)\left(d_1+\bar{d}_1(t)\right)+\cos K(t)\left(d_2+\bar{d}_2(t)\right)\right)= \tag{23}$$
$$=B_1(t)^{0.25}\left\{sinK(t)\left[c_1+\bar{c}_1(t)+\mu\bar{d}_1(t)\right]+cosK(t)\left[c_2+\bar{c}_2(t)+\mu\bar{d}_2(t)\right]\right\}$$

The initial conditions are:

$$\begin{aligned}\varphi(0)&=1\\ \varphi'(0)&=0\end{aligned} \tag{24}$$

In order to obtain hybrid asymptotic solution of initial nonlinear equation we rewrite equation (3) as

$$\varepsilon^2\frac{d^2f}{dt^2}+B_1(t)f=\bar{Q}_0(t)-\mu\left(B_2(t)f^2+B_3(t)f^3\right)=\bar{Q}_0(t)-\mu N(t) \tag{25}$$

Approximate analytical (P-WKB-G) solutions of equation (3) are given in the following forms (26) - (29).

### 3.1 Hybrid asymptotic solution of linear homogeneous problem

$$f_0^0(t)=exp(\varphi(t))\left(cos(K(t))(-0.5973)+sin(K(t))(-0.8508)\right) \tag{26}$$

where according to [5]:

$$\varphi(t)=\int\frac{B_1(0)-B_1(1)}{4\int_0^1 B_1(t)^{1.5}\,dt}B_1(t)^{0.5}\,dt$$
$$K(t)=\int\sqrt{\frac{1}{\varepsilon^2}-\frac{B_1(1)-B_1(0)}{4\int_0^1 B_1(t)^{1.5}\,dt}}B_1(t)^{0.5}\,dt \tag{27}$$

Some numerical calculations for the shell with variable in time parameters and comparison of approximate analytical solutions with direct numerical integration of initial nonlinear nonhomogeneous equation with variable coefficients are given on Figures 1-7.

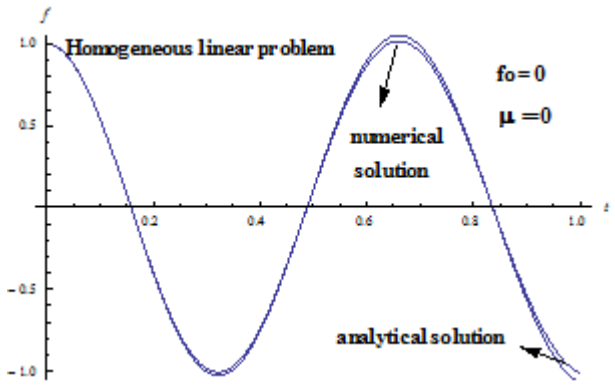



Figure 1.Comparison of hybrid asymptotic and numerical solutions for homogeneous linear problem

## 3.2 Hybrid asymptotic solution of linear nonhomogeneous problem

$$f_0(t) = f_0^0(t) + f_0^p(t) = exp(\varphi(t))\left(cos(K(t))\left(c_1 + \vec{c}_1(t)\right) + sin(K(t))\left(c_2 + \vec{c}_2(t)\right)\right) \tag{28}$$

$$c_1 = -0.5522$$

$$c_2 = -0.8826$$

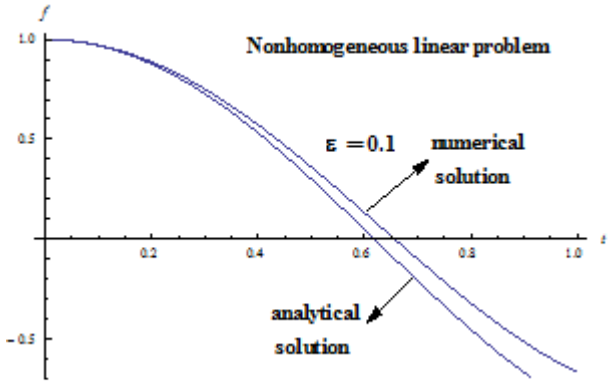



Figure 2.Comparison of hybrid asymptotic and numerical solutions for nonhomogeneous linear problem (one - term WKB-approximation)

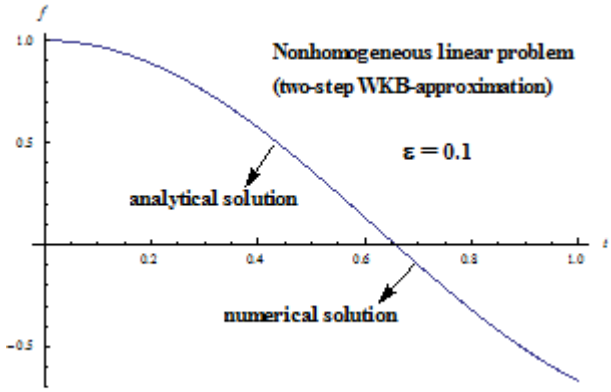



Figure 3. Comparison of analytical and numerical solutions for nonhomogeneous linear problem (two - terms WKB-approximation)

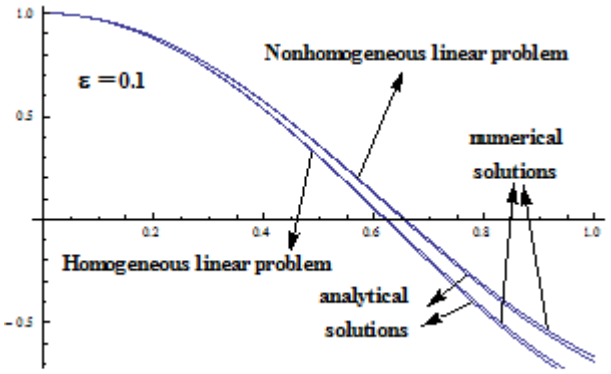



Figure 4. Comparison of analytical and numerical solutions

### 3.3 Hybrid asymptotic solution of nonlinear nonhomogeneous problem

$$f(t)=exp(\varphi(t))\left(cos(K(t))\left(c_1+\bar{c}_1(t)+\mu\bar{d}_1(t)\right)+sin(K(t))\left(c_2+\bar{c}_2(t)+\mu\bar{d}_2(t)\right)\right) \tag{29}$$

$$c_1=-0.541$$

$$c_2=-0.8566$$

where

$$\varphi(t)=-0.03872(1+0.1t)^{1.5},\ K(t)=16.6644(1+0.1t)^{1.5} \tag{30}$$

$$\begin{aligned}\bar{c}_1(t)&=-\int\frac{\bar{Q}_0(t)sinK(t)}{K'(t)}=\int series\left[-0.2(1+0.1t)^{0.5}cosK(t)\right]_0^8dt\\ \bar{c}_2(t)&=\int\frac{\bar{Q}_0(t)cosK(t)}{K'(t)}=\int series\left[0.2(1+0.1t)^{0.5}sinK(t)\right]_0^8dt\end{aligned} \tag{31}$$

$$\begin{aligned}\bar{d}_1(t)&=-\int\frac{N(t)sinK(t)}{K'(t)}=-\int series\left[\frac{30f_0^2-10f_0^3}{2.5(1+0.1t)^{0.5}}sinK(t)\right]_0^{22}dt\\ \bar{d}_2(t)&=\int\frac{N(t)cosK(t)}{K'(t)}=\int series\left[\frac{30f_0^2-10f_0^3}{2.5(1+0.1t)^{0.5}}cosK(t)\right]_0^{22}dt\end{aligned} \tag{32}$$

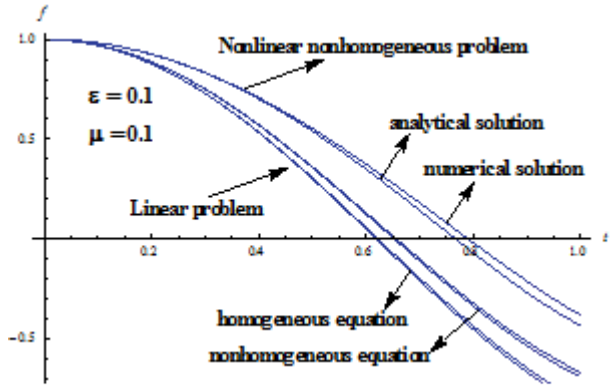



Figure5. Comparison of analytical and numerical solutions

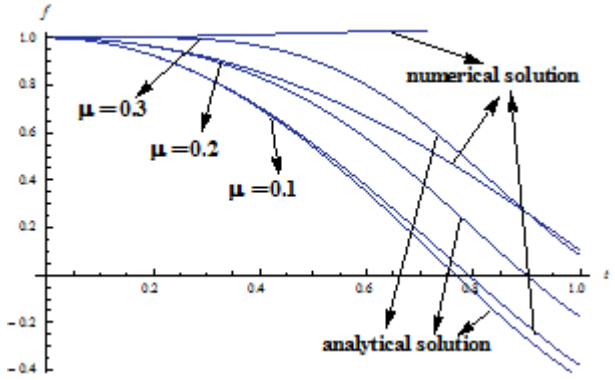



Figure 6. Influence of nonlinear parameter μ. (Comparison of analytical and numerical solutions at ε = 0.1)

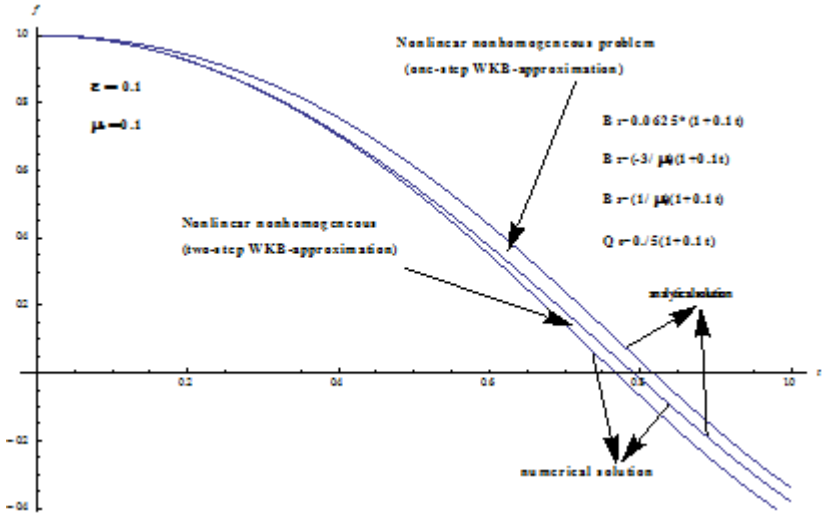



Figure7. Comparison of analytical and numerical solutions for nonhomogeneous linear and nonlinear problems (one-step and two-step WKB-approximations)

## 4 CONCLUSIONS

An approximate analytical solution for forced oscillations of geometrically non-linear FGM imperfect shallow cylindrical shells with time dependent parameters on the basis of hybrid perturbation-two-terms WKB approximation method are obtained. For some particular parameters of structure an analytical solutions are in a good enough correlations with direct numerical solutions of initial singular nonlinear differential equations with variable coefficients. In some cases one-term WKB-approximation gives good enough results for the practical purpose.

Researches will be devoted in future to study of the values of perturbation  and singular $\varepsilon$ parameters influence on dynamics of FGM shell behavior according to three – step (P-WKB-G) approach, different thickness function in time and functions of external forces as well.

## REFERENCES

[1] Dao Huy Bich; Yu Do Long: Nonlinear Dynamical Analysis of Imperfect Functionally Graded Materials Shallow Shells. *Vietnam Journal og Mechanics, VAST*, 32, 1 (2010), 1-14.

[2] Vu Thi Thuy Anh; Nguyen Dinh Duc: Nonlinear Axisymmetric Response of Thin FGM Shallow Spherical Shells with Ceramic-Metal-Ceramic Layers under External Pressure and Temperature. *VNU Journal of Mathematics – physics*, 29, 2 (2013), 1-15.

[3] Katarzhyna Kowal-Michalska; Radoslaw J. Mania: Static and Dynamic Thermomechanical Buckling Load of Functionally Graded Plate. *Mechanics and Mechanical Engineering*, 17, 1 (2013), 99-112.

[4] Khaire, P.D.; Ambhore, N.H.; Jagtap, K.R.: Nonlinear Free Vibration Response of FGM Cylindrical Shell in Thermal Environment. *Int. Journal on Mech. Engineering and Robotics*, 2 (2014).

[5] Geer, J.F.; Andersen, C.V.: Natural Frequency Calculations Using a Hybrid Perturbation Galerkin Technique. *Pan American Congress on Appl. Mech.* (1991), 571-574.

[6] Gristchak, V.; Dmitrijeva, Ye.: A Hybrid WKB-Galerkin Method and its Application. *Technische Mechanik*, 15, 4, (1995), 281-294.

[7] Gristchak, V.; Kabak, V.: Double Asymptotic Method for Nonlinear Forced Oscillations Problem of Mechanical Systems with Time Dependent Parameters. *Technische Mechanik*, 4, (1996), 285-296.

[8] Gristchak,V.Z.; Ganilova,O.A.: Application of a Hybrid WKB-Galerkin Methodin Control of the Dynamic Instability of a Piezolaminated Imperfect Column. T*echnische Mechanik*, 26, 2,(2006), 106-116.

[9] Gristchak,V.Z.; Ganilova,O.A.: Tothedynamic deformation problem of piezoelectric sandwich plates on the basic of hybrid WKB-Galerkin method. *Reportsof National Academy of Science of Ukraine*, 5, (2008), 13-20.

[10] Gristchak,V.Z.; Ganilova,O.A.: A Hybrid WKB-Galerkin Method Applied to a Piezoelectric Sandwich Plate Vibration Problem Considering Shear Force Effects. *Journal of Sound and Vibration*, 317, 1-2, (2008), 366-377.

[11] Gristchak,V.Z.; Ganilova,O.A.: Application of a Hybrid WKB- GalerkinMethodtoa Nonlinear Plate Dynamic Problem with Time Dependent Damping Coefficient. *Proc. VI Int. Conf. EUROMECH ENOC*, 80Min / 700MB, Sankt-Petersburg(2008).

[12] Cartmell,M.P.;Ganilova, O.A.; Gristchak, V.Z.: ControllingPlateVibrationsbyMeansofEmbeddedActiveShapeMemoryAlloyElements.*Euromech 498 Colloquium: international conference*: *Proceedings of Euromech 498 Colloquium*,498, (2008), 112–118.

[13] Kurpa,L.V.;Osetrov, A.A.; Shmatko, T.V.: Determination of eigen frequencies offunctionally graded material shallow shells with using of R-function and splineapproximation. *Announcer of NTU CHPI*, 6, (2014), 99-111.

[14] Stephen Wolfram: Mathematica. A System for Doing Mathematica by the Computer. *Wolfram Research Inc.* (1988).

[15] Gristchak, V.Z.; Fatieieva, Ju.A.: An approximate nonlinear dynamic problem solution of functionally graded material shallow shell structure with in time thickness variation. *Announcer of ZNU*, 3, (2014).

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# Uncertainty of models in intelligent systems under stochastic loading

**A. Moutsopoulou[1*], G.E. Stavroulakis[2], and A. D. Pouliezos[3]**

[1]Department of Civil Engineering, Technological Educational Institute of Crete

Estavromenos, Heraklion Crete, Greece, amalia@staff.teicrete.gr

[2,3]Department of Production Engineering and Management,

Technical University of Crete, Kounoupidianna GR-73100 Chania, Greece,

gestavr@dpem.tuc.gr, tasos@dpem.tuc.gr

**Keywords:** Intelligent structures, robust active control, uncertain structural systems, $H_\infty$ performance

**Abstract.** *In this paper, we address the problem of vibrations of intelligent structures. Stimuli may come from external perturbations of the system, disturbances or excitation that may cause structural vibrations, such as wind loading or earthquake. First, an accurate model of a piezocomposite intelligent structure with special boundary conditions is derived by using of FEM analysis. Then the robustness of the uncertain closed-loop model performances has been investigated. Obtained results show the higher performance of $H_{infinity}$ design approach in rejection of disturbances.*

## 1 INTRODUCTION

The study of algorithms for active vibrations control in intelligent structures became an area of enormous interest, mainly due to the countless demands of an optimal performance of mechanical systems. Many researchers are investigated in the field of intelligent structures [1, 2, 3, 4, 5, 6 ]. A smart structure in one that monitors itself and its environment in order to respond to changes in its conditions. [7,8] Smart structures, formed by a structure base, coupled with piezoelectric actuators and sensor are capable to guarantee the conditions demanded through the application of several types of controllers. This article shows some steps that should be followed in the design of a smart structure. In our paper a cantilever slender beam with rectangular cross-sections is considered. Thirty six pairs of piezoelectric patches are embedded symmetrically at the top and the bottom surfaces of the beam. The beam is from graphite- epoxy T300 − 976 and the piezoelectric patches are PZT G1195N. The top patches act like sensors and the bottom like actuators. The resulting composite beam is modelled by means of the classical laminated technical theory of bending. Let us assume that the mechanical properties of both the piezoelectric material and the host beam are independent in time. The thermal effects are considered to be negligible as well [8, 9].

The beam has length L, width W and thickness h. The sensors and the actuators have width $b_S$ and $b_A$ and thickness $h_S$ and $h_A$, respectively. The electromechanical parameters of the beam of interest are given in the table 1.

| **Parameters** | **Values** |
|---|---|
| Beam length, L | 0.7m |
| Beam width, W | 0.07m |
| Beam thickness, h | 0.0096m |
| Beam density, $\rho$ | 1800kg/m$^3$ |
| Young s modulus of the beam, E | 1.5 X 10$^{11}$ N/m$^2$ |
| Piezoelectric constant, $d_{31}$ | 254 X 10$^{-12}$ m/V |
| Electric constant, $\xi_{33}$ | 11.5 X 10$^{-3}$ V m/N |
| Young's modulus of the piezoelectric element | 1.5 X 10$^{11}$ N/m$^2$ |
| Width of the piezoelectric element | $b_S$= $b_a$= 0.07m |
| Thickness of the piezoelectric element | $h_S$= $h_a$= 0.0002m |

Table 1: Parameters of the composite beam.

In order to derive the basic equations for piezoelectric sensors and actuators [1, 2], we assume that:

- The piezoelectric sensors actuators (S/A) are bonded perfectly on the host beam;

- The piezoelectric layers are much thinner then the host beam;
- The piezoelectric material is homogeneous, transversely isotropic and linearly elastic;
- The piezoelectric S/A are transversely polarized [1, 2, 7].

## 2 SYSTEM MODELLING

This classical finite element procedure leads to the approximate discretized variation problem. For a finite element the discrete differential equations are obtained by substituting the discretized expressions into the first variation of the kinetic energy and strain energy [8, 10]. Integrating over spatial domains and using the Hamiltons principle [8], the equation of motion for a beam element are expressed in terms of nodal variable q as follows,

$$M\ddot{q}(t) + D\dot{q}(t) + Kq(t) = f_m(t) + f_e(t) \tag{1}$$

where M is the generalized mass matrix, D the viscous damping matrix, K the generalized stiffness matrix, $f_m$ the external loading vector and $f_e$ the generalized control force vector produced by electromechanical coupling effects. The independent variable q(t) is composed of transversal deflections $\omega_1$ and rotations $\psi_1$, i.e., [10, 12]

Furthermore n is the number of nodes used in analysis and vectors ω and $f_m$ are positive upwards. To transform to state-space control representation, let (in the usual manner),

$$\dot{x}(t) = \begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} \tag{2}$$

Furthermore to express $f_e(t)$ as $Bu(t)$ we write it as $f_e^* u$ where $f_e^*$ the piezoelectric force is for a unit applied on the corresponding actuator, and u represents the voltages on the actuators. Furthermore, $d(t) = f_m(t)$ is the disturbance vector [10].
Then,

$$\dot{x}(t) = \begin{bmatrix} O_{2n\times 2n} & I_{2n\times 2n} \\ -M^{-1}K & -M^{-1}D \end{bmatrix} x(t) + \begin{bmatrix} O_{2n\times 2n} \\ M^{-1} f_e^* \end{bmatrix} u(t) + \begin{bmatrix} O_{2n\times 2n} \\ M^{-1} \end{bmatrix} \tag{3}$$

$$= Ax(t) + Bu(t) + Gd(t) = Ax(t) + \begin{bmatrix} B & G \end{bmatrix} \begin{bmatrix} u(t) \\ d(t) \end{bmatrix} = Ax(t) + \tilde{B}\tilde{u}(t) \tag{4}$$

The previous description of the dynamical system will be augmented with the output equation (displacements only measured) [5,11],

$$y(t) = \begin{bmatrix} x_1(t) & x_3(t) & \cdots & x_{n-1}(t) \end{bmatrix}^T = Cx(t) \tag{5}$$

The units used are compatible for instance m, rad, sec and N. [7,8]

## 3 DESIGN OBJECTIVES AND SYSTEM SPECIFICATION

The structured singular value of a transfer function matrix is defined as,

$$\mu(M)=\begin{cases}\dfrac{1}{\min\limits_{k_m}\{\det(I-k_m M\Delta)=0,\bar{\sigma}(\Delta)\le 1\}}\\ 0,\text{if no such structured exists}\end{cases} \tag{6}$$

In words it defines the smallest structured $\Delta$ (measured in terms of $\bar{\sigma}(\Delta)$) which makes $\det(I-M\Delta)=0$: then $\mu(\mathrm{M})=1/\bar{\sigma}(\Delta)$. It follows that values of $\mu$ smaller than 1 are desired (the smaller the better: a larger variation is allowed) [13, 14].

## 4 SYSTEM UNCERTAINTY

The main sources of uncertainty are:

* Nonlinearity and/or dynamic aspects of the system that are ignored at the modeling phase. The error introduced in modal analysis by using only a few significant eigenmodes leads to an uncertainty of the type discussed here. [15, 16]

* Incomplete knowledge of model values and parameters and/or natural fluctuation of those values during system operation.

* Influence of the system's environment, in the form of disturbances.

Assume uncertainty in the $M$ and $K$ matrices of the form,

$$K=K_0(I+k_p I_{2n\times 2n}\delta_K) \tag{7}$$

$$M=M_0(I+m_p I_{2n\times 2n}\delta_M) \tag{8}$$

Also, since, $D=0.0005(K+M)$, an appropriate form for $D$ is,

$$D=0.0005[K_0(I+k_p I_{2n\times 2n}\delta_K)+M_0(I+m_p I_{2n\times 2n}\delta_M)]=$$
$$D_0+0.0005[K_0 k_p I_{2n\times 2n}\delta_K+M_0 m_p I_{2n\times 2n}\delta_M] \tag{9}$$

Alternatively, by adopting the well-known Rayleigh damping assumption,

$$D=\alpha K+\beta M \tag{10}$$

$D$ could be expressed similarly to $K$, $M$, as,

$$D=D_0(I+d_p I_{2n\times 2n}\delta_D) \tag{11}$$

In this way we introduce uncertainty in the form of percentage variation in the relevant matrices. Uncertainty is most likely to arise from terms outside the main matrices (since length can be adequately measured).

Here it will be assumed,

$$\|\Delta\|_\infty \overset{\text{def}}{=} \left\| \begin{bmatrix} \mathrm{I}_{n\times n}\delta_{\mathrm{K}} & 0_{n\times n} \\ 0_{n\times n} & \mathrm{I}_{n\times n}\delta_{\mathrm{M}} \end{bmatrix} \right\|_\infty < 1 \tag{12}$$

hence $m_p$, $k_p$ are used to scale the percentage value and the zero subscript denotes nominal values.

(it is reminded that for matrix $A_{n\times m}$ the norm is calculated through $\|A\|_\infty = \max_{1\le j\le m} \sum_{j=1}^{n} |a_{ij}|$)

With these definitions Eq. 3 becomes,

$$M_0(I+m_pI_{2n\times 2n}\delta_M)\,\ddot{q}(t) + K_0(I+k_pI_{2n\times 2n}\delta_K)q(t) + \\ +[D_0+0.0005[K_0k_pI_{2n\times 2n}\delta_K+M_0m_pI_{2n\times 2n}\delta_M]]\,\dot{q}(t) = f_m(t)+f_e(t) \tag{13}$$

$$\Rightarrow M_0\,\ddot{q}(t)+D_0\,\dot{q}(t)+K_0q(t) = \\ -[M_0m_pI_{2n\times 2n}\delta_M\,\ddot{q}(t)+0.0005[K_0k_pI_{2n\times 2n}\delta_K+M_0m_pI_{2n\times 2n}\delta_M]\,\dot{q}(t) \\ +K_0k_pI_{2n\times 2n}\delta_K q(t)]+f_m(t)+f_e(t) \tag{14}$$

$$\Rightarrow M_0\,\ddot{q}(t)+D_0\,\dot{q}(t)+K_0q(t) = \tilde{D}q_u(t)+f_m(t)+f_e(t) \tag{15}$$

where,

$$q_u(t) = \begin{bmatrix} \ddot{q}(t) \\ \dot{q}(t) \\ q(t) \end{bmatrix} \tag{16}$$

$$\tilde{D} = -\begin{bmatrix} M_0m_p & K_0k_p \end{bmatrix} \begin{bmatrix} I_{2n\times 2n}\delta_M & 0_{2n\times 2n} \\ 0_{2n\times 2n} & I_{2n\times 2n}\delta_{\mathrm{K}} \end{bmatrix} \begin{bmatrix} I_{2n\times 2n} & 0.0005I_{2n\times 2n} & 0_{2n\times 2n} \\ 0_{2n\times 2n} & 0.0005I_{2n\times 2n} & I_{2n\times 2n} \end{bmatrix} = \\ = G_1\cdot\Delta\cdot G_2 \tag{17}$$

Writing in state space form, gives,

$$\dot{x}(t) = \begin{bmatrix} 0_{2n\times 2n} & I_{2n\times 2n} \\ -M^{-1}K & -M^{-1}D \end{bmatrix} x(t) + \begin{bmatrix} 0_{2n\times n} \\ M^{-1}f_e^* \end{bmatrix} u(t) + \begin{bmatrix} 0_{2n\times 2n} \\ M^{-1} \end{bmatrix} d(t) + \begin{bmatrix} 0_{2n\times 6n} \\ M^{-1}G_1\cdot\Delta\cdot G_2 \end{bmatrix} q_u(t) \\ = Ax(t)+Bu(t)+Gd(t)+G_uG_2q_u(t) \tag{18}$$

In this way we treat uncertainty in the original matrices as an extra uncertainty term.

To express our system consider in the frequency domain Fig. 6

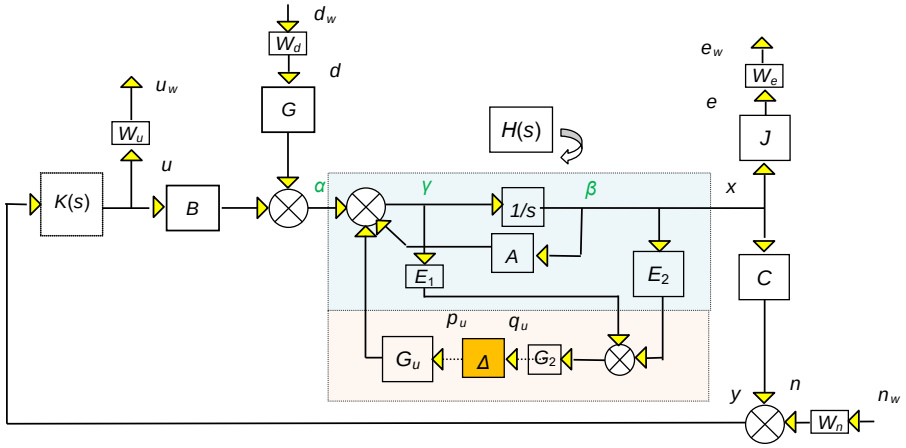



Fig. 6 Uncertainty block diagram

This diagram is the weighted block diagram in the frequency domain. Wd, We, Wu, Wn are the weight of the disturbances, errors, control, noise. *H(s)* is our system, *K(s)*, is the controller and Δ define the uncertainties. [16, 17]

The matrices $E_1$, $E_2$ are used to extract,

$$q_u(t) \overset{\text{def}}{=} \begin{bmatrix} \ddot{q}(t) \\ \dot{q}(t) \\ q(t) \end{bmatrix} \tag{19}$$

Since,

$$\gamma = \begin{bmatrix} \dot{q}(t) \\ \ddot{q}(t) \end{bmatrix} \quad \text{and} \quad \beta = \int \begin{bmatrix} \dot{q}(t) \\ \ddot{q}(t) \end{bmatrix} \mathrm{d}t = \begin{bmatrix} q(t) \\ \dot{q}(t) \end{bmatrix} \tag{20}$$

appropriate choices for $E_1$, $E_2$ are,

$$E_1 = \left[\begin{array}{c:c} 0_{2n\times 2n} & I_{2n\times 2n} \\ \hdashline I_{2n\times 2n} & 0_{2n\times 2n} \\ \hdashline 0_{2n\times 2n} & 0_{2n\times 2n} \end{array}\right], \quad E_2 = \left[\begin{array}{c:c} 0_{2n\times 2n} & 0_{2n\times 2n} \\ \hdashline 0_{2n\times 2n} & 0_{2n\times 2n} \\ \hdashline I_{2n\times 2n} & 0_{2n\times 2n} \end{array}\right] \tag{21}$$

The idea is to find an *N* such that,

$$\begin{bmatrix} q_u \\ e_w \\ u_w \end{bmatrix} = N \begin{bmatrix} p_u \\ d_w \\ n_w \end{bmatrix}, \quad N = \left[\begin{array}{c:c:c} N_{p_u q_u} & N_{d_w q_u} & N_{n_w q_u} \\ \hdashline N_{p_u e_w} & N_{d_w e_w} & N_{n_w e_w} \\ \hdashline N_{p_u u_w} & N_{d_w u_w} & N_{n_w u_w} \end{array}\right] = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} \tag{22}$$

or in the notation of Fig. 6,

$$\begin{bmatrix} q_u \\ w \end{bmatrix} = N \begin{bmatrix} p_u \\ z \end{bmatrix} \tag{23}$$

We'll use a methodology known as "pulling out the Δ's". To this end, break the loop at points $p_u$, $q_u$ (which will be used as additional inputs/outputs respectively) and use the auxiliary signals $\alpha$, $\beta$ and $\gamma$. [18,19]

To get the transfer function $N_{d_w q_u}$ (from $d_w$ to $q_u$):

$$q_u = G_2(E_2\beta + E_1\gamma) = G_2(E_2\frac{1}{s} + E_1)\gamma \tag{24}$$

$$\gamma = GW_d d_w + Bu + A\frac{1}{s}\gamma = GW_d d_w + BKC\frac{1}{s}\gamma + A\frac{1}{s}\gamma$$

$$\Rightarrow \gamma = (I - BKC\frac{1}{s} - A\frac{1}{s})^{-1} GW_d d_w \tag{25}$$

Hence,

$$N_{d_w q_u} = G_2(E_2\frac{1}{s} + E_1)(I - BKC\frac{1}{s} - A\frac{1}{s})^{-1} GW_d$$

Now, $N_{p_u q_u}$, $N_{p_u e_w}$, $N_{p_u u_w}$, are similar to $N_{d_w q_u}$, $N_{d_w e_w}$, $N_{d_w u_w}$ with $GW_d$ replaced by $G_u$, i.e.,

$$N_{p_u q_u} = G_2(E_2\frac{1}{s} + E_1)(I - BKC\frac{1}{s} - A\frac{1}{s})^{-1} G_u$$

$$N_{p_u e_w} = W_y JH[I + B[K(I - CHBK)^{-1} CH]G_u \tag{26}$$

$$M_{p_u u_w} = W_u K(I - CHBK)^{-1} CHG_u$$

Finally to find $N_{n_w q_u}$,

$$q_u = G_2(E_2\beta + E_1\gamma) = G_2(E_2\frac{1}{s} + E_1)\gamma \tag{27}$$

$$\gamma = Bu + A\frac{1}{s}\gamma = BK(W_n n_w + y) + A\frac{1}{s}\gamma = BKW_n n_w + BKC\frac{1}{s}\gamma + A\frac{1}{s}\gamma$$

$$\Rightarrow \gamma = (I - BKC\frac{1}{s} - A\frac{1}{s})^{-1} BKW_n n_w \tag{28}$$

Hence,

$$N_{n_w q_u} = G_2(E_2\frac{1}{s} + E_1)(I - BKC\frac{1}{s} - A\frac{1}{s})^{-1} BKW_n \tag{29}$$

Collecting all the above yields $N$:

$$\begin{bmatrix} G_2\left(E_2\frac{1}{s}+E_1\right)\left(I-BKC\frac{1}{s}-A\frac{1}{s}\right)^{-1}G_u & G_2\left(E_2\frac{1}{s}+E_1\right)\left(I-BKC\frac{1}{s}-A\frac{1}{s}\right)^{-1}GW_d & G_2\left(E_2\frac{1}{s}+E_1\right)\left(I-BKC\frac{1}{s}-A\frac{1}{s}\right)^{-1}BKW_n \\ W_eJH\left[I+BK(I-CHBK)^{-1}CF\right]G_u & W_eJ(I-HBKC)^{-1}HGW_d & W_eJ(I-HBKC)^{-1}HBKW_n \\ W_uK(I-CHBK)^{-1}CFG_u & W_u(I-KCHB)^{-1}KCHGW_d & W_u(I-KCHB)^{-1}KW \end{bmatrix} \quad (30)$$

Having obtained $N$ for the beam problem, all proposed controllers $K(s)$ can be compared using the structured singular value relations. [18, 19]

## 5 INPUTS

A typical wind load (Fig. 7) acting on the side of the structure. The wind load is a real life wind speed measurements in relevance with time that took place in Estavromenos of Heraklion Crete. We transform the wind speed in wind pressure with,

Loading corresponds to the wind excitation. The function $f_m(t)$ has been obtained from the wind velocity record, through the relation

$$f_m(t)=\frac{1}{2}\rho C_u V^2(t) \quad (31)$$

where $V$=velocity, $\rho$=density and $C_u$=1.2.

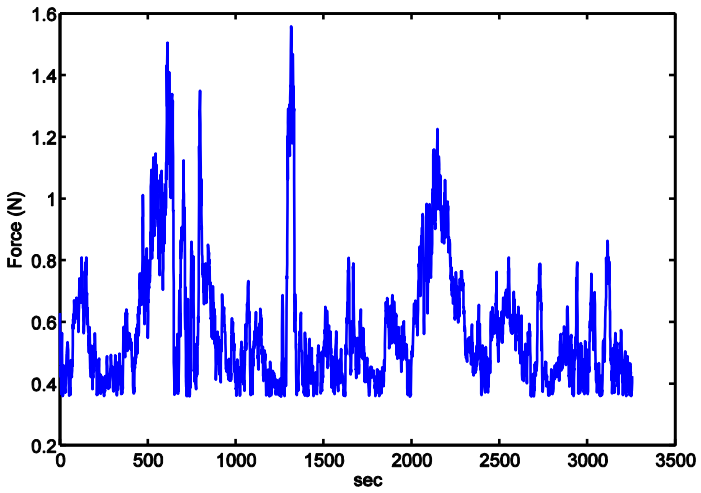



Fig.7 External load

Moreover, in all simulations, random noise has been introduced to measurements at system output locations within a probability interval of ±1%. Due to small displacements of system nodal points, noise amplitude is taken to be small, of the order of $5 \times 10^{-5}$. On the other hand, the signal is introduced at each node of the beam by a different percentage, that percentage being lower at the first node due to the fact that the beam end point is clamped. The controller obtained by applying $H_\infty$ control has an order equal to 36. For this controller, $\gamma = 0.074<1$.

## 6 RESULTS

Robust analysis is carried out through the relations:

$$\sup_{\omega \in \mathrm{P}} \mu_{\Delta}\left(N_{11}(\mathrm{j}\omega)\right) < 1$$

for robust stability, and,

$$\sup_{\omega \in \mathrm{P}} \mu_{\Delta_a}\left(N(\mathrm{j}\omega)\right) < 1$$

for robust performance

For the $H_\infty$ found, robust analysis was performed for the following values of *mp*,*kp*.
1. *mp* = 0, *kp* = 0.9. This corresponds to a ±90% variation from the nominal value of the stiffness matrix *K*. In Fig.8 are shown the bounds on the $\mu$ values. As seen the system remains stable and exhibits robust performance, since the upper bounds of both values remain below 1 for all frequencies of interest. This result is validated in Fig.9, where the displacement of the free end and the voltage applied are shown at the extreme uncertainty. Comparison with the open loop response for the same plant shows the good performance of the $H_\infty$ controller. Results are very good, and the beam remains in equilibrium even under realistic wind conditions. Reduction of vibrations is observed, while piezoelectric add-ons produce voltage within their tolerance limits (±500volt)

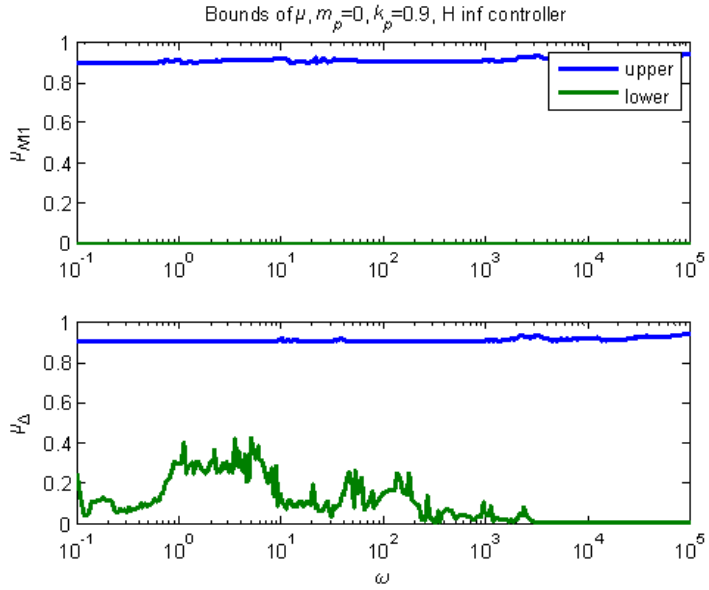



Fig. 8 $\mu$-bounds o $H_\infty$ the controller for $m_p$=0, $k_p$=0.9

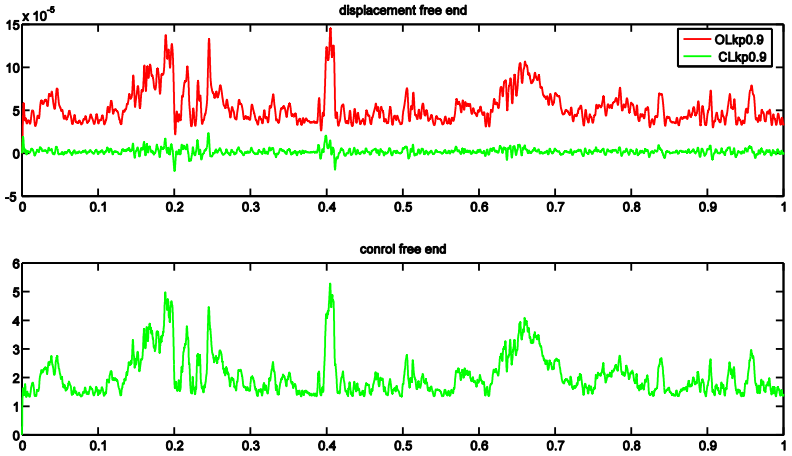



Fig.9 Displacement and control at free end for the $H_\infty$ controller with $m_p = 0$, $k_p = 0.9$

2. *mp* = 0.9, *kp* = 0. This corresponds to a ±90% variation from the nominal value of the mass matrix *M*.

In Fig. 10 are shown the bounds on the $\mu$ values. As seen the system remains stable and exhibits robust performance, since the upper bounds of both values remain below 1 for all frequencies of interest. This result is validated in Fig. 11, where the displacement of the free end and the voltage applied are shown.

Comparison with the open loop response for the same plant shows the good performance of the controller. By employing the $H_\infty$ control, vibration reduction is achieved, while the voltage applied is significantly lower that 500 V.

.

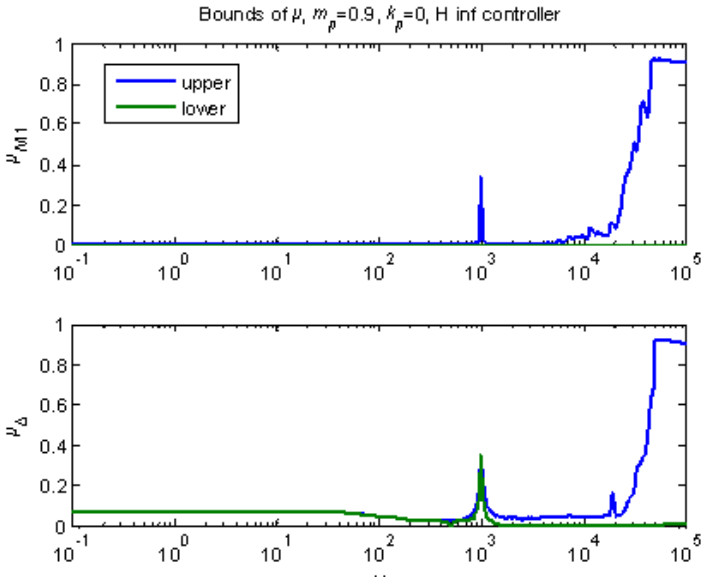



Fig. 10 $\mu$-bounds of the $H_\infty$ controller for $m_p = 0.9$, $k_p = 0$.

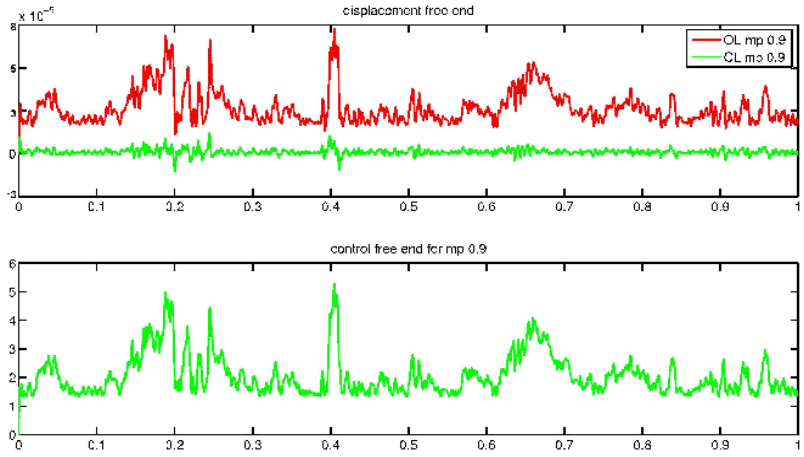



Fig. 11 Displacement and control at free end for the $H_\infty$ controller with $m_p$ =0.9, $k_p$ =0

3. *mp* = 0.9, *kp* = 0.9. This corresponds to a ±90% variation from the nominal values of both the mass and stiffness matrices *M*, *K*.

In Fig. 12 are shown the bounds on the $\mu$ values. As seen the system remains stable and exhibits its robust performance, since the upper bounds of both values remain below 1 for all frequencies of interest. This result is validated in Fig. 13, where the displacement of the free end and the voltage applied are shown.

Comparison with the open loop response for the same plant shows the good performance of the controller. Results are very good, and the beam remains in equilibrium even under realistic wind conditions. Reduction of vibrations is observed, while piezoelectric add-ons produce voltage within their tolerance limits.

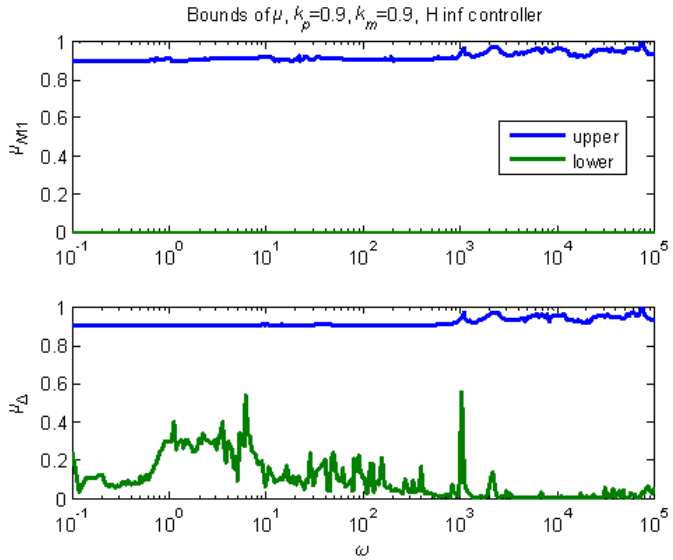



Fig. 12 Displacement and control at free end for the $H_\infty$ controller with $m_p$ =0.9, $k_p$ =0

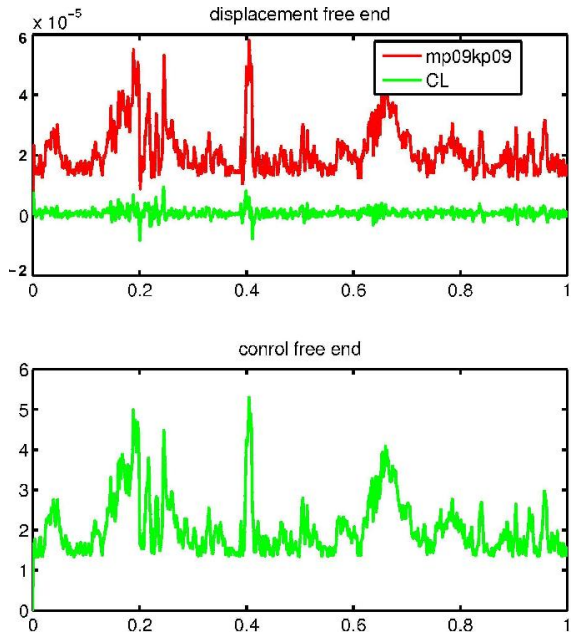



Fig.13 Displacement and control at free end for the $H_\infty$ controller with $m_p$ =0.9, $k_p$ =0

Furthermore, we control the structure with variations of the nominal values of the mass matrix M, stiffness matrix K and matrices A and B. We take into consideration nonlinearities and system dynamics neglected in modeling, incomplete knowledge of disturbances, environment influence in the form of disturbances, and unreliability of system sensor measurements.

## 7 CONCLUSIONS

This paper describes an integrated approach to design and implement robust controllers for intelligent structures. The mathematical model derived using robust control is compared with models obtained by more conventional and well known methods. Using this model, a $H_{infinity}$ ($H_\infty$) controller is designed for vibration suppression purposes. This robust controller accommodates the limited control effort produced by actuators. These designs are all then realized as digital controllers and their closed-loop performances have been compared. In particular, the robustness properties of the controller have been verified for variations in the mass of the test article and the sampling time of the controller. Complete vibration reduction was achieved even for variations of beam mass and stiffness up to 90%. $H_\infty$ controller results were very satisfactory and prove that $H_\infty$ control can reduce smart structures vibrations and deal with modeling uncertainty, external disturbances, and noise in measurements.

## REFERENCES

[1] Bandyopadhyay B., Manjunath TC. & Unapathy M., *Modeling Control and Implementation of Smart Structures,* Springer ISBN-10 3-540-48393-4, 2007.

[2] Burke JV, Henron D, Kewis AS and Overton ML, Stabilization via Nonsmooth, Nonconvex Optimization, *Automatic Control IEE*. Vol. 5, Issue 11, pp. 1760–1769, 2006.

[3] Doyle J C, Glover K, Khargoneker P & Francis B., *State space solutions to standard h2 and h∞ control problems*, IEE Trans. Automatic Control 34, 831–847.

[4] Francis B A., *A course on H∞ control theory*, Springer - Verlag. 1987.

[5] Friedman J and Kosmatka K., An improved two node Timoshenko beam finite element, *J. Computer and Structures*, Vol 47, pp. 473-481, 1993.

[6] Kimura H., Robust stability for a class of transfer functions, *IEE Transactions o Automatic Control* 29, 788–793, 1984.

[7] Miara B, Stavroulakis G and Valente V., *Topics on mathematics for smart systems*, Proceedings of the European Conference. Rome, Italy, 26-28 October 2006.

[8] Moutsopoulou A, Pouliezos A. and Stavroulakis GE. Modelling with Uncertainty and Robust Control of Smart Beams. Paper 35, *Proceedings of the Ninth International Conference on Computational Structures Technology,* B.H.V. Topping and M. Papadrakakis, (Editors), Civil Comp Press, 2008.

[9] Packard A, Doyle J & Balas G. Linear, multivariable robust control with a μ perspective. ASME *Journal of Dynamic Systems, Measurement and Control,* 50th Anniversary Issue, 115(2b), 310–319, 1993.

[10] Stavroulakis GE, Foutsitzi G, Hadjigeorgiou E, Marinova D and Baniotopoulos CC., Design and robust optimal control of smart beams with application on vibrations suppression, *Advances in Engineering Software*, Vol.36, Issues 11-12, pp. 806–813, 2005.

[11] Tiersten HF., *Linear Piezoelectric Plate Vibrations*, Plenum Press New York, 1969.

[12] Yang SM and Lee YJ., Optimization of non collocated sensor, actuator location and feedback gain and control systems, *Smart materials and structures J.* , 8, 96-102, 1993.

[13] Zames G., Feedback minimax sensitivity and optimal robustness, *Autom. Control*, 28, 585–601, 1983.

[14] Zhang N and Kirpitchenko I., Modelling dynamics of a continuous structure with a piezoelectric sensor/actuator for passive structural control, *Journal of Sound and Vibration*, Vol. Comp Press, Stirlingshire Scotland, Civil Comp Press, 2002.

[15] Zhang X, Shao C, Li S and Xu D., Robust H∞ vibration control for flexible linkage mechanism systems with piezoelectric sensors and actuators, *Journal of Sound and Vibration*,243(1),pp. 145-155, 2001.

[16] Zhang N and Kirpitchenko I. Modelling dynamics of a continuous structure with a piezoelectric sensor/actuator for passive structural control, *Journal of Sound and Vibration*, Vol. 249, pp. 251–261, 2002.

[17] Kwakernaak H. ,Robust control and H∞ optimization, *Automatica*, 29(2), pp.255-273, 2003.

[18] Chandrashekara K. & Varadarajan S. , Adaptive shape control of composite beams with piezoelectric actuators, *Intelligent Materials Systems and Structures*, pp.112–124, 1993.

[19] Lim Y. H., Senthil Gopinathan V., Vasundara Varadhan V. & Vijay Varadan K., Finite element simulation of smart structures using an optimalutput feedback controller for vibration and noise control, *Journal of Smart Materials and Structures* 8(8), pp.324–337, 1999.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPUTATIONAL MODELLING OF COLD-FORMED STEEL SCREWED CONNECTIONS AT AMBIENT AND ELEVATED TEMPERATURES

**Luís Mesquita[1], Rui Dias[2], Armandino Parente[2] and Paulo Piloto[3]**

[1] ISISE, Polytechnic Institute of Braganca
Campus de Santa Apolónia, 5300-253 Bragança, Portugal
lmesquita@ipb.pt

[2] Polytechnic Institute of Braganca
Campus de Santa Apolónia, 5300-253 Bragança,, Portugal
{rui_paulinho_93, armandino_parente}@hotmail.com

[3] LAETA-Inegi Polytechnic Institute of Braganca
Campus de Santa Apolónia, 5300-253 Bragança,, Portugal
ppiloto@ipb.pt

**Keywords:** Screwed connections, Cold-formed steel sheet, Elevated temperatures, Experimental tests, Finite element modelling.

**Abstract.** *The application of cold-formed steel structural members in steel construction, and in particular building construction, has a number of advantages that includes its high structural efficiency compared to the member weight. In recent years the increasing safety requirements reflected in the different design standards, boosted the behaviour of materials under extreme conditions, in particular in conditions of high temperatures such as those arising from fires.*

*This work presents a study of the behaviour of cold-formed thin steel sheeting screwed connections at room temperature and elevated temperatures. The shear and bearing failure modes are analysed experimentally by means of a parametric analysis, considering: (i) different elevated temperature values; (ii) cold-formed steel grades; (iii) board effect of the screw position; (iv) and different steel sheet thicknesses. The set of experimental results are used for calibration and verification of the numerical model developed by the finite element method in the software Ansys.*

*The experimental tests presented allowed to determine the load resistance of self-drilling screwed connections of thin steel sheets and the influence of the sheet thickness, edge distance (e1) and the connection temperature. The results show that for the same connection, keeping the sheet thickness and edge distance e1, a change on the failure mode with the temperature increase occur. The finite element numerical model presented allow to study the behaviour of self-drilling screwed connections and determine the collapse load. The cases analysed give results close to the characteristic resistance of the screws manufacturer.*

## 1 INTRODUCTION

The application of cold-formed steel structural members in steel construction, and in particular building construction, has a number of advantages that includes its high structural efficiency compared to the member weight. These sections are usually obtained by thin cold-formed sheets and have slender cross-sections. Although the member resistance is affected by local and distortional buckling instability phenomena, one can reduce these effects by local restraints to the partition walls panels or the building envelop panels, by means of self-drilling or self-tapping screws. Additionally the building erection phase is faster due to the ease steel-to-steel connection between structural members by means of self-drilling screws.

In recent years the increasing safety requirements reflected in the different design standards, boosted the behaviour of materials under extreme conditions, in particular in conditions of high temperatures such as those arising from fires.

The design of cold-formed sections and their connection elements must be performed at room temperature and also in an accidental fire situation, for which the connection strength design value must be determined at elevated temperatures. At room temperature, bolted shear connections from bearing type category must be verified against shear and bearing failure modes. This methodology differs from that presented in EN1993-1- 3, which provides the design resistances of connections for cold-formed members and sheeting. At elevated temperatures the Annex D from EN 1993-1-2 specifies the fire design resistance of bolts loaded in shear as a function of its ambient temperature resistance (from EN1993-1-8) and a reduction factor determined for the appropriate bolt temperature.

In the case of thin steel sheets with thicknesses that typically vary between 0.5 and 3 [mm] screwed connections of these elements are carried out with self-drilling screws without the need of opening a hole, as is shown in Figure 1.



Figure 1 - Self-drilling screws in cold formed profiles.

These screwed connections behave differently than the conventional bolted connections. In the case of self-drilling screwed connections of thin steel sheets, the joint collapse load is usually determined by the bearing resistance of the plates, and only in thicker plates by the screw shear resistance. This distinction is even more noticeable in the case of structures under fire conditions. The heating rate of the thin sheets is higher than the heating rated of the screws, due to their higher thermal capacitance. Therefore, with the increase of temperature, the decrease of the bearing resistance is higher than the screw shear resistance, and may lead to a change of the collapse mode at ambient temperature.

This behaviour has already been studied by some researchers, through experimental tests and numerical simulations at ambient and elevated temperatures. Yan and Young conducted tests on bolted connections of thin sheet steels at elevated temperatures in steady state heating conditions, [1], in which the mechanical load is applied after temperature equilibrium, and also in transient state, [2], in which the connection is subject to a mechanical load and subsequent increase of temperature at constant rate. The authors analysed the effect of the bolt position and

the influence of the bolt diameter. In addition to the test temperature and the sheets thickness, they studied the influence of one and two bolts with different positions and different bolt diameters. The results show four failure modes: namely the bolt shear, bearing, tear out and net section failure.

Lu et al conducted an experimental and numeric study on screwed connections at different temperature levels: 20ºC, 200ºC, 400ºC e 600ºC. The results show two types of failure: (i) thin sheet bearing failure for tests under 200ºC and (ii) screw shear failure for tests at 400ºC and 600ºC. From this results, the authors proposed a design formula based on the EC3-1-8 rules, when the ratio between the distance of the hole to end and the diameter of the hole is between 1.00-1.75 ($1.00 \leq e_2/d \leq 1.75$), [3]. The same authors also performed a numerical study of the connections using the finite element method in Abaqus software, considering an non-linear explicit analysis, [4]. The simulations consider the material and geometry nonlinearity and contact elements between the various surfaces. The bolt geometry is simplified by a simulated thread with three circumferential segments.

Chung and IP conducted a numeric study on thin sheet steel connections, form G300 and G550 sheets, using the finite element software Ansys, [5]. The numerical analyses include the elastoplastic behaviour of the material obtained experimentally and contact elements with a friction coefficient equal to 0.2. The numerical results show that the contact stiffness and frictional coefficient between element interfaces, and clamping force in bolt shanks are important parameters for accurate prediction of bolted connections load-displacement behaviour.

This work presents a study of the behaviour of cold-formed thin steel sheeting screwed connections at room temperature and elevated temperatures. The shear and bearing failure modes are analysed experimentally by means of a parametric analysis, considering: (i) different elevated temperature values; (ii) cold-formed steel grades; (iii) board effect of the screw position; (iv) and different steel sheet thicknesses. The set of experimental results are used for calibration and verification of the numerical model developed by the finite element method in the software Ansys. The numerical method is used for a wider parametric analysis about cold-formed screwed connections at elevated temperatures. The numerical and experimental results are compared with the simplified calculation method presented in the European standards.

## 2 SAFETY VERIFICATION OF BOLTED CONNECTIONS

The design of screwed connections must be performed at room temperature and also in an accidental fire situation, in which it is necessary to know the resistance design value in function of exposure temperature. At room temperature the bolted connections loaded in shear of the bearing type, according to EN1993-1-8, should comply with the safety verifications in relation to their design shear resistance and design bearing resistance. This methodology differs from the one presented in the EN1993-1- 3 for cold-formed member connections, which provides rules for the design of self-drilling screwed connections. At elevated temperatures, the annex D of EN1993-1-2 presents the methodology to design bolted connections for shear and bearing resistance, updating the ambient temperature design resistance by a reduction factor for the appropriate bolt temperature.

### 2.1 Design of bolted connections at ambient temperature.

The design methodology of bolted connections in shear, defined as category A according to EC3 part 1.8, [6], stats that the design ultimate shear load should not exceed the design shear resistance ($F_{v,Ed} < F_{v,Rd}$), nor the design bearing resistance, ($F_{v,Ed} < F_{b,Rd}$).

$$F_{v,rd} = \frac{\alpha_v f_{ub} A}{\gamma_{M2}} \tag{1}$$

$$F_{b,Rd} = \frac{k_1 \alpha_b f_u dt}{\gamma_{M2}} \tag{2}$$

Where $\alpha_v$ represents the factor associated with the bolt class ($\alpha_v$ = 0.6 for classes 4.6, 5.6 e 8.8, $\alpha_v$ = 0.5 for classes 4.8, 5.8 e 10.9). If the shear plane passes through the threaded part of the bolt is assumed $\alpha_v$ = 0.6, $f_{ub}$ represents the bolt ultimate of tensile strength, A the resistant area of the bolt and $\gamma_{M2}$ the partial safety factor for joints. The constant $\alpha_b$ is obtained by the smallest of $f_{ub}/f_u$ , $\alpha_d$ or 1. The value of $\alpha_d$ must be calculated for end bolts and inner bolts independently, by the following expressions, respectively:

$$\alpha_d = \frac{e_1}{3d_0} \tag{3}$$

$$\alpha_d = \frac{p_1}{3d_0} - \frac{1}{4} \tag{4}$$

The factor $k_1$ considers the bolt position perpendicular to the direction of load transfer. For edge bolts assume the smallest of equations (5) and (6) values and 2.5. In the case of inner bolts k1 is given by the smallest value between the equation (6) and 2.5.

$$2.8\frac{e_2}{d_0} - 1.7 \tag{5}$$

$$1.4\frac{p_2}{d_0} - 1.7 \tag{6}$$

Eurocode 3 part 1-3, [7], provides additional rules for safety verification of cold-formed thin steel sheets screwed connections (applied to self-drilling screws). In this case the connections bearing resistance is determined by the following equation.

$$F_{b,Rd} = \frac{\alpha f_u dt}{\gamma_{M2}} \tag{7}$$

The value of α is determined in function of the thickness of the connected plates, by the following expressions.

$$\alpha = 3.2\sqrt{t/d} \le 2.1 \quad for \quad t = t_1 \tag{8}$$

$$\alpha = 3.2\sqrt{t/d} \le 2.1 \quad for \quad t_1 \ge 2.5t \quad and \quad t < 1.0mm \tag{9}$$

$$\alpha = 2.1 \quad for \quad t_1 \ge 2.5t \quad and \quad t \ge 1.0mm \tag{10}$$

$t$ represents the thickness of the thinner plate and $t_1$ the thickness of thicker plate. In the cases not mentioned, α is obtained by interpolation ($t < t_1 < 2.5t$ ). The net section resistance is determined by equation 11.

$$F_{n,Rd} = \frac{A_{net} f_u}{\gamma_{M2}} \tag{11}$$

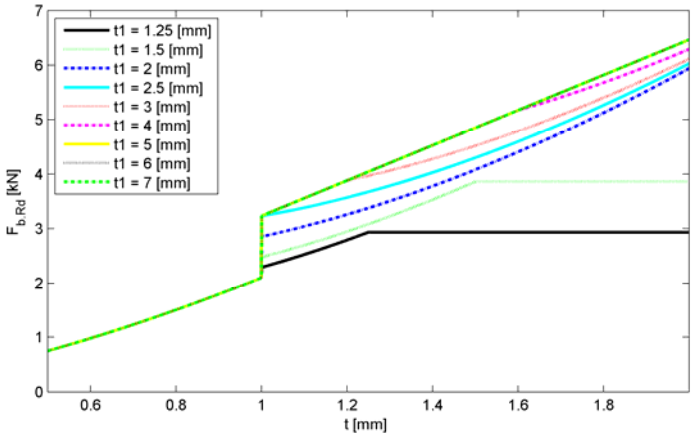



Figure 2 - Bearing resistance according to EN1993-1-3 for different steel sheet thicknesses.

Additionally, it is necessary to check the screw shear resistance. However, this part of the Eurocode defines that this resistance is determined through experimental testing, dividing the characteristic shear resistance ( $F_{v,Rk}$ ) by the partial safety factor. This experimentally determined value, in the case of a single screw connection, is subject to the conditions presented in equation 13.

$$F_{v,Rd} = \frac{F_{v,Rk}}{\gamma_{M2}} \tag{12}$$

$$F_{v,Rd} \geq 1.2 F_{b,Rd} \text{ or } F_{v,Rd} \geq 1.2 F_{n,Rd} \tag{13}$$

The experimental values are usually determined through experimental tests conducted by the screw manufacturers when the product certification is based on an European technical approval (ETA). In the case of the screws used in this study, the shear resistance characteristic values, at room temperature, are presented by SFS INTEC in ETA-10/0198 [8].

## 2.2 Design of screwed connections at elevated temperatures.

The annex D of EN 1993-1-2, [9], presents the methodology to determine the resistance of screwed connections in shear and bearing at at elevated temperatures. The net-section failure at screw holes do not need to be considered if there is a screw in each hole, because the steel temperature is lower at the connection due to the presence of additional material. The shear and bearing resistances are determined by the following equations, in which $F_{v,Rd}$ and $F_{b,Rd}$ are relative to the screw resistance at room temperature.

$$F_{v,t,Rd} = F_{v,Rd} k_{b,\theta} \frac{\gamma_{M2}}{\gamma_{M,fi}} \tag{14}$$

$$F_{b,t,Rd} = F_{b,Rd} k_{b,\theta} \frac{\gamma_{M2}}{\gamma_{M,fi}} \tag{15}$$

$k_{b,\theta}$ represents the reduction coefficient for the appropriate screw temperature. This reduction value is independent of the screw strength and steel class of the connecting plates. Consequently the connection bearing resistance is determined in function of this parameter and its resistance at room temperature. Figure 3 shows the reduction coefficient temperature variation

and its comparison with the ultimate strength ($k_{u,\theta}$) reduction coefficient for the material G250, with $f_{y,0.2} = 292MPa$ and $f_u = 361MPa$, determined by Kankanamge et al [10]. The next figure shows that in temperature range 400-600 [ºC] the reduction coefficient $k_{b,\theta}$ increases the bearing resistance in a non-conservative way.

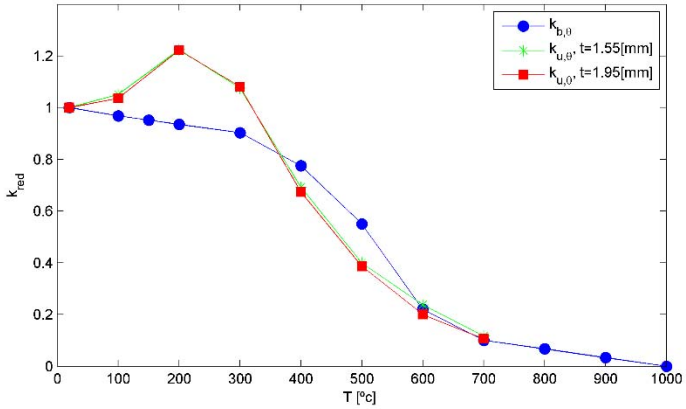



Figure 3 - Bearing resistance reduction value and comparison with the G250 ultimate strength temperature variation. [10].

## 3 EXPERIMENTAL TESTS

### 3.1 Thin steel sheets screwed connections.

This paper presents a set of experimental tests at elevated temperatures on the behaviour of screwed connections of thin steel sheets in simple shear. The galvanized steel sheets are from class DX51D +Z (EN10342) with $f_y = 284[MPa]$ and $f_u = 355[MPa]$. The connections are made with self-drilling carbon steel screws, with the reference SFS SD6-H15 Ø5,5 x 22 mm and a diameter of D=5,5 [mm]. The screw is placed along the middle axis of the sheet and the different distances from the end edge to the hole centre (e1): 10mm, 15mm and 20mm, as presented in Figure 4. A parametric study is made considering the steel sheets thickness (1.5+1.5, 2.0+1.5 and 2.0+2.0 [mm]) and different temperature levels, representative of a fire event.

The tests presented allow the analysis of this parameters on the connection resistance and on the collapse mode.

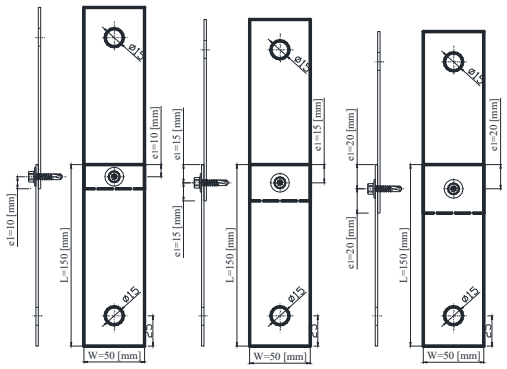



Figure 4 - Dimensions of test specimens for the screwed connections.

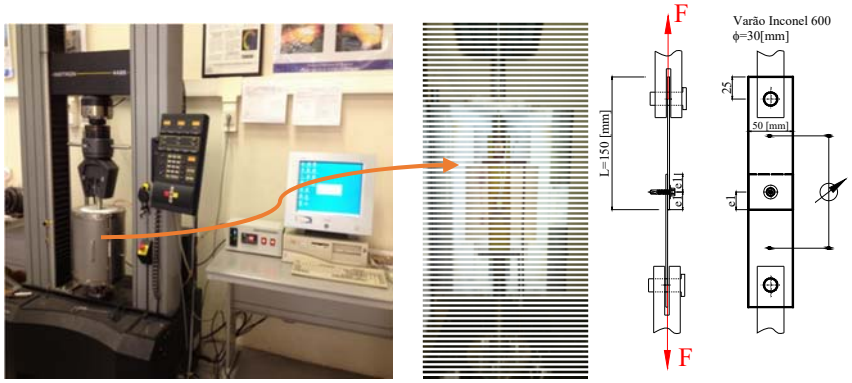



Figure 5 - Experimental setup of the screwed connections.

The experimental setup consists of a universal tensile testing machine to apply the mechanical load, (Figure 5), resulting in a shear effort to the screw. The elements are placed inside of an electric furnace, whose temperature is measured by the thermocouple type K and controlled by a PID system. After the set point temperature is attained, and a small initial load of approximately 0.1 [KN] is applied, the test starts with a displacement control at the velocity equal to 1 [mm/min], following the reference to the technical documentation of the ECCS, [11]. According to this procedure, the failure resistance must be defined as the peak load in a deformation of 3[mm], as represented in Figure 6, using an extensometer in a reference length of 150 [mm]. To measure the reference length extension two auxiliary rods, fixed to steel plates are used, allowing the mechanical extensometer be placed outside the furnace, as can be seen in Figure 5.

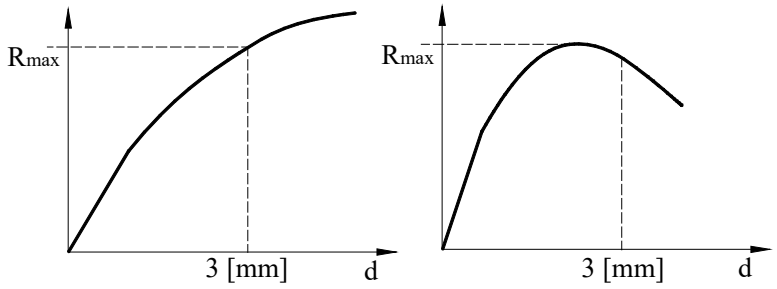



Figure 6 – Experimental tests failure limits, [11].

The failure modes obtained from self-drilling screwed connections are usually one or a combination of two failure modes presented in Figure 7. This figure presents the mechanisms of failure by shearing of the screw (SFM), bearing failure mode (BFM), which can occur with or without tearing of the steel sheet, and screw pull-out and tilting due to the screw slipping from the connection hole (TFM). In the case of thick steel sheets the expected failure mode is due to the screw shear, while for very thin steel sheets the sheet tearing or bearing is verified. Other situations may include two or more combinations of this failure modes.

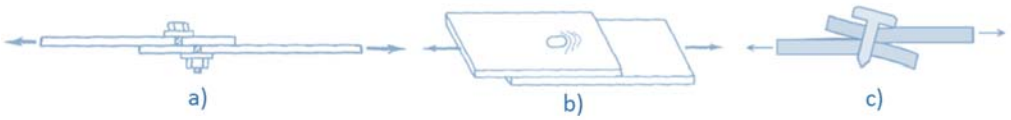



Figure 7 - Failure modes: Shear failure mode (SFM), b) Bearing failure mode (BFM), c) Tilting failure mode (TFM).

Figure 8 presents the experimental results for a combination of sheet thicknesses 1.5+1.5 [mm], considering the applied load in function of the displacement, for different edge distances of the connection and temperature values.

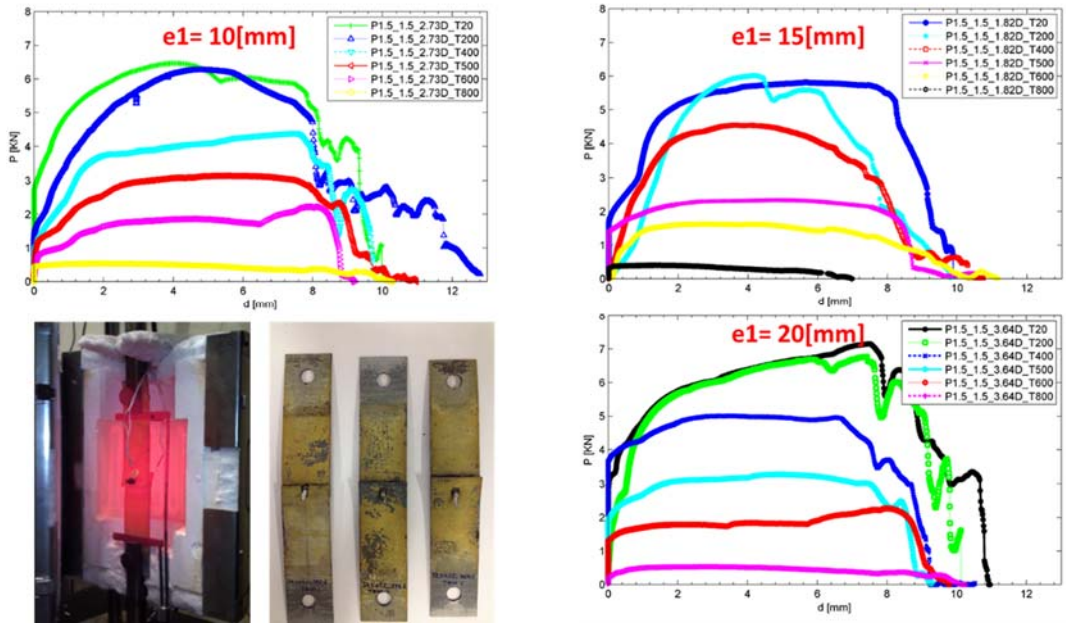



Figure 8 - Experimental results of plates 1.5+1.5 [mm] at a temperature of 20, 200, 400, 500, 600 e 800 [ºC].

In the Table 1 the experimental tests results are presented for the maximum load, the load value for 3 [mm] displacement and the failure mode or combination of modes visualized in each test. The table also present the resistance of each connection determined from the simplified equations EC3-1-8 and EC3-1-3, presented in the previous section, considering unit partial safety factors.

The results show that, as expected, the connection resistance decreases with the temperature increase, with the exception of the 200 [ºC] tests, requiring some additional tests. For the same sheet combination, 1.5+1.5 or 2.0+1.5, the resistance increases with the edge distance e1, independently of the test temperature. However, the table shows that the failure mode is mostly influenced by the edge distance (e1) and the test temperature. For thinner steel sheets (1.5+1.5), when an edge distance e1=10[mm] is used, the failure mode is always due to bearing, while for e1=20 [mm] the most predominant is the tilting failure mode, where the screw rotation and sliding is attained. For the combination of steel sheets 2.0+1.5 with an edge distance e1=20[mm] the predominant failure mode is due to the screw shear, with the exception of the room temperature tests. For steel sheet thicknesses 2.0+2.0 at a temperature of 500[ºC], the failure mode is always do to shear, regardless of the edge distance value.

### 3.2 Mechanical strength of the screws steel

The usual process of design bolted connections is based on the mechanical strength of bolts and plates material. In the case of conventional bolts, the minimum value of the bolts nominal strength (yield strength and ultimate strength) is defined by its material class, established by the standard ISO898-1:1999 [12]. In the case of self-drilling screws SFS INTEC SD6-H15 Ø5,5 x 22 mm a classification from this standard is unknown. According to the manufacturer, these screws are produced by cold forming a length of a steel wire from steel 18B2 according to the

EN10263. The threaded length is opened by cold working followed by a hardening heat treatment and stress relief. Finally, there is a surface treatment and dehydrogenation.

According to the manufacturer's experimental tests of the screws, the bolt tension resistant and shear resistance are 16 and 10 [KN], respectively, [13].

| Reference | $P_{MAX}$ [kN] | $P_{[d=3mm]}$ [kN] | Failure mode | EC3-1-8 | | EC3-1-3 | |
|---|---|---|---|---|---|---|---|
| | | | | Fb,t,Rd | Fv,t,Rd | Fb,t,Rd | Fv,t,Rd |
| P1.5+1.5_1.82d_T20 | 5.820 | 5.514 | BFM | 4.438 | 6.808 | 4.894 | 3.430 |
| P1.5+1.5_1.82d_T200 | 6.008 | 5.659 | BFM | 5.432 | 8.332 | 5.991 | 4.198 |
| P1.5+1.5_1.82d_T400 | 4.537 | 4.435 | BFM | 3.075 | 4.718 | 3.392 | 2.377 |
| P1.5+1.5_1.82d_T500 | 2.336 | 2.303 | BFM | 1.771 | 2.716 | 1.953 | 1.369 |
| P1.5+1.5_1.82d_T600 | 1.622 | 1.616 | BFM+TFM | 1.052 | 1.613 | 1.160 | 0.813 |
| P1.5+1.5_2.73d_T20 | 6.459 | 6.266 | TFM+SFM | 6.656 | 6.808 | 4.894 | 3.430 |
| P1.5+1.5_2.73d_T200 | 6.260 | 5.686 | BFM | 8.147 | 8.332 | 5.991 | 4.198 |
| P1.5+1.5_2.73d_T400 | 4.387 | 3.871 | TFM | 4.613 | 4.718 | 3.392 | 2.377 |
| P1.5+1.5_2.73d_T500 | 3.141 | 2.964 | TFM | 2.656 | 2.716 | 1.953 | 1.369 |
| P1.5+1.5_2.73d_T600 | 2.223 | 1.777 | SFM | 1.578 | 1.613 | 1.160 | 0.813 |
| P1.5+1.5_3.64d_T20 | 7.152 | 6.051 | TFM | 7.322 | 6.808 | 4.894 | 3.430 |
| P1.5+1.5_3.64d_T200 | 6.765 | 6.003 | TFM | 8.962 | 8.332 | 5.991 | 4.198 |
| P1.5+1.5_3.64d_T400 | 5.004 | 4.999 | TFM | 5.074 | 4.718 | 3.392 | 2.377 |
| P1.5+1.5_3.64d_T500 | 3.281 | 3.125 | TFM | 2.921 | 2.716 | 1.953 | 1.369 |
| P1.5+1.5_3.64d_T600 | 2.266 | 1.766 | SFM | 1.735 | 1.613 | 1.160 | 0.813 |
| P2.0+1.5_1.82d_T20 | 5.750 | 5.101 | BFM | 4.438 | 6.808 | 5.173 | 4.310 |
| P2.0+1.5_1.82d_T200 | 6.631 | 6.325 | BFM | 5.432 | 8.332 | 6.332 | 5.275 |
| P2.0+1.5_1.82d_T400 | 4.977 | 4.940 | BFM | 3.075 | 4.718 | 3.585 | 2.987 |
| P2.0+1.5_1.82d_T500 | 3.001 | 2.996 | BFM | 1.771 | 2.716 | 2.064 | 1.720 |
| P2.0+1.5_1.82d_T600 | 1.670 | 1.664 | BFM+SFM | 1.052 | 1.613 | 1.226 | 1.021 |
| P2.0+1.5_2.73d_T20 | 8.086 | 6.636 | SFM | 6.656 | 6.808 | 5.173 | 4.310 |
| P2.0+1.5_2.73d_T200 | 8.666 | 7.452 | SFM | 8.147 | 8.332 | 6.332 | 5.275 |
| P2.0+1.5_2.73d_T400 | 5.702 | 5.380 | TFM+SFM | 4.613 | 4.718 | 3.585 | 2.987 |
| P2.0+1.5_2.73d_T500 | 3.511 | 3.485 | TFM+SFM | 2.656 | 2.716 | 2.064 | 1.720 |
| P2.0+1.5_2.73d_T600 | 1.863 | 1.788 | TFM+SFM | 1.578 | 1.613 | 1.226 | 1.021 |
| P2.0+1.5_3.64d _T20 | 8.274 | 6.481 | TFM | 7.322 | 6.808 | 5.173 | 4.310 |
| P2.0+1.5_3.64d _T200 | 7.501 | 7.114 | SFM | 8.962 | 8.332 | 6.332 | 5.275 |
| P2.0+1.5_3.64d _T400 | 4.730 | 4.515 | TFM+SFM | 5.074 | 4.718 | 3.585 | 2.987 |
| P2.0+1.5_3.64d _T500 | 3.458 | 3.426 | TFM+SFM | 2.921 | 2.716 | 2.064 | 1.720 |
| P2.0+1.5_3.64d _T600 | 1.885 | 1.681 | TFM+SFM | 1.735 | 1.613 | 1.226 | 1.021 |
| P1.5+1.5_1.82d_T800 | 0.397 | 0.344 | BFM | 0.186 | 0.285 | 0.205 | 0.144 |
| P1.5+1.5_2.73d_T800 | 0.537 | 0.526 | TFM | 0.279 | 0.285 | 0.205 | 0.144 |
| P1.5+1.5_3.64d_T800 | 0.532 | 0.532 | TFM | 0.307 | 0.285 | 0.205 | 0.144 |
| P2.0+2.0_1.82d_T20 | 8.059 | 6.196 | SFM | 5.917 | 6.808 | 7.535 | 4.310 |
| P2.0+2.0_2.73d_T20 | 8.977 | 5.981 | SFM | 8.875 | 6.808 | 7.535 | 4.310 |
| P2.0+2.0_3.64d_T20 | 9.364 | 7.060 | SFM | 9.763 | 6.808 | 7.535 | 4.310 |
| P2.0+2.0_1.82d_T500 | 3.474 | 3.474 | SFM | 2.361 | 2.716 | 3.007 | 1.720 |
| P2.0+2.0_2.73d_T500 | 3.640 | 3.452 | SFM | 3.541 | 2.716 | 3.007 | 1.720 |
| P2.0+2.0_3.64d_T500 | 3.839 | 3.650 | SFM | 3.895 | 2.716 | 3.007 | 1.720 |

Table 1 - Characteristics and results of the experimental tests (maximum load, load at 3 [mm] of deformation and failure mode).

With the aim of establish the classification of the screw material strength, a set of tensile experimental tests were carried out until the material rupture, using a universal testing machine. The screws were previously machined to remove the threads and to create a constant reference length with 6 [mm] of length and 3 [mm] of diameter. An increasing load was applied to the screw between its head and a drilled 4 [mm] thick sheet, as shown in Figure 9. For the tests performed at elevated temperatures this setup was placed inside an electric ceramic furnace.

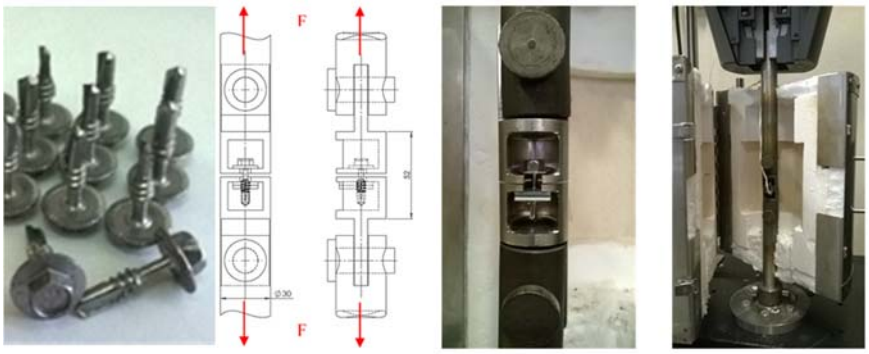



**Figure 9 -** Experimental setup for traction test.

Table 2 presents the average values of the experimental results of maximum load and its tensile strength for different temperature levels. The results show a noticeable reduction of the screw ultimate strength with temperature. The average results at room temperature (from a set of three) allow us to conclude that the screw meets the requirements of the minimum tensile strength for the class 10.9.

| Temp. [ºC] | $P_{MAX}$ [kN] | Ultimate Strength [Mpa] |
|---|---|---|
| 20 | 8,003 | 1135,67 |
| 200 | 7,952 | 1125,00 |
| 400 | 5,326 | 753,50 |
| 500 | 3,157 | 443,9 |

Table 2 – Average value of the screws tensile tests.

## 4 NUMERICAL MODELLING OF THIN SHEET SCREWED CONNECTIONS

The structural model used in the numerical simulations, by the finite elements method, is presented in Figure 10. This model follows the experimental setup used in the previous sections, in which one end of the steel sheet is fixed and the other is subjected to an increasing mechanical load until the collapse load is achieved. The simulations are performed in the software Ansys® considering geometric and material non-linear analysis. The structural steel elements, including the screw, are modelled using three-dimensional solid finite elements (Solid186 and Solid187) and the contact surfaces between screw-sheets and sheet-sheet modelled with contact elements (Conta174 and Targe170). All contact surfaces consider frictional contact with a friction coefficient equal to 0.2, with an Augmented Lagrange Formulation, and a refined mesh to obtain the required numeric precision of the contact between the contact and target finite elements.
The material nonlinear behaviour of the steel sheets was considered using the stress-strain curves presented in the Eurocode 3 part 1.2 at high temperatures, [9]. These curves are determined in function of the material characteristic yield stress and the temperature reduction coefficients therein presented.

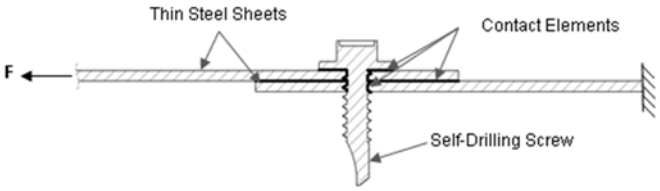



Figure 10 - Geometric model used in the numerical simulations.

In the case of thin steel sheets, being considered of class 4, the design yield strength and its variation with temperature must be considered equal to the proportional limit strength at 0.2% deformation. For this class the temperature reduction coefficients of yield stress at 2% are not presented. This requires a modification of stress-strain curves design presented by Eurocode 3 part 1-2, consisting in an iterative process for calculating the 2% reduction coefficients, and its yield strength in function of the temperature, so that the proportional limit strength at 0.2% match the ones from EN1993-1-2 annex E, [9].

The nonlinear material analysis performed in Ansys defined by an elastic-plastic material law, needs the stress-strain curves to be converted into true stress-true strain curves. This transformation is accomplished by the following equations, where the nominal stresses and strain values are determined by the nominal curves provided by EC3-1-2.

$$\begin{aligned} \varepsilon_{true} &= ln(1+\varepsilon_{nom}) \\ \sigma_{true} &= \sigma_{nom}(1+\varepsilon_{nom}) \end{aligned} \tag{16}$$

The Figure 11 presents the true stress-strain curves for different temperature values of galvanized steel sheets S350GD + Z, with a yield stress at 0.2% equal to 350 [MPa], and considering strain hardening up to 400 [ºC].

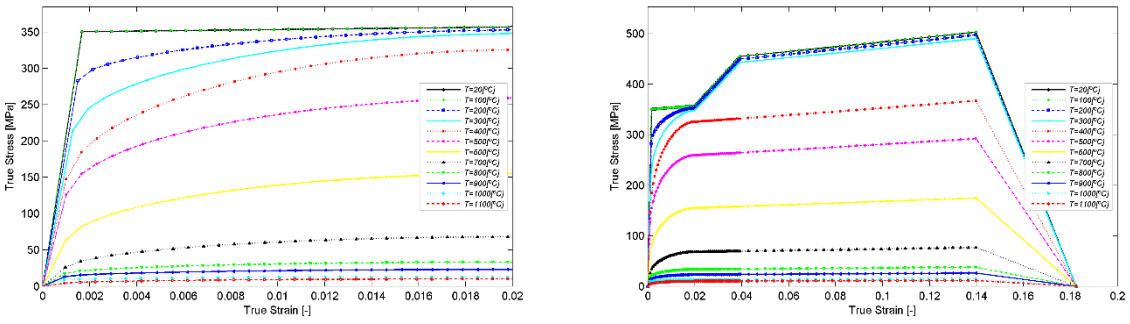



Figure 11 - True stress-strain curves for different temperatures.

As this self-drilling screws are not classified as the conventional bolts, to allow the material characterization a set of tensile tests were done to the screws in tension. The average ultimate strength value obtained was 1135.67 [MPa], allowing the screw to be classified as class 10.9, from ISO898-1:1999 [12]. The screw material stress strain curves follow a bilinear model defined by the yield strength and a tangent modulus equal to E/100.

### 4.1 Numerical simulations.

The simulations performed at room temperature follow a two step analysis. Firstly, without mechanical load (F=0), the equilibrium conditions of the contact pressure and contact conditions between the various surfaces are determined. As the experimental screw preload was not

determined, this was not included in the numerical analysis. The second step consists in applying a linearly increasing load, with minimum increments of 1 [N], while the equilibrium condition are attained, obtaining the connection collapse load at the end of the simulation.

Figure 12 presents the finite element model of two thin steel sheets (1+2 [mm]) screwed connection considering an edge distance equal to 20 [mm]. To reduce the finite element model and its computational time a symmetry condition was used. The Von Mises equivalent stress distribution show the yielding of the sheets due to bearing, with stress values higher than the yield stress, and a small plastic area along the screw thread, with stresses over 900 [MPa]. The connection collapse load applied at this instant is 4693.4 [N]. The numerical results from simulations performed with a steel sheets combination equal to 2+2 [mm] allow to obtain a collapse load of 5600 [N].

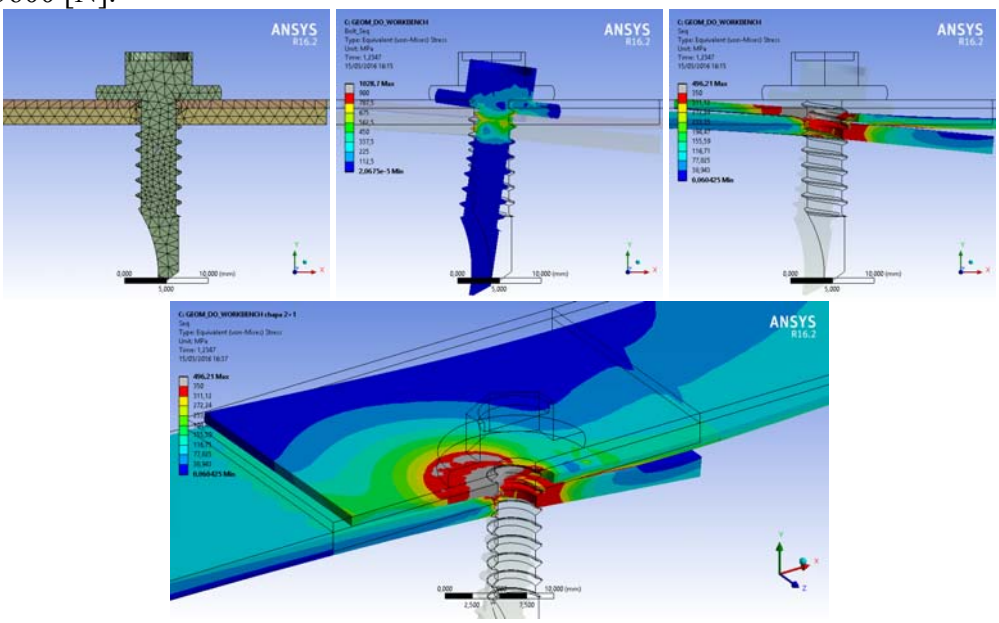



Figure 12 - Finite element model of steel sheets 2+1 [mm]. Steel sheets and screw von Mises equivalent stress distribution at the collapse load.

## 5 CONCLUSIONS

- The experimental tests presented allowed to determine the load resistance of self-drilling screwed connections of thin steel sheets and the influence of the sheet thickness, edge distance (e1) and the connection temperature.
- For the same connection, keeping the sheet thickness and edge distance e1, we can observe a change on the failure mode with the temperature increase. The three failure modes were detected, being mainly the screw shear failure, steel sheets bearing failure and a combination of both.
- Connections with double steel sheets of 2.0 [mm] have a failure mode due to screw shear at ambient temperature and at 500 [ºC]. Connections with double steel sheets of 1.5 [mm], the failure mode varies with the edge distance e1 and changes from steel sheets bearing failure to a combination of tilting and screw shearing at higher temperatures.
- The finite element numerical model shown allow to study the behaviour of self-drilling screwed connections and determine the collapse load. The cases analysed give results close to the characteristic resistance of the screws manufacturer.
- Further experimental tests are being planned for a parametric analysis of the steel sheet material class (SGD220, SGD280, SGD320 and SGD350) with different sheet thicknesses.

## ACKNOWLEDGEMENTS

The authors acknowledge the support and the supply of materials used in the study from the companies SFS INTEC, LUSOSIDER and IRMALEX.

## REFERENCES

[1] Yan, S. and B. Young, Tests of single shear bolted connections of thin sheet steels at elevated temperatures—Part I: Steady state tests. Thin-Walled Structures, 2011. **49**(10): p. 1320-1333.

[2] Yan, S. and B. Young, Tests of single shear bolted connections of thin sheet steels at elevated temperatures—Part II: Transient state tests. Thin-Walled Structures, 2011. **49**(10): p. 1334-1340.

[3] Lu, W., et al., Design of screwed steel sheeting connection at ambient and elevated temperatures. Thin-Walled Structures, 2011. **49**(12): p. 1526-1533.

[4] Lu, W., et al., Behaviour of shear connectors in cold-formed steel sheeting at ambient and elevated temperatures. Thin-Walled Structures, 2012. **61**: p. 229-238.

[5] Chung, K.F. and K.H. Ip, Finite element modeling of bolted connections between cold-formed steel strips and hot rolled steel plates under static shear loading. Engineering Structures, 2000. **22**(10): p. 1271-1284.

[6] Portugal. Instituto Português da Qualidade, Eurocódigo 3 projecto de estruturas de aço Parte 1-8 projecto de ligações NP EN 1993-1-8: 2010. 2010, Caparica: IPQ. 146 p.

[7] CEN, EN1993-1-3 Eurocode 3 : Design of steel structures, Part 1-3: General rules Supplementary rules for cold-formed members and sheeting. 2004, European Committee for Standardization: Brussels, Belgium.

[8] EOTA, ETA-10/0198 Fastening screws for metal members and sheeting, Fastening screws SFS, SFS intec AG. 2013: European Organisation for Technical Approvals.

[9] IPQ, Eurocódigo 3 projecto de estruturas de aço Parte 1-2 regras gerais. Verificação da resistência ao fogo NP EN 1993-1-2: 2010, ed. Instituto Português da Qualidade. 2010, Caparica: IPQ. 87 p.

[10] Kankanamge, N.D. and M. Mahendran, Mechanical properties of cold-formed steels at elevated temperatures. Thin-Walled Structures, 2011. **49**(1): p. 26-44.

[11] ECCS-TC7, ECCS TC 7: The Testing of Connections with Mechanical Fasteners in Steel Sheeting and Sections. Vol. Nº 124. 2009, Brussels: Brussels.

[12] Standardization, E.C.f., EN ISO 898-1:1999 - Mechanical properties of fasteners made of carbon steel and alloy steel - Part 1: Bolts, screws and studs. 1999.

[13] SFS-INTEC, Technical performance values for SFS fasteners, SFS-INTEC, Editor. 2000: Prestbury.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# MONITORING AND ANALYSIS OF STRESS FIELD FOR ORTHOTROPIC STEEL DECK OF DASHENGGUAN YANGZTE RIVER BRIDGE (ECCOMAS CONGRESS 2016)

**Y. Wang[1], Y.S. Song[2]**

[1] Jiangsu Key Laboratory of Engineering Mechanics, Southeast University
2# Sipailou Rd., Xuanwu District, Nanjing 210096, China
Civil_wangying@seu.edu.cn

[2]Jinling Institute of Technology,
99 Hongjing Avenue, Jiangning District, Nanjing 211129, China
646814724@qq.com

**Keywords:** Stress Field, Fatigue Life, High-speed Railway Bridge, Structural Health Monitoring

**Abstract:** *Driving security of trains with high speed is one of pivotal issues influenced by the smoothness of rail lines significantly. For high-speed railway steel bridges, fatigue cracking of typical positions under the rail lines will reduce stiffness of the deck plate inducing the smooth of the rail lines. In this paper, the strain data recorded by the structural health monitoring system is analyzed and a new theoretical approach is proposed by integrating fatigue assessment of typical welded joint for high-speed railway steel bridge in service. The theoretical approach includes various factors including stress concentration, environmental corrosion and train flow. Having applied in an actual high-speed railway bridge –Dashengguan Yangzte River Bridge, the approach was converted to conduct determinative and reliable fatigue-life evaluation of the bridge. Static-load experiment was conducted to obtain the stress distribution around the weld, as well as the distinction of stress between the positions of weld and strain gauge. The influence of environmental corrosion was integrated by two aspects: increase of stress and reduce of the fatigue resistance. Without consideration of train-weight growth, in conclusion, the fatigue life of typical welded joint will be infinite even if the train flow is with growth constantly. In addition, fatigue life of typical welded joints, determinative or reliable, may be less than the designed service period when growth rate of the train weight arrives at 5%.*

## 1 INTRODUCTION

High-speed train emerges to satisfy passengers' need of shorter time between distant cities. Trains in high speed are likely to suffer cruel accidents resulting in mayor loss of life and property [1,2]. A mayor concern about the railway transportation system is whether these as-built high-speed railway bridges are robust enough to carry the ever-increasing transportation. Considerable railway bridges were built in steel for lighter weight and easier construction [3,4]. However, steel material and structure are vulnerable to suffering corrosion and fatigue issues especially in welded positions. Thus, subjected to repeated load with ever increasing, there is an urgent need to assess the fatigue performance, remaining fatigue life or reliability combining environmental corrosion.

Fatigue performance assessment requires actual information on structural response of strain at welded position [5-7]. Although Structural Health Monitoring System (i.e. SHMS) has provided strain historical information, proper consideration has to be given involving stress concentration in welds, environmental corrosion and consistent growth of train flow in the future [8,9]. Limited by the feasible positions for strain gauges, measured strain data is incapable of acquiring the strain closing to the weld. So, laboratory test has to be conducted to obtain the stress field around the weld, solving the problems of distinction between the in-situ measurement and actual stress state. In addition, various investigations have been conducted on reduce of cross section by environmental corrosion. But, it is not remarkable to consider the influence of corrosion on fatigue resistance of steel material.

Besides the determinative fatigue-life evaluation of welded joints, reliable fatigue-life evaluation has become much more noticeable, with the growing attention on uncertainty of material, geometry, environmental and loading actions [10,11]. Because of the variation of influencing factors with time, time-dependent reliability assessment becomes essential for full-cycle evaluation of fatigue performance [12]. SHMS has been proved to be one of essential measures to conduct reliability evaluation for actual projects [13]. However, additional investigations require being carried out for the restriction of sensor-system scale and environmental conditions. Mechanical-property and material-resistance experiments are likely to help completion and improvement for the in-situ evaluation of projects.

In this paper, a new theoretical approach is proposed by integrating fatigue assessment of typical welded joint for high-speed railway steel bridge in service. The approach is developed by considering multiple influencing factors - stress concentration, environmental corrosion and growth of train flow. Structural health monitoring system is introduced to provide essential data for fatigue evaluation. Static-load experiment is adopted to obtain the information of stress concentration between the welds and strain-sensor locations. In addition, the relation of the stress reducing and environmental corrosion is derived to coincide with the actual property of plate structure. Finally, determinate and reliable fatigue-life evaluations are conducted by adopting the recommended approach, applying to a large high-speed railway bridge – the Dashengguan High-speed Railway Bridge in China.

## 2 STRAIN MONITORING SYSTEM OF DASHENGGUAN HIGH-SPEED RAILWAY BRIDGE

Dashengguan High-speed Railway Bridge (Fig. 1 (a))is located in Nanjing city of Jiangsu province in China, spanning the Yangtze River which is the longest river in China. Two national railway lines were conducted and run through the bridge. These two railway lines are respectively Beijing-Shanghai High-Speed Railway and Shanghai-Hankou-Chengdu Railway

and the design speed is 300 km/h and 250km/m. Two kinds of trains serve for these two main rail lines with respectively and 16 carriages. As one kind of steel trussed arch bridge, the vertical view of the bridge is presented in Fig. 1(b) with the main span of 336m. To obtain higher stiffness reducing the risk of train derailing, orthotropic steel deck is applied with internal crossbeams distributed longitudinally.

To acquire structural performance of the railway bridge, structural health monitoring system was installed in construction period. Strain gauges were equipped near rib-to-deck-joint locations for further evaluation of fatigue performance as the orthotropic steel deck is prone to fatigue cracking. The locations of strain gauges are shown in Fig. 1(c). There are two strain gauges in the middle cross section of orthotropic steel deck, which are labeled as DYB-11-23 and DYB-11-24. The strain gauge DYB-11-23 belongs to the side of Beijing-Shanghai High-Speed Railway and DYB-11-24 is on the side of Shanghai-Hankou-Chengdu Railway.



Fig. 1(a) Landscape of the whole bridge

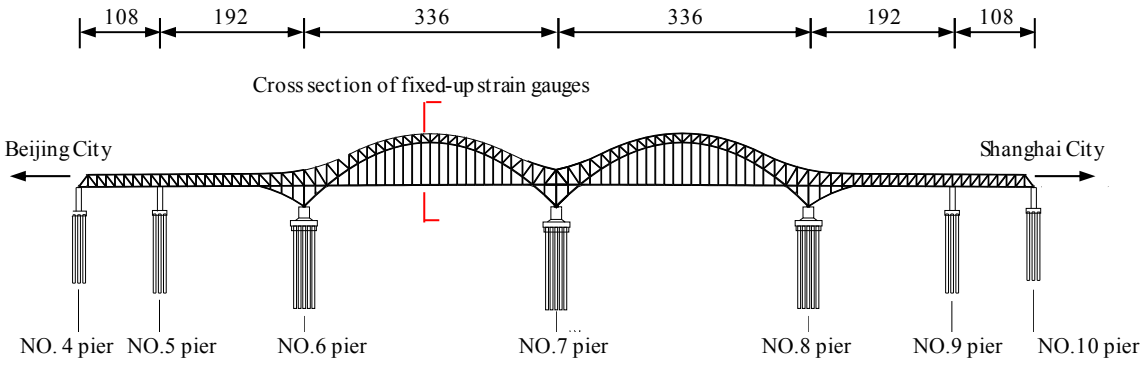



Fig. 1(b) Side elevation of the main bridge (unit: m)

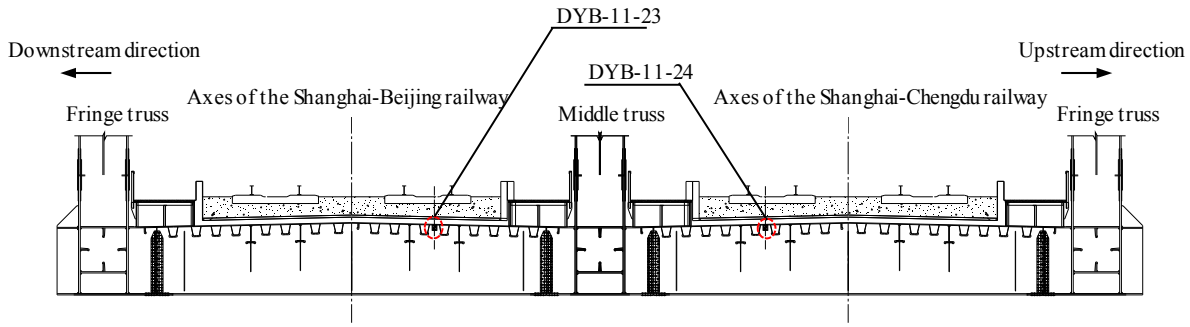



Fig. 1(c) Cross-section diagram of strain gauges

Fig. 1: Layout diagram of strain gauges for steel deck

# 3 MONITORED STRAIN DATA AND TWO CALCULATED FATIGUE EFFECTS

## 3.1 Strain history curve under typical trains

In this paper, strain history data is introduced for fatigue evaluation from February to July in 2013. It is presented by Fig. 1 (c) that the strain-gauge locations of DYB-11-23 and DYB-11-24 are symmetrical. Thus, the monitored strain data of DYB-11-24 was merely applied to conduct fatigue analysis. Fig. 2 presents two typical strain history curves of DYB-11-24 under single train. These two curves were induced by typical trains with respectively 8 and 16 carriages. The maximum value of changes in strain is within 2-6 με during the passing of train with8 and 16 carriages. For the train with 8 carriages, the number of the strain peaks is 16 while the same value for the train with 8 carriages is about 32. Considering the number of wheels for each carriage is 2, the number of the strain peaks corresponds to the number of total wheels.

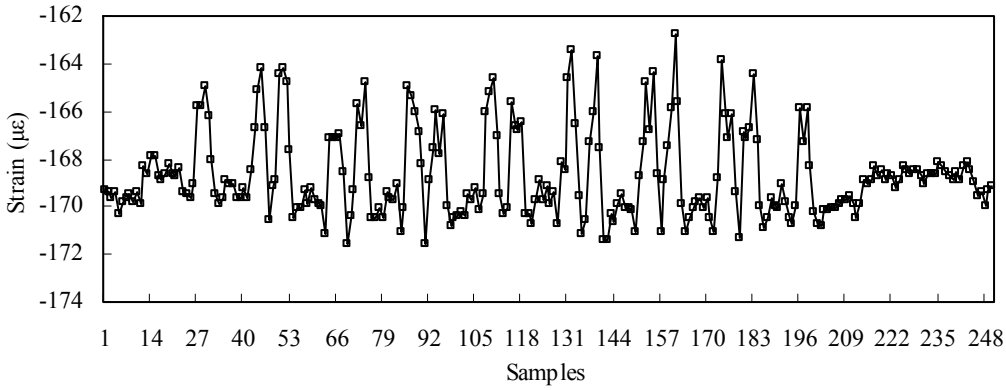



Fig. 2 (a) Typical strain history curve during the passing of train with 8 carriages

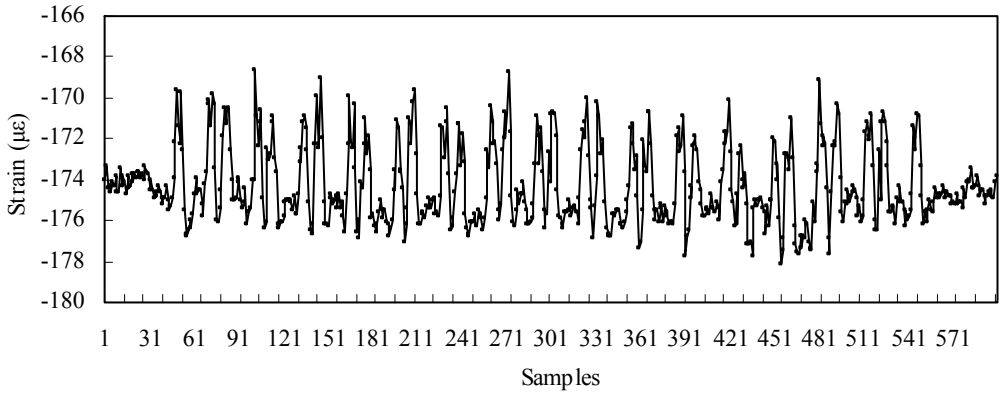



Fig. 2(b) Typical strain history curve during the passing of train with 16 carriages

Fig. 2: Typical strain history curves of strain gauge DYB-11-24 during the passing of typical trains

## 3.2 Stress amplitude and stress cycle number

Firstly, stress curves were obtained by multiplying0.206 to strain data, in which the units of stress and strain are respectively MPa and με. In addition, 0.206 is the value of elastic modulus of steel in unit of GPa. Stress amplitude and stress cycle number are two typical parameters for evaluating fatigue performance of welded joint. Secondly, rain-flow method was introduced to process the calculated stress history data and stress amplitude spectrum was obtained for further investigation. Fig. 3 presents the stress amplitude spectrum of two typical trains with 8 and 16 carriages. In the figure, the scope of stress amplitudes smaller than 0.3MPa occupying the vast majority is induced by interferences with randomness. Thus, the

stress amplitudes smaller than 0.3MPa were eliminated for further calculation conduction. Finally, two fatigue parameters of Seq (equivalent stress amplitude) and Ns (stress cycle number) under single train were calculated as follows based on the Palmgren and Miner theory combined with standard of Eurocode 3:

$$S_{eq} = \left( \frac{n_i S_i^5}{\sum n_i} \right)^{1/5} \tag{1}$$

$$N_s = \sum n_i \tag{2}$$

where Si is the ith stress amplitude, ni is the number of stress cycle corresponding to Si and 1/5 is slope value of logS-logN curve corresponding the scope of stress amplitude in Eurocode 3. Fig. 4 presents the series of fatigue S-N curves recommended by Eurocode 3. In the figure, the type of rib-to-deck joints corresponds to Type 71 with ΔσD equals to 52 MPa.

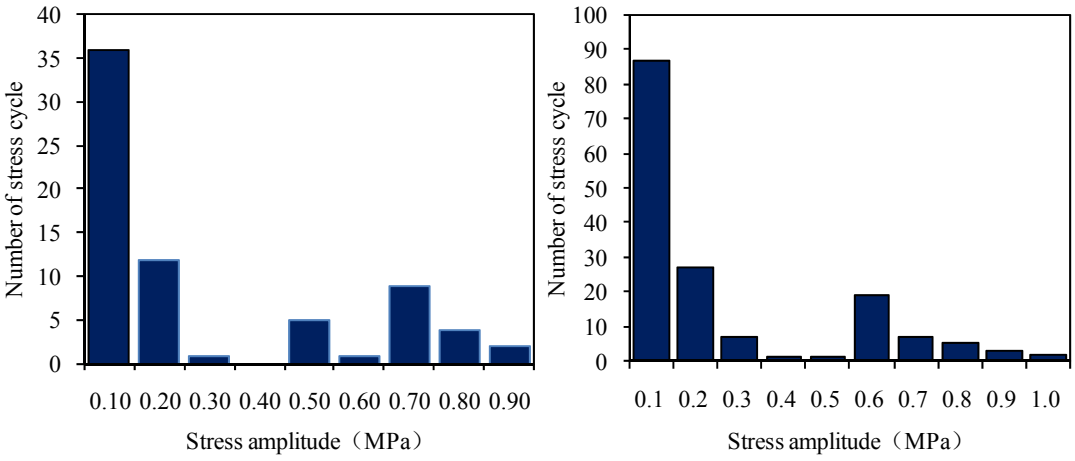



Fig. 3 (a) 8-carriage train Fig. 3 (b) 16-carriage train

Fig. 3: Stress amplitude spectrum of DYB-11-24 under two typical trains

Table 1 presents the calculated results of Seq and Ns for typical stress history curve under single train with 8 and 16 carriages. The values of Seq are similar by comparison of trains with 8 and 16 carriages. The difference ratio between these two trains is only 5.6%. However, the difference of Ns between these two types of trains is much more significant. As the ratio of Ns between two trains is about 2.0, the number of stress cycle is certainly linearly relative with the number of train carriages.

| **Type of train** | ***Seq* (MPa)** | ***Ns*** |
|---|---|---|
| 8-carriages | 0.75 | 22 |
| 16- carriages | 0.71 | 45 |

Table 1**:** Comparative results of fatigue effects under two typical trains

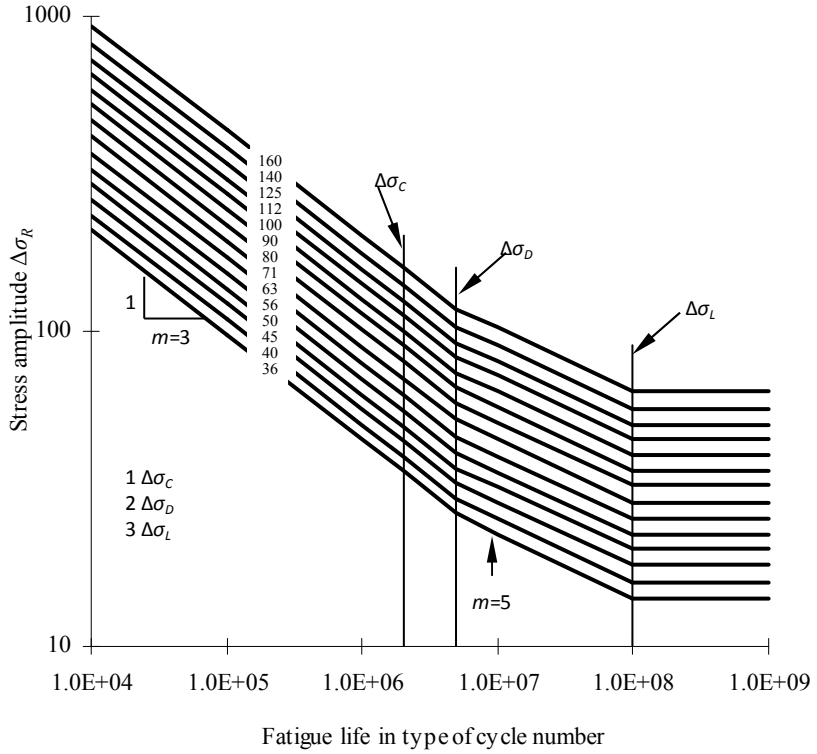



Fig. 4: Series of fatigue S-N curves recommended by Eurocode 3

Time-dependent rule of two fatigue effects is investigated by processing the strain data from February to July in 2013. The processing time unit of the strain data is defined as one day. The treating procedure is similar to presentation in the previous paragraph. Thus, the resultant fatigue effects are so-called daily Seq and daily Ns. Fig. 5 presents the curves of daily Seq and daily Ns from February to July in 2013. The value of daily Seq is similar between various days as the distribution is confined to the scope between 0.67 MPa and 0.80 MPa. However, daily Ns in the first half of the February is much greater than other months. In this period, the train flow is much larger because of the nationally shot rush called Spring Festival Travel in China.

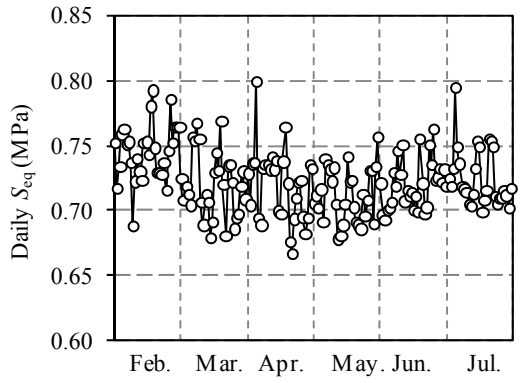



Fig. 5 (a) Daily stress amplitude

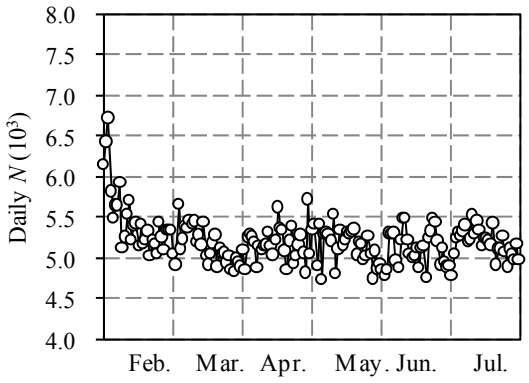



Fig. 5 (b) Daily stress cycle number

Fig. 5: Time-varying curves of two fatigue effects from February to July in 2013

# 4 MUTI-FACTORS INFLUENCING FATIGUE PERFORMANCE

## 4.1 Stress concentration

It is presented in the section 3 that the monitored data of strain is capable of obtaining the typical factors accounting fatigue performance. However, accurate evaluation is merely achieved when sufficient causes are taken account into consideration (Ma et al. 2010). Restricted from the field environment and sensor's capability, strain gauges have to be only installed from the weld with certain distance. Thus, the recorded monitoring strain data is much smaller than the actual where approaching the weld toe. Near the weld toe, the stress is called hot-spot stress and the recorded monitoring stress is regard as the nominal stress. Currently, the difference between them is profiled by the parameter called Stress Concentration Factor (SCF) which is described as follows:

$$SCF = S_{hot} / S_{nom} \tag{3}$$

where Shot is hot-spot stress and Snom is nominal stress.

Static-load experiment is conducted to obtain SCF for rib-to-deck welded joint. The geometry of experimental specimen is uniform with actual steel deck of Dashengguan Railway Bridge. The specific dimension of the experimental specimen is presented in Fig. 6 and Table 2. In addition, the WJ-X# indicates the rib-to-deck welded joints of the specimen. Resistance strain gauges are applied to measure the stress distribution around the deck plate under typical load level. The layout of the strain gauges is presented in Fig. 7. The label of DTX and DBX indicates the stain gauges located on the deck top and deck bottom. In addition, the label of RBX indicates the stain gauges on the rib bottom approaching the weld. Typical two-point bending experiment is adopted as static-load mechanism. Fig.8 presents the loading equipment and mechanism for static-load experiments. During the region between the loading points, the bending moment keeps constant and the deck-top/bottom stress at the middle position between neighboring welds is regard as the nominal stress. Then, the stress approaching the weld toe is considered to be the hot-spot stress.

| Group | Specimen Number | Deck plate | | | U-shaped rib | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Length | Width | Thickness | Opening width | Flange width | Height | Thickness | Spacing |
| 1 | OSD1-OSD4 | 2100 | 300 | 16 | 300 | 160 | 280 | 8 | 600 |

Table 2: Typical geometrical parameters of experimental specimen (unit: mm)

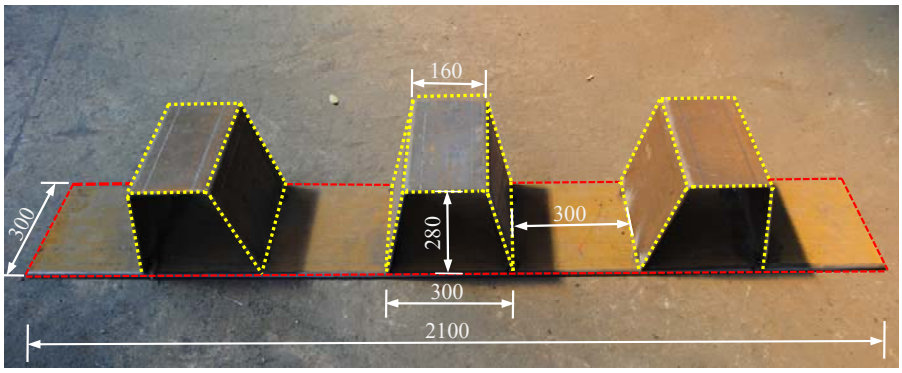



Fig. 6: Geometric parameters of experimental member (unit: mm)

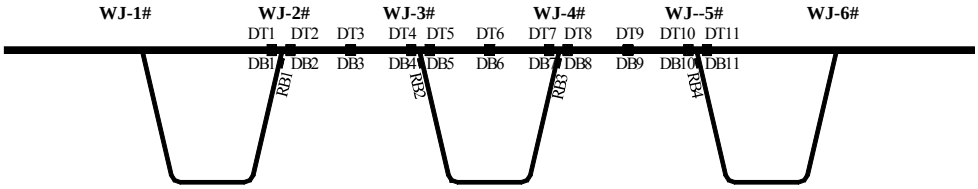



Fig. 7: Layout of strain gauges for experimental specimen

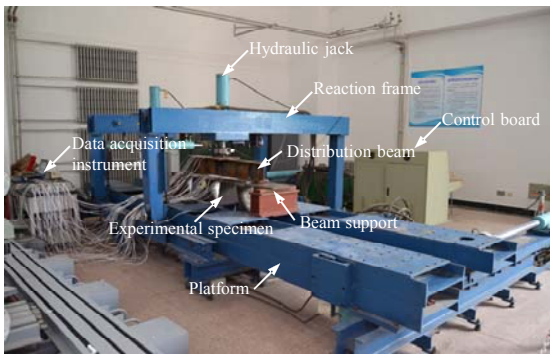



Fig.8 (a) Experimental device of static test

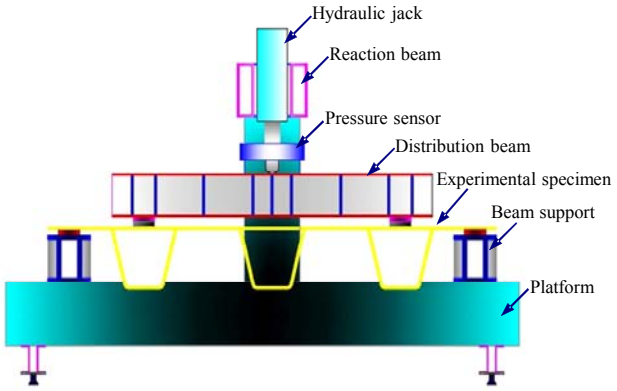



Fig.8 (b) Loading diagram of static test

Fig.8: Experimental diagram of static test for typical experimental specimen

9 load levels are adopted to conduct the static-load experiment. The minimum and maximum loads are 2kN and 18kN with load interval of 2kN. Taking OSD5 for instance, typical stress distribution is presented in Fig. 9 under the loads of 6 kN, 12kN and 18kN. The deck-top stress and deck-bottom stress is similar in absolute value with reverse direction. Thus, the specimen is certainly in the state of the pure moment conforming to the mechanical model. In addition, the stress is nearly equality during the middle region between neighboring welds. However, the stress increases significantly with the locations approaching the welded joint. The SCFs are calculated and obtained by Equation (2). Among the equation, the deck-bottom stress of middle locations between neighboring welds is applied as Snom while the deck-bottom stress approaching the weld toe is adopted as Shot. Various SCFs are presented in Table 3 with two welds for each specimen. It appears that SCFs at different weld location differs significantly because of the distinction from the geometrical dimension of the weld. Thus, fatigue evaluation requires taking into account of the discreteness of SCF.

| Specimen | Weld number | Value of SCF | Specimen | Weld number | Value of SCF |
|---|---|---|---|---|---|

| number | | | number | | |
|---|---|---|---|---|---|
| OSD1 | WJ-2# | 1.12 | OSD2 | WJ-2# | 1.25 |
| | WJ-3# | 1.46 | | WJ-3# | 1.11 |
| | WJ-4# | 1.06 | | WJ-4# | 1.55 |
| | WJ-5# | 1.32 | | WJ-5# | 1.56 |
| OSD3 | WJ-2# | 1.11 | OSD4 | WJ-2# | 1.29 |
| | WJ-3# | 1.17 | | WJ-3# | 1.13 |
| | WJ-4# | 1.11 | | WJ-4# | 1.22 |
| | WJ-5# | 1.24 | | WJ-5# | 1.21 |

Table 3: Data summary of experimental values of SCF around weld toe

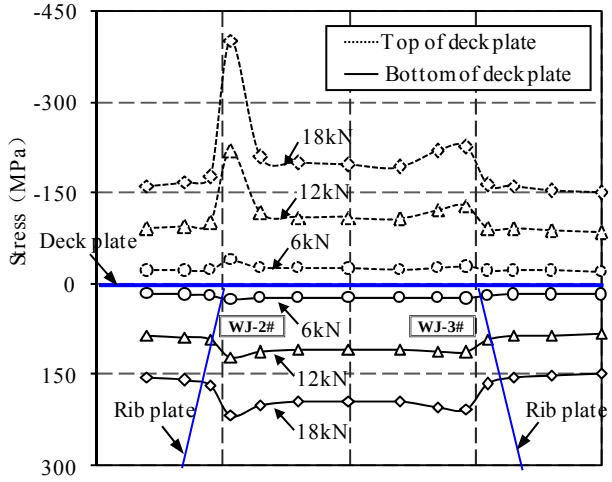



Fig.9: Typical stress distribution of deck plate for OSD 5under various loading steps

## 4.2 Environmental corrosion

Metal material and structure are subjected to environmental actions such as temperature, wetness and chemical substance in the standing service period [7,8]. Among these causes, fatigue performance of steel structure is affected significantly by the environmental corrosion. The influence of the corrosion is mainly highlighted on two aspects acting on fatigue performance. Firstly, fatigue resistance of steel material degrades gradually under the action of corrosion. Secondly, cross area of steel member decreases with the erosion of the corrosion environment. So, the stress or the stress amplitude magnify with the reducing of the cross area under uniform loads. Thus, above two aspects require taking account into consideration for achieve more accurate evaluation of steel-structural fatigue performance.

Assumptions are proposed to profile the action of corrosion on the cross area. Firstly, depth of the steel plate is merely influenced by the corrosion while the other two geometrical dimensions are consistence with no variation. Secondly, the depth of steel plate is presumed to satisfy the function with time as follows [14]:

$$\delta(t) = bt^r \tag{4}$$

where $\delta(t)$ denotes etch depth of steel plate, tis exposure time of steel structure, b and r is parameters related with material property.

For steel-plate composed structures, stress is influenced more significantly by the etch depth in through-thickness direction. For rib-to deck welded joint belonging to orthotropic steel deck, the stress around deck-plate region is crucial for the performance of fatigue. Previous investigations indicate that flexure stress of deck plate approaching the weld toe is sig-

nificant to fatigue performance. Therefore, the loss rate of section modulus ($\eta(t)$) is described as follows corresponding to the flexure capacity:

$$\eta(t) = \frac{W - \bar{W}(t)}{W} = \frac{2\delta(t)}{B} - \frac{\delta^2(t)}{B^2} \tag{5}$$

where $W$ denotes initial section modulus, $\bar{W}(t)$ is effective section modulus and B is width of cross section.， Bis thickness of steel plate. Assuming the stress is within elastic range around the rib-to-deck joint for deck plate, the bending moment is described as follows:

$$E \cdot \varepsilon \cdot W = E \cdot \bar{\varepsilon}(t) \cdot \bar{W}(t) \tag{6}$$

where $E$ denotes elastic modulus of steel material, $\varepsilon$ and $\bar{\varepsilon}(t)$ are initial strain and the strain at time of t respectively.

Substituting equation (5) into (6), then:

$$\bar{\varepsilon}(t) = \frac{1}{1 - \frac{2\delta(t)}{B} + \frac{\delta^2(t)}{B^2}} \varepsilon \tag{7}$$

Finally, effective stress amplitude (S(t)) is described as follows:

$$S(t) = \frac{1}{1 - \frac{2\delta(t)}{B} + \frac{\delta^2(t)}{B^2}} S_n \tag{8}$$

where $Sn$ denotes nominal stress amplitude which is calculated by initial area of cross sections.

Compared to the aggravation of near-weld stress for typical members, material property of fatigue resistance is much harder to profile under the action of environmental. Typical resistance models for fatigue of the specific welded joint are described by series of S-N curves. Most equations corresponding to S-N curves are shown as:

$$\lg N = \lg C - m \lg S \tag{9}$$

where $N$ denotes fatigue life represented by cycle number, $C$ is model parameter relating to material property, m is slop of lg$S$-lg$N$ curves, and $S$ is stress amplitude of welded joints. Plenty of investigations have pointed out that the value of m keeps consistent by action of the environmental corrosion. In addition, the parameter $C$ indicates time-dependent fatigue resistance of steel material acted by corrosion. Thus, $C$ is transferred as $C(t)$ to account for its time-dependent property. Some fitting functions have been proposed to profile the time-dependent property of $C(t)$. In further, exponential function has been recommended for application described as follows [14]:

$$C(t) = C_0 \varphi(t) = C_0 e^{-\alpha t} \tag{10}$$

Where $C_0$ is initial parameter of welded joint without environmental corrosion, $\varphi$(t) is degenerate function of $C$, and $\alpha$ is fitting parameter of $C(t)$ relative with environment condition and types of welded joints.

# 5 DETERMINATIVE FATIGUE- LIFE EVALUATIONS INTEGRATING MULTIPLE FACTORS

## 5.1 Primary evaluation method

Fatigue-resistance model is the basis to conduct fatigue- life evaluation. Typical codes for fatigue evaluation of steel bridge include BS5400, Eurocode 3and AASHTO. Compared to the ancient code of BS5400, Eurocode3and AASHTO involve the influence of stress threshold, which means that fatigue life of typical welded joints is infinite when stress amplitude is less than a certain value. Such kinds of certain value were called fatigue life constant amplitude fatigue limit. In addition, different slope value is taken into account at various scope of stress amplitude. Thus, Eurocode 3 has been applied to conduct the fatigue design and evaluation of steel bridge in mayor nations and regions.

Similarly with typical fatigue-resistance codes, fatigue life profiled by Eurocode 3 in shape of series of *S*-*N* curves. Fig. 4 presents the series of *S*-*N* curves for various types of welded joints. In the figure, $\Delta\sigma_C$ presents type of welded joints defined by constant stress range $\Delta\sigma$R ($\Delta\sigma_R$=2.0×*S*) in condition of stress cycle number equals to two million. $\Delta\sigma_C$ and $\Delta\sigma_L$ denote respectively constant fatigue limit and fatigue cut-off. The value of constant slope m of lg*S*-lg*N* curves equals to zero in condition of $\Delta\sigma_R<\Delta\sigma_L$. In such range, fatigue life of welded joints is infinite. In range of $\Delta\sigma_L\leq\Delta\sigma_R<\Delta\sigma_D$, the value of m varies from zero to five while the value of m becomes three when $\Delta\sigma_R \geq \Delta\sigma_C$.

Fatigue damage requires being calculated and obtained to confirm subsequently the fatigue life of typical welded joint. Plamgren-Miner methodology is adopted in mayor scopes including fatigue design and real-time evaluation. The methodology is based on the assumption of linear damage accumulation described as follows:

$$D = \sum_{S_i \geq \Delta\sigma_D} \frac{n_i S_i^3}{K_C} + \sum_{\Delta\sigma_L \leq S_j \leq \Delta\sigma_D} \frac{n_j S_j^5}{K_D} \tag{11}$$

where $K_C$ and $K_D$ keynote parameters related with types of welded joint.

Suffered from variable stress amplitudes, the most significant issue is how the fatigue life in confirmed by corresponding to the accumulated fatigue damage. Although the linear fatigue-damage methodology of Plamgren-Miner has been admitted widely, the ultimate value of accumulated fatigue damage, which indicates the ultimate fatigue life, is still in dispute. The reason for these disputes is mainly because linear fatigue-damage methodology does not include the nonlinear influence of fatigue damage by variable stress amplitudes. Therefore, fatigue life may achieve when fatigue damage *D* is less or more than 1.0. For simplification, deterministic fatigue life is confirmed and obtained by principle of fatigue damage *D* equaling to 1.0.

## 5.2 Fatigue-life evaluation considering multiple factors

Having identifying the general principle of evaluating fatigue life, fine factors are then required being taken account including the influence of stress concentration and environmental corrosion. Firstly, influence of stress concentration is represented by SCF. According to the result of static-load experiment, the value of SCF scatters in scope from 1.11 to 1.56. The mean value of SCF equaling to1.24 is adopted to conduct fatigue evaluation for simplification. Secondly, fatigue capacity is reflected in two aspects of $\Delta\sigma_C$ and $\Delta\sigma_L$ described by Equation (10). A is set as 0.006 and $C_0$ is selected as $\Delta\sigma_L$, $\Delta\sigma_R$ , $\Delta\sigma_D$, $K_C$ and $K_D$.As recommended by Eurocode 3 for rib-to-deck welded joint, $\Delta\sigma$L, $\Delta\sigma_R$,, $\Delta\sigma_D$, $K_C$ and $K_D$ are respectively 70MPa, 52MPa, 29MPa, 7.16×1011 and1.90×1015. Thirdly, the weakness of section corresponding to

corrosion is presented by Equation (4) to (8). In the equation, b and r is made as 60 and 0.48 [14].

Daily Seq is consequently obtained with the increase of service time, presented by Fig. 10. With the growth of service time, the daily $S_{eq}$ increases linearly with the service time. When the service time arrives at 100 year, the daily $S_{eq}$ reaches at 1.04 MPa. Even through the service time arrives at 300 year, three times of the designed service period, the daily $S_{eq}$ is only about 1.09 MPa. Fig. 11 presents curves of $\Delta\sigma L$ decreasing with commission period nonlinearly. $\Delta\sigma_L$ equals to 15.9, 8.7 and 4.8 MPa, when commission period arrives at 100, 200 and 300 years. Without consideration of increase of train weight, Se,d is far less than $\Delta\sigma L$ indicating that fatigue life of rib-to-deck welded joints is beyond 300 years. As the designed service life of bridge is 100 years, fatigue life of orthotropic steel deck of Dashengguan Bridge can be considered infinite.

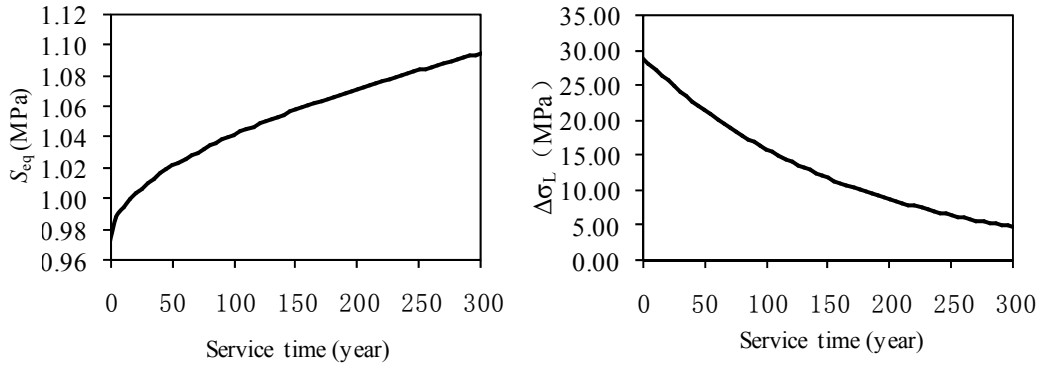



Fig.10: Curve of equivalent stress amplitude with service time Fig.11: Curve of fatigue limit with service time

### 5.3 Discussion on influence of train weight, train flow and number of carriages

However, fatigue life is perhaps not infinite when taking account of sustained development of train weight. If the stress of orthotropic steel deck is within elastic range under train load actions, the stress amplitude is linearly increased with the growth of train weight. Annual growth rate of train weight is introduced to considering the sustained development of train weight. The value of 1%, 3% and 5% are adopted to investigate the influence of train weight growth on fatigue life. Fig. 12 presents the curves of $S_{e,d}$ under growth rate of train weight of 1%, 3% and 5%, as well as the variation curves of $\Delta\sigma_L$ and $\Delta\sigma_D$. When the train weight is increased by annual growth rate of 1% and 3%, $S_{e,d}$ approaches to $\Delta\sigma L$ at the time of 95th year by far. It indicates that fatigue life is much greater than the designed service period of 100 years.

Yet, $S_{e,d}$ is perhaps beyond $\Delta\sigma_L$ and $\Delta\sigma_D$ at 61th and 72th year when annual growth rate of train weight reaches at 5%.It is known from Equation (11) that there is no fatigue damage within the first 61 years and fatigue damage is accumulated from 61th and index of stress amplitude transform from 5.0 to 3.0 at 72th. Fatigue life is consequently 77 years when accumulated fatigue damage *D* reach at 1.0. It means that fatigue failure perhaps occurs within the designed service life (100 years).

However, above investigations merely includes the increase of train weight without consideration of annual train flow and number of train carriages. If these two factors are taken account of, the calculated fatigue life becomes much shorter, less than 77 years.

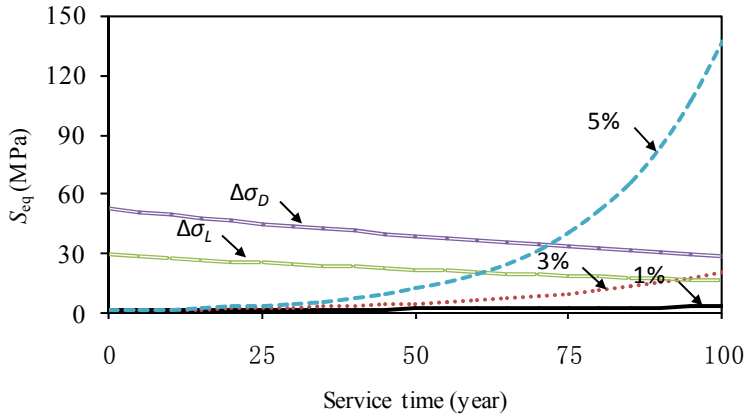



Fig.12: Curves of equivalent stress amplitude, constant-amplitude fatigue limit and cut-off fatigue limit with the service time

## 6 TIME-DEPENDENT FATIGUE RELIABILITY ASSESSMENT

### 6.1 Limit state equation of fatigue failure

When stress amplitude exceeds the constant fatigue limit, fatigue damage is accumulated induced by action of sustained train load and environmental corrosion. Fatigue failure perhaps occurs in condition that the fatigue damage arrives at critical value. Thus, the limit state equation of fatigue failure for welded joint can be described as follows:

$$g(X)=\Delta-e\cdot D=\Delta-e\cdot\left(\sum_{S_i\geq\Delta\sigma_D}\frac{n_iS_i^3}{K_C}+\sum_{\Delta\sigma_L\leq S_j\leq\Delta\sigma_D}\frac{n_jS_j^5}{K_D}\right) \tag{12}$$

where△ denotes ultimate value of fatigue damage, $e$ is correction factor caused by error measurement of strain gauges and $D$ is accumulated fatigue damage calculated by Equation (11).

When g(X)<0, the welded joint is under fatigue-failure state in probability sense. Thus, fatigue failure probability is described as $p_f=p(g(X)<0)$ and the fatigue reliability index $\beta$ is calculated as follows:

$$\beta=\Phi^{-1}(1-p_f)=-\Phi^{-1}(p_f) \tag{13}$$

where $\Phi^{-1}(\cdot)$ denotes inverse function of standard normal distribution.

### 6.2 Probability-distribution function of critical factors

There are four variables in the fatigue-failure limit state equation, i.e. △, *e*, *Si (Sj)*. Above determinative fatigue-life calculation is based on the assumption that fatigue life arrives when fatigue damage $D$ equals to 1.0. However, fatigue failure may occur when the $D$ is unequal to 1.0 with random distribution. According to related investigation from article [15], e is described statistically as normal distribution whose mean value and standard deviation are respectively 1.0 and 0.03.

Fig. 13 presents the calculated histogram distribution of two fatigue effects, i.e. daily *Seq* and N. In the figure, Nominal, Log-nominal and t-distribution functions are applied for profiling. By contrast, t-distribution function fits much better than the two others. Thus, t-distribution functions are applied to profile statistical property of daily *Seq* and *N*. Table 4 presents the parameters of t-distribution function for daily *Seq* and *N*.

| Fatigue effect | Distribution function | Mean value | Standard deviation | Variable cofficient |
|---|---|---|---|---|
| Daily Seq | t-distribution | 0.72 | 0.025 | 0.03 |
| Daily N | t-distribution | 5179 | 267 | 0.05 |

Table 4. Parameters of probabilistic models for daily $S_{eq}$ and daily $N$

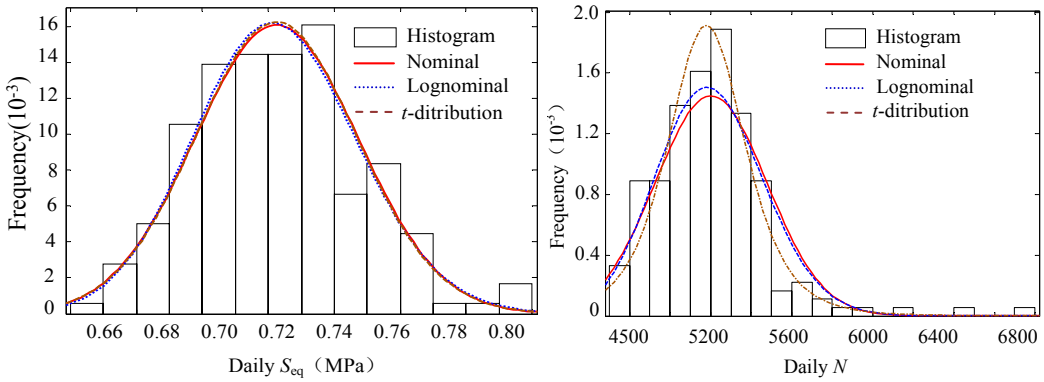



Fig. 13: Probabilistic models of daily equivalent stress amplitude and daily stress cycle number

### 6.3 Time-variant fatigue-reliability index

Because of the randomness of influencing factors, reliability evaluation is constantly applied to assess in-service performance of engineering structure. With the service time going on, the stochastic behavior of influencing factors varies inducing the transformation of structural reliability. The characteristic of structural reliability corresponding to time is so-called time-dependent reliability.

Constantly, reliability index βis adopted as the parameter for evaluation the reliability. Various methods has been carried out to calculate the reliability index β. These presentative methods include First Order Second Moment Method, Response Surface Method and Monte-Carlo Simulation Method. Typically, Monte-Carlo Simulation Method is much more applicable in situation when influencing factor does not fit the Nominal distribution function. Monte-Carlo Simulation Method conducts random sampling directly based on statistical model of actual events. Then, the occurrence frequency of actual events is acquired such as the fatigue-failure frequency of welded joints.

In this paper, Monte-Carlo Simulation Method is applied to acquire the approximate solution of fatigue failure probability according to Equation (12). Influencing factors are taken account of including the stress concentration, environmental corrosion as well as the growth of train weight. The action rule of above influencing factors is considered by Section 4 incorporating the time effect. Fatigue reliability index β is finally commutated by Equation (12).

The value of 100,000 is selected as the sampling number in each condition for typical Monte-Carlo Simulation process. Including the influencing factors mentioned above, three kinds of growth rate of train weight, i.e. 1%, 3% and 5%, are adopted as typical conditions. Fig. 14 presents the curves of fatigue reliability index $\beta$ corresponding to time under above three typical conditions. The target reliability index is selected as 3.0 according to engineering experience [16]. In the figure, the fatigue reliability index β maintains 3.30 until the designed service period of 100 years under the condition of growth rate of train weight with 1% and 3%.

In contrast, the fatigue reliability index $\beta$ maintains 3.30 until the 84th year when the growth rate of train weight equals to 5%. Then, the fatigue reliability index β declines rapidly and reaches 3.0 at the time of 85th year. After short-time volatility for about ten years, the fatigue reliability index β maintains at constant level equaling to 0.0. It indicates that fatigue failure may occur with significantly high probability.

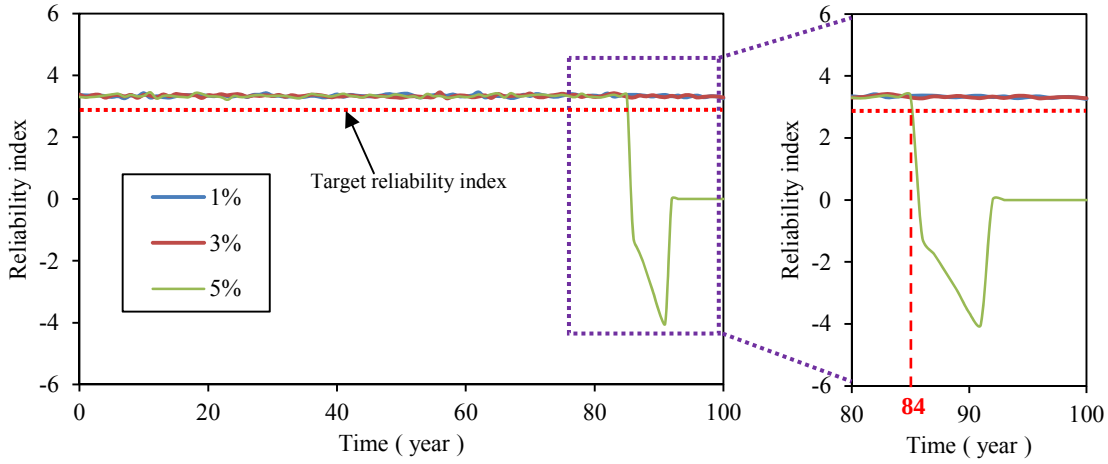



Fig.14: Curve of reliability index with time under different growth rate of train weight

## 7 CONCLUSIONS

Fatigue issue is typical structural performance for high-speed railway bridge. Once fatigue cracking is happened in the deck plate under rail lines, the ride performance of the high-speed train is reduced inducing potential major traffic accident. Obtained from the structural monitoring system, strain history data provides chances to conduct the real-time fatigue evaluation for typical high-speed railway bridge. In this paper, the fatigue performance is evaluated for Dashengguan High-speed Railway Bridge in China by assessing its fatigue life and time-dependent fatigue reliability. In addition, multiple factors are taken into account including stress concentration, environmental corrosion and growth of train weight. Conclusions are presented as follows:

- Under the passage of trains with different number of carriages, the value of equivalent stress amplitude is approaching with each other. However, the stress cycle number is proportional to the number of train carriages, indicating the linear relation between stress cycle number and train axle number.
- Without consideration of the randomness of influencing factors, the fatigue life of rib-to-deck welded joint is infinite taking only account of environmental corrosion and stress concentration. If consistent development of train weight is considered, the fatigue life may decrease to the value less than the designed service period. For instance, fatigue life is only about 132 years and 77 years when the annual growth rate of train weight reaches to 3% and 5%.
- Similarly to the fatigue-life evaluation without regardless of randomness, fatigue reliability of rib-to-deck welded joint remains at relatively higher level taking no account of the consistent development of train weight. If the growth rate of train weight is considered as 5%, the fatigue reliability index is less than the target reliability index when arriving at the 84$^{th}$ year. Thus, the reliable fatigue life is 84 years when the growth rate of train weight equaling to 5%. Consequently, it indicates that fatigue performance of rib-to-deck welded joint is likely to be unreliable within the design-life period.

## ACKNOWLEDGEMENTS

The work described in this paper was substantially supported by the grant from the National Natural Science Foundation of China (No. 51008069), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), and the Top-notch Academic Programs Project of Jiangsu Higher Education Institutions (TAPP), which are gratefully acknowledged.

## REFERENCES

[1] D. Baxter, D. Nemovitz. Sensitivity analysis of rail–structure interaction force effects for direct-fixation, *In Annual Conference & Exposition AREMA*. 2012.

[2] Y. Bin, G.L Dai, and H.P. Zhang, Beam–track interaction of high-speed railway bridge with ballast track, *J Cent S Univ Technol* , 1447–1453, 2012

[3] Z.G. Xiao, K. Yamada, S. Ya, et al. Stress analyses and fatigue evaluation of rib-to-deck joints in steel orthotropic decks. *International Journal of Fatigue,* 1387–1397, 2008.

[4] M.S. Pfeil, R.C. Battista and A.J.R. Mergulhao. Stress concentration in steel bridge orthotropic deck. *J Construct Steel Res,* 1172–1184, 2005.

[5] T. Guo, A.Q. Li, and H. Wang. Influence of ambient temperature on the fatigue damage of welded bridge decks, *International Journal of Fatigue,* 1092-1102. 2008

[6] Y. Wang, Z.X. Li, and A.Q. Li, Combined use of SHMS and finite element strain data for assessing the fatigue reliability index of girder components in long-span cable-stayed bridge, *Theoretical and Applied Fracture Mechanics,* 127-136, 2010

[7] Y.S. Song, and Y.L Ding. Fatigue monitoring and analysis of orthotropic steel deck considering traffic volume and ambient temperature, *Sci China Tech Sci,* 1758-1766, 2013

[8] R. Rahgozar, and Y. Sharlifi. Remaining fatigue life of corroded steel structural members, *Advances in Structural Engineering,* 881-890, 2011

[9] Y.T. Ma, Y. Li and F.H Wang. The atmospheric corrosion kinetics of low carbon steel in a tropical marine environment, *Corrosion Science,* 1796-1800, 2010

[10] X.W. Ye, Y.Q. Ni and J.M. Ko. Experimental evaluation of stress concentration factor of welded steel bridge T-joints, *Journal of Constructional Steel Research,* 78-85, 2012

[11] X.W. Ye, Y.Q. Ni, K.Y. Wong, et al. Statistical analysis of stress spectra for fatigue life assessment of steel bridges with structural health monitoring data, *Engineering Structures,* 166-176, 2012

[12] T. Guo, and Y.W. Chen, Field Stress/Displacement Monitoring and Fatigue Reliability Assessment of Retrofitted Steel Bridge Details, *Engineering Failure Analysis,* 354-363, 2011

[13] A.Q. Li, Y.L. Ding and H. Wang. Analysis and assessment of bridge health monitoring mass data—Progress in research development of “Structural Health Monitoring”, *Sci China Tech Sci,* 2212–2224, 2012

[14] J.X. Gong, and G.F. Zhao. Fatigue reliability analysis for corroded reinforced concrete structures, *China civil engineering journal,* 50-56, 2000

[15] D.M. Frangopol, A. Strauss and S. Kim. Bridge Reliability Assessment based on Monitoring, *Journal of Bridge Engineering,* 258-270, 2008

[16] P.H. Wirsching, and Y.N. Chen, Fatigue design criteria for TLP tendons. *Journal of Structural Engineering,* 1398-1414, 1987.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# A MORE COMPREHENSIVE MODELING OF CONTACT FORCE DURING SHEAR TESTING USING DEM

**Varvara Roubtsova**[1] **and Mohamed Chekired**[1]

[1] Institut de recherche d'Hydro-Québec (IREQ),
Expertise Sciences des matériaux
Varennes, Québec, Canada
e-mail: roubtsova.varvara@ireq.ca

**Keywords:** Discrete Element Method (DEM), Shear Test, Energy-Work Balance, High Performance Parallel Computing (HPPC).

**Abstract.** *The subject of this paper is a visualization tool which may prove helpful for understanding the highly complex responses of granular systems during shear tests. The motivation of such approach is to face the challenge related to the micro scale contact forces in a granular system. The Discrete Element Method (DEM) was used to simulate shear testing on an assembly of spherical glass beads. All simulations were performed using the 3D virtual laboratory SiGran. The macro-scale response of the granular system is presented, including energy and force distributions. Two particle-size distributions were considered in the study to validate the virtual approach. The influence of the contact model is also a part of this study.*

# 1 INTRODUCTION

In this paper, we present a first attempt to assess the different types of energy involved in granular systems at a particle scale by using the Discrete Element Method (DEM). Continuum approaches do not take into account the particulate nature of the granular system, in which overall deformation is essentially the effect of two non-linear and irreversible processes: relative sliding between particles, and rotation of particles. Because of their particulate nature, granular systems exhibit highly complex responses to applied load. For practical purposes, then, it is important to consider this particulate nature in order to overcome the limitations of continuum approaches. Being discrete in nature, DEM [1] is able to model the motion of a particle assembly without losing or neglecting the individual characteristics of each particle. The granular system could either be saturated with water or have other fluids or/and gases in the voids. A virtual laboratory named SiGran was developed by Hydro-Quebec's research institute (IREQ) to mimic the different experimental tests that are standard for soil sciences and also to better understand the different phenomena that occur at the micro level during those tests. The visualization features of this laboratory help to access information that is very difficult or impossible to study in real experiments. The description of work conducted with SiGran can be found in the proceedings of international conferences [2-4]. SiGran is based on coupling of DEM and MAC (Marker-And-Cell [5]) methods. Additionally, we assume a dry particle assembly, excluding any kind of attraction forces. A description of the MAC method is absent here but can be found in [6].

In the present study, a validation was first performed by comparing the DEM simulation results with the experimental ones obtained by [7] at a macro-scale on a sheared granular system made up of mono-disperse idealized spherical particles. In the second phase, the model was applied to poly-disperse idealized spherical particles.

# 2 DISCRETE ELEMENT METHOD

The Discrete Element Method (DEM) describes the motion of particle assemblies and has been widely used as a mathematical tool studying their behavior. DEM was introduced by Cundall in 1971 [1] for the analysis of rock mechanics and was applied to soils by Cundall and Strack in 1979 [8]. It is based on Newton's second law of motion. An explicit time stepping scheme is used for integration in time. The behavior of the particle assembly under external loading at a given time step can be described using information from the previous time step, including the dimensions of each particle, its position relative to adjacent particles and the relative velocities (displacement and rotation) of neighboring particles. The contact forces acting on the particle are calculated from this data, and the acceleration, velocity and position at the new time step are determined. The concept of interactive forces between particles is illustrated in Figure 1.

The fundamental equations of motion for the $i^{\text{th}}$ particle are as follows (the subscript $j$ refers to the $j^{\text{th}}$ neighboring particle, and $k$ is the number of neighboring particles):

$$m_i \frac{d\vec{V}_i}{dt} = m_i \vec{g} + \sum_{j=1}^{k} \left( \vec{F}_{cn,ij} + \vec{F}_{dn,ij} + \vec{F}_{ct,ij} + \vec{F}_{dt,ij} \right) \tag{1}$$

$$I_i \frac{d\vec{\omega}_i}{dt} = \sum_{j=1}^{k} \left( \vec{T}_{ij} + \vec{M}_{ij} \right) \tag{2}$$

Where:

$m_i$, $I_i$, are the mass (kg) and moment of inertia (kg·m$^2$) of particle *i* respectively;

$\vec{V}_i$ and $\vec{\omega}_i$ are the translational (m/s) and rotational (s$^{-1}$) velocities of particle *i*;

$\vec{F}_{cn,ij} = -K_n \delta_n^{\frac{3}{2}} \vec{n}$ is the normal force of contact (N);

$\vec{F}_{dn,ij} = -C_n \vec{V}_{n,ij}$ and $\vec{F}_{dt,ij} = -C_t \vec{V}_{t,ij}$ are the normal and tangential damping forces (N) respectively;

$\vec{F}_{ct,ij}$ is the tangential force of contact (N); this force will be discussed in detail later;

$\vec{T}_{ij} = R_i \left( \vec{F}_{ct,ij} + \vec{F}_{dt,ij} \right)$ is the rolling torque (N·m);

$\vec{M}_{ij} = -\mu_r \left| \vec{F}_{cn,ij} \right| \frac{\vec{\omega}_i}{\left| \vec{\omega}_i \right|}$ is the friction torque (N·m);

$K_n = \frac{4}{3} E^* \sqrt{R^*}$, $C_n = D_n \sqrt{6 m^* E^* \sqrt{R^* \delta_n}}$, $C_t = D_t \sqrt{6 m^* E^* \sqrt{R^* \delta_n}}$

$\delta_n$, $\delta_t$ are the normal and tangential contact displacement respectively;

$\frac{1}{E^*} = \frac{1-\nu_i^2}{E_i} + \frac{1-\nu_j^2}{E_j}$ and $\frac{1}{R^*} = \frac{1}{|R_i|} + \frac{1}{|R_j|}$

Where $E$ is Young's modulus (Pa), $\nu$ is the Poisson ratio, $R_i$ is the particle radius (m), $\mu_r$ is the rolling (m) friction coefficient, and $D_n$ and $D_t$ are the normal and tangential damping coefficients respectively.

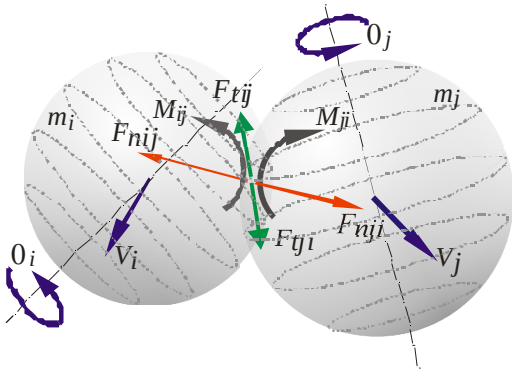



Figure 1. Interactive forces between two particles

## 3 TANGENTIAL FORCE MODELS

A wide range of tangential forces models are available for DEM [9]. The theoretical basis for a complex approach was proposed by Mindlin and Deresiewicz [10] for elastic materials based on Hertz theory. According this theory, the actual tangential force depends on the whole loading history and also on the instantaneous rate of change of the normal force or displacement. The tangential force is given by:

$$\frac{dF_{ct,ij}}{dt} = K_t \nu_t \qquad (3)$$

Where: $K_t = 2\sqrt{R^* \delta_n} \left( \frac{G_i}{2-\nu_i} + \frac{G_j}{2-\nu_j} \right)$

G is the shear modulus; $v_t$ is a relative tangential velocity between surfaces in contact.
A complete loading force-displacement path can be built through successive incremental steps. A limit on the maximum tangential stress exists as stated by Coulomb's law of friction and is expressed as follows:

$$\left|F_{ct,ij}\right| \leq \mu_s F_{cn,ij} \quad (4)$$

Where $\mu_s$ is the sliding friction coefficient.

More details and analysis of loading and unloading curves for specific cases can be found in [11] and [12].

The SiGran programming code uses a high level of data parallelism on NVIDIA Graphics Processing Units (GPU) based on Tesla High Performance Computing (HPC) hardware. The forces acting on each inter-particle contact at each step of time are computed on its own GPU-kernel. Thousands of contacts are thus processed at the same time. In addition, one contact usually lasts several time steps, and several contacts can begin or end during a given time step. In theory, each particle can have more than one contact with all of the particles in the assembly during an experiment. For calculation purposes, only the history for the acting contact must be stored. This history begins with the first contact of specific particles, and it can be ended after loss of contact. A special algorithm, optimized for GPU calculations, was introduced into SiGran to maintain the history of each contact. This algorithm takes into account the maximum possible contacting particles, which depends on the relative particle size (i.e., the biggest and the smallest) in the assembly and a memory model of the GPU to minimize calculation time.

The simplest tangential forces model uses only the maximum force obtained by Coulomb's law. In this case, no history of contact is considered, which simplifies the algorithm and significantly reduces calculation time and used memory. In the case of SiGran, this is overall shared GPU memory.

# 4 DIRECT SHEAR TEST

The direct shear test is one of the oldest soil strength tests performed in the laboratory. Even though it has some disadvantages, as documented in ASTM D 3080 – 98 [13], it is routinely used by geotechnical engineers to determine the shear strength parameters of soil essential for stability assessment. The first direct shear apparatus was built by Alexander Collin in 1846 to measure the strength of a clay soil shear for purposes of slope stability assessment [14, 13]. The current version of the direct shear apparatus was designed by Casagrande in 1932. Figure 2 shows a schematic representation of the shear box.

A normal load is applied to the specimen before shearing across the predetermined horizontal plane between the two halves of the shear box. Measurements of shear load, shear displacement and normal displacement are recorded. From these results, the shear strength parameters can be determined.

## 4.1 Physical direct shear test on spherical mono-disperse chrome steel particles

The shear tests performed on spherical chrome steel particles are described in [7]. A square metal box, cross-section 60 x 60 mm, was filled with 11,700 chrome steel particles with a sphere radius of 0.992mm. The material density was 7,800 kg/m$^3$, the shear modulus was 7.9 x $10^{10}$Pa, and the Poisson's ratio was 0.28. The average friction coefficient of 0.096 was measured in [16], and the sphere-boundary friction coefficient for the shear box was 0.175. The force required to maintain the fixed section of the box in a stationary position was measured using a load cell. The normal stress was 55 kPa.

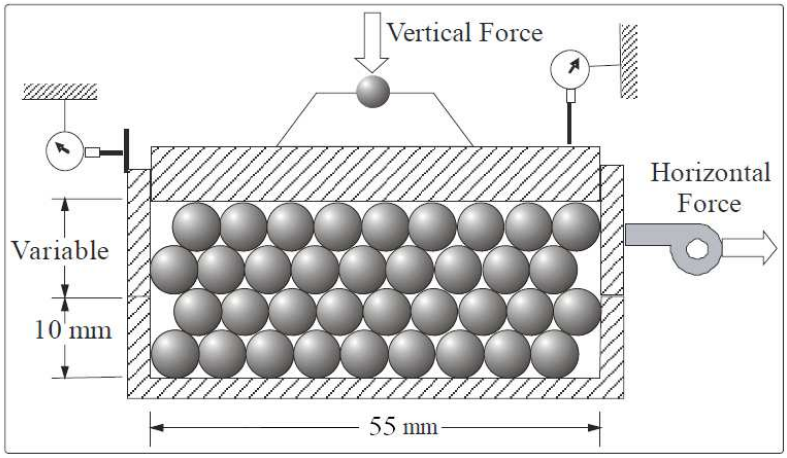
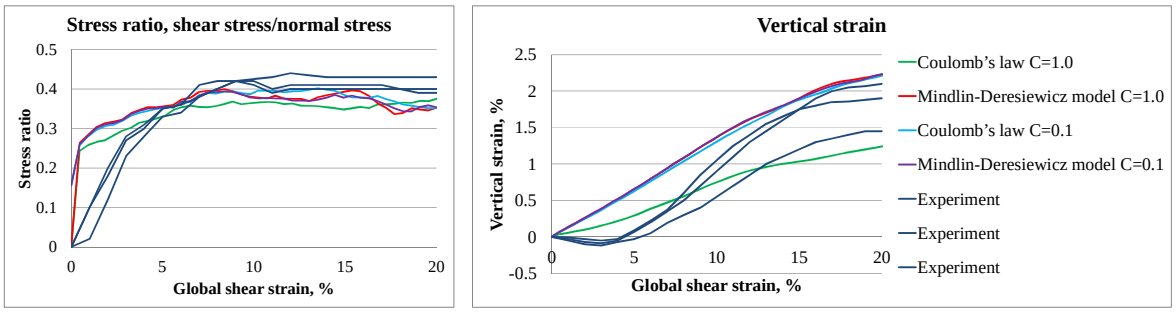



Figure 2: Schematic representation of a shear box

## 4.2 Numerical simulation of direct shear test with spherical mono-disperse chrome steel particles

The shear test simulation was conducted with SiGran. The forces applied on each box wall were calculated by summing the contact forces during interaction between particles and boundaries. The comparison between the experimental data and numerical results is shown in Figure 3. The shear stress was determined as in the experiment: the total of particle forces applied on the fixed section of the box was divided by the original cross-section area. The overall shear strain, as proposed in [7], is considered to be the displacement of the moving section of the box divided by the initial height of the sample. The vertical strain is taken to be the ratio of the vertical displacement of the upper half of the shear box to the original height of the samples. The difference in vertical stress results between the laboratory tests and the DEM simulations is most likely due to initial preloading and better arrangement for the simulation. The increase in the inter-particle friction coefficient during a shear test also contributed to the difference in the results.




(a) (b)

Figure 3. Comparison of numerical and experimental data

The numerical simulations for two tangential force models were conducted with two values of Courant's number that differed by a factor of 10 (1.0 and 0.1). Courant's number is known for its connection with the Courant-Friedrichs-Lewy convergence condition for solving partial differential equations. The time step is calculated using the following formula:

$$\Delta t \leq C \min \sqrt{\frac{m}{K_n}} \tag{4}$$

Where:
C is a Courant number
m is the mass of the particle
$K_n$ is a normal stiffness coefficient (see Formula 1)

The results for the Mindlin-Deresiewicz model are independent of the number of iterations during a particle contact, in contrast with Coulomb's low, which is dependent on that value.

As mentioned before, numerical simulation gives access to more parameters than physical experiments. The real tangential force can be obtained from the force equilibrium on the shear box in the shear direction. All forces acting on the boundaries in the shear direction are shown in figure 4.

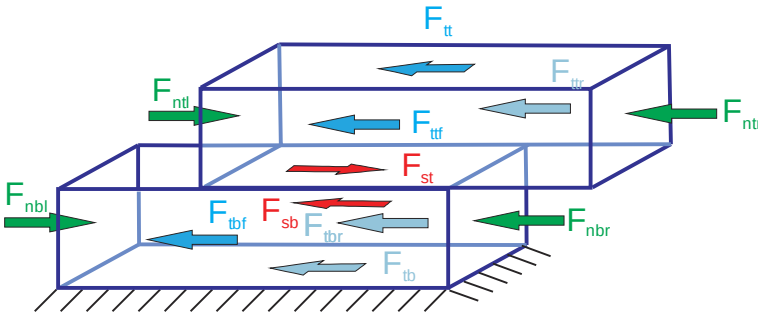



Figure 4. Forces acting in shear direction

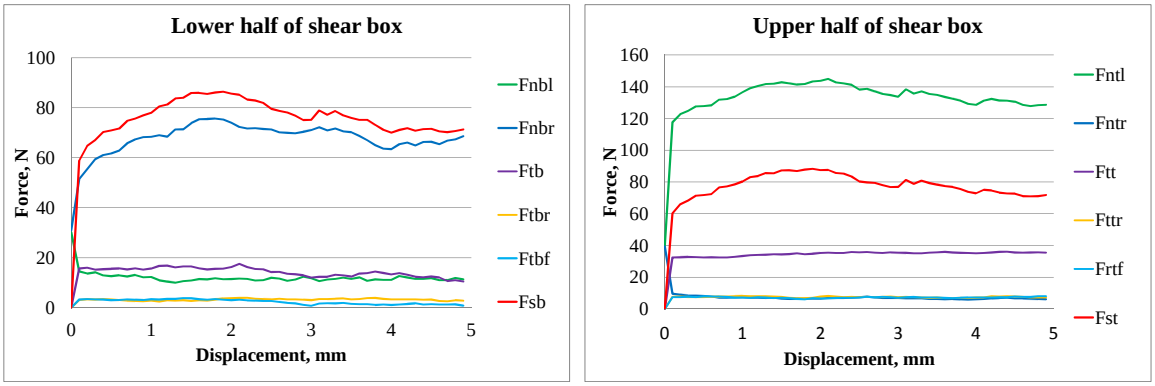
The following expressions can be written to find the shear forces involved in the upper and lower halves of the shear box:

$$\begin{aligned} F_{st} &= F_{ntl} - F_{ntr} - F_{tt} - F_{ttf} - F_{ttr} \\ F_{sb} &= F_{nbr} - F_{nbl} + F_{tb} + F_{tbf} + F_{tbr} \end{aligned} \quad but\ F_{st} = F_{sb} \tag{5}$$

Where:

$F_{st}$ : Shear force obtained from the equilibrium of the upper half of the shear box (N)
$F_{sb}$ : Shear force obtained from the equilibrium of the lower half of the shear box (N)
$F_{ntl}$ : Normal force on the left boundary of the upper half of the shear box (N)
$F_{nbl}$ : Normal force on the left boundary of the lower half of the shear box (N)
$F_{ntr}$ : Normal force on the right boundary of the upper half of the shear box (N)
$F_{nbr}$ : Normal force on the right boundary of the lower half of the shear box (N)
$F_{tt}$ : Tangential (friction) force on the top surface of the upper half of the shear box (N)
$F_{tb}$ : Tangential (friction) force on the down surface of the lower half of the shear box (N)
$F_{ttf}$ : Tangential (friction) force on the front surface of the upper half of the shear box (N)
$F_{tbf}$ : Tangential (friction) force on the front surface of the lower half of the shear box (N)
$F_{ttr}$ : Tangential (friction) force on the rear surface of the upper half of the shear box (N)
$F_{tbr}$ : Tangential (friction) force on the rear surface of the lower half of the shear box (N)

Figure 5 shows the evolution of all forces on the upper and lower part of the shear box separately. These forces are obtained by simulations.




(a) (b)

Figure 5. Evolution of forces acting during the shear test

Figure 6 shows the comparison between the shear forces calculated from Formula 5 and the normal forces. The shear force can be measured during physical experiments. The shear force values calculated at the upper and lower halves are almost identical, which demonstrates a good equilibrium of forces during all simulations. This figure shows that the force needed to move the upper half of the shear box is different from the force needed to keep the lower half stationary. This difference is a result of friction between the particles and the box.

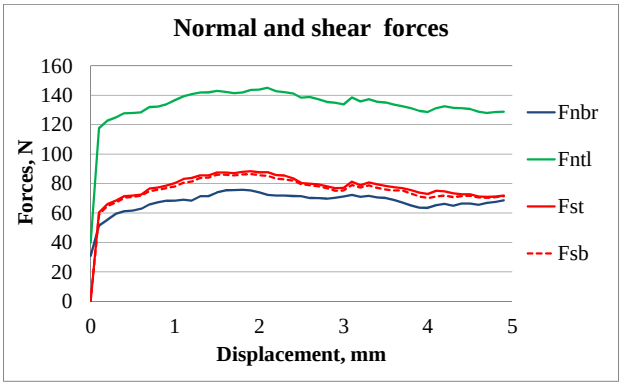



Figure 6. Normal and shear forces obtained during the shear test

### 4.3 Work-energy balance during shear test

The force required to move the upper section of the box accomplished work. The work done was calculated by integrating the force vector along the path traveled by the force. The total work done by all forces must be equal to the energy variation. The following types of energy were used:

- Kinetic energy, which has two parts: translation ($E_{kt}$) and rotation ($E_{kr}$):

$$E_k = E_{kt} + E_{kr} = \sum_{i=1}^{n} \frac{m_i v_i^2}{2} + \sum_{i=1}^{n} \frac{I_i w_i^2}{2}$$

Where *n* is the number of particles in the shear box.

- Potential energy:

$$E_p = \sum_{i=1}^{n} m_i h_i g$$

Where $h_i$ is the distance from the bottom of the box to the particle center.

- Internal elastic energy:

$$E_s = 0.5 \cdot \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\delta} F_{cn} \Delta\delta = 0.5 \cdot \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\delta} K_n \delta^{\frac{3}{2}} \Delta\delta = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{1}{5} K_n \delta^{\frac{5}{2}}$$

Where $k$ is a number of contacts for particle $i$.

- Normal dissipative energy:

$$E_{dn} = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\delta} F_{dn} \Delta\delta$$

- Friction energy:

$$E_f = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{L} F_{ct} \Delta l$$

Where $l$ is a tangential overlap during of contact ( m).

- Tangential dissipative energy

$$E_{dt} = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{L} F_{dt} \Delta l$$

- Rotation energy:

$$E_M = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\varphi} T_i \Delta\varphi = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\varphi} r_i \left(\overline{F}_{ct} + \overline{F}_{dt}\right) \Delta\varphi$$

Where φ is a rotational angle of particle $i$ (rad).

Note that this type of energy can be included in tangential energy by including rotation in the tangential overlap.

- Dissipative rotational energy:

$$E_{dM} = \sum_{i=1}^{n} \sum_{j=1}^{k} \int_0^{\varphi} M_i \Delta\varphi$$

More details about the calculation of energy using SiGran can be found in [17].

Figure 7 shows the work-energy balance and also the contribution of the normal forces applied to move the half box and accomplish the total work.

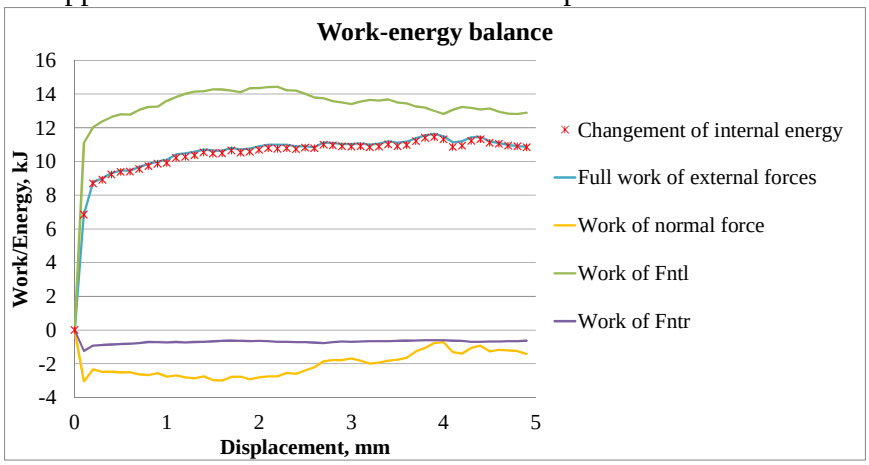
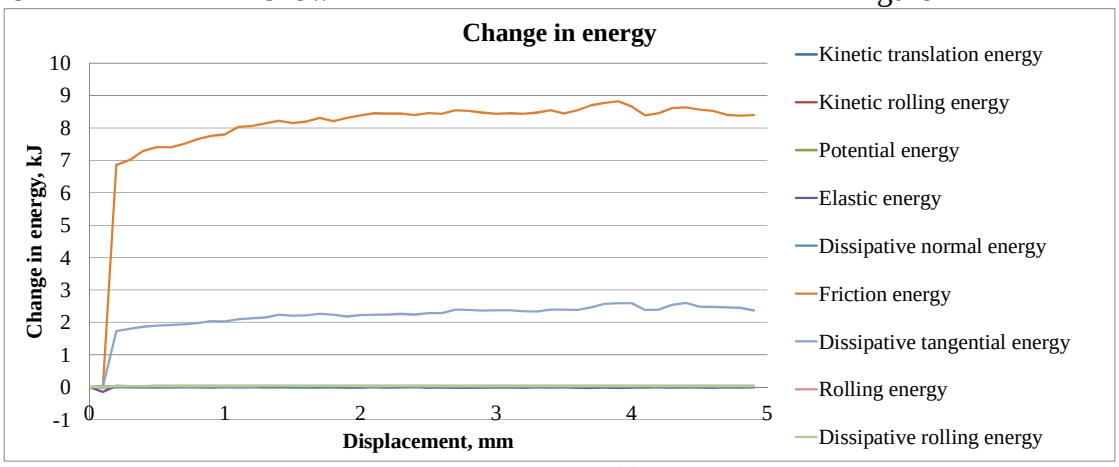


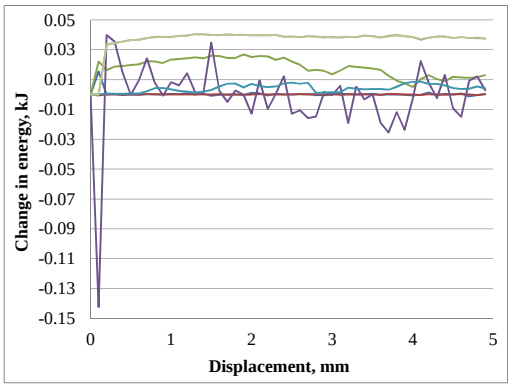

Figure 7. Work-Energy balance during the shear test

Figure 8 shows the contribution of each type of energy in the total energy balance of the system. It is clear that the main work of external forces is expended by the dissipation of energy during inter-particle and particle-wall friction. The others type of energies are shown

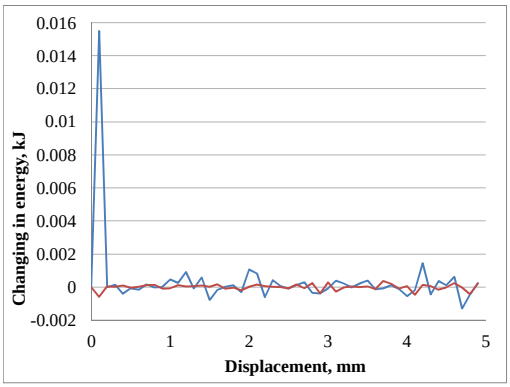
in Figure 8 b and 8c in which the scale of the vertical axe has been changed. Figure 9 shows the distribution of different energies among the particles. The dissipative types of energy were calculated cumulatively from the beginning of the shear experiment without considering the energy scattered during sample formation. Inter-particle force distribution during shear testing is shown in Figure 10.




(a)

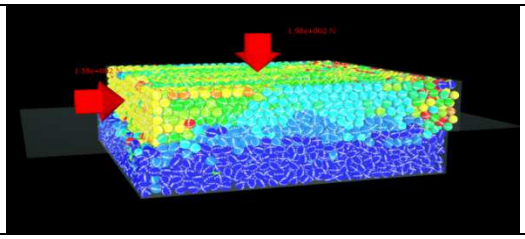



(b)

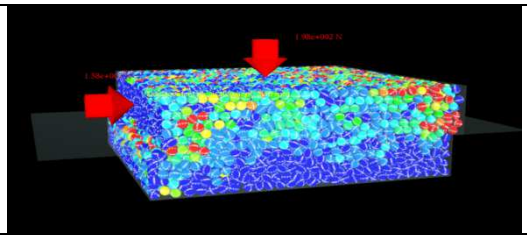



(c)

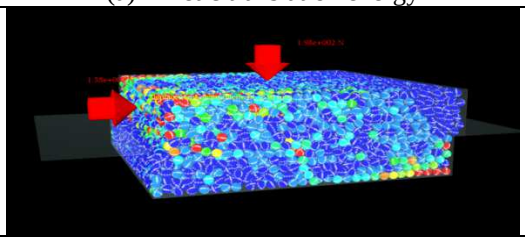
Figure 8. Change in energy during shear testing



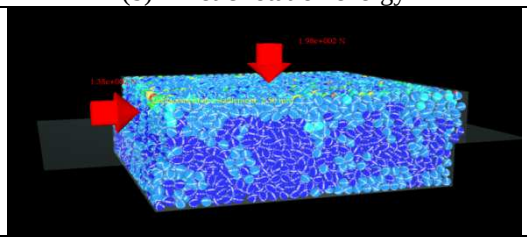
(a) Kinetic translation energy



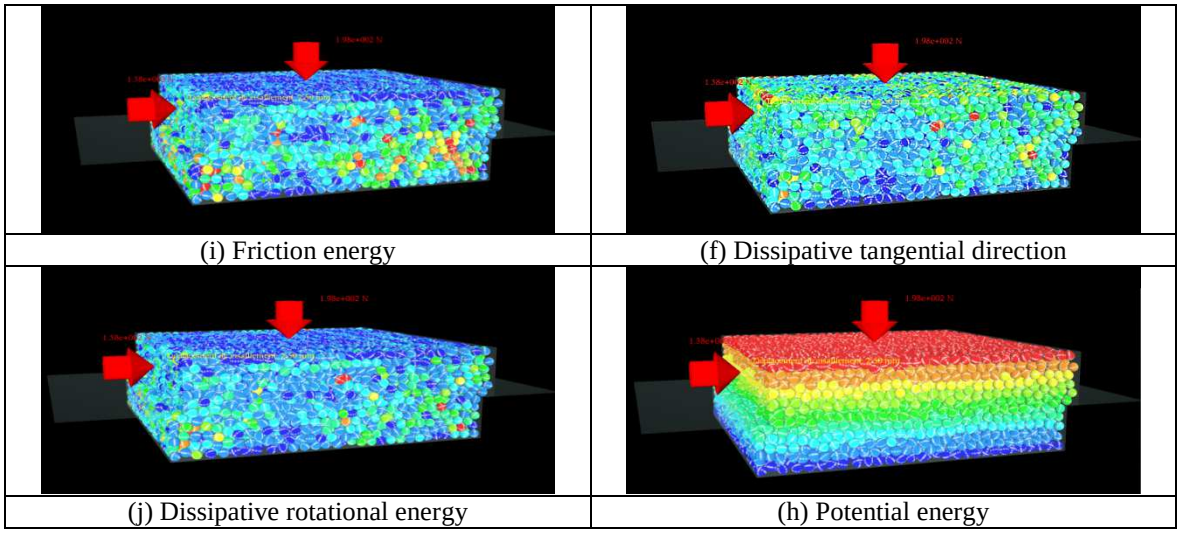
(b) Kinetic rotation energy



(c) Elastic energy



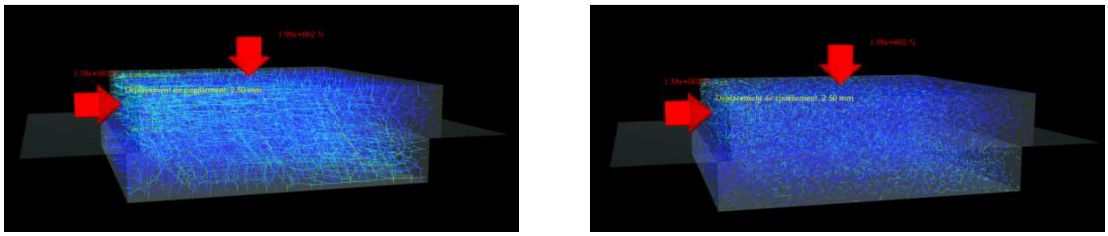
(d) Dissipative normal direction energy



(i) Friction energy | (f) Dissipative tangential direction

(j) Dissipative rotational energy | (h) Potential energy

Figure 9. Distribution of different energies among particles (shear displacement = 2.5 mm)



(a) Normal forces (b) Tangential forces

Figure 10. Inter-particle force distribution (shear displacement = 2.5 mm)

### 4.4 Simulation of shear test on poly-disperse particle assembly

The next simulation was carried out on a particle assembly having the particle size distribution presented in Figure 11. The details about this test and its comparison with experimental data can be found in [17]. Only the influence of the friction coefficient will be considered in this case. The shear tests were carried out in a box with internal dimensions of 55 x 55 mm. The samples were spherical glass beads.

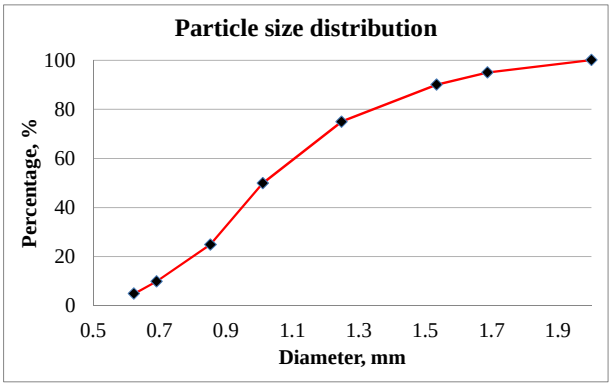
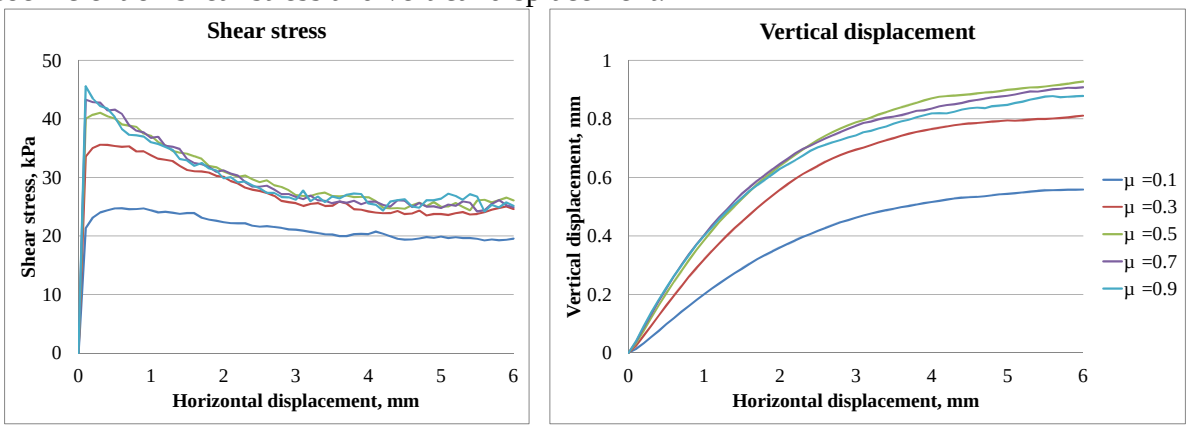



Figure 11. Particle size distribution of idealized granular media used in numerical and laboratory simulations

Young's Modulus was 68.9 GPa, Poisson's coefficient was 0.21 and the particle density was 2430 kg/m$^3$. To minimize the influence of wall friction, which can change the results as shown in the previous example, the friction coefficient between the walls of the shear box and

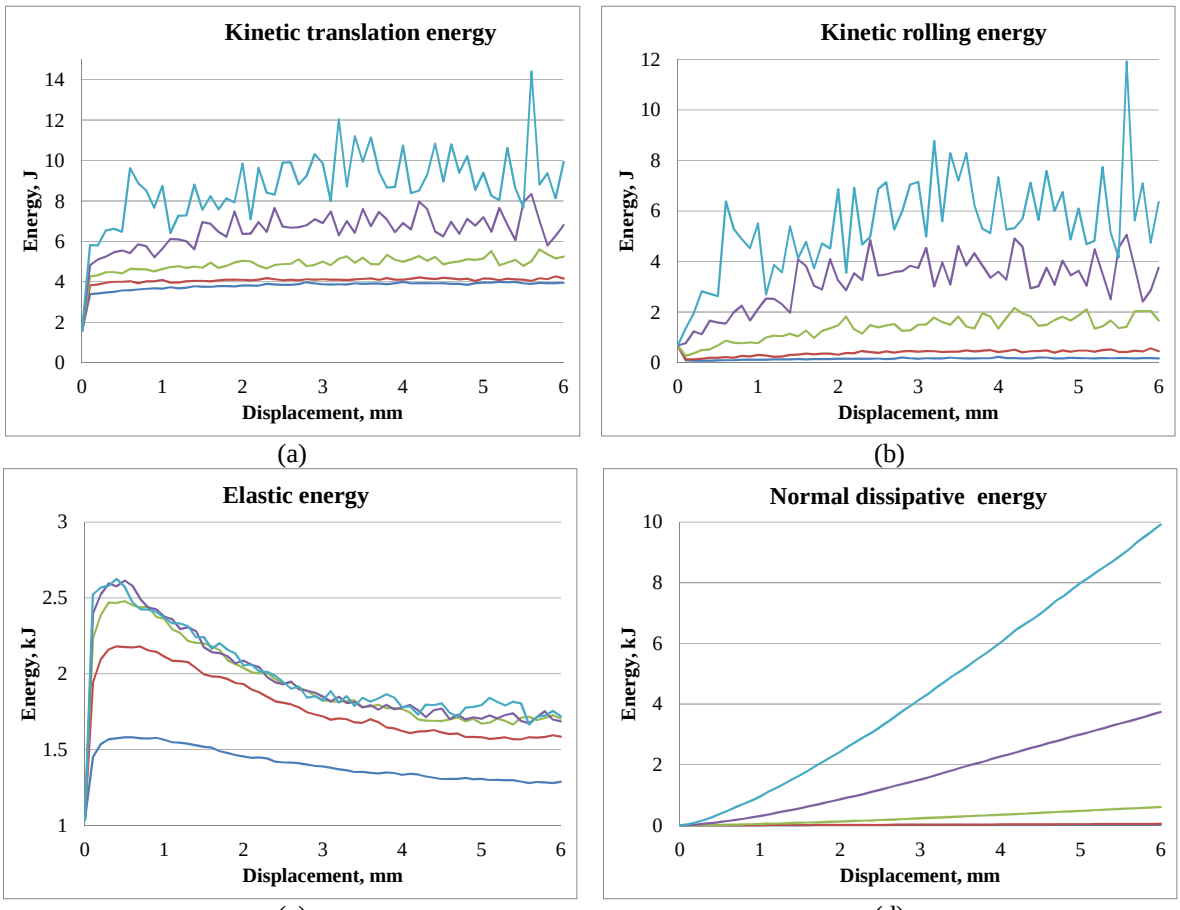
the particles was disregarded. This simplification allows us a better understanding of the inter-particle friction effect. Figure 12 shows the influence of the inter-particle friction coefficient on shear stress and vertical displacement.




(a) (b)

Figure 12. Shear stress and vertical displacement with different friction coefficients

The influence of friction is more important at low coefficient values because the particles can slide more easily on each other. Rolling friction predominates for the highest coefficients of friction. This phenomenon can be studied by analysis of the different energy types as shown in Figure 13.




(a) (b)

(c) (d)

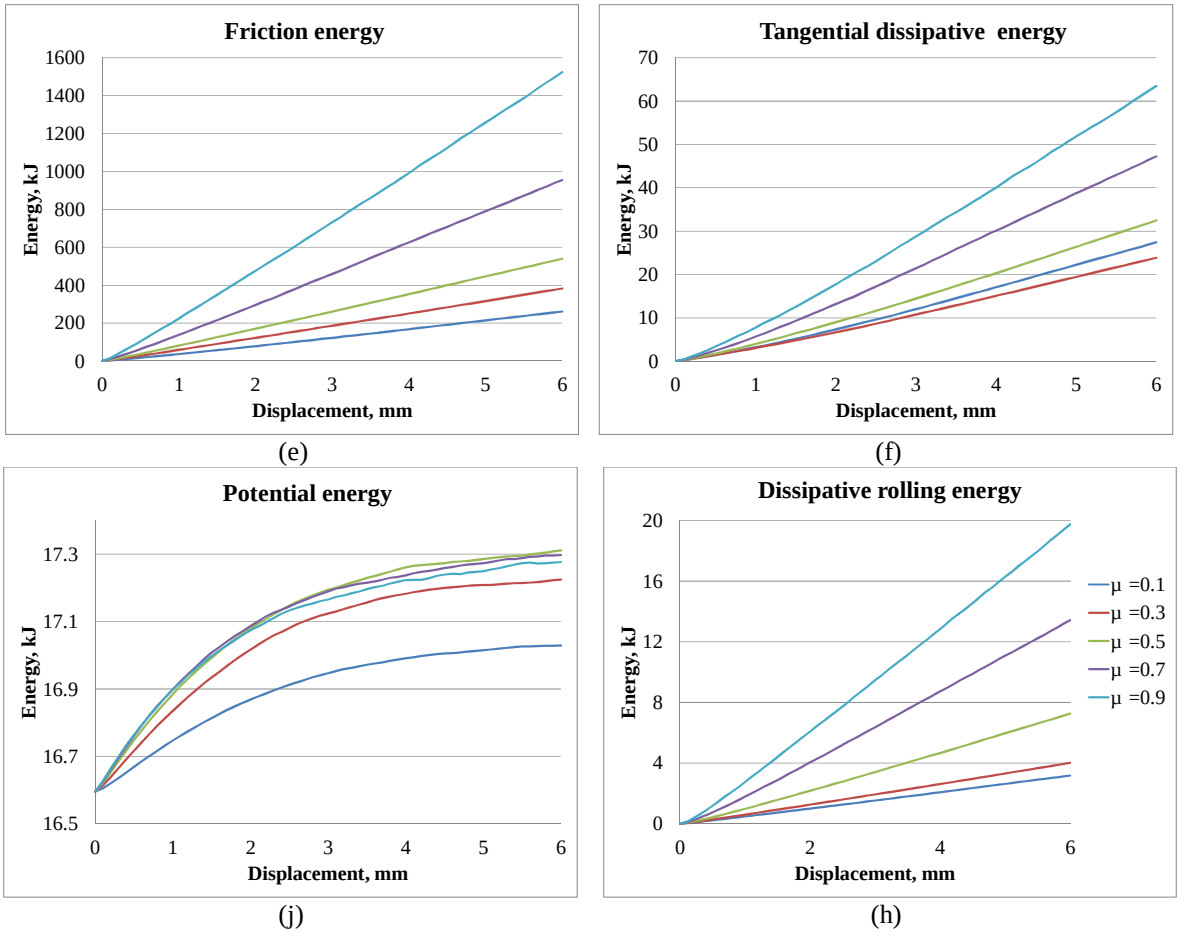
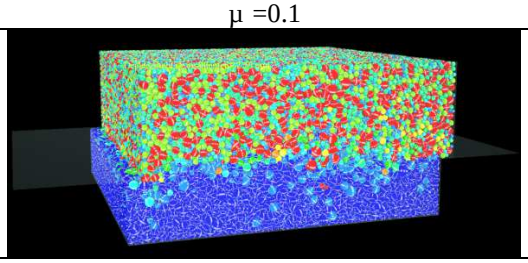
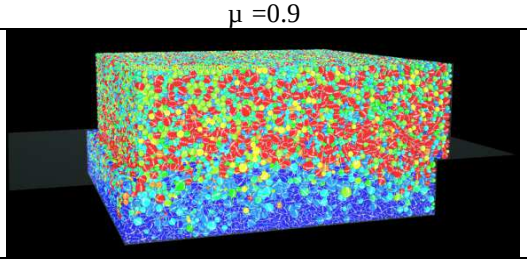
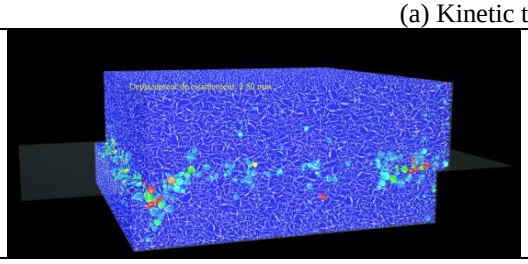
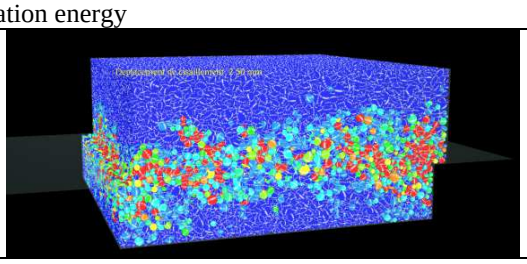
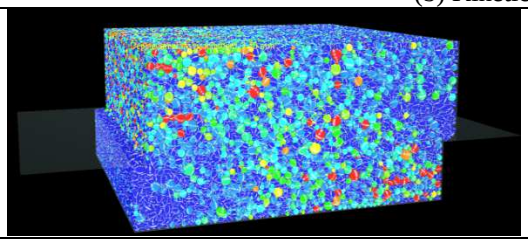
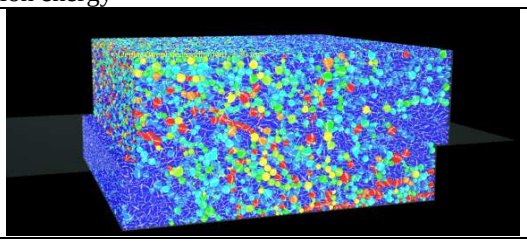
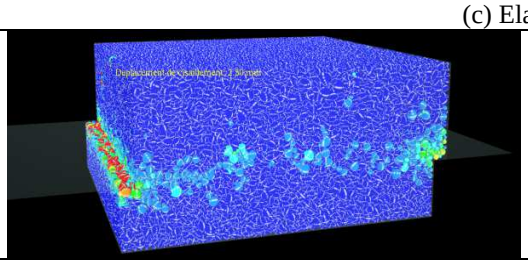
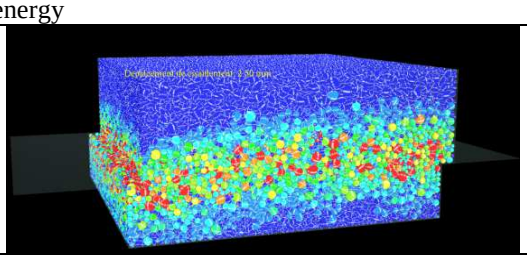
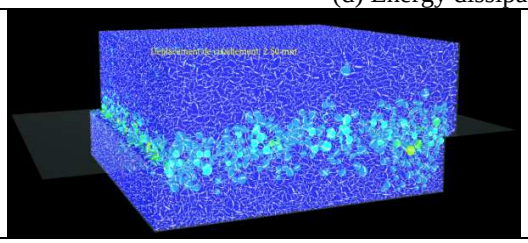
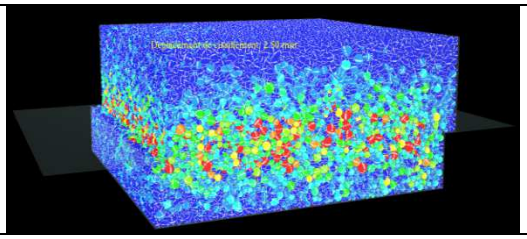
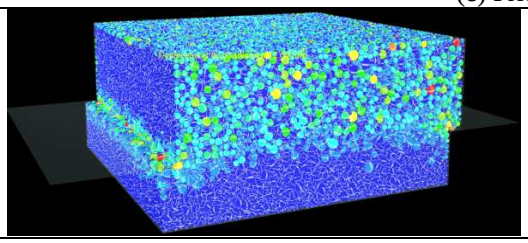
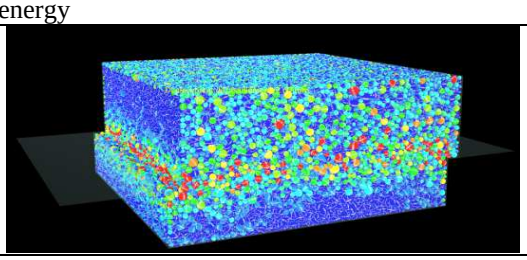
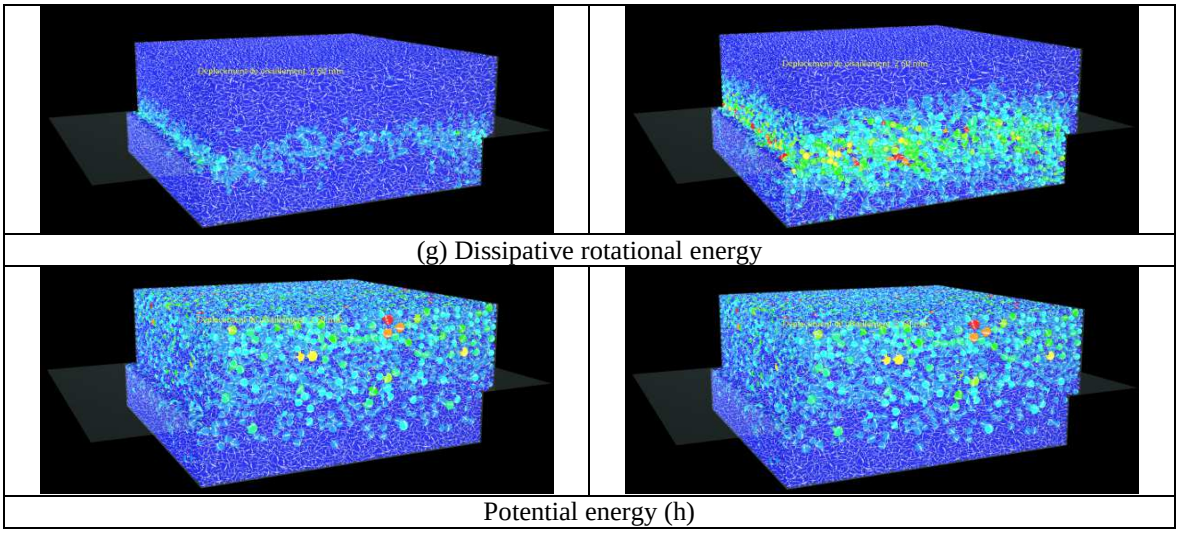



Figure 13. Variation in energies during a shear test

Figure 13 shows that the energy during the shear process increases with the friction coefficient. Due to the difficulty of sliding (see Figure 14a, b), more particles are drawn into movement, which in turn causes the dissipative energy to increase in two directions (see Figure 13d, f). The rolling motion prevails under sliding for high friction coefficients (see Figure 13b); therefore more rolling energy dissipates (see Figure 13h). The value of tangential force also increases with the friction coefficient, and thus more energy dissipates because of friction (see Figure 13e). The change in elastic energy follows the shear stress curve because the normal interaction force between the particles is the main force in this experiment. Some of the external work was spent during the movement by the interlocking of the particles on each other. There is little variation in potential energy because the vertical particle displacement is very small. The curve of potential energy follows of the change in vertical position of the top surface (see figures 12b and 13j)

Figure 14 shows the distribution of energy among particles for the maximum and minimum friction coefficients (0.1and 0.9, respectively). The area where the shear is greatest is clearly visible, and is characterized by larger relative displacements. The dissipation of energy is at its maximum at this surface. The distribution of maximum deformation of particles (elastic energy) is along the line between the upper left surface and lower right surface (Fig. 14c).

| μ =0.1 | μ =0.9 |
|---|---|
|  |  |
| (a) Kinetic translation energy | |
|  |  |
| (b) Kinetic rotation energy | |
|  |  |
| (c) Elastic energy | |
|  |  |
| (d) Energy dissipation in normal direction | |
|  |  |
| (e) Friction energy | |
|  |  |
| (f) Energy dissipation in tangential direction | |



(g) Dissipative rotational energy

Potential energy (h)

Figure 14. Energy for maximum and minimum friction coefficents (displacement of 2.5 mm)

The work of external forces during the shear test is shown in Figure 15.

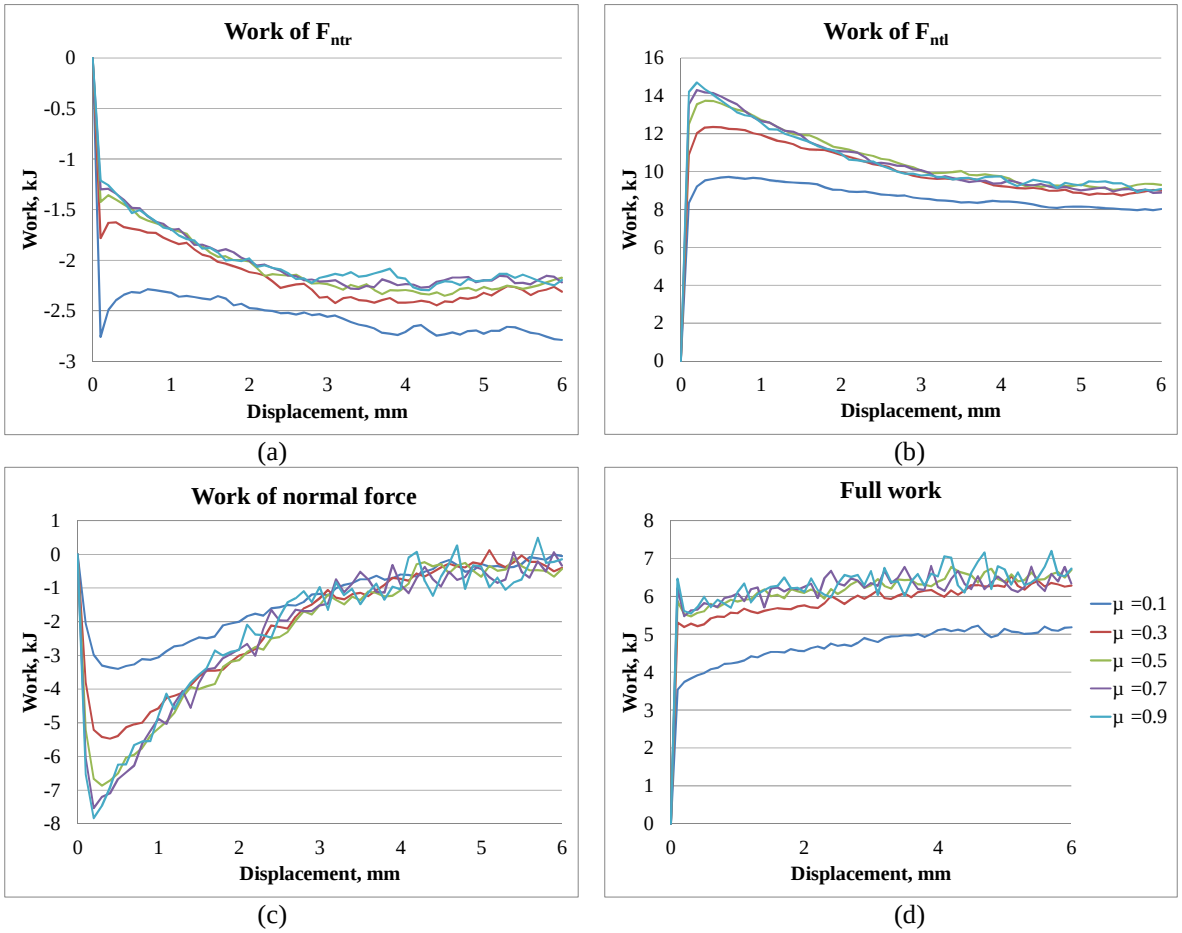



Figure 15. Work of external forces during shear test

Figure 16 shows the inter-particle force distribution before and during the shear test for the minimum and maximum friction coefficients.

## 5 DISCUSSION

We present a model that provides complete microscopic contact and force vector information from a 3D granular system made up of spherical particles subjected to a shear test and, consequently, all the energies can be evaluated. The obtained microscopic data allows us to quantitatively assess the relationship between the macroscopic mechanical response of the granular system and the microscopic measures such as the forces at the contact of each particle. This virtual approach has great potential for improving understanding of other granular systems, including granular response to flow, effects of particle shape and so on. This paper also presents the capability of the model to show, at each time step, images in 3D configurations. From these images, we can see the geometric properties of the particles, their contacts and also all the energy distributions in the granular assembly. Furthermore, 3D tracking of each particle can be extracted.

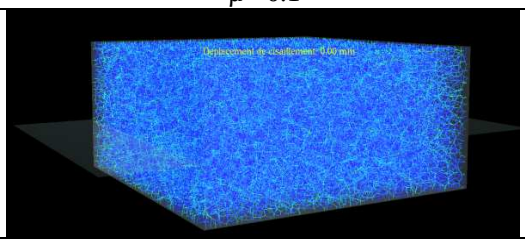
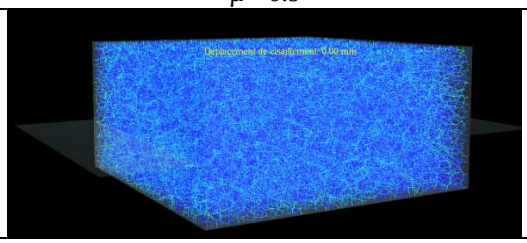
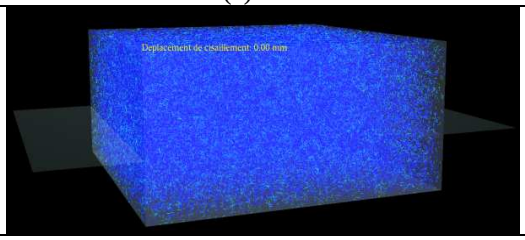
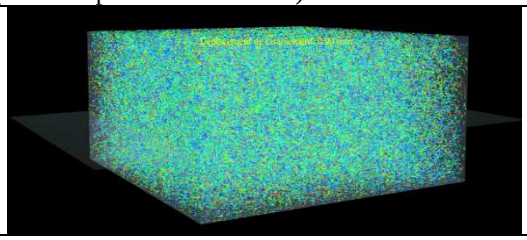
## 6 CONCLUSION

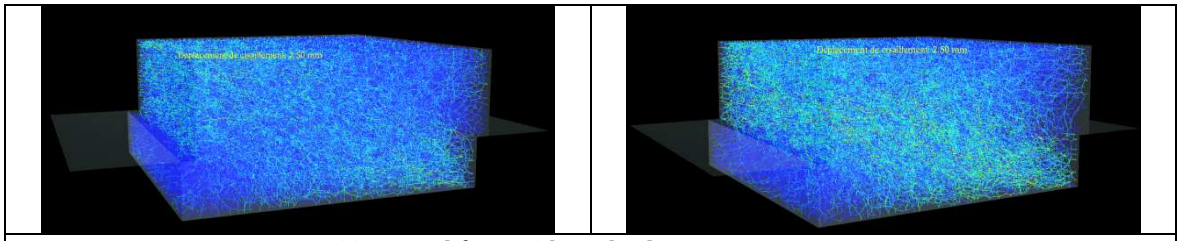
SiGran, a 3D model developed for quantitatively assessing energies involved in a granular system at a particulate scale, is presented in the current work. This model has demonstrated the ability to evaluate the change in different energies during a shear test. This powerful tool offers a new opportunity to link the energy applied to a granular system to the energies at the micro-scale.

A shear test was simulated, and the obtained results clearly indicate that the proposed model is able to solve significant problems associated with the granular system. In fact, it gives access to parameters that are impossible to reach in physical experiments. All the phenomena involved in a granular system can be captured, and the quantification of the different energies at the micro-scale can be assessed. Several findings can be used for different granular fields.

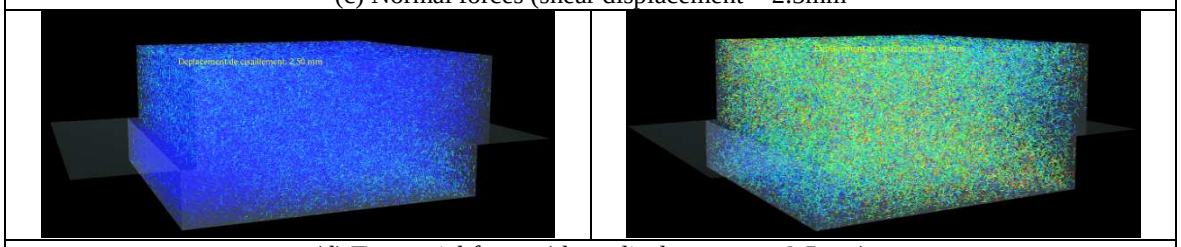
The simulation results at the macro-scale agreed well with those of physical tests. However, further work must be done to reduce the differences between simulation results and physical ones.

Extension of the current approach to soils liquefaction phenomena is underway.

| μ =0.1 | μ =0.9 |
|---|---|
|  |  |
| (a) Normal forces before the test (shear displacement = 0 mm) | |
|  |  |
| (b) Tangential forces (shear displacement = 0 mm) | |



(c) Normal forces (shear displacement = 2.5mm



(d) Tangential forces (shear displacement = 2.5mm)

Figure 16. Forces during the shear test at different displacements

## 7 REFERENCES


[1] P. Cundall, "A computer model for simulating progressive large scale movements in blocky rock systems". *Proc. Symp. Int. Soc. Rock Mechanics*, Nancy, France, vol. 1, paper, No. II-8. (1971)

[2] V. Roubtsova, M. Chekired, B. Morin and M. Karray. "3D virtual laboratory for geotechnical applications: another perspective". Particles 2011, II International Conference on Particle-Based Methods. Fundamentals and Applications, Barcelona, Spain (2011).

[3] M. Chekired and V. Roubtsova "Virtual Reality 3D Interaction of Fluid-Particle Simulations". *V International Conference on Coupled Problems in Science and Engineering*, Ibiza, Spain, June 17-19 (2013).

[4] M. Chekired and V. Roubtsova "Pore-scale study of permeability and tortuosity for flow through particulate media using virtual approach". *ICSE-7, 7th International Conference on Scour and Erosion*, Perth, Australia, December 2–4, 2014.

[5] H. Harlow and J. E. Welch. "Numerical calculation of time-dependent viscous incompressible flow of fluid with a free surface", *The Physics of Fluids*, vol. 8, pp. 2182–2189 (1965).

[6] V. Roubtsova, M. Chekired, Y. Ethier and F. Avendano "SIMSOLS: A 3D virtual laboratory for geotechnical applications", *ICSE-6, 6th International Conference on Scour and Erosion*, Paris, France, August 27–31, (2012)

[7] L. Cui and C. O'Sullivan. 2006. "Exploring the macro- and micro-scale response of an idealized granular material in the shear apparatus", *Géotechnique*, Vol. 56, No. 7, pp. 455–468. 2006

[8] P. Cundall and O. Strack. "A discrete numerical model for granular assemblies", *Geotechnique*, Vol. 29(1), pp. 47–65. (1979).

[9] H. Kruggel-Emden, S. Wirtz and V. Scherer. . "A study on tangential force laws applicable to the discrete element method (DEM) for materials with viscoelastic or plastic behavior", *Chemical Engineering Science*, Vol. 63, Issue 6, pp. 1523–1541

(2008)

[10] R.D. Mindlin and H. Deresiewicz. “Elastic spheres in contact under varying oblique forces” *Transactions of ASME*, Series E., *Journal of Applied Mechanics*, Vol. 20, pp. 327–344. (1953)

[11] Loc Vu-Quoc and Xiang Zhang “An accurate and efficient tangential force-displacement model for elastic frictional contact in particle-flow simulations”, *Mechanics of Materials*, Vol. 31, (1999) pp. 235–269.

[12] Alberto Di Renzo and Francesco Paolo Di Maio “Comparison of contact-force models for the simulation of collisions in DEM-based granular flow codes”, *Chemical Engineering Science*, Vol. 59, pp. 525 – 541(2004)

[13] ASTM D3080 -98 ASTM D3080: Standard test method for direct shear test of soils under consolidated drained conditions (2004).

[14] Collin, “Recherches expérimentales sur les glissements spontanés des terrains argileux” Carilian-Goeurley et Delmont, Paris (1846).

[15] W. Skempton, “A slip in the west bank of the Eau Brink Cut". *Jour. Inst. Civil Engrs.*, Vol. 24, pp. 267-287 (1945).

[16] C. O’Sullivan, J.D. Bray and M.F. Riemer. “An examination of the response of regularly packed specimens of spherical particles using physical tests and discrete element simulations”, ASCE, *J. Engng Mech*., Vol. 130, No. 10, pp. 1140-1150, 2004.

[17] V. Roubtsova, M. Chekired, M. Karray and S. Amirpour. “Work-energy balance for discrete element method using shear stress tests”, *Congress on Numerical Methods in Engineering*, Lisbon, Portugal, 29 June-2 July, 2015.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THREE-DIMENSIONAL BE–FE MODEL OF BUCKET FOUNDATIONS IN POROELASTIC SOILS

**Jacob D. R. Bordón , Juan J. Aznárez, Orlando F. Maeso**

Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numricas en Ingeniera
Universidad de Las Palmas de Gran Canaria
Edificio Central del Parque Cientfico y Tecnolgico del Campus Universitario de Tafira, 35017 Las Palmas de Gran Canaria, Spain
e-mail: {jdrodriguez,jaznarez,omaeso}@iusiani.ulpgc.es

**Keywords:** bucket foundations, poroelasticity, impedances, Dual BEM, BEM–FEM coupling

**Abstract.** *A three-dimensional BE–FE model for the time harmonic analysis of bucket foundations in poroelastic soils is presented. The soil follows the Biot's poroelasticity and is discretized using the BEM. The skirt of the bucket is modeled as a degenerated shell finite element. The soil-structure interaction is taken into account assuming a crack-like boundary from the soil point of view, where the Dual BEM is applied. It is shown that this simple representation is accurate and efficient. This model is applied to an analysis of the impedances of bucket foundations, where the influences of the foundation geometry and soil properties are studied. The study shows that, when considering realistic seabed soils, a poroelastic model should be used for the low-frequency range (*$< 1 - 6$ Hz *depending on the seabed soil). It is shown that this is particularly true for bucket foundations with small length to diameter ratios.*

# 1 INTRODUCTION

Bucket foundations (or suction caisson foundations) are used as anchors and foundations of offshore platforms, and more recently as foundations of offshore wind turbines when suitable water depths and soil conditions are encountered [1]. Foundations of offshore wind turbines experience important horizontal and moment loadings, which are larger for deeper waters. Single bucket or monopod foundations are used for wind turbines installed at moderate water depths. When monopod foundations are not enough to carry these loads, three or four small buckets can be combined to form what are known as tripod or tetrapod foundations. In general, wind turbines with bucket foundations are well suited for water depths between 20 to 50 meters [2].

Despite the experience gained from oil and gas industries, their application to wind turbines faces several new challenges [1, 3]. They must be designed to withstand large horizontal forces and overturning moments, and in addition these are of dynamic nature. These loads mainly comes from steady-state operation of the machine (rotor rotation), wind field, water current field, water waves, tidal effects, and earthquakes. Furthermore, the installation process and the soil conditions of the seabed near the foundation introduce several uncertainties. These designs should be able to operate under such conditions for a number of years in order to be economically viable. Therefore, it is necessary to advance towards the development of rigorous models able to take into account realistic conditions.

Many aspects of the installation and design of bucket foundations have been studied, and the literature is large. A very complete review about bearing capacity and installation was published by Foglia and Ibsen [4]. In the context of dynamics, a recent work of Kourkoulis et al. [5] uses a non-linear FEM model to study the behaviour of bucket foundations of offshore wind turbines under lateral monotonic, cyclic, and earthquake loading. They give an interesting discussion about the interface conditions between soil and foundation. Liingaard et al. [6] studied the impedances of bucket foundations in elastic soils, including the variation of these under changes of geometry and soil properties.

In the present work, dynamic stiffnesses of bucket foundations buried in poroelastic soils are studied. A simple but accurate boundary element – finite element (BE–FE) model is developed to this aim. Bucket foundations with different skirt length to diameter ratios buried in different realistic seabed soils are considered. Also, the effect of different contact conditions between the lid and the seabed is studied.

The rest of the paper is organized as follows. In Section 2, the boundary element – finite element model is described. In Section 3, impedances of bucket foundations buried in different poroelastic seabeds are obtained and discussed. Finally, in Section 4 the main conclusions are given.

# 2 BE–FE MODEL

## 2.1 Conventional and Dual BEM for three-dimensional Biot's poroelasticity

The soil is considered to be a homogeneous poroelastic half-space following the Biot's theory of poroelasticity. Given its unbounded nature, the BEM is used to numerically treat it. A particular feature of the proposed model is that, unlike Liingaard et al. [6], the problem can be handled directly without needing any artificial interfaces. This is achieved thanks to the usage of the Dual BEM.

The governing equations of Biot's poroelasticity [7] in the time domain can be written as:

$$\mu\boldsymbol{\nabla}^2\mathbf{u}+\boldsymbol{\nabla}\left[\left(\lambda+\mu+Q^2/R\right)(\boldsymbol{\nabla}\cdot\mathbf{u})+Q\,(\boldsymbol{\nabla}\cdot\mathbf{U})\right]+\mathbf{X}=\rho_{11}\ddot{\mathbf{u}}+\rho_{12}\ddot{\mathbf{U}}+b\left(\dot{\mathbf{u}}-\dot{\mathbf{U}}\right) \quad (1)$$

$$\boldsymbol{\nabla}\left[Q\,(\boldsymbol{\nabla}\cdot\mathbf{u})+R\,(\boldsymbol{\nabla}\cdot\mathbf{U})\right]+\mathbf{X}'=\rho_{12}\ddot{\mathbf{u}}+\rho_{22}\ddot{\mathbf{U}}-b\left(\dot{\mathbf{u}}-\dot{\mathbf{U}}\right) \quad (2)$$

and the stress-strain relationships are:

$$\tau_{ij}=\delta_{ij}\left[\left(\lambda+Q^2/R\right)(\boldsymbol{\nabla}\cdot\mathbf{u})+Q\,(\boldsymbol{\nabla}\cdot\mathbf{U})\right]+\mu\left(u_{i,j}+u_{j,i}\right),\quad i,j=1,2,3 \quad (3)$$

$$\tau=Q\,(\boldsymbol{\nabla}\cdot\mathbf{u})+R\,(\boldsymbol{\nabla}\cdot\mathbf{U}) \quad (4)$$

where $u_i$, $\tau_{ij}$ and $\mathbf{X}$ are respectively displacements, stresses and body forces in the solid phase, and $U_i$, $\tau$ and $\mathbf{X}'$ are respectively displacements, equivalent stress and body forces in the fluid phase. The poroelastic medium has the following properties: $\lambda$ and $\mu$ are the Lamé's parameters of the solid phase, $Q$ and $R$ are the Biot's coupling parameters, $b$ is the dissipation constant, and $\rho_{11}=(1-\phi)\rho_{\mathrm{s}}+\rho_{\mathrm{a}}$, $\rho_{12}=-\rho_{\mathrm{a}}$, $\rho_{22}=\phi\rho_{\mathrm{f}}+\rho_{\mathrm{a}}$, being $\phi$ the porosity, $\rho_{\mathrm{s}}$ the solid phase density, $\rho_{\mathrm{f}}$ the fluid phase density, and $\rho_{\mathrm{a}}$ the additional apparent density. The additional apparent density $\rho_{\mathrm{a}}$ is obtained from $\rho_{\mathrm{a}}=(\alpha-1)\phi\rho_{\mathrm{f}}$, where $\alpha$ is the tortuosity [8]. The dissipation constant $b$ is related to the hydraulic conductivity $k$ by the relationship $b=\rho_{\mathrm{f}}g\phi^2/k$, where $g$ is the gravitational acceleration [9]. In the frequency domain $\omega$, Equations (1-2) can be written as:

$$\mu\boldsymbol{\nabla}^2\mathbf{u}+\boldsymbol{\nabla}\left[\left(\lambda+\mu+Q^2/R\right)(\boldsymbol{\nabla}\cdot\mathbf{u})+Q\,(\boldsymbol{\nabla}\cdot\mathbf{U})\right]+\mathbf{X}=-\omega^2\hat{\rho}_{11}\mathbf{u}-\omega^2\hat{\rho}_{12}\mathbf{U} \quad (5)$$

$$\boldsymbol{\nabla}\left[Q\,(\boldsymbol{\nabla}\cdot\mathbf{u})+R\,(\boldsymbol{\nabla}\cdot\mathbf{U})\right]+\mathbf{X}'=-\omega^2\hat{\rho}_{12}\mathbf{u}-\omega^2\hat{\rho}_{22}\mathbf{U} \quad (6)$$

where $\hat{\rho}_{11}=\rho_{11}-ib/\omega$, $\hat{\rho}_{22}=\rho_{22}-ib/\omega$ and $\hat{\rho}_{12}=\rho_{12}+ib/\omega$. It is well known that three bulk modes exists: two longitudinal modes (P1 and P2) and one transverse mode (S); and their associated wavenumbers are denoted respectively as $k_1$, $k_2$ and $k_3$.

The BEM is based on the usage of Boundary Integral Equations (BIE), which are used to build a solvable linear system of equations after its discretization. Domínguez [10, 11] presented a BEM based on a Singular BIE (SBIE) for two-dimensional Biot's poroelasticity, and Maeso et al. [12] extended it to three-dimensions. One of the advantages of this family of BIEs over others is the reduction of the variables related to the fluid phase to two: the equivalent stress $\tau$ and the normal displacement $U_n$. Another advantage is the possibility of writing the fundamental solution and its derivatives in a way that resembles the fundamental solutions of acoustics and elastodynamics and their derivatives.

Let $\Omega$ be a poroelastic region, and $\Gamma=\partial\Omega$ its boundary with outward unit normal $\mathbf{n}$. Using the weighted residual formulation proposed by Domínguez [10, 12], the SBIE at a collocation point $\mathbf{x}^{\mathrm{i}}$ can be written as:

$$\begin{bmatrix} Jc^{\mathrm{i}} & 0 \\ 0 & c^{\mathrm{i}}_{lk}\end{bmatrix}\begin{Bmatrix}\tau^{\mathrm{i}}\\ u^{\mathrm{i}}_k\end{Bmatrix}+\oint_\Gamma\begin{bmatrix}-(U^*_{n00}+JX'^*_j n_j)\,t^*_{0k} \\ -U^*_{nl0} \qquad\qquad t^*_{lk}\end{bmatrix}\begin{Bmatrix}\tau\\ u_k\end{Bmatrix}\mathrm{d}\Gamma=\int_\Gamma\begin{bmatrix}-\tau^*_{00} & u^*_{0k}\\ -\tau^*_{l0} & u^*_{lk}\end{bmatrix}\begin{Bmatrix}U_n\\ t_k\end{Bmatrix}\mathrm{d}\Gamma$$

$$\mathbf{C}^{\mathrm{i}}_{\mathrm{S}}\mathbf{u}^{\mathrm{i}}+\oint_\Gamma\mathbf{T}^*\mathbf{u}\,\mathrm{d}\Gamma=\int_\Gamma\mathbf{U}^*\mathbf{t}\,\mathrm{d}\Gamma \quad (7)$$

where $l,k=1,2,3$, and $J=1/(\hat{\rho}_{22}\omega^2)$. The vector $\mathbf{u}$ contains the primary variables: fluid equivalent stress $\tau$ and solid displacements $u_k$; and $\mathbf{t}$ contains the secondary variables: fluid normal displacement $U_n=U_jn_j$ and solid traction $t_k=\tau_{kj}n_j$. The superscript $\square^{\mathrm{i}}$ is not an

index, and indicates that the corresponding variable is related to the collocation point. The free-term matrix $\mathbf{C}^{\mathrm{i}}_{\mathrm{S}}$ contains the potential free-term $c^{\mathrm{i}}$ and the elastostatic free-term $c^{\mathrm{i}}_{lk}$, which are $c^{\mathrm{i}} = 1$ and $c^{\mathrm{i}}_{lk} = \delta_{lk}$ for an interior collocation point ($\mathbf{x}^{\mathrm{i}} \in \Omega$), and $c^{\mathrm{i}} = 1/2$ and $c^{\mathrm{i}}_{lk} = 1/2\delta_{lk}$ for a smooth boundary collocation point ($\mathbf{x}^{\mathrm{i}} \in \Gamma$, $\Gamma(\mathbf{x}^{\mathrm{i}}) \in \mathcal{C}^1$). As usual, the notation $\oint$ stands for an integral in the Cauchy Principal Value (CPV) sense, which is evaluated as described in [12].

The Hypersingular BIE (HBIE) for two-dimensional problems was obtained by Bordón et al. [13]. Likewise, the three-dimensional HBIE is built by establishing the secondary variables at the collocation point:

$$U^{\mathrm{i}}_n = U^{\mathrm{i}}_j n^{\mathrm{i}}_j = -J\tau^{\mathrm{i}}_{,j} n^{\mathrm{i}}_j - Z u^{\mathrm{i}}_j n^{\mathrm{i}}_j \tag{8}$$

$$t^{\mathrm{i}}_l = \tau^{\mathrm{i}}_{lj} n^{\mathrm{i}}_j = \left[\lambda u^{\mathrm{i}}_{m,m}\delta_{lj} + \mu\left(u^{\mathrm{i}}_{l,j} + u^{\mathrm{i}}_{j,l}\right)\right] n^{\mathrm{i}}_j + \frac{Q}{R}\tau^{\mathrm{i}} n^{\mathrm{i}}_l \tag{9}$$

where $Z = \hat{\rho}_{12}/\hat{\rho}_{22}$, $\mathbf{n}^{\mathrm{i}}$ is the unit normal vector at the collocation point, and the comma derivative notation denotes $\partial/\partial x^{\mathrm{i}}_k$. Therefore, a mix of the SBIE and its derivatives with respect to the collocation point is required to build the HBIE. After carrying out all the required operations, the HBIE at a collocation point $\mathbf{x}^{\mathrm{i}}$ with unit normal $\mathbf{n}^{\mathrm{i}}$ can be written as:

$$\begin{gathered}\begin{bmatrix} c^{\mathrm{i}} & 0 \\ 0 & c^{\mathrm{i}}_{lk} \end{bmatrix} \begin{Bmatrix} U^{\mathrm{i}}_n \\ t^{\mathrm{i}}_k \end{Bmatrix} + \fint_\Gamma \begin{bmatrix} -s^*_{00} & s^*_{0k} \\ -s^*_{l0} & s^*_{lk} \end{bmatrix} \begin{Bmatrix} \tau \\ u_k \end{Bmatrix} \mathrm{d}\Gamma = \oint_\Gamma \begin{bmatrix} -d^*_{00} & d^*_{0k} \\ -d^*_{l0} & d^*_{lk} \end{bmatrix} \begin{Bmatrix} U_n \\ t_k \end{Bmatrix} \mathrm{d}\Gamma \\ \mathbf{C}^{\mathrm{i}}_{\mathrm{H}}\mathbf{t}^{\mathrm{i}} + \fint_\Gamma \mathbf{S}^*\mathbf{u}\,\mathrm{d}\Gamma = \oint_\Gamma \mathbf{D}^*\mathbf{t}\,\mathrm{d}\Gamma\end{gathered} \tag{10}$$

where the notation $\fint$ stands for an integral in the Hadamard Finite Part (HFP) sense. The presence of a HFP integral imposes that the primary variables at the collocation point must have continuous first derivatives, i.e. $\tau(\mathbf{x}^{\mathrm{i}}),\ u_k(\mathbf{x}^{\mathrm{i}}) \in \mathcal{C}^1$. Using this fact, a meaningful HBIE can be obtained once a regularization process based on the work of Domínguez et al. [14] is performed.

Equations (7) and (10) correspond to BIEs for interior collocation points, or boundary collocation points at ordinary boundaries. When the collocation point is located at a crack-like boundary, both BIEs have to be modified. A crack-like boundary has two boundaries geometrically coincident but with opposite orientations, denoted as positive $+$ and negative $-$ faces. Hence, the SBIE and HBIE when the collocation point $\mathbf{x}^{\mathrm{i}}$ is located at a crack-like boundary can be written as:

$$\frac{1}{2}\begin{bmatrix} J & 0 \\ 0 & \delta_{lk} \end{bmatrix} \left(\mathbf{u}^{\mathrm{i}+} + \mathbf{u}^{\mathrm{i}-}\right) + \oint_\Gamma \mathbf{T}^*\mathbf{u}\,\mathrm{d}\Gamma = \int_\Gamma \mathbf{U}^*\mathbf{t}\,\mathrm{d}\Gamma \tag{11}$$

$$\frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & \delta_{lk} \end{bmatrix} \left(\mathbf{t}^{\mathrm{i}+} - \mathbf{t}^{\mathrm{i}-}\right) + \fint_\Gamma \mathbf{S}^*\mathbf{u}\,\mathrm{d}\Gamma = \oint_\Gamma \mathbf{D}^*\mathbf{t}\,\mathrm{d}\Gamma \tag{12}$$

where it has been assumed that $\Gamma(\mathbf{x}^{\mathrm{i}}) \in \mathcal{C}^1$. Both Equations (11) and (12) have to be used simultaneously in order to solve problems where crack-like boundaries are present. When considered this way, they are known as Dual BIEs, and their application to the BEM is called the Dual BEM [15, 16]. As explained before, the HBIE requires that $\tau(\mathbf{x}^{\mathrm{i}}),\ u_k(\mathbf{x}^{\mathrm{i}}) \in \mathcal{C}^1$, thus the collocation at crack-like boundaries must be performed carefully. Aliabadi and co-workers [15, 16] use discontinuous boundary elements with nodes already located at points where this condition is fulfilled. Another approach is that of Domínguez et al. [14], where standard continuous boundary elements with multiple non-nodal collocation is used. The latter is considered in the present work since, as it will become clear in the next section, continuous boundary elements are much more appropriate for the proposed coupling.

### 2.2 Degenerated shell finite element

The bucket foundation is considered to be a massless linear elastic solid. It is modelled with degenerated shell finite elements. The degenerated shell FE has been formulated and implemented following Oñate [17], and a robust 8-noded quadrilateral element with reduced integration is used.

### 2.3 BE–FE coupling

The BE–FE coupling is done at the level of discretized equations. The boundary element mesh and the finite element mesh must be conforming. A perfect bonding between the mid-surface of the shell and the crack-like boundary of the soil is considered. It means that rotations are not taken into account in the coupling. Given the small thickness to length ratio of the bucket skirt, their contribution can be neglected. Furthermore, since the bucket is made of steel, an impermeable interface between the shell and the poroelastic soil is assumed.

Consider $n_j^+$, $U_n^+$, $u_i^+$, $\tau^+$ and $t_i^+$ as respectively the unit normal, fluid normal displacement, solid displacement, fluid equivalent stress and solid traction of the positive face of the crack-like boundary (soil), and analogously for the negative face. Also, consider $u_i^{\mathrm{s}}$ as the displacement of the mid-surface, and $t_i^{\mathrm{s}}$ as the distributed load on the mid-surface. Then, the compatibility and equilibrium coupling conditions between the crack-like boundary and the mid-surface of the shell can be written as:

$$U_n^+ = u_j^+ n_j^+,\ U_n^- = u_j^- n_j^-,\ u_i^+ = u_i^{\mathrm{s}},\ u_i^- = u_i^{\mathrm{s}} \tag{13}$$

$$\tau^+ n_i^+ + t_i^+ + \tau^- n_i^- + t_i^- + t_i^{\mathrm{s}} = 0 \tag{14}$$

## 3 IMPEDANCES OF BUCKET FOUNDATIONS

A bucket foundation is composed of a rigid lid with diameter $D$, and a flexible skirt of length $L$ and thickness $t$. Using six degrees of freedom at the center of the lid, it is possible to build an impedance matrix $\mathbf{S}$ relating the forces and moments $\mathbf{R}$ produced by unitary displacements and rotations $\mathbf{U}$. Since buckets are axisymmetric, the impedance matrix has five different impedance functions: horizontal ($S_{\mathrm{HH}}$), vertical ($S_{\mathrm{VV}}$), rocking ($S_{\mathrm{MM}}$), horizontal-rocking coupling ($S_{\mathrm{MH}}$), and torsional ($S_{\mathrm{TT}}$). In the present paper, all impedance functions are studied except the torsional one. For the sake of brevity, the same notation and normalization procedure as Liingaard et al. [6] is used.

Elastic soils can be defined by a small set of properties, for example shear modulus $\mu$, Poisson's ratio $\nu$, density $\rho$ and a hysteretic damping ratio $\xi$ ($\mu^* = \mu(1 + i2\xi)$). Hence, fully dimensionless studies can be carried out by defining some shape factors of the structure, a dimensionless frequency $a_0$ with the help of a length of the structure and a wave velocity of the soil, and setting the Poisson's ratio and damping ratio of the soil. In the case of poroelastic soils, this task becomes impractical due to the number of properties involved, and the difficulties of knowing if a given set of values of the properties represents a realistic soil or not. For these reasons, we have decided to use realistic seabed soils taken from Buchanan and Gilbert [18], see Table 1. All results are shown using a dimensionless frequency $a_0 = \omega R/c_{\mathrm{S}}^{\mathrm{u}}$, where $R$ is the radius of the bucket, and $c_{\mathrm{S}}^{\mathrm{u}} = \sqrt{\mu/(\phi\rho_{\mathrm{f}} + (1-\phi)\rho_{\mathrm{s}})}$ is the undrained S-wave velocity.

The bucket foundation is considered massless ($\rho = 0\ \mathrm{kg/m^3}$), with a Young's modulus $E = 210\ \mathrm{GPa}$, Poisson's ratio $\nu = 1/4$ and hysteretic damping ratio $\xi = 0.01$ ($E^* = E(1 + i2\xi)$). The diameter is $D = 10\ \mathrm{m}$, and the thickness $t = 0.05\ \mathrm{m}$. Because of the nature of the BE-FE coupling presented, the mass distribution through the soil-structure interface is continuous according to the density of the soil despite the structure is considered massless.

| **Property, symbol and units** | **Coarse sand and fine gravel (sb1)** | **Coarse sand (sb2)** | **Fine sand (sb3)** | **Silty clay (sb4)** | **Silty sand (sb5)** |
|---|---|---|---|---|---|
| Frame shear modulus $\mathrm{Re}\,(\mu^*)$ [MPa] | 12.50 | 74.00 | 7.12 | 0.79 | 41.00 |
| Frame shear modulus $\mathrm{Im}\,(\mu^*)$ [MPa] | 4.50 | 4.70 | 0.23 | 0.03 | 7.90 |
| Frame bulk modulus $\mathrm{Re}\,(K^*)$ [MPa] | 27.10 | 52.00 | 9.49 | 3.67 | 29.00 |
| Frame bulk modulus $\mathrm{Im}\,(K^*)$ [MPa] | 0.90 | 0.74 | 0.30 | 0.12 | 1.30 |
| Poisson's ratio $\nu$ [–] | 0.30 | 0.02 | 0.20 | 0.40 | 0.02 |
| Porosity $\phi$ [–] | 0.30 | 0.38 | 0.43 | 0.68 | 0.65 |
| Fluid bulk modulus $K_\mathrm{f}$ [GPa] | 2.38 | 2.40 | 2.39 | 2.38 | 2.40 |
| Biot's coupling paramater $Q$ [GPa] | 1.666 | 1.488 | 1.362 | 0.762 | 0.840 |
| Biot's coupling paramater $R$ [GPa] | 0.714 | 0.912 | 1.028 | 1.618 | 1.560 |
| Fluid density $\rho_\mathrm{f}$ [kg/m$^3$] | 1000 | 1000 | 1000 | 1000 | 1000 |
| Solid density $\rho_\mathrm{s}$ [kg/m$^3$] | 2680 | 2710 | 2670 | 2680 | 2670 |
| Tortuosity $\alpha$ [–] | 1.25 | 1.25 | 1.25 | 3.00 | 3.00 |
| Additional apparent den. $\rho_\mathrm{a}$ [kg/m$^3$] | 75 | 95 | 107.5 | 1360 | 1300 |
| Fluid viscosity $\eta$ [mPa $\cdot$ s] | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 |
| Permeability $\kappa$ [m$^2$] | $2.6 \cdot 10^{-10}$ | $7.5 \cdot 10^{-11}$ | $3.1 \cdot 10^{-14}$ | $5.2 \cdot 10^{-14}$ | $6.3 \cdot 10^{-15}$ |
| Hydraulic conductivity $k$ [m/s] | $2.5 \cdot 10^{-3}$ | $7.3 \cdot 10^{-4}$ | $3.0 \cdot 10^{-7}$ | $5.1 \cdot 10^{-7}$ | $6.2 \cdot 10^{-8}$ |
| Disipation constant $b$ [N $\cdot$ s/m$^4$] | $3.52 \cdot 10^{5}$ | $1.95 \cdot 10^{6}$ | $5.99 \cdot 10^{9}$ | $8.98 \cdot 10^{9}$ | $6.74 \cdot 10^{10}$ |
| Undrained Poisson's ratio $\nu^\mathrm{u}$ [–] | 0.4992153 | 0.4942119 | 0.4993609 | 0.4998878 | 0.4945113 |
| Bulk density $\rho$ [kg/m$^3$] | 2176 | 2060 | 1952 | 1538 | 1585 |
| Undrained S-wave velocity $c_\mathrm{S}^\mathrm{u}$ [m/s] | 75.8 | 189.5 | 60.4 | 22.6 | 160.9 |

Table 1: Properties of seabed soils taken from Buchanan [18]. Top: poroelastic medium. Bottom: undrained solid.

Impedances are calculated using a BE-FE model based on the methodology described in the previous section. Figure 1 shows a mesh used in the calculations. Taking into account the symmetric nature of the geometry, only one-quarter of the domain is discretized. The soil region $\Omega_\mathrm{soil}$ has three BE boundaries: the seabed free-surface $\Gamma_{\mathrm{free-surface}}$, the soil-skirt interface $\Gamma_{\mathrm{soil-skirt}}$ (a crack-like boundary), and the bucket lid $\Gamma_\mathrm{lid}$. The skirt region $\Omega_\mathrm{skirt}$ is a mesh of degenerated shell FE. For the sake of clarity in Figure 1, the skirt $\Omega_\mathrm{skirt}$ is not located at its real position, which is exactly in the position of $\Gamma_{\mathrm{soil-skirt}}$. The seabed free-surface $\Gamma_{\mathrm{free-surface}}$ is a permeable traction-free boundary, i.e. $\tau = 0$ and $t_k = 0$. The bucket lid $\Gamma_\mathrm{lid}$ has prescribed fluid and solid displacements according to the impedance that is being being calculated. Shell FE nodes in ($x > 0$, $y > 0$, $z = 0$) and in $zx$ and $yz$ symmetry planes are 6 DOF shell nodes, while the rest are 5 DOF nodes. By doing so, it is easy to establish the prescribed displacements and rotations to the 6 DOF nodes according to the impedance that is being calculated and the symmetric/anti-symmetric conditions imposed by the displacement field. Both $\Gamma_{\mathrm{soil-skirt}}$ and $\Omega_\mathrm{skirt}$ are discretized with conforming meshes of 8-noded quadrilateral elements. Boundaries $\Gamma_\mathrm{lid}$ and $\Gamma_{\mathrm{free-surface}}$ are discretized with 6-noded triangular elements. The size of the elements of the foundation and its surroundings is at least of 6 elements per wavelength, while at least 4 elements per wavelength is used beyond it.

### 3.1 Validation

In order to check the validity of the formulation and the models, a comparison between several results of Liingaard et al. [6] and results from our BE-FE model is done. Figure 2 shows impedances (normalized magnitude and angle) for bucket foundations with several length to diameter ratios $L/D = \{1/4, 1, 2\}$. The given elastic soil properties are $\mu = 1$ MPa, $\nu = 1/3$

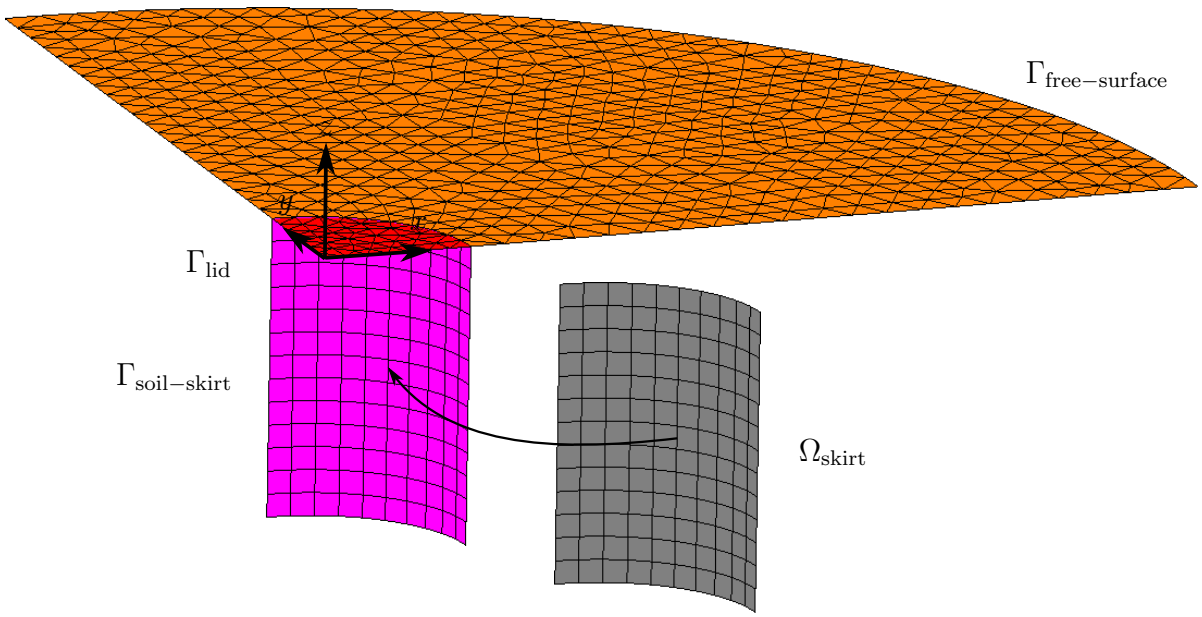



Figure 1: Description of a mesh used in the calculations ($L/D = 1$)

and $\xi = 0.025$. In our model, the same properties are used for the solid, the fluid is considered to be air, and a small porosity is used $\phi \to 0$. Figure 2 demonstrates complete agreement between results. Although not shown here, all other static and dynamic results presented in [6] also agree with results obtained by our model.

### 3.2 Results and discussion

Seabed soils taken from Buchanan and Gilbert [18], see Table 1, cover a wide range of possible realistic soils, from gravels, sands, silts, to clays. These soils are denoted as "sb1" to "sb5" in the following tables and graphs. Three length to diameter ratios $L/D = \{1/4, 1, 2\}$ are studied. Table 2 shows the dimensionless quasi-static stiffnesses for all cases, where they are calculated for a small dimensionless frequency $a_0 = 10^{-6}$. Nondimensionalization of impedances is performed using the shear modulus $\mu$ of the soil and the radius $R$ of the bucket. Figures 3 to 5 show the impedances for all cases, where in the low-frequency range ($a_0 = [10^{-6}, 1]$) only their magnitudes are analysed, and in a broader frequency range ($a_0 = [0, 6]$) also their angles are shown. Taking into account the definition of the dimensionless frequency $a_0$, the low-frequency range corresponds approximately to frequencies below $1 - 6$ Hz depending on the seabed soil. Also, the broader frequency range corresponds approximately to frequencies between $1 - 6$ Hz and $40$ Hz depending on the seabed soil.

Dimensionless quasi-static stiffnesses are similar in magnitude to those obtained by Liingaard et al [6] for elastic soils, considering the seabed as a drained elastic soil. In fact, Table 2 includes the results using an elastic solid with the drained conditions of the porous medium, and the discrepancy is small. Differences are due to a not sufficiently small dimensionless frequency for the calculation of the quasi-static stiffness.

As can be seen in the left hand side graphs of Figures 3 to 5, impedance functions are almost constant and approximately equal to the quasi-static value in the low-frequency range. This is characteristic of any elastic soil, which is even more smooth. In the case of poroelastic soils, the smaller length to diameter ratio the less regular behaviour at low-frequencies. In the case of buckets with $L/D = 1/4$, it is very noticeable the variation of impedances when $a_0 \to 0$.

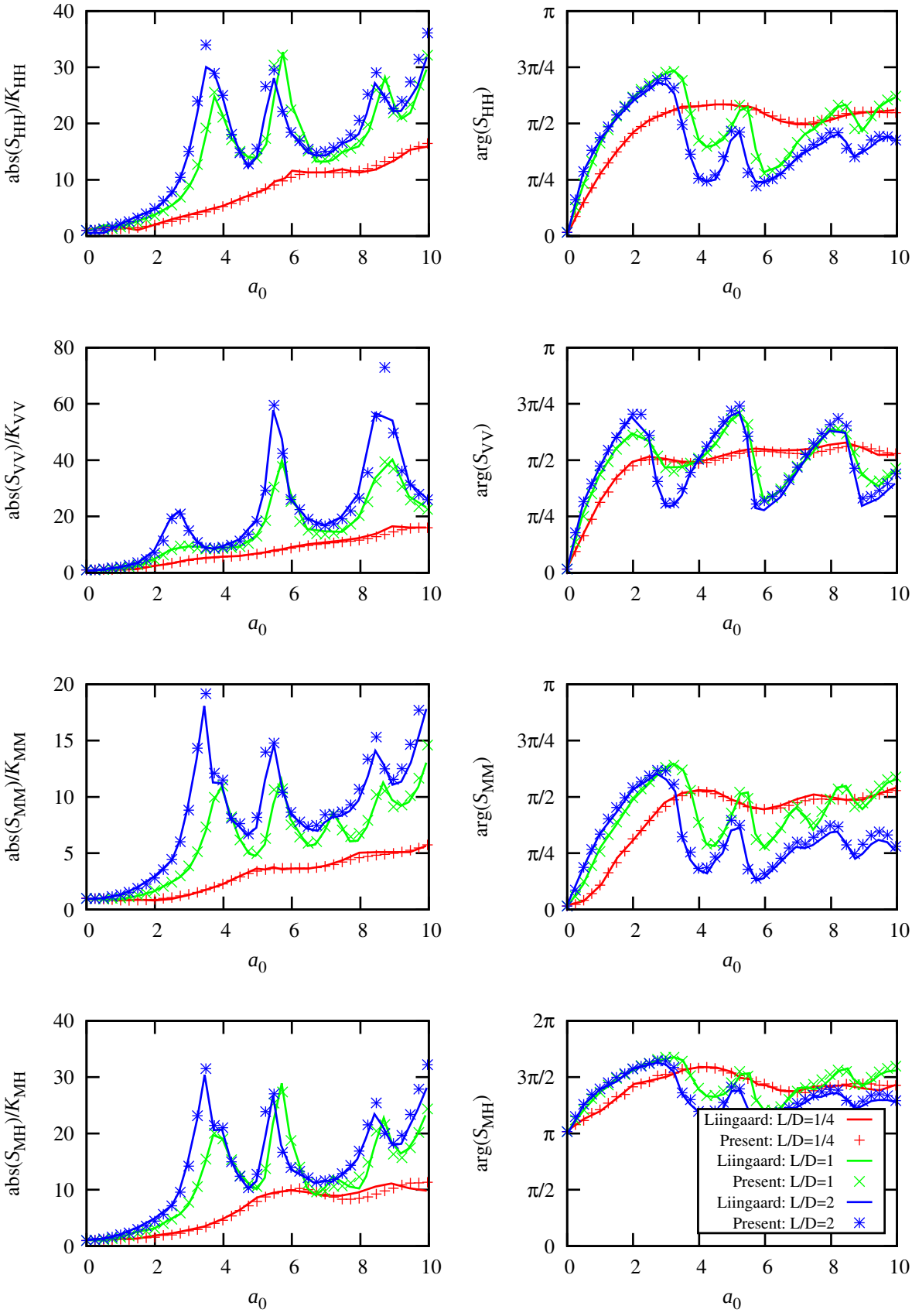



Figure 2: Comparison between Liingaard et al. [6] and the present approach. From top to bottom: normalized horizontal, vertical, rocking, and horizontal-rocking coupling impedances.

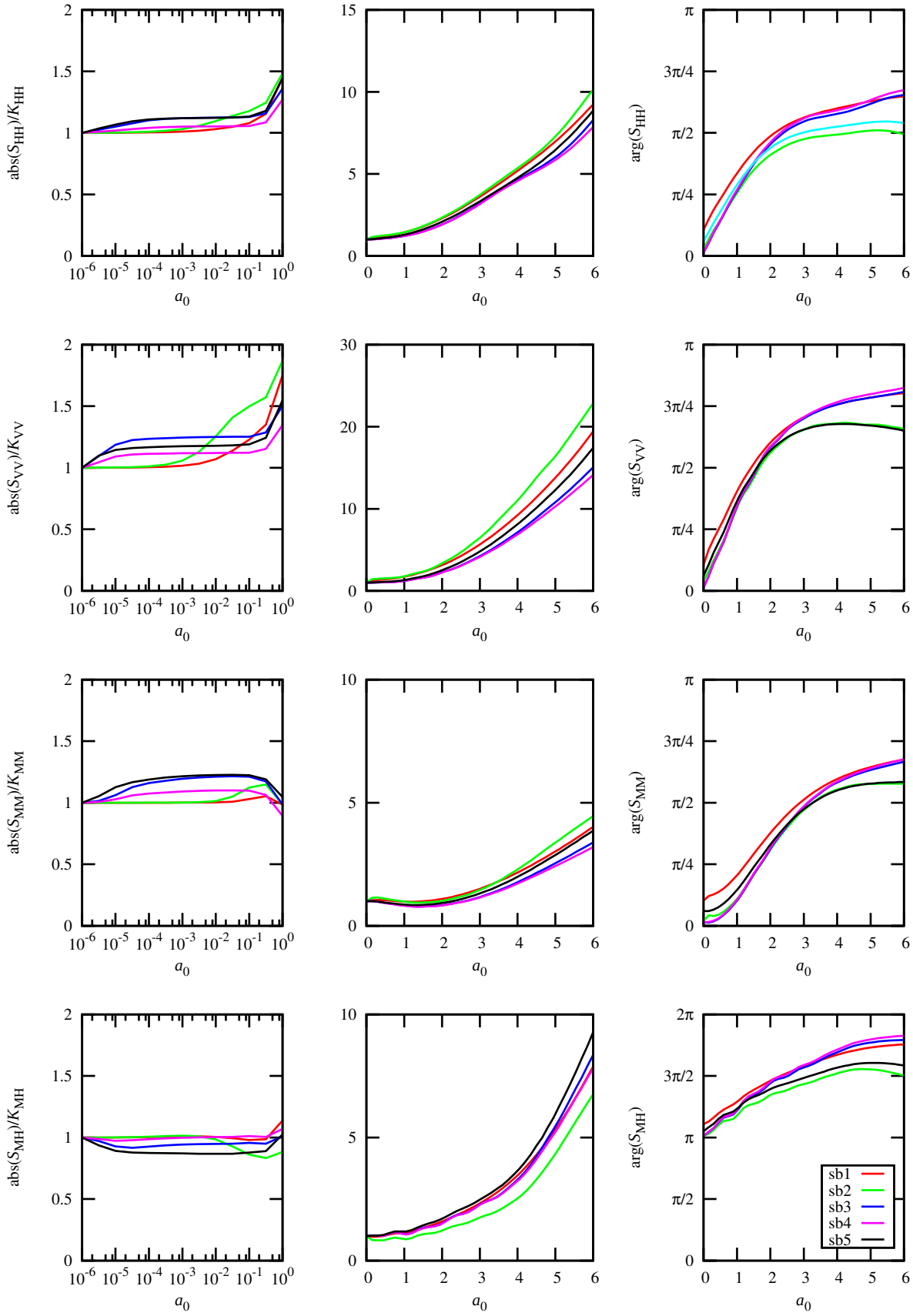



Figure 3: Impedances of bucket foundations with $L/D = 1/4$ in poroelastic soils. From top to bottom: horizontal, vertical, rocking, and horizontal-rocking coupling impedances normalized with respect to the corresponding quasi-static stiffness.

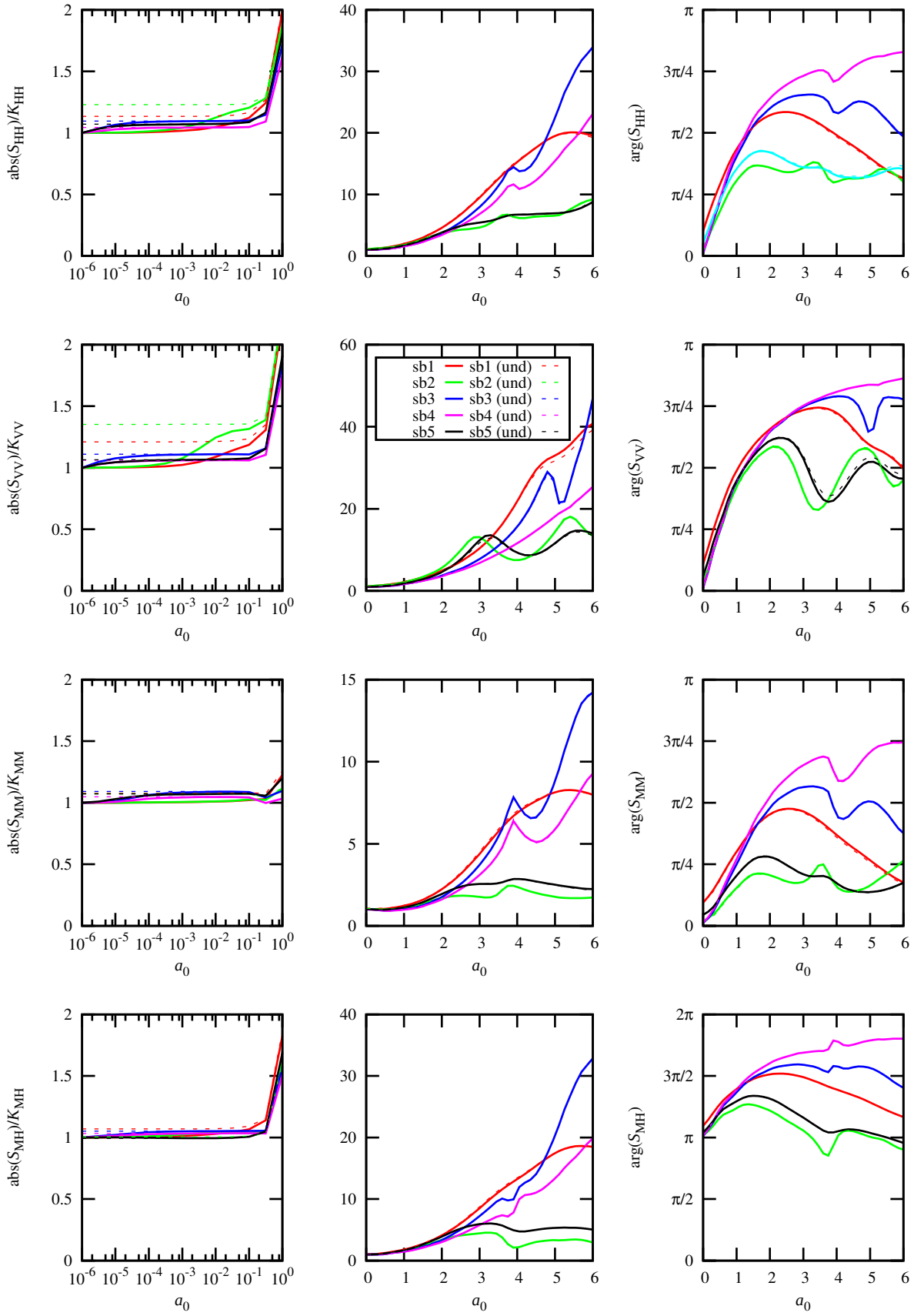



Figure 4: Impedances of bucket foundations with $L/D = 1$ in poroelastic soils and corresponding undrained elastic soils (dashed lines) . From top to bottom: horizontal, vertical, rocking, and horizontal-rocking coupling impedances normalized with respect to the corresponding quasi-static stiffness using the poroelastic soil.

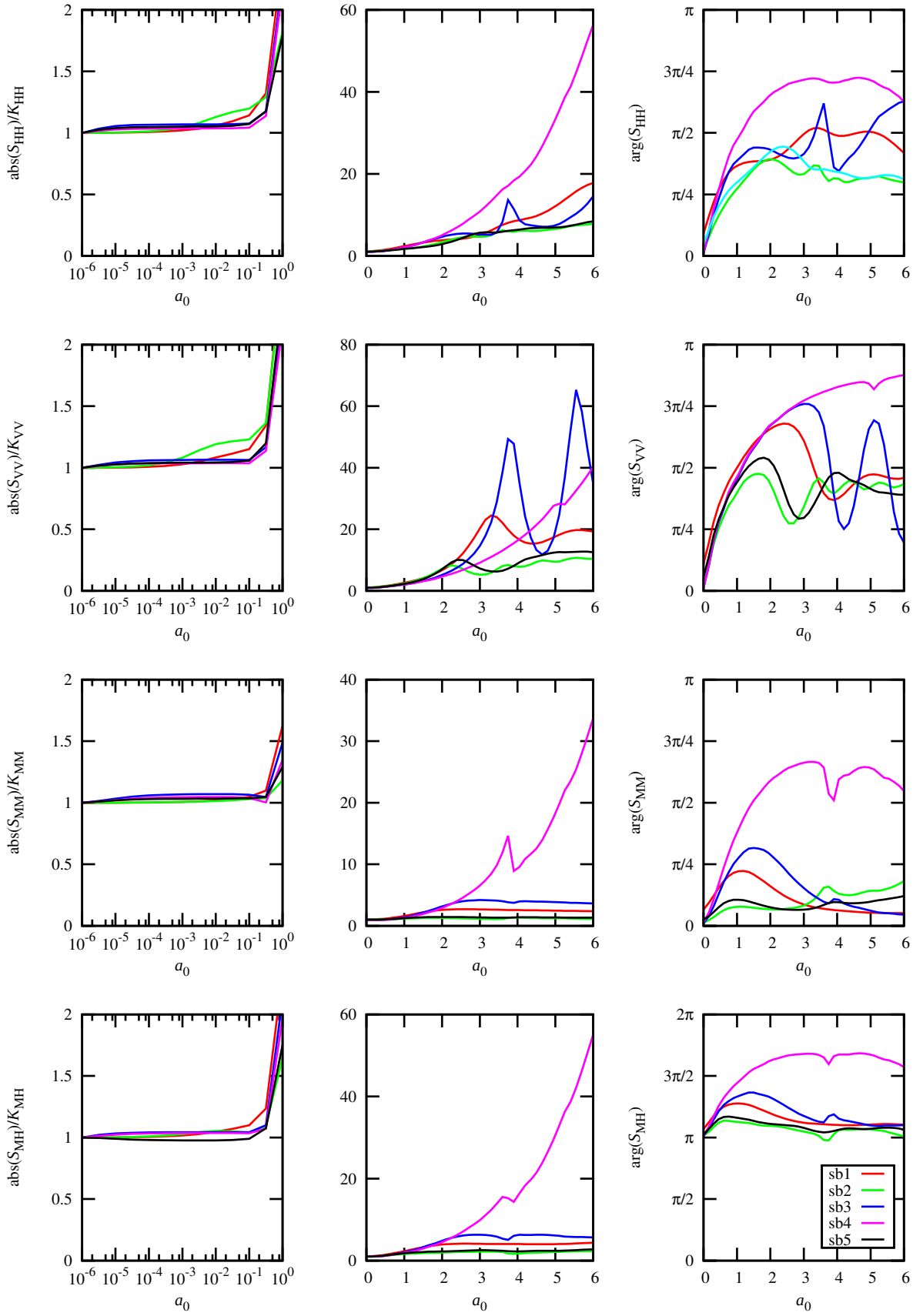



Figure 5: Impedances of bucket foundations with $L/D = 2$ in poroelastic soils. From top to bottom: horizontal, vertical, rocking, and horizontal-rocking coupling impedances normalized with respect to the corresponding quasi-static stiffness.

| **Quasi-static stiffness** (porous: $a_0 = 10^{-6}$) | **Seabed soil** | $\frac{L}{D} = \frac{1}{4}$ | $\frac{L}{D} = 1$ | | $\frac{L}{D} = 2$ |
|---|---|---|---|---|---|
| | | Porous | Porous | Drained | Porous |
| $K_{\mathrm{HH}}$ | sb1 | 7.774 | 13.073 | 13.116 (0.3%) | 16.137 |
| | sb2 | 6.186 | 8.900 | 8.892 (0.1%) | 9.385 |
| | sb3 | 7.445 | 13.198 | 12.175 (7.8%) | 17.754 |
| | sb4 | 8.065 | 14.516 | 14.069 (3.1%) | 21.614 |
| | sb5 | 7.216 | 11.525 | 9.956 (13.6%) | 12.723 |
| $K_{\mathrm{VV}}$ | sb1 | 7.557 | 11.288 | 11.336 (0.4%) | 15.502 |
| | sb2 | 5.822 | 8.954 | 8.946 (0.1%) | 11.952 |
| | sb3 | 7.483 | 11.662 | 10.124 (13.2%) | 16.013 |
| | sb4 | 8.403 | 12.321 | 11.58 (6.0%) | 16.731 |
| | sb5 | 7.904 | 11.849 | 9.283 (21.7%) | 15.407 |
| $K_{\mathrm{MM}}$ | sb1 | 8.739 | 47.368 | 47.48 (0.2%) | 131.429 |
| | sb2 | 7.066 | 28.100 | 28.096 (0.0%) | 44.571 |
| | sb3 | 8.003 | 46.973 | 46.246 (1.5%) | 153.538 |
| | sb4 | 8.993 | 53.106 | 52.728 (0.7%) | 217.156 |
| | sb5 | 7.581 | 35.139 | 34.53 (1.7%) | 68.016 |
| $K_{\mathrm{MH}}$ | sb1 | −2.778 | −15.539 | −15.572 (0.2%) | −30.881 |
| | sb2 | −2.464 | −8.816 | −8.806 (0.1%) | −10.950 |
| | sb3 | −2.729 | −16.036 | −15.307 (4.5%) | −37.250 |
| | sb4 | −2.700 | −17.923 | −17.561 (2.0%) | −51.545 |
| | sb5 | −2.550 | −11.751 | −11.112 (5.4%) | −17.182 |

Table 2: Quasi-static stiffnesses of the studied bucket foundations and seabed soils

The effect is due to the permeability of the porous medium, the smaller permeability the more pronounced variation. It is more relevant for buckets with smaller length to diameter ratios because of the relevance of the compressional interaction of the bucket lid with respect to the total impedance.

In Figure 4, results of the corresponding undrained elastic soils are included, and they are normalized with respect to the quasi-static stiffnesses of the correspoding porous media. Along the low-frequency range (except when $a_0 \to 0$), it is quite clear that neither the drained nor the undrained elastic soil is able to reproduce the real poroelastic behaviour.

The right hand side and central graphs of Figures 3 to 5 show impedance functions for a broader frequency range ($a_0 = [0, 6]$). By comparing these graphs and those obtained by Liingaard et al. [6] for elastic soils, the same qualitative behaviour is observed. For small length to diameter ratios, results tend to the solution of a disc foundation, while for larger ratios results tend to the solution of an infinite hollow cylinder. As shown in Figure 4, the behaviour is not only qualitatively similar, but also numerically if the corresponding undrained elastic soil is used. The difference between the real poroelastic soil and the undrained elastic soil is very small.

## 4 CONCLUSIONS

In this paper, a simple, efficient and accurate three-dimensional BE-FE dynamic model able to directly manage bucket foundations in poroelastic soils is presented. The model makes use of the Dual Boundary Element Method in order to avoid using any artificial boundary in the discretization when thin open structures are buried in soils.

In this work, results of impedance functions for horizontal, vertical, rocking and horizontal-rocking coupling modes of bucket foundations buried in poroelastic soils are presented. A

realistic set of seabed soils are used to obtain the impedances. It is shown that the poroelastic nature of the seabed soil should be considered when studying a problem in the low-frequency range ($< 1-6$ Hz depending on the seabed soil). This is particularly true for bucket foundations with small length to diameter ratios.

## ACKNOWLEDGEMENTS

This work was supported by the Subdirección General de Proyectos de Investigación of the Ministerio de Economía y Competitividad (MINECO) of Spain and FEDER through Research Project BIA2014-57640-R and also by the Agencia Canaria de Investigación, Innovación y Sociedad de la Información (ACIISI) of the Government of the Canary Islands and FEDER through Research Project ProID20100224. J.D.R. Bordón is recipient of the research fellowship FPU13/01224 from the Ministry of Education, Culture and Sports of Spain. The authors are grateful for this support.

## REFERENCES

[1] G.T. Houlsby, B.W. Byrne, Suction caisson foundations for offshore wind turbines and anemometer masts. *Wind Engineering*, **24**, 249–255, 2000.

[2] Det Norske Veritas, *Design of offshore wind turbine structures*. Offshore Standard DNV-OS-J101, 2010.

[3] G.T. Houlsby, L.B. Ibsen, B.W. Byrne, Suction caissons for wind turbines. S. Gourvenec, M. Cassidy eds. *Frontiers in Offshore Geotechnics: ISFOG2005*, Perth, September 19-21 2005.

[4] A. Foglia, L.B. Ibsen, Bucket foundations: a literature review. *Department of Civil Engineering, Aalborg University*, DCE Technical Reports No **176**, 2014.

[5] R.S. Kourkoulis, P.C. Lekkakis, F.M. Gelagoti, A.M. Kaynia, Suction caisson foundations for offshore wind turbines subjected to wave and earthquake loading: effect of soil–foundation interface. *Géotechnique*, **64**, 171–185, 2014.

[6] M. Liingaard, L. Andersen, L.B. Ibsen, Impedance of flexible suction caissons. *Earthquake Engineering and Structural Dynamics*, **36**, 2249–2271, 2007.

[7] M.A. Biot, Theory of propagation of elastic waves in a fluid-saturated porous solid. I. Low-frequency range. *Journal of Acoustical Society of America*, **28**, 168–178, 1956.

[8] J.G. Berryman, Confirmation of Biot's theory. *Applied Physics Letters*, **37**, 382-384, 1980.

[9] S. Bougacha, J.L. Tassoulas, Seismic analysis of gravity dams I: modeling of sediments. *Journal of Engineering Mechanics*, **117**, 1826–1837, 1991.

[10] J. Domínguez, An integral formulation for dynamic poroelasticity. *Journal of Applied Mechanics*, **58**, 588–591, 1991.

[11] J. Domínguez, Boundary element approach for dynamic poroelastic problems. *International Journal for Numerical Methods in Engineering*, **35**, 307–324, 1992.

[12] O. Maeso and J.J. Aznárez and F. García, Dynamic impedances of piles and groups of piles in saturated soils. *Computers and Structures*, **83**, 769–782, 2005.

[13] J.D.R. Bordón, J.J. Aznárez, O. Maeso, Two-dimensional numerical approach for the vibration isolation analysis of thin walled wave barriers in poroelastic soils. *Computers and Geotechnics*, **71**, 168–179, 2016.

[14] J. Domínguez and M.P. Ariza, R. Gallego, Flux and traction boundary elements without hypersingular or strongly singular integrals. *International Journal for Numerical Methods in Engineering*, **48**, 111–135, 2000.

[15] A. Portela, M.H. Aliabadi, D.P. Rooke, The Dual Boundary Element Method: effective implementation for crack problems. *International Journal for Numerical Methods in Engineering*, **33**, 1269–1287, 1992.

[16] Y. Mi, M.H. Aliabadi, Dual Boundary Element Method for three-dimensional fracture mechanics analysis. *Engineering Analysis with Boundary Elements*, **10**, 161–171, 1992.

[17] E. Oñate, *Structural Analysis with the Finite Element Method. Linear Statics. Volume 2: Beams, Plates and Shells*. Lecture notes on Numerical Methods in Engineering and Sciences. CIMNE - Springer, 2013.

[18] J.L. Buchanan, R.P. Gilbert, Transmission Loss in the Far Field over a One-layer Seabed assuming the Biot sediment model. *Journal of Applied Mathematics & Mechanics*, **2**, 121–135, 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# ON THE NUMERICAL INFLUENCES OF INERTIA REPRESENTATION FOR RIGID BODY DYNAMICS

**Xiaoming Xu and Wanxie Zhong**

Department of Engineering Mechanics, Dalian University of Technology
No.2 Linggong Road, Dalian 116024, P.R.China
e-mail: xxm020201@163.com, zwoffice@dlut.edu.cn

**Keywords:** Mass matrix, numerical integration, Computational Methods, Preconditioning, Higher accuracy.

**Abstract.** *Inertia plays a crucial role in rigid body dynamics, and the associated mass matrix is of various forms in representation. The influences of accuracy of different representation, however, have not drawn enough attention in previous researches about numerical simulation of rigid body dynamics. In the paper, the inertia representation is intensively investigated for rigid body dynamics and a modified formulation is derived through splitting the kinetic energy into two parts: a square term of velocity and a quadratic form in the derivatives with quadratic coefficients in generalized displacement, of which the proportion is controlled by a scaling parameter. Although the kinetic energy with different scaling parameters is theoretically equivalent in dynamics, error estimation demonstrates that accuracy of numerical scheme crucially depends on the particular value of scaling parameter if only rotational coordinates are expressed in pseudo vectors. This attractive feature distinguishes the modified formulation from others in numerical significance. According to the modified representation of inertia, a variational integrator is derived for rigid body dynamics in pseudo vectors. Numerical results demonstrate that the variational integrator, of which the scaling parameter is selected as the arithmetic mean of three principal moments of inertia tensor, is of impressively higher accuracy in simulation, especially compared with the integrations derived with the original formulation of mass matrix.*

## 1 INTRODUCTION

Rigid body dynamics is applied frequently in engineering, physics, chemistry, molecular dynamics, etc. As well known, the motion of a single body can be split into two separate parts [1], of which, one is the pure translation of a reference point, and the other is pure rotation about the reference point. Although the description and integration of the translation part has found a fairly conclusive form, the rigid body rotation is still actively investigated because a variety of representation coexists and presents their respective advantages in different aspects of the application. The most common and widely used type of rotational coordinates for describing the rotation motion of a rigid body is Euler angles, that one of the minimal representation for rigid body rotation. It has advantage that the motion of rigid body is governed by ordinary differential equations in the description of Euler angle. In the last few decades, the theory of numerical methods for general (non-stiff and stiff) ordinary differential equations has reached a certain maturity [2], excellent general-purpose codes, mainly based on Runge-Kutta methods or linear multistep methods, have become available for the numerical simulation of rigid body dynamics.

Nevertheless, with steady increases in both size and complexity of the systems investigated, pseudo-vectors, such as unit quaternion and convected base vectors, are found new attraction due to the simplicity in mathematical formulation and the possibility of avoiding singularities which may occur in the framework of Euler angles. Because of the non-independence of the parameters in pseudo-vectors, algebraic constraints are usually included in the equations of motion, which essentially extends the equations by holonomic constraints and yields a set of differential-algebraic equations (DAEs) of Index 3, instead of ordinary differential equations (ODEs). According to a formal definition of the index of DAEs [3], the early researches [4-6] underline that higher indices result in more arduous solving process, especially numerical difficulties associated with the solution of these index-3 DAEs. Due to these reasons, a large amount of effort [7-18] has been devoted to the study of computational methods in pseudo-vectors for handling the motion of rigid body.

One of the more remarkable aspects of pseudo-vector is that the inertia representation can greatly influence the numerical accuracy of integrations in simulation. The inertia representation, associated with mass matrix, expresses the connection between velocity of a system and the kinetic energy of that system. Xu and Zhong [19] first found this numerical phenomenon in their research about quaternion-based rigid body dynamics. They present that the quaternion-based mass matrix can be split into two parts: an identity matrix with a scaling parameter, denoted as $\sigma$, and a displacement-dependent matrix with a partial inertia tensor, denoted as $\boldsymbol{J}_{\sigma}$, and the proportion of two components is controlled by the scaling parameter. They further demonstrate that geometric integrations with a proper value of scaling parameter is of much higher accuracy than the others. Here we aim to present that the new type of inertia representation are available no matter what kind of rotational coordinates are used to describe the motion of rigid body. In the continuous case these inertia representation with different scaling parameter $\sigma$ are theoretically equivalent, however, while the pseudo-vector is considered as the rotational coordinates, accuracy of numerical scheme crucially depends on the particular value of scaling parameter that is being chosen as starting point for the discretization process. This attractive feature distinguishes the modified formulation from others in numerical significance. We derive the variational integrator with the aid of the new type of inertia representation. The arithmetic average of three principal moments of inertia, denoted as $\sigma_{\mathrm{m}}$ is recommended as a reasonable preconditioning value of the scaling parameter. Numerical re-

sults demonstrate that the variational integrator with $\sigma = \sigma_{\mathrm{m}}$ is of impressively high accuracy, especially compared with the integrator with the original formulation of mass matrix.

## 2 KINEMATICS

The motion of a single rigid body is illustrated in Figure 1. A body fixed coordinate frame, denoted by *O-XYZ* is centered at point *O* with translation vector $\boldsymbol{x}_0$. A point located inside the rigid body with coordinates $\boldsymbol{X} = [\, X, Y, Z \,]^T$ has the global components [18]

$$\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{R}\boldsymbol{X} \tag{1}$$

where $\boldsymbol{R}$ is the rotational matrix satisfying the orthonormal condition

$$\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_3 \tag{2}$$

and $\boldsymbol{I}_3$ is the $3\times 3$ identity matrix.

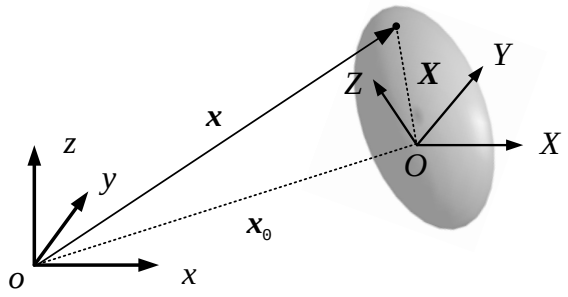



Figure 1: A single body described by translation vector $\boldsymbol{x}_0$ and rotational matrix $\boldsymbol{R}$.

The kinetic energy of a body with volume $V_0$ can be expressed by the integral

$$T = \frac{1}{2}\int_{V_0} \rho \dot{\boldsymbol{x}}^T \dot{\boldsymbol{x}} \mathrm{d}V_0 \tag{3}$$

where $\rho$ denotes the mass density. Substituting (1) into (3) and selecting the reference point $O$ as the center of mass, the kinetic energy takes a particularly simple form where the translational kinetic energy of the center of mass $T_{\mathrm{t}}$ decouples from the rotational kinetic energy of the body $T_{\mathrm{r}}$ as

$$T = T_{\mathrm{t}} + T_{\mathrm{r}} \tag{4}$$

First, the translational part of kinetic energy of the mass center is expressed as

$$T_{\mathrm{t}} = \frac{1}{2} m \dot{\boldsymbol{x}}_0^T \dot{\boldsymbol{x}}_0 \tag{5}$$

where $m$ denotes the mass of the body, and $\dot{\boldsymbol{x}}_0$ is the velocity of the mass center. Similarly, the rotational part of the kinetic energy can be expressed in terms of the velocity $\boldsymbol{v}_{\mathrm{r}}$ due to rotation in the form

$$T_{\mathrm{r}} = \frac{1}{2}\int_{V_0} \rho \boldsymbol{v}_{\mathrm{r}}^T \boldsymbol{v}_{\mathrm{r}} \mathrm{d}V_0 \tag{6}$$

### 2.1 Rigid body motion

In the present formulation of rotational part of kinetic energy, the vector $\boldsymbol{v}_{\mathrm{r}}$ represents the rotation velocity of a mass point of rigid body in the body fixed coordinate frame *O-XYZ*. Considered the rigid body is rotating with angular velocity $\boldsymbol{\Omega}$, the rotation velocity can therefore be expressed in terms of the angular velocity $\boldsymbol{\Omega}$ as,

$$\boldsymbol{v}_{\mathrm{r}} = \boldsymbol{\Omega} \times \boldsymbol{X} = \hat{\boldsymbol{\Omega}}\boldsymbol{X} = -\hat{\boldsymbol{X}}\boldsymbol{\Omega} \tag{7}$$

where the symbol (^) denotes the vector product via the skew-symmetric local component matrix

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{\Omega}\times = \begin{bmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{bmatrix} \tag{8}$$

Substituting (7) into (6) yields

$$T_{\mathrm{r}} = \frac{1}{2}\boldsymbol{\Omega}^T \left( \int_{V_0} \rho \hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}} dV_0 \right) \boldsymbol{\Omega} = \frac{1}{2}\boldsymbol{\Omega}^T \boldsymbol{J} \boldsymbol{\Omega} \tag{9}$$

where $\boldsymbol{J}$ is the inertia tensor, defined by the volume integral

$$\boldsymbol{J} = \int_{V_0} \rho \hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}} dV \tag{10}$$

and conveniently expressed in body fixed coordinate frame, whereby $\boldsymbol{J}$ is constant.

Suppose that the angular velocity can be formally expressed as

$$\boldsymbol{\Omega} = \boldsymbol{L}(\boldsymbol{q})^T \dot{\boldsymbol{q}} \tag{11}$$

where $\boldsymbol{q}$ denotes the *n* generalized coordinates described the three dimensional rotation motion and $\dot{\boldsymbol{q}}$ is its velocity; $\boldsymbol{L}(\boldsymbol{q})^T$ represents a geometric mapping between $\boldsymbol{\Omega}$ and $\dot{\boldsymbol{q}}$. While $n = 3$, $\boldsymbol{q}$ represents the minimal coordinate representation for rigid body rotation. while $n > 3$, $\boldsymbol{q}$ represents pseudo-vector, of which the *n* components are not independent, and generally *n*-3 constraint equations, defined by

$$\boldsymbol{\psi}(\boldsymbol{q}) = \boldsymbol{0} \tag{12}$$

are imposed on the system to preserve the one to one correspondence relation between $\boldsymbol{\Omega}$ and $\dot{\boldsymbol{q}}$.

The kinetic energy of the rigid body motion can now be expressed in terms of the translation velocity, the generalized coordinates $\boldsymbol{q}$ and the time derivative $\dot{\boldsymbol{q}}$, by substitution of the angular velocity from (11) into the expressions (9), whereby

$$T = \frac{1}{2} m \dot{\boldsymbol{x}}_0^T \dot{\boldsymbol{x}}_0 + \frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{M} \dot{\boldsymbol{q}} \tag{13}$$

where

$$\boldsymbol{M} = \boldsymbol{L}(\boldsymbol{q}) \boldsymbol{J} \boldsymbol{L}(\boldsymbol{q})^T \tag{14}$$

is defined as the mass matrix with respect to the rotation motion.

## 2.2 Rotational coordinates for rigid body rotation

As presented in the above, the rotational part of kinetic energy can be expressed formally by the generalized coordinates $\boldsymbol{q}$ and the time derivative $\dot{\boldsymbol{q}}$, and however, not every generalized coordinates preserving the one to one mapping relationship between $\boldsymbol{\Omega}$ and $\dot{\boldsymbol{q}}$ can be used to describe the rotation motion of a rigid body. Actually, it can be derived from the infinitesimal rotation that the rotation motion is governed by the differential equations [1]

$$\dot{\boldsymbol{R}} = \boldsymbol{R}\hat{\boldsymbol{\Omega}} \tag{15}$$

Considered that $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_3$, we have

$$\hat{\boldsymbol{\Omega}} = \boldsymbol{R}^T\dot{\boldsymbol{R}} \tag{16}$$

which implies a one to one mapping relationship between $\boldsymbol{\Omega}$ and $\dot{\boldsymbol{R}}$, and consequently, it requires one-to-one mapping between $\boldsymbol{q}$ can $\boldsymbol{R}$ to describe the rotation motion of a rigid body with $\boldsymbol{q}$.

*1) Euler angles*

The most common and widely used type of rotational coordinates is Euler angles. In the description of Euler angles[1], the rotational motion is expressed in terms of the angle of nutation $\theta$, the angle of precession $\phi$, and the spin angle $\psi$, whereby

$$\boldsymbol{R} = \begin{bmatrix} \cos\psi\cos\phi - \cos\theta\sin\phi\cos\psi & \cos\psi\sin\phi + \cos\theta\cos\phi\sin\psi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\theta\sin\phi\cos\psi & -\sin\psi\sin\phi + \cos\theta\cos\phi\cos\psi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{bmatrix} \tag{17}$$

Substituting (17) into (16) leads to

$$\begin{aligned} \Omega_1 &= \dot{\phi}\sin\theta\sin\psi + \dot{\theta}\cos\psi \\ \Omega_2 &= \dot{\phi}\sin\theta\cos\psi - \dot{\theta}\sin\psi \\ \Omega_3 &= \dot{\phi}\cos\theta + \dot{\psi} \end{aligned} \tag{18}$$

In combination of (18) with (11), we have the generalized coordinate $\boldsymbol{q} = [\psi, \theta, \phi]^T$ and

$$\boldsymbol{L}(\boldsymbol{q}) = \begin{bmatrix} \sin\theta\sin\psi & \cos\psi & 0 \\ \sin\theta\cos\psi & -\sin\psi & 0 \\ \cos\theta & 0 & 1 \end{bmatrix}^T \tag{19}$$

*2) Unit quaternion*

Unit quaternion is another popular way to describe the rotation of a rigid body. In the (real) matrix representation [16], unit quaternion can be thought of as a 4-dimensional vector

$$\boldsymbol{q} = [q_0, q_1, q_2, q_3]^T \tag{20}$$

with the unit length constraint

$$\boldsymbol{\psi}(\boldsymbol{q}) = \boldsymbol{q}^T\boldsymbol{q} - 1 = 0 \tag{21}$$

and then the rotation matrix can be expressed in terms of unit quaternion

$$\boldsymbol{R}(\boldsymbol{q})=\begin{bmatrix} q_0^2+q_1^2-q_2^2-q_3^2 & 2(q_1q_2-q_0q_3) & 2(q_1q_3+q_0q_2) \\ 2(q_1q_2+q_0q_3) & q_0^2-q_1^2+q_2^2-q_3^2 & 2(q_2q_3-q_0q_1) \\ 2(q_1q_3-q_0q_2) & 2(q_2q_3+q_0q_1) & q_0^2-q_1^2-q_2^2+q_3^2 \end{bmatrix} \tag{22}$$

Substituting (22) into (16) leads to

$$\begin{aligned} \Omega_1 &= 2(-q_1\dot{q}_0+q_0\dot{q}_1+q_3\dot{q}_2-q_2\dot{q}_3) \\ \Omega_2 &= 2(-q_2\dot{q}_0-q_3\dot{q}_1+q_0\dot{q}_2+q_1\dot{q}_3) \\ \Omega_3 &= 2(-q_3\dot{q}_0+q_2\dot{q}_1-q_1\dot{q}_2+q_1\dot{q}_4) \end{aligned} \tag{23}$$

and correspondingly

$$\boldsymbol{L}(\boldsymbol{q})=2\begin{bmatrix} -q_1 & q_0 & q_3 & -q_2 \\ -q_2 & -q_3 & q_0 & q_1 \\ -q_3 & q_2 & -q_1 & q_0 \end{bmatrix}^T \tag{24}$$

*3) Convected base vectors*

As an alternative way to describe the rotation of a rigid body, the rotation matrix can be expressed directly by

$$\boldsymbol{R}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\boldsymbol{q}_3] \tag{25}$$

where $\boldsymbol{q}_1,\boldsymbol{q}_2,\boldsymbol{q}_3$ are named as the convected base vectors [15, 18]. Define the 9-dimensional vector

$$\boldsymbol{q}=\begin{bmatrix} \boldsymbol{q}_1^T & \boldsymbol{q}_2^T & \boldsymbol{q}_3^T \end{bmatrix}^T \tag{26}$$

and then the orthonormal constraints presented by (2), can be rewritten as

$$\boldsymbol{\psi}(\boldsymbol{q})=\begin{bmatrix} \frac{1}{2}(\boldsymbol{q}_1^T\boldsymbol{q}_1-1) \\ \frac{1}{2}(\boldsymbol{q}_2^T\boldsymbol{q}_2-1) \\ \frac{1}{2}(\boldsymbol{q}_3^T\boldsymbol{q}_3-1) \\ \frac{1}{2\sqrt{2}}(\boldsymbol{q}_2^T\boldsymbol{q}_3+\boldsymbol{q}_3^T\boldsymbol{q}_2) \\ \frac{1}{2\sqrt{2}}(\boldsymbol{q}_3^T\boldsymbol{q}_1+\boldsymbol{q}_1^T\boldsymbol{q}_3) \\ \frac{1}{2\sqrt{2}}(\boldsymbol{q}_1^T\boldsymbol{q}_2+\boldsymbol{q}_2^T\boldsymbol{q}_1) \end{bmatrix}=\boldsymbol{0} \tag{27}$$

Substituting (26) into (16) yields the expression of angular velocity

$$\begin{aligned} \Omega_1 &= (-\boldsymbol{q}_2^T\dot{\boldsymbol{q}}_3+\boldsymbol{q}_3^T\dot{\boldsymbol{q}}_2)/2 \\ \Omega_2 &= (-\boldsymbol{q}_3^T\dot{\boldsymbol{q}}_1+\boldsymbol{q}_1^T\dot{\boldsymbol{q}}_3)/2 \\ \Omega_3 &= (-\boldsymbol{q}_1^T\dot{\boldsymbol{q}}_2+\boldsymbol{q}_2^T\dot{\boldsymbol{q}}_1)/2 \end{aligned} \tag{28}$$

Correspondingly, the projection matrix is expressed as

$$\boldsymbol{L}(\boldsymbol{q})=\frac{1}{2}\begin{bmatrix} \boldsymbol{0} & -\boldsymbol{q}_3 & \boldsymbol{q}_2 \\ \boldsymbol{q}_3 & \boldsymbol{0} & \boldsymbol{q}_1 \\ -\boldsymbol{q}_2 & \boldsymbol{q}_1 & \boldsymbol{0} \end{bmatrix} \tag{29}$$

## 3 DYNAMICS AND INERTIA REPRESENTATIONS

In rigid body dynamics, the motion of a single body is governed by the Lagrange's equations of first kind. Suppose that the translation vector $\boldsymbol{x}_0 = \boldsymbol{0}$, and the equations of rotation motion can be expressed in a unified form

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{\boldsymbol{q}}}\right) - \frac{\partial T}{\partial \boldsymbol{q}} = \boldsymbol{Q} + \boldsymbol{Q}^c \tag{30}$$

In the above expression, $\boldsymbol{Q}$ denotes the generalized torque vector in terms of the generalized coordinates $\boldsymbol{q}$. Suppose that the rigid body is rotating in gravitational fields, the generalized torque vector can be expressed as

$$\boldsymbol{Q} = -\partial U / \partial \boldsymbol{q} \tag{31}$$

where $U$ is the potential energy. The vector $\boldsymbol{Q}^c$ denotes the constrained forces. While minimal representation is implemented (e.g. Euler angles), $\boldsymbol{Q}^c$ is identically equal to zero, whereas the constrained forces are defined as

$$\boldsymbol{Q}^c = \boldsymbol{A}(\boldsymbol{q})^T \boldsymbol{\lambda} \tag{32}$$

where $\boldsymbol{A}$ is the constraint jacobian matrix, defined by

$$\boldsymbol{A}(\boldsymbol{q}) = \partial \boldsymbol{\psi}(\boldsymbol{q}) / \partial \boldsymbol{q}^T \tag{33}$$

and $\boldsymbol{\lambda}$ is the Lagrange's multiplier which preserves the path of pseudo-vector satisfying the constraints (12). Define the generalized momentum

$$\boldsymbol{p} = \partial T / \partial \dot{\boldsymbol{q}} = \boldsymbol{M}\dot{\boldsymbol{q}} \tag{34}$$

and the generalized force

$$\boldsymbol{f} = \partial T / \partial \boldsymbol{q} + \boldsymbol{Q} \tag{35}$$

and then Expression (30) can be expressed in a simplified form

$$\dot{\boldsymbol{p}} = \boldsymbol{f} + \boldsymbol{Q}^c \tag{36}$$

### 3.1 The modified inertia representation for rotation motion

The inertia representation, associated with mass matrix, establishes the connection between velocity of a system and the kinetic energy of that system. The mass matrix of a system is of various forms in representation: Firstly, it can be observed from the above that inertia representations of rotation motion are different if only different rotational coordinates, such as Euler angles, unit quaternion and convected base vectors, are implemented in practice; Secondly, the inertia representation is non-unique, even if the system is described under the same generalized coordinate. The former is a clear and logical conclusion, and the latter would be intensively investigated in the following, though it is not terribly intuitive in understanding.

Without loss of generality, suppose that the body-fixed coordinate axes *O-XYZ* are aligned along the principal axes of inertia of the rigid body, and then the inertia tensor can be expressed as $\boldsymbol{J} = \mathrm{diag}(J_1, J_2, J_3)$. Pre-multiply Expression (11) by itself, and then we have

$$\boldsymbol{\Omega}^T \boldsymbol{\Omega} = \dot{\boldsymbol{q}}^T \boldsymbol{L}(\boldsymbol{q}) \boldsymbol{L}(\boldsymbol{q})^T \dot{\boldsymbol{q}} \tag{37}$$

Assuming that

$$\dot{\boldsymbol{q}}^T \boldsymbol{L}(\boldsymbol{q})\boldsymbol{L}(\boldsymbol{q})^T \dot{\boldsymbol{q}} = \dot{\boldsymbol{q}}^T \dot{\boldsymbol{q}} \tag{38}$$

and then it can be derived that

$$\boldsymbol{\Omega}^T \boldsymbol{\Omega} = \dot{\boldsymbol{q}}^T \dot{\boldsymbol{q}} \tag{39}$$

Reformulate the rotational part of kinetic energy in an equivalent form

$$T = \frac{1}{2}\sigma \boldsymbol{\Omega}^T \boldsymbol{\Omega} + \frac{1}{2}\boldsymbol{\Omega}^T \boldsymbol{J}_\sigma \boldsymbol{\Omega} \tag{40}$$

where

$$\boldsymbol{J}_\sigma = \boldsymbol{J} - \sigma \boldsymbol{I}_3 \tag{41}$$

and $\sigma > 0$, and then substituting (11) and (39) into (40) yields a new type of formulation of kinetic energy

$$T = \frac{1}{2}\sigma \dot{\boldsymbol{q}}^T \dot{\boldsymbol{q}} + \frac{1}{2}\dot{\boldsymbol{q}}^T \boldsymbol{L}(\boldsymbol{q})\boldsymbol{J}_\sigma \boldsymbol{L}(\boldsymbol{q})^T \dot{\boldsymbol{q}} \tag{42}$$

In the modified inertia representation, the kinetic energy is generally split into a square term of velocity and a quadratic form in the derivatives with quadratic coefficients in generalized coordinates, of which the proportion is controlled by a scaling parameter, denoted as $\sigma$. Based on the Legendre transformation, we can derive the generalized momentum,

$$\boldsymbol{p} = \boldsymbol{M}_\sigma \dot{\boldsymbol{q}} \tag{43}$$

where

$$\boldsymbol{M}_\sigma = \sigma \boldsymbol{I} + \boldsymbol{L}(\boldsymbol{q})\boldsymbol{J}_\sigma \boldsymbol{L}(\boldsymbol{q})^T \tag{44}$$

and $\boldsymbol{I}$ is the identity matrix. The mass matrix $\boldsymbol{M}_\sigma$, presented by (44), is distinguished with the mass matrix $\boldsymbol{M}$ since an identity matrix with respect to the scaling parameter $\sigma$ is separated from the original mass matrix, and while $\sigma = 0$, the two matrices are of the same formulations. Consequently, substituting (43) into (36), we can obtain a more generalized formulation of Lagrange's equations for rotation motion.

### 3.2 The inertia representations under different rotational coordinates

In the proceeding section, a newly inertia representation are derived by addition and subtraction of the homogeneous isotropic form $\frac{1}{2}\sigma \boldsymbol{\Omega}^T \boldsymbol{\Omega}$. It can be observed that the assumption presented by (38), plays an important role in the derivational process. However, this restriction is not a necessary condition to make sure the existence of the inertia representation, but it provides a clear way to reach the ultimate formulations, and the mass matrix presented by (44) can be considered as the standard formulation of mass matrix in the modified representation. Here we aim to develop the application of the modified inertia representation to three different rotational coordinates presented above.

*1) Euler angles*

While Euler angles is considered as the rotational coordinates, Pre-multiplying Expression (19) by itself and considering trigonometric identities yield

$$\boldsymbol{L}\boldsymbol{L}^T = \begin{bmatrix} 1 & 0 & \cos\theta \\ 0 & 1 & 0 \\ \cos\theta & 0 & 1 \end{bmatrix} \tag{45}$$

Substituting (45) into (37) yields

$$\boldsymbol{\Omega}^T\boldsymbol{\Omega} = \dot{\boldsymbol{q}}^T\dot{\boldsymbol{q}} + 2\dot{\psi}\dot{\phi}\cos\theta \tag{46}$$

Substitute (11) and (46) into (40) , and finally the mass matrix can be formally expressed as

$$\boldsymbol{M}_\sigma = \sigma\begin{bmatrix} 1 & 0 & \cos\theta \\ 0 & 1 & 0 \\ \cos\theta & 0 & 1 \end{bmatrix} + \boldsymbol{L}(\boldsymbol{q})\boldsymbol{J}_\sigma\boldsymbol{L}(\boldsymbol{q})^T \tag{47}$$

*2) Unit quaternion*

While unit quaternion is considered as the rotational coordinates, the unit length constraint leads to

$$\dot{\boldsymbol{q}}^T\boldsymbol{q} = 0 \tag{48}$$

Pre-multiplying Expression (24) by itself and considering $\boldsymbol{q}^T\boldsymbol{q} = 1$ gives

$$\boldsymbol{L}(\boldsymbol{q})\boldsymbol{L}(\boldsymbol{q})^T = 4\boldsymbol{I}_4 - 4\boldsymbol{q}\boldsymbol{q}^T \tag{49}$$

where $\boldsymbol{I}_4$ is 4-dimentional identity matrix. Substituting (48) and (49) into (37) yields

$$\boldsymbol{\Omega}^T\boldsymbol{\Omega} = 4\dot{\boldsymbol{q}}^T\dot{\boldsymbol{q}} \tag{50}$$

Substitute (11) and (50) into (40) , and finally the mass matrix can be expressed as

$$\boldsymbol{M}_\sigma = 4\sigma\boldsymbol{I}_4 + \boldsymbol{L}(\boldsymbol{q})\boldsymbol{J}_\sigma\boldsymbol{L}(\boldsymbol{q})^T \tag{51}$$

*3) Convected base vectors*

While convected base vectors are considered as the rotational coordinates, the orthonormal constraints lead to the following constraint jacobian matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \frac{1}{\sqrt{2}}\boldsymbol{q}_3 & \frac{1}{\sqrt{2}}\boldsymbol{q}_2 \\ \boldsymbol{0} & \boldsymbol{q}_2 & \boldsymbol{0} & \frac{1}{\sqrt{2}}\boldsymbol{q}_3 & \boldsymbol{0} & \frac{1}{\sqrt{2}}\boldsymbol{q}_1 \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{q}_3 & \frac{1}{\sqrt{2}}\boldsymbol{q}_2 & \frac{1}{\sqrt{2}}\boldsymbol{q}_1 & \boldsymbol{0} \end{bmatrix}^T \tag{52}$$

The constraint jacobian matrix satisfies the following orthogonal relation

$$\boldsymbol{A}\dot{\boldsymbol{q}} = \boldsymbol{0} \tag{53}$$

A direct calculation reveals that

$$[\sqrt{2}\boldsymbol{L},\ \boldsymbol{A}^T]\begin{bmatrix} \sqrt{2}\boldsymbol{L}^T \\ \boldsymbol{A} \end{bmatrix} = \begin{bmatrix} \sqrt{2}\boldsymbol{L}^T \\ \boldsymbol{A} \end{bmatrix}[\sqrt{2}\boldsymbol{L},\ \boldsymbol{A}^T] = \boldsymbol{I}_9 \tag{54}$$

where $\boldsymbol{I}_9$ is 9-dimentional identity matrix. Substituting (53) and (54) into (37) gives

$$\boldsymbol{\Omega}^T\boldsymbol{\Omega} = \dot{\boldsymbol{q}}^T\dot{\boldsymbol{q}}/2 \tag{55}$$

and hence, the mass matrix can be expressed as

$$\boldsymbol{M}_\sigma = \sigma\boldsymbol{I}/2 + \boldsymbol{L}(\boldsymbol{q})\boldsymbol{J}_\sigma\boldsymbol{L}(\boldsymbol{q})^T \tag{56}$$

### 3.3 A glance about the error estimation

In the above, three different formulations of mass matrix has been derived for rotational motion under the modified representation, and a scaling parameter $\sigma$ is formally, extracted from the original formulations. In the continuous case these formulations with different scaling parameter $\sigma$ are theoretically equivalent, that specifically the scaling parameter can be eliminated if only one admits that the geometric relations, which should be obeyed by rotational coordinates for describing rotation motion, are precisely satisfied. However, this is not always reached by every kind of rotational coordinate in numerical simulation.

Suppose that $\boldsymbol{q} = \boldsymbol{q}(t)$ denotes the real solution of rotation motion of a single body, and $\boldsymbol{q}^+ = \boldsymbol{q}^+(t)$ is its numerical approximated solution. Correspondingly, we can define

$$\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{q}), \quad \boldsymbol{L}^+ = \boldsymbol{L}(\boldsymbol{q}^+) \tag{57}$$

and

$$\boldsymbol{\Omega} = \boldsymbol{L}^T\dot{\boldsymbol{q}}, \ \boldsymbol{\Omega}^+ = (\boldsymbol{L}^+)^T\dot{\boldsymbol{q}}^+, \ \delta\boldsymbol{\Omega} = \boldsymbol{\Omega}^+ - \boldsymbol{\Omega} \tag{58}$$

Then the approximation error of kinetic energy can be expressed as

$$\delta T = T(\boldsymbol{q}^+, \dot{\boldsymbol{q}}^+) - T(\boldsymbol{q}, \dot{\boldsymbol{q}}) \tag{59}$$

Without loss of generality, we assume that the relation presented by (38), is satisfied, and then the kinetic energy can be expressed in the standard form under the modified representation of inertia. In combination of (57), (58) and (59) with (42), it can be derived that

$$\delta T = K_s\sigma + \delta T_0 \tag{60}$$

where

$$\delta T_0 = \delta\boldsymbol{\Omega}^T\boldsymbol{J}\boldsymbol{\Omega} + \delta\boldsymbol{\Omega}^T\boldsymbol{J}\delta\boldsymbol{\Omega}/2 \tag{61}$$

is just the numerical error of the kinetic energy with the original formulation of kinetic energy, and the slope $K_s$ is

$$K_s = \tfrac{1}{2}(\dot{\boldsymbol{q}}^{+T}\dot{\boldsymbol{q}}^+ - \boldsymbol{\Omega}^{+T}\boldsymbol{\Omega}^+) = \tfrac{1}{2}\dot{\boldsymbol{q}}^{+T}[\boldsymbol{I} - \boldsymbol{L}^+(\boldsymbol{L}^+)^T]\dot{\boldsymbol{q}}^+ \tag{62}$$

It can be observed that the scaling parameter $\sigma$ has no influence on the numerical error of kinetic energy if and only if $K_s = 0$, whereas the discretization error of kinetic energy is a linear function with respect to the scaling parameter.

While rotational coordinates are expressed in minimal representation, $K_s = 0$ is equivalent to $\boldsymbol{L}^+(\boldsymbol{L}^+)^T = \boldsymbol{I}$. Consider that all the geometric relations are satisfied automatically in minimal representation, and this would not be broken by any perturbation about the rotational coordinates. Consequently, $\boldsymbol{L}^+(\boldsymbol{L}^+)^T = \boldsymbol{I}$ if only $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{I}$, and this assumption is admitted at the beginning of the discussion. As a result, we can conclude that the scaling parameter $\sigma$ has no influence on the numerical error in minimal representation. For instance, we consider the Eu-

ler angles, that one of the minimal representation, as the rotational coordinates. In this case, the representation is not in standard form. However, a direct calculation reveals that

$$\boldsymbol{L}^{+}(\boldsymbol{L}^{+})^{T}=\begin{bmatrix}1 & 0 & \cos\theta^{+}\\ 0 & 1 & 0\\ \cos\theta^{+} & 0 & 1\end{bmatrix} \tag{63}$$

and this leads to

$$\boldsymbol{M}_{\sigma}^{+}=\sigma\boldsymbol{L}^{+}(\boldsymbol{L}^{+})^{T}+\boldsymbol{L}(\boldsymbol{q}^{+})\boldsymbol{J}_{\sigma}\boldsymbol{L}(\boldsymbol{q}^{+})^{T}=\boldsymbol{L}(\boldsymbol{q}^{+})\boldsymbol{J}\boldsymbol{L}(\boldsymbol{q}^{+})=\boldsymbol{M} \tag{64}$$

Hence the scaling parameter can be directly eliminated from the expression of mass matrix in minimal representation.

While rotational coordinates are expressed in non-minimal representation (i.e. the pseudo-vector), the geometric relations, which should be obeyed by rotational coordinates, are directly embodied as the constraints $\boldsymbol{\psi}(\boldsymbol{q})=\boldsymbol{0}$ and the implicit orthogonal relations $\boldsymbol{A}\dot{\boldsymbol{q}}=\boldsymbol{0}$. In continuous case, the scaling parameter $\sigma$ can be eliminated from the formulation of kinetic energy through the geometric relation $\boldsymbol{\psi}(\boldsymbol{q})=\boldsymbol{0}$ and $\boldsymbol{A}\dot{\boldsymbol{q}}=\boldsymbol{0}$. Nevertheless, the geometric relations $\boldsymbol{\psi}(\boldsymbol{q})=\boldsymbol{0}$ can only be satisfied strictly at every time-grid points in the discretization process, following which the scaling parameter $\sigma$ cannot be eliminated from the formulation of kinetic energy (i.e. $K_s\neq 0$), and consequently, the numerical accuracy of integration can be improved by determining a proper value of the scaling parameter $\sigma$. This distinguishes the modified representation of inertia with others in numerical significance. This numerical phenomenon is first discovered by Xu and Zhong [19] in their research about unit quaternion. They further recommend that

$$\sigma_{\mathrm{m}}=(J_1+J_2+J_3)/3 \tag{65}$$

is a reasonable preconditioning value for the scaling parameter. In the following, we will first develop the variational integrator for rigid body dynamics in pseudo-vector by considering the modified inertia representation, and further investigate the numerical performance of the scaling parameter $\sigma$ through two numerical examples.

## 4 A DIRECT DISCRETIZING APPROACH FOR VARIATIONAL INTEGRATOR

Considered the generalized momentum defined by

$$\boldsymbol{p}=\boldsymbol{M}_{\sigma}(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{66}$$

where $\boldsymbol{M}_{\sigma}$ is expressed specifically by (51) for unit quaternion and (56) for convected base vectors. Define an equidistant time grid, consisting of a number $N+1$ of discrete points $t_k$, defined by

$$t_k=k\Delta t \qquad k\in[0,1,\ldots,N] \tag{67}$$

Then differential part of the Lagrange's equations, presented by (36) can be formally written as

$$\dot{\boldsymbol{p}}_k=\boldsymbol{f}_k+\boldsymbol{Q}_k^{c} \tag{68}$$

for every discrete time $t_k=[0,N\Delta t]$. The approximation of the above equations leads to different difference schemes in Lagrange's frame. Here we define the difference approximations

$$\dot{\boldsymbol{q}}_{k+1/2} = (\boldsymbol{q}_{k+1} - \boldsymbol{q}_k)/\Delta t, \quad \boldsymbol{q}_{k+1/2} = (\boldsymbol{q}_k + \boldsymbol{q}_{k+1})/2 \tag{69}$$

and correspondingly, define

$$\boldsymbol{p}_{k+1/2} = \boldsymbol{M}_\sigma(\boldsymbol{q}_{k+1/2})\dot{\boldsymbol{q}}_{k+1/2}, \ \boldsymbol{f}_{k+1/2} = \boldsymbol{f}(\boldsymbol{q}_{k+1/2}, \dot{\boldsymbol{q}}_{k+1/2}) \tag{70}$$

Then the vectors $\dot{\boldsymbol{p}}_k$ and $\boldsymbol{f}_k$ are approximated as

$$\dot{\boldsymbol{p}}_k \approx (\boldsymbol{p}_{k+1/2} - \boldsymbol{p}_{k-1/2})/\Delta t \ , \quad \boldsymbol{f}_k \approx (\boldsymbol{f}_{k-1/2} + \boldsymbol{f}_{k+1/2})/2 \tag{71}$$

Substituting (32) and (71) into (68) and considering the constraint conditions (12) yield the following schemes

$$(\boldsymbol{p}_{k+1/2} - \boldsymbol{p}_{k-1/2})/\Delta t \ = (\boldsymbol{f}_{k-1/2} + \boldsymbol{f}_{k+1/2})/2 + s\boldsymbol{A}(\boldsymbol{q}_k)^T \tilde{\boldsymbol{\lambda}}_k, \quad \boldsymbol{\psi}(\boldsymbol{q}_{k+1}) = \boldsymbol{0} \tag{72}$$

where $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}/s$ are scaled Lagrange multipliers and the scaling factor $s$ can be set to any constant and is recommended as $s = \Delta t^{-2}$ to avoid ill-conditioning problem of iteration matrices [20]. The three-term recursion (72) is proposed by Leyendecker et al [21] through discrete variational principle (also see Wendlandt and Marsden [22]), and in the first step, it can be simplified to the term recursion

$$(\boldsymbol{p}_{1/2} - \boldsymbol{p}_0)/(\tfrac{1}{2}\Delta t) = \boldsymbol{f}_{1/2} + \tfrac{1}{2} s\boldsymbol{A}(\boldsymbol{q}_0)^T \tilde{\boldsymbol{\lambda}}_0, \ \boldsymbol{\psi}(\boldsymbol{q}_1) = \boldsymbol{0} \tag{73}$$

The specific algorithm is summarized in pseudo-code format in Table 1.

| | |
|---|---|
| 1) | Initial condition: $\boldsymbol{q}_0$, $\boldsymbol{p}_0$, $\tilde{\boldsymbol{\lambda}}_{-1} = \boldsymbol{0}$. |
| 2) | Prediction step: $\boldsymbol{g}_{k+1} = \boldsymbol{g}_k$ where $\boldsymbol{g}_k = [\boldsymbol{q}_k^T \quad \tilde{\boldsymbol{\lambda}}_{k-1}]^T$ |
| 3) | Residual calculation: $\boldsymbol{r} = [\boldsymbol{r}_q \quad \boldsymbol{r}_\lambda]^T$<br>$\begin{cases} \boldsymbol{r}_q = (\boldsymbol{p}_{1/2} - \boldsymbol{p}_0)/(\frac{1}{2}\Delta t) - \boldsymbol{f}_{1/2} - \frac{1}{2} s\boldsymbol{A}(\boldsymbol{q}_0)^T \tilde{\boldsymbol{\lambda}}_0, k = 0 \\ \boldsymbol{r}_q = (\boldsymbol{p}_{k+1/2} - \boldsymbol{p}_{k-1/2})/\Delta t - (\boldsymbol{f}_{k-1/2} + \boldsymbol{f}_{k+1/2})/2 - s\boldsymbol{A}(\boldsymbol{q}_k)^T \tilde{\boldsymbol{\lambda}}_k, \ k \geq 1 \end{cases}$<br>$\boldsymbol{r}_\lambda = \boldsymbol{\psi}(\boldsymbol{q}_{k+1})$ |
| 4) | Update incremental rotation parameters:<br>$\delta \boldsymbol{g} = -\boldsymbol{K}^{-1}\boldsymbol{r}, \ \boldsymbol{g}_{k+1} = \boldsymbol{g}_k + \delta \boldsymbol{g}$ ,<br>where $\boldsymbol{K} = \partial \boldsymbol{r}/\partial \boldsymbol{g}_{k+1}^{\mathrm{T}}$ and If $\|\boldsymbol{r}\| > \varepsilon_r$ , repeat from 3). |
| 5) | Return to 2) for new step, or stop. |

Table 1: Variational integrator algorithm

## 5 NUMERICAL EXAMPLES

The numerical examples deal with a top spinning in gravitational field. As is shown in Figure 2, a top with the mass $m$ is considered. The distance between mass center $A$ and fixed point $O$ is denoted by $l$, and the gravitational acceleration $g$ is equal to 9.81 in the negative z-direction. While unit quaternion is considered, the gravitational potential energy of the top can be written as

$$U(\boldsymbol{q}) = mgl\cos\theta = mgl\boldsymbol{q}^T\boldsymbol{K}\boldsymbol{q} \tag{74}$$

where $\boldsymbol{K} = \mathrm{diag}(1,-1,-1,1)$ and correspondingly, we have

$$\partial U/\partial \boldsymbol{q} = 2mgl\boldsymbol{K}\boldsymbol{q} \tag{75}$$

While convected base vectors are considered, the gravitational potential energy of the top can be written

$$U(\boldsymbol{q}) = mgl\cos(\theta) = ml\boldsymbol{g}^T\boldsymbol{q}_3 \tag{76}$$

where the gravitational vector is given as $\boldsymbol{g}^T = [0,0,g]$. Hence external force can be expressed as

$$\partial U/\partial \boldsymbol{q} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad ml\boldsymbol{g}^T]^T \tag{77}$$

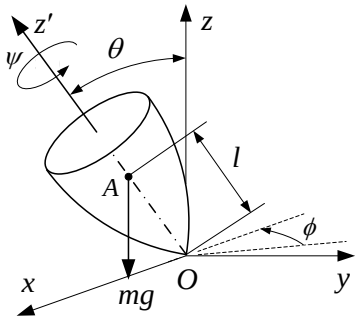



Figure 2: A heavy top.

### 5.1 Fast spinning top

Firstly, the fast spinning of a symmetric top is considered. The parameters are corresponding to $m=1$, $l=0.04$, and the principal moments of inertia tensor with respect to the fixed point $[J_1, J_2 = J_3] = [0.002, 0.0008]$. The following initial condition is considered

$$[\phi_0, \theta_0, \psi_0] = [0, \pi/6, 0], \quad \boldsymbol{\Omega}_0^T = [0, 0, 40\pi] \tag{78}$$

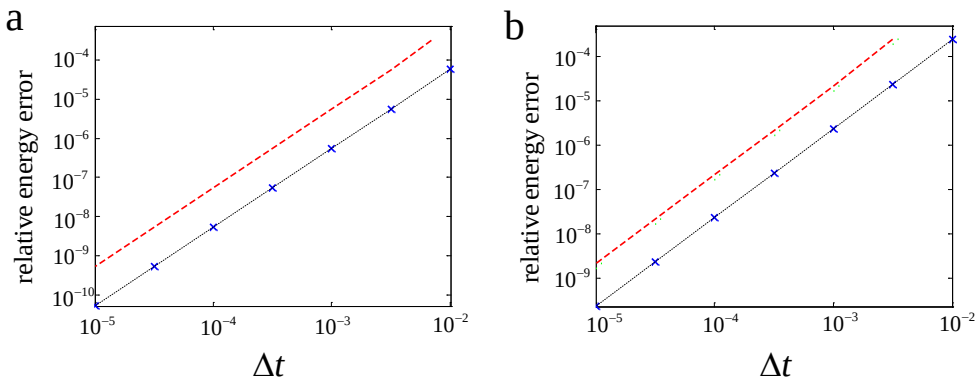



Figure 3: the maximum of periodic energy error with time step increasing. (a) unit quaternion, (b) convected base vectors. $\sigma = 0$ ( --- ), $\sigma = \sigma_m$ ( × ).

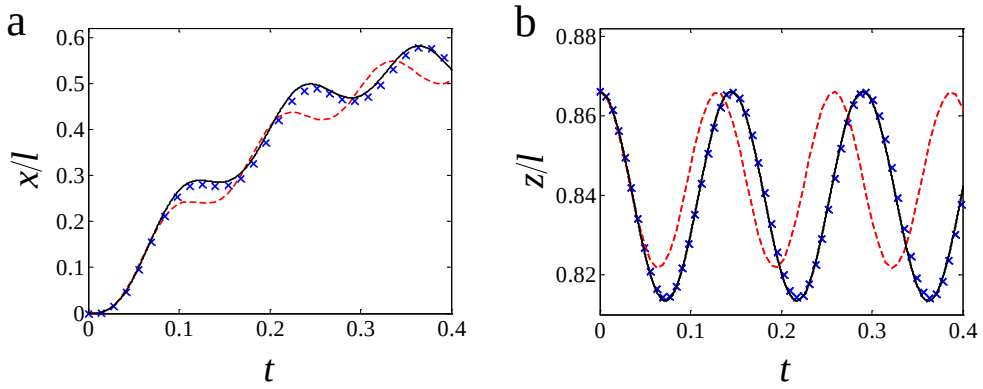



Figure 4: Trajectory errors of variational integrator with unit quaternion. (a) the $x$ - component of mass center, (b) the $z$ - component of mass center. $\sigma = 0$ (---), $\sigma = \sigma_{\mathrm{m}}$ (×), analytical (—), time step $\Delta t = 0.007$.

Figures 3-5 show numerical comparison of variational integrators with different values of scaling parameter $\sigma$. Figure 3 presents that the variational integrator is a two-order algorithm. Figures 3-5 further present that the variational integrator with $\sigma = \sigma_{\mathrm{m}}$ is of impressively higher accuracy than that with $\sigma = 0$. Considered that $\boldsymbol{M}_{\sigma}(\sigma = 0) = \boldsymbol{M}$, the value of the scaling parameter $\sigma$, in modified inertia representation, can greatly influence the numerical accuracy of integrations , although it can be determined arbitrarily in theoretical analysis.

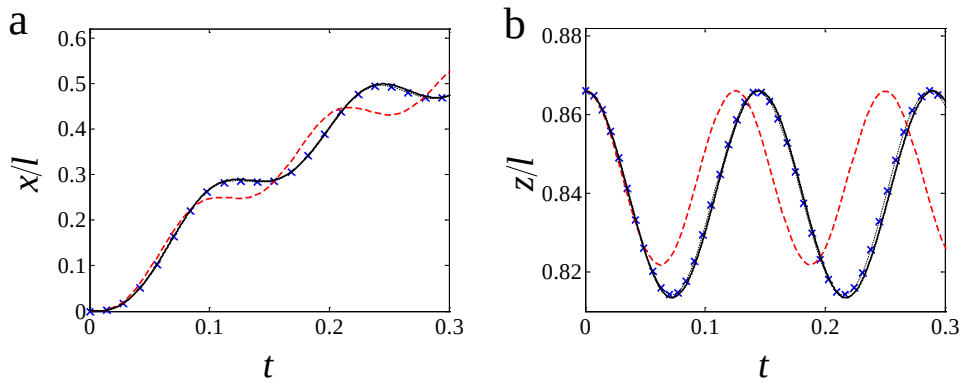



Figure 5: Trajectory errors of variational integrator with convected base vectors. (a) the $x$ - component of mass center, (b) the $z$ - component of mass center. $\sigma = 0$ (---), $\sigma = \sigma_{\mathrm{m}}$ (×), analytical (—), time step $\Delta t = 0.0035$

## 5.2 Regular precession

We secondly consider the regular precession of a heavy top in gravitational field. The top is represented as a cone with dimensions equivalent to those used in [16-18]. As illustrated in Figure 2, the parameters are height $h = 0.1$, $l = \frac{3}{4}h$, radius $r = h/2$ and the mass $m = \rho\pi^2 h/3$ with the mass density $\rho = 2700$. The principal values of the inertia moments with respect to fixed point are given by

$$J_1 = J_2 = 0.6\, m(0.25r^2 + h^2), \quad J_3 = 0.3\, mr^2 \tag{79}$$

The regular precession requires that nutation angle $\theta$, the velocity components $\dot{\phi}$ and $\dot{\psi}$ satisfy the relation

$$\dot{\psi} = (J_3\dot{\phi})^{-1} mgl + J_3^{-1}(J_1 - J_3)\dot{\phi}\cos(\theta) \tag{80}$$

see e.g. [1]. The initial conditions correspond to those used in [16-18], i.e. a precession rate $\dot{\phi} = 10$, an initial nutation angle $\theta_0 = \pi/3$ and the initial angular velocity vector

$$\boldsymbol{\Omega}_0 = [0, \dot{\phi}\sin(\theta_0), \dot{\psi} + \dot{\phi}\cos(\theta_0)]^T \tag{81}$$

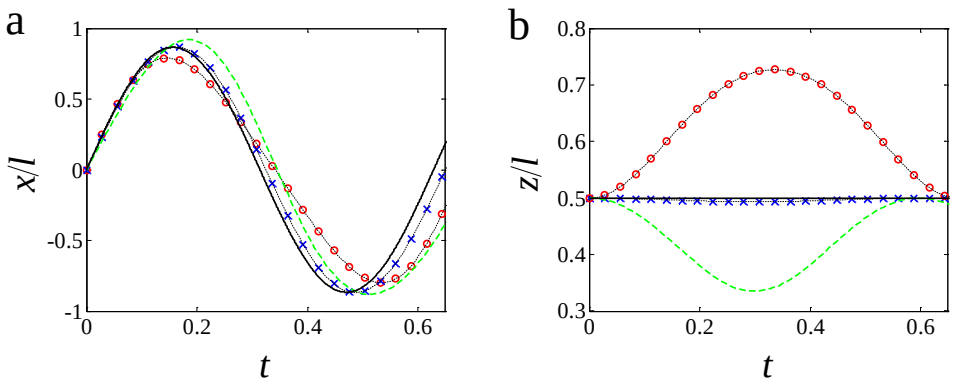



Figure 6: Numerical integrations with unit quaternion. (a) the $x$ - component of mass center, (b) the $z$ - component of mass center. Variational integrator with $\sigma = 0$ ( ∘ ) and $\sigma = \sigma_m$ ( × ), energy-momentum conserving scheme ( --- ), analytical solution ( — ), time step $\Delta t = 0.01$.

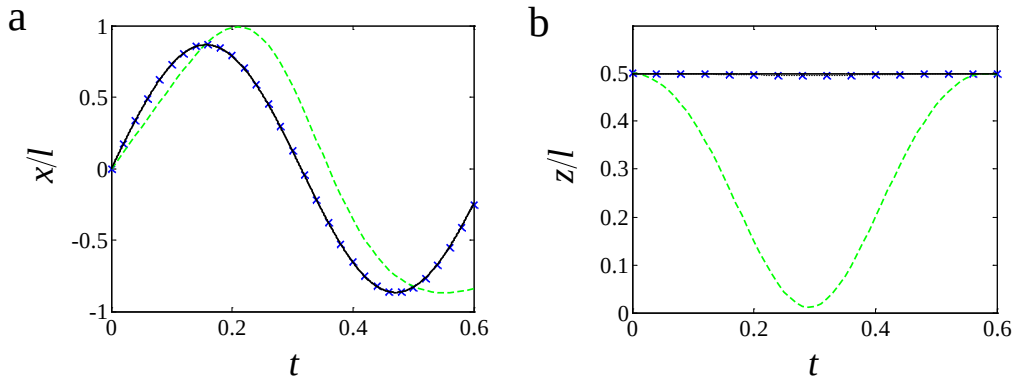



Figure 7: Numerical integrations with convected base vectors. (a) the $x$ - component of mass center, (b) the $z$ - component of mass center. Variational integrator with $\sigma = \sigma_m$, energy-momentum conserving scheme ( --- ), analytical solution ( — ), time step $\Delta t = 0.01$.

Recently, regular precession top is detailed discussed by Krenk and Nielsen, in their researches about energy-momentum conserving integrations of rigid body dynamics [17,18]. They observed that the numerical integrations are of significant nutation error in simulation of regular precession, whether the integrations are implemented in terms of unit quaternion or convected base vectors. Figures 6-7 compare the numerical results of regular precession between the variational integrator and energy-momentum conserving schemes proposed by Krenk and Nielsen [17, 18]. It should be mentioned that the convected-base-vectors-based variational integrator with $\sigma = 0$ and $\Delta t = 0.01$ presents almost wrong numerical results in simulation, and hence is not shown in Figure 7. In addition, the energy-momentum schemes are implemented in terms of unit quaternion and convected base vector in Figures 6 and 7, respectively. It can be observed that the variational integrator with $\sigma = 0$ and energy-momentum conserving integration both present significant numerical errors, whereas the vari-

ational integrator with $\sigma = \sigma_{\mathrm{m}}$ present impressively high accuracy in simulation. In combination of Figures 6-7 with Figures 3-5, we can conclude that the numerical accuracy of integrations can be improved in modified representation of inertia, associated with the scaling parameter $\sigma$, and the mean value of three principal moments of inertia tensor can be considered as a reasonable value of the scaling parameter, with which the numerical integrations are expected to present much higher accuracy than others.

## 6 CONCLUSIONS

A modified inertia representation is proposed for rigid body dynamics, through splitting the kinetic energy into two parts, that a square term of velocity and a quadratic form in the derivatives with quadratic coefficients in generalized displacement. The proportion of the two components is controlled by a scaling parameter, denoted as $\sigma$. The associated mass matrix with different values of scaling parameter, are theoretically equivalent in dynamics. However, it presents different features while discretizing the govern equations of the system. Error estimation demonstrates that the discretization error of kinetic energy is a linear function with respect to the scaling parameter, of which the slope is not equal to zero if only the rotational coordinates are expressed in pseudo-vector, such as unit quaternion, convected base vectors, etc.

Error estimation implies that the numerical accuracy of integration can be improved by optimizing the scaling parameter. Based on the modified representation of inertia, a variational integrator is proposed for rigid body dynamics in terms of pseudo vectors. Two examples that the fast spinning top and the regular precession, are considered to testing the numerical performance of the scaling parameter. Numerical results demonstrate that the variational integrator with $\sigma=\sigma_{\mathrm{m}}$ presents impressively higher accuracy in simulation than other integrations. Consequently, the arithmetic mean value of the principal moments of inertia tensor (i.e. $\sigma_{\mathrm{m}}$), is a reasonable preconditioning value of the scaling parameter, with which the numerical integrations are expected to present much higher accuracy than others.

## REFERENCES

[1] H. Goldstein, C. Poole, J. Safko, *Classical mechanics, 3rd Edition*. Addison Wesley, 2002.

[2] E. Hairer, S.P. Norsett, G. Wanner, *Solving ordinary differential equation, Vol. 1 and Vol. 2*. Springer Verlag.

[3] C.W. Gear, Differential-algebraic equation index transformations. *SIAM Journal on Scientific and Statistical Computing*, **9**, 39-47, 1988.

[4] L.R. Petzold, Differential/algebraic equations are not ODE's. *SIAM Journal on Scientific and Statistical Computing*, **3**, 367-384, 1982.

[5] C.W. Gear, L.R. Petzold, ODE methods for the solution of differential/algebraic systems. *SIAM Journal On Numerical Analysis*, **21**, 716-728, 1984.

[6] K.E. Brenan, S.L. Campbell, L.R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations, Vol. 14*. Siam, 1996.

[7] R. Featherstone, *Rigid body dynamics algorithms*. Springer, 2014.

[8] E. Hairer, C. Lubich, G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, 2006.

[9] J.M. Wendlandt, J.E. Marsden, Mechanical integrators derived from a discrete variational principle. *Physica D: Nonlinear Phenomena*, **106**, 223-246, 1997.

[10] S. Leyendecker, J.E. Marsden, M. Ortiz, Variational integrators for constrained dynamical systems. *ZAMM‐Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, **88**, 677-708, 2008.

[11] A. Dullweber, B. Leimkuhler, R. McLachlan, Symplectic splitting methods for rigid body molecular dynamics. *The Journal of Chemical Physics*, **107**, 5840-5851, 1997.

[12] T.F. Miller, M. Eleftheriou, P. Pattnaik, etc., Symplectic quaternion scheme for biophysical molecular dynamics. *The Journal of Chemical Physics*, **116**, 8649-8659, 2002.

[13] R.I. McLachlan, C. Scovel, Equivariant constrained symplectic integration. *Journal of Nonlinear Science*, **5**, 233-256, 1995.

[14] J.C. Simo, K.K. Wong, Unconditionally stable algorithms for rigid body dynamics that exactly preserve energy and momentum. *International Journal for Numerical Methods in Engineering*, **31**, 19-52, 1991.

[15] P. Betsch, P. Steinmann, Constrained integration of rigid body dynamics. *Computer Methods in Applied Mechanics and Engineering*, **191**, 467-488, 2001.

[16] P. Betsch, R. Siebert, Rigid body dynamics in terms of quaternions: Hamiltonian formulation and conserving numerical integration. *International Journal for Numerical Methods in Engineering*, **79**, 444-473, 2009.

[17] M.B. Nielsen, S. Krenk, Conservative integration of rigid body motion by quaternion parameters with implicit constraints. *International Journal for Numerical Methods in Engineering*, **92**, 734-752, 2012.

[18] S. Krenk, M.B. Nielsen, Conservative rigid body dynamics by convected base vectors with implicit constraints. *Computer Methods in Applied Mechanics and Engineering*, **269**, 437-453, 2014.

[19] X. Xu, W. Zhong, On the numerical influences of inertia representation for rigid body dynamics in terms of unit quaternion. *ASME Journal of Applied mechanics(accepted)*, 2016.

[20] C.L. Bottasso, O.A. Bauchau, A. Cardona, Time-step-size-independent conditioning and sensitivity to perturbations in the numerical solution of index three differential algebraic equations. *SIAM Journal On Scientific Computing*, **29**, 397-414, 2007.

[21] S. Leyendecker, J.E. Marsden, M. Ortiz, Variational integrators for constrained dynamical systems. *ZAMM Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, **88**, 677-708, 2008.

[22] J.M. Wendlandt, J.E. Marsden, Mechanical integrators derived from a discrete variational principle. *Physica D: Nonlinear Phenomena*, **106**, 223-246, 1997.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# APPROACHES TO THE CREATION OF MULTIBODY MODELS OF THE VVER 1000 NUCLEAR REACTOR CONTROL ASSEMBLY

**Pavel Polach**[1,2]**, and Michal Hajžman**[2]

[1] Section of Materials and Mechanical Engineering Research, Research and Testing Institute Plzeň
Tylova 1581/46, CZ-301 00 Plzen, Czech Republic
e-mail: polach@vzuplzen.cz

[1,2] Faculty of Applied Sciences, University of West Bohemia
Univerzitní 8, CZ-306 14 Plzeň, Czech Republic
ppolach@ntis.zcu.cz, mhajzman@kme.zcu.cz

**Keywords:** Nuclear Reactor, Control Assembly, Multibody Model, Drop, Contact.

***Abstract.*** *Multibody systems modelling and analysis are efficient tools utilizable in general tasks of nonlinear dynamics. One of the branches suitable for the application of multibody approaches is the modelling of nonlinear motion and dynamical analysis in nuclear engineering. The motion of important parts of nuclear reactors called control assemblies has to be analyzed. The purpose of these control systems is to control the power of a nuclear reactor and possibly to stop the reaction in case of an emergency state. With respect to different types of nuclear reactors, different control systems composed of various mechanical parts and transmissions can be distinguished. Generally, they can be simplified to the typical problem of a long thin rod moving through guide tubes and driven by a motor. Then the modelling approach depends on the chosen mode of operation, which can be operation (regulation) under normal conditions or a certain emergency state. The paper is focused on the control assembly of the VVER 1000 nuclear reactor. The VVER 1000 reactor is a pressurized water-cooled and water-moderated reactor consisting of a reactor pressure vessel with an interior structure and a reactor upper block with 61 control assembly drives. Under the vessel head with nozzles of the control assembly drives there is a block of protective tubes, which is above the core with 163 fuel assemblies. A moveable part of the control assembly, i.e. a rod control cluster assembly (this part is the necessary condition for stopping the nuclear reaction), is composed of a suspension bar, a spider and 18 long thin absorber rods. Two modular spatial multibody models of the LKP-M/3 control assembly of the VVER 1000 nuclear reactor were created in the alaska simulation tool and they differ in their kinematic scheme and in the number of contact positions. Both models are meant for the simulations of the control assembly moving parts drop. The first one is intended for the simulations of the rod control cluster assembly drop during the seismic event, where the total time of the drop is the significant measured and computed quantity, and the second one for investigating limit curves of deformations of fuel assembly guide tubes, at which the absorber rods still reach the lower part of the core.*

## 1 INTRODUCTION

Multibody systems modelling and analysis are efficient tools utilizable in general tasks of nonlinear dynamics. They are used for solving many different mechanical problems in various branches of industry. Multibody approaches are based on the concepts of classical mechanics with different complicated couplings and applied force effects. They are useful due to their ability to handle complex systems of bodies with nonlinear behaviour and they can be applied in different multidisciplinary applications.

One of the branches suitable for the application of multibody approaches is the modelling of nonlinear motion and dynamical analysis in nuclear engineering. Besides the necessary seismic analyses of reactor buildings and structures, which belong rather to structural dynamics, the motion of important parts of nuclear reactors called control assemblies has to be analyzed [1]. The purpose of these control systems is to control the power of a nuclear reactor (i.e. to control nuclear reaction) and possibly to stop the nuclear reaction in case of a defined emergency state. With respect to different types of nuclear reactors, different control systems composed of various mechanical parts and transmissions can be distinguished [2]. Generally, they can be simplified to the typical problem of a long thin rod moving through guide tubes and driven by a motor. The influences of contacts and influences of a coolant (water) surrounding the control assembly in the nuclear reactor would be considered in the mathematical models (note: the relevant sources for their introduction are given in the following chapters). The modelling approach depends on the chosen mode of operation, which can be operation (regulation) under normal conditions or a certain emergency state, in addition.

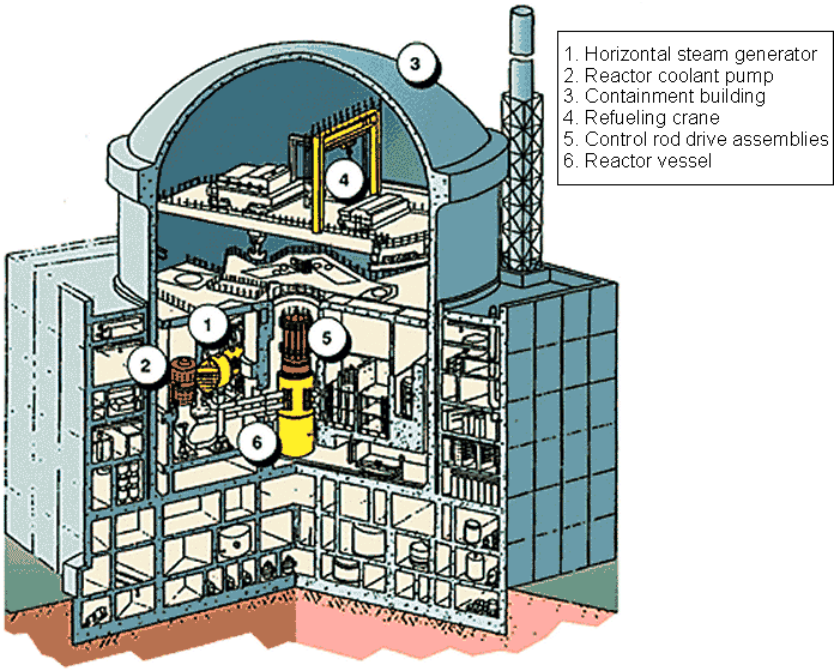



Figure 1: Scheme of the VVER 1000 plant layout.

The problem of the mechanical behaviour of a control assembly during the emergency drop is not commonly studied in available literature and therefore the investigation of this problem is a challenge. Some basic studies can be found in [1] or [3], but they are not focused on mechanical behaviour in a sufficient detail. Authors of this paper presented several approaches to modelling the flexible falling rod in [2, 4, 5, 6, 7] and more detailed models of control assemblies in [8, 9, 10, 11].

The paper is focused on multibody modelling of the control assembly of the VVER 1000 reactor. The scheme of the VVER 1000 plant layout is given in Fig. 1.

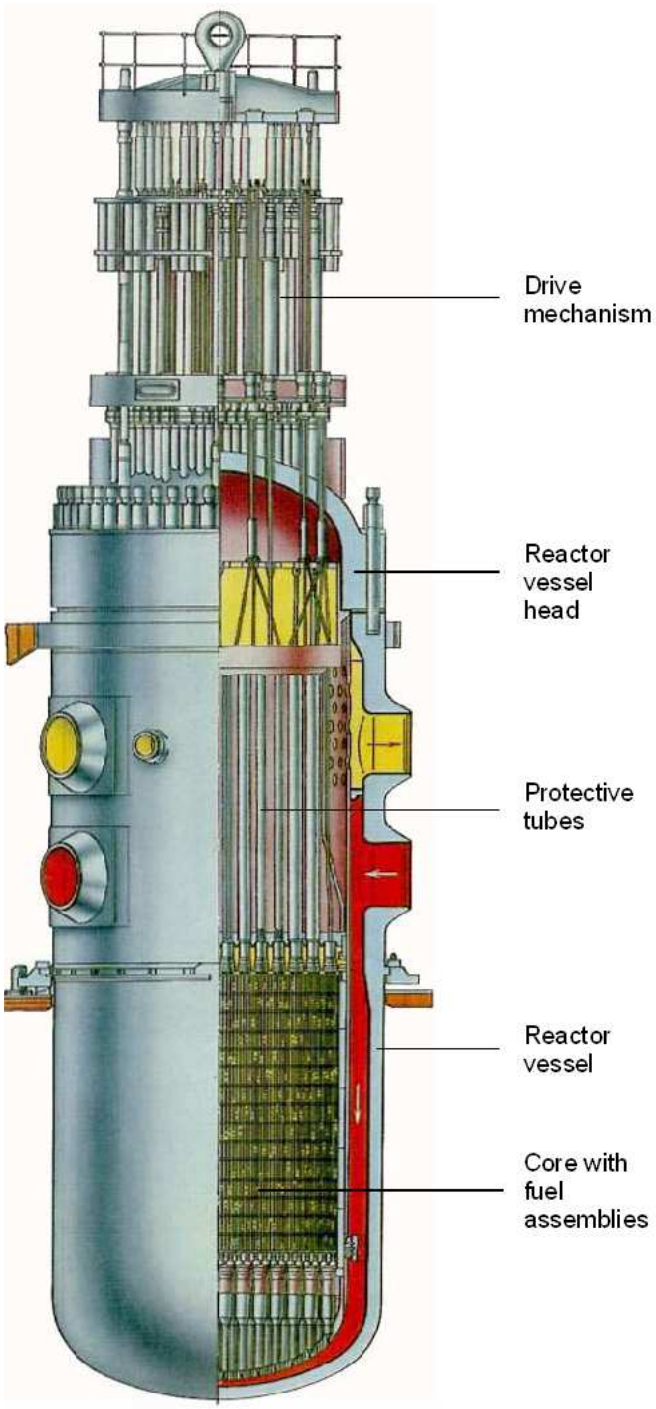



Figure 2: Scheme of the VVER 1000 reactor.

Two modular spatial multibody models of the LKP-M/3 control assembly of the VVER 1000 nuclear reactor were created in the *alaska* 2.3 simulation tool [12] and they differ in their kinematic scheme and in the number of contact positions. Both models are meant for the simulations of the control assembly moving parts drop. The first one is intended mainly for the simulations of the rod control cluster assembly drop during the seismic event [11], where the significant measured and computed quantity is the total time of the drop, and the second

one for investigating limit curves of deformations of fuel assembly guide tubes [13], at which the absorber rods still reach the lower part of the core.

## 2 VVER 1000 NUCLEAR REACTOR

The VVER 1000 nuclear reactor is a pressurized water-cooled and water-moderated reactor developed in the Soviet Union and is used in many countries (e.g. the Czech Republic, the Ukraine, Bulgaria, China etc.). This reactor type is very similar to the PWR reactors produced in the USA and some west European countries.

The VVER 1000 reactor consists of a reactor pressure vessel with an interior structure and a reactor upper block with control assembly drives (see scheme in Fig. 2).

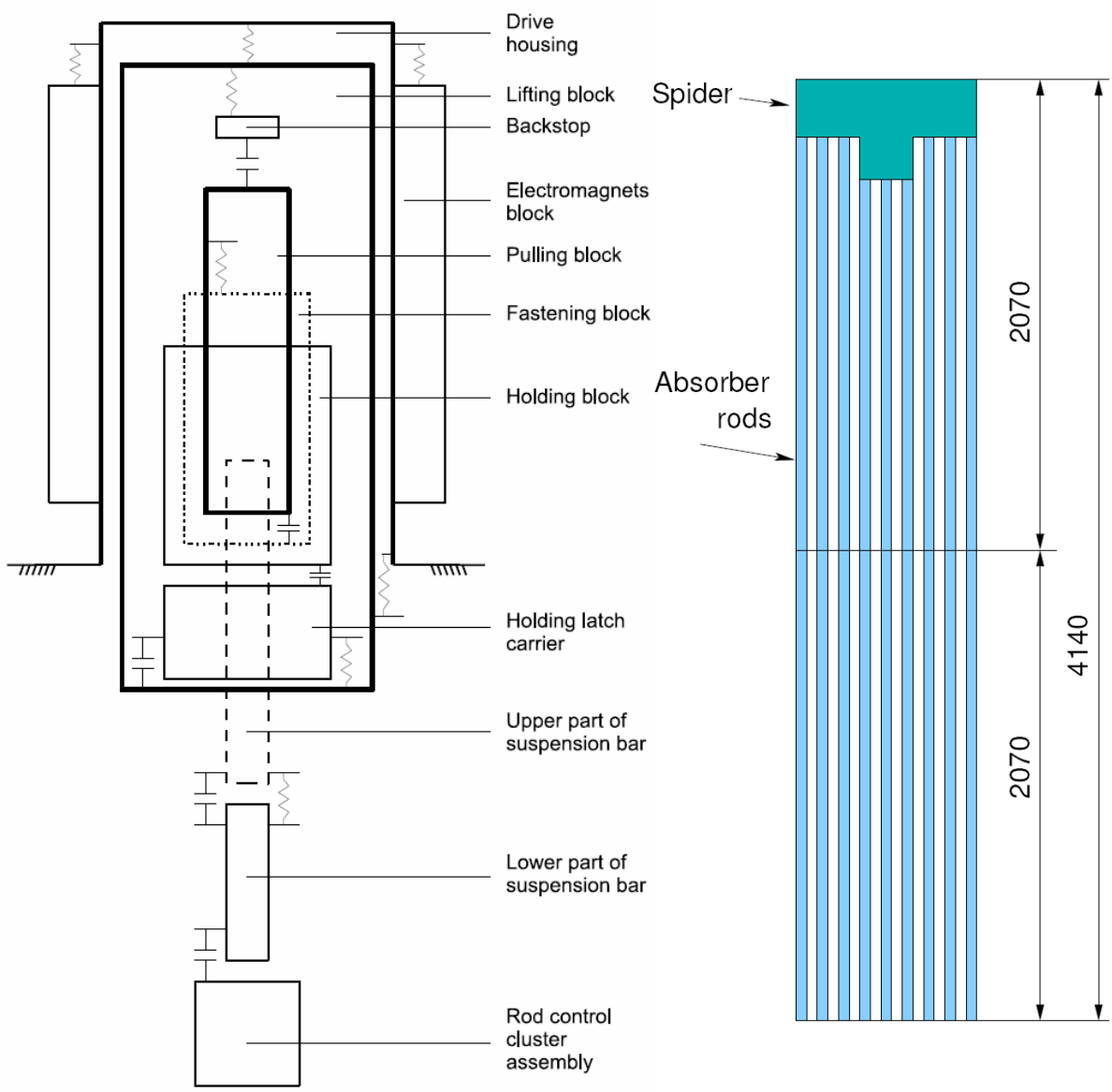



Figure 3: Scheme of the control assembly without casing (not in a real scale factor) and scheme of the rod control cluster assembly.

Under the vessel head with nozzles of the control assembly drives there is a block of protective tubes, which is above the core with fuel assemblies (so called active zone). Fuel assemblies are placed on the core support plate.

The VVER 1000 reactor contains 163 fuel assemblies and 61 control assemblies (e.g. [14]). A moveable part of the control assembly called a rod control cluster assembly (this part is the necessary condition for the stopping of nuclear reaction), is composed of a suspension bar, a spider and 18 long thin absorber rods.

The detailed scheme of the control assembly is shown in Fig. 3. It has to be noted that the objects in this figure are not displayed in a real scale factor, especially the rod control cluster assembly. The suspension bar is divided into an upper and a lower part with a bayonet joint for connecting with a spider of the control element (i.e. the rod control cluster assembly). The suspension bar of the rod control cluster assembly is driven by a linear stepper motor, the mechanism of which is based on the usage of electro-magnets. In case of an emergency state the lifting system mechanism is set off and the rod control cluster assemblies can drop down through the protective tubes to the reactor core with fuel assemblies and stop the nuclear reaction. Other functional subsystems plotted in Fig. 3 do not influence the dynamic behaviour in the course of the studied rod control cluster assembly drop. The spider and the absorber rods of the rod control cluster assembly are schematically drawn in Fig. 3.

When the absorber rods reach the lower part of the core, they pass through the guide channel narrowing, which have the function of hydraulic shock absorbers for the stop of the rod control cluster assemblies drop.

## 3 MULTIBODY MODELS OF THE CONTROL ASSEMBLY

The spatial multibody models of the control assembly were created in the *alaska* 2.3 simulation tool [12] on the basis of the technical documentation and drawings provided by the control assembly producer ŠKODA JS a.s. (the Czech Republic). As it has been already mentioned, the first one is intended mainly for the simulations of the rod control cluster assembly drop during the seismic event [11] (model 1), where the significant measured and computed quantity is the total time of the drop, and the second one for investigating limit curves of deformations of fuel assembly guide tubes [13] (model 2), at which the absorber rods still reach lower part of the core. Basic summary data of both multibody models are given in Tab. 1.

| | Model intended for simulations of the rod control cluster assembly drop | |
|---|---|---|
| | during the seismic event (model 1) | at considering deformations of fuel assembly guide tubes (model 2) |
| Number of (rigid) bodies | 23 | 22 |
| Number of kinematic joints | 23 | 22 |
| Number of degrees of freedom | 56 | 45 |
| Number of contact points in vertical direction | 10 | 11 |
| Number of contact positions in horizontal plane | 40 | 70 |

Table 1: Basic summary data about multibody models of the control assembly.

Two different computational models of the control assembly were created due to the complex including of all effects influencing the rod control cluster assembly drop, which results in relatively long computational times of simulations (in the order of tens of hours). In model 1 intended mainly for the simulations of the rod control cluster assembly drop during the seismic event [11] it was necessary to consider more degrees of freedom due to introducing a

seismic excitation and in model 2 it was necessary to consider more contact positions in a horizontal plane for investigating limit curves of deformations of fuel assembly guide tubes [13] (see Tab. 2). Dynamical aspects of the control assembly that do not influence the rod control cluster assembly drop were neglected in both computational models.

Problems of dynamical mechanical systems in nuclear engineering are of a strong multidisciplinary character. The presented multibody models are the intermediate control assembly models that should be the initial state to start the detailed study of the control assembly behaviour. The multibody models include coolant influences, impacts, contacts and, in case of model 1, seismic excitation. Not all influences and specific behaviour could be considered in detail in the models, but some aspects of the control assembly are and will be studied in more detail in future works.

| Contact between bodies (see Fig. 4) | | Number of contact positions | |
|---|---|---|---|
| First body | Second body | Model 1 | Model 2 |
| upper part of suspension bar | lifting block | 1 | 1 |
| upper part of suspension bar | pulling block | 2 | 2 |
| upper part of suspension bar | protective tube, fuel assembly | 1 | - |
| upper part of suspension bar | frame | - | 1 |
| cluster spider | protective tube, fuel assembly | 6 | - |
| cluster spider | frame | - | 6 |
| upper parts of 1st to 6th threes of absorber rods | protective tube, fuel assembly | 18 | - |
| upper parts of 1st to 6th threes of absorber rods | frame | - | 24 |
| lower parts of 1st to 6th threes of absorber rods | protective tube, fuel assembly | 12 | - |
| lower parts of 1st to 6th threes of absorber rods | frame | - | 36 |

Table 2: Contact positions of bodies in horizontal plane in the multibody models.

### 3.1 Kinematic scheme, contact positions and seismic excitation

Kinematic scheme of the control assembly multibody model is shown in Fig. 4, where circles represent kinematic joints (BUNC – unconstrained, PRI – prismatic in vertical axis, UNI – universal around horizontal axes) and quadrangles represent rigid bodies. In the computational model, the rigid bodies are described by their mass and inertia properties computed analytically for simple shapes or by means of the COSMOS/M software [15] based on the finite element method.

Due to virtual division of the rod control cluster assembly into thirteen rigid bodies (the purpose of this virtual division is partial considering the elastic properties of the rod control cluster assembly) certain distinguishing names of its parts were introduced (cluster spider, upper parts of the 1st to 6th threes absorber rods and lower parts of upper parts of the 1st to 6th threes absorber rods). The terminology is evident from Fig. 4 (or partly from Fig. 3 left). This modelling enables, using appropriately chosen torsional stiffnesses in the kinematic joints, to “tune” the values of the lowest natural frequencies corresponding to bending vibration modes to the natural frequencies obtained from the FEM model of the rod control cluster assembly.

The best way of introducing the seismic excitation in the computational model is presented in [8, 9, 10] (and in internal research report [16]), the introducing of fuel assembly guide tubes deformations is given in available publication first time ([13] is internal research report).

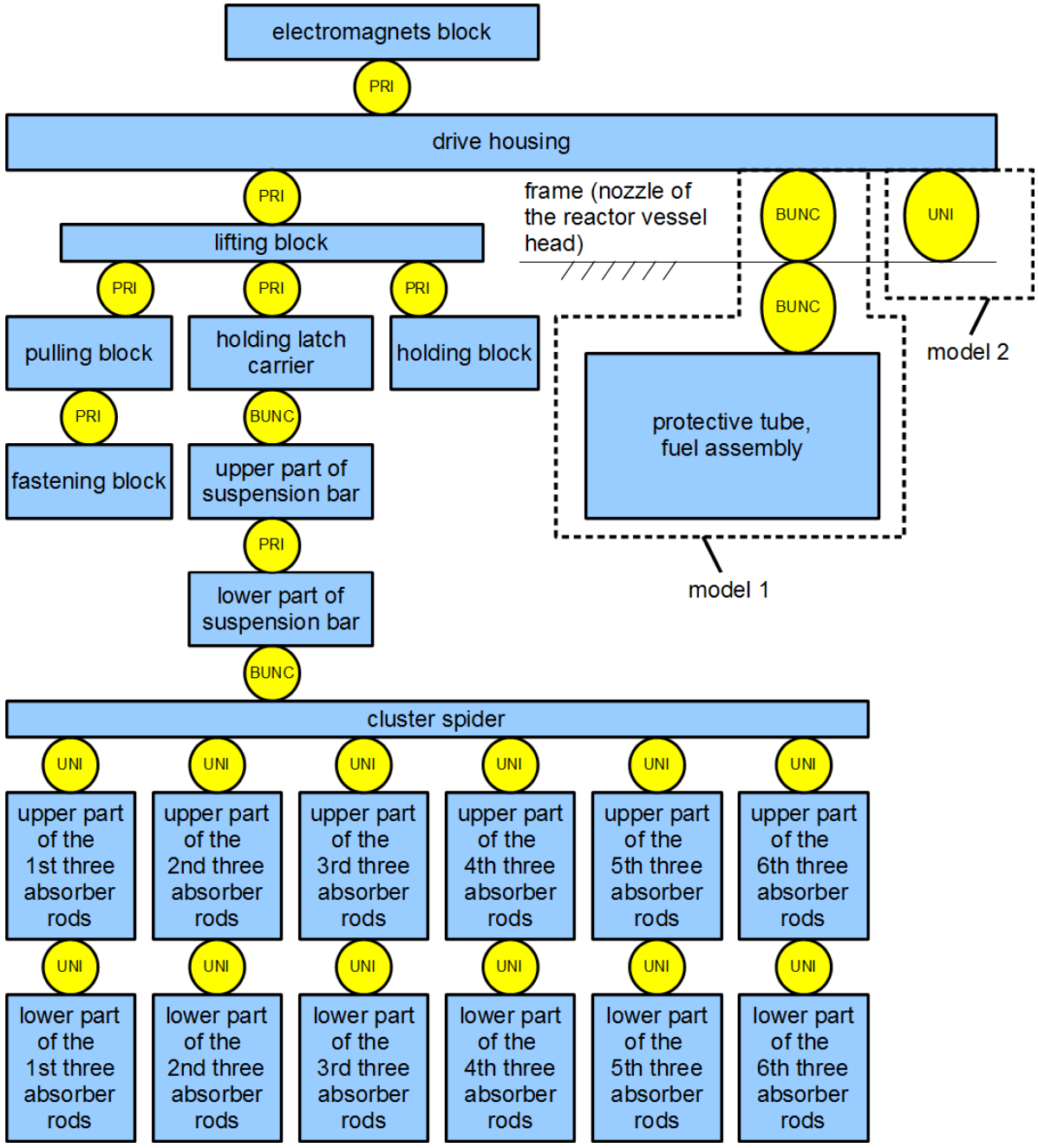



Figure 4: Kinematic scheme of the control assembly multibody models.

In multibody model 1 (i.e. model intended mainly for the simulations of the rod control cluster assembly drop during the seismic event), a special rigid body substituting interior parts in the reactor ("protective tube, fuel assembly" – see Fig. 4) is important to specify contact and impact conditions in the course of the rod control cluster assembly drop. Motion of this body and body "drive housing" are defined by seismic excitations, which are specified by time histories of the absolute displacements in three directions (see example of time history of seismic excitation in Fig. 5). In multibody model 2 (i.e. model intended for investigating limit

curves of deformations of fuel assembly guide tubes), protective tube and fuel assembly are modelled as a part of the frame (see Fig. 4). Both considered time histories of normalized deformations of fuel assembly guide tubes are given in Fig. 6.

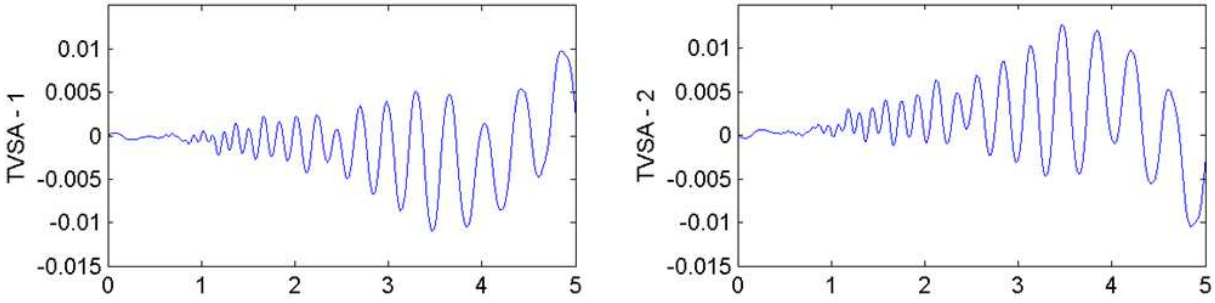



Figure 5: Time histories [s] of the absolute displacements [m] in horizontal directions in place of connection of nozzle of the reactor vessel head and drive housing during seismic event (prescribed motion of the body "drive housing").

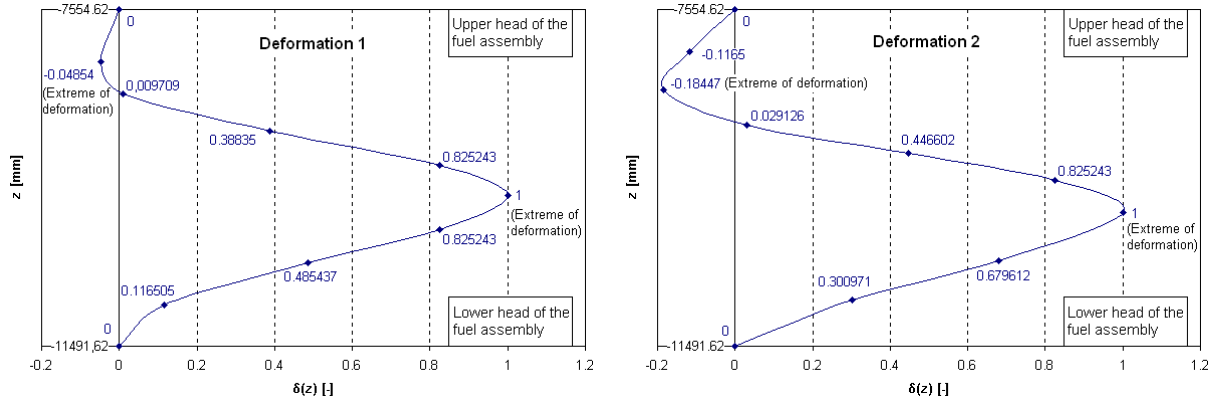



Figure 6: Normalized deformations of fuel assembly guide tubes.

## 3.2 Modelling of contacts and impacts

Possible contacts and impacts of the moving parts with adjacent structural parts inside the reactor are very important and significant aspects of the rod control cluster assembly drop modelling. Many publications studying contact problems were released, e.g. [17, 18, 19, 20]. However, it was necessary to use some simple contact-impact model of the rigid bodies applicable in the *alaska* 2.3 simulation tool.

The problem can be divided into two steps – the first one is the determination of the contact occurrence and contact position and the second one is the calculation of the impact force acting on the bodies. The possible bodies in contact were specified on the basis of technical documentation and drawings. Since the clearances between the falling bodies (parts of the rod control cluster assembly) and adjacent structures (protective tubes, fuel assembly guide tubes) are relatively small – 1 to 7.5 millimetres – the contacts occur frequently. In all cases the contact of the body of circular cross section with the circular hole (the body moves through the hole) can occur in the multibody model. The simple geometrical consideration can decide if the contact of the body cross section boundary curves occurs or not (see Fig. 7). It holds

$$R-\sqrt{x_S^2+y_S^2}-r\le 0 \ , \tag{1}$$

if the body is in contact with the adjacent structure. Coordinate system axes are designated by letters $x$ and $y$, $R$ is the hole radius, $r$ is the radius of the body cross section, $x_S$ and $y_S$ are

horizontal displacements of the body centre. If the condition (1) is fulfilled angle $\alpha = \arctan\frac{y_S}{x_S}$ can be computed and relative deformation $d$ of the contact surfaces can be expressed

$$d = R - \sqrt{x_K^2 + y_K^2}\,, \tag{2}$$

where $x_K = x_S + r \cdot \cos\alpha$ and $y_K = y_S + r \cdot \sin\alpha$ are coordinates of contact point K.

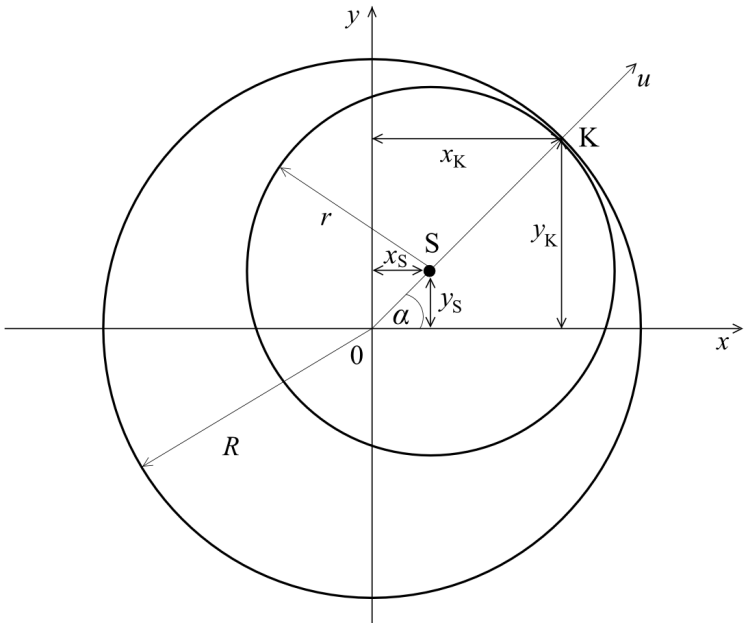



Figure 7: Contact of the body of the circular cross section moving through the circular hole.

Relative deformation $d$ is used in the impact force evaluation. E.g. the simple nonlinear Hertzian law (e.g. [21])

$$F = k \cdot d^n\,, \tag{3}$$

can be used for the impact force evaluation, where $k$ is the generalized stiffness and $n$ is the constant coefficient for given materials.

The contacts are related to vertical motion and therefore the friction force

$$F_t = f \cdot F\,, \tag{4}$$

where $f$ is the friction coefficient and $F$ is the impact (normal) force acting in the contact, is introduced.

The impact (normal) forces are introduced in the control assembly multibody model as the applied forces. Their instantaneous value is (instead of formula (3)) given by the *bistop* function [12]. After introducing axis $u$, direction of which is given by connecting point 0 (centre of the circular hole) and instantaneous position of contact point K, and independent variable $u$ in its direction, impact (normal) force is calculated using the *bistop* function

$$F = bistop\left(u,\dot{u},u_1,u_2,c,d_b,e,b\right) = \begin{cases} -c\cdot\left(u_1-u\right)^e + step\left(u,u_1-d_b,b,u_1,0\right)\cdot\dot{u}\,, & u \le u_1 \\ 0\,, & u_1 < u < u_2 \\ c\cdot\left(u-u_2\right)^e + step\left(u,u_2,0,u_2+d_b,b\right)\cdot\dot{u}\,, & u_2 \le u\,, \end{cases} \tag{5}$$

where $u$ is the independent variable, $\dot{u}$ is the first-order derivative of $u$ and $c$, $d_b$, $e$, $b$ are the coefficients determining the course of the function when one of the boundary values $u_1 = r - R$ or $u_2 = R - r$ is exceeded.

Function *step* approximates a step function by evaluating a cubic polynomial, which starts at the initial ordinate value $h_0$ at $u = u_0$ and ends at the final ordinate value $h_1$ at $u = u_E$

$$step(u, u_0, h_0, u_E, h_1) = \begin{cases} h_0 , & u \le u_0 \\ h_0 + (h_1 - h_0) \cdot \left( \dfrac{u - u_0}{u_E - u_0} \right)^2 \cdot \left( 3 - 2 \cdot \dfrac{u - u_0}{u_E - u_0} \right), & u_0 < u < u_E \\ h_1 , & u_E \le u . \end{cases} \quad (6)$$

### 3.3 Influence of the coolant

The most problematic factor in the control assembly multibody model is the influence of the pressurized water (coolant) that flows through the reactor interior structure and acts against the rod control cluster assembly motion. Like in the case of the contacts and impacts, this issue is frequently studied in the theoretical way (e.g. [22]) but the simple practical approach was necessary in case of modelling such a complex system in the *alaska* 2.3 simulation tool. It is evident that the main influence of coolant is the hydraulic resistance and friction slowing down the rigid body motion.

If $m_F = \rho_F \cdot V$ is the fluid mass pushed up by the body of volume $V = \dfrac{m_B}{\rho_B}$ static uplift pressure is respected in the corrected gravity force acting on the body

$$G = m_B \cdot g - m_F \cdot g = m_B \cdot g \cdot \left( 1 - \frac{\rho_F}{\rho_B} \right), \quad (7)$$

where $m_B$ is the body mass, $g$ is the gravity acceleration, $\rho_B$ is the body material density and $\rho_F$ is the coolant density.

The influence of the fluid on the horizontal vibrations is involved into the control assembly multibody model considering the hydraulic resistance force in the form (e.g. [23])

$$F_H = \frac{1}{2} \cdot C_H \cdot \rho_F \cdot S_H \cdot v_H^2 . \quad (8)$$

The relative horizontal velocity of the body is designated $v_H$, $S_H$ is the body effective surface and $C_H$ is the hydraulic coefficient depending on the shape of the body.

Presence of the flowing medium has essential influence on the vertical motion of the rod control cluster assembly. This effect is significant due to the falling bodies motion in the fluid filling the limited volume of the protective tubes and the fuel assemblies guide tubes and because of the fact that the fluid cannot leak out in the horizontal directions. The vertical resistance force is introduced in the control assembly multibody model using formula

$$F_V = \frac{1}{2} \cdot C_V \cdot \rho_F \cdot S_V \cdot w^2 . \quad (9)$$

The total vertical velocity of the body $w$ is the vector sum of the falling body velocity and the fluid velocity flowing against the falling body, $S_V$ is the body effective surface and $C_V$ is

the hydraulic coefficient depending on the shape of the body. Due to the complexity of the cluster surface the hydraulic resistance coefficient $C_V$ is difficult to determine accurately by the simple computation. The fact that rod control cluster assembly falls in the limited volume also makes the estimation more difficult. The way of the hydraulic resistance coefficient $C_V$ determination is given in [16].

## 4 CONCLUSIONS

The paper presents two modular spatial multibody models of the LKP-M/3 control assembly of the VVER 1000 nuclear reactor created in the *alaska* 2.3 simulation tool. Both models are meant for the simulations of the control assembly moving parts drop. Model 1 is intended mainly for the simulations of the rod control cluster assembly drop during the seismic event [11], where the significant measured and computed quantity is the total time of the drop, and model 2 for investigating limit curves of deformations of fuel assembly guide tubes [13], at which the absorber rods still reach the lower part of the core. The multibody models include the influences of the fluid and the contacts and impacts of the rod control cluster assembly with the adjacent structural parts inside the reactor.

The models with the considered influences should be understood to be of the intermediate level in the framework of this topic. Various problems arising from the solution of this task will be studied in more detail in the future work. Mainly the problems of the falling body that is in contact with other bodies and interacts with fluid (coolant) in limited space have to be investigated. The rod control cluster assembly is composed of many absorber rods and consideration of their flexibility seems to be useful. Introduction to modelling the flexible rods has already been given in [2, 5, 6, 7]. The advanced modelling of flexible rods based on the absolute nodal coordinate formulation (ANCF) is given in [24].

The paper has originated in the framework of solving LG 15058 project the Ministry of Education, Youth and Sports of the Czech Republic entitled “Presidency of CEACM and related posts” and institutional support for the long-time conception development of the research institution provided by the Ministry of Industry and Trade of the Czech Republic to Research and Testing Institute Plzeň.

## REFERENCES

[1] D. Bosselut, H. Andriambololona, E. Longatte, J. Pauthenet, Insertion and drop of control rod in assembly simulations and parametric analysis. *Technical meeting on structural behaviour of fuel assemblies for water cooled reactors, IAEA-TECDOC-1454*, Cadarache, France, November 22-26, 2004.

[2] M. Hajžman, P. Polach, Nonlinear Dynamic Analysis of Flexible Control Rods in Nuclear Reactors. *The Third Asian Conference on Multibody Dynamics 2006*, Tokyo, Japan, August 1-4, 2006.

[3] B. Collard, RCCA Drop Kinetics Test, Calculation and Analysis, Abnormal Friction Force Evaluation. *7th International Conference on Nuclear Engineering*, Tokyo, Japan, April 19-23, 1999.

[4] R. Bulín, M. Hajžman, P. Polach, Basic Model of a Control Assembly Drop in Nuclear Reactors. *Transactions of the VŠB – Technical University of Ostrava, Mechanical Series*, **59**, 1-7, 2013.

[5] M. Hajžman, P. Polach, Modelling of Flexible Rods Falling in Fluid with Possible Contacts. C. A. Mota Soares, J. A. C. Martins, H. C. Rodrigues, J. A. C. Ambrósio, C. A. B. Pina, C. M. Mota Soares, E. B. R. Pereira, J. Folgado eds. *III European Conference on Computational Mechanics: Solids, Structures and Coupled Problems in Engineering*, Lisbon, Portugal, June 5-8, 2006.

[6] M. Hajžman, P. Polach, Two Approaches to The Modelling of The Control Rods in Nuclear Reactors. *The Fourth Asian Conference on Multibody Dynamics ACMD2008*, Jeju, Republic of Korea, August 20-23, 2008.

[7] M. Hajžman, P. Polach, Modelling of flexible bodies in the framework of multibody systems. B. Skočilasová ed. *6th International Conference Dynamics of Rigid and Deformable Bodies 2008*, Ústí nad Labem, Czech Republic, September 17-19, 2008.

[8] M. Hajžman, P. Polach, Modelling and Seismic Response of the Control Assembly for the VVER 440/V213 Nuclear Reactor. J. M. Goicolea, J. Cuadrado, J. C. García Orden eds. *ECCOMAS Thematic Conference Multibody Dynamics 2005 on Advances in Computational Multibody Dynamics*, Madrid, Spain, June 21-24, 2005.

[9] M. Hajžman, P. Polach, Multibody analysis of the control assembly drop in the VVER 440/V213 nuclear reactor. J. Náprstek, C. Fischer eds. *National Conference with International Participation Engineering Mechanics 2006*, Svratka, Czech Republic, May 15-18, 2006.

[10] M. Hajžman, P. Polach, Seismic Analysis of the Control Assembly of the IRT-200 Research Nuclear Reactor by Means of a Multibody Approach. M. Papadrakakis, D.C. Charmpis, N.D. Lagaros, Y. Tsompanakis eds. *ECCOMAS Thematic Conference COMPDYN 2007 on Computational Methods in Structural Dynamics and Earthquake Engineering*, Rethymno, Crete, Greece, June 13-16, 2007.

[11] P. Polach, M. Hajžman, Multibody Model of the Control Assembly of the VVER 1000 Nuclear Reactor. B. Skočilasová ed. *7th International Conference Dynamics of Rigid and Deformable Bodies*, Ústí nad Labem, Czech Republic, September 9-11, 2009.

[12] P. Maißer, C.-D. Wolf, A. Keil, K. Hendel, U. Jungnickel, H. Hermsdorf, P. A. Tuan, G. Kielau, O. Enge, U. Parsche, T. Härtel, H. Freudenberg, *alaska, User Manual, Version 2.3*. Institute of Mechatronics, Chemnitz, Germany, 1998.

[13] P. Polach, M. Hajžman, *Multibody model of the LKP-M/3 control assembly of the VVER 1000 nuclear reactor for investigating limit curves of deformations of fuel assembly guide tubes, research report VYZ 0880/06*. ŠKODA VÝZKUM s.r.o., Plzeň, Czech Republic, 2006. (in Czech)

[14] V. Zeman, Z. Hlaváč, Dynamic Response of VVER 1000 Type Reactor Excited by Pressure Pulsation. *Engineering Mechanics*, **15**, 435-446, 2008.

[15] *COSMOS/M, Finite Element Analysis System, User Guide, Version 2.9*. SRAC, Los Angeles, USA, 2006.

[16] P. Polach, M. Hajžman, *Multibody model of the LKP-M/3 control assembly of the VVER 1000 nuclear reactor and calculation of time of the fall of the rod control cluster assembly at seismic event, research report VYZ 0755/2005*. ŠKODA VÝZKUM s.r.o., Plzeň, Czech Republic, 2005. (in Czech)

[17] M. K. Vukobratovic, V. Potkonjak, Dynamics of contact tasks in robotics. Part I: general model of robot interacting with environment. *Mechanism and Machine Theory*, **34**, 923-942, 1999.

[18] T. Klisch, Contact Mechanics in Multibody Systems. *Multibody System Dynamics*, **2**, 335-354, 1998.

[19] W. Schiehlen, R. Seifried, Three Approaches for Elastodynamic Contact in Multibody Systems. *Multibody System Dynamics*, **12**, 1-16, 2004.

[20] A. Kecskeméthy, Ch. Lange, G. Grabner, Analysis of Impact Responses Using the Regularized Model Approach. *EUROMECH Colloquium 397*, Grenoble, France, June 30 – July 2, 1999.

[21] P. Flores, J. C. Pimenta Claro, J. Ambrósio, H. M. Lankarani, C. Sony Koshy, Study of Contact-Impact Force Models in Multibody Mechanical Systems. *10th Conference on Nonlinear Vibrations, Stability and Dynamics of Structures*, Blacksburg, Virginia, USA, July 25-28, 2004.

[22] E. Feireisl, On the motion of rigid bodies in a viscous fluid. *Applications of Mathematics*, **47**, 463-484, 2002.

[23] R. V. Giles, *Fluid Mechanics and Hydraulics*. McGraw-Hill, 1962.

[24] R. Bulín, M. Hajžman, P. Polach, Nonlinear dynamic analysis of a single control rod in nuclear reactors. L. Pešek ed. *National Colloquium with International Participation Dynamics of Machines 2014*, Prague, Czech Republic, February 4-5, 2014.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# THE UPLIFT CAPACITY OF HORIZONTAL PLATE ANCHORS IN SPATIALLY VARIABLE CLAY USING SPARSE POLYNOMIAL CHAOS EXPANSIONS

**Tom Charlton**[1] **and Mohamed Rouainia**[2]

[1] School of Civil Engineering and Geosciences, Newcastle University
Newcastle upon Tyne, NE1 7RU
e-mail: t.s.charlton@newcastle.ac.uk

[2] School of Civil Engineering and Geosciences, Newcastle University
Newcastle upon Tyne, NE1 7RU
e-mail: mohamed.rouainia@newcastle.ac.uk

**Keywords:** Offshore Geotechnics, Random Field, Finite Element Analysis, Polynomial Chaos.

**Abstract.** *As the offshore energy industry moves towards deepwater installations, plate anchors are increasingly used to moor floating production facilities. The ultimate holding capacity of a plate anchor in undrained clay has been widely investigated in scenarios where the undrained shear strength is a deterministic parameter, uniform or linearly increasing across the soil mass. However, it has been shown that bearing capacity of footings can be overestimated if spatial variability is not taken into account. In this paper, a least angle regression-based sparse polynomial chaos expansion is used to efficiently study the uplift capacity of horizontal plate anchors in spatially variable clay represented by a high-dimensional random field. The coefficients of the expansion are obtained from a set of finite element analyses and a range of anchor embedment ratios are modelled to investigate both shallow and deep anchor behaviour. The limiting cases of an attached and vented anchor, where the anchor is either fixed to or separable from the soil, are also considered. It is found that the probability of failure of vented anchors reduces with embedment depth due to a decrease in the variability of anchor capacity as shear planes lengthen. In the attached case, the probability of failure is dependent upon whether the anchor fails by a shallow or deep mechanism.*

## 1 INTRODUCTION

The recent move towards deepwater energy production has led to an increased interest in the analysis and design of anchoring systems. Plate anchors are commonly deployed to moor floating production facilities [1]. The holding capacity of plate anchors in undrained clay has been widely investigated, using both physical modelling and numerical analysis [2-4]. In these studies, the undrained shear strength is uniform or linearly increasing according to a defined trend. In reality, natural clay is a highly variable material and the values of engineering parameters fluctuate across the soil mass [5]. Spatial variability, generally represented by random fields, has been shown to influence mechanical behaviour in a range of geotechnical scenarios [6]. In footing problems, bearing capacity can be overestimated if spatial variability is not taken into account. This has recently been demonstrated by Li *et al.* [7] for the case of buried footings, representative of applications such as spudcan foundations for offshore drilling platforms. The effect on uplift capacity is less well-studied but clearly has important design implications.

In this paper, a least angle regression (LAR)-based sparse PCE [8] is used to efficiently study the uplift capacity of horizontal plate anchors in spatially variable undrained clay represented by a random field. LAR enables automatic selection of only the most influential terms of the expansion, reducing the number of model evaluations required to ensure a well-posed least squares regression problem and a good approximation. A finite element model is used to obtain the uplift capacity across a range of embedment ratios and the limiting cases of an 'attached' and 'vented' anchor, where the anchor is either fixed to or separable from the soil, are considered. The use of the PCE enables an accurate assessment of the probability of failure in comparison with current offshore design practice, where the mean undrained shear strength value is applied in combination with a partial safety factor.

## 2 SPARSE POLYNOMIAL CHAOS EXPANSIONS

Polynomial chaos expansions are a method for quantifying uncertainty in complex numerical models with input parameters represented by random variables. The model output can be approximated by expanding the response quantity onto a basis of orthogonal multivariate polynomials. If the model is denoted $\Gamma$, and is a function of $M$ independent input random variables $\boldsymbol{\xi} = \{\xi_1, \dots, \xi_M\}^T$, this can be written as follows:

$$\Gamma(\boldsymbol{\xi}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \tag{1}$$

where $\psi_{\boldsymbol{\alpha}}$ is a multivariate polynomial basis and $a_{\boldsymbol{\alpha}}$ are deterministic coefficients which must be computed, for example by regression from a set of model evaluations. The sequence of non-negative integers $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_M\}$ is a multi-index representing the polynomial order associated with each random variable. The series is known as a polynomial chaos expansion and provides an analytical expression for the model response. For optimal convergence, the multivariate polynomial basis is chosen to be orthogonal with respect to the joint probability density function (PDF) of the input random variables. The multivariate, $M$-dimensional basis is constructed as a product of one-dimensional polynomials. In this paper, Gaussian input random variables are used with the corresponding Hermite polynomials.

The truncation, for computational purposes, of the infinite series in Eq. (1) was originally undertaken by keeping only those polynomials with degree less than or equal to the current PCE order $p$, i.e. $\sum_{i=1}^{M} \alpha_i \leq p$. The size of this basis is $P = \binom{M+p}{p}$ and the number of retained

coefficients, and so the computational cost, grows dramatically as either the number of random variables *M* or polynomial order *p* is increased.

In a *sparse* polynomial chaos expansion (SPCE), only significant terms are retained in the polynomial basis, thus reducing the number of coefficients that must be computed. Firstly, the candidate set of terms can be reduced prior to analysis to remove high-order interactions likely to be insignificant. Blatman and Sudret [8] proposed a 'hyperbolic' PCE in which a q-norm of the multi-indices should be smaller than the current order as follows:

$$\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M} (\alpha_i)^q\right)^{1/q} \leq p \tag{2}$$

where $0 < q \leq 1$. This is a stricter requirement than the classical truncation scheme, which is recovered by setting $q = 1$.

The PCE is finally obtained as:

$$\Gamma(\boldsymbol{\xi}) \cong \hat{\Gamma}_p(\boldsymbol{\xi}) = \sum_{\|\boldsymbol{\alpha}\|_q \leq p} \hat{a}_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) \tag{3}$$

However, even after the size of the truncation set has been reduced, not all remaining terms will be significant. An efficient solution is to use least angle regression (LAR) [9] to select the basis functions that have most effect on the model response. In LAR, the predictors are progressively activated based on their correlation with the set of model outputs until either all predictors are active or, if the number of model evaluations $n \leq P$, $n - 1$ predictors are active.

The LAR-SPCE method [8] does not actually use the coefficients computed by LAR but instead uses the predictors retained in each step along the LAR path in a least squares regression. Hence a series of SPCEs are produced and their approximation performance is assessed by a corrected leave-one-out cross-validation error, denoted $Q^{2*}$, in order to select the best expansion for subsequent use. Details of the error estimate $Q^{2*}$ can be found in [8]. To minimise the number of deterministic model evaluations an adaptive method is implemented in this study, with $q = 0.7$ and a maximum expansion order of 4.

## 3 PLATE ANCHOR UPLIFT CAPACITY

### 3.1 Description

Figure 1 shows the layout and notation of the plate anchor scenario analysed in this study. The dimensionless ratio H/B is used to describe the embedment depth at which the plate anchor is installed. Loads are applied perpendicular to the longitudinal axis of the anchor, with the ultimate pullout capacity denoted $Q_u (= q_u B)$. The pullout capacity in an undrained clay is generally expressed in terms of a dimensionless factor:

$$N_c = \frac{Q_u}{A s_u} \tag{4}$$

where A is the area of the plate and $s_u$ is the undrained shear strength. In the deterministic analysis the clay is weightless and uniform, with $s_u$ constant across the soil mass. If the anchor is embedded to such a depth that the failure mechanism becomes localised, the anchor can be described as 'deep'. In contrast, the failure mechanism of a shallow anchor will extend to ground level as the anchor is pulled out of the soil. It should be noted that this distinction is only relevant if soil and anchor are attached. Vented anchors only reach the ultimate capacity once the failure mechanism reaches the surface.

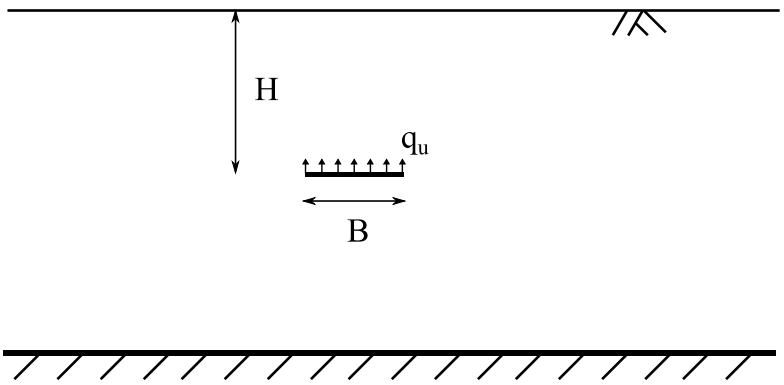



Figure 1. Layout and notation.

### 3.2 Representation of spatial variability

The spatial variability of undrained shear strength, $s_u$, is modelled as a lognormal random field. The mean ($\mu$) of $s_u$ is 10kPa and the coefficient of variation (COV) is taken to be 0.2, representing a typical variability based on results reported by Lacasse and Nadim [5]. In addition a constant rigidity index of $E/s_u$ = 500 is assumed, where E is the elastic modulus. In statistical terms, E is therefore perfectly correlated with $s_u$ and is generated from the same set of random variables.

An anisotropic square exponential autocorrelation function is adopted for $s_u$, with correlation distance 10m and 1m in the horizontal and vertical directions respectively. The lognormal random field can be generated as follows:

$$s_u(x,y) = \exp\left(\mu_{\mathrm{L,r}} + \sigma_{L,r} G(x,y)\right) \tag{5}$$

where $\mu_{L,s_u}$ is the mean of $ln(s_u)$, $\sigma_{L,s_u}$ is the standard deviation of $ln(s_u)$, and $G(x,y)$ is a correlated Gaussian random field of zero mean and unit variance. The expansion optimal linear estimation (EOLE) method [10] is used to discretise the random field $G(x,y)$ on a rectangular grid, henceforth referred to as the stochastic mesh to indicate its independence from the deterministic model. The expansion is truncated to include $M$ random variables, chosen such that at least 90% of the variance of $G(x,y)$ is captured.

### 3.3 Finite element model

The geotechnical FE software Plaxis 2D [11] is used as the deterministic model. The plate anchor, of width B = 2m, is modelled in plane strain and a range of embedment ratios are considered (H/B = 1, 2, 3, 6, and 10) in order to analyse both shallow and deep anchor behaviour.

Figure 2 shows a typical mesh, consisting of 15-node triangular elements, for an anchor embedded at H/B = 6. The anchor is modelled by a stiff plate element and the analysis is displacement-controlled. The clay is undrained and behaves according to the Mohr Coulomb model with friction angle φ = 0° and cohesion $c = s_u$. In the vented condition, an interface element is applied along the underside of the anchor with extensions at either end of 0.25B to avoid stress concentrations at the anchor tips. The section of interface adjacent to the anchor has no tensile strength so that separation of clay and plate occurs immediately when the anchor is displaced in the pullout direction. An interpolation procedure is used to transfer information from the stochastic mesh to the FE mesh. The deterministic capacity factors ($N_{c,det}$), using the mean value of $s_u$ in a uniform soil profile, showed an overestimation of less than 2% compared with those reported by Yu *et al*. [4].

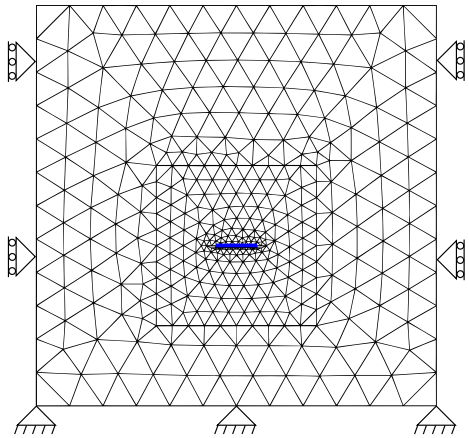


Figure 2. Typical FE mesh (H/B = 2).

## 4 RESULTS AND DISCUSSION

### 4.1 Performance of the LAR-SPCE method

Figure 3 shows the SPCE prediction of anchor capacity at H/B = 6 for the regression (or 'training') set and a test set of an additional 100 FE simulations not used in the regression. In this case, the random field is discretised using 50 standard Gaussian variables. To achieve the target accuracy, 1100 FE model evaluations were necessary in the attached condition, with 400 needed for a vented anchor. In both cases, a 3rd order expansion was found to be optimal, meaning the terms in the retained expansion have a maximum order of 3.

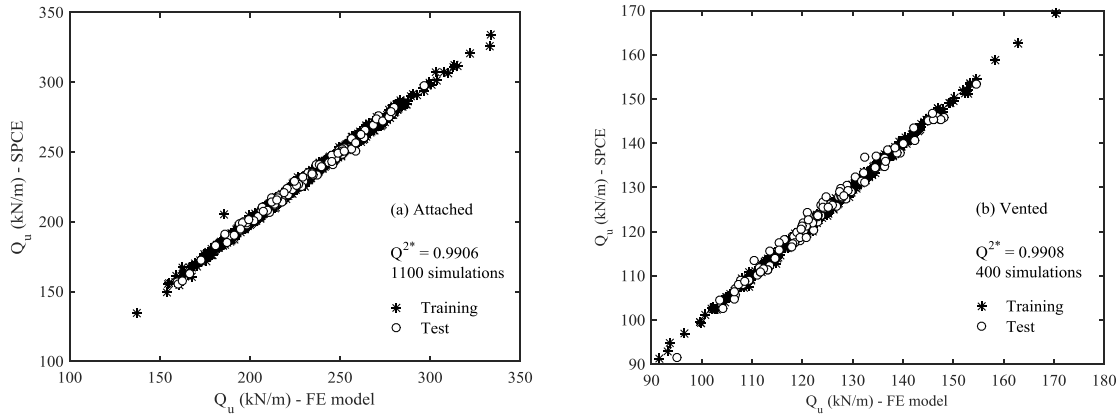



Figure 3. Retained SPCE approximation of FE model for a horizontal anchor with H/B = 6: training and test data for (a) attached and (b) vented anchor.

The strong linear relationship between the output of the FE and SPCE models demonstrates that the SPCE is able to produce an accurate approximation of the pullout capacity. This ensures that valid conclusions can be drawn about the probability distribution of the plate anchor capacity using an SPCE in place of the expensive FE model. The target accuracy of $Q^{2^*} = 0.99$ also provides acceptable performance for the number of simulations required; the rate of convergence tends to slow as the error reduces.

The number of FE model evaluations required to construct an SPCE of target accuracy is given in Table 1. In general, the computational effort is related to the number of variables necessary to discretise the field. However, in certain cases the nature of the failure mechanism

can slow convergence. For example, the target value of $Q^{2*}$ was not achieved for the attached anchor at H/B = 2. A very slow convergence was observed and the number of simulations was limited to 5000 for practical reasons. This slow convergence rate is a result of the anchor failing in a variety of different modes due to the spatial variability of the clay.

| H/B | Random variables | Interface | $Q^{2*}$ | FE simulations | Max. order |
|---|---|---|---|---|---|
| 1 | 50 | Attached | 0.9915 | 800 | 3 |
| | | Vented | 0.9908 | 400 | 3 |
| 2 | 60 | Attached | 0.9863 | 5000* | 4 |
| | | Vented | 0.9921 | 300 | 3 |
| 3 | 60 | Attached | 0.9908 | 1200 | 3 |
| | | Vented | 0.9910 | 300 | 3 |
| 6 | 50 | Attached | 0.9906 | 1100 | 3 |
| | | Vented | 0.9908 | 400 | 3 |
| 10 | 120 | Attached | 0.9910 | 3300 | 3 |
| | | Vented | 0.9903 | 800 | 3 |

Table 1. Details of the retained SPCEs (*indicates target $Q^{2*}$ was not reached).

In uniform (i.e. not spatially variable) clay, a deep mechanism forms with the shear plane localised around the anchor. In contrast, as shown in Figure 4, if the clay is spatially variable the failure mechanism can either be deep, being fully localised around the anchor, or shallow, involving a reverse end bearing mechanism, depending on the particular realisation of the random field. For this case the function $\Gamma(\boldsymbol{\xi})$ is not smooth, resulting in slow convergence of the SPCE. Note that 4$^{th}$ order terms were retained whereas in all other expansions only 3$^{rd}$ order polynomials were necessary. It is also clear that the critical embedment ratio, when the anchor transitions from a shallow to deep mechanism, is difficult to define exactly in spatially variable soil.

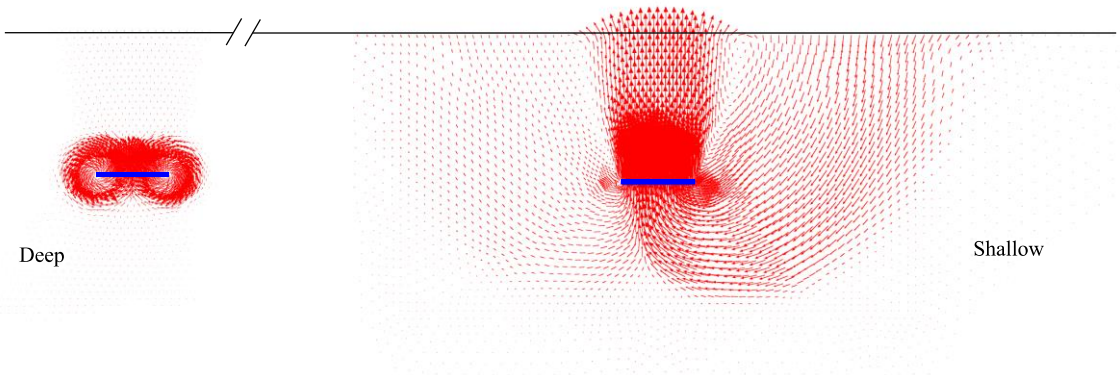



Figure 4. Failure mechanisms for an anchor embedded at H/B = 2 for two different random field realisations.

### 4.2 Statistics of the uplift capacity

The mean and standard deviation of the anchor capacity are obtained analytically from each SPCE. Figure 5 shows the mean, $\hat{\mu}_{Nc}$, and standard deviation, $\hat{\sigma}_{Nc}$, of the capacity factor in spatially variable clay across a range of embedment ratios. For comparison, the figure also shows the deterministic capacity factors. In both interface conditions, it can be seen that the mean capacity factor has a similar relationship with the embedment ratio as in the deterministic case and tends to be marginally (no more than 5%) lower than the equivalent deterministic capacity factor.

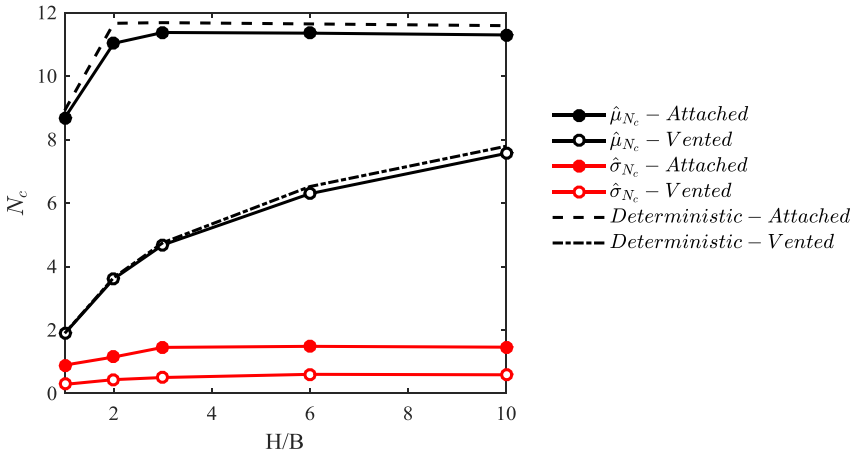



Figure 5. Mean, standard deviation, and deterministic capacity factors, $N_c$.

The variability of the pullout capacity for the different anchor configurations can be compared by considering the COV, as presented in Figure 6. For the vented case, the COV reduces with increasing embedment ratio. Li *et al*. [7] show that, for buried footings, a longer shear plane lowers the COV of the bearing capacity due to a greater spatial averaging effect. The same conclusion can be drawn from the failure mechanisms of the vented anchors. When anchor and soil are separable, the ultimate load is reached once the shear plane reaches the ground surface, regardless of anchor orientation. As H/B increases, the length of the shear plane is necessarily longer and the COV reduces. When the anchor and soil are attached, there is a distinct difference between the COV of 'shallow' and 'deep' anchors. If H/B is $\geq$ 3, the embedment ratio no longer affects the failure mechanism and COV is relatively constant.

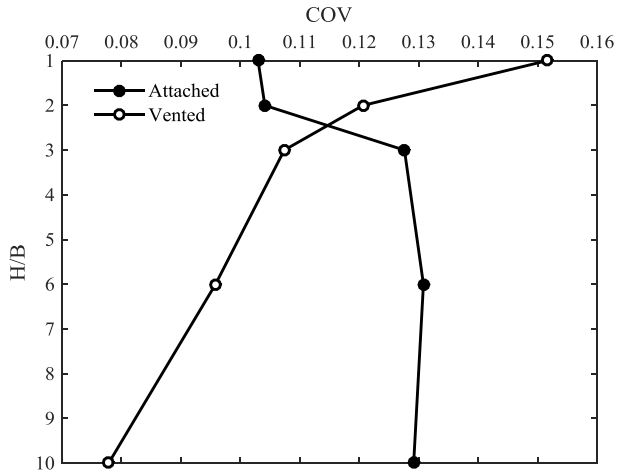



Figure 6. Coefficient of Variation (COV) against embedment ratio, H/B.

### 4.3 Probability of failure

The probability of failure can be defined as:

$$P_f = P\left(N_c < \frac{N_{c,det}}{FS}\right) \tag{6}$$

where FS is a factor of safety. This represents the probability that the capacity factor in spatially variable clay will be less than that predicted in a conventional numerical analysis with deterministic soil parameters.

Monte Carlo realisations of the SPCE are used to compute the probability of failure. To ensure a COV ≤ 0.1 for a probability of failure of $10^{-5}$, which tends to be the lowest value of $P_f$ considered in practice, the number of samples $(n_{SPCE})$ is $10^7$. The estimated probability of failure is then:

$$\hat{P}_f = \frac{1}{n_{SPCE}} \sum_{i=1}^{n_{SPCE}} I\left(N_c^{(i)} < \frac{N_{c,det}}{FS}\right) \tag{7}$$

where $I$ is the indicator function.

Figure 7 shows the probability of failure across a range of embedment depths as the factor of safety is increased from 1 to 3. If no factor of safety is used (FS = 1), the probability of failure ranges from 0.53 to 0.71 and is relatively independent of the interface condition. When a factor of safety is applied to $N_{c,det}$, the probability of failure can change greatly depending on the anchor configuration. If the anchor is vented, the probability of failure reduces as the anchor is embedded deeper into the clay. This is a direct result of the decreasing COV with depth observed for vented anchors. For the attached case, a distinction is again observed between deep and shallow anchors. As H/B increases, a constant probability of failure is reached as the failure mechanism is localised around the anchor.

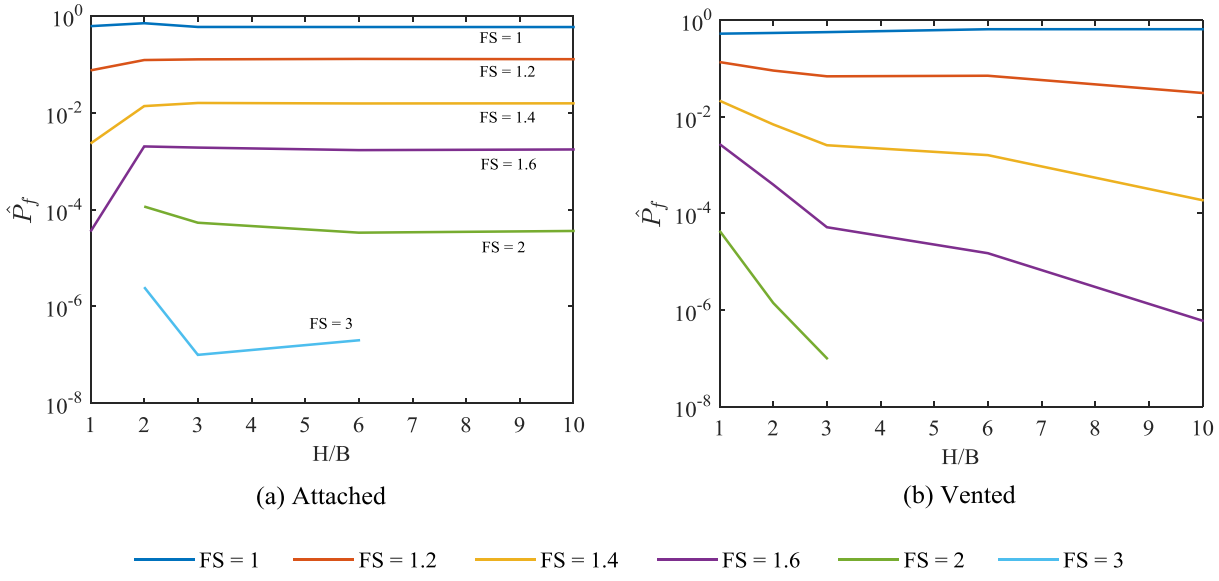



Figure 7. Probability of failure for (a) attached and (b) vented anchors when different factors of safety (FS) are applied to the deterministic capacity factor $(N_{c,det})$.

The recommended practice for the design of plate anchors DNV-RP-E302 [12] suggests a partial safety factor of 1.4 to account for uncertainty in "$s_u(z)$ *as it affects* $R_s(z)$" (where $z$ indicates depth and $R_s$ is the static resistance), as well as epistemic uncertainties resulting from, for example, the analytical model. The $N_c$ values recommended by the design code correspond to attached anchors, and the mean of $s_u$ is used as the characteristic value. The analysis presented here suggests that this factor may not be sufficient to account for natural spatial variability if the intended probability of failure is less than $10^{-2}$. However, further investigation into the effect of COV and the autocorrelation structure of $s_u$ is needed before concluding that the current design method is not conservative.

## 5 CONCLUSIONS

A study of the uplift capacity of a horizontal plate anchor in spatially variable clay has been conducted. An LAR-SPCE method was used to efficiently obtain the statistical moments of the anchor capacity and the probability of failure with respect to a conventional numerical analysis employing deterministic soil parameters. The probability of failure of vented anchors reduces with embedment depth due to a decrease in the variability of anchor capacity as shear planes lengthen. In the attached case, the probability of failure becomes relatively constant once the embedment ratio is large enough to ensure that the failure mechanism is localised around the anchor. The results of this study suggest that the partial factor used in current design practice may be not be sufficient to account for the spatial variability of undrained shear strength. Further research is necessary in order to better understand the effect of spatial variability on the ultimate capacity of plate anchors.

## REFERENCES

[1] M.F. Randolph and S. Gourvenec, *Offshore geotechnical engineering*. Spon Press, 2011.

[2] R.K. Rowe and E.H. Davis, The behaviour of anchor plates in clay. *Géotechnique*, **32**(1), 9-23, 1982.

[3] R.S. Merifield, S.W. Sloan, and H.S. Yu, Stability of plate anchors in undrained clay. *Géotechnique*, **51**(2), 141-153, 2001.

[4] L. Yu, et al., Numerical study on plate anchor stability in clay. *Géotechnique*, **61**(3), 235-246, 2011.

[5] S. Lacasse and F. Nadim, *Uncertainties in Characterising Soil Properties*, in *Uncertainty in the Geologic Environment: from Theory to Practice*, C.D. Shackelford, P.P. Nelson, and M.J.S. Roth, Editors, New York: ASCE. 49-75. 1996.

[6] G.A. Fenton and D.V. Griffiths, *Risk Assessment in Geotechnical Engineering*. John Wiley & Sons, 2008.

[7] J. Li, Y. Tian, and M. Cassidy, Failure Mechanism and Bearing Capacity of Footings Buried at Various Depths in Spatially Random Soil. *Journal of Geotechnical and Geoenvironmental Engineering*, **141**(2), 2015.

[8] G. Blatman and B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, **230**(6), 2345-2367, 2011.

[9] B. Efron, et al., Least angle regression. *Annals of Statistics*, **32**(2), 407-499, 2004.

[10] C. Li and A. Der Kiureghian, Optimal Discretization of Random Fields. *Journal of Engineering Mechanics*, **119**(6), 1136-1154, 1993.

[11] Plaxis, PLAXIS 2D 2012 - Reference Manual: Delft. 2012.

[12] DNV, *Design and Installation of Plate Anchors in Clay*, Det Norske Veritas (DNV): Norway. 2002.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# FATIGUE RELIABILITY OF AGEING RAILWAY BRIDGES: FEASIBILITY OF PROBABILISTIC APPROACH

**Nirosha D. Adasooriya**

Department of Structural, Mechanical Engineering and Material Science, University of Stavanger, Norway.
mudiyan.n.adasooriya@uis.no

**Keywords:** Railway bridge, Fatigue life, Loading sequence effect, Reliability index.

**Abstract.** *Miner's rule employed deterministic or probabilistic fatigue assessment approaches are generally used to predict remaining fatigue life of ageing railway bridges. Under many variable amplitude loading conditions, life predictions have been found to be unreliable since Miner's rule does not properly take account the loading sequence effect. Therefore, this paper presents a comparison of a new probabilistic fatigue assessment approach with deterministic approach consisting of a new damage indicator, which captures the loading sequence effect of variable amplitude loads more precisely than the Miner's rule. The comparison is performed by applying both fatigue assessment approaches to predict the remaining fatigue life of an ageing railway bridge. Initially the paper presents a probabilistic stress-life approach. Then the proposed approach is applied to predict the remaining fatigue lives of the ageing railway bridge. Finally, predicted fatigue lives are compared with fatigue lives predicted by the deterministic approach. Hence applicability, significance and validity of the proposed approach is discussed.*

## 1 INTRODUCTION

Majority of the railway bridges in the world are exceeding their design lives and bridge authorities are working on precious life extension methods [1-3]. As a result, a significant amount of research are ongoing for development of precious structural health monitoring and life assessment methods [2-12]. As railway bridges are vulnerable for time-dependent fatigue damage due to cyclic nature of traffic loads, the assessment of remaining fatigue life of railway bridges for continuing services has become more important than ever, especially when making decisions regarding structure replacement and other major retrofits. However, this task is difficult due to the increase of axel load and corrosion deterioration on bridges.

The fatigue assessment of structures mainly done by either deterministic or probabilistic approach. Most of the present day deterministic fatigue assessment approaches of railway bridges are generally based on the combination of measured stress histories under actual traffic load [12,13], Miner's rule [14] and railway code provided fatigue curve (also referred S-N or Wöhler curve). Although the mentioned deterministic approach predicts the remaining fatigue life, the uncertainties inherent in the fatigue evaluation process are not captured. These uncertainties are found in the process of determination of stress histories (i.e. structural analysis, field measurements, load testing, loading sequence and respective histories), selecting detail categories, choosing fatigue damage theories [15, 16].

The probabilistic fatigue assessments have been originated to capture the effect of these uncertainties more precisely. This approach is generally based on probability of fatigue failure associated reliability index. Fatigue reliability index provides a tool for predicting the remaining fatigue life [16]. A number of studies on the reliability analysis have been done for fatigue life prediction of bridges. Imam et al. [17] proposed a probabilistic fatigue assessment methodology for riveted railway bridges under historical and present-day train loading. Kwon and Frangopol [18] performed fatigue reliability assessment of steel bridges using the probability density function (PDF) of the equivalent stress ranges obtained by field measurement data. Ni et al. [19] proposed a fatigue reliability model for fatigue life and reliability evaluation of steel bridges with long-term monitoring data, which integrates the probability distribution of hot spot stress range with a continuous probabilistic formulation of Miner's damage cumulative rule. Recently, Kwon et al. [15] and Soliman et al. [16] proposed a probabilistic bilinear stress-life approach for better fatigue assessment of steel bridges. Miner's rule has been used as the fatigue damage theory for mentioned probabilistic models.

The Miner's rule is the simplest and the most commonly used fatigue life prediction technique. One of its interesting features is that life calculation is simple and reliable when the detailed loading history is unknown. However, under many variable amplitude loading conditions, life predictions have been found to be unreliable since it does not properly take account the loading sequence effect [20-22]. Therefore, it is uncertain to use the Miner's rule for remaining fatigue life estimation of railway bridges because most of the railway bridges are subjected to variable amplitude loadings. None of research studies confirmed about the consideration of the loading sequence effect on probabilistic fatigue assessment approaches.

To overcome this problem to some extent, objective of this paper is to compare probabilistic fatigue assessment approach with deterministic approach consisting of a new damage indicator (i.e. damage stress model) [22], which captures the loading sequence effect of variable amplitude loads more precisely than Miner's rule. The comparison is performed by applying both fatigue assessment approaches to predict the remaining fatigue life of an ageing railway bridge. This comparison provides an indication of feasibility of probabilistic stress-life fatigue approach for ageing railway bridges.

## 2 FATIGUE RELIABILITY INDEX

This section proposes a method to determine fatigue reliability index of bridges based on probabilistic bilinear *S-N* approach. The reliability index provides a measure of fatigue damage of the considered detail category of the bridge. Reliability index defines as the probability of violating fatigue limit state. The fatigue reliability index is defined as,

$$\beta = \emptyset^{-1}(1 - P_f) \tag{1}$$

where $\emptyset^{-1}$ is the inverse the standard normal distribution function. The corresponding fatigue limit state function can be derived as,

$$g(t) = \Delta - D \tag{2}$$

where $\Delta$ is Miner's critical damage accumulation index, which is assumed to be lognormal distribution with a mean value of 1.0 and coefficient of variation (COV) of 0.3 and $D$ is the Miner's damage accumulation index, which can be derived as,

$$D = \begin{cases} \frac{N(t)}{A_1}(S^{\,L}_{re})^{m_1} & for\ N(t) \le \frac{A_1}{CAFT^{m_1}} \\ \frac{N(t)}{(CAFT^{m_2-m_1}.A_1)}(S^{\,B}_{re})^{m_2} & for\ N(t) > \frac{A_1}{CAFT^{m_1}} \end{cases} \tag{3}$$

where $S^{\,L}_{re}$ and $S^{\,B}_{re}$ are equivalent constant amplitude stress ranges calculated using linear and bilinear *S-N* approach respectively as shown in Eq (4). The *CAFT* designated as constant amplitude fatigue threshold. The $m_1$ and $m_2$ are slopes of stress-life fatigue curve above and below the *CAFT*, respectively. The $A_1$ is the fatigue detail coefficient above the *CAFT* of the fatigue curve. The $A_2 = A_1 CAFT^{m_2-m_1}$, which is the fatigue detail coefficient bellow the *CAFT* of the fatigue curve. The *N(t)* is the number of cycles that considered detail category has subjected at the life time of *t*. The $m_1$, $m_2$, *CAFT* and *N(t)* are considered as the deterministic parameters. The stress range $S_{re}$, fatigue detail coefficient $A_1$ are considered as random variables.

As bridges are generally subjected to variable amplitude stress cycles, the equivalent constant amplitude stress range $S_{re}$ can be calculated for bilinear *S-N* approach as [16],

$$S_{re} = \left[\frac{\sum(n_i^o S_{ri}^{m_1}) + (CAFT^{m_1-m_2}).\sum(n_j^o S_{rj}^{m_2})}{\sum(n_i^o) + \sum(n_j^o)}\right]^{1/m_1} \tag{4}$$

where $n_i^o$ is the number of cycles in stress range bin $S_{ri}$ greater than the *CAFT* and $n_j^o$ is the number of cycles in stress range bin $S_{rj}$ less than the *CAFT*. The $\sum(n_i^o) + \sum(n_j^o)$ is the total number of cycles. Alternatively, the equivalent constant amplitude stress range $S_{re}$ can be calculated using the PDF of the stress ranges as follows [16],

$$S_{re} = \left[\int_0^{CAFT}(CAFT^{m_1-m_2}).S^{m_2}.f_s(s).ds + \int_{CAFT}^{\infty} S^{m_1}.f_s(s).ds\right]^{1/m_1} \tag{5}$$

The Eqs. (2) and (3) can be used to calculate fatigue reliability index ($\beta$) by using Monte Carlo simulation employed some of the softwares *R*, *RELSYS*, *CALREL* or etc. The fatigue reliability index ($\beta$) versus lifetime of bridge should be plotted and compared with target reliability index ($\beta_{target}$) to determine the fatigue life of each detail category.

## 3 CONSIDERED RAILWAY BRIDGE

The selected bridge is one of the longest railway bridges in Sri Lanka spanning 160m (Fig.1). It is a six span-riveted bridge with double lane rail tracks having warren type semi through trusses supported on cylindrical piers. The bridge deck is made of wrought iron and the piers are made of cast iron casings with infilled concrete. The bridge was constructed in 1885 and is located in marine environment. The bridge components have been categorized to several groups



Fig.1 General views of the considered bridge

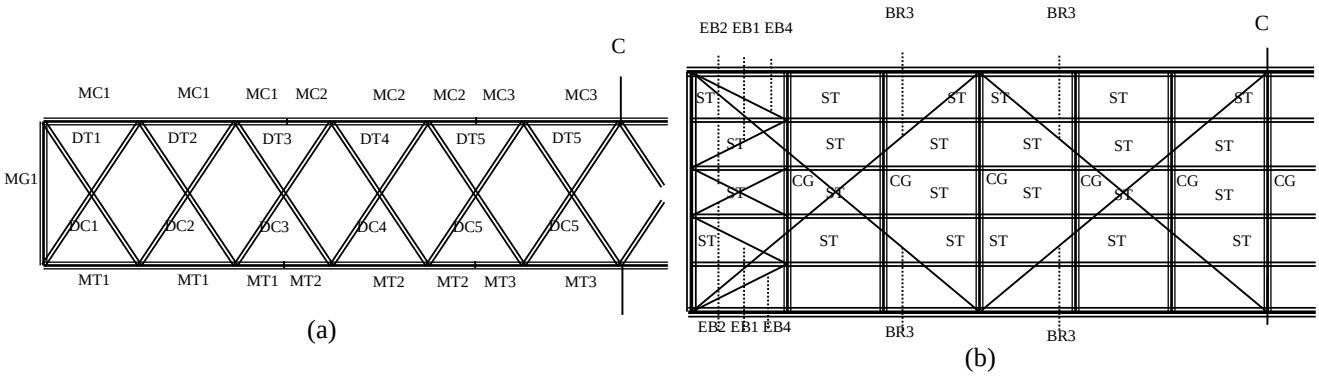



Fig. 2. Member set categorization: (a). Main truss girder, (b). Horizontal bridge deck [3]

entitled “member set” by considering similar cross sectional properties as shown in Fig. 2. Details of trains carried by the bridge and their frequencies illustrate that the bridge is subjected to variable amplitude loading [23].

The laboratory testing concluded that the bridge super structure material is wrought iron and the obtained values for elastic modulus, yield strength, ultimate strength in tension, fatigue strength and density are 195 GPa, 240 MPa, 383 MPa, 155 MPa and 7600 kg/m$^3$ respectively [3].

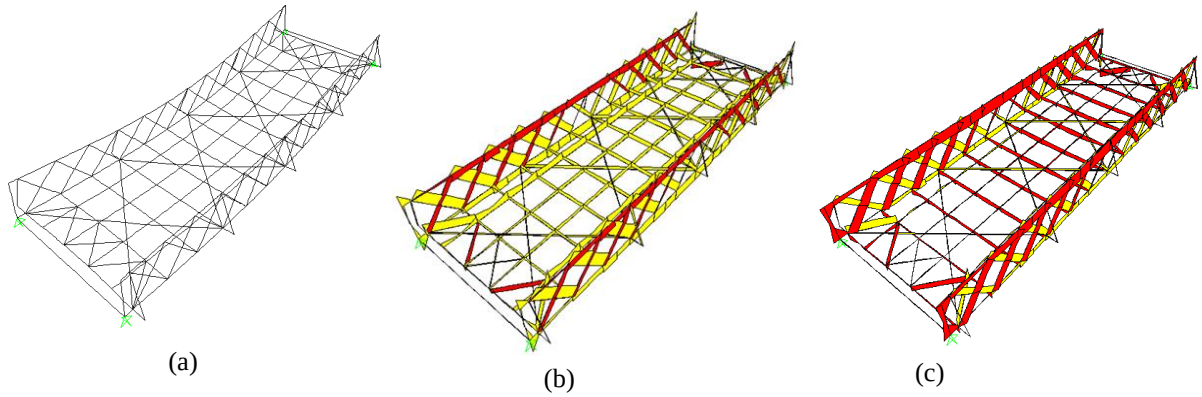



Fig. 3. The FE analysis results for moving train load: (a) Vertical displacement when the train is in the middle of the bridge (b) Maximum stress taken over all stress points at each cross sections when train is in the middle of the bridge (c) Minimum stress taken over all stress points at each cross sections when train is in the middle of the bridge cross sections when the train just before leave the bridge
Yellow color: Tensile stress- Red color: compressive stress

## 4 STRUCTURAL ANALYSIS

The bridge deck was analyzed using the general-purpose software SAP 2000. A three-dimensional (3D) model (Fig. 3) of one complete middle span of the bridge was analysed under actual loading to determine stresses in members and deflections, as well as variations of stresses under moving loads. The material properties recorded in section 3 and calculated section properties were utilized for this analysis. The bridge deck was modelled with 3D frame elements and the riveted connections were assumed to be fully-fixed [5]. Dynamic analysis was conducted for each different past and present passage of trains specified by the owner. The validation of FE model was done by comparing the results of time history dynamic analysis with those from measured time histories during the structural appraisal in year 2001 [3, 23]. These comparisons show that there are good agreement among analytical results of the FE model and the measurement of the actual bridge. Finally, the model is used to obtain past and present stress histories of each members due to each passage of trains.

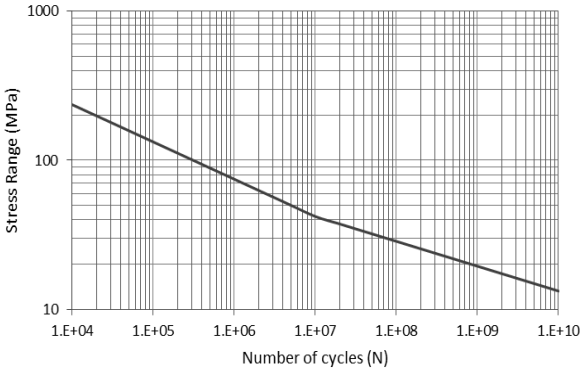



Fig. 4. S-N curves for WI-rivet detail

## 5 FATIGUE STRENGTH

The *S-N* curve of detail category (also referred to as detail class) are generally used with nominal stress histories to capture the fatigue damage due to the local stresses near the connection. The detail category is determined by considering the current condition of the connection.

Field investigations revealed that the riveted wrought iron connections of the bridge represent lapped or spliced connection behaviour with normal clamping force. Therefore, riveted connections were classified as WI-rivet (i.e. WI-rivet detail category or class), which is proposed by the UK railway assessment code [1,24]. The different mean and design *S-N* curves for WI-rivet detail class have been proposed by previous researchers based on results of experiments on full scale riveted members [1]. The above design *S-N* curve of WI-rivet detail (i.e. mean minus two standard deviations, which has 2.3% probability of failure), which is shown in Fig. 4, was used for fatigue reliability assessment of this bridge. The corresponding slopes of *S-N* fatigue curve $m_1$, $m_2$, fatigue detail coefficients $A_1$, $A_2$ and constant amplitude fatigue threshold *CAFT* are 4,6, $3.117\times10^{13}$, $5.489\times10^{16}$ and 42MPa respectively.

## 6 FATIGUE RELIABILITY ASSESSMENT

The stress ranges and average number of cycles per day at each members were calculated for each period using the rainflow counting algorithm. The stress range histograms for critical members and its probability density functions are plotted as shown Fig. 5. The Fig. 5 illustrates that the stress ranges of almost all critical members follow the log-normal distribution. Hence equivalent constant amplitude stress ranges ($S_{re}$) for each critical members of each member sets were calculated by Eq (5).

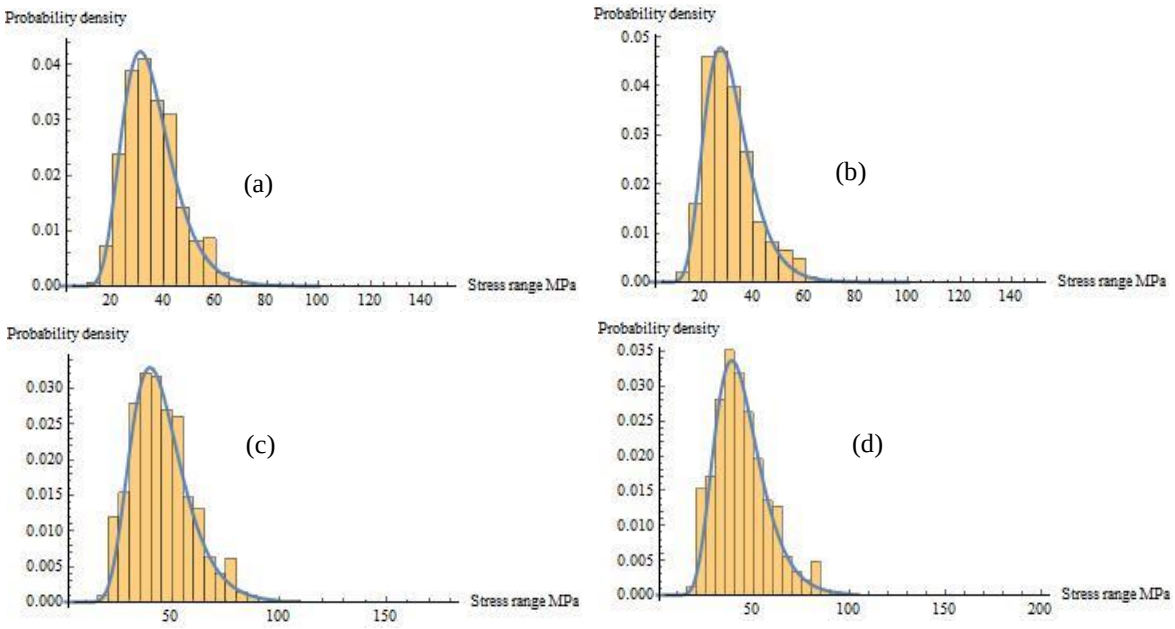



Fig. 5. Stress range histogram and its probability distribution function: (a) for critical member in cross girder set CG; (b) for critical member in stringer set ST; (c) for critical members in main girder set MT2; (d) for critical members in truss diagonal set DT3

The COV of $S_{re}$ is considered as 0.1 [15,25]. The parameter $A_1$ and $\Delta$ are random variables and corresponding COV's are 0.45 and 0.3 respectively as discussed in section 2 [18,25]. Other parameters such as $m_1$, $m_2$, *CAFT* and *N(t)* are considered as the deterministic parameters. As all the random variables follow the log normal distribution, based on Eqs. (2) and (3), fatigue reliability index, $\beta$ can be derived as follows:

$$\beta(t) = \begin{cases} \dfrac{\lambda_\Delta + \lambda_{A_1} - m_1.\lambda_{S_{re}^L} - \ln N(t)}{\sqrt{\zeta_\Delta^2 + \zeta_{A_1}^2 + (m_1.\zeta_{S_{re}^L})^2}} & for\ N(t) \le \dfrac{A_1}{CAFT^{m_1}} \\ \dfrac{\lambda_\Delta + \lambda_{A_2} - m_2.\lambda_{S_{re}^B} - \ln N(t)}{\sqrt{\zeta_\Delta^2 + \zeta_{A_2}^2 + (m_2.\zeta_{S_{re}^B})^2}} & for\ N(t) > \dfrac{A_1}{CAFT^{m_1}} \end{cases} \quad (6)$$

where λ and ζ are lognormal parameters of the various random variables.

The cumulative number of cycles *N(t)*, lognormal parameters of $S_{re}$ , $A_1$, $A_2$ and $\Delta$ are substituted to Eq. (6) and hence the fatigue reliability profiles (i.e. variation of fatigue reliability index with the age of the bridge) of the critical members of each member set of the bridge are generated and plotted in Fig.6. A target reliability index is defined to evaluate probability of limit state failure and corresponding fatigue life. Based on survival probability of 95%, target reliability indices was calculated as 1.65. The calculated fatigue lives are shown in Table 1.

The sequential law associated method [22], obtained nominal stress ranges in section 4 were used together to obtain remaining fatigue lives of critical members of each critical member sets of the bridge. The obtained fatigue lives of fatigue critical members of each member sets (i.e. which are possible to fatigue damage) are shown in Table 1. It is assumed that future sequence of passage is similar to that of the present day.

## 7 CONCLUSIONS

A probabilistic fatigue assessment approach and a deterministic approach consisting of a new damage indicator, which capture the loading sequence effect of variable amplitude loads more precisely than Miner's rule, were introduced to assess the fatigue life of an ageing railway

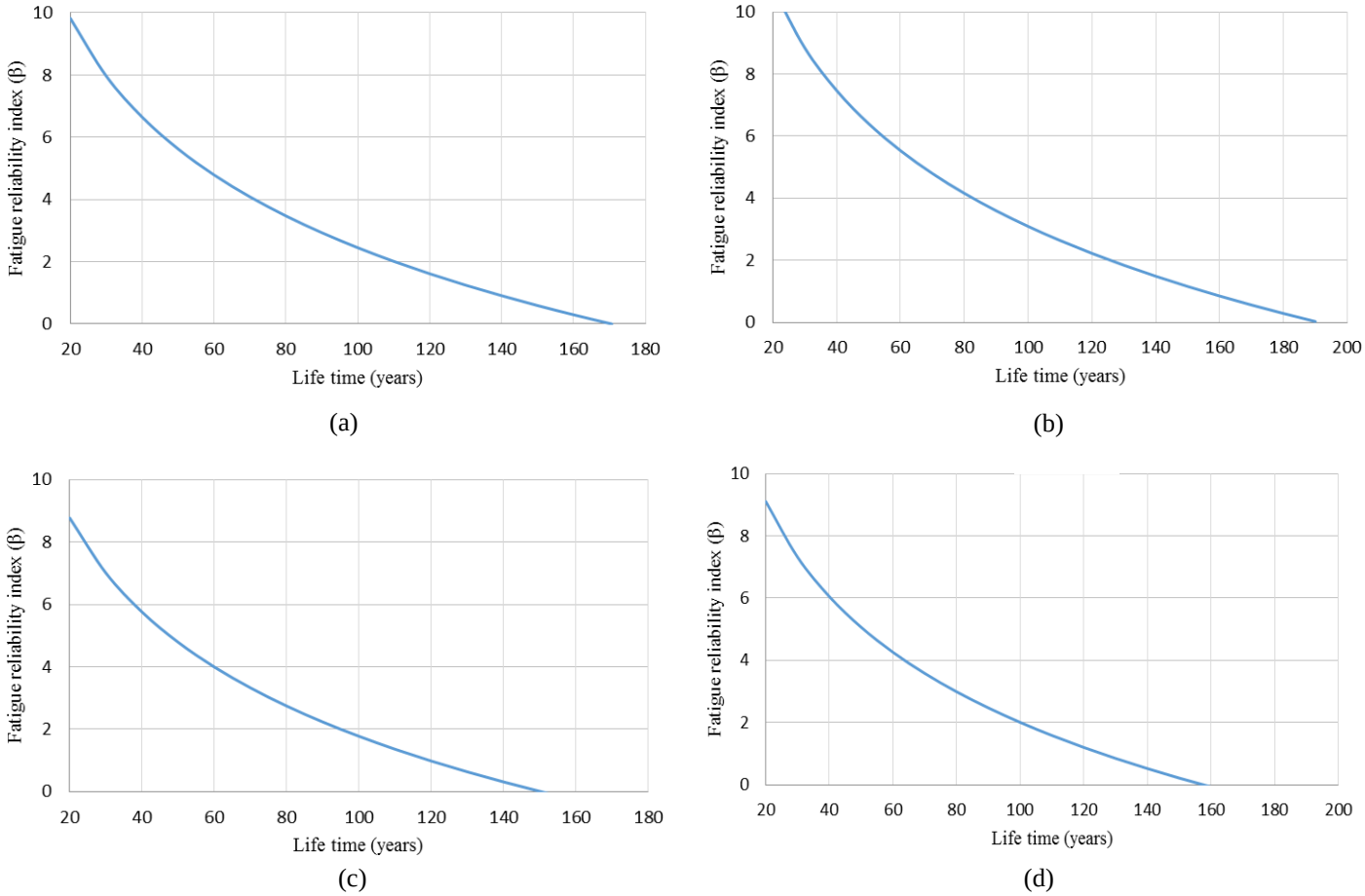



Fig. 6. Fatigue reliability index versus life of the bridge: (a) for critical member in cross girder set CG, (b) for critical member in stringer set ST, (c) for critical member in main girder set MT2, (d) for critical member in truss diagonal set DT3

bridge. Obtained fatigue lives were compared for critical members of each member sets as shown in Table 1. The Table shows that both deterministic and probabilistic approaches provide almost closer fatigue lives for bridge deck members (i.e. cross girders CG and stringers ST). However, it is opposite for the main girder truss members (i.e. main girder chords and truss diagonals).

Table shows that highly stressed member of main girder bottom chord MT2 is the most vulnerable to fatigue failure and the vulnerable members are located in the main girder consisting of truss members. Further, it seems that there are no more remaining lives for majority of vulnerable members of main girder truss under the 95% of survival probability. However, bridge is still in service without any recorded damage or fracture. The deterministic approach

Table 1. Summary of fatigue lives for critical members of each member sets

| Bridge component | Member set | Fatigue life (years) | |
|---|---|---|---|
| | | Deterministic Approach | Probabilistic Approach |
| | | Damage stress model | $\beta_{target}$=1.65 |
| Cross girders | CG | 133 | 119 |
| Stringers | ST | 134 | 135 |
| Main girder bottom chord | MT2 | 286 | 102 |
| Truss diagonal (tension member) | DT3 | 259 | 108 |

predicts most vulnerable member for fatigue failure as the critical member in cross girder member set CG. According to the deterministic approach, the remaining fatigue life of the considered bridge is three more years.

The deviations of fatigue lives of both approaches illustrate that introduced probabilistic fatigue assessment approach may not preciously capture the loading sequence effect. However, it can be concluded that application of introduced probabilistic model provides a conservative fatigue assessment for railway bridges. Therefore, it is doubtful to conclude that this introduced probabilistic model and corresponding modal parameters provides a precise remaining life for ageing railway bridges.

## ACKNOWLEDGMENT

The authors wish to express their sincere gratitude to Emeritus Professor M.P. Ranaweera and the team of experts who worked in the Railway Bridge project, for their great advice, which laid the foundation for this research. The kind support given by the Railway department is also appreciated.

## REFERENCES

[1] B.M. Imam, T.D. Righiniotis, Fatigue evaluation of riveted railway bridges through global and local analysis. J. Constr. Steel. Res. 66 (2011) 1411–1421.

[2] H. Rosemarie, K. Bertram, N. Alain, Assessment of existing steel structures: A guideline for estimation of the remaining fatigue life, Structure and Infrastructure Engineering.3 (2007) 245-255.

[3] N.D. Adasooriya, S.C. Siriwardane, Remaining Fatigue Life Estimation of Corroded Bridge Members, Fatigue & Fracture of Engineering Materials & Structures. 37 (2014) 603-622.

[4] B. Imam, T.D. Righiniotis, M.K. Chryssanthopoulos, Fatigue assessment of riveted railway bridges, Int. J. Steel Struct. 5 (2005) 485-494.

[5] Z.X. Li, T.H.T. Chan, J.M. Ko, Fatigue analysis and life prediction of bridges with structural health monitoring data- Part 1: methodology and strategy, Int. J Fatigue. 23 (2001) 45-53.

[6] Z.X. Li, T.H.T. Chan, J.M. Ko, Fatigue analysis and life prediction of bridges with structural health monitoring data- Part II: application, Int. J Fatigue. 23 (2001) 55-64.

[7] Z.X. Li, T.H.T. Chan, J.M. Ko, Fatigue damage model for bridge under traffic loading: application made to Tsing Ma Bridge, Theo. & App. Fracture Mechanics. 35 (2001) 81-91.

[8] Z.X. Li, T.H.T. Chan, J.M. Ko, Determination of effective stress range and its application on fatigue stress assessment of existing bridges, Int. J Solid & Struct. 39 (2002) 2401-2417.

[9] Z.X. Li, T.H.T. Chan, R. Zheng, Statistical analysis of online strain response and its application in fatigue assessment of a long span steel bridge, Eng. Struct. 25 (2003) 1731-1741.

[10] Z.X. Li, T.H.T. Chan, Fatigue criteria for integrity assessment of long span steel bridge health monitoring, Theo. & App. Fracture Mechanics. 46 (2006) 114-127.

[11] B. Sherif, M. Shuichi, O. Toshiyuki, Nondestructive damage detection scheme for steel bridges, J App. Mech. 9 (2006) 63-74.

[12] L.Köröndi A. Szittner, M. Kálló, L. Krisróf, Determination of fatigue safety and remaining fatigue life on a riveted railway bridge by measurement, J. Constr. Steel. Res. 46 (1998) 430.

[13] C.S. Constantine, G.R. Ioannis, Ch.E. John, Condition assessment and retrofit of historic steel-truss railway bridges, J. Constr. Steel. Res. 60 (2004) 1213-1225.

[14] M.A. Miner, Cumulative damage in fatigue, J App. Mech. 12 (1945) 159-164.

[15] K. Kwon, D. M. Frangopol, M. Soliman, Probabilistic fatigue life estimation of steel bridges by using a bilinear S-N approach, J Bridge Eng. 17 (2012) 58–70.

[16] M. Soliman, D. M. Frangopol, K. Kown, Fatigue assessment and service life prediction of existing steel bridges by integrating SHM into a probabilistic bilinear S-N approach, J Struc. Eng. 139 (2013) 1728– 1740.

[17] B. M. Imam, T. D. Righiniotis, M. K. Chryssanthopoulos, Probabilistic fatigue evaluation of riveted railway bridges, J Bridge Eng. 13 (2008) 237–244.

[18] K. Kwon, D. M. Frangopol, Bridge fatigue reliability assessment using probability density functions of equivalent stress range based on field monitoring data," Int. J Fat. 32 (2010) 1221–1232.

[19] Y. Q. Ni, X. W. Ye, J. M. Ko, Monitoring-based fatigue reliability assessment of steel bridges: analytical model and application, J Str. Eng, 136 (2010) 1563–1573.

[20] S. Suresh, Fatigue of materials, Cambridge University Press, UK, 1998.

[21] V. Dattoma, S. Giancane, R. Nobile, F.W. Panelle, Fatigue life prediction under variable loading based on a new non-linear continuum damage mechanics model, Int. J Fat. 28 (2006) 89-95.

[22] A. Aid, A. Amrouche, B.B Bouiadjra, G. Benguediab, G. Mesmacque, Fatigue life prediction under variable loading based on a new damage model, Mat. & Des. (2011) 183-191.

[23] S.C. Siriwardane, Vibration measurement-based simple technique for damage detection of truss bridges: A case study, Case Studies Eng. Failure Anys. 4 (2015) 50-58.

[24] Network Rail (2001). The Structural Assessment of Underbridges, RT/CE/C/025, Rail track Line Code of Practice, Rail track.

[25] B. Ayyub, I.A. Assakkaf, D.P. Khil, M.W. Sieve, Reliability-based design guidelines for fatigue of ship structures, J Nav. Eng. 114 (2002) 113-138.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# AN EFFICIENT AERODYNAMIC SHAPE OPTIMIZATION FRAMEWORK FOR ROBUST DESIGN OF AIRFOILS USING SURROGATE MODELS

**D. Maruyama, D. Liu, and S. Görtz**

DLR, Institute of Aerodynamics and Flow Technology
Lilienthalplatz 7, 38108 Braunschweig, Germany
{daigo.maruyama,dishi.liu,stefan.goertz}@dlr.de

**Keywords:** Robust Design Optimization, Geometrical Uncertainties, Surrogate Model, Aerodynamic Shape Optimization, Computational Fluid Dynamics (CFD).

**Abstract.** *This paper deals with developing an efficient Robust Design Optimization (RDO) framework. The goal is to obtain an aerodynamic shape that is less sensitive to small random geometry perturbations and to uncertain operational conditions. The initial shape is the RAE2822 airfoil which is parameterized with 10 design variables. The robust design formulation used is based on an expectation measure. The goal was to minimize the sum of the mean and standard deviation of the drag coefficient of the RAE 2822 airfoil for a given nominal lift coefficient. Here, we focus on improving the methods used for computing the statistics of the aerodynamic performance of the airfoil in every optimization cycle. A relatively small number of samples is evaluated with CFD and used to construct surrogate models based on Kriging and gradient-enhanced Kriging. The aerodynamic performance statistics, which are used to evaluate the robust objective function, are estimated by using quasi Monte Carlo (QMC) sampling with many samples evaluated on the surrogate models. A large number of geometrical uncertainties is parameterized by using a truncated Karhunen-Loève expansion, which enables a significant reduction of the dimensionality of the problem and thus of the surrogate models. By varying the number of samples used to build the surrogate model and by comparing the two types of surrogate modeling methods, it is confirmed that the robust objective function can be evaluated accurately with at most 30 CFD computations and corresponding adjoint computations.*

## 1 INTRODUCTION

Computational Fluid Dynamics (CFD) has become a sophisticated tool and can be used for complex aerodynamic analysis and design of aircraft. Many traditional shape optimization problems have been solved by high fidelity CFD and the optimized configurations feature the desired aerodynamic performance under perfect conditions. However, the performance may degrade due to environmental and operational uncertainties, or manufacturing tolerances. Recently, robust design of aircraft under uncertainties is gaining interest in aerodynamics. In general, the high-fidelity CFD models used in deterministic shape optimization are expensive in terms of computational cost if applied to robust design.

In this paper, we are concerned with robust design optimization (RDO) of the transonic RAE2822 airfoil, which is a test case in the FP7 European project UMRIDA [1]. The functional of interest is the drag coefficient ($C_d$). Random variations of the flight conditions such as angle of attack, $\alpha$, and Mach number, $M$, may occur during flight due to changing atmospheric conditions. The airfoil shape may also be subject to manufacturing tolerances and to degradation during operations. These factors are irreducible and considered as realistic uncertainties (aleatory uncertainties). On the other hand, numerical and modeling errors are an additional source of uncertainty that has an impact on the RDO results. These are reducible (epistemic) uncertainties. Figure 1 shows a simple depiction of these uncertainties in numerical flow simulations. Here, we focus on simulating the effect of irreducible uncertainties on the design, which are modeled using stochastic methods. We consider non-intrusive approaches, where we treat the flow solver as a black box and only use it to calculate the outputs that are required to evaluate the objective function. In the whole process, the flow solver settings, the mesh topology, and the geometry parameterization are carefully chosen and fixed. We consider the realistic uncertainties as input uncertainties with normal distributions when we evaluate the robustness of the design and we introduce our efficient RDO approach which combines a shape optimization framework with uncertainty quantification (UQ) methods.

There are two main issues that need to be considered when developing an RDO framework. One is how to evaluate stochastic values, such as the mean and standard deviation, accurately and efficiently. For example, in transonic flow, because of the strong non-linearity of the governing equations, the results from full Monte Carlo (MC) simulations and those derived from approximations by stochastic modeling could be different. The MC method is one of the most straightforward UQ methods, but comes at a much higher cost than other non-intrusive methods. The method of moments, polynomial chaos expansion and stochastic collocation methods are widely used non-intrusive approaches for evaluating stochastic values [2,3,4,5]. We propose a RDO framework based on low-discrepancy sampling methods combined with Kriging-based surrogate models to compute the stochastic values, which are used to evaluate the objective function to be optimized, accurately and efficiently. The other issue is how to deal with a large number of uncertain parameters. Here, a large number of geometrical uncertainties is parameterized by a Karhunen-Loève expansion to achieve a significant reduction of the number of uncertain inputs [6,7].

There are several measures of robustness [5] that can be used in RDO and reliability-based design optimization. In this paper, we limit ourselves to a robustness measure that is based on the sum of the mean and standard deviation of the drag coefficient as caused by the operational and geometrical uncertainties. Here we mainly discuss the accuracy of the robust objective function and how we can obtain the statistics efficiently.

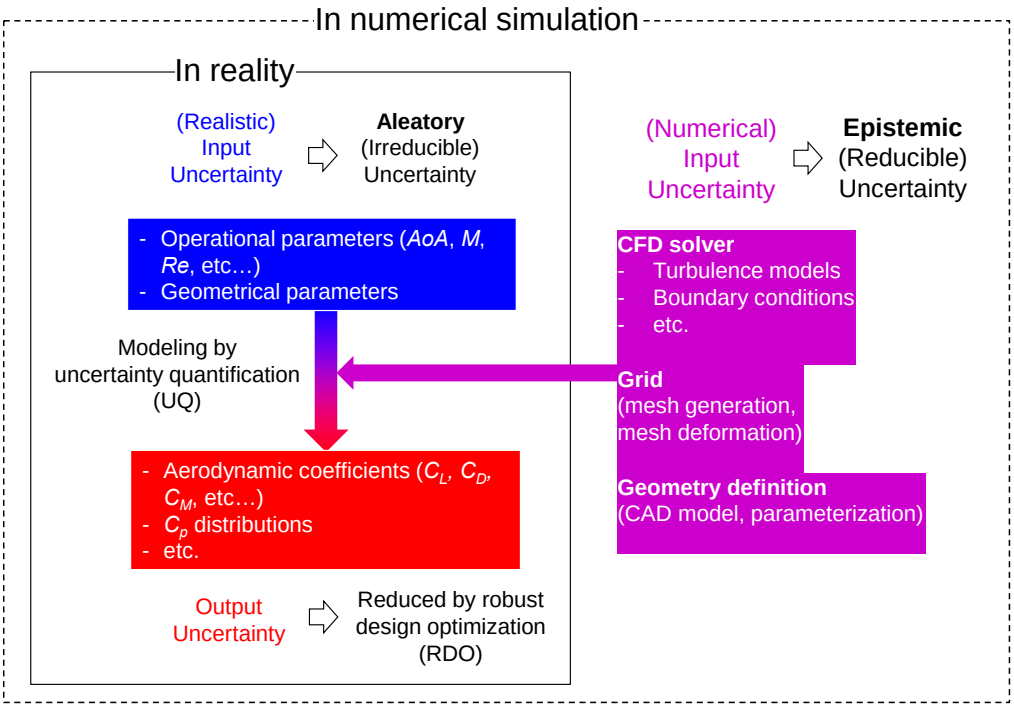



Figure 1: Overview of aleatory and epistemic uncertainties.

## 2 UNCERTAINTY QUANTIFICATION FOR ROBUSTNESS MODELING

### 2.1 Description of test case and problem setting

RDO is applied to the RAE2822 airfoil, which is one of the basic test cases in the UMRIDA project [1]. We consider the angle of attack, $\alpha$, the Mach number, $M$, and the geometry to be uncertain. The nominal condition of the angle of attack and Mach number are $\alpha$=2.79 deg. and M=0.734, respectively. The standard deviations of these operational parameters are $\alpha$=0.1 deg. and M=0.005 for the angle of attack and Mach number, respectively. These uncertainties are assumed to have a normal distribution. The Reynolds number is Re=6.5 million and is not considered uncertain. The operational uncertainties are therefore expressed as $\alpha \sim N(\mu_\alpha, \sigma_\alpha)$ and $M \sim N(\mu_M, \sigma_M)$. The geometrical uncertainties are described in section 2.2.

To evaluate the aerodynamic performance of the airfoil, DLR's flow solver TAU [8,9,10] was used. Fully turbulent computations were performed with the negative Spalart-Allmaras turbulence model [11]. A quasi two-dimensional hybrid unstructured grid with prisms and tetrahedral elements was used for the RANS simulations. Figure 2 shows a detailed view of the grid around the RAE2822 airfoil, which is used as the baseline configuration in this paper. The number of grid nodes is approximately 29,000. During the RDO the geometry changes of the airfoil are realized by using a radial basis function (RBF) based mesh deformation tool developed at DLR [12]. In this paper, all 380 surface nodes are considered uncertain. This yields the issue of how to treat a large number of geometrical uncertainties.

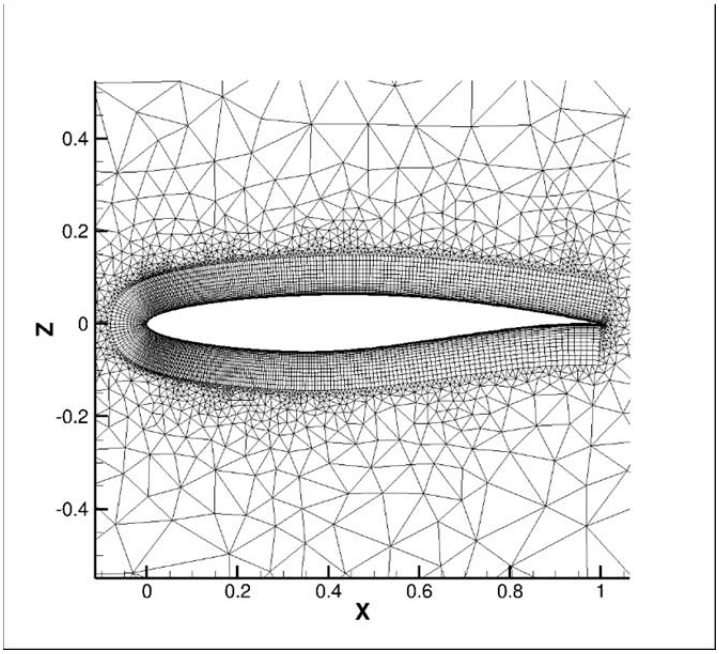



Figure 2: Detailed view of hybrid unstructured grid around RAE2822 airfoil.

## 2.2 Modeling of a large number of geometrical uncertainties

As briefly mentioned in the introduction, our approach to evaluate the cost function (the robust objective function) is based on a combination of design of experiment (DOE) methods and surrogate modeling. A reduction of the rather large number of geometrical uncertainties using a truncated Karhunen-Loève expansion (tKLE) [6,7] enables us to use much less samples in the DOE. We assume that random perturbations of the airfoil arise from manufacturing tolerances, for example, resulting in the surface to be smooth, i.e., the nodes on each surface are regarded to be correlated to each other.

The Karhunen-Loève expansion (KLE), which is also known as proper orthogonal decomposition (POD), is used to decompose correlated random fields into a linear combination of orthogonal functions, which are represented by uncorrelated random variables. In practice the orthogonal functions are carefully truncated to keep their number small. We assume a random perturbation $\theta(x)$ of each surface grid node in the direction normal to the airfoil surface on both the upper and lower surfaces with zero-mean normal distributions, i.e., $\theta(x) \sim N(0, \sigma_\theta(x))$. Here, the standard deviation of the perturbation, $\sigma_\theta(x)$, is represented by the distribution represented in Fig. 3(i). This setting means that perturbations are largest at the mid-chord and zero at the leading and trailing edges. Thus, the maximum perturbation of the upper and lower surfaces is 0.0025 in total in this case, which is around half of the UMRIDA test case definition. Here, it is assumed that the random perturbations are spatially correlated by a Gaussian correlation function. Then, the KLE can decompose these correlated variables into uncorrelated ones. Only a relatively small number of the first several significant modes is retained. Finally, the correlated random perturbation $\theta(x)$ is transformed into a smaller number of independent standard Gaussian variables as $\xi \sim N(0,1)$, which are the newly modeled geometrical uncertainties. In our test case the number of random geometrical variables is reduced from 380 to 10 using the KLE. More details of this process can be found in [6]. Three

realizations of the randomly perturbed geometry of the RAE2822 airfoil based on this parametrization are shown in Fig. 3(ii) as examples.

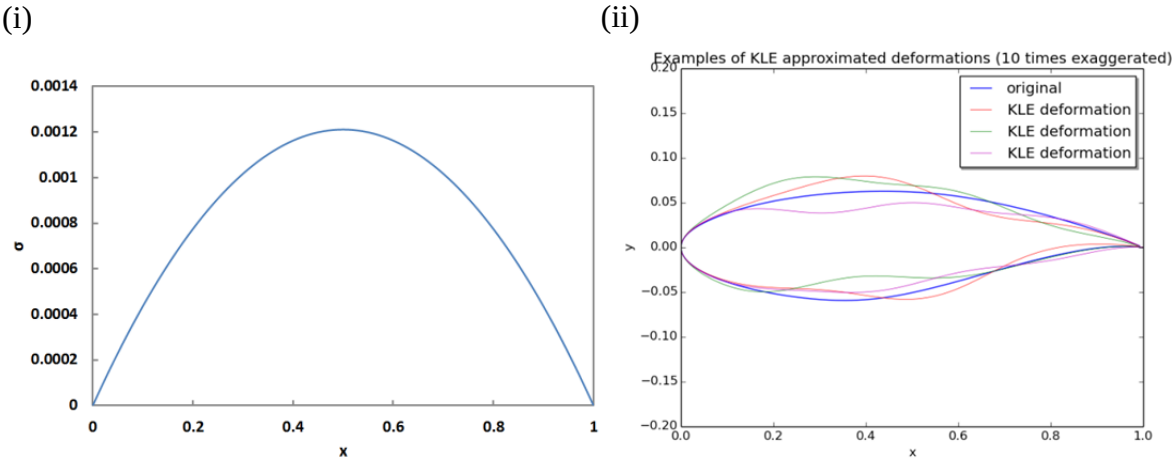



Figure 3: Distribution of the standard deviation of each surface node (left), and examples of the perturbed airfoil configurations of RAE2822 parameterized by truncated Karhunen-Loève expansion (tKLE) (right).

## 2.3 Design of experiments and surrogate models

The statistics, which are used to evaluate the robustness measure in every optimization cycle, are approximated using a combination of a DOE method and a surrogate model. The Sobol sequence based quasi Monte Carlo (QMC) [13,14,15] was used as a DOE method to generate samples of the above uncertain input parameters in the domain of $\mu \pm 3\sigma$. The samples are used to construct Kriging-based surrogate models of the functional of interest [2,3,16]. A further reduction of the computational cost is achieved by constructing gradient-assisted surrogate models, so called gradient-enhanced Kriging (GEK) [16], such that the gradient of the function of interest with respect to the uncertain input parameters are used as secondary DOE samples. Both Kriging and GEK are response surface methods used to statistically evaluate the function values belonging to an unsampled combination input parameters based on Bayesian inference. In both methods, maximum likelihood estimation (MLE) is adopted to optimize parameters of the correlation function. In this study, a Gaussian correlation function was used. Both Kriging and GEK as implemented in DLR's surrogate and reduced-order modeling toolbox SMARTy were used. The gradients with respect to the operational and geometrical parameters supplied to GEK were obtained by an adjoint method available in the DLR TAU-code [17]. The advantage of the adjoint method in this context is that all partial gradients of one cost function can be computed with a single adjoint solve.

Once a surrogate model is constructed, it can be used to evaluate the statistic of the functional of interest. The errors in the statistics computed with the help of the surrogate model can be made small by applying a MC method with a large number of samples such as $10^6$. For example, the mean and standard deviation can be evaluated with an error to the true value of around 0.01% by using the MC. Note that it is feasible to evaluate such a large number of samples using the surrogate model, while this is not feasible when using the CFD solver itself.

## 3 ROBUST DESIGN OPTIMIZATION FRAMEWORK

The UQ approaches described in Section 2 are applied to evaluate the robust objective function. Once the objective function value is determined, existing deterministic optimization algorithms are applicable. Figure 4 shows the framework of the deterministic optimization part. We use an in-house aerodynamic shape optimization toolbox called Pyranha [18] as the basis in this part. As in deterministic optimization, the objective functions, the design variables, and the constraints can be introduced. In this case the constraints are a constant thickness-chord ratio ($t/c$) and a constant lift coefficient ($C_l$). Therefore, the parameterization to define the nominal shape is applied to the camber line. In this paper, we used Bernstein polynomials. The objective or cost function, $f$, is a measure of the robustness of the design:

$$\left.\begin{aligned} &\min_{\boldsymbol{\chi}} \; f(\boldsymbol{\chi}) \\ &\text{s.t.} \quad t/c = const. \\ &\qquad C_l = const. \end{aligned}\right\} \tag{1}$$

where $\boldsymbol{\chi}$ denotes the deterministic design variables. The design variables are the output of the ROD. The computation of the cost function at each optimization cycle is described later and depicted in Fig. 5.

Compared to deterministic design optimization (DDO), an additional target-lift CFD calculation is required before the cost function can be evaluated to determine the mean angle of attack $\mu_\alpha$ that satisfies the nominal $C_l$ constraint. Here, $C_l$ is fixed to 0.791. The design variables, which are denoted by $\boldsymbol{\chi}$, are represented by 10 Bernstein polynomials that are applied to the camber line. The Subplex algorithm [18] was used in this paper to optimize the cost function $f$ according to Eq. (1). It is based on the Simplex and Nelder–Mead algorithms. Different from the Simplex algorithm, the Subplex searches for the convex hull in several decomposed low dimensional subspaces. Thus, the Subplex algorithm is generally more efficient than the Simplex for many design variables. Moreover, it is applicable to noisy cost functions [19].

Figure 5 shows more details of the stochastic modeling part concerned with computing the cost function. In this paper, the cost function $f$ is the sum the mean, $\mu$, and the standard deviation, $\sigma$, of the drag coefficient ($C_d$):

$$f(\boldsymbol{\chi}, \mathbf{a}, \boldsymbol{\xi}) \equiv \mu_{Cd}(\boldsymbol{\chi}, \mathbf{a}, \boldsymbol{\xi}) + \sigma_{Cd}(\boldsymbol{\chi}, \mathbf{a}, \boldsymbol{\xi}) \tag{2}$$

where $\boldsymbol{\chi}$, $\mathbf{a}$, and $\boldsymbol{\xi}$ are the design variables, the operational uncertainties, and the geometrical uncertainties given by the tKLE parameters, respectively. The number of parameters $\boldsymbol{\chi}$, $\mathbf{a}$, and $\boldsymbol{\xi}$ are 10, 2, and 10, respectively. Note that the design variables, $\boldsymbol{\chi}$, are deterministic and describe the nominal airfoil geometry at each iteration of the optimization process.

As described in Section 2, $\mu_{Cd}(\boldsymbol{\chi}, \mathbf{a}, \boldsymbol{\xi})$ and $\sigma_{Cd}(\boldsymbol{\chi}, \mathbf{a}, \boldsymbol{\xi})$ are calculated by constructing a surrogate model of $C_d$ based on a limited number of CFD computations. This is detailed in the flowchart in Fig. 5. At first, QMC sampling with a limited number in the range $\mu \pm 3\sigma$ for the uncertainty parameters $\mathbf{a}$ and $\boldsymbol{\xi}$ is performed. This covers 99.97% of the cumulative density function (cdf) of the normally distributed input uncertainties. Here, the uncertainty parameters are angle of attack, Mach number and 10 geometrical uncertainties given by 10 tKLE parameters. Therefore, the dimension is 12. In this research, the number of QMC samples was set to 12 or 30, respectively.

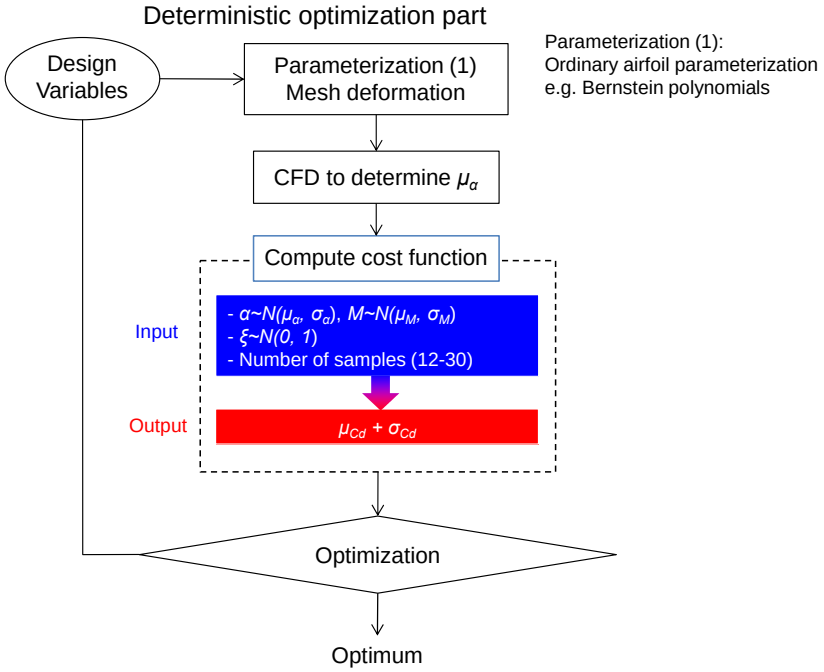



Figure 4: Overview of robust design optimization framework (deterministic optimization part).

Next, to construct a surrogate model of $C_d$, CFD calculations with the TAU-code are conducted on the samples. When a gradient-enhanced surrogate model is considered, additional adjoint calculations are run with the TAU-code to obtain gradients of $C_d$ with respect to the uncertainty parameters $\mathbf{a}$ and $\xi$. The computation of the samples values and their gradients is obviously the most time-consuming part of the entire process. Therefore, the number of DOE samples has the largest impact on the framework in terms of efficiency. Note, however, that each sample is independent of all other samples. Thus, all samples can be computed in parallel.

Finally, a surrogate model is constructed by using Kriging or GEK. Then the mean, $\mu$, and standard deviation, $\sigma$, of the drag coefficient are calculated based on the surrogate model by direct MC quadrature as:

$$\begin{cases} \mu_{Cd}(\boldsymbol{\chi}, \mathbf{a}, \xi) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(C_d(\boldsymbol{\chi}, \mathbf{a}, \xi)\right) \\ \sigma_{C_d}(\boldsymbol{\chi}, \mathbf{a}, \xi) = \sqrt{\dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}\left(C_d(\boldsymbol{\chi}, \mathbf{a}, \xi) - \mu_{C_d}(\boldsymbol{\chi}, \mathbf{a}, \xi)\right)^2} \end{cases} \tag{3}$$

where $N$ is a sufficiently large number of samples ($N$=$10^6$ in this paper). As described in Section 2.3, the error to the true value on a surrogate model is negligible when the number of samples is large. Therefore the quality of the surrogate models is the most important factor in terms of accuracy of the statistics.

In the next section, both Kriging and GEK are investigated in this respect by varying the number of samples used to construct the surrogate models.

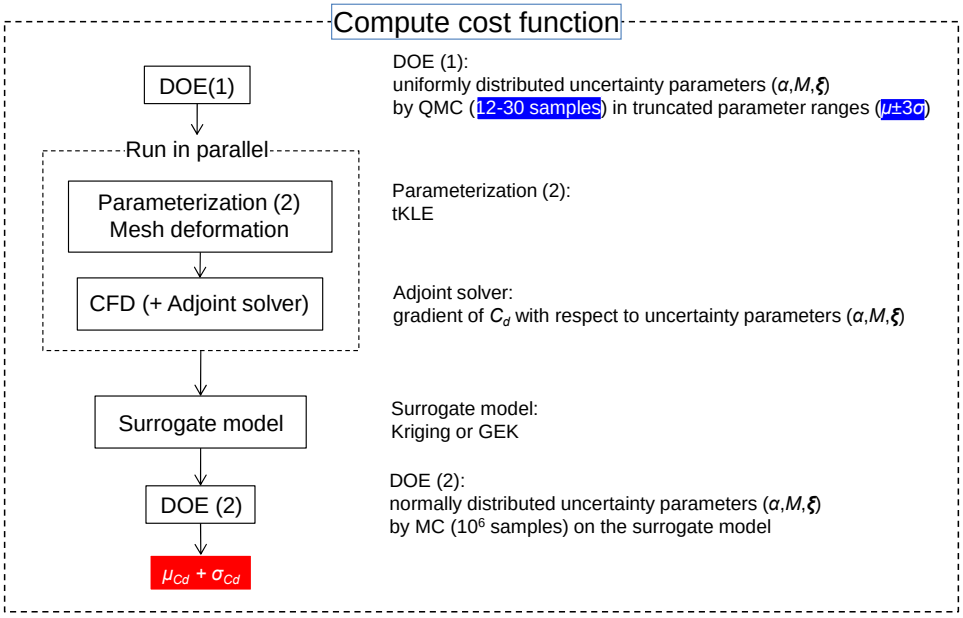



Figure 5: Overview of robust design optimization framework (evaluation of stochastic cost function).

# 4 RESULTS AND DISCUSSION

## 4.1 Accuracy of cost function

In this section, the cost function $f$, which is defined by Eq. (2), calculated for different surrogate models is compared with the true cost function $\hat{f}$ for the baseline airfoil configuration The true cost function $\hat{f}$, which acts as a reference, was calculated according to Eq. (3) by the direct QMC quadrature based on $10^5$ samples computed with TAU for RAE2822 as the given nominal airfoil shape. Note that in general in RDO, the true cost function $\hat{f}$ for a designed nominal airfoil shape at each optimization cycle is unknown.

As mentioned in the previous sections, the accuracy of the statistics at every optimization cycle depends on that of the surrogate models. As mentioned in Section 2.3, the cost function $f$ is almost uniquely determined once a surrogate model is constructed by Eq. (3). As for the construction of a surrogate model, once the sample points are fixed, the Kriging-based surrogate model is uniquely determined by its Gaussian correlation function and the likelihood function. When constructing a Kriging model, model parameters, so-called hyper-parameters, need to be tuned based on the sampled data using MLE, which is obtained in this study by the differential evolution as a global optimization method.

In our RDO framework, a surrogate model is constructed at every optimization cycle to obtain the cost function $f$. The surrogate model needs to be accurately constructed many times to obtain the cost function $f$, which is unknown but should be robustly evaluated to sample points $\hat{f}$ for a better guide to the next steps in whole the optimization cycles.

Therefore as a first step of estimation of the cost function $f$, we focus here on the initial guess, that is, the distribution of sample points. In practice, different 100 sets of distributions of sample points are derived from the Sobol sequence. Then, the dependency of accuracy of $f$ (in surrogate models) on the sample points are investigated by using two kinds of the num-

ber of sample points: 12 and 30, and by using two types of surrogate models: Kriging and GEK. Therefore, we investigated the following four approaches: (a) 12 samples with Kriging, (b) 30 samples with Kriging, (c) 12 samples with GEK, (d) 30 samples with GEK.

As defined in the previous section, the number of uncertainties is fixed to 12. Therefore, 12×12 or 30×12 sequences are generated to construct one surrogate model. These different sets of sequences are extracted 100 times from the Sobol sequences. Then 100 sets of differently distributed sample points are generated. The mean $\mu_f$ and standard deviation $\sigma_f$ of these 100 cost functions $f_1 \sim f_{100}$ in the 100 surrogate models are calculated to investigate how accurate the approach to construct surrogate models is by comparison with the true cost function $\hat{f}$.

Figure 6(i) shows distributions of the 100 predicted cost functions $f_1 \sim f_{100}$ obtained by 100 different sets of sample points by the four approaches (a)~(b). The stochastic values of these cost functions $f_1 \sim f_{100}$ are then calculated. The standard deviation $\sigma_f$ and mean $\mu_f$ of $f_1 \sim f_{100}$ is summarized in Fig. 6(ii). The mean values $\mu_f$ are depicted by adjusting to the relative error to the true cost function $\hat{f}$ as $\left\| \left( \mu_f - \hat{f} \right) / \hat{f} \right\|$. In this figure a correlation can be observed between $\sigma_f$ and $\left\| \left( \mu_f - \hat{f} \right) / \hat{f} \right\|$, those of which express the accuracy in terms of the distribution of sample points. The result using the strategy (d) 30 samples with GEK, can provide the mean of cost function $\mu_f$ with only 0.0003% relative error to the true value $\hat{f}$ and with 0.9 count (here one count=$10^{-4}$) for the standard deviation $\sigma_f$ as the error bound. Further improvement of efficiency by using less sample points can be expected by applying adaptive sampling strategies, which will be the next step of estimation of the cost function $f$.

Finally, the averaged computational costs to construct a surrogate model on one core on a cluster are summarized in Table 1. They are proportional to between the square and the cube of the samples points for both Kriging and GEK. For instance, the GEK using 30 samples costs around 10 times more than that using 12 samples. However, they are all quite small compared with one general CFD computational time. In this study, we fixed the optimizer for the MLE to a global one. We note that it was pointed out in [20] that global optimizers can provide a more accurate surrogate model in terms of not only the MLE but also the sample points.

### 4.2 Application to robust design optimization (RDO)

The developed RDO framework was applied to the transonic RAE2822 airfoil, which is one of the URMIDA test cases, by taking into account an uncertain angle of attack and Mach number, and a large number of geometrical uncertainties. Based on the results in the previous section, the GEK surrogate model used to evaluate the robust objective function at each optimization cycle is constructed by computing 30 samples and the corresponding partial gradients with TAU and its adjoint flow solver. All the other settings are described in Section 3. Figure 7 shows the history of the cost function $f$ defined in Eq. (2) and the nominal design variables $\chi$. We can observe that there is a relatively large amplitude of oscillations of the cost function $f$ after 350 optimization cycles, while the design variables $\chi$ do not vary any more. However, this amplitude is the smallest in the four types of surrogate models mentioned

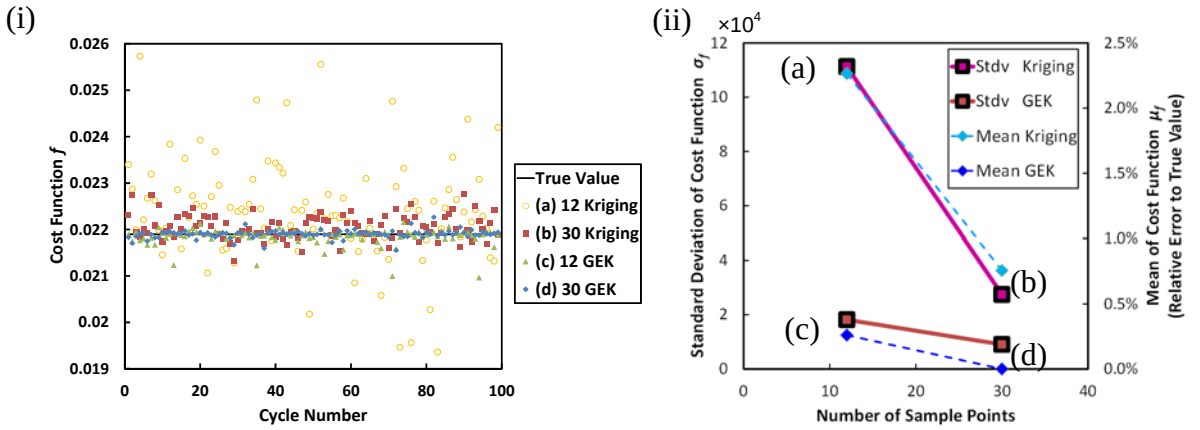



Figure 6: Dependency of accuracy of cost function on number of sample points and surrogate methods used to construct a surrogate model (left), and comparison of mean and standard deviation of the cost function based on (a) 12 samples with Kriging, (b) 30 samples with Kriging, (c) 12 samples with GEK, (d) 30 samples with GEK (right).

| | 12 samples | 30 samples |
|---|---|---|
| Kriging | (a) 0.4 sec. | (b) 0.4 sec. |
| GEK | (c) 1.7 sec. | (d) 14.6 sec. |

Table 1: Averaged computational costs for construction of a surrogate model on one core on a cluster.

in the previous sub-section. Figure 8 depicts a comparison of configurations of the initial airfoil (RAE2822) the robust airfoil. A designed airfoil configuration by the deterministic design optimization (DDO) is also shown in this figure. Table 2 shows $\mu_{Cd}$ and $\sigma_{Cd}$ of the initial, robust, and the deterministic designed airfoils. These stochastic values $\mu_{Cd}$ and $\sigma_{Cd}$ were re-evaluated by using the same strategy as that in Section 4.1. They are calculated by taking the average of 100 stochastic cost functions $f$ predicted by 30 sample points and the GEK surrogate model. It can be observed that both $\mu_{Cd}$ and $\sigma_{Cd}$ were successfully decreased. The total iteration number of the Supblex-based RDO for 10 design variables was 564. All computations were performed on a single core apart from the CFD computations which were ran in parallel on 24 cores. In terms of computational time, the total wall clock time was about 160 hours. The efficiency of the total RDO optimization could be improved by a two-step optimization strategy. For instance, at first deterministic design optimization (DDO) is conducted to the initial configuration and then a local optimizer such as gradient-based optimization algorithm is applied to the deterministic designed airfoil for RDO.

In this paper, we focused on accuracy and efficiency of construction of surrogate models in terms of stochastic values such as mean and standard deviation. The developed framework can also be applied for investigating the Pareto front in this robust optimization setting. Moreover, this surrogate-assisted cost function evaluation can also be applied to reliability-based design optimization by considering the probability density function of $C_d$.

As can be confirmed by Fig. 6 in the previous sub-section, there is a correlation between the amplitude of the oscillations of the cost function and the accuracy of the surrogate models.

The amplitude of the oscillations can be further reduced by improving the accuracy of the underlying surrogate models, e.g., by employing adaptive sampling strategies.

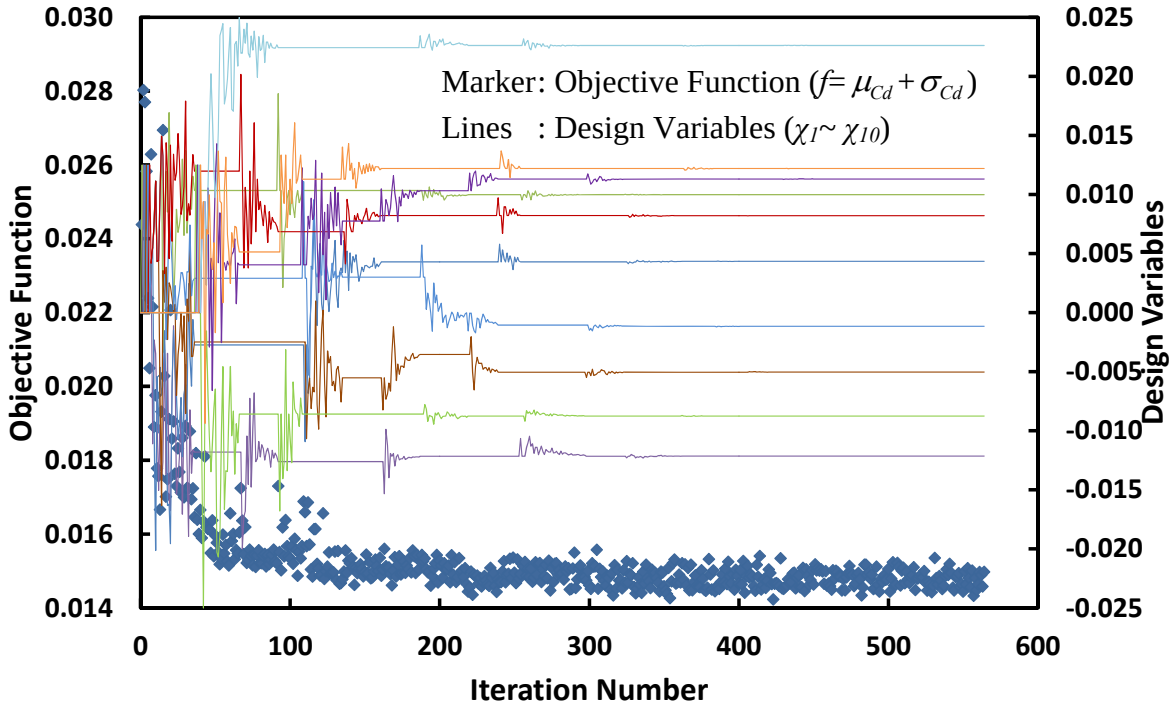



Figure 7: History of objective function and design variables in robust design optimization (RDO).

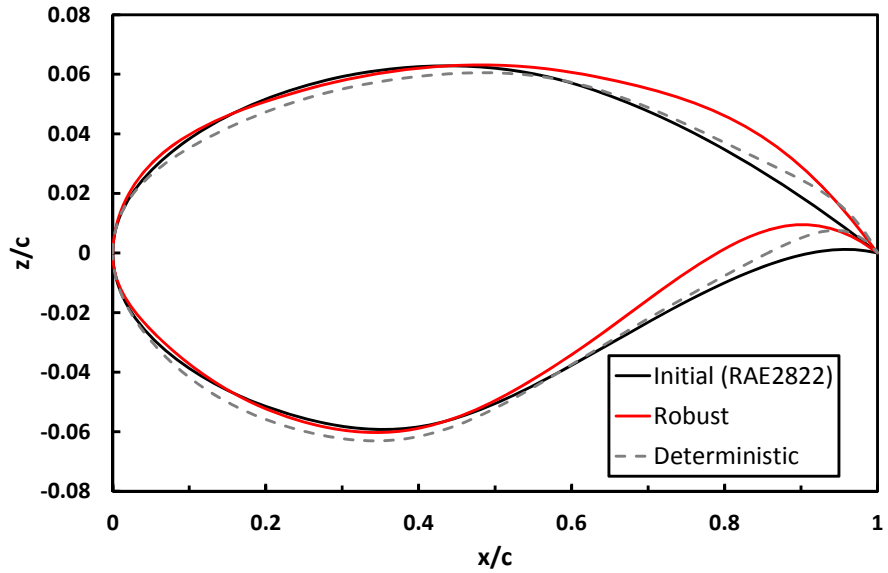



Figure 8: Comparison of initial (RAE2822), robust and deterministic designed airfoil geometries.

| | Initial | Robust | Deterministic |
|---|---|---|---|
| $\mu_{Cd}$ | 192.4 cts. | 137.9 cts. | 134.2 cts. |
| $\sigma_{Cd}$ | 26.6 cts. | 8.8 cts. | 12.6 cts. |

Table 2: Mean and standard deviation of $C_d$ of initial, robust, and deterministic designed airfoils (1cts.=$10^{-4}$).

## 5 CONCLUSIONS

A robust optimization framework for the aerodynamic design of airfoils was developed. To efficiently calculate the stochastic values which are used to evaluate the robustness measure in every optimization cycle, we employ design of experiments methods and Kriging-based surrogate models. A large number of geometrical uncertainties was modeled using a Karhunen-Loève expansion. The stochastic output values due to two operational uncertainties and 10 geometrical uncertainties were evaluated quite accurately based on at most 30 CFD computations and corresponding adjoint computations, the latter providing gradient information with respect to the 12 uncertainty parameters. These stochastic methods were integrated into an in-house framework for deterministic shape optimization. The developed framework was used to perform robust optimization of a transonic RAE2822 airfoil, which is one of the test cases in the EU project UMRIDA. The stochastic values such as the mean and standard deviation of the drag coefficient were successfully minimized and a new, more robust nominal configuration was proposed by this efficient robust design optimization framework.

## ACKNOWLEDGEMENTS

Part of the research leading to this work was supported by the UMRIDA (Uncertainty Management for Robust Industrial Design in Aeronautics) project, funded by the European Commission under grant number ACP3-GA-2013-605036, and by the Digital-X project at DLR. The authors would like to thank their colleagues Niklas Karcher, Thomas Franz, and Dr. Matteo Ripepi for discussions and help with DLR's SMARTy toolbox and Caslav Ilic and Mohammad Abu-Zurayk for help with DLR's Pyranha toolbox.

## REFERENCES

[1] UMRIDA official website: http://www.umrida.eu.

[2] R. Duvigneau, Aerodynamic Shape Optimization with Uncertain Operating Conditions using Metamodels. Research Report 6143, INRIA, March, 2007.

[3] M. Martinelli, R. Duvigneau, On the use of second-order derivatives and metamodel-based Monte-Carlo for uncertainty estimation in aerodynamics. *Computers & Fluids*, Volume 39, Issue 6, 953-964, 2010.

[4] M. Dodson, T.P. Geoffrey, Robust Aerodynamic Design Optimization Using Polynomial Chaos. *Journal of Aircraft*, Vol. 46, No. 2, 635-646, 2009.

[5] C. Schillings, V. Schulz, On the influence of robustness measures on shape optimization with stochastic uncertainties. *Optimization and Engineering*, Volume 16, Issue 2, 347-386, 2014.

[6] D. Liu, Efficient Quantification of Aerodynamic Uncertainties Using Gradient-Employing Surrogate methods. *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, Volume 122, Management and Minimisation of Uncertainties and Errors in Numerical Aerodynamics, 283-296, 2013.

[7] D. Liu, S. Görtz, Influence of reduced-order modelling of geometrical uncertainties on statistics. 7. Dresdner-Probabilistik-Workshop, Dresden, October, 2014.

[8] M. Galle, T. Gerhold, J. Evans, Parallel computation of turbulent flows around complex geometries on hybrid grids with the DLR-TAU code, in: A. Ecer, D.R. Emerson (Eds.), Proc. 11th Parallel CFD Conf., Williamsburg, VA, North-Holland, May 23–26, 1999.

[9] T. Gerhold, V. Hannemann, D. Schwamborn, On the validation of the DLR-TAU code, in: W. Nitsche, H.-J. Heinemann, R. Hilbig (Eds.), New Results in Numerical and Experimental Fluid Mechanics, in: Notes on Numerical Fluid Mechanics, vol. 72, Vieweg, ISBN 3-528-03122-0, 1999, pp. 426–433.

[10] D. Schwamborn, T. Gerhold, R. Heinrich, The DLR TAU-code: Recent applications in research and industry, invited lecture, in: P. Wesseling, E. Oate, J. Priaux (Eds.), Proceedings of the European Conference on Computational Fluid Dynamics (ECCOMAS CFD 2006), The Netherlands, 2006.

[11] S.R. Allmaras, F.T. Johnson, P.R. Spalart, Modifications and Clarifications for the Implementation of the Spalart-Allmaras Turbulence Model. Seventh International Conference on Computational Fluid Dynamics (ICCFD7), ICCFD7-1902, Hawaii, July, 2012.

[12] R. Heinrich, K. Reimer, A. Michler, Multidisciplinary simulation of maneuvering aircraft interacting with atmospheric effects using the DLR TAU code. RTO AVT-189 Specialists' Meeting on Assessment of Stability and Control Prediction Methods for Air and Sea Vehicles, Portsdown West, October 12-14, 2011.

[13] I.M. Sobol, Distribution of points in a cube and approximate evaluation of integrals. *Zh. Vychisl. Mat. Mat. Fiz.*, Volume 7, Number 4, 784–802, 1967.

[14] S. Joe, F. Y. Kuo, Remark on Algorithm 659: Implementing Sobol's quasirandom sequence generator, *ACM Trans. Math.* Softw. 29, 49-57, 2003

[15] S. Joe, F. Y. Kuo, Constructing Sobol sequences with better two-dimensional projections, *SIAM J. Sci. Comput.* 30, 2635-2654, 2008.

[16] Z.H. Han, S. Görtz, R. Zimmermann, Improving Variable-Fidelity Surrogate modeling via Gradient-Enhanced Kriging and a Generalized hybrid Bridge Function. *Journal of Aerospace Science and Technology*, Vol. 25, Issue 1, 2013.

[17] R. Dwight, J. Brezillon, Effect of approximations of the discrete adjoint on gradient-based optimization, *AIAA Journal*, Vol. 44, Issue 12, 3022–3071, 2006.

[18] J. Brezillon, M. Abu-Zurayk, Aerodynamic Inverse Design Framework Using Discrete Adjoint Method, Notes on Numerical Fluid Mechanics and Multidisciplinary Design, 121. Springer Verlag. pp. 489-496. January 2013.

[19] T. Rowan, Functional Stability Analysis of Numerical Algorithms, Ph.D. thesis, Department of Computer Sciences, University of Texas at Austin, 1990.

[20] L. Zhao, K.K. Choi, I. Lee, A Metamodel Method Using Dynamic Kriging and Sequential Sampling. AIAA-2010-9391, Fort Worth, September, 2010.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# QUANTIFICATION OF SPATIAL VARIABILITY FOR TRANSVERSE ELASTIC MODULUS OF SPRUCE WOOD

**Alireza Farajzadeh Moshtaghin[1], Steffen Franke[2], Thomas Keller[1], and Anastasios P. Vassilopoulos[1]**

[1] Composite Construction Laboratory (CCLab),
Ecole Polytechnique Fédérale de Lausanne
Station 16, Bâtiment BP, CH-1015 Lausanne, Switzerland
e-mail: alireza.farajzadehmoshtaghin@epfl.ch, thomas.keller@epfl.ch,
anastasios.vasilopoulos@epfl.ch

[2] Institute for Timber Construction, Structures and Architecture
Bern University of Applied Sciences
Solothurnstrasse 102, CH-2500 Biel 6, Switzerland
steffen.franke@bfh.ch

**Keywords:** Clear Spruce Wood, Transverse Elastic Modulus, Tensile Tests, Random Spatial Variability, Mesostructure.

**Abstract.** *Different factors such as age, location of timber within the tree, structural imperfections, load history such as wind and snow etc. can affect the material properties of timber. Consequently, there is a high variability in the mechanical properties which is sometimes referred to as 'random spatial variability'. In this work, the spatial variability of the transverse elastic modulus of clear spruce wood is quantified by conducting mechanical tests. Specimens of 120 mm nominal length with a small cross-sectional area were prepared and their quasi-static behavior was experimentally investigated under tensile loading. In addition to the global displacement monitoring, the local deformations along the length of each specimen were measured. Using these data, the spatial variability of the elastic modulus was experimentally characterized in both longitudinal and transverse directions. Variability coefficients were defined to quantify the variability in both directions. Also, the effect of the mesostructure of the clear timber on the local elastic modulus was examined. The results show a very significant spatial variability in the transverse elastic modulus, more 1000% within some specimens. The variability of the transverse elastic modulus is much higher in the transverse direction. The grain angle has the most important effect on the value of the local elastic modulus. The results of this work can be considered as part of establishing advanced stochastic models with ransom spatial variability for timber structures.*

## 1 INTRODUCTION

Different factors such as age, location of timber within the tree, structural imperfections, load history such as wind and snow etc. can affect the material properties of timber taken from the same species, and grown in the same geographical location and local growth conditions. Consequently, there is a high inherent variability in the timber mechanical properties [1]. This variability is both spatial and random, and is sometimes referred to as 'random spatial variability' [2, 3].

A few experimental works have been devoted to the characterization of the spatial variability in the transverse mechanical properties of timber and timber products. Regarding the clear timber, Xavier et al. [4] used the unnotched Iosipescu test for the study of the radial variability of stiffness parameters of maritime pine wood. Pereira et al. [5] conducted tensile tests on cubic specimens cut in different radial positions and heights of Pinus pinaster tree. The results show that the radial transverse modulus and the shear transverse modulus decrease from tree center outwards. Also, Brandner and Schickhofer [6] used EN standard-type specimen (cubic bulk material) to investigate the spatial correlation in the transverse tensile strength and elastic modulus along the tree stem.

Nevertheless, the characterization of spatial variability in the transverse local elastic modulus at mesoscale (a few millimeters) has not been studied in detail in the literature. Consequently, the spatial variability is frequently neglected in the probabilistic simulations of timber structures [7-9]. The knowledge of the random spatial variability in elastic parameters can lead to predicting more accurate stress fields within the material. This can improve structural designs with a required reliability levels.

The mesostructure of clear wood is mainly characterized by earlywood-latewood arrangements. Concerning the transverse plane, the mechanical properties are superior in the radial direction [10]. Therefore, when the latewood are locally perpendicular to the nominal axis of a specimen, which is cut in the transverse plane, the local mechanical properties are higher.

Specimens of 120 mm lengths were cut in the transverse direction of clear spruce wood and their quasi-static behavior was experimentally investigated under tensile loading. In addition to the global displacement monitoring, the local deformations along the length of each specimen were measured. The spatial variability of the elastic modulus was experimentally characterized. The effect of the mesostructure of the clear timber on the local elastic modulus was examined.

## 2 EXPERIMNETAL INVESTIGATION

Clear Norway spruce wood was used for the specimens' preparation in this study. All specimens were conditioned to 12% moisture content and were tested at the laboratory temperature of 22 ± 3 C°. Three boards were cut from the same batch of spruce wood in the radial-longitudinal plane. These were used for cutting regularly positioned transverse specimens and designated as REB1, REB2 and REB3. Figure 1 shows REB1.

The nominal length of the specimens was 120 mm with a square cross section of 4×4 $mm^2$. A total of 52 specimens were tested on a 5 kN electromechanical Walter+Bai testing machine. Quasi-static tensile tests were performed in displacement-control mode. A stroke rate of 2 mm/min for the specimens was used on the basis of previous preliminary experiments so that the final failure occurred within 180 ± 60 s throughout the testing program.

A video extensometer system composed of a 10-bit Sony XCLU1000 CCD connected to a Fujinon HF35SA-1, 35 mm, f 1.4-22 lens with an accuracy of ± 0.005 mm was used during the experiments to measure the axial deformation. Prior to the tests, black target dots of 1.1 mm diameter were applied on the specimen surfaces. The distance between each two con-

secutive dots was 4 mm. A typical specimen mounted in the testing rig is shown in Fig. 2. The axial coordinates of the dots were recorded at a frequency of 5 Hz by the video extensometer throughout loading. Using these data, the engineering strain between each two consecutive dots was calculated, designated as the local strain. These data were used for calculation of the local elastic modulus. Also, using the displacements of the first and last dots on each specimen, an overall strain for each nominal length was obtained. These data were used for calculating the effective elastic modulus, for each specimen. Nominal axial stresses were calculated by using the load measurements and the initial cross-sectional areas.

## 3 SPATIAL VARIABILITY IN THE MECHANICAL PROPERTIES

The spatial variability of the transverse elastic modulus along the transverse direction are shown in Fig. 3 for the three REBs. The three boards are different in terms of local elastic modulus variations. On average, the third board is stiffer than the second board and the second board is stiffer than the first board. In most specimens cut from these boards, the local modulus value gradually increases from the one side, reaches a maximum value somewhere in the middle, and then decreases. Also, the scatter in the results is more significant in the second

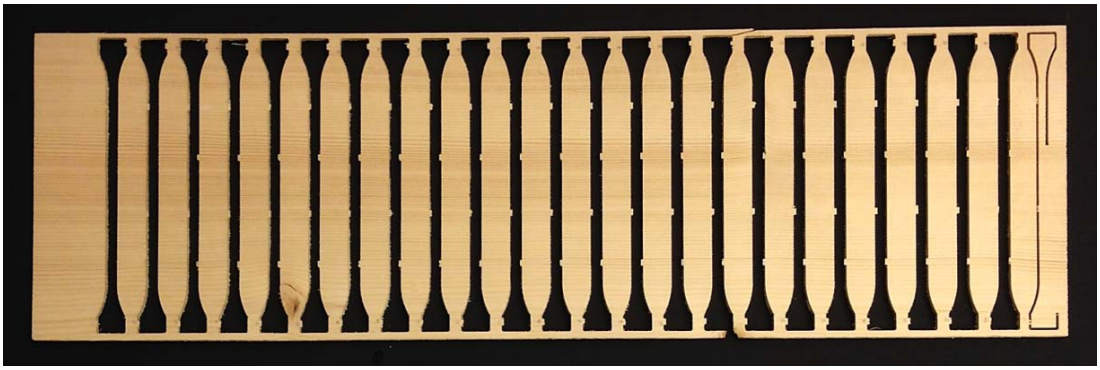


Figure 1: Board REB1 with regular arrangement of transverse specimens.



Figure 2: Specimen of 120 mm nominal length with applied dots inside machine grips.

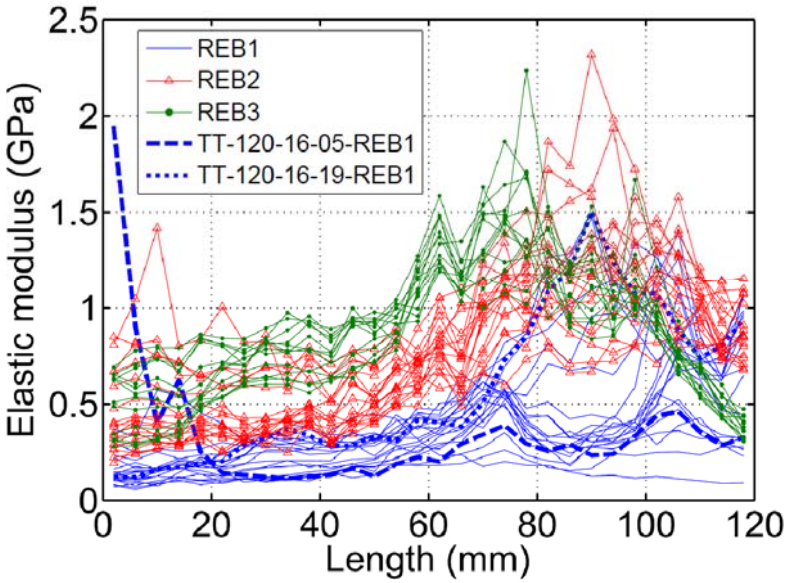



Figure 3: Spatial variability of the transverse elastic modulus in the transverse direction in 3 REBs.

board compared to the other two boards. There are major changes in the trend of the modulus variability in the first board in a few specimens compared to the others such as the highlighted dashed and dotted curves for specimens TT-120-16-05-REB1 (TT is transverse tensile, 120 is length, 16 is cross section, 05 is specimen number and REB1 is the board name) and TT-120-16-19-REB1, respectively, in Fig. 3. In the first case, first four 4 mm segments or the specimen were affected by the presence of a small nearby knot, which increased the local modulus very significantly. In the second case, the presence of a small knot in the board changed the mesostructure of the specimens 16 to 21 in REB1 at the range between 70 mm to 120 mm.

The variability of the local elastic modulus due to changing the spatial position within individual specimens (within-specimen variability or WSV) and its variability due to switching between specimens (between-specimen variability or BSV) contribute to the total variability of the local elastic modulus. In Table 1, the COVs for total variability of the local elastic modulus, the average COV for WSV and the COV for BSV are given. It is seen that the contribution of the WSV is higher than that of the BSV.

Variability of the transverse modulus in the longitudinal direction in the three REBs is plotted in Figs. 4-6. In each board, 30 paths longitudinal paths exist. Four curves have been highlighted in each figure in order to show how the longitudinal variability of the transverse modulus differs from lower paths to upper paths. On average, the longitudinal variability is higher in the first board.

| Parameter | Data set size | COV (%) |
|---|---|---|
| Local modulus (total variability) | 1560 | 60.65 |
| Local modulus of individual specimens (WSV) | 30 (each specimen) | 48.02 (average) |
| Effective modulus (BSV) | 52 | 44.06 |

Table 1: Variability of local and effective elastic moduli of 120 mm specimens.

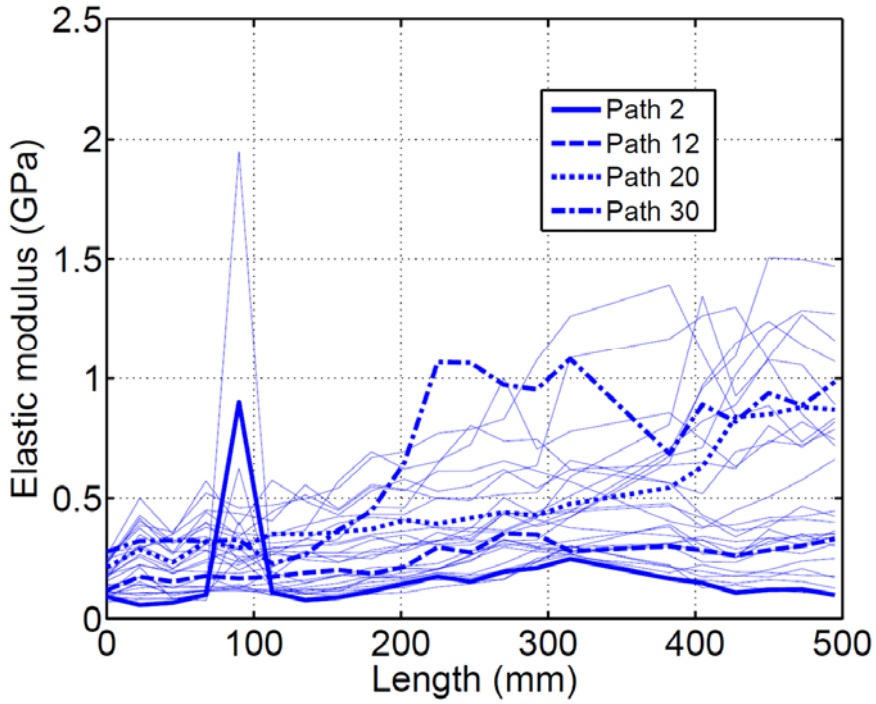



Figure 4: Spatial variability of the transverse elastic modulus in the longitudinal direction in REB1

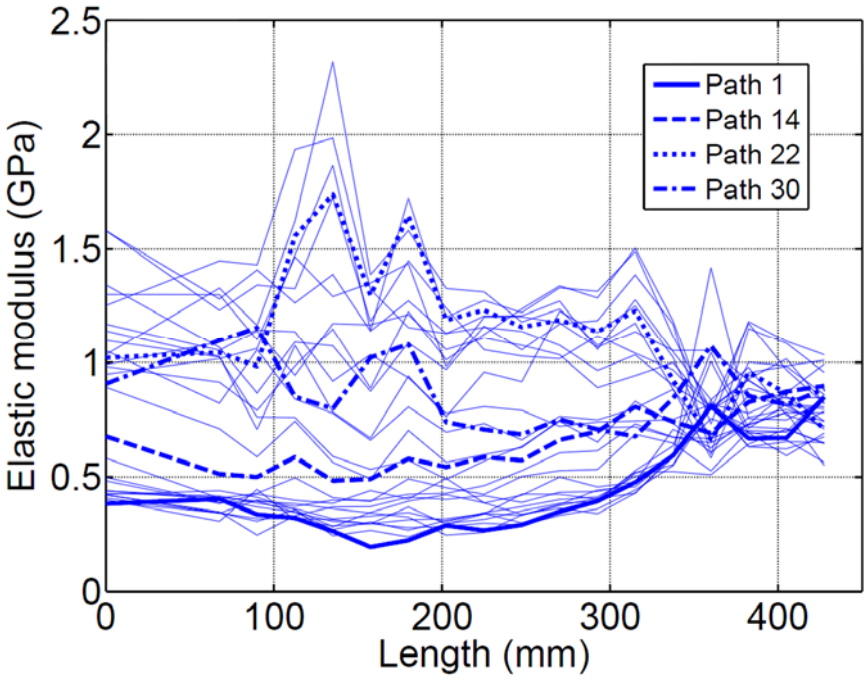



Figure 5: Spatial variability of the transverse elastic modulus in the longitudinal direction in REB2

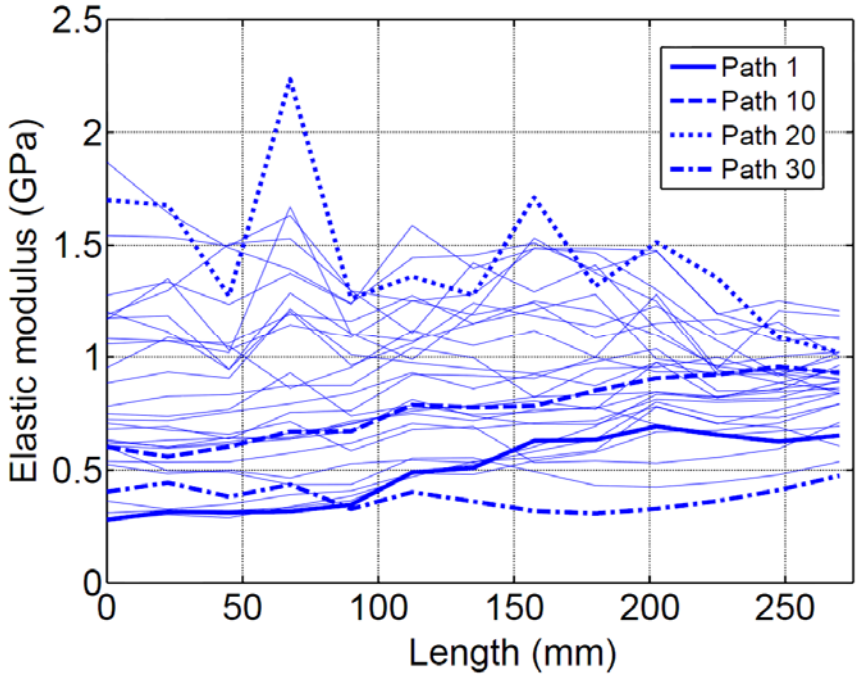



Figure 6: Spatial variability of the transverse elastic modulus in the longitudinal direction in REB3

As a general trend in REB1, the upper paths have higher values of the modulus compared to the lower paths. The paths are closer to each other in the left half length (lower scatter) of the board compared to the right half length (higher scatter). The unusual increase in the value of the transverse modulus, at the longitudinal position of 90 mm, as mentioned above for specimen TT-120-16-05-REB1, is due to a small nearby knot. Excluding this unusual variation at 90 mm position, the local modulus variation within individual paths ranges from 183.4 MPa to 1340.5 MPa, with a mean value of 581.3 MPa. The corresponding values for the

transverse paths in this board are 208.9, 1372.8 and 869.0 MPa. The mean variation in the longitudinal direction is significantly lower than in the transverse direction.

In the second board, the upper paths have again higher values. The difference between paths after the position of 320 mm from the left side is considerably lower than before that position. The minimum and maximum variation in the individual paths is 394.1 MPa and 1650.8 MPa, with a mean value of 686.4 MPa. These values for transverse paths are 484.0 MPa, 2075.5 MPa and 1064.5 MPa, respectively. Again, the mean variation in the transverse direction is higher than in the longitudinal direction.

In the third board, the longitudinal variability of the transverse elastic modulus is more uniformly distributed, although the scatter is slightly reduced from left to right. Generally, middle paths are stiffer. In this case, the minimum, the maximum and the mean variation in individual paths are 167.0 MPa, 1216.1 MPa and 421.9 MPa, respectively, which are considerably lower than the corresponding values for the transverse paths, namely 733.6 MPa, 1918.5 MPa and 1205.4 MPa. Similar to previous cases, the mean variation in the transverse direction is significantly higher than in the longitudinal direction.

The spatial variations of the elastic modulus in the three boards are shown in Fig. 7 as 2D contours. In order to compare the degree of the variability in the boards, three coefficients are defined for each board. The first is the ratio between the mean variation in the transverse direction and the transverse distance, the second is the ratio between the mean variation in the longitudinal direction and the longitudinal distance and the third is the ratio of these two coefficients, designated as anisotropy ratio, in the sense of different levels of variation in different directions. These values are provided in Table 2. It can be seen that the longitudinal variability coefficients are much lower than the transverse variability coefficient.

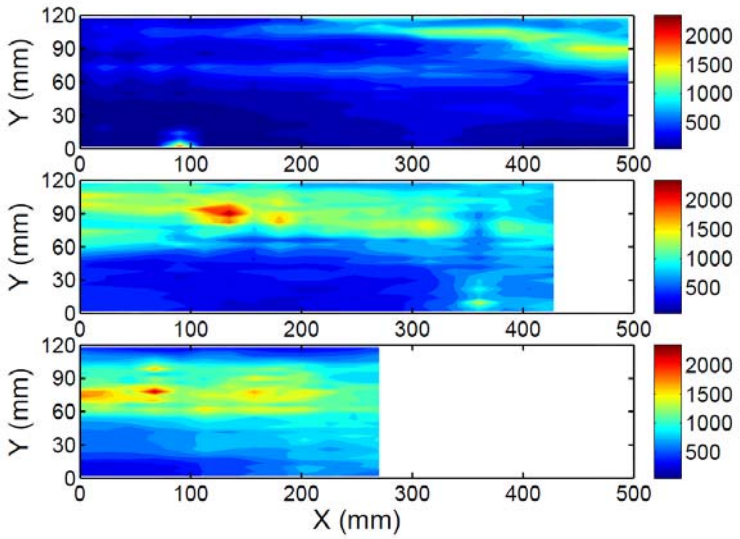



Figure 7: 2D contours of the spatial variability of the transverse elastic modulus in REBs

| Board number | Longitudinal variability coefficient (MPa/mm) | Transverse variability coefficient (MPa/mm) | Anisotropy ratio |
|---|---|---|---|
| 1 | 1.17 | 7.24 | 6.19 |
| 2 | 1.61 | 8.87 | 5.51 |
| 3 | 1.56 | 10.0 | 6.41 |

Table 2: Longitudinal variability coefficients, transverse variability coefficients and anisotropy ratios in REBs.

## 4 EFFECT OF TIMBER MESOSTRUCTURE ON LOCAL ELASTIC MODULUS

Typical examples of the correspondence between the local elastic modulus and the local mesostructure of spruce wood are given in Fig. 8, considering three specimens of 120 mm length. In specimen the TT-120-16-11-REB1 starting from the left, the grain angle, defined as the angle between the latewood and the vertical axis in Fig. 8, decreases up to 70 mm of the length. This is associated by a gradual increase in the local elastic modulus. The angle increases from 70 mm to 95 mm and after that decreases again, causing the local modulus to decrease and then increase, as a general trend. The maximum value happened to be at end of the specimen, where the grain angle has its lowest value. There is also a localized decrease in the value of the elastic modulus in the part of the specimen between 110 mm and 115 mm which is attributed to a decrease in the thickness of the latewood strips in this zone.

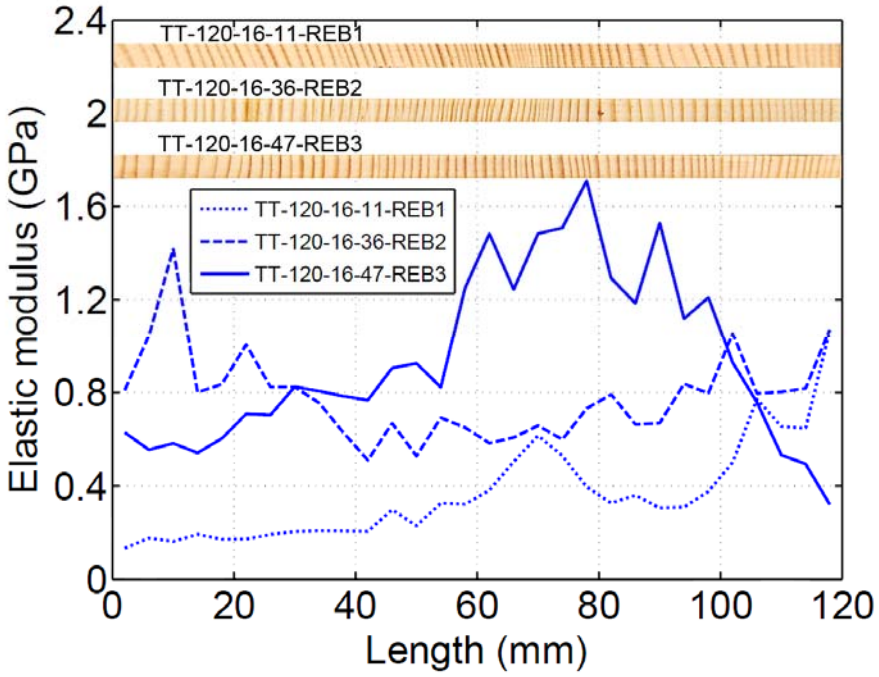



Figure 8: Correspondence between mesostructure of spruce and local transverse elastic modulus

In specimen the TT-120-16-36-REB2, over the first 10 mm of the specimen length, there is a decrease in the grain angle, then from 10 mm to about 60 mm the angle increases and after that the angle decreases. Correspondingly, the value of the elastic modulus first increases, then decreases and finally increases again, as a general trend. The effect of the grain angle on the local elastic modulus is even clear at the last 20 mm of the specimen length. In this segment of the specimen, the angle first increases a little and then decreases a little and, accordingly, there is a localized minimum in the value of the local elastic modulus in the middle of this segment. Also, a local maximum for the local modulus has happened in about 80 mm from the left end of the specimen. The grain angle has a local minimum in this position of the specimen length.

The local elastic modulus in specimen TT-120-16-47-REB3 has generally higher values compared the other two discussed specimens. The main reason is that, in this specimen, the grain angle has generally lower values. Only, in the first 30 mm and last 20 mm of the specimen length where the grain angle has higher values, the value of the modulus of this specimen is less than those of the other specimens. Also, in this specimen, latewood strips are thicker.

The presence of higher volume fractions of latewood tends to increase the value of the local elastic modulus. This can either happen by presence of a higher number of latewood strips or thicker latewood strips. For example, in both specimens TT-120-16-11-REB1 and TT-120-16-47-REB3 in the segment from 60 mm to 80 mm the number of latewood strips are higher than in the rest of segments in the corresponding specimens. Consequently, the transverse elastic modulus has a local peak in those segments.

## 5 CONCLUSIONS

- A highly significant spatial variability was observed in the transverse tensile elastic modulus of clear spruce wood. A difference of more than 1000% for the local elastic modulus was observed within some specimens.
- The variability of the transverse modulus is higher in the transverse direction than its variability in the longitudinal direction.
- The main reason for the spatial variability of the local elastic modulus is irregular changes in the mesostructure of the wood. The change in the grain angle is the most important factor influencing the local elastic modulus.
- Knots can substantially increase the transverse modulus of the clear wood around them.

## REFERENCES

[1] J.D. Barrett, Effect of size on tension perpendicular-to-grain strength of Douglas-fir. *Wood and Fiber*, **6**, 126-143, 1974.

[2] M. Vořechovský, Interplay of size effects in concrete specimens under tension studied via computational stochastic fracture mechanics. *International Journal of Solids and Structures*, **44**, 2715-2731, 2007.

[3] S. Sriramula, M.K. Chryssanthopoulos, An experimental characterisation of spatial variability in GFRP composite panels. *Structural Safety*, **42**, 1-11, 2013.

[4] J. Xavier, S. Avril, F. Pierron, J. Morais, Variation of transverse and shear stiffness properties of wood in a tree. *Composites: Part A*, **40**, 1953-1960, 2009.

[5] J. Pereira, J. Xavier, J. Morais, J. Lousada, Assessing wood quality by spatial variation of elastic properties within the stem: Case study of Pinus pinaster in the transverse plane. *Canadian Journal of Forest Research*, **44**, 107-117, 2014.

[6] R. Brandner, G. Schickhofer, Spatial correlation of tensile perpendicular to grain properties in Norway spruce timber. *Wood Science and Technology*, **48**, 337-352, 2014.

[7] P.L. Clouston, F. Lam, A stochastic plasticity approach to strength modeling of strand-based wood composites. *Composites Science and Technology,* **62**, 1381-1395, 2002.

[8] P.L. Clouston, F. Lam, Computational modeling of strand-based wood composites. *Journal of Engineering Mechanics*, **127**, 844-851, 2001.

[9] T. Tannert, F. Lam, T. Vallée, Strength prediction for rounded dovetail connections considering size effects. *Journal of Engineering Mechanics*, **136**, 358-366, 2010.

[10] Forest Product Laboratory, Wood handbook: Wood as an engineering material. *USDA, Madison, Wis*, 1999.

ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

# COMPARISON BETWEEN A POLYNOMIAL CHAOS SURROGATE MODEL AND MARKOV CHAIN MONTE CARLO FOR INVERSE UNCERTAINTY QUANTIFICATION BASED ON AN ELECTRIC DRIVE TEST BENCH

**Philipp Glaser[1], Michael Schick[1], Kosmas Petridis[1], and Vincent Heuveline[2]**

[1]Robert Bosch GmbH
Robert-Bosch-Campus 1, 71272 Renningen, Germany
e-mail: {philipp.glaser, michael.schick3, kosmas.petridis}@de.bosch.com

[2] Engineering Mathematics and Computing Lab (EMCL)
University Heidelberg, 69117 Heidelberg, Germany
e-mail: vincent.heuveline@uni-heidelberg.de

**Keywords:** Polynomial Chaos Expansion, Sparse Grid, Bayes Inference, Electric Drive Test Bench, Markov Chain Monte Carlo

**Abstract.** *The development of uncertainty quantification schemes has been pushed forward due to the increasing demands for complex physical and computational simulation models. In industrial applications, distributions on model parameters play a crucial role and quantifing them is a big challenge.*

*In this work, a test bench hardware is presented, which is designed to measure the motor characteristic of an electric drive. The special aspect of this setup is that parameter distributions, which in general are unknown, can be defined a priori. The obtained measurements serve as a reference to analyse the convergence of Polynomial Chaos (PC) and Markov Chain Monte Carlo (MCMC) in context of Bayesian inference.*

*Our focus is on analysing the feasibility of the PC approach as a surrogate model to replace the forward model in the Bayesian inference. In comparison to the classical approach, which directly uses the simulation model, we investigate the number of simulations needed to obtain a good estimation of the parameter distribution. In addition we use different orders for the PC expansion to fit the surrogate model. In our benchmark, we show that the PC expansion is able to significantly reduce the computational cost compared to a pure MCMC approach.*

## 1 INTRODUCTION

Many industrial applications involve uncertain parameters, for which correct values are rarely available in a precise way and which have a significant impact on simulation results. Therefore, the quantification of such uncertainties plays a crucial role, for example, in face of unknown component tolerances or measurement errors. Among many others, one challenging task is to infer knowledge on parameter distributions in light of experimental data.

In this context, Bayesian inference [6, 7, 16] offers a framework to combine numerical simulations with observed experimental data in order to gain better knowledge on model parameter uncertainties. One popular method is the so-called Markov Chain Monte Carlo (MCMC) approach in which the posterior of the model parameters is updated using correlated Monte Carlo samples. However, in order to achieve a high numerical accuracy, usually many MCMC samples are required. This can become a large drawback, especially if the numerical cost of a single simulation becomes prohibitive. In contrast, surrogate models, like, for example, Polynomial Chaos (PC) expansions [9, 18] provide a meta description of model output variations with respect to the random input, making the exploration of the posterior parameter distribution numerically cheap. PC is a spectral approach which expands the stochastic solution in an orthogonal polynomial basis defined in terms of the input parameters. The corresponding coefficients in this expansion, which account for the major part of the numerical cost, may be determined by many different methods. In our work, we consider a projection based approach [11, 21] to compute the coefficients using a Smolyak sparse grid numerical cubature [15].

Furthermore, we built a test bench hardware, which is used to measure a motor characteristic of an electric drive with uncertain physical parameters. In our test bench hardware it is possible to define physical reference parameter distributions a priori and store an according set of measurements. This can be seen as a replacement of the electric drive with another one from series-production for every measurement. An alternative method to obtain those measurement data would be to pick out a large amount of electric drives from the production line. For each sample drive a motor characteristic has to be recorded under a standardized measurement setup. This would increase significantly the time and cost needed. Afterwards we use this reference data to compare the posterior distribution obtained from a PC surrogate with the posterior of the MCMC approach starting from various prior distributions, which are different to the distributions taken to generate our reference data. Our focus is on the analysis of the truncation error resulting from restricting the total polynomial degree of the PC expansion along with a comparison of the numerical effort for computing the PC coefficients to the number of samples required for a pure MCMC approach.

The remainder of this article is organized as follows. Section 2 introduces the used methods for the approximation of the posterior distributions. In addition the PC surrogate model is presented and a method for fitting this model is addressed. The summary of this part is given by a combination of the surrogate model with the Bayesian inference. In the following section 3 the benchmark problem is introduced. This includes on the one hand the test bench hardware and on the other hand the underlying simulation model. The identification and verification between those is one main challenge and analyzed in a subsection. Besides the simulation model in total, one focus is on the modeling of physical parameter at the input. This is based on the defined hyperpriors for the mean

and standard derivation which have to be mapped into one prior. The numerical results are presented in section 4. We conclude this work and provide an outlook on open research questions in Section 5.

## 2 UNCERTAINTY QUANTIFICATION

In the following, a model with a given uncertain parameter space is considered. For convenience the input and output connection of this forward problem can be represented by a deterministic mapping

$$\boldsymbol{y} = \mathcal{M}(\boldsymbol{x}),$$

where $\boldsymbol{x} = \{x_1, \ldots, x_N\}^T \in \mathbb{R}^N$ is a vector of the model parameters uncertainties. The vector $\boldsymbol{y} = \mathbb{R}^Q$ represents the model outputs. As the parameter vector $\boldsymbol{x}$ is assumed to be uncertain, we change notation from $\boldsymbol{x}$ to $\boldsymbol{X}$. The probability density function (PDF) $p_{\boldsymbol{X}}(\boldsymbol{x})$ is assumed to be with independent, scalar components $x_i$ (finite noise assumption). Hence the model outputs are represented by a random vector, too, and are denoted as follows:

$$\boldsymbol{Y} = \mathcal{M}(\boldsymbol{X}).$$

### 2.1 BAYESIAN INFERENCE

The main issue of the Bayesian paradigm is to model uncertainties, e.g., arising from a lack of knowledge in a probabilistic way. In this work the focus is on model parameters $\boldsymbol{X}$ and any additional information about those can be described with a prior probability density $p_{\boldsymbol{X}}(\boldsymbol{x})$. In the absence of information, one may select a prior that is uninformative, e.g., a uniform distribution. Our goal is to update the knowledge about $\boldsymbol{X}$ by performing experiments and using the obtained observations $\boldsymbol{y}^{obs}$ to infer knowledge about the according pdf of $\boldsymbol{X}$. The connection between the model parameters and the gathered observations is expressed by the likelihood function $p(\boldsymbol{y}^{obs}|\boldsymbol{X})$, which requires the evaluation of a forward model. Using Bayes' rule, the posterior density for the model parameters can be written as

$$p(\boldsymbol{X}|\boldsymbol{y}^{obs}) = \frac{p(\boldsymbol{y}^{obs}|\boldsymbol{X})p(\boldsymbol{X})}{\int p(\boldsymbol{y}^{obs}|\boldsymbol{X})p(\boldsymbol{X})d\boldsymbol{X}}. \tag{1}$$

For convenience the likelihood function can be defined as $L(\boldsymbol{X}) := p(\boldsymbol{y}^{obs}|\boldsymbol{X})$, which may be viewed as a function of the model parameters $\boldsymbol{X}$. A classical approach is than to model the results from the forward problem with an additive noise:

$$\boldsymbol{y}^{obs} = \mathcal{M}(\boldsymbol{X}) + \boldsymbol{\epsilon},$$

where the components of $\boldsymbol{\epsilon}$ are independent and identically distributed random variables with density $p_\epsilon$ and the pdf of $\boldsymbol{\epsilon}$ is assumed to be normal distributed with mean $\mu = 0$ and variance $\sigma = \boldsymbol{\Sigma}$. In this simple case $\boldsymbol{\epsilon}$ includes both the model error and the measurement error, that can arise for example from sensor noise. The likelihood function can now be rewritten as follows:

$$L(\boldsymbol{X}) = \prod_i p_\epsilon(\boldsymbol{y}_i^{obs} - \mathcal{M}_i(\boldsymbol{X})), \tag{2}$$

where $\mathcal{M}_i(\boldsymbol{X}) := \mathcal{M}(\boldsymbol{X}^{(i)})$ denotes the model outputs at the $i$-th sample $\boldsymbol{X}^{(i)}$.The posterior distribution in equation (1) can be approximated by a Markov Chain Monte Carlo (MCMC) method. Therefore many simulation runs have to be evaluated, which often are expensive to compute due to the involved cost of evaluating the forward model $\mathcal{M}_i(\boldsymbol{X})$. Using a surrogate model, e.g. a Polynomial Chaos expansion, which is described in subsection 2.2, can provide a means to decrease the computational costs.

## 2.2 POLYNOMIAL CHAOS EXPANSION

The original polynomial chaos was first introduced by Wiener [18]. It uses the Hermite polynomials with respect to Gaussian random variables. It can also be extended to non-Gaussian measures [2]. Considering the given model $\mathcal{M}(\boldsymbol{X})$, the output is represented as a random vector $\boldsymbol{Y}$. Without loss of generality, a scalar output $Y$ is assumed for simplicity in the following. The vector case can be obtained by applying the approach componentwise to the model outputs using the same polynomial basis for each component.

### 2.2.1 DEFINITION IN ONE-DIMENSIONAL CASE

Assume the probability space $(\Theta, 2^\Theta, \mathbb{P})$, where $\Theta$ is the sample space and $\theta \in \Theta$ is a sample of it. $2^\Theta$ is the $\sigma$-algebra and $\mathbb{P}$ is some probability measure. Now, let $Y(\theta)$ be a random process, treated as a function of $\theta$. The spectral stochastic representation considered in this work can be seen as an Fourier-like decomposition [12]:

$$Y(\theta) = \sum_i y_i \phi_i(\xi(\theta)),$$

where $y_i$ are the spectral coefficients, which have to be determined. The hyperparameter $\xi$ is defined on the support $\Theta$, corresponding to type of basis polynomials. $\{\psi_i\}_{i=0}^{\infty}$ denote the basis functions, which form an orthogonal basis with the following orthogonality relation:

$$\langle \phi_j(\xi(\theta)), \phi_k(\xi(\theta)) \rangle = \mathbb{E}\left[\phi_j(\xi(\theta))\phi_k(\xi(\theta))\right] = \mathbb{E}\left[\phi_k^2(\xi(\theta))\right]\delta_{jk}, \quad j,k \in \mathbb{N}_0, \tag{3}$$

with the Kronecker delta $\delta_{jk}$ which is equal 1 if $j = k$ and 0 otherwise. $\mathbb{E}\left[\phi_k^2(\xi(\theta))\right]$ is a normalization factor. To simplify the notation, we shall drop the dependence of the hyperparameter $\xi$ on $\theta$. Suppose that the cumulative distribution function $F(\xi)$ of $Y$ is absolutely continuous w.r.t. the Lebesque-measure, then the corresponding PDF exists, such that $dF(\xi) = \rho(\xi)d\xi$. The inner product from equation (3) can be written as:

$$\langle \phi_j(\xi), \phi_k(\xi) \rangle = \mathbb{E}\left[\phi_j(\xi)\phi_k(\xi)\right] = \int \phi_j(\xi)\phi_k(\xi)\rho(\xi)d\xi, \quad j,k \in \mathbb{N}_0.$$

Some examples on the correspondence between common continuous distributions on $Y$ and the type of the generalized PC basis polynomials and the support $\Theta$ for the hyperparameter $\xi$ [21] can be found in Table 1.

### 2.2.2 DEFINITION IN MULTI-DIMENSIONAL CASE

Now we assume the hyperparameter $\xi$ to be a random vector $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_N\}$ with mutually independent components. The orthogonality relation from equation (3) immediately takes the form (due to independence):

$$\langle \phi_j(\boldsymbol{\xi}), \phi_k(\boldsymbol{\xi}) \rangle = \mathbb{E}\left[\phi_j(\boldsymbol{\xi})\phi_k(\boldsymbol{\xi})\right] = \int \phi_j(\boldsymbol{\xi})\phi_k(\boldsymbol{\xi})\boldsymbol{\rho}(\boldsymbol{\xi})d\boldsymbol{\xi}, \quad 0 \leq j,k \leq \infty, \tag{4}$$

Table 1: Correspondence between the type of generalized PC polynomial basis and probability distribution.

| distribution | generalized PC basis polynomials | support |
|---|---|---|
| Beta | Jacobi | $[a, b]$ |
| Gamma | Laguerre | $[0, \infty]$ |
| Gaussian | Hermite | $(-\infty, \infty)$ |
| Uniform | Legendre | $[a, b]$ |

where $\boldsymbol{\rho}(\boldsymbol{\xi}) = \prod_{i=1}^{N} \rho_i(\xi_i)$. Let $\boldsymbol{i} = \{i_1, \ldots, i_N\} \in \mathbb{N}_0^N$ denote a multi-index with $|\boldsymbol{i}| = i_1 + \cdots + i_N$. The $N$-variate generalized PC basis functions can be obtained by the product of the univariate generalized PC polynomials in equation (4):

$$\psi_{\boldsymbol{i}}(\boldsymbol{\xi}) = \prod_{j=1}^{N} \phi_{i_j}(\xi_j).$$

Then, the generalized PC expansion reads:

$$Y(\boldsymbol{\xi}) = \sum_{|\boldsymbol{i}|=0}^{\infty} y_{\boldsymbol{i}} \psi_{\boldsymbol{i}}(\boldsymbol{\xi}),$$

where $y_{\boldsymbol{i}}$ are the unknown PC coefficients. In addition the multi-index $\boldsymbol{i}$ can be mapped into a one dimensional index by a one-to-one correspondence, which results in the following relation:

$$Y = \sum_{j=0}^{\infty} y_j \psi_j(\boldsymbol{\xi}). \tag{5}$$

For computational reasons the given polynomials have to be truncated at order $p$. The approximation is denoted as $\tilde{Y}$ and can be written as:

$$\tilde{Y} = \sum_{j=0}^{P} \tilde{y}_j \psi_j(\boldsymbol{\xi}), \quad \text{with } P + 1 = \frac{(N+p)!}{N!p!}. \tag{6}$$

In a second step the PC coefficients $\tilde{y}_j$ for $j = 0, \ldots, P$ in equation (6) have to determined. The PC coefficients can be recast each as a multidimensional integral which can be computed by multivariate numerical quadrature rule. This approach is called the projection method and will be used in the following.

Starting from the expansion of the solution in equation (6) the projection method projects the solution $\tilde{Y}$ against each basis function $\{\psi_i\}$ using inner product:

$$\left\langle \tilde{Y}, \psi_i(\boldsymbol{\xi}) \right\rangle = \left\langle \sum_{j=0}^{P} \tilde{y}_j \psi_j(\boldsymbol{\xi}), \psi_i(\boldsymbol{\xi}) \right\rangle. \tag{7}$$

By applying the orthogonality property of the basis polynomials in equation (3), the

right-hand side of the equation (7) can be modified as follows:

$$\begin{aligned}\mathbb{E}\left[\tilde{Y}\psi_j(\boldsymbol{\xi})\right] = \left\langle \tilde{Y}, \psi_i(\boldsymbol{\xi})\right\rangle &= \left\langle \sum_{j=0}^{P} \tilde{y}_j \psi_j(\boldsymbol{\xi}), \psi_i(\boldsymbol{\xi})\right\rangle \\ &= \sum_{j=0}^{P} \tilde{y}_j \mathbb{E}\left[\psi_j(\boldsymbol{\xi})\psi_i(\boldsymbol{\xi})\right] \\ &= \sum_{j=0}^{P} \tilde{y}_j \mathbb{E}\left[\psi_j^2(\boldsymbol{\xi})\right] \delta_{ij}.\end{aligned}$$

Thereby the PC coefficients result in

$$\tilde{y}_j = \frac{1}{\mathbb{E}\left[\psi_j^2(\boldsymbol{\xi})\right]} \mathbb{E}\left[\tilde{Y}\psi_j(\boldsymbol{\xi})\right], \quad \text{for } j = 0, \ldots, P,$$

with $\tilde{Y}$, that can be seen as an approximation of the random process. If we assume that $\boldsymbol{\xi}$ is continuous then the orthogonality can be written as:

$$\tilde{y}_j = \frac{1}{\mathbb{E}\left[\psi_j^2(\boldsymbol{\xi})\right]} \int_{\mathbb{R}^N} \tilde{\mathcal{M}}(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi})\rho(\boldsymbol{\xi})d\boldsymbol{\xi}, \quad \text{for } j = 0, \ldots, P, \tag{8}$$

where $\rho(\boldsymbol{x})$ is the multivariate distribution density. The PC coefficients in equation (8) can be estimated numerically based on discretization. The integration can be done e.g. using a full tensor quadrature rule with order $r$ and weights $w_i$. Hence the total number of quadrature points computes to:

$$R = (r)^N \tag{9}$$

and the approximation of the PC coefficients can be expressed as:

$$\tilde{y}_j \approx \hat{y}_j = \frac{1}{\mathbb{E}\left[\psi_j^2(\boldsymbol{\xi})\right]} \sum_{i=1}^{R} \tilde{Y}(\xi_i)\psi_j(\xi^{(i)})w_i, \quad \text{for } j = 0, \ldots, P. \tag{10}$$

### 2.3 SPARSE GRID

One limitation of the full grid approach is the severe exponential growth of the required simulations related to the number of dimensions in the parameter space. Given that, modifications has to be done to make the method applicable to complex models. One approach considers the choice of the nodes where the model is evaluated by selecting 'important' points. This is called sparse grid and it can reduce the number of evaluations appreciably.

Based on a spectral projection method described in the previous section 2.2 it is essential to give a selection of nodes to approximate the multidimensional integral in equation (8) with a discrete sum given in equation (10). We use a set of points generated by Gauss quadrature. This rule approximates best for the one-dimensional case ($N = 1$) [19, 21]. In the multi-dimensional case ($N > 1$) the total number of simulation points grows fast in high dimensional random parameter space.

It is obvious from equation (9) that even for a proportionally small parameter space $N \gg 1$ and an adequate PC order of $p = 3$, it can be assumed that $Q = 4^N \gg 1$. With

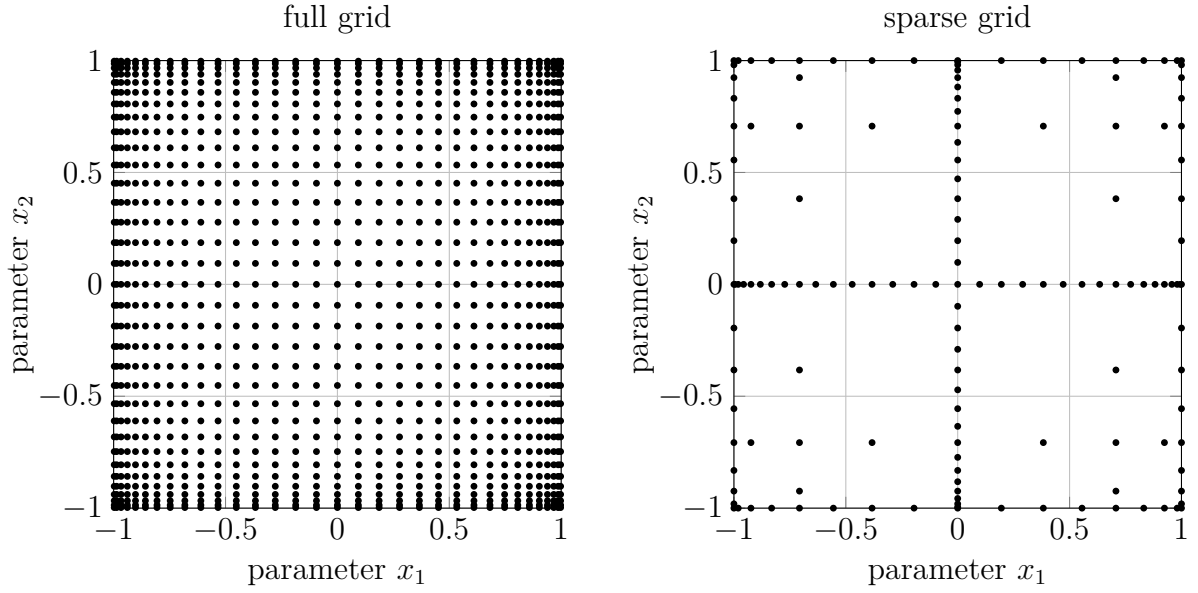



Figure 1: Comparison of two-dimensional grids based on the same one-dimensional nodes based on a two-dimensional parameter space. Left: full grid with 1089 nodes. Right: sparse grid with 145 nodes.

respect to this property the full grid rule is only used at lower dimensions. To reuse the spectral projection method sparse grids are proposed in this context [20]. Those were introduced in [15] and they are first used in the context of multivariate integration [8, 14].

The Smolyak algorithm is a linear combination of product formulas and based on a one-dimensional interpolation formula for an univariate function $f$

$$Q_j^1 f := \sum_{i=1}^{n_j^1} w_{ji} f(x_{ji}),$$

with the nodal sets $\Theta_j^1 = \left(x_{j1}, \ldots, x_{jn_j^1}\right)$. The Smolyak algorithm [8] is given:

$$Q_j^k f = \sum_{j \leq |\boldsymbol{i}|_1 \leq j+k-1} (-1)^{j+k-|\boldsymbol{i}|_1-1} \begin{pmatrix} k-1 \\ |\boldsymbol{i}|_1 - j \end{pmatrix} (Q_{i_1}^1 \otimes \cdots \otimes Q_{i_k}^1) f, \tag{11}$$

where $\boldsymbol{i} = (i_1, \ldots, i_N) \in \mathbb{N}$. The points of the multivariate Smolyak formula in equation (11) form a sparse grid given by the union over the parwise disjoint nodal sets:

$$\Theta_N = \bigcup_{j \leq |\boldsymbol{i}|_1 \leq j+k-1} (\Theta_{i_1}^1 \otimes \cdots \otimes \Theta_{i_k}^1).$$

Figure 1 shows the comparison of a full and sparse grid based on a two dimensional parameter space. It can be seen that the sparse grid based on the Smolyak algorithm [15] is a subset of the full grid and it consists of less simulation points than the full tensor grid. This opens up the possibility of considering higher dimensional random parameter spaces in combination with the PC expansion [1, 5, 11].

## 2.4 BAYESIAN INFERENCE WITH A PC SURROGATE MODEL

The integration of the PC surrogate model with a suitable sparse grid approach and the Bayesian inference can be done in the following way. The assumptions on the parameter

distributions can be used to define priors $p(\boldsymbol{X})$ for the uncertain model parameters. With knowledge of the PC coefficients obtained in equation (10), an approximation for the density function of model output $\hat{Y} = \hat{\mathcal{M}}_p(\boldsymbol{\xi}(\boldsymbol{X}))$ with order $P$, which can be seen as an surrogate model.To simplify the notation, we shall drop the dependence of the hyperparameter $\boldsymbol{\xi}$ on $\boldsymbol{X}$ and the PC surrogate model reads:

$$\hat{\mathcal{M}}_P(\boldsymbol{\xi}) = \sum_{j=0}^{P} \hat{y}_j \psi_j(\boldsymbol{\xi}).$$

The likelihood function $L(\boldsymbol{X})$ from equation (2) results in the following approximation:

$$\hat{L}(\boldsymbol{X}) = \prod_i p_\epsilon \left( y_i^{obs} - \hat{\mathcal{M}}_P(\boldsymbol{\xi}) \right) = \prod_i p_\epsilon \left( y_i^{obs} - \sum_{j=0}^{P} \hat{y}_j \psi_j(\boldsymbol{\xi}) \right). \tag{12}$$

Applying the obtained result to equation (1) and excluding the normalization in the denominator for an instant, one achieve the following:

$$p(\boldsymbol{X}|\boldsymbol{y}^{obs}) \propto L(\boldsymbol{X}) p(\boldsymbol{X}) = \prod_i p_\epsilon \left( y_i^{obs} - \sum_{j=0}^{P} \hat{y}_j \psi_j(\boldsymbol{\xi}) \right) p(\boldsymbol{X}). \tag{13}$$

As mentioned above, one approach to estimate the posterior parameter distributions $p(\boldsymbol{X}|\boldsymbol{y}^{obs})$ is to employ MCMC methods, which construct a Markov chain on the parameter space $\boldsymbol{X}$, whose steady state distribution is the posterior distribution of interest.

## 3 BENCHMARK PROBLEM

One main challenge is the validation of the resulting posterior distributions by applying the methods mentioned in section 2 to a real application case. One reason is that parameter distributions can not easy obtained in many cases or determining them is very expensive. On the other hand measurements of the outputs can be gathered simply.

For this purpose a test bench hardware is built up and there it is possible to vary the voltage supply and the winding resistance in an automatic way. This offers the possibility to define parameter distributions and run the test bench on them. For the resulting measurements the based parameter variations are known. Besides that, different motor characteristics can be chosen and recorded. The given model of an electric drive consists of a motor and a worm gear. Besides the electrical and mechanical relations, a detailed thermal model is available. To obtain a good fitting between the test bench and the model, a parameter identification, described in section 3.3, was performed.

### 3.1 TEST BENCH HARDWARE

The electric drive (cf. figure 2, ①) contains the motor and the gearing and is attached to a plate, which is connected to the mounting (cf. figure 2, ⑥) of the test bench. Several temperature sensors are fixed to the motor and the gear housing. The motor shaft is connected with a torque sensor (cf. figure 2, ②). In addition the rotational speed can be measured. The electromagnetic powder break (cf. figure 2, ③) enables the test bench to perform different series of measurements, which can range from static operating mode through motor characteristics. To compensate the resulting shaft misalignments, metal bellow-type couplings (cf. figure 2, ⑤) are used for the link between the components. The electric drive is connected with a power supply (cf. figure 2, ④).

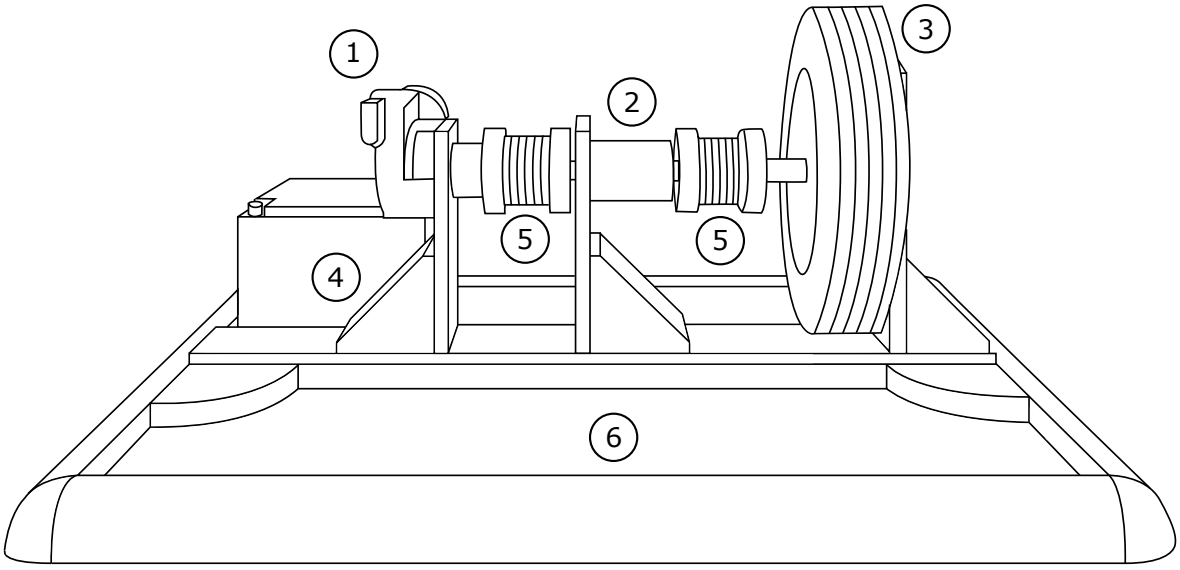



Figure 2: Schematic view on the test bench hardware.

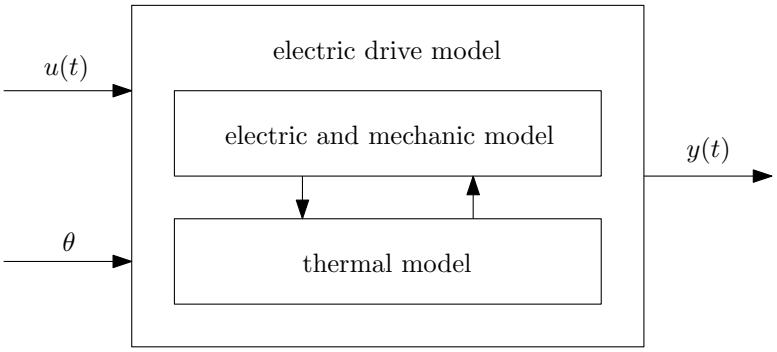



Figure 3: Schematic view on the electric drive model with the input/output connections.

The special feature of the test bench is the possibility to add uncertainties, which influences the power supply voltage of the drive and the characteristics of the electric motor. Hence, this can be done in an automatic way without replacing the electric drive or the power supply. For this reason many measurements based on given parameter variations can be generated.

### 3.2 SIMULATION MODEL

The given model of an electric drive consists of a motor and a worm gear. Beside the electrical and mechanical relations, a detailed thermal model is provided. They are strongly connected in both directions by internal states, cf. Fig. 3. The electric drive has eight model parameters and one input which is considered as uncertain. We assume that it is independent and identically distributed with a Gaussian distribution. To get a general view on the model, simplified equations are representing the correlations of the parts. The electric and mechanic part is characterized by the two states which are the current $I(t)$ and motor speed $\omega(t)$. The ordinary differential equations are given as

follows:

$$\frac{d}{dt}I(t) = \frac{1}{L}\Big[ - RI(t) - c_m\omega(t) + \Big(U(t) - U_{drop}(I(t))\Big)\Big], \tag{14}$$

$$\frac{d}{dt}\omega(t) = \frac{1}{J}\Big[ - \tau_{load}(t)i_g + \Big(\Big(\tau_{motor}(I(t), c_m)\Big) - \tau_{fric}\Big(\omega(t), \eta\Big)\Big)\eta\Big]. \tag{15}$$

The inputs of the simulation model are the voltage $U(t)$ and the load torque $\tau_{load}(t)$ which are summarized as $\boldsymbol{u}(t)$. The parameters are defined as the winding resistance $R$, the motor constant $c_m$, the inductance $L$, the gear ratio $i_g$ and the gear efficiency $\eta$. The voltage drop $U_{drop}(I(t))$ is approximated by a polynomial of high order depending on the current $I(t)$.

The motor torque $\tau_{motor}$ can be characterized by the following equation:

$$\tau_{motor}(I(t), c_m) = c_m I - \tau_{loss}(\epsilon_{iron}),$$

where the losses in the copper, the iron and by hysteresis are summarized. They can be adapted by the parameter $\epsilon_{iron}$. The mechanical friction losses can be described by the following nonlinear function:

$$\tau_{fric} = f(\omega(t), \eta, \mu_A, \mu_B),$$

which only shows the dependences for the sake of simplicity. The losses in the bearings A and B can be modified by the parameters $\mu_A$ and $\mu_B$.

The thermal model of the electric drive is subdivided into several parts related to the mechanical structure. The link with the two states given in equations (14) and (15) are provided with the thermal power losses. The additional states are the temperatures at

- the magnet $T_{magnet}$,
- the bearings A $T_{bearingA}$ and B $T_{bearingB}$,
- the minus brush $T_{brush}$,
- the worm $T_{worm}$ and
- the coil $T_{coil}$.

The interconnection with the electrical and mechanical model are as follows:

$$\begin{aligned}
R &= R_0\left[1 + \alpha\left(T_{coil} - T_{amb}\right)\right], \\
c_m &= \sum_{n=0}^{6} k_p(n) I^n, \quad \text{with } k_p(n) = \sum_{i=0}^{6} b_n(i) T_{magnet}^i, \\
\eta &= g(T_{worm}), \quad g \in \text{nonlinear function}, \\
\tau_{fric} &= h(T_{beraingA}, T_{beraingB}) \quad h \in \text{nonlinear function}.
\end{aligned}$$

### 3.3 OBTAINING MEASUREMENT DATA AND PARAMETER IDENTIFICATION

On main part of the benchmark problem is the parameter identification of the introduced model in subsection 3.2 with the measurements obtained from the test bench mentioned in subsection 3.1. In this paper, the temperatures of the system supposed to be at a stationary level and therefore the measurements have to be taken in a bounded temperature range:

$$T_{brush} = T_{brush,stat} \pm 1[K], \quad T_{magnet} = T_{magnet,stat} \pm 0.5[K],$$
$$T_{winding} = T_{winding,stat} \pm 0.5[K], \quad T_{worm} = T_{worm,stat} \pm 1.5[K].$$

This can be obtained by monitoring the temperature levels and run a warm-up sequence before the measurements are recorded. A measurement $i$ starts at time $t_0^i$ and records the time responses of the outputs. After a fixed time $\Delta t$ the measurement is finished. The resulting data can be called measurement run $i$ and contains $I_{measured,i}(t)$, $\omega_{measured,i}(t)$ within the range $t \in [t_0^i, t_0^i + \Delta t]$. Moreover, the voltage $U(t)$ is set to a stationary value in run $i$ using the Heaviside step function:

$$U(t) = \hat{U}_i H(t), \quad t \in [t_0^i, t_0^i + \Delta t].$$

Before the next run is started, the electric drive is switched off by setting the voltage $U(t > t_0^i + \Delta t)$ to zero. The next run $i+1$ will then start at time $t_0^{i+1} > t_0^i + \Delta t$ with stationary voltage $\hat{U}_{i+1}$. This offers the opportunity to vary the model input $\hat{\boldsymbol{U}} = [\hat{U}_1, \hat{U}_2, \dots]$ over the measurement runs $1, 2, \dots$, or to put it another way, the electric drive is switched on in every run and the variation can be seen in the different stationary input voltages.

For the identification of the simulation model a measurement run at $\hat{U}_{ident}$ is performed and the obtained data over time is used to fit the physical model parameters:

$$\theta = [R, c_m, L, J, \eta_{iron}, \mu_A, \mu_B, \mu_{worm}] .$$

This nonlinear optimization problem is solved using the Matlab Optimization Toolbox [17] with the function lsqnonlin which includes a large-scale algorithm. This algorithm is a subspace trust-region method and is based on the interior-reflective Newton method described in [3] and [4]. Each iteration involves the approximate solution of a large linear system using the method of preconditioned conjugate gradients. The estimated parameters are denoted as $\hat{\theta}$ in the following.

### 3.4 MODELING OF PHYSICAL PARAMETER

Instead of using a non-parametric Bayesian model, it is assumed that the distribution of the physical parameter $U$ belongs to a Normal distribution. This leads to a parametric model which has to be extended in the following way:

$$\tilde{U}(\mu_U, \sigma_U, \xi_U) = \mu_U + \sigma_U \xi_U,$$

with the hyperprior $\mu_U$ and the hyperprior $\sigma_U$. The parameter $\xi_U$ is Normal distributed with expected value 0 and standard derivation 1. The interconnection of the parameter and the simulation model is shown in figure 4.

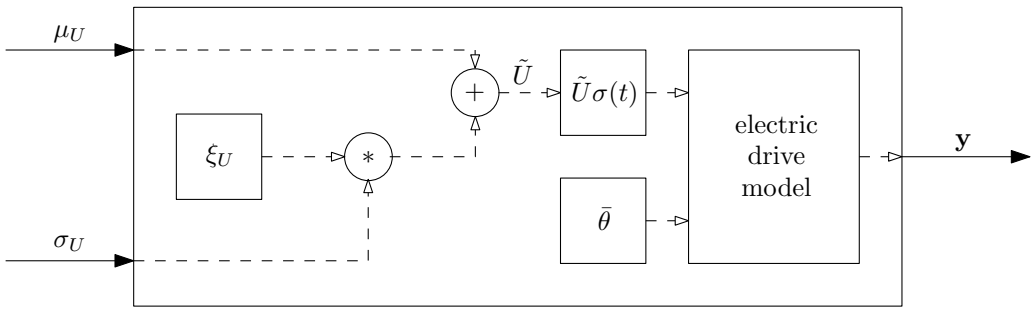



Figure 4: Parameter mapping.

Using the parametric model the likelihood function in equation (13) can be rewritten in the following way:

$$\hat{L}(\boldsymbol{X}) = \prod_i p_\epsilon \left( y_i^{obs} - \sum_{j=0}^{M} \hat{y}_{i,j} \psi(\xi_\mu(\mu_U), \xi_\sigma(\sigma_U), \xi_U) \right),$$

where $\xi_\mu(\mu_U)$ and $\xi_\sigma(\sigma_U)$ are hyperparameters of the PC surrogate model with the mapping:

$$\xi_\mu(\mu_U) = \frac{\mu_U - E[P(\mu_U)]}{\sqrt{Var(P(\mu_U))}}, \quad \xi_\sigma(\sigma_U) = \frac{\sigma_U - E[P(\sigma_U)]}{\sqrt{Var(P(\sigma_U))}}.$$

## 4 NUMERICAL RESULTS

We assume a constant load scenario with $\tau_{load} = 3Nm$ and a stationary input voltage $\hat{U}$. For this reason the observed motor speed $\omega$ will reach a steady state. In the following all given quantities are normalized with a fixed maximum value.

The physical parameter of the stationary voltage $\hat{U}$ is assumed to be Normal distributed with the expected value $\mu_{U,real} = 0.9$ and the standard derivation $\sigma_{U,real} = 0.047$. The number of measurement runs is set to 100 and $\Delta t$ is fixed to 10 [s]. The data can be obtained using $\hat{\boldsymbol{U}} = [\hat{U}_1, \ldots, \hat{U}_{100}]$, which are drawn from the assumed distribution. Based on the setting a steady state of the motor speed $\omega_{measured,i}(t)$ will be reached after time $t = t_{steady}$ with $0 < t_{steady} \leq \Delta t$ for each run $i$ and a stationary value for the motor speed $\omega_{steady,i} = \omega_{measured,i}(t_{steady})$ can be obtained. This means that for every measurement run a scalar value for the motor speed is taken into account, instead of a time behavior.

Based on the measurement runs $i = 1, \ldots, 100$ an error $\epsilon_i$ between the data $\omega_{meas,i} = \omega_{steady,i}$ and the simulation model $\omega_{sim,i} = \mathcal{M}(\hat{U}_i, \hat{\theta})$ can be defined. This is illustrated in figure 5. The resulting error $\epsilon$ can be described by the stochastic moments:

$$\mu_\epsilon = 0.00015, \quad \sigma_\epsilon = 0.00603.$$

Applying the MCMC method to the described benchmark problem and the measured data, the Metropolis-Hastings algorithm [10, 13] is used for the estimation of posterior distribution. The priors are assumed to be Uniform distributed with the following stochastic moments:

$$P(\mu) = \mathcal{U}(0.65, 1), \quad P(\sigma) = \mathcal{U}(0, 0.065).$$

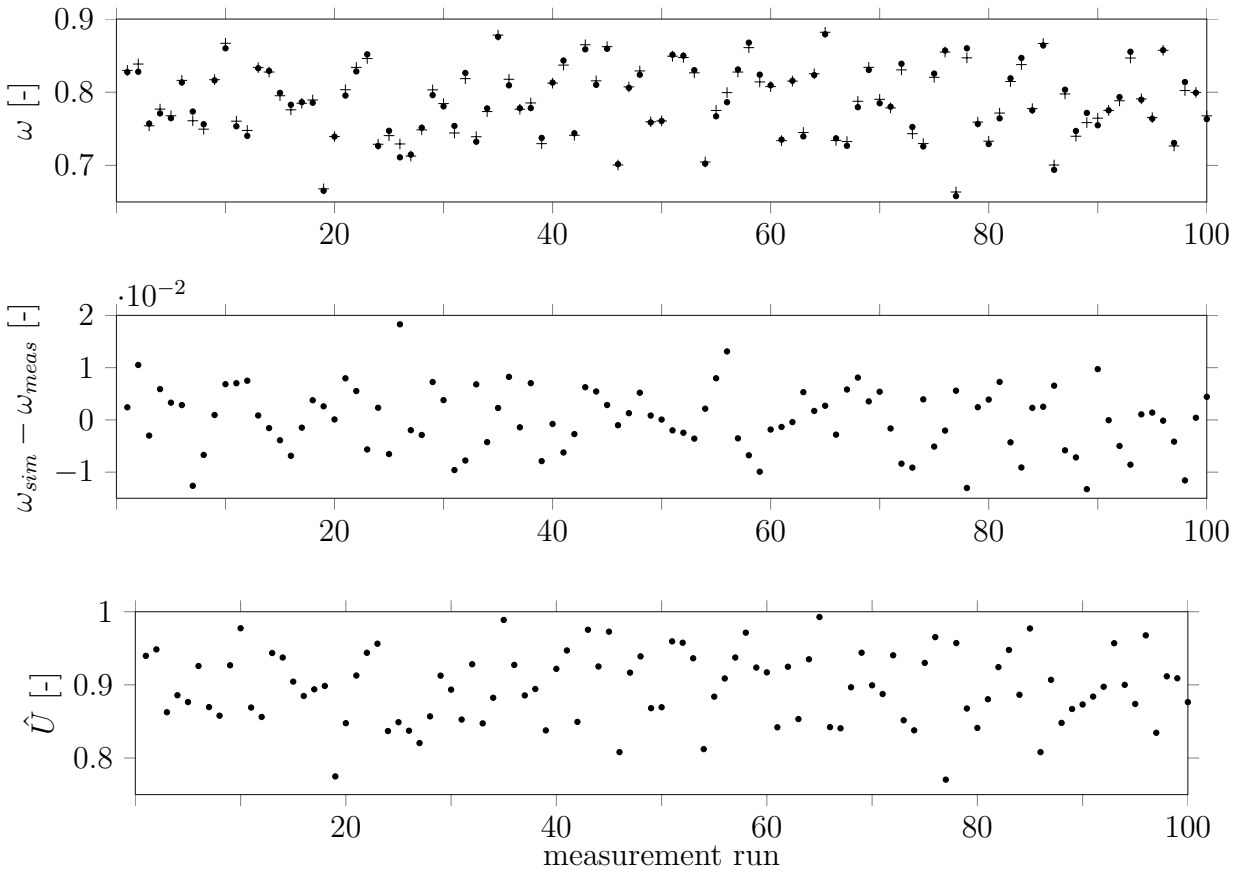



Figure 5: The first plot shows the normalized model output $\omega$ at a stationary time point $t_{steady}$ for the measurement ($\bullet$) and the simulation result ($+$). The second plot shows the error $\epsilon$ between the measurement and the simulation output. The third plot shows the corresponding normalized voltage for the model evaluation.

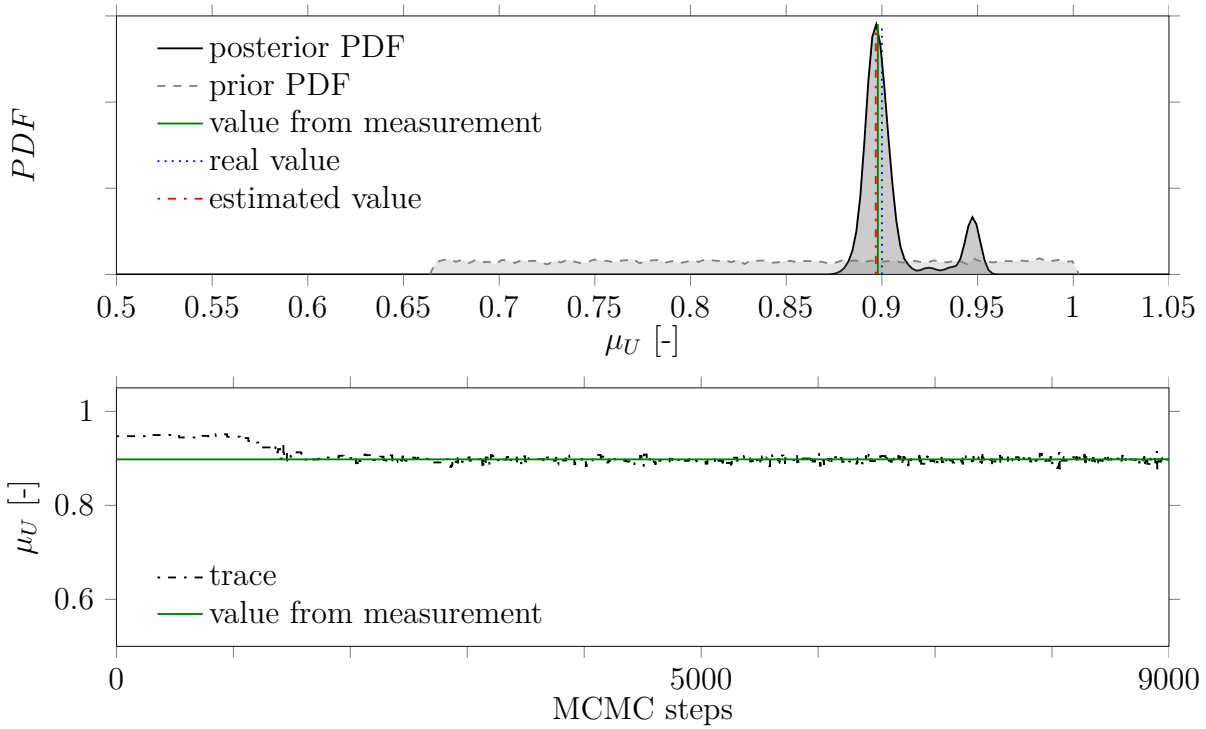



Figure 6: This figure shows the result of the MCMC estimation. The first plot contains the prior and posterior PDF of the hyperparameter $\mu_U$. The second plot visualize the trace.

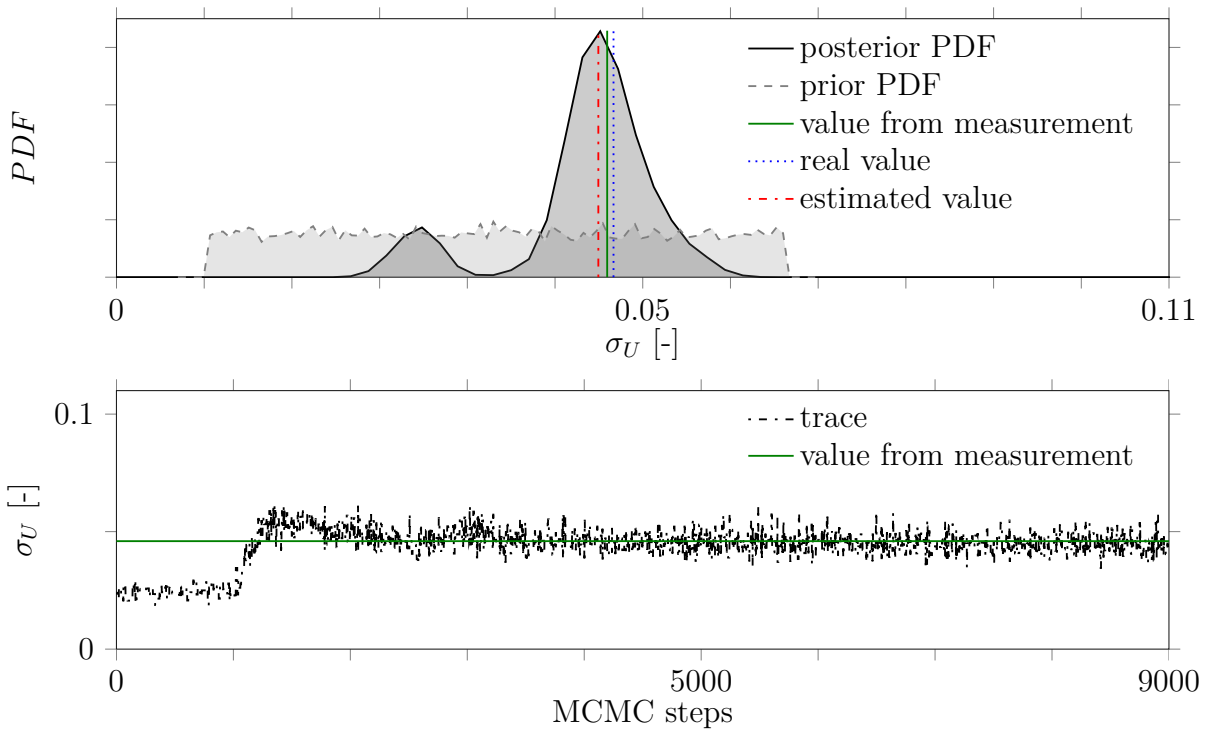



Figure 7: This figure shows the result of the MCMC estimation. The first plot contains the prior and posterior PDF of the hyperparameter $\sigma_U$. The second plot visualize the trace.

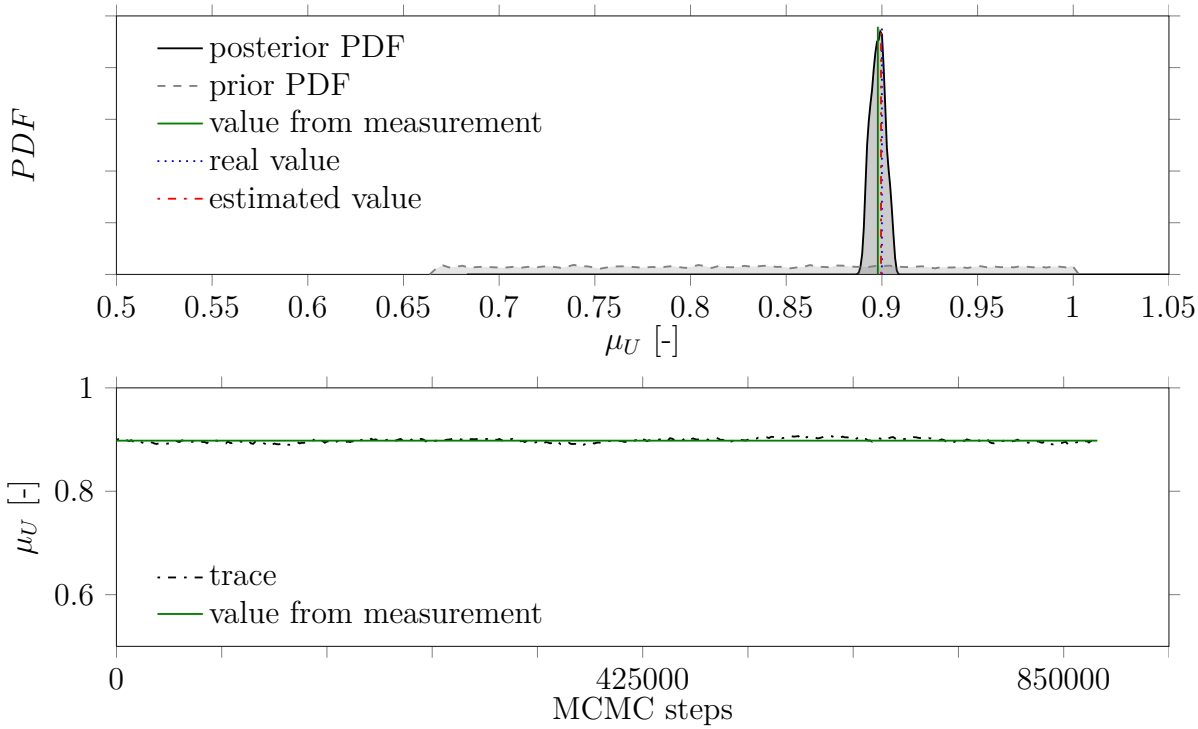



Figure 8: This figure shows the result of the PC surrogate approach. The first plot contains the prior and posterior PDF of the hyperparameter $\mu_U$. The second plot visualize the trace.

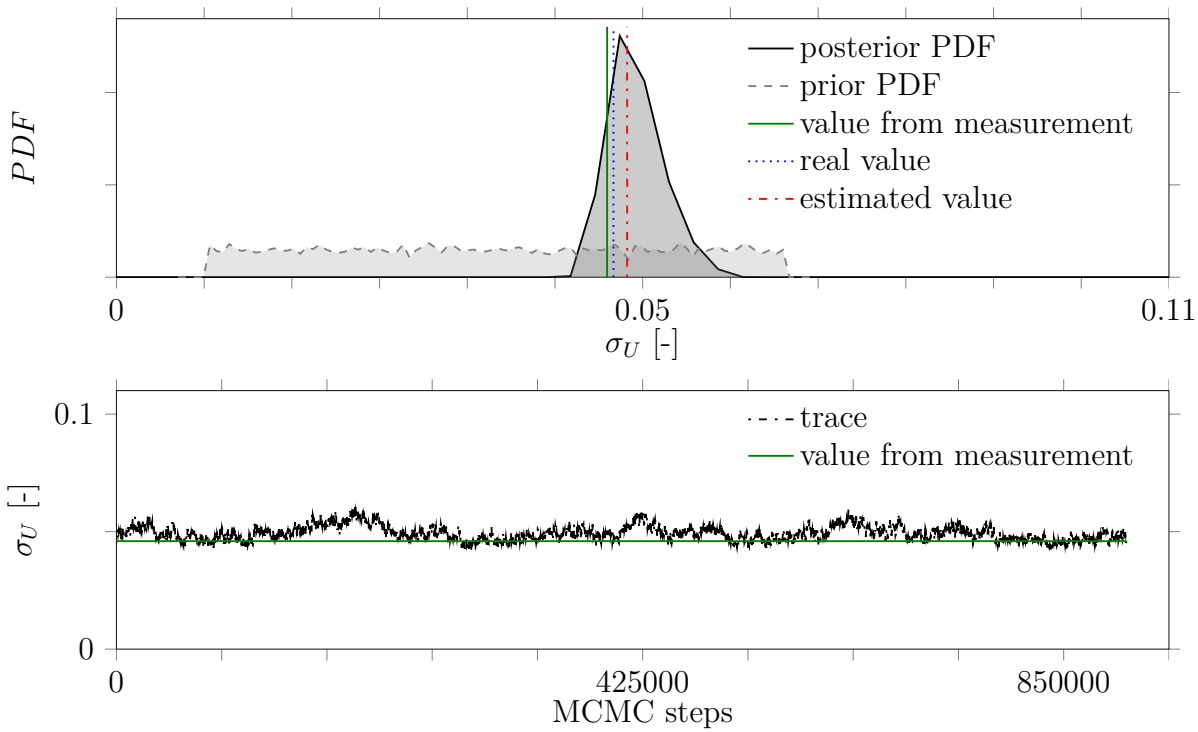



Figure 9: This figure shows the result of the PC surrogate approach. The first plot contains the prior and posterior PDF of the hyperparameter $\sigma_U$. The second plot visualize the trace.

Note that the first 10% of the MCMC runs are thrown away to give the Markov Chain time to reach its equilibrium distribution. The results of the posterior distribution using a pure MCMC approach is shown in figure 6 and figure 7. The expected values for the mean and standard derivation of the voltage are well approximated. However, we stopped the iteration at $10,000$ MCMC samples, since the involved computational cost of evaluating the forward model became prohibitive.

The results of the posterior distribution using a MCMC approach based on a PC surrogate model is shown in figure 8 and figure 9. As before the expected values are well approximated and compared to the method without a surrogate model the duration for the estimation was reduced significant. In this case a polynomial order of $p = 3$ was assumed and $Q = 111$ evaluations on the simulation model were performed. Note that increasing the polynomial order in the PC expansion did not provide any further improvement in the approximation of the mean and standard derivation. The remaining uncertainty represents the effect of the discrepancy between simulation model and reality in combination with measurement noise.

## 5 CONCLUSIONS

We presented a comparison between a PC surrogate model and a MCMC approach for inverse UQ. The used benchmark problem is based on an electric drive test bench hardware. The advantage of this setting is the possibility to define a distribution for an uncertain parameter at the system input. Based on the design of experiment measurement data can be recorded under real-world conditions. This approach gives us the ability to compare the gained results from the method with the assumed, or to put it another way, predefined distributions, which are in general not known.

We focused on the implementation of a MCMC approach, which uses the recorded data with 100 measurement runs. The estimation of the posterior distributions was done using a PC surrogate model in addition to run all evaluations on the simulation model. Due to the fact, that the time for a single simulation run is quite high and therefore the number of evaluations is limited in an industrial context. The duration, which is needed for the estimation of the posterior distributions, can be reduced significantly by using a PC surrogate model. For the pure MCMC run $1,000,000$ evaluations were needed. In contrast, 111 runs on the simulation model were performed to fit the PC surrogate model. Afterwards $100,000,000$ evaluations on the surrogate model for the MCMC approach were done. In addition, the obtained results for both approaches are satisfying.

Our current work is to investigate different types of distributions. Beside that, more uncertain parameters are take into account.

## REFERENCES

[1] Barthelmann, V., Novak, E., Ritter, K.: High dimensional polynomial interpolation on sparse grids. Advances in Computational Mathematics **12**(4), 273–288 (2000). DOI 10.1023/A:1018977404843

[2] Cameron, R.H., Martin, W.T.: The Orthogonal Development of Non-Linear Functionals in Series of Fourier-Hermite Functionals. The Annals of Mathematics **48**(2), 385–392 (1947). DOI 10.2307/1969178

[3] Coleman, T.F., Li, Y.: On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds. Mathematical Programming **67**(2), 189–224 (1994)

[4] Coleman, T.F., Li, Y.: An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. SIAM Journal on Optimization **6**(2), 418–445 (1996). DOI 10.1137/0806023

[5] Eldred, M., Webster, C., Constantine, P.: Evaluation of nonintrusive approaches for wiener-askey generalized polynomial chaos. 49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference pp. 1–22 (2008)

[6] Gelfand, A.E., Smith, A.F.M.: Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association **85**(410), 398–409 (1990)

[7] Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images **6**, 721–741 (1984)

[8] Gerstner, T., Griebel, M.: Numerical integration using sparse grids. Numerical Algorithms **18**(3/4), 209–232 (1998). DOI 10.1023/A:1019129717644

[9] Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach. Springer New York, New York, NY (1991)

[10] Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**(1), 97–109 (1970). DOI 10.1093/biomet/57.1.97

[11] Le Maître, O.P., Knio, O.M.: Spectral methods for uncertainty quantification: With applications to computational fluid dynamics. Scientific computation. Springer, Dordrecht and London (2010)

[12] Le Maître, O.P., Knio, O.M., Najm, H.N., Ghanem, R.G.: Uncertainty propagation using Wiener–Haar expansions. Journal of Computational Physics **197**(1), 28–57 (2004). DOI 10.1016/j.jcp.2003.11.033

[13] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics **21**(6), 1087 (1953). DOI 10.1063/1.1699114

[14] Novak, E., Ritter, K.: High dimensional integration of smooth functions over cubes. Numerische Mathematik **75**(1), 79–97 (1996). DOI 10.1007/s002110050231

[15] Smolyak, S.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Soviet Mathematics, Doklady **4**, 240–243 (1963)

[16] Tanner, M.A., Wong, W.H.: The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association **82**(398), 528–540 (1987). DOI 10.1080/01621459.1987.10478458

[17] The MathWorks Inc.: MATLAB and Optimization Toolbox: Release 2012b, Natick, Massachusetts, United States.

[18] Wiener, N.: The Homogeneous Chaos. American Journal of Mathematics **60**(4), 897–936 (1938). DOI 10.2307/2371268

[19] Xiu, D.: Fast numerical methods for stochastic computations: A review. Communications in Computational Physics **5**(2-4), 242–272 (2009)

[20] Xiu, D., Hesthaven, J.S.: High-Order Collocation Methods for Differential Equations with Random Inputs. SIAM Journal on Scientific Computing **27**(3), 1118–1139 (2005). DOI 10.1137/040615201

[21] Xiu, D., Karniadakis, G.E.: The Wiener-Askey Polynomial Chaos for Stochastic Differential Equations. SIAM Journal on Scientific Computing **24**(2), 619–644 (2002). DOI 10.1137/S1064827501387826

# ECCOMAS Congress 2016

## *VII European Congress on Computational Methods in Applied Sciences and Engineering*

**ECCOMAS Congress 2016**

**Proceedings of the VII European Congress on Computational Methods in Applied Sciences and Engineering**

M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (Eds)

First Edition, September 2016

ISBN: 978-618-82844-0-1



**Institute of Structural Analysis and Antiseismic Research**
**National Technical University of Athens, Greece**