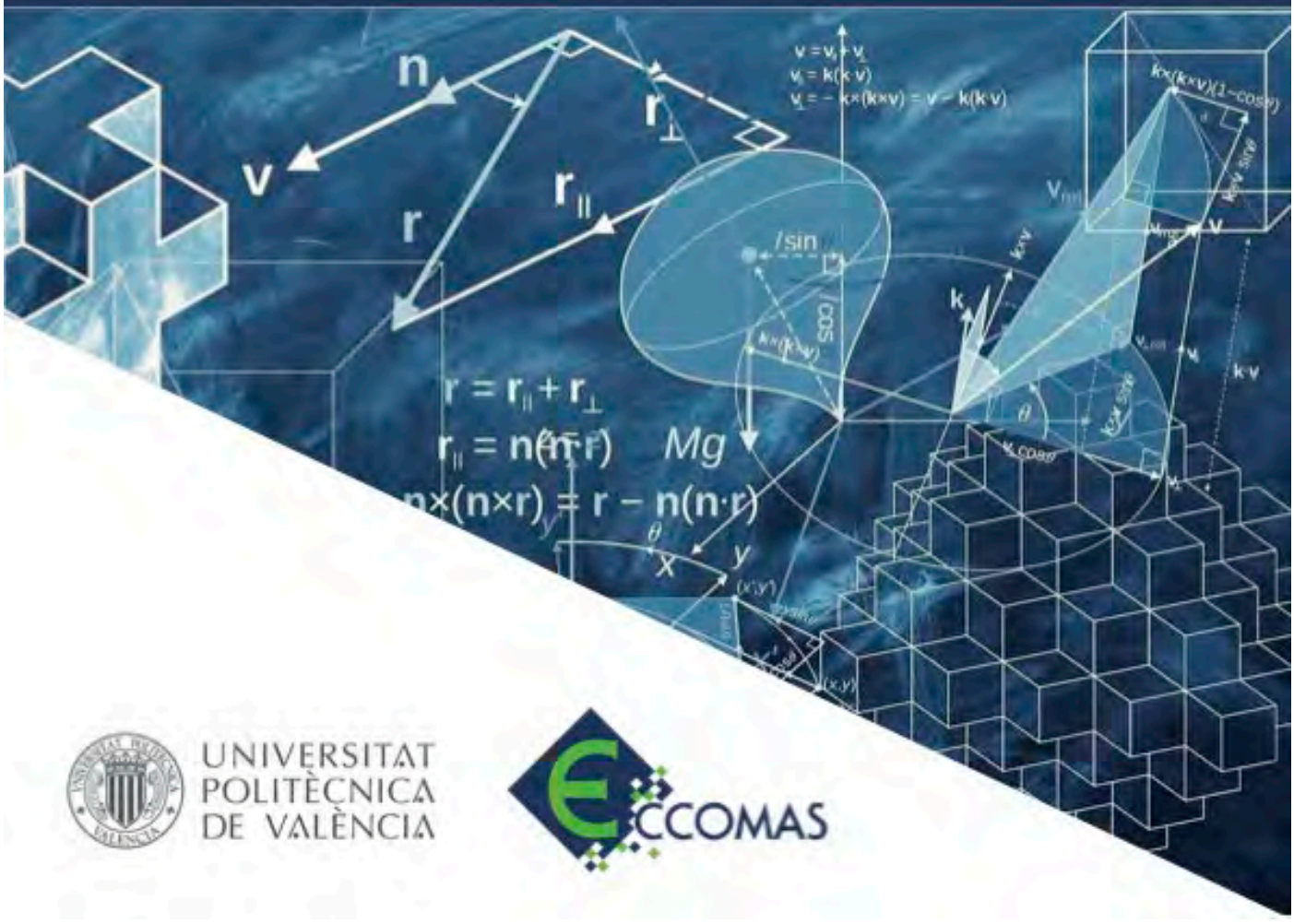
YIC 2021

BOOK OF EXTENDED ABSTRACTS

July, 7-9 2021



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ECCOMAS

Congresos UPV
Proceedings of the YIC 2021 - VI ECCOMAS Young Investigators Conference

The contents of this publication have been evaluated by the Scientific Committee which it relates and the procedure set out http://ocs.editorial.upv.es/index.php/YIC/YIC2021/about/editorialPolicies

© **Publisher**
Editorial Universitat Politècnica de València, 2021
www.lalibreria.upv.es / Ref.: 6651_01_01_01

ISBN: 978-84-9048-969-7

DOI: http://dx.doi.org/10.4995/YIC2021.2021.15320

PREFACE

The 6th ECCOMAS Young Investigators Conference YIC 2021 belongs to a series of International Conferences that have been organized since 2012 in different European cities, under the auspices of the European Community on Computational Methods in Applied Sciences (ECCOMAS). YIC 2021 has been jointly organized by the two Spanish associations of ECCOMAS: the Spanish Association for Numerical Methods in Engineering (SEMNI) and the Spanish Society of Applied Mathematics (SEMA), as a joint effort to promote the collaboration between young researchers from both organizations with other European young investigators and to reinforce the social impact of computational engineering science and computational applied mathematics. Although YIC2021 originally was going to be held at Universitat Politécnica de València (UPV), one of the leading universities in Spain, due to the pandemic situation caused by COVID-19 the conference was carried out finally in an online format from 7th to 9th July 2021.

The main goal of the YIC 2021 conference was to bring together young investigators and experienced researchers, and provide a forum for discussing the current state-of-the-art on Computational Methods and Applied Sciences. This successful combination was already seen in the Scientific Committee of the event which included distinguished professors of internationally recognized prestige and brilliant young researchers with a promising career.

The scientific program of YIC 2021 combined four plenary lectures, mainly given by outstanding young researchers, with more than two hundred presentations included in nineteen specific minisymposium sessions, proposed by the participants, and one more general session of contributed talks. In addition, the ECCOMAS PhD Olympiad also took place during the celebration of the conference. This latter consisted of scientific presentations of some of the finalists of the ECCOMAS PhD Thesis Awards nominated by each ECCOMAS member association, and resulted in two prizes for the best presentations. In occasion of the YIC 2021 conference, the ECCOMAS Young Investigators Committee (EYIC) also organized different activities for the participants of the conference. In particular, these activities comprised three short introductory courses on recent research topics, a Career Forum to discuss issues regarding pursuing an academic career, and a panel discussion on European research funding opportunities. Finally, to encourage the interaction between the participants and to promote a pleasant virtual environment, several dynamization activities were proposed to enjoy all together.

We were delighted to welcome all participants of the 6th ECCOMAS Young Investigators Conference YIC 2021, and we hope that you found in this meeting valuable information and inspiring ideas for your scientific work. We would like to take the opportunity to thank you for taking part in the YIC2021 Conference, that has been a success thanks to your interesting contributions and your active participation.

On behalf of the Conference Organizers

Enrique Nadal Soriano, UPV, Spain
Carmen Rodrigo Cardiel, UZ, Spain
José Martínez Casas, UPV, Spain

ORGANIZATION

CONFERENCE ORGANIZERS

ENRIQUE NADAL SORIANO

Enrique Nadal is member of the European Committee of Young Investigators of ECCOMAS (EYIC) since March 2019. PhD in Mechanical Engineering at Polytechnic University of Valencia in 2014. After the PhD, he went to École Centrale de Nantes for a 2-years postdoctoral research stay. Since September 2015 he is Associate Professor at Polytechnic University of Valencia. His main area of research covers different branches like error estimators, topology optimization (in solid mechanics and acoustics), fictitious domain methods for image-based simulation, reduced order modeling and domain decomposition techniques for the Proper Generalized Decomposition.

CARMEN RODRIGO CARDIEL

Carmen Rodrigo obtained her PhD in Applied Mathematics at the University of Zaragoza in 2010. Her PhD was also selected by SEMA (Spanish Society of Applied Mathematics) as finalist of the 2010 ECCOMAS PhD Award. Currently, she has an Associate Professor position at the Department of Applied Mathematics at the University of Zaragoza. Dr. Rodrigo's main area of research is numerical methods for partial differential equations, primarily the development and analysis of iterative methods for the solution of the systems of algebraic equations that are obtained after discretization. Her research interests include the study of flow problems in rigid, deformable and fractured porous media, with an emphasis on stable discretizations and efficient solvers for this type of problems. Due to her research career, in 2015 she was honored with the SEMA Award «Antonio Valle» to Young Researchers. Since 2017 she is Group Member of the European Committee of Young Investigators of ECCOMAS (EYIC).

JOSÉ MARTÍNEZ CASAS

José Martínez-Casas obtained his PhD in Mechanical Engineering at the Polytechnic University of Valencia in 2013 where, nowadays, he holds an Associate Professor position at the Department of Mechanical and Materials Engineering. The main area of research of Dr. Martínez-Casas is railway dynamics. More particularly, his main efforts are devoted to dynamic modelling of car body and bogie in the low frequency domain, development of the dynamic models for flexible and rotatory wheelset and flexible railway track in the high frequency domain, railway vehicle-track interaction modelling, as well as the study of different phenomena like rail corrugation, rolling and squeal noise.

ORGANIZING COMMITTEE

LOCAL ORGANIZING COMMITTEE AT UPV

- Víctor Tomás ANDRÉS RUIZ
- Ricardo BELDA GONZÁLEZ
- Javier CARBALLEIRA MORADO
- Borja FERRÁNDIZ CATALÁ
- Xavier GARCÍA ANDRÉS
- Jaime GIL ROMERO
- Juan GINER NAVARRO
- Jorge GÓMEZ BOSCH
- Santiago GREGORI VERDÚ
- María Amparo GUERRERO ALONSO
- Jorge GUTIÉRREZ GIL
- Raquel MEGÍAS DÍAZ
- David MUÑOZ PELLICER
- Héctor NAVARRO GARCÍA
- Ana María PEDROSA SÁNCHEZ
- Águeda SONSECA OLALLA

SCIENTIFIC COMMITTEE

NATIONAL SCIENTIFIC COMMITTEE

- José ALBELDA VITORIA (UPV)
- Icíar ALFARO (University of Zaragoza)
- Andrés ARRARÁS (Public University of Navarra)
- Marino ARROYO (UPC)
- Joan BAIGES (UPC)
- Miguel CERVERA (UPC)
- Tomás CHACÓN REBOLLO (University of Sevilla)
- Ramón CODINA (UPC)
- Elías CUETO (University of Zaragoza)
- Francisco David DENIA GUZMAN (UPV)
- Pedro DÍEZ (UPC)
- Rosa DONAT (University of Valencia)
- Antonio FALCÓ (CEU)
- José Ramón FERNÁNDEZ (University of Vigo)
- Francisco Javier FUENMAYOR FERNÁNDEZ (UPV)
- Francisco José GASPAR LORENZ (University of Zaragoza)
- Matteo GIACOMINI (UPC)
- Eugenio GINER MARAVILLA (UPV)
- Luca GIOVANNELLI (CIBER-BBN)
- David GONZÁLEZ (University of Zaragoza)
- Lucía HILARIO (CEU)
- Ángel LEÓN COLLADO (IMDEA)
- Cristina MONTALVO (UPM)
- Francisco MONTANS (UPM)
- Rafael MONTENEGRO (University of Las Palmas de Gran Canaria)
- Nicolás MONTES (CEU)
- Marta Covadonga MORA AGUILAR (UJI)
- Eugenio OÑATE (UPC-CIMNE)
- Carlos PARÉS (University of Málaga)
- Núria PARÉS (UPC)
- Antonio PÉREZ (UJI)
- María Ángeles PÉREZ (University of Zaragoza)
- Peregrina QUINTELA ESTÉVEZ (University of Santiago de Compostela)
- Luis RÁNDEZ GARCÍA (University of Zaragoza)
- Juan José RÓDENAS GARCÍA (UPV)
- Riccardo ROSSI (CIMNE)
- María José RUPÉREZ MORENO (UPV)
- Eva María SÁNCHEZ ORGAZ (UPV)
- Javier SEGURADO (IMDEA)
- Manuel TUR VALIENTE (UPV)
- Emilio VELASCO (Miguel Hernández University)
- Francesc VERDUGO (CIMNE)

INTERNATIONAL SCIENTIFIC COMMITTEE

- José Vicente AGUADO (ECN)
- Olivier ALLIX (LMT-CACHAN)
- Cosmin ANITESCU (WEIMAR)
- Leevi ANNALA (University of Jyväskylä)
- Albertino ARTEIRO (Porto University)
- Stéphane P.A. BORDAS (University of Luxembourg)
- Domenico BORZACCHIELLO (ECN)
- Víctor CALO (Curtin University)
- Ludovic CHAMOIN (LMT-CACHAN)
- Francisco CHINESTA (ENSAM)
- Susanne CLAUS (ONERA)
- William M. COOMBS (Durham University)
- Alexey DUBEN (Keldysh Institute of Applied Mathematics of RAS)
- Stefanie ELGETI (Vienna University of Technology)

- Antonio GIL ANDRADE-CAMPOS (Aveiro University)
- Héctor GÓMEZ (Purdue University)
- Xiaozhe HU (TUFTS University)
- Barry KOREN (Eindhoven University of Technology)
- Kundan KUMAR (Karlstad University)
- Pierre LADEVÈZE (LMT-CACHAN)
- Antonia LARESE (Universita' degli Studi de Padova)
- Leonardo LEONETTI (Universita' della Calabria)
- Carina NISTERS (Duisburg-Essen University)
- Jan Martin NORDBOTTEN (University of Bergen)
- Nicolas MOËS (ECN)
- Jose Paulo MOITIHNO DE ALMEIDA (Instituto Superior Técnico – Lisbon)
- Simone MORGANTI (University of Pavia)
- Jose Manuel NAVARRO (ECN)
- David NÉRON (LMT-CACHAN)
- Bastian OESTERLE (Stuttgart University)
- Simona PEROTTO (Politecnico of Milano)
- Konrad PERZYNSKI (AGH)
- Alexander POPP (UniBw Munich)
- Serge PRUDHOMME (Polytechnique Montréal)
- Timon RABCZUK (WEIMAR)
- Ernst RANK (Technical University of Munich)
- Alessandro REALI (University of Pavia)
- Gianluigi ROZZA (SISSA)
- Ulrich RÜDE (Friedrich-Alexander-University of Erlangen-Nürnberg)
- Marcus RÜTER (UCLA)
- Rubén SEVILLA (Swansea)
- Ole SIGMUND (Technical University of Denmark)
- Lorenzo TAMELLINI (CNR-IMATI Pavia)
- Deepesh TOSHNIWAL (TU Delft)
- Karen VEROY-GREPL (Aachen University)

CONTENTS

BONE TISSUE CHARACTERIZATION AND ITS STRUCTURAL SIMULATION

ADVANCED DISCRETIZATIONS AND SOLVERS FOR COUPLED SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS

RECENT ADVANCES IN SPACE-TIME METHODS

MATHEMATICAL MODELING OF COMPLEX SURFACE PROPERTIES

NOVEL COMPUTATIONAL METHODS FOR DESIGN, MODELLING, AND HOMOGENIZATION OF METAMATERIALS AND FUNCTIONAL SMART MATERIALS

NOVEL NUMERICAL METHODS FOR FLUID-STRUCTURE INTERACTION PROBLEMS

NUMERICAL METHODS FOR CHARACTERIZATION OF RAILWAY DYNAMICS AND VIBRO-ACOUSTICS

RECENT ADVANCES IN STABILISED METHODS FOR FLOW PROBLEMS

FLOW AND MECHANICS IN POROUS MEDIA

CHALLENGES IN SEA ICE MODELING, HIGH-RESOLUTION SIMULATION AND VALIDATION

MODEL REDUCTION AND ARTIFICIAL INTELLIGENCE TECHNIQUES FOR SURROGATE AND DATA-ASSISTED MODELS IN COMPUTATIONAL ENGINEERING

PHD OLYMPIADS

CONTRIBUTED SESSION

A coupled lattice Boltzmann/finite volume method for turbulent gas-liquid bubbly flows

Daniel Lauwers[*], **Matthias Meinke**[*,†] **and Wolfgang Schröder**[*,†]

[*] Chair of Fluid Mechanics and Institute of Aerodynamics
RWTH Aachen University
Wüllnerstr. 5a, 52062 Aachen, Germany
e-mail: {d.lauwers, m.meinke, office}@aia.rwth-aachen.de

[†] JARA Center for Simulation and Data Science
RWTH Aachen University
Seffenter Weg 23, 52074 Aachen, Germany

Key words: Eulerian-Eulerian model, bubbly flow, lattice Boltzmann, finite volume, LES

Abstract: *A simulation method for large eddy simulations (LES) of dispersed gas-liquid bubbly flows based on an Eulerian-Eulerian (E-E) model is presented. A volume averaging approach is used resulting in a set of conservation equations for each phase. The liquid phase is predicted using a lattice Boltzmann method, while the gas phase is modeled by a finite volume method. Interface terms between the phases result in a two-way coupled system. Both methods are formulated on a shared Cartesian grid similar to the concept in [1], which facilitates the exchange of coupling terms between the two solvers and an efficient implementation on high-performance computing (HPC) hardware. This coupled multiphase approach combines the advantages of the lattice Boltzmann (LB) method as an efficient prediction tool for low Mach number flows with those of a finite volume method used for the modeling of the phase with larger density changes by solving the Navier-Stokes equations. To accurately model the turbulent motion of the liquid phase on all resolved scales, a cumulant-based collision step for the lattice Boltzmann method [2] is combined with a Smagorinsky sub-grid-scale turbulence model. In the finite volume solver, the effect of the sub-grid-scale turbulence is incorporated according to the MILES approach. For the validation of the new method, large-eddy simulations of turbulent bubbly flows are performed. The accuracy of the predictions is evaluated comparing the results to experimental reference data for a generic test case, for which good agreement is found. The applicability of the method will be demonstrated for a bubbly turbulent channel flow, which mimics the phenomena in the electrochemical machining (ECM) process.*

1 INTRODUCTION

The study of gas-liquid multiphase flows has been an active research topic for many decades. They occur in processes belonging to industries including chemical, pharmaceutical, food, energy, and machinery industries. The quest for an improved design of these processes, generates an increasing demand for the accurate prediction and detailed analysis of such two-phase flows. Multiphase gas-liquid flows can be classified in many categories, mainly depending on the gas-liquid volume ratio and the bubble size. Here, we consider a dispersed phase in a carrier phase, such as small gas bubbles in liquids or liquid droplets in a gas.

The technical application for which the current method is developed, is an electrochemical machining (ECM) process, in which gas bubbles are generated in a liquid electrolyte during the electrochemical removal of material. Since the local removal rate of material depends on the electrical current and the gas is non-conductive, the local gas concentration greatly influences the process speed and the geometry of the final workpiece. Therefore, a detailed prediction of the highly unsteady gas transport phenomena in the electrolyte flow is important for a good design of the ECM process. The simulation results can be used to identify process parameters

https://doi.org/10.4995/YIC2021.2021.12211

and a tool geometry for an improved accuracy of the workpiece geometry without deviations caused by, e.g., local gas agglomerations.

The turbulence modeling in E-E models of gas-liquid bubbly flows can be based on the Reynolds-averaged Navier-Stokes (RANS) or large eddy simulation (LES) approaches. In the LES, the larger turbulent scales are fully resolved, whereas in RANS methods all turbulent motions are filtered from the flow field and modeled by additional transport equations. This leads to less strict spatial resolution requirements compared to LES but limits the level of details that can be predicted. In LES, the turbulent motion is explicitly resolved up to a certain spatial filter width, which is usually on the order of the local spatial step. LES typically provides more accurate results where the assumptions of the RANS models do not hold, i.e., typically in flows with strong streamline curvature, adverse pressure gradients or in large scale vertical motion in separated flow regions or wakes. Since such flows might occur in the ECM process, an LES method should be used.

In the following, a brief overview over existing E-E modeling approaches for gas-liquid bubbly flow is given. Early examples of LES methods for E-E models are the works of Milelli et al. [3] and Deen et al. [4]. In both cases, the E-E model is implemented by extending the commercial ANSYS CFX code, that is based on the finite volume (FV) method. Many authors improved on these works by studying certain aspects of the earlier models in more detail [5]. In particular, the interface forces between the phases and the sub-grid scale modeling of turbulence in the liquid phase are discussed. Recent studies include [6], [7], and [8], in which different solver types based on a finite volume formulation are used for the two sets of conservation equations. While the finite volume method is capable of producing excellent results, it is computationally more expensive compared to the lattice Boltzmann (LB) method. The LB method has been successfully used for direct numerical simulations (DNS) of gas-liquid flows [9]. DNS, however, becomes too expensive for dispersed bubbles in high Reynolds number flow. Sungkorn et al. [10] performed LES of bubble columns using the Eulerian-Lagrangian (E-L) approach with the LB method. Recently, an E-E model solved with coupled FV and LB solvers was presented by Shu et al. [11], which was, however, based on the RANS approach and which used a coupling strategy of the solvers different from that presented in this study.

This paper is organized as follows. The details of the method will be described in Sections 2 and 3. Validation results for the method are shown in Section 4. Finally, the application to an ECM setup is shown in Section 5, where results of the LES of turbulent gas-liquid channel flows, similar to the electrolyte flow in the ECM process are presented.

2 PHYSICAL MODELING

The phase averaged Eulerian-Eulerian conservation equations for mass and momentum are given by [12, 5]

$$\frac{\partial \alpha_k \bar{\bar{\rho}}_k}{\partial t} + \nabla \cdot (\alpha_k \bar{\bar{\rho}}_k \hat{\boldsymbol{v}}_k) = \Gamma_k \tag{1}$$

$$\frac{\partial \alpha_k \bar{\bar{\rho}}_k \hat{\boldsymbol{v}}_k}{\partial t} + \nabla \cdot (\alpha_k \bar{\bar{\rho}}_k \hat{\boldsymbol{v}}_k \hat{\boldsymbol{v}}_k) = -\alpha_k \nabla \hat{p} - \nabla \cdot (\alpha_k \hat{\boldsymbol{\tau}}_k) + \alpha_k \bar{\bar{\rho}}_k \boldsymbol{g} + \boldsymbol{M}_k \ . \tag{2}$$

The quantity k represents the gas or liquid component of the fluid, i.e., g for gas and l for liquid, and α_k is the local void fraction of the phase, which represents the probability of finding the corresponding phase at a certain location in time and space. For the gas-liquid flow, the condition $\alpha_g + \alpha_l = 1$ holds. The other variables are the density ρ, velocity v, pressure p, viscous fluxes $\boldsymbol{\tau}$, and the gravity vector $\boldsymbol{g}$. The mass and momentum transfer terms between the two phases are denoted Γ and $\boldsymbol{M}$. The symbols $\bar{\bar{\cdot}}$ and $\hat{\cdot}$ define the phase averaging and mass weighted averaging operators [12].

2.1 Model simplifications

The Eulerian-Eulerian model equations Eq. (1) and (2) can become stiff and difficult to solve due to possibly large two-way coupling terms resulting from the interfacial force terms, which are discussed in Section 2.2, and the influences of the coupled void fractions α_k. Some models avoid these difficulties by completely neglecting the influence of the dispersed phase on the liquid phase. These one-way coupled models are valid only for cases, where the motion of the liquid phase is dominated by external forces and the volume fraction of the gas phase is very low. A more accurate approach can be derived for the liquid phase via the mixture balance equations [13, 12]. Exploiting the fact that the density difference between gas and liquid is large, the influence of the void fraction α_k can be removed from all terms of the liquid momentum equation for moderate void fractions except for the gravity term. This simplification is similar to the Boussinesq approximation for the single-phase momentum balance equations. In that case, changes in density due to temperature changes are also neglected in all terms except for the gravity term. Additionally, no mass transfer between the phases is present in our case, hence $\Gamma_g = \Gamma_l = 0$. This leads to the following conservation equations for the liquid phase

$$\frac{\partial \bar{\bar{\rho}}_l}{\partial t} + \nabla \cdot (\bar{\bar{\rho}}_l \hat{\boldsymbol{v}}_l) = 0 \tag{3}$$

$$\frac{\partial \bar{\bar{\rho}}_l \hat{\boldsymbol{v}}_l}{\partial t} + \nabla \cdot (\bar{\bar{\rho}}_l \hat{\boldsymbol{v}}_l \hat{\boldsymbol{v}}_l) = -\nabla \hat{p} - \nabla \cdot \hat{\boldsymbol{\tau}}_l + \alpha_l \bar{\bar{\rho}}_l \boldsymbol{g} \ . \tag{4}$$

These equations only differ from their single-phase counterparts by the last term of the momentum equation. This makes it possible to use a standard solution procedure for a single phase with only minor changes for the solution of the liquid phase flow.
The balance equations for the gas phase resemble the base E-E equations Eq. (1) and (2) much more closely. Like in [13], an additional diffusive term is added to the mass equation of the gas phase to model the bubble path dispersion. This diffusion effect is caused by the interaction of bubbles with the turbulent wake of other bubbles. This effect causes a diverging flow pattern of the ascending bubbles in locally aerated bubble columns. It can also be described using a drifting velocity $\boldsymbol{v}_{drift}$ that reads [13]

$$\boldsymbol{v}_{drift} = -\frac{1}{Sc} \frac{\mu_{l,turb}}{\bar{\bar{\rho}}_l} \boldsymbol{I} \cdot \frac{1}{\alpha_g} \nabla \alpha_g \ . \tag{5}$$

The Schmidt number Sc is usually assumed to be $Sc = 1$ for the bubble path dispersion. The quantity $\mu_{l,turb}$ is the turbulent viscosity of the liquid phase that is representative for the wake of the bubbles. The resulting conservation equations of the gas phase are

$$\frac{\partial \alpha_g \bar{\bar{\rho}}_g}{\partial t} + \nabla \cdot (\alpha_g \bar{\bar{\rho}}_g \hat{\boldsymbol{v}}_g) = \frac{1}{Sc} \nabla \cdot (\mu_{l,turb} \frac{\bar{\bar{\rho}}_g}{\bar{\bar{\rho}}_l} \nabla \alpha_g) \tag{6}$$

$$\frac{\partial \alpha_g \bar{\bar{\rho}}_g \hat{\boldsymbol{v}}_g}{\partial t} + \nabla \cdot (\alpha_g \bar{\bar{\rho}}_g \hat{\boldsymbol{v}}_g \hat{\boldsymbol{v}}_g) = -\alpha_g \nabla \hat{p} - \nabla \cdot (\alpha_g \hat{\boldsymbol{\tau}}_g) + \alpha_g \bar{\bar{\rho}}_g \boldsymbol{g} + \boldsymbol{M}_g \ . \tag{7}$$

2.2 Interfacial forces

The interfacial forces represent the forces that the individual bubbles experience during the movement through the liquid phase. The relevant forces are the drag force $\boldsymbol{F}_D$, the lift force $\boldsymbol{F}_L$, the virtual mass force $\boldsymbol{F}_{VM}$, and the turbulent dispersion force $\boldsymbol{F}_{TD}$ [8]

$$\boldsymbol{M}_g = \boldsymbol{F}_D + \boldsymbol{F}_L + \boldsymbol{F}_{VM} + \boldsymbol{F}_{TD} \ . \tag{8}$$

The corresponding reaction forces on the liquid phase are disregarded in the liquid momentum equation Eq. (4), due to the large difference in density between the phases. For the closure of the force terms, the equations described in [8] are used.
The drag force is modeled as

$$\boldsymbol{F}_D = \frac{3}{4}\alpha_g C_D \frac{\bar{\bar{\rho}}_l}{d_B}|\hat{\boldsymbol{v}}_l - \hat{\boldsymbol{v}}_g|(\hat{\boldsymbol{v}}_l - \hat{\boldsymbol{v}}_g) \tag{9}$$

with the bubble diameter d_B and the coefficient of drag C_D. For the distorted bubble regime, the drag coefficient C_D can be estimated [12]

$$C_D = \frac{2}{3}\sqrt{Eo} \tag{10}$$

with the Eötvös number $Eo = |\boldsymbol{g}|(\rho_l - \rho_g)d_B^2/\sigma$. The quantity σ represents the surface tension of the liquid phase. The distorted bubble regime is applicable for the bubble column in Section 4. For the simulation in Section 5, Stokes' drag law is assumed since the bubbles are sufficiently small to be considered spherical [12]

$$C_D = \frac{24}{Re_B} = \frac{24\ \nu_l}{|\hat{\boldsymbol{v}}_l - \hat{\boldsymbol{v}}_g|\ d_B}\ . \tag{11}$$

The lift force is modeled by

$$\boldsymbol{F}_L = C_L \alpha_g \bar{\bar{\rho}}_l (\hat{\boldsymbol{v}}_g - \hat{\boldsymbol{v}}_l) \times (\nabla \times \hat{\boldsymbol{v}}_l)\ . \tag{12}$$

The correct coefficient of lift C_L for bubbly flows has been a controversial topic. In the review paper [5], the values range from -0.05 to 0.5. In principle, the value and sign of the lift force highly depends on the shape of the bubble and the flow conditions around it. A widely adopted variable model for the coefficient of lift is the Tomiyama lift force model [14]. This leads to $C_L = 0.288$ for the validation case studied in this work, when the Morton number is extrapolated for an air-water mixture. Tomiyama's model, however, was obtained for single bubbles under static shear flow conditions, that differ from turbulent dispersed bubbly flows. In recent studies, Shu et al. [11] have shown that simulations of buoyancy driven bubbly flows can be performed without lift force. Because of the uncertainties regarding the value of the lift coefficient for the different flow regimes studied in this work, the lift force is taken here as zero.
The virtual mass force is modeled by

$$\boldsymbol{F}_{VM} = C_{VM}\alpha_g \bar{\bar{\rho}}_l \left[\left(\frac{\partial \hat{\boldsymbol{v}}_l}{\partial t} + (\hat{\boldsymbol{v}}_l \cdot \nabla)\hat{\boldsymbol{v}}_l\right) - \left(\frac{\partial \hat{\boldsymbol{v}}_g}{\partial t} + (\hat{\boldsymbol{v}}_g \cdot \nabla)\hat{\boldsymbol{v}}_g\right)\right] \tag{13}$$

with the virtual mass coefficient $C_{VM} = 0.5$.
The turbulent dispersion force is

$$\boldsymbol{F}_{TD} = -\frac{3}{4}\frac{C_D}{Sc}\frac{\mu_{l,eff}}{d_B}|\hat{\boldsymbol{v}}_l - \hat{\boldsymbol{v}}_g|\nabla\alpha_g \tag{14}$$

with the effective liquid viscosity $\mu_{l,eff}$ (see Section 2.3).

2.3 Turbulence modeling

Since in LES not all the scales of turbulent motion are resolved in the continuous phase, the influence of the sub-grid scale (SGS) motions have to be modeled. In this work, a Smagorinsky SGS model is chosen due to its previous successful application to bubble columns [5]. In addition

to the influence of the SGS turbulence, bubble induced turbulence (BIT) must be accounted for in the simulation of bubbly flows. Thus, the effective viscosity of the liquid phase is the sum of the molecular viscosity, the viscosity due to the SGS of turbulence and the BIT contribution

$$\mu_{l,eff} = \mu_{l,M} + \mu_{l,SGS} + \mu_{l,BIT} \ . \tag{15}$$

The equation for the SGS viscosity μ_{SGS} is

$$\mu_{l,SGS} = \bar{\bar{\rho}}_l (C_S \Delta)^2 \sqrt{2 \boldsymbol{S}_{ij} \boldsymbol{S}_{ij}} \tag{16}$$

as a function of the Smagorinsky constant C_S, the rate-of-strain tensor $\boldsymbol{S}_{ij}$, and the grid filter width Δ. In this study, $C_S = 0.1$ is used for all simulations.
For the BIT viscosity, the model from [3, 8] is used

$$\mu_{l,BIT} = C_S \Delta \bar{\bar{\rho}}_l \alpha_g |\hat{\boldsymbol{v}}_g - \hat{\boldsymbol{v}}_l| \ . \tag{17}$$

The influence of the turbulent motion in the gas phase is incorporated according to the MILES approach [15].

3 NUMERICAL METHOD

The Eulerian-Eulerian model described in Section 2 is implemented in the multiphysics solver mAIA – formerly denoted ZFS – developed by the Institute of Aerodynamics of RWTH Aachen University [16]. One of the strengths of mAIA is the coupling concept that allows multiple solvers of different type to share simulation data in a common data structure. In this study, a lattice Boltzmann (LB) solver representing the liquid phase is coupled to a finite volume (FV) solver for the gas phase. This approach combines the advantages of the LB method as an efficient prediction tool for low Mach number flows with those of the FV method for the phase with higher density changes. Both solvers are implemented for the efficient parallel execution on HPC hardware, which enables the simulation of problems requiring a large number of mesh cells as the simulations discussed in Sections 4 and 5. Both methods are discretized on hierarchical Cartesian meshes, which allows a straightforward local mesh refinement with dynamic load balancing. For the generic cases simulated in this paper, however, local grid refinement was not necessary.

3.1 Description of the solvers

The conservation equations for the liquid phase Eq. (3) and (4) are identical to their single-phase counterparts except for the buoyancy term. Therefore, only minor modifications to the single-phase LB solver are necessary. The LB method is based on the discrete Boltzmann equation with the Bhatnagar-Gross-Krook approximation [17]. The particle probability distribution functions (PDFs) are discretized in this study according to the D3Q27 model [18] in a cell-centered approach. A cumulant based collision step [2] is used that is capable of producing accurate results across a wide range of Reynolds numbers. This is important due to the highly turbulent nature of the studied gas liquid bubbly flows. In contrast to the multi-relaxation time (MRT) approach, the cumulant based collision step does not require the tuning of model parameters. The buoyancy term of the momentum equation Eq. (4) is implemented according to [19], adding the force components to the PDFs before the propagation step.

The finite volume method solving the gas conservation equations combines an advective upstream splitting method (AUSM) with the second-order accurate monotone upstream centered scheme for conservation laws (MUSCL) approach for the computation of the inviscid fluxes. The viscous fluxes are discretized by a second-order accurate centered scheme as well. Time

integration is accomplished by an explicit, low-storage 5-step Runge-Kutta method. The gas momentum equations become stiff due to the momentum exchange terms in Eq. (8). To maintain numerical stability, the explicit time-integration scheme is modified for the momentum equations. The influence of the drag, turbulent dispersion and the local term of the virtual mass forces are incorporated by an implicit Crank–Nicolson scheme, while the explicit formulation is kept for the remaining terms.

3.2 Solution procedure

The time steps of the two solvers are synchronized by using the constant time step of the LB solver also for the finite volume method. Due to the strongly two-way coupled nature of the conservation equations of the two phases, the flow solvers operate in a staggered approach in time direction. First, the liquid phase solver completes a time-step using the gas void fraction field of the previous time step. The resulting updated liquid velocity and pressure field is then transferred to the gas phase solver. Secondly, the gas solver completes its time step, consisting of the following procedure. The density distribution is updated with the liquid pressure field. The changes in gas velocity and gas void fraction are obtained by the solution of the gas mass and momentum equations in each Runge-Kutta step. Finally, the updated gas flow variables are transferred to the liquid flow solver for the next time step.

4 VALIDATION OF THE METHOD

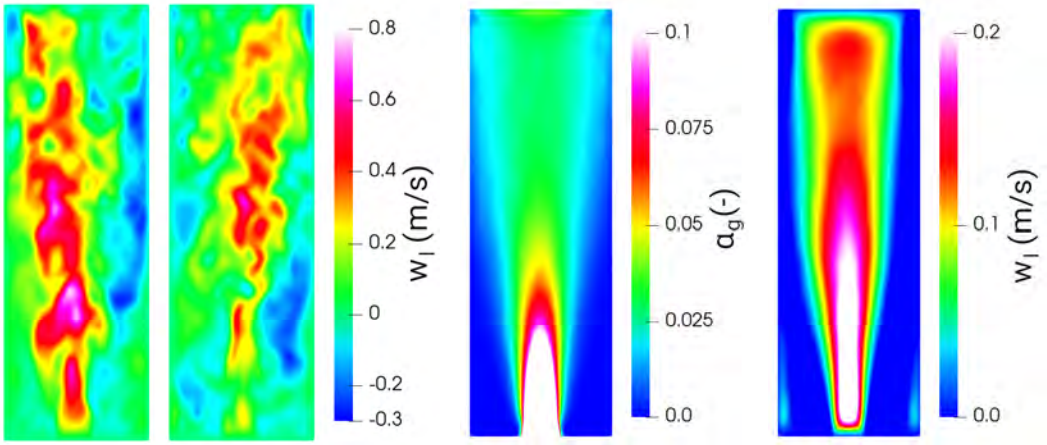
For the validation of the numerical method, a standard test case for turbulent, buoyancy driven bubbly flow is simulated. It was first studied experimentally and numerically by Deen et al. [20, 4]. Air is injected at the bottom surface into a water column with a height H of $H = 0.45$ m and a square cross-section $W \times D$ of 0.15×0.15 m^2. The bubbles enter the duct geometry through a perforated plate with 49 holes at the center of the bottom surface. The holes have a diameter of 1 mm and are arranged in a square pattern of 7×7 holes with a pitch of 6.25 mm. At the top, the water forms a free surface through which the injected air escapes. The gas velocity above the water surface is specified as 4.9×10^{-3} m s^{-1}, which leads to a gas flow rate of 1.1×10^{-4} m^3 s^{-1} at ambient pressure. The diameter of the resulting bubbles d_B is 4.0 mm [4]. This test case has been extensively studied for the validation of numerical models for bubbly flows. Recent studies that feature simulation results of this setup are [6, 7, 8, 11].

For the present simulations, a uniform, unstructured Cartesian grid with $44 \times 44 \times 128$ cells is used to discretize the bubble column. For the liquid phase, no-slip boundary conditions are applied at the column walls and bottom surface. The interpolated bounce back following the BFL rule is used for these boundaries [21]. The water surface is modeled by a slip-wall boundary condition. The gas flow rate is enforced in the inflow boundary condition. Furthermore, a no-slip condition is applied to the walls and the remaining bottom surface. For the outflow, a pressure outflow boundary condition is imposed to the top surface. The time step of the simulation is constant at 6.77×10^{-4} s.

The bubble diameter d_B is assumed to be 4 mm, bubble coalescence or break-up are not accounted for in this study. The original experiment was designed to minimize coalescence of bubbles by adding salt to the water [20]. The change in bubble diameter due to the varying hydrostatic pressure is also neglected. This approximation is justified since the increase in bubble diameter over the height of the column only amounts to about 1.4 %.

4.1 Instantaneous results

The flow field of the bubble column is mainly determined by a bubble plume meandering in the column. The liquid flow field is turbulent and strongly influenced by the location of the



(a) (b) (c)

Figure 1: Liquid axial velocity at the times $t = 67.7$ s and $t = 135.3$ s (a), average gas void fraction (b) and average liquid axial velocity (c) at the mid-plane of the column.

bubble plume. Close to the plume, the liquid moves predominantly in the upward direction due to the buoyancy of the gas phase. The axial component of the liquid velocity at the mid-plane is shown in Figure 1(a) for two time levels, $t = 67.7$ s and $t = 135.3$ s.

4.2 Time-averaged results

For the generation of time-averaged results and turbulence statistics, 100,000 time steps are performed first, to obtain a fully developed flow field before the time averaging begins. The averaging is based on an additional 400,000 time steps. This corresponds to an averaging window from 68 s to 271 s in physical time.

In Figures 1(b) and 1(c), the averaged liquid axial velocity and gas void fraction in the mid-plane of the column are displayed. In the bottom part of the column, the gas void fraction is confined to a narrow, but diverging area. Due to the movement of the bubble plume, this differs from the top part of the column, where gas can be found over the full width of the column. The highest liquid velocity values are found in the region close to the gas inlet. Near the walls, a recirculation region with negative velocity values is visible. In the bottom corners, vortices are formed in the liquid phase.

In Figures 2(a) and 2(b), the liquid and gas axial velocities at the mid-plane are plotted along a line at the height of 25 cm above the bottom of the column. The liquid axial velocity agrees well with the experimental data of Deen [4]. The gas axial velocity is somewhat underestimated, especially in the center region of the column.

In Figures 2(c) and 2(d), the root-mean-square of the fluctuating liquid axial velocity w_l' and the turbulent kinetic energy $TKE_l = 0.5\ (u_l'u_l' + v_l'v_l' + w_l'w_l')$ are plotted. The turbulent fluctuations of the liquid phase flow are slightly overestimated by the present method except for the region near the walls, where the intensity must vanish. Both graphs exhibit a dent near the center of the column, that is visible in the simulation results and the experiments. The asymmetry of the results indicates that the time averaging interval is not sufficient, which can be attributed to the highly unsteady character of the test case featuring low frequency variations from the meandering bubble column. This may also explain a part of the visible deviations of the numerical solution.

5 SIMULATIONS OF A TURBULENT CHANNEL FLOW

The presented method is applied to a generic setup of the gas-liquid flow as it occurs in the ECM process in the gap between the tool and the work piece. This multiphase flow is studied

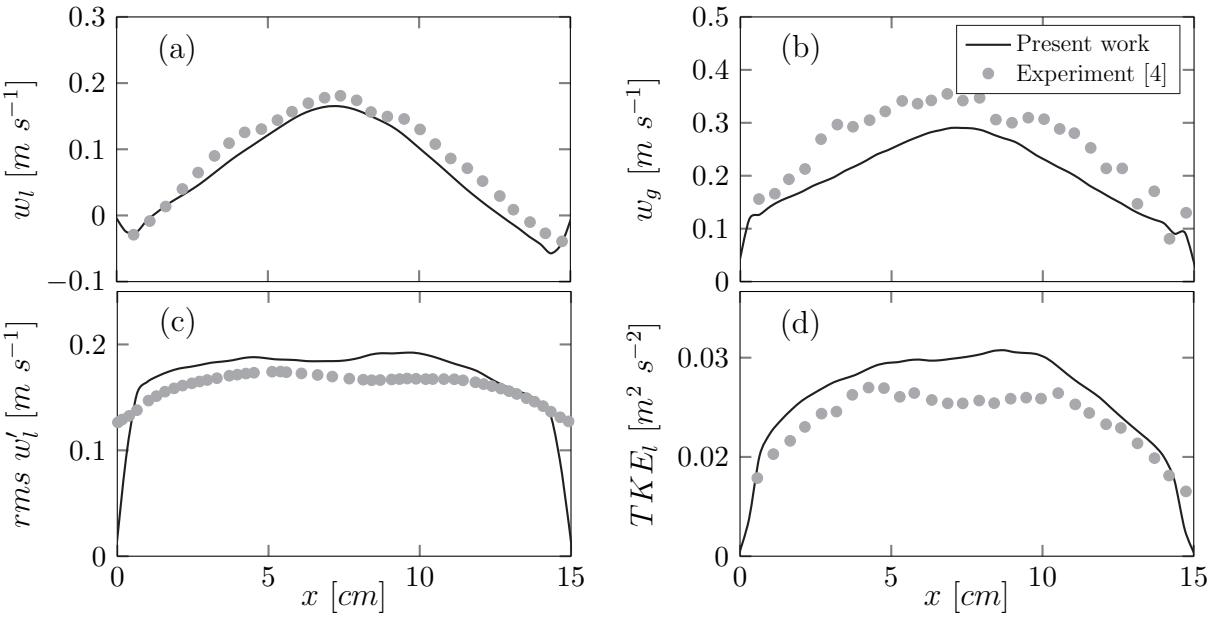


Figure 2: Averaged liquid (a) and gas (b) axial velocity, root-mean-square of the fluctuating liquid axial velocity (c) and liquid turbulent kinetic energy (d) at a height of $z = 25$ cm in comparison with measurements from [4].

experimentally and numerically in the research project described in [22]. For the experimental investigations, the flow is upscaled from a gap height of 1×10^{-4} m to a gap height of 1×10^{-2} m, keeping the Reynolds number constant. This upscaled machining gap is modeled as a turbulent channel flow with a Reynolds number of $Re = 3000$ based on the channel height and the mean flow velocity. The machining process is imitated by the introduction of gas bubbles through the bottom or top wall of the channel with a constant flow rate. As a starting condition, a fully developed turbulent channel flow is generated. The friction velocity based Reynolds number Re_τ of this single-phase flow is approximately $Re_\tau = 100$. To capture the large turbulent structures that are present in such a low Re_τ flow, a simulation domain of $1H \times 3H \times 12H$ channel heights is used in the wall normal, spanwise, and streamwise direction. The no-slip boundary condition is applied at the top and bottom boundary, and periodic conditions for the remaining boundaries. A volume force in the main flow direction is applied to keep a constant volume flow rate. The simulation domain is discretized with a uniform, Cartesian grid using $100 \times 300 \times 1200$ cells leading to a Δy^+ of 2.07. The gas flow rate per channel height unit is chosen to be 1/500 of the average liquid flow rate in the channel. The liquid is modeled as water with a density of 1000 kg m^{-3} and a kinematic viscosity of 1×10^{-6} m^2 s^{-1}. The gas is air with a density of 1.2 kg m^{-3} at atmospheric pressure and a kinematic viscosity of 1.52×10^{-5} m^2 s^{-1}. Gravity acts in the downward direction with the acceleration of 9.81 m s^{-2}. The bubble diameter is estimated to be 10^{-4} m, which is equal to the mesh cell size.

Figure 3 shows instantaneous gas void fraction fields for the gas injection from the bottom and the top of the channel. The effect of the buoyancy is clearly visible. In the case of the injection from the bottom wall, the gas is much more distributed throughout the channel. The large turbulent structures of the channel flow lead to areas of severely varying gas void fraction along the axis of the channel. In the case of the injection from the top of the channel, the buoyancy prevents the gas bubbles to be distributed throughout the channel. A nearly uniform layer of higher gas void fraction in formed at the top wall.
The averaged liquid velocity profiles in Figure 4 reflect this difference in the gas distribution. For the injection from the bottom wall, the velocity profile stays much more symmetric after the initial distribution of the bubbles. Compared to the single-phase profile, the velocity gradients

at the walls increase after the gas injection due to the increased momentum exchange normal to the walls. In the case of the injection from the top of the channel, the velocity profile becomes increasingly asymmetric. In this case, the buoyancy of the gas bubbles prevent momentum transfer in the top region of the channel enabling higher axial velocities in the top half of the channel.

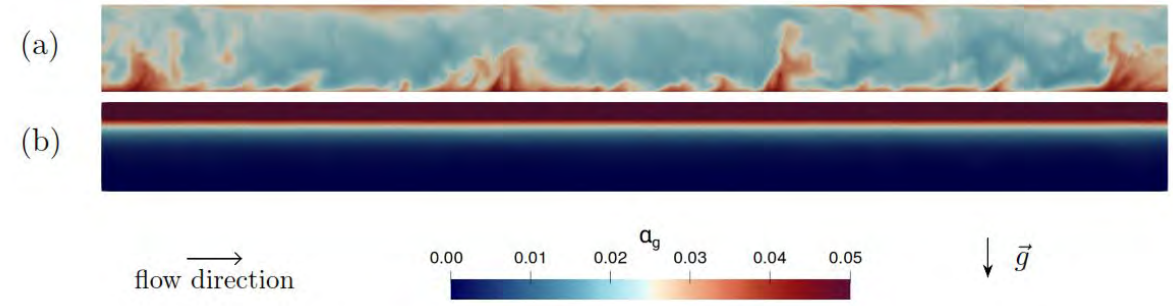


Figure 3: Instantaneous gas void fraction at t = 0.385 s after gas injection from the bottom (a) and the top (b) walls into the turbulent channel flow.

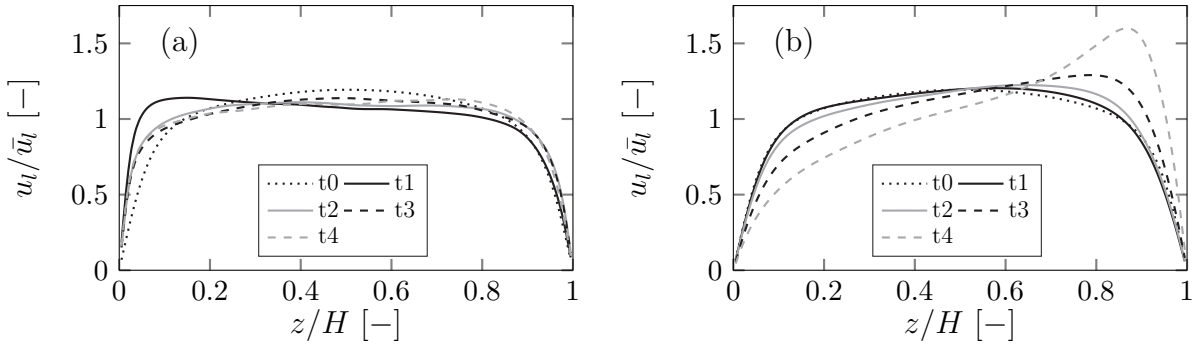


Figure 4: Averaged relative liquid velocity distribution plotted over the distance from the channel bottom wall for the times t1 = 0.192 s, t2 = 0.385 s, t3 = 0.577 s and t4 = 0.770 s after the start of the gas injection at time t0. Gas injection from the bottom wall (a) and the top wall (b).

6 CONCLUSION

A coupled lattice Boltzmann/finite volume method for the Eulerian-Eulerian simulation of gas-liquid bubbly flows is presented. The novel method is validated with LES of Deen's bubble column case. The model is capable of reproducing the meandering of the bubble plume in the column. The averaged results show good agreement with the experimental data. An engineering application of the method is shown with the simulation of a turbulent channel flow, similar to the gas-liquid electrolyte flow during the ECM process. The method predicts the expected large differences in the gas distribution depending on the location of the gas generation. Since the local gas void fraction has a strong influence on the geometry of the workpieces produced with the ECM process, the method can be used to further study the influence of process parameter variations on the resulting workpiece.

REFERENCES

[1] M. Schlottke-Lakemper, A. Niemöller, M. Meinke, and W. Schröder, "Efficient parallelization for volume-coupled multiphysics simulations on hierarchical cartesian grids," *Computer Methods in Applied Mechanics and Engineering*, vol. 352, pp. 461 – 487, 2019.

[2] M. Geier, M. Schönherr, A. Pasquali, and M. Krafczyk, "The cumulant lattice boltzmann equation in three dimensions: Theory and validation," *Comp. & Math. w. Appl.*, vol. 70, no. 4, pp. 507 – 547, 2015.

[3] M. Milelli, *A numerical analysis of confined turbulent bubble plumes*. PhD thesis, ETH Zurich, 2002.

[4] N. G. Deen, T. Solberg, and B. H. Hjertager, "Large eddy simulation of the gas–liquid flow in a square cross-sectioned bubble column," *Chem. Eng. Sci.*, vol. 56, no. 21, pp. 6341–6349, 2001.

[5] M. Dhotre, N. Deen, B. Niceno, Z. Khan, and J. Joshi, "Large eddy simulation for dispersed bubbly flows: A review," *Int. J. of Chem. Engineering*, vol. 2013, 03 2013.

[6] T. Ma, D. Lucas, T. Ziegenhein, J. Fröhlich, and N. Deen, "Scale-adaptive simulation of a square cross-sectional bubble column," *Chem. Eng. Sci.*, vol. 131, pp. 101–108, 2015.

[7] Z. Liu and B. Li, "Scale-adaptive analysis of euler-euler large eddy simulation for laboratory scale dispersed bubbly flows," *Chem. Eng. Journal*, vol. 338, pp. 465–477, 2018.

[8] M. H. Mohammadi, F. Sotiropoulos, and J. R. Brinkerhoff, "Eulerian-eulerian large eddy simulation of two-phase dilute bubbly flows," *Chem. Eng. Sci.*, vol. 208, p. 115156, 2019.

[9] A. Mühlbauer, M. W. Hlawitschka, and H.-J. Bart, "Models for the numerical simulation of bubble columns: A review," *Chemie Ing. Technik*, vol. 91, no. 12, pp. 1747–1765, 2019.

[10] R. Sungkorn, J. Derksen, and J. Khinast, "Modeling of turbulent gas-liquid bubbly flows using stochastic lagrangian model and lattice-boltzmann scheme," *Chem. Eng. Sci.*, vol. 66, pp. 2745–2757, 06 2011.

[11] S. Shu, N. Yang, F. Bertrand, and J. Chaouki, "High-resolution simulation of oscillating bubble plumes in a square cross-sectioned bubble column with an unsteady k-ε model," *Chem. Eng. Sci.*, 11 2020.

[12] M. Ishii and T. Hibiki, *Thermo-fluid dynamics of two-phase flow*. New York: Springer, second ed., 2011.

[13] A. Sokolichin, G. Eigenberger, and A. Lapin, "Simulation of buoyancy driven bubbly flow: Established simplifications and open questions," *AIChE Journal*, vol. 50, no. 1, pp. 24–45, 2004.

[14] A. Tomiyama, H. Tamai, I. Zun, and S. Hosokawa, "Transverse migration of single bubbles in simple shear flows," *Chem. Eng. Sci.*, vol. 57, no. 11, pp. 1849–1858, 2002.

[15] J. P. Boris, F. F. Grinstein, E. S. Oran, and R. L. Kolbe, "New insights into large eddy simulation," *Fluid Dynamics Research*, vol. 10, pp. 199–228, dec 1992.

[16] A. Lintermann, M. Meinke, and W. Schröder, "Zonal flow solver (ZFS): a highly efficient multi-physics simulation framework," *Int. J. of Comp. Fluid Dynamics*, vol. 34, no. 7-8, pp. 458–485, 2020.

[17] P. L. Bhatnagar, E. P. Gross, and M. Krook, "A model for collision processes in gases. I. small amplitude processes in charged and neutral one-component systems," *Phys. Rev.*, vol. 94, pp. 511–525, May 1954.

[18] Y. H. Qian, D. D'Humières, and P. Lallemand, "Lattice BGK models for navier-stokes equation," *Europhysics Letters (EPL)*, vol. 17, pp. 479–484, feb 1992.

[19] D. Hänel, *Molekulare Gasdynamik: Einführung in die kinetische Theorie der Gase und Lattice-Boltzmann-Methoden*. Berlin Heidelberg: Springer, 2004.

[20] N. G. Deen, B. H. Hjertager, and T. Solberg, "Comparison of piv and lda measurement methods applied to the gas-liquid flow in a bubble column," in *10th Int. Symp. on Appl. of Laser Techniques to Fluid Mech, Lisbon, Portugal, 9-13 July 2000*, 2000.

[21] M. Bouzidi, M. Firdaouss, and P. Lallemand, "Momentum transfer of a boltzmann-lattice fluid with boundaries," *Physics of Fluids*, vol. 13, no. 11, pp. 3452–3459, 2001.

[22] B. Rommes, D. Lauwers, T. Herrig, M. Meinke, W. Schröder, and A. Klink, "Concept for the experimental and numerical study of fluid-structure interaction and gas transport in precise electrochemical machining," in *18th CIRP CMMO, Ljubljana, Slovenia, 15-17 June*, 2021.

Shape Optimization for Thermal Insulation Problems

S. Tozza* **and G. Toraldo†**

* Department of Mathematics and Applications "Renato Caccioppoli"
University of Naples Federico II
Naples, Italy
e-mail: silvia.tozza@unina.it

† Department of Mathematics and Physics
University of Campania "Luigi Vanvitelli"
Caserta, Italy
e-mail: gerardo.toraldo@unicampania.it

Key words: Elliptic PDEs, shape optimization, thermal insulation, Robin-Dirichlet boundary conditions, Finite Element Method, heat dispersion

Abstract: *In this work we consider two domains: an external domain whose geometry varies, and an internal fixed one. From the thermal insulation viewpoint, we are considering a body to be insulated, enveloped in a layer of insulator, and we want to find the best shape for the thermal insulator, in terms of heat dispersion. Mathematically, our problem is described by an elliptic partial differential equation with Dirichlet-Robin boundary conditions.*

1 INTRODUCTION

One of the major challenges for environmental improvement is represented by thermal insulation. Problems related to insulation are well-known and widely studied in mathematical physics. Nevertheless, mathematics involved is still hard especially when one looks at shape optimization issues [1, 2], and sometimes the answers are counterintuitive [3]. In this work we focus on the case of an internal domain of circular shape (the body to be insulated), enveloped in a layer of thermal insulator whose geometry varies. Our aim is to explore different shapes for the external domain, in order to find configurations which produce low values in terms of heat dispersion.

The work is organized as follows: In Sect. 2 we formulate our problem, explaining the peculiar behavior of the heat dispersion looking at the case of two concentric circles. In Sect. 3 we move to the numerical part, describing the numerical resolution and the results obtained, ending with final remarks and future perspectives contained in Sect. 4.

2 THE PROBLEM

Let us consider a domain Ω embedded into a domain D. The formulation of the problem we deal with is the following:

$$\begin{cases} \text{Find } D^* \in \mathcal{D} \text{ such that} \\ F_\beta(D^*, \Omega) := \min\limits_{D \in \mathcal{D}} F_\beta(D, \Omega) \\ \text{where } F_\beta(D, \Omega) := \beta \int_{\partial D} u \, dx \\ \text{and } u \text{ is solution of} \\ (PDE) \begin{cases} \Delta u = 0, & \text{in } D \setminus \Omega, \\ \dfrac{\partial u}{\partial n} + \beta u = 0, & \text{on } \partial D, \\ u = 1, & \text{on } \partial\Omega, \end{cases} \end{cases} \tag{1}$$

where $u \in H'(D \cup \Omega)$ represents the temperature, $\mathcal{D}$ is the set of admissible domains, n the exterior normal vector, and $\beta > 0$ a fixed parameter depending on the physical characteristics

https://doi.org/10.4995/YIC2021.2021.12288

of the insulating material. We fix the domain Ω as a unit circle, i.e., $\Omega := B_1(0)$, and $\mathcal{D}$ as the class of polygons, in which the domain D varies. From the thermal insulation viewpoint, the compact connected set Ω represents a conductor of constant temperature fixed to 1, which is thermally insulated by surrounding it with a layer of thermal insulator, denoted by the open set $D \setminus \Omega$, with $\Omega \subset \bar{D}$. The goal of the present work is to find configurations for the domain D which give sufficiently low values for the heat dispersion functional defined as

$$F_\beta(D, \Omega) := \beta \int_{\partial D} u \, dx, \tag{2}$$

comparing the results with the case of two concentric circles, for which we are able to compute the heat dispersion functional analytically. In fact, let us consider Ω and D as two circles of radius r and R, respectively, with $0 < r < R$. The set of solutions to the Laplace's equation in the case of a circular crown is

$$A \log \sqrt{x^2 + y^2} + C = u(x, y), \tag{3}$$

where A and C are two constants. Using for u the expression (3), the functional (2) can be written as

$$\begin{aligned} \beta \int_{\partial D} u &= \beta \int_{\partial D} A \log \sqrt{x^2 + y^2} + C & (4)\\ &= \beta \int_0^{2\pi} (A \log \left(\sqrt{R^2 \cos^2(t) + R^2 \sin^2(t)} \right) + C) R dt & (5)\\ &= \beta 2\pi R (A \log(R) + C). & (6) \end{aligned}$$

Using the boundary conditions of the partial differential equation (PDE) in (1), the constants A and C can be computed solving the following system of two equations:

$$\begin{cases} A \log r + C = 1 \\ \beta(A \log R + C) + A \frac{1}{R} = 0. \end{cases} \tag{7}$$

In that way, we get

$$A = -\frac{1}{\log(\frac{R}{r}) + \frac{1}{\beta R}}, \qquad C = 1 - A \log r = 1 + \frac{1}{\log(\frac{R}{r}) + \frac{1}{\beta R}} \log r \, . \tag{8}$$

Notice that for $r = 1$, the constant C is always equal to one, independently from the admissible values for R and β.

For fixed $\beta > 0$ and $r \in (0, R)$, the dispersion computed according to (6) only depends on R. In particular, it is an increasing function for $R < 1/\beta$, and decreasing for $R > 1/\beta$. Since must be $R > 1$, for $\beta > 1$ the dispersion is a decreasing function (the insulation increases adding insulator), whereas for $0 < \beta < 1$ the dispersion increases for $R < 1/\beta$ and decreases for $R > 1/\beta$. About the increasing phase, it may seem surprising that adding insulator increases the heat dispersion; however this is a well-known phenomenon, that from the physical viewpoint can be explained by the competing effects of the convection and the conduction resistances (see [4], Sect. 3.3.1-3.3.2).

The dependence of the dispersion function on the parameter β vanishes looking at its asymptotic behavior, as visible also in Fig. 1 (see [5] for details).

Considering a geometry which is different from the circle of radius R for the external domain D, we noticed that the qualitative behavior of the dispersion function by varying the parameter β is analogous. As an example, see the plots in Fig. 2 for a comparison between a circle and a regular octagon as external domain.

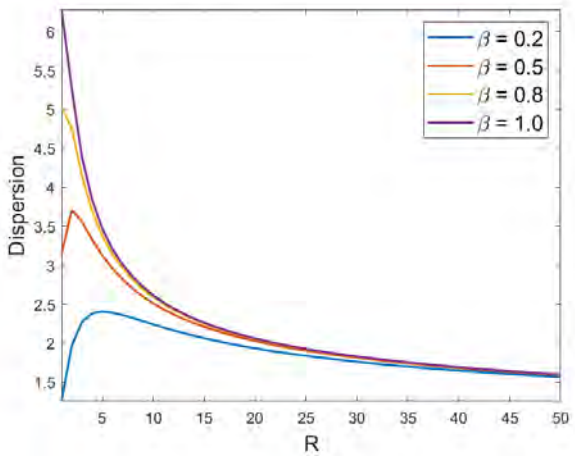


Figure 1: Plot of $(R, F_\beta(B_R(0), B_r(0)))$ for different values of β ($\beta = 0.2, 0.5, 0.8, 1$).

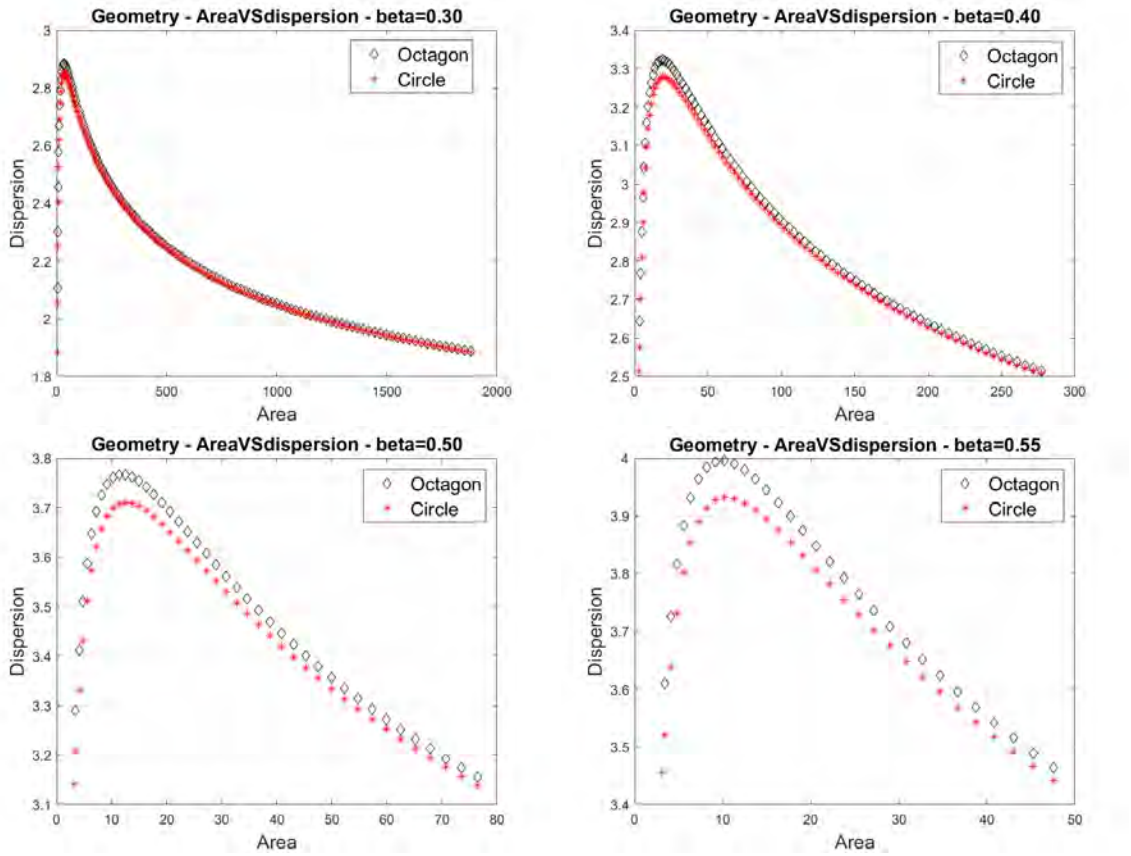


Figure 2: Plots of $(Area, F_\beta(D, \Omega))$ with $\Omega = B_1(0)$, D as a regular octagon (black $\diamond$) or D as a circle (red $*$), for different values of β ($\beta = 0.3, 0.4, 0.5, 0.55$).

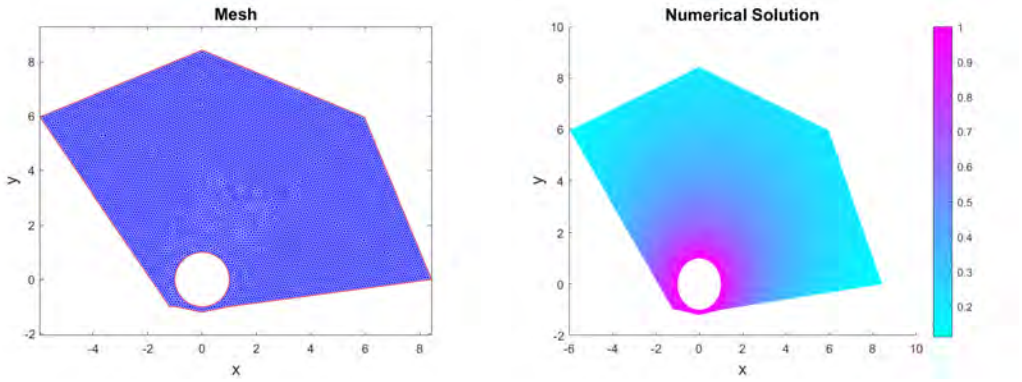
3 NUMERICAL EXPERIMENTS

We discretized the weak formulation of (PDE) in (1), that is

$$\int_D \nabla u \cdot \nabla \phi + \int_{\partial D} \beta u \phi = 0, \tag{9}$$

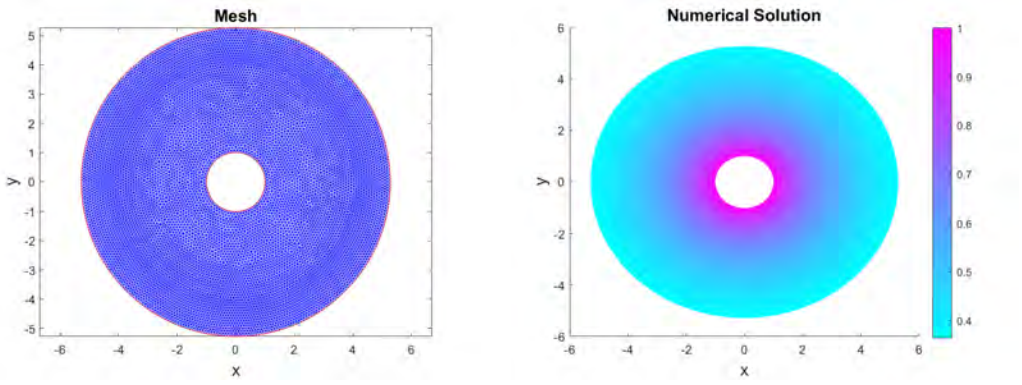
using Finite Element method implemented via Matlab software, in order to calculate the function u. We analyzed the results of the numerical simulations by distinguishing them in the cases summarized below (see [5] for more details):

- We considered circles and polygons with the same area as convex geometries for D. The computational experiments show that the circle seems to be in general the best choice for the external domain, even if, in a few cases, irregular polygons produce a smaller dispersion $F_\beta(D, \Omega)$ for values of $\beta < 1$. An example is depicted in Fig. 3 related to an irregular octagon, compared to the circle visible in Fig. 4 with the same area. However, such an example may be misleading: the example in Figs. 3-4 refers to a case in which the prescribed quantity of insulator represents a technically inadvisable option if we want to have little heat dispersion, since in this case using no insulator at all would be a much better choice (looking at Fig. 1 for $\beta = 0.2$, $R \simeq 5$ corresponds to the maximum possible heat dispersion for the case of concentric circles).



(a) (b)

Figure 3: Plots of (a) the mesh and (b) the numerical solution in the case $\beta = 0.2$, area $A = 87.49$. Dispersion $F_{0.2}(D, \Omega) = 2.32$.



(a) (b)

Figure 4: Plots of (a) the mesh and (b) the numerical solution in the case $\beta = 0.2$, area $A = 87.49$. Dispersion $F_{0.2}(D, \Omega) = 2.41$.

- Motivated by the previous considerations, that seem to advise against a fixed amount of insulator, we considered the problem of minimizing the heat dispersion functional (2) under the constraints $\text{Area}(D) \leq Amax$, $D \in \mathcal{D}$, considering different values of $Amax$. For the numerical minimization, we made use of the `patternsearch` MATLAB routine for global minimization, varying the starting point, and imposing geometrical non linear constraints on the outer polygons. We observed different behaviors of the heat dispersion for $\beta < 1$ and $\beta > 1$. In particular, for $\beta > 1$ all the amount of insulator material is

used (saturation case), whereas for $\beta < 1$ the minimization process may lead to different solutions, depending on the starting point (see [5] for more details).

- Simulations by considering internal domains with shape different from the unit circle have been carried out confirming that the external circle is the domain which gets the lowest values of heat dispersion among all the convex domains considered (see [5]).

Further details and results will be shown during the talk and can also be found in [5].

4 CONCLUSIONS

In this work we have performed a systematic numerical analysis of a mathematical model for thermal insulation problems described by a shape optimization formulation. The results obtained seem to suggest that the most effective thermal insulation for a conductor of constant temperature is obtained by surrounding it with insulating material disposed according to a circular geometry, independently from the shape of the internal body. Counterexamples to that, obtained comparing different geometries which share the same area, do not seem to be of practical interest. Nevertheless, a similar computational approach, possibly with a more complex model, appears extremely useful for the understanding of the physical problem of thermal insulation. In that line, in the future we would like to further explore the thermal insulation problem in order to face in a more accurate way the real life needs coming from engineering applications.

ACKNOWLEDGMENTS

This research has been carried on within the PON R&I 2014-2020 - "AIM: Attraction and International Mobility" (Linea 2.1, project AIM1834118 - 2, CUP: E61G19000050001).
The authors are members of the INdAM Research Group GNCS.

REFERENCES

[1] Della Pietra, F. and Nitsch, C. and Trombetti, C., An optimal insulation problem. *Math. Ann.*, (2020). DOI: https://doi.org/10.1007/s00208-020-02058-6.

[2] Bucur, D. and Buttazzo, G. and Nitsch, C., Two optimization problems in thermal insulation. *Notices Am. Math. Soc.*, **64**(8): 830–835, 2017.

[3] Bucur, D. and Buttazzo, G. and Nitsch, C., Symmetry breaking for a problem in optimal insulation. *J. Math. Pures et Appl.*, **107**(4): 451–463, 2017.

[4] Bergman, Theodore L. and Lavine, Adrienne S. and Incropera, Frank P. and Dewitt, David P., *Introduction to heat transfer*, John Wiley & Sons, 2011.

[5] Tozza, S. and Toraldo, G., Numerical Hints for Insulation Problems, *Applied Mathematics Letters*, **123**, 2022. DOI: https://doi.org/10.1016/j.aml.2021.107609.

Modelling delamination of a DCB test using non-linear truss interface elements and plate elements with assumed shear strain

I.Hlača*, D. Ribarić*, L. Škec* and M. R. Zefreh*

* Faculty of Civil Engineering
University of Rijeka
R. Matejčić 3, 51000 Rijeka, Croatia
e-mail: {ivan.hlaca, dragan.ribaric, leo.skec, maedeh.ranjbar}@gradri.uniri.hr

Key words: Fracture Mechanics, Delamination, Cohesive Zone Model, Finite Element Analysis.

Abstract: *In this work we are investigating mode I delamination of plate-like specimens, where the width is comparable to the length. In such cases anticlastic bending of the plates takes place on the debonded part and the crack front is a curve, rather than a straight line. We model the interface by means of discrete non-linear truss elements with embedded exponential traction-separation law. Such choice is justified because in this test, only pure mode I (opening) displacements occur at the interface, which in our case will cause axial elongation of the truss elements. The plates are modelled using 4-node plate finite elements derived by the assumed shear strain approach that pass the general constant-bending patch test. Cohesive-zone interface parameter identification is performed by a direct method (J-integral) and by virtual experiments regression. Numerical tests have been performed and the exponential cohesive-zone interface model compared against the bi-linear in terms of precision, robustness and computational time. The results confirm the experimentally observed behavior with anticlastic bending of the arms and the curved crack front.*

1 INTRODUCTION

Delamination is one of the most important and severe failure modes of composite structures. Resistance to delamination is essentially resistance to fracture of the interlayer connection, which is expressed in terms of fracture-mechanics parameters such as the critical energy release rate (G_C), the stress intensity factor (K_C) or the J integral.

Although in general there are three basic modes of delamination, as well as the combination of the basic modes (so-called mixed-mode delamination), in this work we will focus only on mode I (opening) delamination. Experimental studies of mode I delamination are commonly performed by so-called double cantilever beam (DCB) test [1,2]. For structural joints and composites, crack is introduced by inserting a thin film in otherwise glued interface and by pulling the specimen apart one is able to monitor the crack propagation.

Fracture mechanics is divided into discipline of linear elastic (LEFM) fracture mechanics and elastic-plastic fracture mechanics. In contrast to limit load analysis, fracture mechanics allows for modelling of inelastic behaviour and drop in load-carrying capabilities. Currently, this is usually done by using the so-called cohesive zone model (CZM) which was introduced by Dugdale and Barenblatt [3,4] in the early '60s of the last century. CZM is used in combination with finite element method by making the constitutive behaviour of material nonlinear. This nonlinearity requires iterative solver (e.g. Newton-Raphson method) and definition of so-called traction-separation law (TSL). By using the J-integral approach, one can experimentally determine the TSL from DCB experiment [5].

https://doi.org/10.4995/YIC2021.2021.12587

2 PROBLEM DESCRIPTION AND DCB TEST FOR PLATE-LIKE SPECIMEN

Double cantilever beam (DCB) test is the standard test for determining the fracture resistance in mode I. Typical geometry, boundary conditions and loading of the test is illustrated in Figure 1. Test specimens are made by gluing two equal adherends together in order to expose them to a symmetric opening load during the experiment, thus creating crack propagation along the bonded surface. Applied load, load-line displacement and crack length are continuously measured during the experiment. While the first two parameters can be obtained directly from the tensile-testing machine, for the measurement of the crack length, additional optical measuring equipment is required. The data obtained from the experiment is then used to compute the fracture toughness of the adhesive using methods known as data-reduction schemes [6]. Geometry and material properties used are given in Table 1.

Table 1: Geometry and material used.

Aluminium layer	L = 250 mm
	B = 120 mm
	h = 6 mm
	a_0 = 45 mm
	E = 70 GPa
	ν = 0.33
SikaPower®-4720 interface	t = 0.5 mm
	G_{IC} = 1.15 N/mm^2
	δ_0 = 0.02 mm
	σ = 21.15 MPa

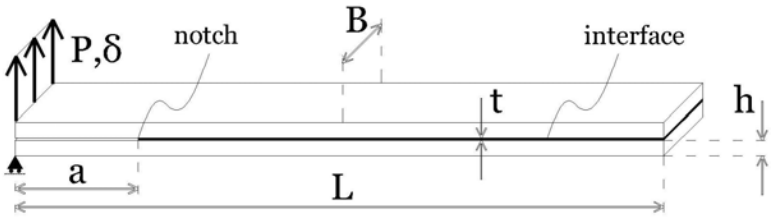


Figure 1: Geometry of a DCB specimen with corresponding boundary conditions and loading.

2.1 Critical energy release rate

Crack propagation occurs when G reaches critical value G_C. Furthermore, it was shown that this value equals,

$$G_C = \frac{P^2 dC}{2Bda} \tag{1}$$

By using the well-known beam theory expressions for cantilever beam deflection we can calculate the compliance and its derivative over change of the fractured area, i.e. we can derive the critical energy release rate G_C. Calculation of G_C is usually done in accordance with international standards for determining the fracture resistance in mode I, namely ISO 25217:2009 and ASTM-D3433-99. Experimentally determined fracture resistance is used as a parameter in CZM model.

2.2 DCB test for plate-like specimen

To the best of authors' knowledge, there is no available method to determine fracture resistance of wide specimens where width is comparable to the length. In this scenario beam theory does not apply and crack front is not a straight line i.e. $dA \neq B \cdot da$ as defined in expression (2). Experimental findings show that the crack front is curved and that the crack length at the specimen's edges is shorter than at its center (see Figure 2). Earlier research [7] reported that curved crack front has a parabolic shape which is verified in this work.

International standards give instruction on how to measure the crack length along the edge of the specimen assuming that the crack front does not vary along the width of the specimen. However, this assumption becomes very questionable for relatively wide specimens. By using digital image correlation (DIC) it is possible to measure the crack without restricting it to the edges only. Measurement method takes advantage of the experiment symmetry.

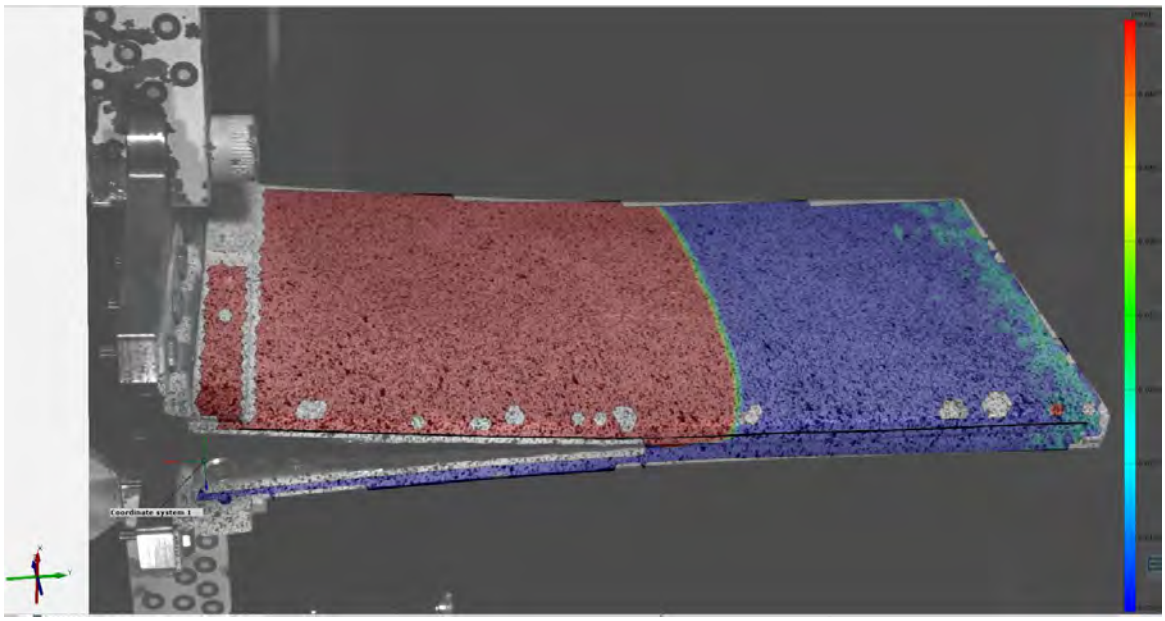


Figure 2: DCB test of wide specimens and observable crack front from the DIC measurement (debonded area is given in red).

3 MODELLING DELAMINATION

3.1 Cohesive zone model

In a cohesive zone model, a non-linear relationship with softening is introduced between the relative displacements at the interface and the corresponding tractions. Interface surfaces are able to lose cohesion and separate from one another as the interface traction σ approach zero. At first, material behaves elastically, as it would in linear calculation, but after reaching the maximum traction, softening at the interface takes place.

3.2 Interface elements

Finite element analysis can use specific TSL for various interface elements, such as 1D spring or multi-node interface, e.g. INT-4, INT-8 [8]. Traction-separation laws can have various shapes, such as bi-linear, trapezoid, exponential, etc. This report will show the application of 1D truss/spring element with material described by exponential and bi-linear [9] TSL. Comparison between the two is shown on Figure 3. Exponential law used here is equivalent to Needleman's

mixed mode law [10] but for spring element application, where only mode I is sufficient, TSL is defined by the authors as follows,

$$\sigma\left(\delta\right) = \frac{G_C}{\delta_0^2}\delta \cdot e^{-\frac{\delta}{\delta_0}}, \tag{2}$$

$$K\left(\delta\right) = \frac{G_C}{\delta_0^2} \cdot e^{-\frac{\delta}{\delta_0}} \cdot \left(1 - \frac{\delta}{\delta_0}\right). \tag{3}$$

Area under the curve of TSL (Figure 3) is by definition equivalent to critical energy release rate G_{IC}. Other TSL parameters such as initial stiffness, maximum traction (maximum stress) and maximum elongation might be adopted from simple mechanical experiments but this is not practical nor reliable. Size effect and other influences result in different behavior for bulk material and cohesive material at interface.

Cohesive-zone interface parameter identification is performed by a direct method (J-integral) [5] and by virtual experiments regression. Exponential TSL parameters used here are retrieved from experiments on regular DCB tests [6] and they are mentioned earlier in Table 1. Bi-linear TSL law has two additional parameter δ_C, σ_{MAX} which can be reduced to either one of the two. Bi-linear TSL is chosen to match the maximum traction of exponential TSL by integrating Expression (2) and finding relation,

$$\sigma_{max} = G_C/(\delta_0 \cdot e) = 21.155\ MPa. \tag{4}$$

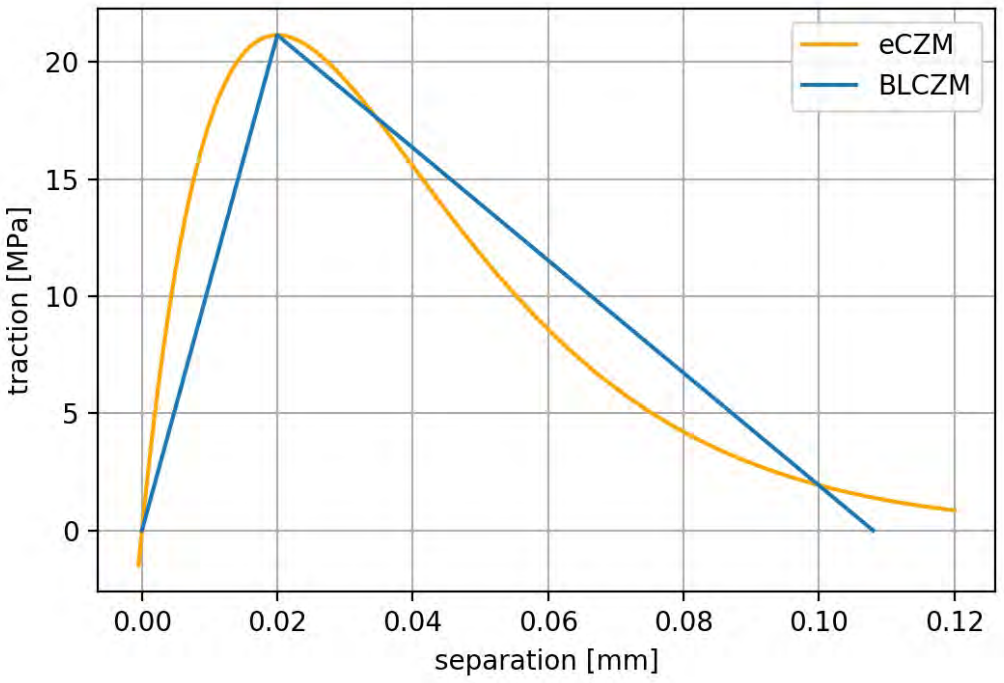


Figure 3: Exponential (eCZM) vs bi-linear (BLCZM) traction-separation law; critical energy release rate and maximum traction are both made equal for the two TSL.

3.3 Layer elements

Q4-U3 is a facet shell element desig ned for the general shell model analysis [11]. It is structured as a joined plate and in-plane (membrane) element whose stiffness matrices are

integrated for a 3D environment. The plate part is based on the cubic linked interpolations for lateral displacement and two section rotations, and since it is cubic it is also, problem dependent, e.g. the material parameters E, ν and thickness t are necessary values in the interpolations. The final expressions for strains (shear strains) are assumed in the simplified form, derived from the differential kinematic equations because they have to match the equilibrium equations too. The membrane part of interpolations is also displacement based and has the cubic form with higher order terms depending of the drilling nodal rotations similar to Allman's cubic interpolations for triangle membrane elements. Membrane part is not needed in delamination simulation considered in this work because in pure mode I delamination there are no in-plane (membrane) stresses.

3.4 Delamination model

The proposed numerical model makes use of truss interface elements with embedded exponential TSL and 4-node linked interpolation plate/shell element for the layers. Truss elements are perpendicular to the layers. Pre-processing and post-processing was done in Python while the main calculations were implemented in FEAP [12] where solution procedure minimizes the residual. At first, it checks the strain energy norm and then the residual norm before stopping. Equation (2) and (3) are implemented in a user material element along with the plate finite element from earlier chapter. Only one half of the DCB specimen is modelled due to the symmetry. Truss interface element has cross-sectional area equal to

$$A = \frac{(L - a_0) \cdot B}{(m-1) \cdot (n-1)}, \tag{5}$$

where m, n are number of nodes across length and width respectively. Interface elements at edges have only half of the area A from (5), while the four elements at vertices have the same area as if they were edge elements for simplicity only. Damage history variable that usually saves the value of maximum separation is not needed because no unloading or reloading occurs during a DCB experiment. One of the problems encountered was the inability to use a computationally lightweight mesh size while achieving a convergence in iterative residual minimization. Higher ductility of the interface parameters (higher δ_0, δ_C) and mesh refinements are the simplest way to improve convergence. It was found that type of finite element for layer also influences the result, especially the oscillations as reported in [8]. Advanced solution procedures, such as arc-length method [8], that can significantly improve convergence of delamination simulation, have not been used in this work. It has been found that the choice of interface TSL can have a strong influence on the convergence, which will be evident from the results in the next chapter.

4 COMPARISON OF RESULTS

4.1 Results

Minimum required finite element mesh for exponential TSL was found to be 50 x 10, which corresponds to element size of 5 x 12 mm. On the other hand, minimum required finite element mesh for bi-linear TSL was found to be 250 x 24 (or more), meaning that one finite element is as small as 1 x 5 mm. Mesh regularity (aspect ratio) could be improved for both cases but this would additionally increase the computational time, in particular for the bi-linear law, while not improving the results noticeably. Figure 4 and 5 show comparison with experimental force-displacement data on the left hand side graph. On the right hand side is the plot of layer separation where the yellow color represents delamination and the purple color represents portion that is still intact while the white color represents area of initial crack.

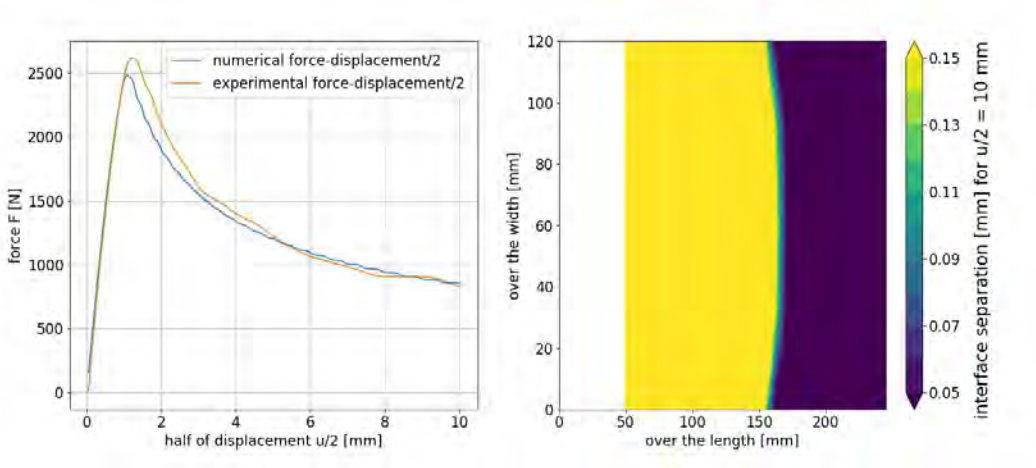


Figure 4: Results for exponential TSL

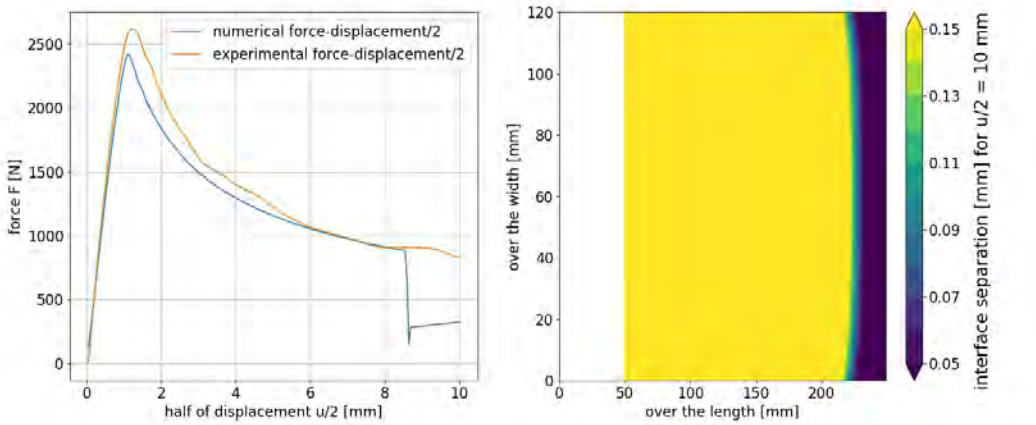


Figure 5: Results for bi-linear TSL

4.2 Performance assessment

In terms of robustness, we found that the exponential TSL, unlike the bi-linear TSL, is capable of converging even with oscillating behavior and large number of iterations. The model with the bi-linear TSL lost convergence near the end of virtual experiment (Figure 5) which suggests that minimum mesh size should in fact be even more refined, e.g. 250 x 48. Results for refined mesh are not presented here for the sake of brevity, but they approximately match the results in Figure 4 although with less oscillations due to increased number of FEM nodes (over 20 000).

Table 2 shows performance in terms of computational cost. It is found that exponential law again outperforms the bi-linear law. Our findings relate to a specific case with truss interface element and this may not translate to other scenarios because this was not observed earlier by other authors [8]. It is worth noting that the number of iterations per increment is approximately the same for different meshes used. Furthermore, exponential law as defined by (2) and (3) is continuous and does not need any IF statements in FEM code in contrast to bi-linear law which require a total of 3 IF statements. As shown in Table 2, the run-time needed to finish simulation is obviously in favor to the eCZM. Important factor that has substantial influence on convergence is the choice of finite element for the layers. All reported

observations were similar for FEAP's original SHELL element which actually showed better convergence properties in all cases. By taking into consideration shear strains of plate (or beam), delamination problems converge harder. Nevertheless, authors expect that the layer elements that account for shear deformation will show better behavior in scenarios with multi-node interface cohesive elements and solution procedures where mesh is sparser as this was the case in simpler benchmarks [11].

Table 2: Geometry and material used.

	eCZM	BLCZM
Minimal mesh (No. of elements)	50 x 10	250 x 24
No. of nodes	1012	11300
Average No. of iterations per increment	4.75	4.68 (before loss of convergence)
Run-time	33 seconds	47 minutes

5 CONCLUSIONS

In this work, the curved crack front of a double cantilever beam specimen with width comparable to its length has been obtained experimentally and numerically. In addition, force vs displacement data for experimental and numerical results are in good agreement. Non-linear truss elements used to model delamination in conjunction with shell elements behave better in all aspects if exponential, rather than bi-linear, TSL is assumed. Benefits of elements with higher order interpolation were not noticed for relatively dense meshes used. On the contrary, accounting for shear strain introduced problems such as oscillations or even loss of convergence. Reasons behind this phenomenon will also be a topic of further study.

REFERENCES

[1] British Standards Institution BS ISO 25217. (2009) Adhesives – Determination of the mode 1 adhesive joints using double cantilever beam and tapered double cantilever beam specimen. British Standard.

[2] ASTM D3433-99. (2012) Standard test method for fracture strength in cleavage of adhesives in bonded metal joints. ASTM International.

[3] Dugdale DS. Yielding of steel sheets containing slits. *JMech Phys Solids 1960*; **8**:100–4.

[4] Barenblatt GI. The mathematical theory of equilibrium cracks in brittle fracture. *Adv Appl Mech* 1962;**7**:55–129.

[5] Gorman, J. M. and Thouless, M. D. (2019) The use of digital-image correlation to investigate the cohesive zone in a double-cantilever beam, with comparisons to numerical and analytical models, *Journal of the Mechanics and Physics of Solids*, **123**, pp. 315–331. doi: 10.1016/j.jmps.2018.08.013.

[6] Hlača, I., Grbac, M. and Škec, L. (2019) Determining Fracture Resistance of Structural Adhesives in Mode-I Debonding Using Double Cantilever Beam Test, *Zbornik radova*, **22**(1), pp. 59–74. doi: 10.32762/zr.22.1.4.

[7] WIT Transactions on Engineering Sciences (1994). ISSN: 1743-3533, 251-258.

[8] Alfano, G. (2006) On the influence of the shape of the interface law on the application of cohesive-zone models, *Composites Science and Technology*, **66**(6), pp. 723–730. doi: 10.1016/j.compscitech.2004.12.024.

[9] Ranjbar, M., Jelenić, G., and Škec, L. (2020). Modelling Adhesive Using Non-Linear Truss Elements in Mode I Delamination Problems, *Zbornik radova*, XXIII(1), 29-40. https://doi.org/10.32762/zr.23.1.2

[10] Xu, X. P. and Needleman, A. (1993) Continuum Modelling of Interfacial Decohesion, *Solid State Phenomena*, 35–36, pp. 287–302. doi: 10.4028/www.scientific.net/ssp.35-36.287.

[11] Ribarić, D. (2016) Problem-dependent cubic linked interpolation for Mindlin plate four-node quadrilateral finite elements, *Structural Engineering and Mechanics*, **59**(6), pp. 1071–1094. doi: 10.12989/sem.2016.59.6.1071.

[12] Taylor, R.L. FEAP - Finite Element Analysis Program Published: 2014 Publisher: University of California, Berkeley URL: http://www.ce.berkeley/feap

UNCERTAINTY QUANTIFFICATION OF DIFFERENTIAL EQUATIONS WITH RANDOM PARAMETERS: METHODS AND APPLICATIONS

An overview of p-refined Multilevel quasi-Monte Carlo Applied to the Geotechnical Slope Stability Problem

Philippe Blondeel[1], Pieterjan Robbe[1], Stijn François[2], Geert Lombaert[2] and Stefan Vandewalle[1]

[1] KU Leuven, Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
{philippe.blondeel,pieterjan.robbe,stefan.vandewalle}@kuleuven.be

[2] KU Leuven, Department of Civil Engineering
Kasteelpark Arenberg 40, 3001 Leuven, Belgium
{stijn.francois,geert.lombaert}@kuleuven.be

Key words: Multilevel Quasi-Monte Carlo, p-refinement, Higher Order Finite Elements

Abstract: *Problems in civil engineering are often characterized by significant uncertainty in their material parameters. Sampling methods are a straightforward manner to account for this uncertainty, which is typically modeled as a random field. A popular sampling method consists of the classic Multilevel Monte Carlo method (h-MLMC). Its most distinctive feature consists of a hierarchy of h-refined meshes, where most of the samples are taken on coarse and computationally inexpensive meshes, and few are taken on finer but computationally expensive meshes. We present an improvement upon the classic Multilevel Monte Carlo, called the p-refined Multilevel quasi-Monte Carlo method (p-MLQMC). Its key features consist of a mesh hierarchy constructed from a p-refinement scheme combined with a deterministic set of samples points (quasi-Monte Carlo points). In this work we show how the uncertainty needs to be accounted for and present results comparing the total computational cost of the h-ML(Q)MC and p-MLQMC method. Specifically, we present two novel approaches in order to account for the uncertainty in case of p-MLQMC. We benchmarking the different multilevel methods on a slope stability problem, and find that p-MLQMC outperforms h-MLMC up to several orders of magnitude.*

1 INTRODUCTION

Problems in the engineering sciences are typically subject to uncertainty. In order to assess the uncertainty on the solution of the considered engineering problem, different steps need to be taken. First, the engineering problem is discretized, i.e., the underlying partial differential equation (PDE) governing the problem is approximated, by for example, the Bubnov–Galerkin Finite Element method. Second, the uncertainty present in the material parameters of the model, is to be represented as accurately as possible. Here, we chose to represent the uncertainty by means of a random field obtained through a Karhunen-Loève expansion (KL). Third, the modeled uncertainty needs to be accounted for in the Finite Element method. We consider two methods to achieve this step, i.e., the midpoint method and the integration point method. Fourth, the uncertainty on the solution is to be assessed. A straightforward manner to accomplish this last step, is by means of a stochastic sampling method. A well-known stochastic sampling method consists of the classic Multilevel Monte Carlo (h-MLMC) method. First developed by Giles, see [1, 2], the h-MLMC method relies on a hierarchy of refined meshes in order to reduce the total computational cost by means of variance reduction. Most of the samples are taken on low resolution and computationally cheap meshes, while a decreasing number of samples are taken on high resolution and computationally expensive meshes. The mesh hierarchy is typically constructed by selecting a coarse Finite Element mesh approximation of the considered problem, and recursively applying the h-refinement scheme. In previous

https://doi.org/10.4995/YIC2021.2021.12236

work, we introduced the p-refined Multilevel Quasi-Monte Carlo method (p-MLQMC), see [3], which essentially combines a mesh hierarchy based on a p-refinement scheme, i.e., increasing the polynomial order of the elements's shape function, together with a quasi-Monte Carlo sampling rule based on a rank-1 lattice sequence, e.g., [4]. This combination yields significant computational cost savings with respect to classic Multilevel (quasi-) Monte Carlo (h-ML(Q)MC). When accounting for the uncertainty in the Finite Element model, we observed a greater challenge with the p-MLQMC method than with the h-ML(Q)MC method. In our implementation, h-ML(Q)MC makes use of the midpoint method, while p-MLQMC makes use of the integration point method. In this work we present two novel approaches in order to implement the integration point method, with respect to our previous work see [3], i.e., the Local Nested Approach (LNA) and the Non-Nested Approach (NNA). In addition to this, we will benchmark the h-ML(Q)MC method against the p-MLQMC method on a slope stability problem where the cohesion of the soil is uncertain. The slope stability problem is a geotechnical engineering problem, where the goal is to assess the stability of natural or man-made slopes.
The paper is structured as follows. First we introduce the considered model problem. Second, we present the theoretical background pertaining to multilevel methods. Third, we discuss how the uncertainty is modeled as a random field and focus on how to account for said uncertainty in the Finite Element model. Last, we present the results obtained for p-MLQMC coupled with LNA and NNA, and h-ML(Q)MC with the midpoint method.

2 MODEL PROBLEM

The model problem we consider for benchmarking the methods, consists of a slope stability problem where the soil's cohesion has a spatially varying uncertainty, see [5]. We will discuss how to model this uncertainty in §4. In a slope stability problem, the safety of the slope can be assessed by evaluating the vertical displacement of the top of the slope when sustaining its own weight. Different discretizations of the slope stability problem are presented in Figure 1.

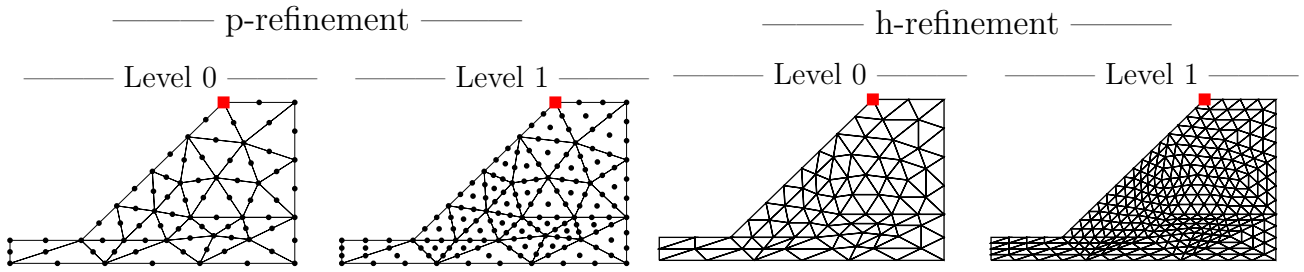


Figure 1: An example of a fine and a coarse mesh used for the slope stability problem with the location of QoI indicated by ■.

We consider the displacement in the plastic domain, which is governed by the Drucker–Prager yield criterion. In the plastic domain, the stress–strain relation has a nonlinear behavior. Therefore, in order to compute a strain increment given a stress increment, an elastic predictor–plastic corrector iterative solver is used. In literature, this is commonly referred to as the 'Return Mapping algorithm', e.g., [6]. The governing partial differential equations are discretized by means of the Bubnov–Galerkin Finite Element method, giving rise to a system of equations. In order to compute the displacement, an incremental load approach is used, i.e., the total load resulting from the slope's weight is added in discrete load steps, starting with a force of 0 N. These load steps are added until the total downward force resulting from the slope's weight is reached. The discretized system of equation, describing the displacement, that needs to be solved iteratively

by a Newton–Raphson solver is given as

$$\mathbf{K}\Delta\mathbf{u} = \mathbf{f} + \Delta\mathbf{f} - \mathbf{k}, \tag{1}$$

where $\Delta\mathbf{u}$ stands for the displacement increment and $\mathbf{K}$ the global stiffness matrix resulting from the assembly of element stiffness matrices $\mathbf{K^e}$. The right hand side of Eq. (1) stands for the residual. Here, $\mathbf{f}$ is the sum of the external force increments applied in the previous steps, $\Delta\mathbf{f}$ is the applied load increment of the current step and $\mathbf{k}$ is the internal force resulting from the stresses. For a more thorough explanation on the methods used to solve the slope stability problem we refer to [7, Chapter 2 §4 and Chapter 7 §3 and §4].

3 SAMPLING AND MESH HIERARCHIES

The expected value of a function P against an s-dimensional probability density function ϕ is defined by

$$\mathbb{E}[P] := \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} P(x_1, \ldots, x_s)\, \phi(x_1, \ldots, x_s)\, \mathrm{d}x_1 \cdots \mathrm{d}x_s = \int_{\mathbb{R}^s} P(\mathbf{x})\, \phi(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{2}$$

In order to approximate the integral in Eq. (2), an equal-weight quadrature rule can be used. An example of such an equal-weight quadrature rules is the Monte Carlo method. In our case, the function P is obtained by means of a Finite Element method on a chosen discretization level L, which leads to a first approximation of the integral, $\mathbb{E}[P] \approx \mathbb{E}[P_\mathrm{L}]$. The computation of the integral itself is performed by defining an estimator, leading to a second approximation, $\mathbb{E}[P_\mathrm{L}] \approx Q_\mathrm{L}^{\mathrm{ML(Q)MC}}$.

3.1 Multilevel Monte Carlo

In multilevel methods, the expected value of $\mathbb{E}[P_\mathrm{L}]$ is written as a telescoping sum

$$\mathbb{E}[P_\mathrm{L}] = \mathbb{E}[P_0] + \sum_{\ell=1}^{\mathrm{L}} \mathbb{E}[P_\ell - P_{\ell-1}]. \tag{3}$$

The resulting MLMC estimator used for the approximation of Eq. (2) is then given as

$$Q_\mathrm{L}^{\mathrm{MLMC}} := \frac{1}{N_0} \sum_{n=1}^{N_0} P_0(\mathbf{x}_0^{(n)}) + \sum_{\ell=1}^{\mathrm{L}} \left\{ \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \left(P_\ell(\mathbf{x}_\ell^{(n)}) - P_{\ell-1}(\mathbf{x}_\ell^{(n)}) \right) \right\}, \tag{4}$$

where $\mathbf{x}_\ell^{(n)}$ stands for the nth sample point. In the MLMC estimator the $\mathbf{x}_\ell^{(n)}$ are (pseudo-)randomly chosen points, which are distributed according to $\phi(\cdot)$, see Eq. (2). The expected value of the quantity of interest on the finest level $\ell = \mathrm{L}$, is expressed as the sample average of the quantity of interest on the coarsest level $\ell = 0$, plus a series of correction terms on levels $\ell = \{1, \ldots, \mathrm{L}\}$, hence the name 'telescoping sum'. The variance of the MLMC estimator is given by

$$\mathbb{V}\left[Q_\mathrm{L}^{\mathrm{MLMC}}\right] = \sum_{\ell=0}^{\mathrm{L}} \mathbb{V}\left[\Delta Q_\ell^{\mathrm{MLMC}}\right] = \sum_{\ell=0}^{\mathrm{L}} \frac{\mathbb{V}\left[\Delta P_\ell\right]}{N_\ell} \approx \sum_{\ell=0}^{\mathrm{L}} \frac{\mathsf{V}_\ell}{N_\ell} = \sum_{\ell=0}^{\mathrm{L}} \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \frac{\left(\Delta P_\ell^{(n)} - \Delta Q_\ell\right)^2}{N_\ell}, \tag{5}$$

where $\Delta Q_\ell := \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \Delta P_\ell^{(n)}$, with $\Delta P_\ell^{(n)} := P_\ell(\mathbf{x}_\ell^{(n)}) - P_{\ell-1}(\mathbf{x}_\ell^{(n)})$ and $P_{-1} := 0$. Multilevel methods rely on a variance reduction across the levels in order to achieve a computational

speedup. This means that the sample variance of the difference for increasing level ℓ continuously decreases, i.e., $\mathbb{V}[\Delta P_1] > \mathbb{V}[\Delta P_2] > \cdots > \mathbb{V}[\Delta P_{\mathrm{L}}]$. This variance reduction is only obtained when a strong positive correlation is achieved between the results of two successive levels, $P_\ell := P_\ell(\mathbf{x}_\ell^{(n)})$ and $P_{\ell-1} := P_\ell(\mathbf{x}_{\ell-1}^{(n)})$, i.e.,

$$\begin{aligned}\mathbb{V}[\Delta P_\ell] &= \mathbb{V}[P_\ell - P_{\ell-1}] \\ &= \mathbb{V}[P_\ell] + \mathbb{V}[P_{\ell-1}] - 2\mathrm{cov}(P_\ell, P_{\ell-1}),\end{aligned} \tag{6}$$

where $\mathrm{cov}(P_\ell, P_{\ell-1}) := \rho_{\ell,\ell-1}\sqrt{\mathbb{V}[P_\ell]\,\mathbb{V}[P_{\ell-1}]}$ is the covariance between P_ℓ and $P_{\ell-1}$ with $\rho_{\ell,\ell-1}$ the correlation coefficient.

3.2 Multilevel quasi-Monte Carlo

The MLQMC estimator is given by

$$Q_{\mathrm{L}}^{\mathrm{MLQMC}} := \frac{1}{R_0}\sum_{r=1}^{R_0}\frac{1}{N_0}\sum_{n=1}^{N_0} P_0(\mathbf{x}_0^{(r,n)}) + \sum_{\ell=1}^{\mathrm{L}}\frac{1}{R_\ell}\sum_{r=1}^{R_\ell}\left\{\frac{1}{N_\ell}\sum_{n=1}^{N_\ell}\left(P_\ell(\mathbf{x}_\ell^{(r,n)}) - P_{\ell-1}(\mathbf{x}_\ell^{(r,n)})\right)\right\}, \tag{7}$$

with its variance given by

$$\mathbb{V}\left[Q_{\mathrm{L}}^{\mathrm{MLQMC}}\right] = \sum_{\ell=0}^{\mathrm{L}}\mathbb{V}\left[\Delta Q_\ell^{\mathrm{MLQMC}}\right]. \tag{8}$$

In order to estimate $\mathbb{V}[\Delta Q^{\mathrm{MLQMC}}]$ we use the sample variance $\mathcal{V}_\ell$ over the R_ℓ independent shifts, see [8]

$$\mathcal{V}_\ell = \sum_{r=1}^{R_\ell}\frac{1}{R_\ell(R_\ell-1)}\left(\frac{1}{N_\ell}\sum_{n=1}^{N_\ell}\Delta P_\ell^{(r,n)} - \Delta Q_\ell\right)^2, \tag{9}$$

where $\Delta Q_\ell := \frac{1}{R_\ell}\sum_{r=1}^{R_\ell}\frac{1}{N_\ell}\sum_{n=1}^{N_\ell}\Delta P_\ell^{(r,n)}$, with $\Delta P_\ell^{(r,n)} := P_\ell(\mathbf{u}_\ell^{(r,n)}) - P_{\ell-1}(\mathbf{u}_\ell^{(r,n)})$ and $P_{-1} := 0$. While the MLMC method is based on (pseudo-)random distributed sample points, the MQLMC method uses deterministic sample points (QMC points), $\mathbf{x}_\ell^{(r,n)}$. More specifically, here we use a rank-1 lattice sequence. In order to recover unbiased estimates of the estimator, the computation of the estimator and its variance include an averaging over a number of shifts $r = 1, 2, ..., R_\ell$ on each level ℓ. The procedure of random shifting consists of adding to each point of the lattice sequence, a uniformly distributed number $\Xi_r \in [0,1)^s$, after which the fractional part is taken. This is illustrated in Figure 2. In our implementation $R_\ell = 10$ for each ℓ, $0 \le \ell \le \mathrm{L}$.
The shifted version of the lattice points is given by

$$\mathbf{x}^{(r,n)} := \Phi^{-1}\left(\mathrm{frac}\left(\phi_2(n)\mathbf{z} + \Xi_r\right)\right), n \in \mathbb{N}, \tag{10}$$

where Φ^{-1} is the inverse of the univariate standard normal cumulative distribution function, $\mathrm{frac}(x) := x - \lfloor x \rfloor, x > 0$, ϕ_2 is the radical inverse function in base 2, and $\mathbf{z}$ is an s-dimensional vector of positive integers. The generating vector $\mathbf{z}$ was constructed with the component-by-component (CBC) algorithm with decreasing weights, $\gamma_j = 1/j^2$, see [9].

3.3 Mesh Hierarchies

In the multilevel setting, the levels $0 \le \ell \le \mathrm{L}$ refer to the meshes in the mesh hierarchy. The coarsest mesh is denoted as level 0, while subsequent refinements of the coarse mesh are denoted as level 1, level 2, ... Classically, the mesh hierarchy in the ML(Q)MC method is constructed starting from a coarse Finite Element mesh, to which h-refinement is recursively applied, see

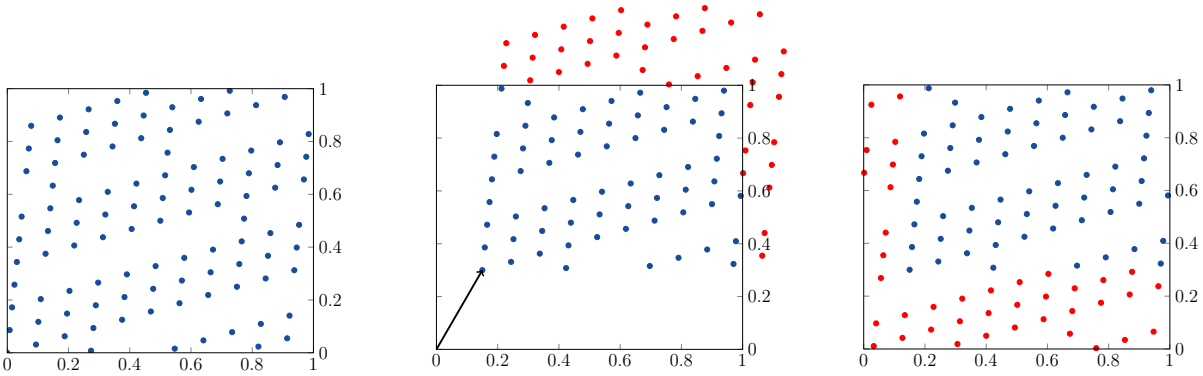


Figure 2: Random shifting procedure applied to points belonging to a rank-1 lattice sequence.

[10]. Here, we use a mesh hierarchy based on a p-refinement approach, i.e., increasing polynomial order of the elements's shape function with increasing level. This mesh hierarchy applied to the slope stability problem, is shown in Figure 1. The Finite Element nodal points are represented as black dots. In Figure 1, we also present the h-refined mesh hierarchy of the slope stability problem.

3.4 Number of Samples

In Multilevel methods, the error is controlled by imposing a tolerance, ε^2, on the Mean Square Error (MSE) of the of the estimator. This MSE is defined as,

$$\begin{aligned} \mathrm{MSE}\left[Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}}\right] &:= \mathbb{E}\left[\left(Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}} - \mathbb{E}[P]\right)^2\right] \\ &= \mathbb{V}[Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}}] + \left(\mathbb{E}\left[Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}}\right] - \mathbb{E}[P]\right)^2 \\ &= \mathbb{V}[Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}}] + \left(\mathbb{E}[P_{\mathrm{L}} - P]\right)^2. \end{aligned} \tag{11}$$

The right-hand side of Eq. (11) consists of two parts, i.e., the variance of the estimator, $\mathbb{V}[Q_{\mathrm{L}}^{\mathrm{ML(Q)MC}}]$, and the squared bias, $(\mathbb{E}[P_{\mathrm{L}} - P])^2$. The stopping criterion for multilevel schemes is typically based on the requirements that both terms are less than $\frac{\varepsilon^2}{2}$. In order to achieve the requested tolerance for the variance of the estimator, the number of samples is increased. In the MLMC method, the optimal number of samples per level is given as

$$N_\ell = \frac{2}{\varepsilon^2}\sqrt{\frac{\mathsf{V}_\ell}{C_\ell}}\sum_{\ell=0}^{\mathrm{L}}\sqrt{\mathsf{V}_\ell C_\ell}, \tag{12}$$

where V_ℓ stands for the sample variance, see Eq. (5), and C_ℓ is the cost to compute one sample on level ℓ, see [2]. However, in the MLQMC method, the number of samples to be taken is determined by means of a 'doubling' algorithm, see [4]. The procedure starts by computing a number of warm-up samples together with a user-defined number of shifts on each level. From these samples $\mathbb{V}\left[\Delta Q_\ell^{\mathrm{MLQMC}}\right]$ is estimated on each level ℓ, see Eq. (9). The iterative step consists of selecting the level τ on which the ratio of the variance of the estimator with the sample cost is maximal, i.e., $\underset{\tau\in\mathrm{L}}{\mathrm{argmax}}\,(\mathcal{V}_\tau/C_\tau)$. On this level τ the number of samples is multiplied with a constant factor. This procedure is repeated until $\mathbb{V}\left[Q_{\mathrm{L}}^{\mathrm{MLQMC}}\right] < \frac{\varepsilon^2}{2}$. In our approach, this constant is chosen as 1.2.

4 UNCERTAINTY MODELING AND INCORPORATION

The uncertainty present in the cohesion of the soil of the slope stability problem is modeled as a lognormal random field, i.e., the exponential of a Gaussian random field. Realizations of the Gaussian random field are computed by means of the truncated Karhunen–Loève (KL) expansion,

$$Z(\mathbf{x},\omega) = \overline{Z}(\mathbf{x}) + \sum_{n=1}^{s} \sqrt{\theta_n}\, \xi_n(\omega)\, b_n(\mathbf{x}), \tag{13}$$

where s is the number of terms in the expansion, i.e., the number of stochastic dimensions. Here, $\overline{Z}(\mathbf{x})$ is the mean of the field and $\xi_n(\omega)$ denote i.i.d. standard normal random variables. The eigenvalues θ_n and eigenfunctions $b_n(\mathbf{x})$ are the solutions of the eigenvalue problem

$$\int_D C(\mathbf{x},\mathbf{y})\, b_n(\mathbf{y})\, \mathrm{d}\mathbf{y} = \theta_n\, b_n(\mathbf{x}), \tag{14}$$

where $C(\mathbf{x},\mathbf{y})$ is a given covariance kernel. The kernel we consider for the random field is the Matérn covariance kernel

$$C(\mathbf{x},\mathbf{y}) := \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, \|\mathbf{x}-\mathbf{y}\|_2}{\lambda} \right)^{\nu} K_\nu \left(\sqrt{2\nu}\, \frac{\|\mathbf{x}-\mathbf{y}\|_2}{\lambda} \right), \tag{15}$$

where ν is the smoothness parameter, $K_\nu(\cdot)$ is the modified Bessel function of the second kind, $\Gamma(\cdot)$ is the gamma function, σ^2 is the variance, λ is the correlation length, and $\|\cdot\|_2$ is the L^2 norm. The integral in Eq. (14) is approximated by means of a numerical collocation scheme. For more information, we refer to [11, Chapter 7 Section 2]. The lognormal representation of the random field is obtained by applying the exponential to the field obtained in Eq. (13), $Z_{\text{lognormal}}(\mathbf{x},\omega) = \exp(Z(\mathbf{x},\omega))$.

In order to incorporate the uncertainty in the Finite Element model, we consider two different methods, the midpoint method and the integration point method. In both methods the uncertainty resides in the elastoplastic constitutive matrix $\mathbf{D}$. This matrix is used for constructing the element stiffness matrices by integrating the following expression,

$$\mathbf{K^e} = \int_{\Omega_e} \mathbf{B}^T \mathbf{D} \mathbf{B} d\Omega_e \approx \sum_{i=1}^{|\mathbf{q}|} \mathbf{B}_i^T \mathbf{D}_i \mathbf{B}_i \mathrm{w}_i. \tag{16}$$

The matrix $\mathbf{B}$ contains the derivatives of the element shape function, and $|\mathbf{q}|$ is the number of quadrature points used for the numerical integration. The assembly of the element stiffness matrices results in the global stiffness matrix, see Eq. (1). In practice, the matrix $\mathbf{K^e}$ is computed by means of a quadrature rule, where $\mathbf{B}_i$ stands for the matrix $\mathbf{B}$ evaluated at quadrature point $\mathbf{q}_i \in \mathbf{q}$, i.e., $\mathbf{B}(\mathbf{q}_i)$, $\mathbf{D}_i$ the matrix $\mathbf{D}$ containing the uncertainty, i.e., $\mathbf{D}(\omega_i)$, and w_i the quadrature weight.

We will now present the two methods used to account for the uncertainty in the Finite Element method. The goal consists of selecting the random field evaluation points $\mathbf{x}$ used for the evaluation of Eq. (13). Because we are considering a multilevel approach, a set of random field evaluation points must be selected for each level, i.e., $\mathbf{x}_\ell$ for $\ell = \{0, \ldots, \mathrm{L}\}$.

4.1 Midpoint Method

The midpoint method is often used in conjunction with the h-ML(Q)MC method. The random field evaluation points are selected as the centroids of the elements, i.e., Eq. (13) is evaluated

at the centroids of the elements and the resulting values are assigned to the elements. This is shown in Figure 3, where ● represent the spatial locations of the centroids of the elements. In case of the midpoint method, the uncertainty inside each element is assumed to be constant, i.e., $\mathbf{D}_1 = \mathbf{D}_2 = \cdots = \mathbf{D}_{|\mathbf{q}|}$, see Eq.(16). Note that the resolution of the random field increases with each level, i.e., $|\mathbf{x}_0| < |\mathbf{x}_1| < \cdots < |\mathbf{x}_\mathrm{L}|$.

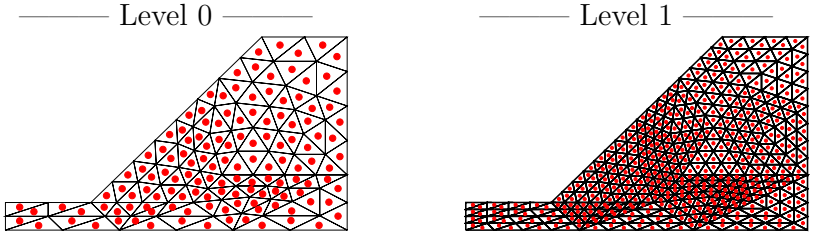


Figure 3: Locations of the random field evaluation points ● for the midpoint method.

4.2 Integration Point Method

In the p-MLQMC method, the number of elements in the hierarchy of mesh discretizations remains the same. Therefore, the midpoint method can not be used if we want the resolution of the random field to increase with increasing level. In order to obtain a higher resolution of the random field per increasing level, we use the integration point method, see [12], with the added condition that the number of quadrature points used to numerically integrate Eq. (16) also increases with increasing level. In the integration point method, Eq. (13) is evaluated at the locations of the quadrature points, or integration points, meaning that the uncertainty varies inside each individual element, i.e., $\mathbf{D}_1 \neq \mathbf{D}_2 \neq \cdots \neq \mathbf{D}_{|\mathbf{q}|}$.

4.2.1 Non-Nested Approach

The Non-Nested Approach is the most simple way to select random field evaluation points. In this approach, the random field evaluation points are chosen equal to the quadrature points used for the numerical integration of Eq. (16). In practice, these quadrature points are first selected on a reference triangular element, see Figure 4, before being mapped to the global coordinates of the mesh. Note that the sets of quadrature points are not nested across the different levels, i.e., $\mathbf{q}_0 \not\subseteq \mathbf{q}_1 \not\subseteq \cdots \not\subseteq \mathbf{q}_\mathrm{L}$. Hence the sets of random field evaluation points are not nested across the levels either, i.e., $\mathbf{x}_0 \not\subseteq \mathbf{x}_1 \not\subseteq \cdots \not\subseteq \mathbf{x}_\mathrm{L}$. The obtained sets of random field evaluations points $\mathbf{x}_\ell$, with $0 \leq \ell \leq \mathrm{L}$, are then used to compute discrete instances of the random field according to Eq. (13). As such, the random field $Z(\mathbf{x}, \omega)$ is approximated on each level by a discrete set of random variables. Defining $\mathbf{Z}_\ell := (Z(\mathbf{x}_\ell, \omega), \mathbf{x}_\ell)$ as the the set of random variables representing the random field and their locations, we see that those are not nested across levels, i.e., $\mathbf{Z}_0 \not\subseteq \mathbf{Z}_1 \not\subseteq \cdots \not\subseteq \mathbf{Z}_\mathrm{L}$. This impacts the variance reduction, see Eq. (6), as it leads to a weak correlation between the solutions on successive levels.

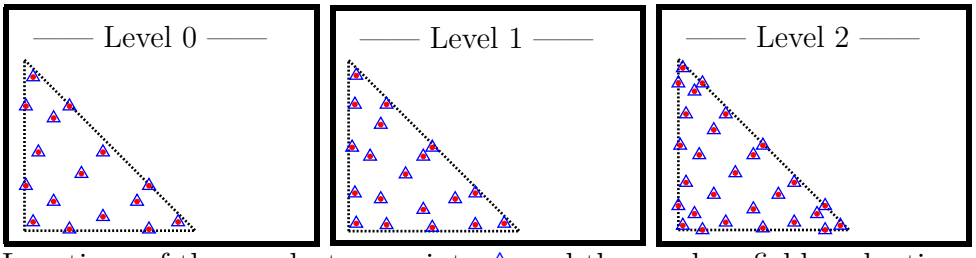


Figure 4: Locations of the quadrature points △ and the random field evaluation points ● on a reference triangular element for NNA.

4.2.2 Local Nested Approach

In the Local Nested Approach we try to improve the correlation between the solutions on different levels. Ideally, one would have $\mathbf{Z}_0 \subseteq \mathbf{Z}_1 \subseteq \cdots \subseteq \mathbf{Z}_\mathrm{L}$, i.e., the random field on each level is represented by using an exact subset of the information used to represent the random field on the finest level. Such an approach has been tried in [13], with limited success. Here, we suggest an alternative in which we only aim at a good correlation between each set of two successive levels in the mesh hierarchy. Such a two-by-two correlation is sufficient for multilevel sampling methods to achieve a rapid reduction of $\mathbb{V}[\Delta P_\ell]$.
Consider the correction $\mathbb{E}[\Delta P_\ell] := \mathbb{E}[P_\ell - P_{\ell-1}]$, which is one of the terms in the telescopic sum, Eq. (3). The integral for computing the element stiffness matrices in P_ℓ makes use of the quadrature point set $\mathbf{q}_\ell$. At those points, we evaluate the random field $Z(\mathbf{x}, \omega)$, i.e., we set $\mathbf{x}_\ell = \mathbf{q}_\ell$. The integral for computing the element stiffness matrices in $P_{\ell-1}$ makes use of the quadrature point set $\mathbf{q}_{\ell-1}$. However, we do not evaluate the random field at those locations, but rather evaluate the random field at points which are a subset of $\mathbf{x}_\ell$, i.e., $\mathbf{x}_{\ell-1,\mathrm{subs}} \subseteq \mathbf{x}_\ell$, such that they have minimal distance with $\mathbf{q}_{\ell-1}$. This is illustrated in Figure 5. (Note that this approximation is done on the level of the reference triangular element, before the mapping to the actual elements of the mesh.)
Define again $\mathbf{Z}_\ell := (Z(\mathbf{x}_\ell, \omega), \mathbf{x}_\ell)$, here with $\mathbf{x}_\ell = \mathbf{q}_\ell$, and $\mathbf{x}_{\ell,\mathrm{main}} := \mathbf{x}_\ell$. The local nested approach ensures that, for each correction $\mathbb{E}[\Delta P_\ell]$ separately, a relation $\mathbf{Z}_{\ell-1,\mathrm{subs}} \subseteq \mathbf{Z}_\ell$ is satisfied. Here, $\mathbf{Z}_{\ell-1,\mathrm{subs}}$ is a 'substitute random field', which approximates $\mathbf{Z}_{\ell-1}$. The substitute field correlates well with the discrete field on the ℓ'th level as it shares part of that field's random variables.

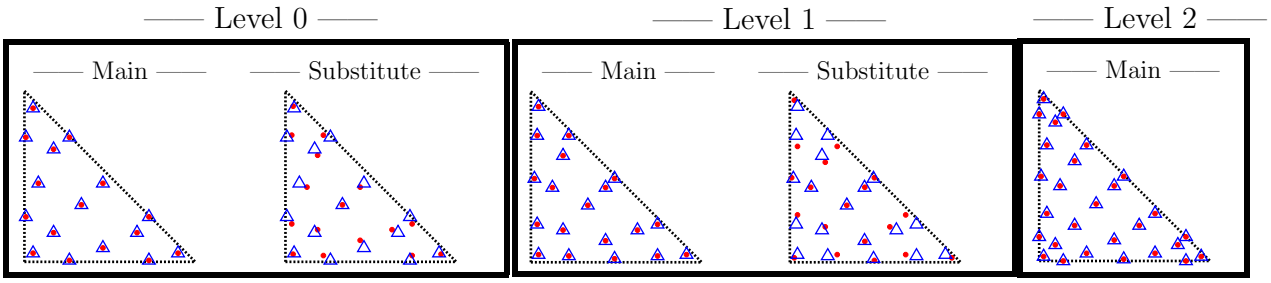


Figure 5: Locations of the quadrature points △ and the random field evaluation points ● on a reference triangular element for LNA.
An important note must be made concerning the LNA approach. While it successfully correlates the solutions of two successive levels, the expected value obtained from the telescoping sum is biased. We have observed a small bias of the order of 10^{-6} with respect to the actual values, an error that is well below the discretization error of the finite element discretization. The reasons behind this additional bias stems from the fact that substitute random fields are used. We are currently investigating how this additional bias can be avoided.

5 RESULTS

In this section we discuss the results obtained with the p-ML(Q)MC-LNA/NNA and the h-ML(Q)MC methods. The quantity of interest (QoI) is taken as the vertical displacement in meters of the upper left node of the model. This location of the QoI is depicted in Figure 1 by ■. The mesh hierarchies shown in Figure 1 are generated by using a combination of the open source mesh generator GMSH, see [14], and MATLAB, see [15]. In this paper we consider two-dimensional Lagrange triangular elements. The random field, computed by means of the Julia package **GaussianRandomFields.jl** [16] has the following parameters $\nu = 0.4$, $\sigma^2 = 1.0$, $\lambda = 1.5$. The characteristics of the lognormal distribution used to represent the uncertainty of the cohesion of the soil are as follows: a mean of 8.02 kPa and a standard deviation of

400 Pa. The spatial dimensions of the slope are: a length of 20 m, a height of 14 m and a slope angle of 30°. The material characteristics are: a Young's modulus of 30 MPa, a Poisson ratio of 0.25, a density of 1330 kg/m^3 and a friction angle of 20°. The number of stochastic dimensions considered for the generation of the Gaussian random field is $s = 400$, see Eq. (13). With a value $s = 400$ at least 99% of the variability of the random field is accounted for. The stochastic sampling was performed with the Julia packages **MultilevelEstimators.jl**, see [17]. The Finite Element code used, is an in-house MATLAB code developed by the Structural Mechanics Section of the KU Leuven. All the results have been computed on a workstation equipped with 2 physical cores, Xeon Gold 6240 CPU's, each with 18 logical cores, clocked at 2.60 GHz, and a total of 192 GB RAM.

5.1 Displacement of the Mesh

In Figure 6 we show the displacement of the mesh and the value of the QoI for four samples of the random field computed on the first four levels.

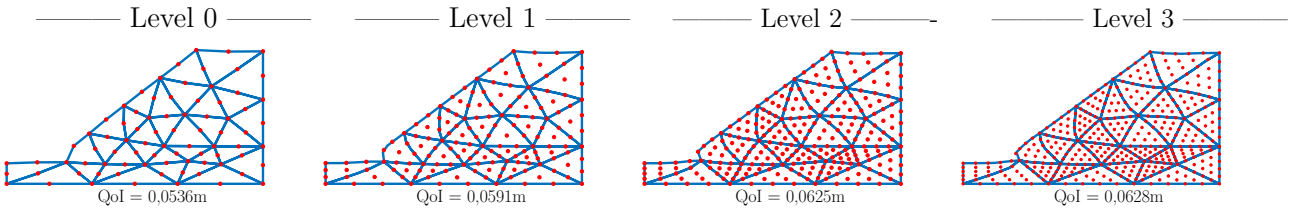


Figure 6: Displacement of the mesh and QoI for different samples of the random field.

5.2 Variance and Expected Value

In Figure 7 we show the sample variance over the levels $\mathbb{V}[P_\ell]$, the sample variance of the difference over the levels $\mathbb{V}[\Delta P_\ell]$, the expected value over the levels $\mathbb{E}[P_\ell]$ and the expected value of the difference over the levels $\mathbb{E}[\Delta P_\ell]$.

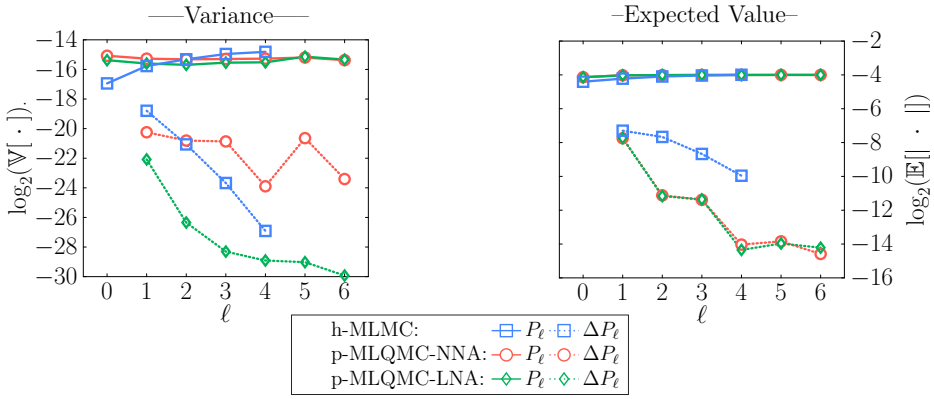


Figure 7: Variance and Expected Value over the levels.

As expected, we observe that $\mathbb{E}[P_\ell]$ remains constant over the levels, while $\mathbb{E}[\Delta P_\ell]$ decreases with increasing level. As explained in §3.1, multilevel methods are based on a variance reduction. In practice this means that the sample variance $\mathbb{V}[P_\ell]$ remains constant across the levels, while the sample variance of the difference over the levels $\mathbb{V}[\Delta P_\ell]$ decreases per increasing level. This is indeed what we observe for p-ML(Q)MC-LNA and h-ML(Q)MC. For p-ML(Q)MC-NNA we observe that $\mathbb{V}[\Delta P_\ell]$ does not decrease, but oscillates. From Figure 7, we can conclude that the choice of the evaluation points for the random field greatly influences the behavior of $\mathbb{V}[\Delta P_\ell]$ in the p-MLQMC method.

5.3 Runtimes

We show the absolute and relative runtime as a function of the user requested tolerance ε on the RMSE in Figure 8.

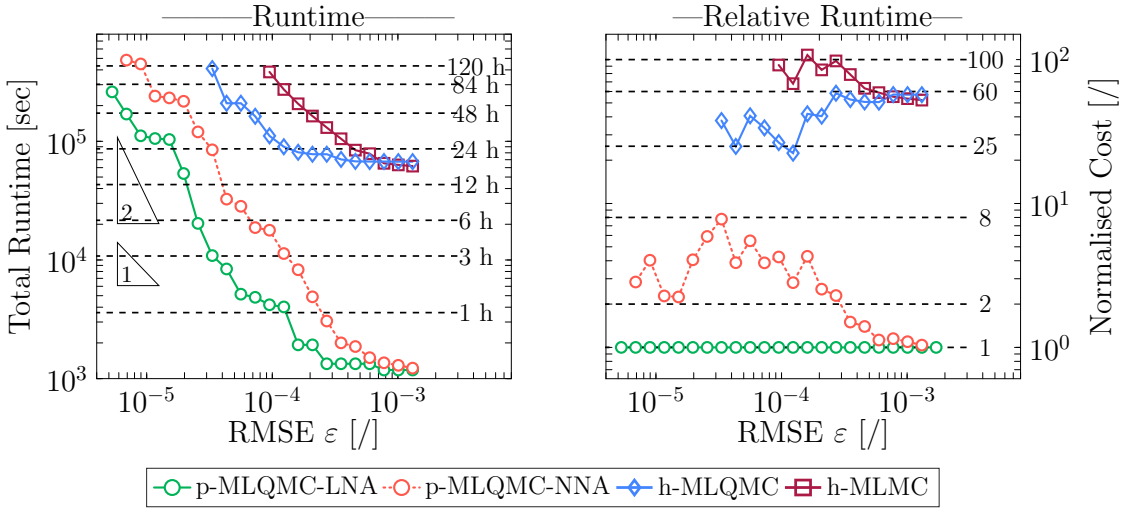


Figure 8: Absolute runtimes in function of requested user tolerance.

The results for the absolute runtime are expressed in seconds. For the relative runtime, we have normalised the computational cost of all the methods such that the results for p-MLQMC-LNA have unity cost for each tolerance. We observe that p-MLQMC combined with the LNA approach outperforms all other considered methods. p-MLQMC-LNA outperforms p-MLQMC-NNA by a factor 2 to 8. In addition, the p-refined Multilevel methods outperform the h-refined Multilevel methods. p-refined MLQMC achieves a speedup up to a factor 60 with respect to h-MLQMC and a factor 100 with respect to h-MLMC.

6 CONCLUSION

In this work, we have benchmarked the p-MLQMC method on a slope stability problem where the soil has a spatially varying uncertainty. We also investigated how the evaluation points of the random field are to be selected in the p-MLQMC method in order to obtain a lower computational cost. We distinguished two different approaches, the Non-Nested Approach and the Local Nested Approach. We showed that the approaches impact the variance reduction over the levels, and thus the total computational cost. p-MLQMC combined with LNA exhibits a much better decrease of $\mathbb{V}[\Delta P_\ell]$ due to a better correlation between the levels than with NNA. This is reflected in the total computational cost where the LNA approach outperforms NNA by a factor between 2 to 8. We also showed that the p-MLQMC-LNA method outperforms h-Multilevel Monte Carlo (h-MLMC) by a factor ranging between 60 and 100, and classic Multilevel quasi-Monte Carlo (h-MLQMC) by a factor 25 to 60. Of the considered approaches, the p-MLQMC-LNA method offers the lowest computational cost for a given tolerance on the RMSE.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support from the Research Council of KU Leuven through project C16/17/008 “Efficient methods for large-scale PDE-constrained optimization in the presence of uncertainty and complex technological constraints”. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

REFERENCES

[1] Giles, M. B. Multilevel Monte Carlo methods. *Acta Num.* (2015)**24**:259–328.
[2] Giles, M. B. Multilevel Monte Carlo path simulation. *Oper. Res.* (2008)**56(3)**:607–617.
[3] Blondeel, P., Robbe, P., Van hoorickx, C., François, S., Lombaert, G., and Vandewalle, S. p-refined Multilevel Quasi-Monte Carlo for Galerkin Finite Element Methods with Applications in Civil Engineering. *Algorithms* (2020)**13(5)**:1–30.
[4] Giles, M. B. and Waterhouse, B. J. Multilevel Quasi-Monte Carlo path simulation. *Rad. Ser. Comp. App.* (2009)**8**:1–18.
[5] Whenham, V., De Vos, M., Legrand, C., Charlier, R., Maertens, J., and Verbrugge, J.-C. Influence of soil suction on trench stability. In T. Schanz, editor, *Experimental Unsaturated Soil Mechanics*. Springer Berlin Heidelberg (2007) pages 495–501.
[6] Simo, J. C. and Taylor, R. L. A return mapping algorithm for plane stress elastoplasticity. *Int. J. Numer. Meth. Eng.* (1986)**22(3)**:649–670.
[7] de Borst, R., Crisfield, M. A., and Remmers, J. J. C. *Non Linear Finite Element Analysis of Solids and Structures*. Wiley, U.K. (2012).
[8] Kuo, F. Y., Scheichl, R., Schwab, C., Sloan, I. H., and Ullmann, E. Multilevel Quasi-Monte Carlo methods for lognormal diffusion problems. *Math. Comput.* (2017)**86(308)**:2827–2860.
[9] Kuo, F. Lattice rule generating vectors (2007). Online `https://web.maths.unsw.edu.au/~fkuo/lattice/index.html` and `https://web.maths.unsw.edu.au/~fkuo/lattice/lattice-32001-1024-1048576.3600`, accessed on 12/04/2019.
[10] Cliffe, K. A., Giles, M. B., Scheichl, R., and Teckentrup, A. L. Multilevel Monte Carlo methods and applications to elliptic pdes with random coefficients. *Comput. Vis. Sci.* (2011)**14(1)**:3.
[11] Lord, G. J., Powell, C. E., and Shardlow, T. *An Introduction to Computational Stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press (2014).
[12] Matthies, H. G., Brenner, C. E., Bucher, C. G., and Guedes Soares, C. Uncertainties in probabilistic numerical analysis of structures and solids-stochastic finite elements. *Struct. Saf.* (1997)**19(3)**:283–336.
[13] Blondeel, P., Robbe, P., François, S., Lombaert, G., and Vandewalle, S. On the selection of random field evaluation points in the p-MLQMC method. *arXiv Preprint* (2020).
[14] Geuzaine, C. and Remacle, J.-F. Gmsh: A 3-d finite element mesh generator with built-in pre- and post-processing facilities. *Int. J. Numer. Meth. Eng.* (2009)**79(11)**:1309–1331.
[15] MATLAB. *version 9.2.0 (R2017a)*. The MathWorks Inc., Natick, Massachusetts (2017).
[16] Robbe, P. Gaussianrandomfields.jl (2017). Online `https://github.com/PieterjanRobbe/GaussianRandomFields.jl`, accessed on 05/11/2020.
[17] Robbe, P. Multilevelestimators.jl (2018). Online `https://github.com/PieterjanRobbe/MultilevelEstimators.jl`, accessed on 05/11/2020.

A Model-based Damage Identification using Guided Ultrasonic Wave Propagation in Fiber Metal Laminates

Nanda. K. Bellam-Muralidhar[1], Dirk Lorenz[1]

[1]Technische Universität Braunschweig, Universitätsplatz 2,
38106 Braunschweig, Germany
e-mail: {n.bellam-muralidhar, d.lorenz}@tu-braunschweig.de

Key words: Guided ultrasonic waves, Fiber metal laminates, damage detection, Bayesian approach, Reduced-order model

Abstract: *Fiber metal laminates (FML) are lightweight hybrid structural materials that combine the ductile properties of metal with high specific stiffness of fiber reinforced plastics. These advantages led to a dramatic increase in such materials for aeronautical structures over the last few years. One of the most common and vulnerable defects in FML is impact-related delamination, often invisible to the human eye. Guided ultrasonic waves (GUW) show high potential for monitoring structural integrity and damage detection in thin-walled structures by using the physical phenomena of wave propagation interacting with the defects. The focus of this research project is on describing an inverse solution for the detection and characterization of defect in FML. Model-based damage analysis utilizes an accurate finite element model (FEM) of GUW interaction with the damage. The FEM is developed by the project partners from mechanics at Helmut-Schmidt-University in Hamburg, Germany, and will be treated as a black-box for further analysis. A Bayesian approach (Markov chain Monte Carlo) is employed to characterize the damage and quantify its uncertainties. This inference problem in a stochastic framework requires a very large number of forward solves. Therefore, a profound investigation is carried out on different reduced-order modeling (ROM) methods in order to apply a suitable technique that significantly improves the computational efficiency. The proposed method is well illustrated on a simpler case study for the damage detection, localization and characterization using 2D elastic wave equation. The damage in this case is modeled as a reduction in the wave propagation velocity. The inference problem utilizes a parameterized projection-based ROM coupled with a surrogate model instead of the underlying high-dimensional model.*

1 INTRODUCTION

Fiber reinforced plastics (FRPs), due to their very high strength to weight ratio, are often the favorite choice of material for engineers in building lightweight structures. Although FRPs possess high specific stiffness, they exhibit a weak bearing behaviour and impact resistance. In order to overcome these disadvantages of FRPs, fiber metal laminates (FMLs) are developed in the late 20th century. FMLs have the ability to demonstrate elastic-plastic behavior, as a corollary, a part of the energy introduced by impacts is absorbed by plastic deformations of the metal layers impeding its failure. The most commonly used FML is glass laminate aluminium reinforced epoxy (GLARE), which has excellent fatigue strength, high specific strength and low weight. However, due to its complex structure with different materials, its application is very challenging in terms of its production as well as the damage detection. Guided ultrasonic waves (GUW) have an immense potential in ensuring integrity of the structure and have been extensively used over the last decade. It has been shown that the propagation behavior of GUW changes when interacting with a damage.[1,2]

Numerical studies like finite element methods (FEM) play a crucial role for a well founded analysis of wave propagation and to assess the suitability of the GUW for damage detection. Furthermore, based on these numerical models, the requirements for sensors and actuators

https://doi.org/10.4995/YIC2021.2021.12684

can be derived with regard to their sensitivity through the solution of an inverse problem. Often high-dimensional FEM analysis will be very expensive which restricts us to use them directly for an inverse problem analysis. To alleviate this burden, projection-based model order reduction techniques are commonly used. There exits two approaches towards solving an inverse problem: the method of maximum likelihood estimation (MLE) and Bayesian estimation. The former results into the best single point estimation of the parameter while the latter models the parameter as a random variable and produces a probability density function (PDF) associated with it. The fact that the likelihood function is often extremely complicated with several local maxima, inhibits the use of MLE approach. Therefore, the inverse optimization problem is reformulated to a stochastic inference problem.[3]

Based on the current status of this research project, we consider a two-dimensional elastic hyperbolic wave equation as a test case, upon which a parameterized reduced order model is developed and Bayesian inference is applied to estimate the damage parameters. The remainder of this paper is organized as follows. Section 2 and section 3 describes the numerical model and the model order reduction approach used in this project respectively. Bayesian stochastic framework for damage identification is described in section 4. Section 5 discusses the results of parametric model reduction and damage characterization. Finally, conclusion and future works are given in section 6.

2 NUMERICAL MODEL

As the FEM model for wave propagation in FML is currently being developed by the project partners from mechanics group at Helmut-Schmidt-University in Hamburg, several potential inverse problem algorithms for damage characterization are simultaneously analyzed at Technical University Braunschweig. This led to the use of a simpler model, a two dimensional elastic hyperbolic wave equation, instead of the FEM model itself.

A 2D plate of isotropic and heterogeneous medium with multiple damages (two damages) is considered and the wave propagation is modeled by the equation:

$$\ddot{u} - c^2 \Delta u = f. \tag{1}$$

Here, $u(\mu, t)$ is displacement of the plate, Δ is the Laplacian in $\mathbb{R}^2$, $c(x, y)$ describes the wave velocity at any given point (x, y) on the plate, and $f(x, y, t)$ is the excitation function. The system is parameterized by $\mu \in \mathbb{R}^{3d}$, where d represents the number of damages and the factor 3 accounts for the number of parameters x, y, c for each of the damages.

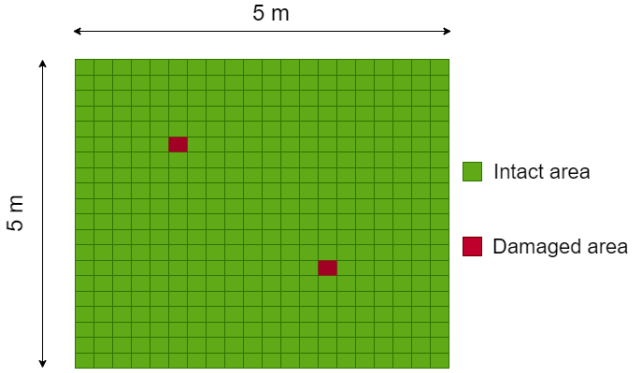


Figure 1: Distribution of wave propagation velocity in the plate

The plate has a side of 5 m and the damage was modeled as a change in the wave propagation velocity. Approximating the spatial derivatives using central difference operators, the

considered hyperbolic wave equation can be written as follows:

$$\ddot{u} - Au = f \tag{2}$$

with $A(\mu) = C(x,y)\Delta$. Here, $A(\mu) \in \mathbb{R}^{N\times N}$ is a parameterized symmetric positive definite matrix, $C(x,y) \in \mathbb{R}^{N\times N}$ is a matrix with squared wave velocities at any given x-y coordinate, and $u(\mu,t)$, $f(x,y,t) \in \mathbb{R}^N$ at any given instant of time $t \in [0,T]$. The plate is discretized with an element size of 0.02 m in both x and y directions. The wave propagation velocity in the intact area is assumed to be 0.5 ms^{-1}. Figure 1 represents the distribution of wave propagation velocity in the plate. The green intact area of the plate has the highest velocity (0.5 ms^{-1}) whereas the brown regions represent the damages with relatively lower propagation velocity. Based on Courant-Friedrichs-Lewy condition[4], the time step for numerical integration of the system is evaluated as 0.02 s in order to avoid the convergence issues.

3 PARAMETRIC MODEL ORDER REDUCTION

The numerical simulation of large-scale engineering problems requires a huge computational effort. To overcome this computational cost, projection-based model reduction techniques are often employed to reduce the model without a considerable loss of accuracy. The order reduction is accomplished by projecting the full order solution to the reduced order space using an orthogonal projection matrix $\Phi \in \mathbb{R}^{N\times n}$ such that,

$$u \approx u_h = \Phi\alpha \qquad\qquad \ddot{u} \approx \ddot{u}_h = \Phi\ddot{\alpha}. \tag{3}$$

where, u_h is the approximation of displacement u. Inserting (3) into (2) and projecting it onto the lower dimensional space leads to the reduced order problem,

$$\begin{aligned} \Phi\ddot{\alpha} - A\Phi\alpha &= f \\ \Phi^T\Phi\ddot{\alpha} - \Phi^T A\Phi\alpha &= \Phi^T f \\ \ddot{\alpha} - A_r\alpha &= f_r \end{aligned} \tag{4}$$

where, $\alpha(\mu,t) \in \mathbb{R}^n$, $A_r(\mu) \in \mathbb{R}^{n\times n}$ and $f_r(x,y,t) \in \mathbb{R}^n$ at any given instant of time t. The projection matrix Φ can be obtained by proper orthogonal decomposition (POD) of adaptively extracted features of the system. The displacements of the system that are numerically evaluated at m discrete time steps are saved in an observation matrix called snapshot matrix $U \in \mathbb{R}^{N\times m}$

$$U = \begin{bmatrix} | & | & & | \\ u(t_1) & u(t_2) & \dots & u(t_m) \\ | & | & & | \end{bmatrix}. \tag{5}$$

The snapshot matrix is then split into its basis and coefficients using singular value decomposition, $U = P\Sigma V^T$. Here, $\Sigma \in \mathbb{R}^{m\times m}$ is a diagonal matrix containing singular values σ_j, $P \in \mathbb{R}^{N\times m}$ is a left singular matrix with proper orthogonal modes (POMs) and $V \in \mathbb{R}^{m\times m}$ is a right singular matrix. The projection error incurred for considering upto σ_k singular values can be measured as

$$E = \frac{\sum_{j=k+1}^{m}\sigma_j^2}{\sum_{j=1}^{m}\sigma_j^2} \tag{6}$$

see Kerschen and Golinval, 2002[5]. Using (6), the required level of accuracy to capture the energy of the system can be chosen and subsequently, the number of POMs that enriches the projection matrix can also be decided

$$\Phi = [p_1, p_2, ..., p_n]. \tag{7}$$

As the governing equation depends on several parameters like x-y coordinates of the central position of the damage(s) and wave propagation velocity c in the damaged area(s), a parametric model order reduction (PMOR) is targeted. Due to its affine parameter dependency of the wave equation, PMOR involves an offline training phase, where the projection matrix Φ is built. This ensures that the projection matrix need not vary with the model parameters during the inverse problem analysis (online phase). After an intensive literature review, it was found that there was only one previous work that studied PMOR for hyperbolic wave equation using classical POD-Greedy approach[6]. However, in this project, an adaptive POD-Greedy procedure with kriging based on the work of Paul-Dubois-Taine[7] is applied to accomplish the PMOR through an optimized exploration strategy. This includes construction of a surrogate model for evaluating the reduced model error estimates, finding the largest error estimate, solving the full order model for the corresponding parameter sample with largest error estimate and subsequently updating the reduced-order model (ROM). The error estimate[6,8,9] at time t used in this procedure is as follows:

$$e_h(\mu) = \|u - u_h\| \leq \sqrt{(\frac{\gamma}{\beta}\|e_{h,0}\|^2 + \frac{1}{\beta}\|\dot{e}_{h,0}\|^2)} + \frac{1}{\sqrt{\beta}}\int_0^t \|r(s)\|\, ds \tag{8}$$

where, $e_{h,0}$ and $\dot{e}_{h,0}$ are the error estimate and its derivative at $t = 0$ respectively. β and γ are the coercivity and continuity constants of $A(\mu)$ and the residual is given by r with $s \in [0, T]$. After each greedy iteration, more error estimates are available to build the surrogate model. As the greedy algorithm proceeds, it eventually makes the error model more accurate and thereby finds a more optimal reduced space. It is essential to ensure that Φ remains orthogonal in this procedure. The offline phase can be terminated whenever the largest error estimate in an iteration is less than the specified threshold error value. Once the projection matrix Φ which is enriched with the required number of POMs is obtained, the solution can be evaluated using (3).

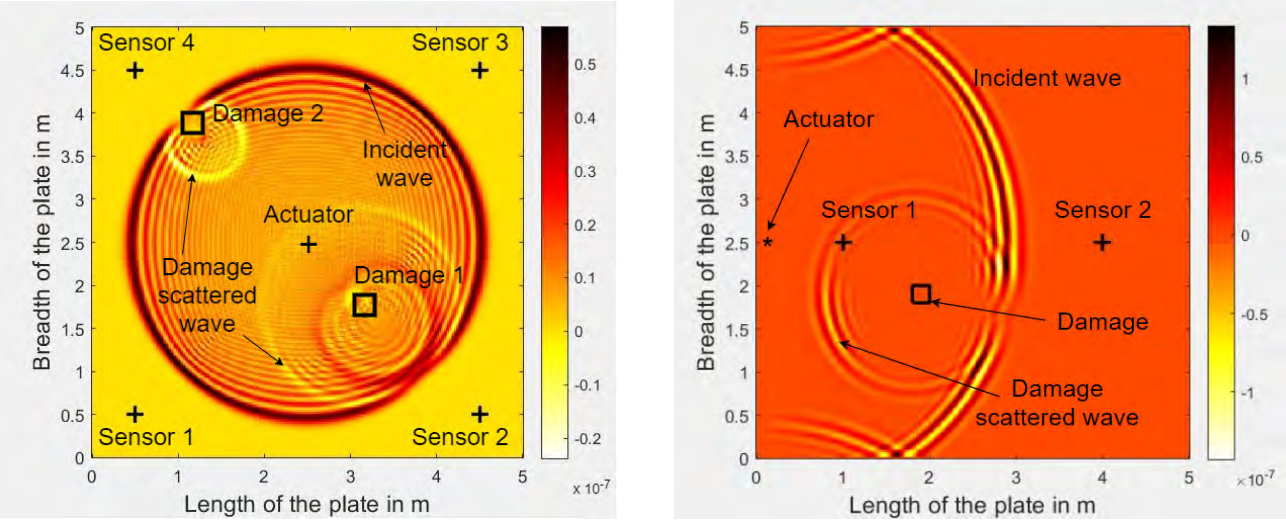
4 BAYESIAN INFERENCE FOR DAMAGE CHARACTERIZATION

Given the shape and size of the damage, the parameters $\mu = \{x, y, c\}$ for a damage are estimated using the Bayesian stochastic framework. The parameter vector μ is represented by a prior probability distribution $P(\mu|I)$ conditioned upon the prior knowledge I on the parameters. The posterior PDF $P(\mu|D, I)$ given data D and prior information is given by the Bayes' formula:

$$P(\mu|D, I) = \frac{P(D|\mu, I)P(\mu|I)}{P(D|I)} \tag{9}$$

where, $P(D|\mu, I)$ is the likelihood function that describes how likely are the candidate parameters to produce the given measurement data. The denominator $P(D|I)$ is called as marginal likelihood or evidence which ensures the integration of the posterior PDF results to 1. Unlike the deterministic approach that yields point estimates of the damage parameters, Bayesian inference method aims to describe posterior distribution for a given set of measurement data D. This allows the researcher to quantify the uncertainties associated with those parameters. The L_2 norm of the residual between the measurements and model output recorded at each sensor is considered to identify the damage. This quantity implicitly signifies the time-of-flight information. In this test case problem, four sensors are located at 4 corners of the plate with an actuator in the center that establishes a pitch-catch configuration to characterize the damage (see Figure 2(a)). The presence of model and measurement errors are described together by the variable ε. For convenience, ε is assumed to be an independent Gaussian variable with its mean at zero and standard deviation of σ_ε, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$. This uncertainty is added to the model output to generate synthetic data D which is used to carry out this inference problem.

The evaluation of posterior distribution is often analytically intractable and hence one tend to draw samples numerically from the posterior. A more commonly used procedure is Markov chain Monte Carlo (MCMC) method, which results into a dependent sequence of samples from a stationary distribution, asymptotically equal to that of the target distribution. Of several existing MCMC variants, we describe Metropolis-Hastings (MH) algorithm[10] and the same is used in this work.



(a) Multiple damages **(b)** Single damage

Figure 2: A snapshot of wave propagation and damage scattering

An arbitrary sample from the prior distribution μ_i is picked, then the algorithm produces a proposal candidate sample μ_i^* using a stochastic model $P(\mu_i^*|\mu_i)$ which denotes the probability of attaining μ_i^* conditioned upon the current sample. Both the samples μ_i^*, μ_i are then used to evaluate the ratio r:

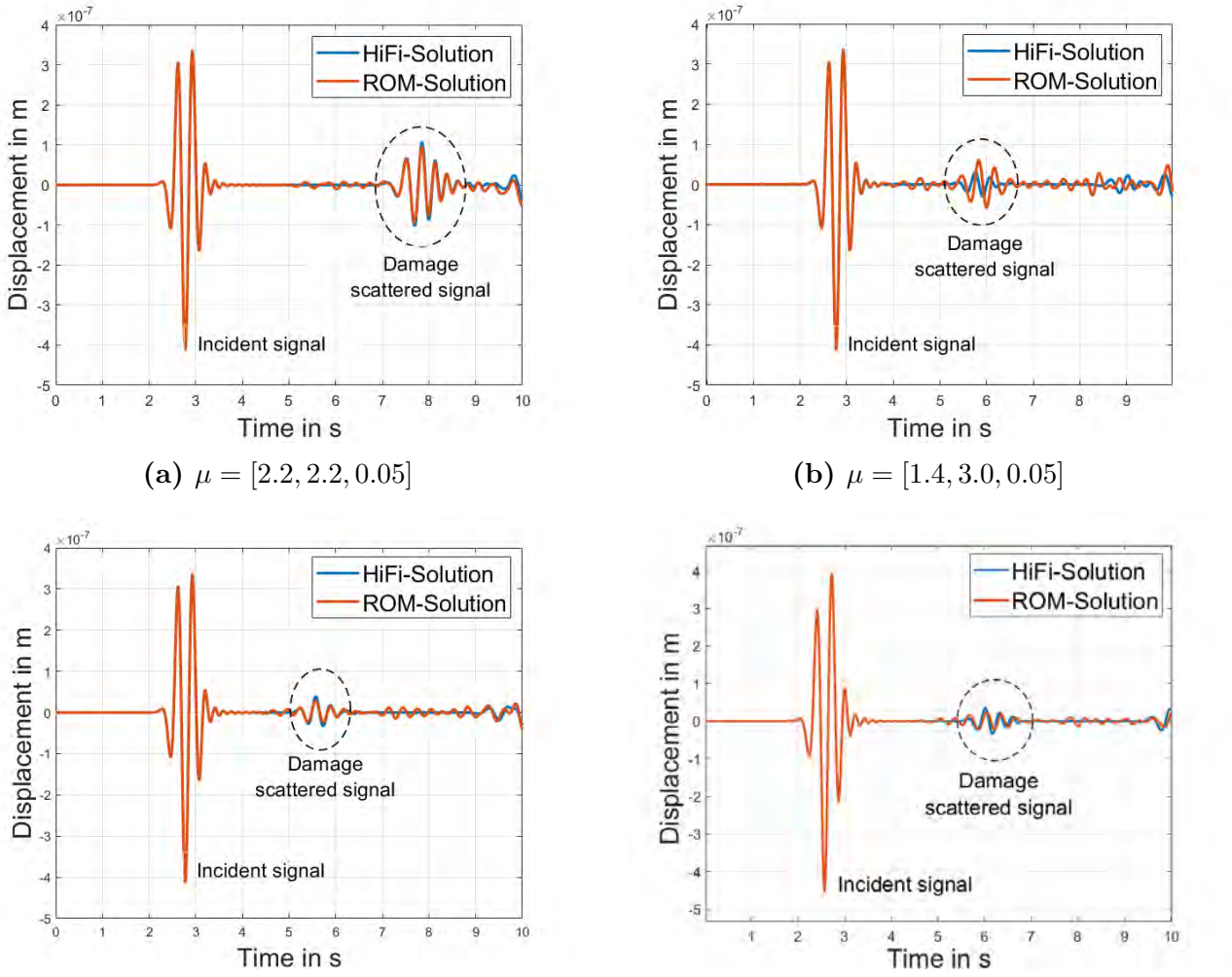
$$r = \frac{P(D|\mu_i^*, I) \times P(\mu_i^*|I) \times P(\mu_i|\mu_i^*)}{P(D|\mu_i, I) \times P(\mu_i|I) \times P(\mu_i^*|\mu_i)} \tag{10}$$

which is nothing but the ratio of their posteriors multiplied by the ratio of the candidate generating stochastic models. The current sample μ_i is updated to the proposal candidate sample μ_i^* if the ratio $r > z$, where z is a random value between 0 and 1. This acceptance-rejection sampling is iteratively carried out for a large specified number of samples, N_T, which ensures that the resulting Markov chain is stationary. Often, when starting from an arbitrary sample, there exist an initial phase of non-stationary period n_B while building the chain. This period is called 'burn-in' period and the samples until n_B have to be discarded to represent the final posterior distribution.

5 RESULTS

For convenience, model reduction is carried out on a slightly different setup with 2 sensors and one damage as shown in figure 2(b). The PMOR is trained in the parametric domain, $\mathcal{P} = \{x \times y \times c \mid [0.5, 4.5] \times [0.5, 4.5] \times [0.05, 0.45]\}$. The application of adaptive POD-Greedy algorithm as described in section 3 produced 800 global modes that could very well capture dynamics of the system for any sample μ from $\mathcal{P}$. Figure 3 depicts the reconstruction of wave signal measured at sensor 1, as shown in figure 2(b), for four randomly selected parameter samples in $\mathcal{P}$. On a 4-core Intel(R) Core(TM) i7-10510U CPU @ 1.80 GHz processor with 16 GB RAM, the evaluation of high-fidelity (HiFi) 2D elastic wave equation took 1.73 s while the

reconstruction using global modes took 1.92 s. Based on the computation time, the first instinct questions the purpose of model reduction. But the actual computational efficiency of PMOR can be realized when applied to a much sophisticated higher-dimensional problem, for example, the FEM-model of composite structures which involves the evaluation of individual element shape functions. However, the application of adaptive POD-Greedy PMOR on hyperbolic wave equation is very well demonstrated through this test case.



(a) $\mu = [2.2, 2.2, 0.05]$ **(b)** $\mu = [1.4, 3.0, 0.05]$

(c) $\mu = [1.68, 3.0, 0.3]$ **(d)** $\mu = [1.68, 2.6, 0.15]$

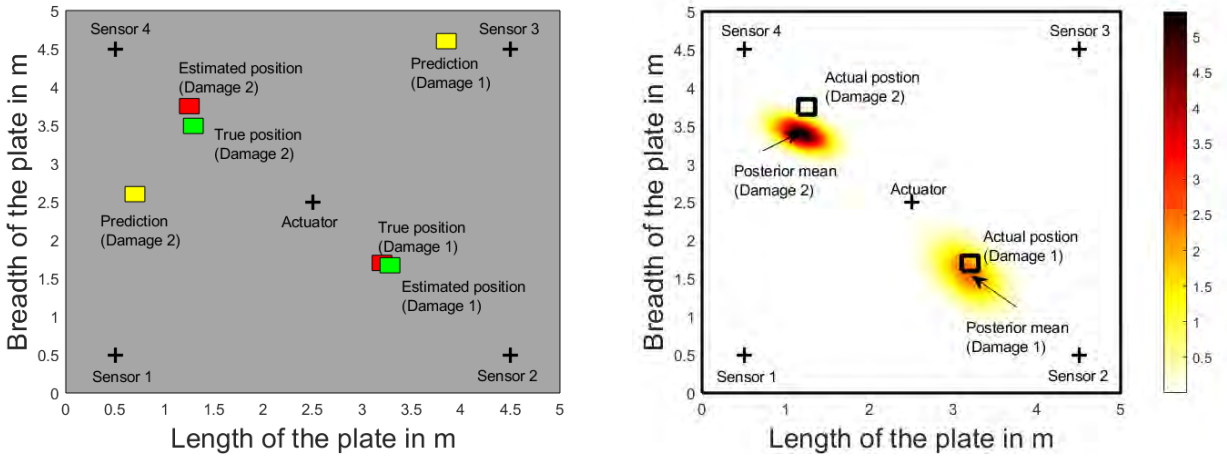
Figure 3: Comparison of reduced-order solution with high-fidelity solution

Bayesian inference for damage characterization was informed by the reduced-order model instead of the high-fidelity model. In order to embed multiple damages, the configuration shown in figure 2(a) is used to estimate the damage parameters and quantify their uncertainties. The MCMC approach described in section 3 is performed to localize and characterize the damages. The measurement data is obtained by adding a zero mean Gaussian-type errors to the model output. The data used for Bayesian inference is generated as follows:

$$D = M(\mu, t) + \varepsilon \tag{11}$$

where, $M(\mu, t)$ is the noise-free model output and ε is the normally distributed measurement error of 5%. The damage localization parameters, i.e., the x-y coordinates are uniformly distributed in $[0.5, 4.5]$ m and the localized wave propagation velocity in the damaged areas are

also uniformly distributed in $[0.05, 0.45]$ ms^{-1}. By MCMC-MH algorithm, 35000 samples from the posterior distribution are drawn. Figure 4(a) illustrates the point estimate and figure 4(b) shows the joint posterior PDF of the x-y coordinates of the center location of the damages in 2D view. The posterior PDF is not normalized with the evidence. The identified center locations of damage 1 and damage 2 are 0.21 m and 0.11 m respectively away from their actual locations accounting for relative errors of 4.2% and 2.2% with respect to the minimum sensor spacing. Similarly, the propagation velocities in damage 1 and damage 2 are estimated to have relative errors of 3.3% and 6.08% respectively. These small quantities of errors in parameter estimation describes the effectiveness of Bayesian inference approach.



(a) Deterministic estimate **(b)** Stochastic estimate

Figure 4: Illustration of the localization of the damages with center locations at $(3.2, 1.7)$ and $(1.25, 3.75)$ using the deterministic and stochastic approaches

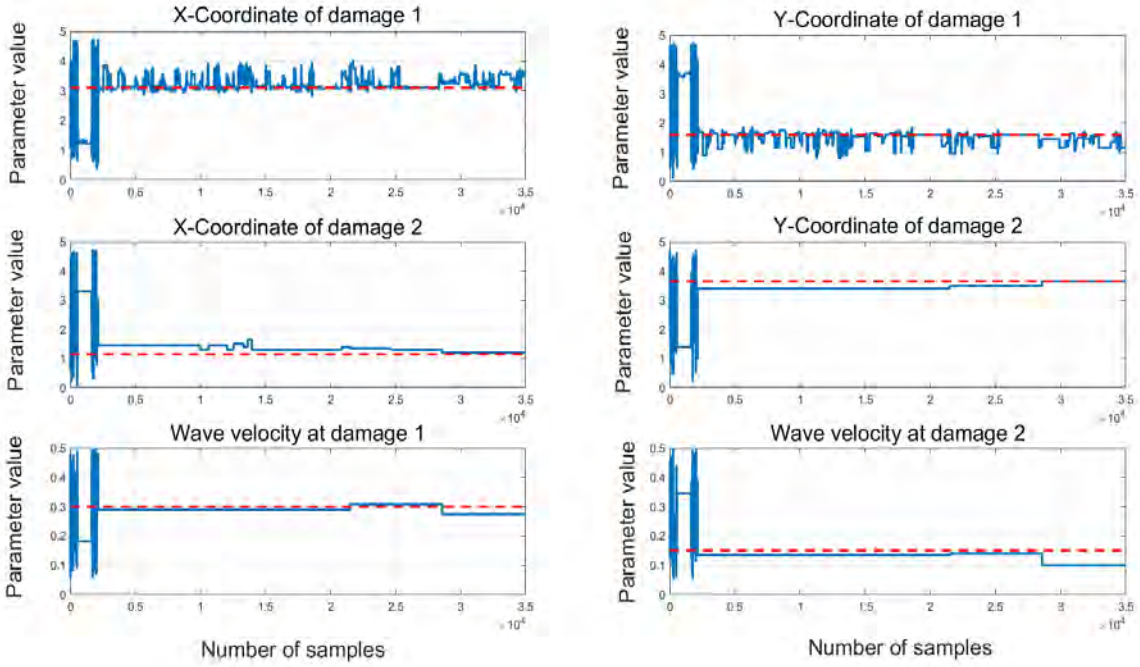


Figure 5: Trace plot of the estimated damage parameters using Bayesian approach with their true values indicated in red dashed lines.

Trace plots and histograms for the damage parameters corresponding to 5% measurement

error are shown in figure 5 and 6 respectively. Trace plots show the Markov chains for each parameter while the histograms represent their marginal posterior distributions. The true values are indicated in red dashed lines in each of these plots. The histograms indicate that all the parameters appear normally distributed with some skewness around their true values.

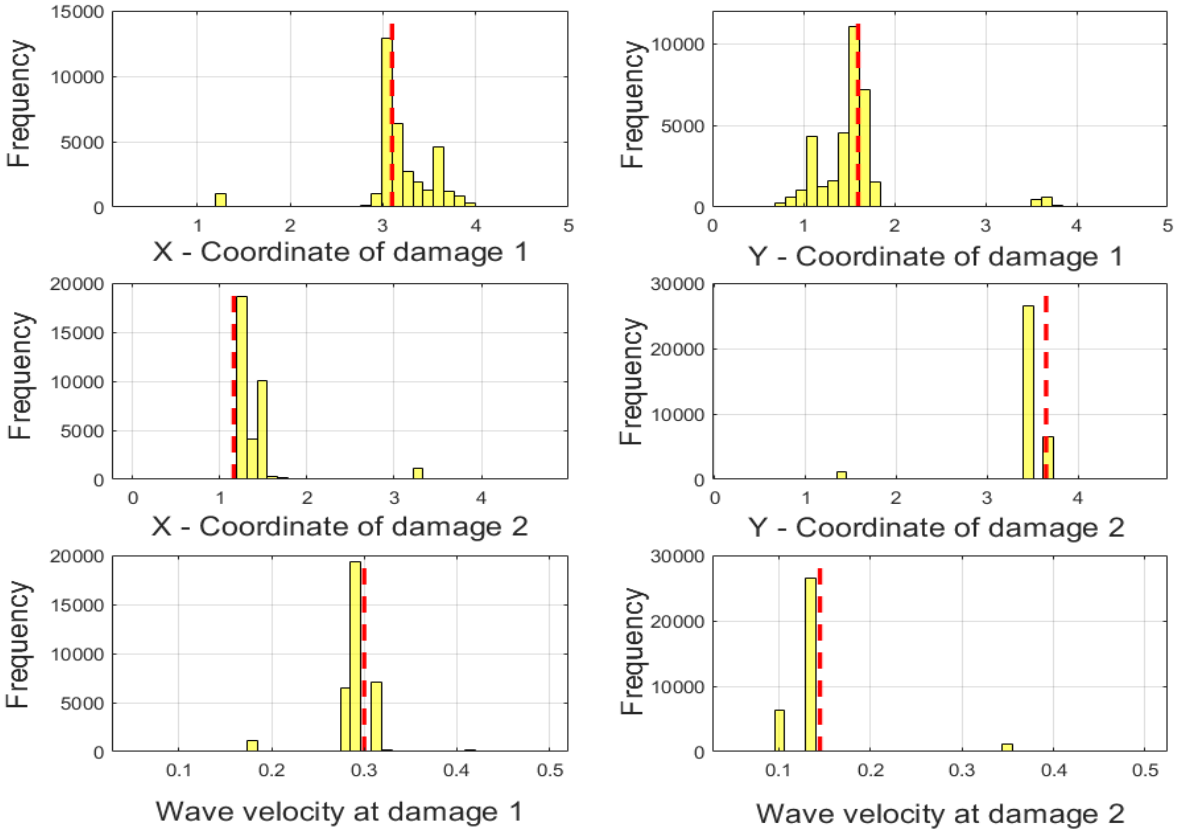


Figure 6: Histograms of the estimated damage parameters using Bayesian approach

Table 1: CoVs of damage parameters for 5% measurement error

Parameters	Damage 1	Damage 2
x	0.144	0.122
y	0.324	0.101
c	0.318	0.374

The uncertainties associated with parameters are analyzed using coefficient of variation (CoV)[11]. CoV is defined as the ratio of the standard deviation to the mean of the distribution. The CoVs for the estimated damage parameters are listed in table 1. The values of CoVs increase as the standard deviation of the error model in (11) increases. This is illustrated in the table 2 containing the CoVs for 7% and 10% error. Therefore, it is crucial to recognize the fact that the estimation uncertainties are positively correlated with the errors, i.e., uncertainties magnifies with the rise in modeling and measurement errors.

Table 2: CoVs of damage parameters for 7% and 10% measurement errors

	7 %		10 %	
Parameters	Damage 1	Damage 2	Damage 1	Damage 2
x	0.287	0.274	0.422	0.421
y	0.382	0.219	0.451	0.407
c	0.412	0.433	0.487	0.574

6 CONCLUSIONS AND FUTURE WORK

This work implemented considerable amount of the future work associated with this research project. An investigation of the applicability of parametric model-order reduction and Bayesian framework for damage identification is presented here. The effectiveness of the proposed approaches for model-order reduction and damage identification is validated by a numerical experiment on a two-dimensional elastic wave equation. The numerical study showed that not only the damages are localized but also the defects are well characterized, i.e., the wave velocities in the damaged areas are also estimated in this case study. The described adaptive POD-Greedy procedure with kriging produced a global projection matrix over the entire parametric domain which is used to evaluate the reduced-order solution. Subsequently, the Bayesian approach for inferring the damage parameters employed the reduced-order model rather than the high-fidelity model. Unlike pinpointing the estimate of parameters through the classical deterministic method, the Bayesian inference produced a distribution for the damage parameters. These distributions are not only used to identify the damage parameters with certain confidence levels but also to quantify their associated uncertainties.

Future work concerns the application of the presented methods on a finite element model of guided wave propagation in fiber metal laminate structures for the damage identification. Obviously, the anisotropic nature of the material could possibly impose challenges which need to be addressed. The number of parameters involved in the constitutive modeling of composite materials is usually large and hence a prior sensitivity analysis should be performed in order to ignore the less influential parameters in damage characterization. Also the embedded sensors and actuators could potentially act as defects, hence novel techniques are required to solve this artefact. It is ultimately aimed to detect, localize and characterize the class and degree of the damage in FML structures. This study can also be further extended to estimate the in-plane residual strength of the structure corresponding to the estimated parameters using a suitable artificial neural network model.

Acknowledgements

This research is funded by the Deutsche Forschungsgemeinschaft Research Unit 3022 under Grant No. LO1436/12-1.

REFERENCES

[1] R. Lammering, U. Gabbert, M. Sinapius, T. Schuster, P. Wierach (Eds)(2018) Lamb-Wave Based Structural Health Monitoring in Polymer Composites, Springer International Publishing.

[2] Li, Guoyi, Rajesh Kumar Neerukatti, and Aditi Chattopadhyay. "Ultrasonic guided wave propagation in composites including damage using high-fidelity local interaction simulation." Journal of Intelligent Material Systems and Structures 29.5 (2018): 969-985.

[3] Martin, James, et al. "A stochastic Newton MCMC method for large-scale statistical

inverse problems with application to seismic inversion." SIAM Journal on Scientific Computing 34.3 (2012): A1460-A1487.

[4] Courant, Richard, Kurt Friedrichs, and Hans Lewy. "Über die partiellen Differenzengleichungen der mathematischen Physik." Mathematische annalen 100.1 (1928): 32-74.

[5] Kerschen, Gaëtan, and Jean-Claude Golinval. "Physical interpretation of the proper orthogonal modes using the singular value decomposition." Journal of Sound and vibration 249.5 (2002): 849-865.

[6] Glas, Silke, Anthony T. Patera, and Karsten Urban. "A reduced basis method for the wave equation." International Journal of Computational Fluid Dynamics 34.2 (2020): 139-146.

[7] Paul-Dubois-Taine A, Amsallem D. An adaptive and efficient greedy procedure for the optimal training of parametric reduced-order models. International Journal for Numerical Methods in Engineering 2014.

[8] Bernardi, Christine, and Endre Süli. "Time and space adaptivity for the second-order wave equation." Mathematical Models and Methods in Applied Sciences 15.02 (2005): 199-225.

[9] Georgoulis, Emmanuil H., Omar Lakkis, and Charalambos Makridakis. "A posteriori L∞(L 2)-error bounds for finite element approximations to the wave equation." IMA Journal of Numerical Analysis 33.4 (2013): 1245-1264.

[10] Yildirim, Ilker. "Bayesian inference: Metropolis-hastings sampling." Dept. of Brain and Cognitive Sciences, Univ. of Rochester, Rochester, NY (2012).

[11] Yan, Gang. "A Bayesian approach for damage localization in plate-like structures using Lamb waves." Smart Materials and Structures 22.3 (2013): 035012.

MULTI-LEVEL SOLVERS FOR LARGE SPARSE LINEAR SYSTEMS

Multigrid Reduced in Time for Isogeometric Analysis

R. Tielen*, M. Möller* and C. Vuik*

* Delft Institute of Applied Mathematics (DIAM)
Delft University of Technology
Delft, the Netherlands
e-mail: r.p.w.m.tielen@tudelft.nl, m.moller@tudelft.nl, c.vuik@tudelft.nl

Key words: Multigrid Reduced in Time, Isogeometric Analysis, Multigrid

Abstract: *Isogeometric Analysis (IgA) can be seen as the natural extension of the Finite Element Method (FEM) to high-order B-spline basis functions. Combined with a time integration scheme within the method of lines, IgA has become a viable alternative to FEM for time-dependent problems. However, as processors' clock speeds are no longer increasing but the number of cores are going up, traditional (i.e., sequential) time integration schemes become more and more the bottleneck within these large-scale computations. The Multigrid Reduced in Time (MGRIT) method is a parallel-in-time integration method that enables exploitation of parallelism not only in space but also in the temporal direction. In this paper, we apply MGRIT to discretizations arising from IgA for the first time in the literature. In particular, we investigate the (parallel) performance of MGRIT in this context for a variety of geometries, MGRIT hierarchies and time integration schemes. Numerical results show that the MGRIT method converges independent of the mesh width, spline degree of the B-spline basis functions and time step size Δt and is highly parallelizable when applied in the context of IgA.*

1 INTRODUCTION

Isogeometric Analysis (IgA) [1] can be seen as the natural extension of the Finite Element Method (FEM) to high-order B-spline basis functions. By using the same building blocks (i.e., B-splines and Non-Uniform Rational B-Splines) as in Computer Aided Design (CAD), IgA tries to bridge the gap between CAD and FEM, resulting in a highly accurate represention of (curved) geometries. Furthermore, the use of high-order B-spline basis functions has shown to be advantageous in many applications [3, 4, 5] and the accuracy per degree of freedom (DOF) compared to FEM is significantly higher [6].

For time-dependent partial differential equations (PDEs), Isogeometric Analysis is often combined with a time integration scheme within the method of lines. However, as with all traditional time integration schemes, the latter part is sequential by design and hence, a bottleneck in numerical simulations. When the spatial resolution is increased to improve accuracy, the time step size has to be reduced accordingly to ensure stability of the overall method. At the same time, processors' clock speeds are no longer increasing, but the core count goes up, which calls for the parallelization of the calculation process to benefit from modern computer hardware. As traditional time integration schemes are sequential by nature, new parallel-in-time methods are needed to resolve this problem.

The Multigrid Reduced in Time (MGRIT) method [2] is a parallel-in-time algorithm based on multigrid reduction (MGR) techniques [7]. In contrast to space-time methods, in which time is considered as an extra spatial dimension, sequential time stepping is still necessary within MGRIT. Space-time methods have been combined in the literature with IgA [8]. Although very successful, a drawback of such methods is the fact that they are more intrusive on existing codes, while MGRIT just requires a routine to integrate the fully discrete problem between two time instances. Over the years, MGRIT has been studied in detail and applied to a variety of problems in the literature [9, 10].

https://doi.org/10.4995/YIC2021.2021.12219

To the best of our knowledge, this is the first publication that reports on combining Isogeometric Analysis and MGRIT and therefore our focus lies on the performance of MGRIT when different multigrid hierarchies, geometries and time integration schemes are considered within an IgA setting.

This paper is structured as follows: Section 2 presents our two-dimensional model problem and its spatial and temporal discretization. The MGRIT algorithm is then described in Section 3. Numerical results, including CPU timings, obtained for different geometries and time integration schemes are presented for different configurations of the MGRIT method in Section 4. Finally, conclusions are drawn in Section 5.

2 MODEL PROBLEM AND DISCRETIZATION

As a model problem, we consider the transient diffusion equation:

$$\partial_t u(\mathbf{x},t) - \kappa \Delta u(\mathbf{x},t) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, t \in [0,T]. \tag{1}$$

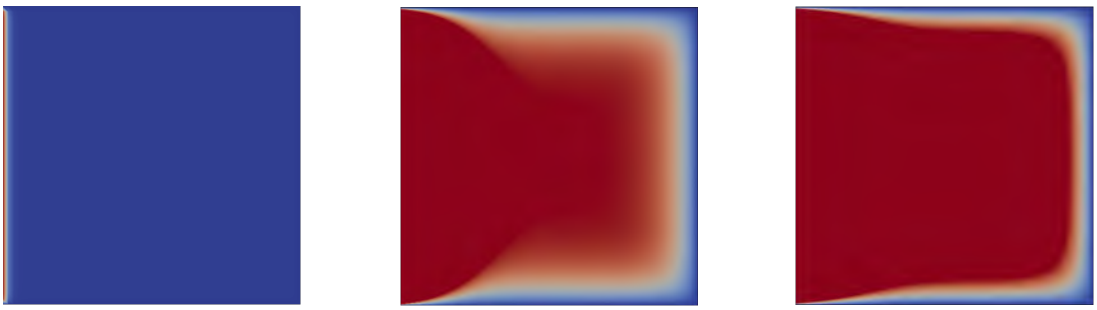
Here, κ denotes a constant diffusion coefficient, Ω the unit square (i.e., $[0,1]^2$) and $f \in L^2(\Omega)$ a source term. The above equation is complemented by initial conditions and both Dirichlet and Neumann boundary conditions:

$$u(\mathbf{x},0) = u^0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \tag{2}$$

$$u(\mathbf{x},t) = 0, \quad \mathbf{x} \in \partial\Omega \setminus \partial\Omega_W, t \in [0,T], \tag{3}$$

$$\frac{\partial u(\mathbf{x},t)}{\partial n} = 1, \quad \mathbf{x} \in \partial\Omega_W, t \in [0,T], \tag{4}$$

where Ω_W denotes the left boundary of Ω. Figure 1 denotes the solution of Equation (1) subject to these initial and boundary conditions at various time instances.



(a) $t = 0$ (b) $t = 0.01$ (c) $t = 0.02$

Figure 1: Solution to Equation (1) on the unit square at different times t.

First, we discretize Equation (1) (in time) by dividing the time interval $[0,T]$ in N_t subintervals of size Δt and applying the θ-scheme to the temporal derivative, which leads to the following equation to be solved at every time step:

$$u(\mathbf{x})^{k+1} - \kappa \Delta t \theta \Delta u(\mathbf{x})^{k+1} = u(\mathbf{x})^k + \kappa \Delta t (1-\theta) \Delta u(\mathbf{x})^k + \Delta t f(\mathbf{x}), \quad \mathbf{x} \in \Omega, k = 0, \ldots, N_t. \tag{5}$$

To obtain the variational formulation, let $\mathcal{V} = H_0^1(\Omega)$ be the space of functions in the Sobolev space $H^1(\Omega)$ that vanish on the boundary $\partial\Omega$. Equation (5) is multiplied with a test function $v \in \mathcal{V}$ and the result is then integrated over the domain Ω:

$$\int_\Omega \left(u^{k+1} v - \kappa \Delta t \theta \Delta u^{k+1} v \right) \mathrm{d}\Omega = \int_\Omega \left(u^k v + \kappa \Delta t (1-\theta) \Delta u^k v + \Delta t f v \right) \mathrm{d}\Omega. \tag{6}$$

Applying integration by parts on the second term on both sides of the equation results in

$$\int_\Omega \left(u^{k+1}v + \kappa\Delta t\theta\nabla u^{k+1}\cdot\nabla v\right)\ \mathrm{d}\Omega = \int_\Omega \left(u^{k+1}v - \kappa\Delta t(1-\theta)\nabla u^{k}\cdot\nabla v + \Delta t f v\right)\ \mathrm{d}\Omega, \tag{7}$$

for $\mathbf{x}\in\Omega, k=0,\dots,N_t$, where the boundary integral integral vanishes since $v=0$ on $\partial\Omega$. To parameterize the physical domain Ω, a geometry function $\mathbf{F}$ is then defined, describing an invertible mapping to connect the parameter domain $\Omega_0=(0,1)^2$ with the physical domain Ω:

$$\mathbf{F}:\Omega_0\to\Omega,\quad \mathbf{F}(\boldsymbol{\xi})=\mathbf{x}. \tag{8}$$

Provided that the physical domain Ω is topologically equivalent to the unit square, the geometry can be described by a single geometry function $\mathbf{F}$. In case of more complex geometries, a family of functions $\mathbf{F}^{(m)}$ $(m=1,\dots,K)$ is defined and we refer to Ω as a multipatch geometry consisting of K patches. For a more detailed description of the spatial discretization in Isogeometric Analysis and multipatch constructions, the authors refer to chapter 2 of [1].

Then, we express u at every time step by a linear combination of multivariate B-spline basis functions. Multivariate B-spline basis functions are defined as the tensor product of univariate B-spline basis functions $\phi_{i,p}$ $(i=1,\dots,N)$, which are uniquely defined on the parameter domain $(0,1)$ by an underlying knot vector $\Xi=\{\xi_1,\xi_2,\dots,\xi_{N+p},\xi_{N+p+1}\}$. Here, N denotes the number of univariate B-spline basis functions and p the spline degree. Based on this knot vector, the basis functions are defined recursively by the Cox-de Boor formula [11], starting from the constant ones

$$\phi_{i,0}(\xi)=\begin{cases}1 & \text{if } \xi_i\le\xi<\xi_{i+1},\\ 0 & \text{otherwise.}\end{cases} \tag{9}$$

Higher-order B-spline basis functions of order $p>0$ are then defined recursively

$$\phi_{i,p}(\xi)=\frac{\xi-\xi_i}{\xi_{i+p}-\xi_i}\phi_{i,p-1}(\xi)+\frac{\xi_{i+p+1}-\xi}{\xi_{i+p+1}-\xi_{i+1}}\phi_{i+1,p-1}(\xi). \tag{10}$$

The resulting B-spline basis functions $\phi_{i,p}$ are non-zero on the interval $[\xi_i,\xi_{i+p+1})$ and possess the partition of unity property. Furthermore, the basis functions are C^{p-m_i}-continuous, where m_i denotes the multiplicity of knot ξ_i. Throughout this paper, we consider a uniform knot vector with knot span size h, where the first and last knot are repeated $p+1$ times. As a consequence, the resulting B-spline basis functions are C^{p-1} continuous and interpolatory at both end points. Figure 2 illustrates both linear and quadratic B-spline basis functions based on such a knot vector.

Denoting the total number of multivariate B-spline basis functions $\Phi_{i,p}$ by N_{dof}, the solution u is thus approximated at every time step as follows:

$$u(\mathbf{x})\approx u_{h,p}(\mathbf{x})=\sum_{i=1}^{N_{\text{dof}}} u_i\Phi_{i,p}(\mathbf{x}),\quad u_{h,p}\in\mathcal{V}_{h,p}. \tag{11}$$

Here, the spline space $\mathcal{V}_{h,p}$ is defined, using the inverse of the geometry mapping $\mathbf{F}^{-1}$ as pull-back operator, as follows:

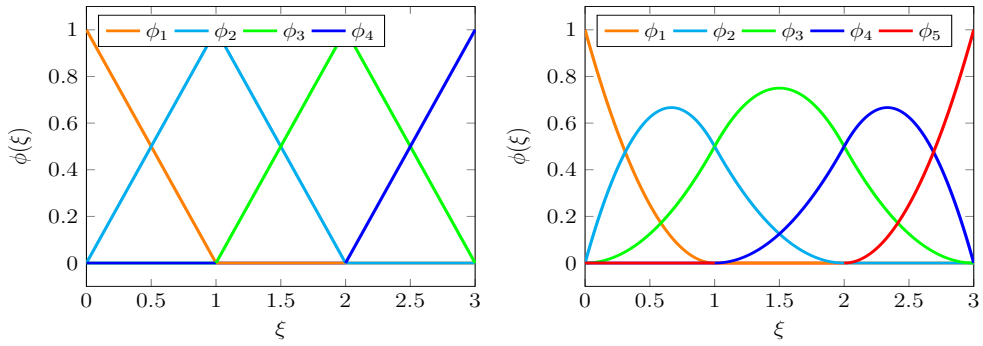
$$\mathcal{V}_{h,p}:=\operatorname{span}\left\{\Phi_{i,p}\circ\mathbf{F}^{-1}\middle|\ \Phi_{i,p}=0 \text{ on } \partial\Omega_0\right\}_{i=1,\dots,N_{\text{dof}}}. \tag{12}$$

By setting $v=\Phi_{j,p}$, Equation (7) can be written as follows:

$$(\mathbf{M}+\kappa\Delta t\theta\mathbf{K})\,\mathbf{u}^{k+1}=(\mathbf{M}-\kappa\Delta t(1-\theta)\mathbf{K})\,\mathbf{u}^{k}+\Delta t\mathbf{f},\ k=0,\dots,N_t, \tag{13}$$

where $\mathbf{M}$ and $\mathbf{K}$ denote the mass and stiffness matrix, respectively:

$$\mathbf{M}_{i,j}=\int_\Omega \Phi_{i,p}\Phi_{j,p}\ \mathrm{d}\Omega,\qquad \mathbf{K}_{i,j}=\int_\Omega \nabla\Phi_{i,p}\cdot\nabla\Phi_{j,p}\ \mathrm{d}\Omega,\quad i,j=1,\dots,N_{\text{dof}}. \tag{14}$$



(a) Linear B-spline basis functions (b) Quadratic B-spline basis functions

Figure 2: Linear and quadratic B-spline basis functions based on the knot vectors (a) $\Xi_1 = \{0, 0, 1, 2, 3, 3\}$ and (b) $\Xi_2 = \{0, 0, 0, 1, 2, 3, 3, 3\}$, respectively.

3 MULTIGRID REDUCED IN TIME

Instead of solving Equation (13) step-by-step directly, we apply the Multigrid Reduced in Time (MGRIT) method. For the ease of notation, we set $\theta = 1$ throughout the remainder of this section. Let $\Psi = (\mathbf{M} + \kappa\Delta t\mathbf{K})^{-1}$ denote the inverse of the left-hand side operator. Equation (13) can then be written as follows:

$$\mathbf{u}^{k+1} = \Psi\mathbf{M}\mathbf{u}^k + \mathbf{g}^{k+1},\ k = 0, \dots, N_t, \tag{15}$$

where $\mathbf{g}^{k+1} = \Psi\Delta t\mathbf{f}$. Setting $\mathbf{g}^0$ equal to the initial condition $u^0(\mathbf{x})$ projected on the spline space $\mathcal{V}_{h,p}$, the time integration method can be written as a linear system of equations:

$$\mathbf{A}\mathbf{u} = \begin{bmatrix} I & & & \\ -\Psi\mathbf{M} & I & & \\ & \ddots & \ddots & \\ & & -\Psi\mathbf{M} & I \end{bmatrix} \begin{bmatrix} \mathbf{u}^0 \\ \mathbf{u}^1 \\ \vdots \\ \mathbf{u}^{N_t} \end{bmatrix} = \begin{bmatrix} \mathbf{g}^0 \\ \mathbf{g}^1 \\ \vdots \\ \mathbf{g}^{N_t} \end{bmatrix} = \mathbf{g}. \tag{16}$$

The two-level MGRIT method combines the use of a cheap coarse-level time integration method with an accurate more expensive fine-level one which can be performed in parallel. That is, Equation (16) can be solved iteratively by introducing a coarse temporal mesh with time step size $\Delta t_C = m\Delta t_F$. Here, Δt_F coincides with the Δt from the previous sections and m denotes the coarsening factor. It can be observed that the solution of Equation (16) at the coarse-level times $T_0, T_1, \dots, T_{N_t/m}$ satisfies:

$$\mathbf{A}_\Delta\mathbf{u}_\Delta = \begin{bmatrix} I & & & \\ -(\Psi\mathbf{M})^m & I & & \\ & \ddots & \ddots & \\ & & -(\Psi\mathbf{M})^m & I \end{bmatrix} \begin{bmatrix} \mathbf{u}_\Delta^0 \\ \mathbf{u}_\Delta^1 \\ \vdots \\ \mathbf{u}_\Delta^{N_t/m} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_\Delta^0 \\ \mathbf{g}_\Delta^1 \\ \vdots \\ \mathbf{g}_\Delta^{N_t/m} \end{bmatrix} = \mathbf{g}_\Delta. \tag{17}$$

Here, $\mathbf{u}_\Delta^j = \mathbf{u}^{jm}$ and the vector $\mathbf{g}_\Delta$ is given by the original vector $\mathbf{g}$ multiplied by a restriction operator:

$$\mathbf{g}_\Delta = \begin{bmatrix} I & (\Psi\mathbf{M})^{m-1} & \dots & \Psi\mathbf{M} & I & & & & \\ & & & & & \ddots & & & \\ & & & & & & (\Psi\mathbf{M})^{m-1} & \dots & \Psi\mathbf{M} & I \end{bmatrix} \begin{bmatrix} \mathbf{g}^0 \\ \mathbf{g}^1 \\ \vdots \\ \mathbf{g}^{N_t} \end{bmatrix}. \tag{18}$$

A two-level MGRIT method solves the coarse system given by Equation (17) iteratively and computes the fine-level values in parallel within each interval (t_{jm}, t_{jm+m-1}). The coarse system is solved using the following residual correction scheme:

$$\mathbf{u}_{\Delta}^{(n+1)} = \mathbf{u}_{\Delta}^{(n)} + \mathbf{B}_{\Delta}^{-1}\left(\mathbf{g}_{\Delta} - \mathbf{A}_{\Delta}\mathbf{u}_{\Delta}^{(n)}\right), \tag{19}$$

where $\mathbf{B}_{\Delta}$ is the coarse-level equivalent of the matrix $\mathbf{A}$ based on Δt_C instead of Δt_F. More precisely, solving for $\mathbf{B}_{\Delta}$ gives the solution on the coarse mesh by coarse time stepping (using Δt_C), while solving for $\mathbf{A}_{\Delta}$ results in the solution on the coarse mesh by fine time stepping (using Δt_F). Here, the fine-level values are computed in parallel, denoted by the action of operator $\mathbf{A}_{\Delta}$. This is in contrast to the action of $\mathbf{B}_{\Delta}$ which typically is performed on a single processor.

The two-level MGRIT algorithm can be seen as a multigrid reduction (MGR) method that combines a coarse time stepping method with (parallel) fine time stepping within each coarse time interval. Here, the time stepping from a coarse point C to all neighbouring fine points is also referred to as F-relaxation [2]. On the other hand, time stepping to a C-point from the previous F-point is referred to as C-relaxation. It should be noted that both types of relaxation are highly parallel and can be combined leading to so-called CF- or FCF-relaxation.

3.1 Multilevel MGRIT method

Next, we consider the true multilevel MGRIT method. First, we define a hierarchy of L temporal meshes, where the time step size for the discretization at level l ($l = 0, 1, \ldots, L$) is given by $\Delta t_F m^l$. The total number of levels L is related to the coarsening factor m and the total number of fine steps Δt_F by $L = log_m(N_t)$. Let $\mathbf{A}^{(l)}\mathbf{u}^{(l)} = \mathbf{g}^{(l)}$ denote the linear system of equations based on the considered time step size at level l. The MGRIT method can then be written as follows:

Algorithm 1 MGRIT

if $l =$ L **then**
  Solve $\mathbf{A}^{(L)}\mathbf{u}^{(L)} = \mathbf{g}^{(L)}$
else
  Apply FCF-relaxation on $\mathbf{A}^{(l)}\mathbf{u}^{(l)} = \mathbf{g}^{(l)}$
  Restrict the residual $\mathbf{g}^{(l)} - \mathbf{A}^{(l)}\mathbf{u}^{(l)}$ using injection
  Call **MGRIT** setting $l \rightarrow l + 1$
  Update $\mathbf{u}^{(l)} \rightarrow \mathbf{u}^{(l)} + P\mathbf{u}^{(l+1)}$
end if

Here, the prolongation operator P is based on ordering the F-points and C-points, starting with the F-points. The matrix $\mathbf{A}$ can then be written as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{ff} & \mathbf{A}_{fc} \\ \mathbf{A}_{cf} & \mathbf{A}_{cc} \end{bmatrix}. \tag{20}$$

and the operator P is then defined as the "ideal interpolation" [2]:

$$P = \begin{bmatrix} -\mathbf{A}_{ff}\mathbf{A}_{fc} \\ \mathbf{I}_c \end{bmatrix}. \tag{21}$$

The recursive algorithm described above leads to a so-called V-cycle. However, as with standard multigrid methods, alternative cycle types (i.e., W-cycles, F-cycles) can be defined. At all levels of the multigrid hierarchy, the operators are obtained by rediscretizing Equation (1) using a different time step size.

4 NUMERICAL RESULTS

To assess the effectiveness of the MGRIT method when applied in combination with Isogeometric Analysis, we solve Equation (1) on the time domain $T = [0, 0.1]$, where the initial condition is chosen equal to zero and a right-hand side equal to one. This initial condition is adopted as well as initial guess at all times $t > 0$. The MGRIT method is said to have reached convergence if the relative residual at the end of an iteration is smaller or equal to 10^{-10}, unless stated otherwise.

Throughout this section, the MGRIT hierarchy, the domain of interest Ω and the time integration scheme are varied. The MGRIT hierarchies that will be adopted are two-level methods, a V-cycle and an F-cycle. As a domain, we consider the unit square (i.e., $\Omega = [0, 1]^2$), a quarter annulus defined in the first quadrant with inner radius of 1 and an outer radius of 2 and a multipatch geometry, see Figure 3. As a time integration scheme, we consider a value of θ of 0, 0.5 and 1 for the θ-scheme throughout this section, which corresponds to forward Euler, Crank-Nicolson and backward Euler, respectively.

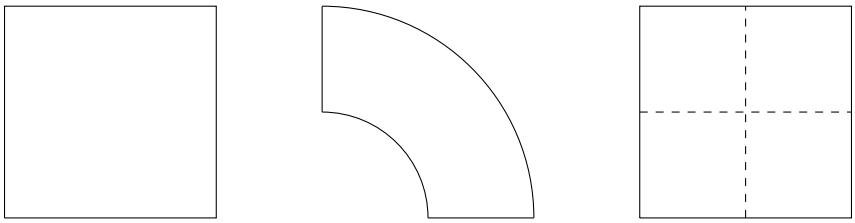


Figure 3: Spatial domains Ω considered throughout this section.

4.1 MGRIT hierarchies

First, we consider the MGRIT method using different hierarchies for the implicit case (i.e., backward Euler). At each time step, the linear system (Equation (13)) is solved by the Conjugate Gradient method. Table 1 shows the number of MGRIT iterations for different values of h and p when a two-level method, V-cycles or F-cycles are considered. Here, F-relaxation is applied at all levels of the MGRIT hierarchy. The number of time steps N_t for all configurations equals 100. For all three hierarchies, the number of MGRIT iterations needed to reach convergence is independent of h and p. The results obtained with a two-level method or F-cycles are identical and lead to a lower number of iterations compared to the use of V-cycles for all configurations.

	$p=2$			$p=3$			$p=4$			$p=5$		
	TL	V	F	TL	V	F	TL	V	F	TL	V	F
$h=2^{-4}$	7	9	7	7	9	7	7	9	7	7	9	7
$h=2^{-5}$	7	9	7	7	9	7	7	9	7	7	9	8
$h=2^{-6}$	8	9	8	8	9	8	8	9	8	8	9	8
$h=2^{-7}$	8	9	8	8	9	8	8	9	8	8	9	8

Table 1: Number of MGRIT iterations for solving the model problem when adopting a two-level (TL) method, V-cycles (V) or F-cycles (F).

Instead of varying the mesh width, the number of time steps can be increased as well. This is particularly interesting as MGRIT is a parallel-in-time method, where speed-ups will primarily come from parallelization in the temporal component. Table 2 shows the number of MGRIT iterations adopting different hierarchies for different numbers of time steps, different values of p

and $h = 2^{-6}$. For all configurations, the use of a two-level method or F-cycles leads to a lower number of iterations compared to the use of V-cycles. In particular, the number of iterations are independent of the number of time steps for all MGRIT hierarchies and comparable to the ones obtained when considering different values of the mesh width.

	$p=2$			$p=3$			$p=4$			$p=5$		
	TL	V	F	TL	V	F	TL	V	F	TL	V	F
$N_t = 250$	7	10	7	7	10	7	7	10	7	7	10	7
$N_t = 500$	7	10	7	7	10	7	7	10	7	7	10	7
$N_t = 1000$	7	11	7	7	11	7	7	11	7	7	11	7
$N_t = 2000$	7	11	7	7	11	7	7	11	7	7	11	7

Table 2: Number of MGRIT iterations for solving the model problem when adopting a two-level (TL) method, V-cycles (V) or F-cycles (F).

4.2 Varying geometries

Next, we apply MGRIT on a curved and multipatch geometry, respectively. Table 3 shows the number of V-cycles needed with MGRIT, using backward Euler for the time integration, for both geometries. Results can be compared to the ones presented in Table 2, showing identical iteration numbers for all geometries.

	Quarter Annulus				Multipatch			
	$p=2$	$p=3$	$p=4$	$p=5$	$p=2$	$p=3$	$p=4$	$p=5$
$N_t = 250$	10	10	10	10	10	10	10	10
$N_t = 500$	10	10	10	10	10	10	10	10
$N_t = 1000$	11	11	11	11	11	11	11	11
$N_t = 2000$	11	11	11	11	11	11	11	11

Table 3: Number of MGRIT iterations for solving Equation (1) on a quarter annulus and multipatch geometry when adopting V-cycles for varying time step sizes.

Table 4 shows the results when the number of time steps is kept constant ($N_t = 100$) for the quarter annulus and multipatch geometry when adopting V-cycles. Results can be compared to Table 1 and are (again) identical for all three geometries.

	Quarter Annulus				Multipatch			
	$p=2$	$p=3$	$p=4$	$p=5$	$p=2$	$p=3$	$p=4$	$p=5$
$h = 2^{-4}$	9	9	9	9	9	9	9	9
$h = 2^{-5}$	9	9	9	9	9	9	9	9
$h = 2^{-6}$	9	9	9	9	9	9	9	9
$h = 2^{-7}$	9	9	9	9	9	9	9	9

Table 4: Number of MGRIT iterations for solving Equation (1) on a quarter annulus and multipatch geometry when adopting V-cycles for varying mesh widths.

4.3 Time integration schemes

Next to the implicit backward Euler scheme, we have considered alternative time integration schemes as well. In this subsection, we consider the forward Euler and (second-order accurate)

Crank-Nicolson method. The use of explicit time integration schemes in the context of parallel-in-time integration is on the one hand highly relevant, as the required number of time steps needed to ensure stability is relatively high. On the other hand, coarsening with respect to the time step size might still exhibit stability issues at coarser levels. Therefore, explicit-implicit methods are often considered, where explicit time integration is applied on the fine level problem, while implicit methods are adopted at the coarser levels. The question remains to which extent the resulting MGRIT algorithm remains robust in the mesh width and/or spline degree.

Table 5 shows the number of MGRIT iterations for different numbers of time steps when adopting V-cycles and a mesh width of $h = 2^{-4}$. Here, forward Euler/Crank-Nicolson is applied at the fine level, while backward Euler is applied at the coarse levels. For some of the considered configurations, the resulting MGRIT method does not converge for forward Euler (indicated by '$*$'). It should be noted, however, that for these configurations, forward Euler applied as a sequential time integration scheme does not converge either, which is a direct consequence of the CFL condition. When the Crank-Nicolson method is applied the resulting MGRIT method converges in a relatively low number of iterations.

	forward Euler				Crank-Nicolson			
	$p=2$	$p=3$	$p=4$	$p=5$	$p=2$	$p=3$	$p=4$	$p=5$
$N_t = 250$	$*$	$*$	$*$	$*$	11	11	14	24
$N_t = 500$	13	$*$	$*$	$*$	11	11	11	12
$N_t = 1000$	13	13	$*$	$*$	11	11	11	11
$N_t = 2000$	13	13	13	$*$	11	11	11	11

Table 5: Number of MGRIT iterations for solving Equation (1) on the unit square using forward Euler and Crank-Nicolson when adopting V-cycles.

Table 6 shows the number of MGRIT iterations for a varying mesh width and 1000 time steps for both time integration methods. For many configurations, MGRIT using forward Euler does not convergence, while the Crank-Nicolson method converges for all configurations. A small dependency on h and p is, however, visible.

	forward Euler				Crank-Nicolson			
	$p=2$	$p=3$	$p=4$	$p=5$	$p=2$	$p=3$	$p=4$	$p=5$
$h = 2^{-3}$	13	13	13	14	11	11	11	12
$h = 2^{-4}$	13	13	$*$	$*$	11	11	11	11
$h = 2^{-5}$	$*$	$*$	$*$	$*$	11	11	13	23
$h = 2^{-6}$	$*$	$*$	$*$	$*$	13	28	52	88

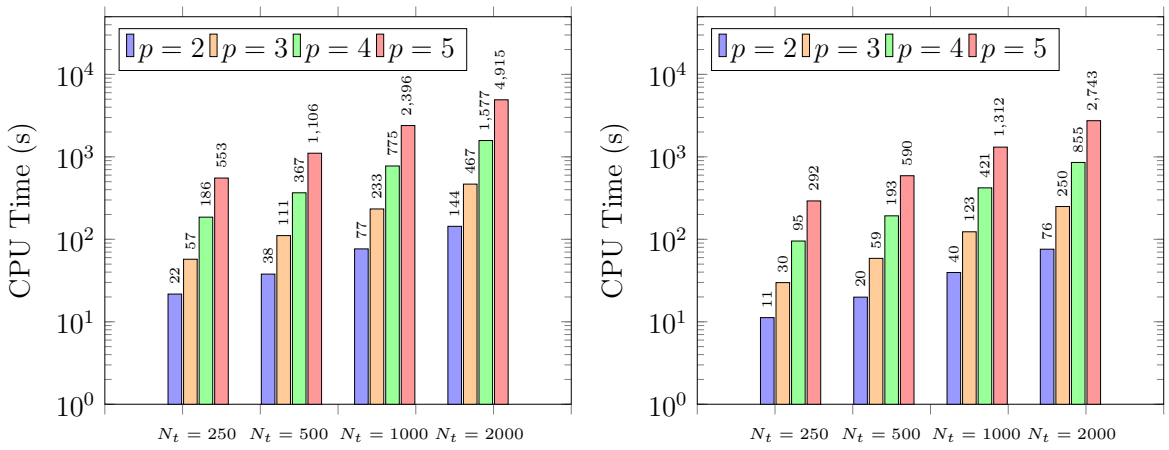
Table 6: Number of MGRIT iterations for solving Equation (1) on the unit square using forward Euler and Crank-Nicolson when adopting V-cycles.

4.4 CPU timings

Next to investigating the iteration numbers needed with MGRIT to reach convergence, CPU timings have been obtained as well. Here, we adopt V-cycles, a mesh width of $h = 2^{-6}$ and the unit square as our domain of interest. Note that the corresponding iteration numbers can be found in Table 2. The computations are performed on three nodes, which consist each of an Intel(R) i7-10700 (@ 2.90GHz) processor.

As shown in Figure 4a, doubling the number of time steps roughly doubles the time needed to reach convergence for all values of p. Furthermore, the CPU times significantly increase for higher values of p which is related to the spatial solves at every time step. It is known from the literature that standard iterative solvers have a detoriating performance for increasing values of p, leading to an increased number of CG iterations and, hence, higher computational costs.

In Figure 4b, results obtained adopting six cores can be found. In general, the same behavior can be observed with respect to the number of time steps and the spline degree. It should be noted, however, that doubling the number of cores significantly reduces the CPU time needed to reach convergence. More precisely, a reduction of $45 - 50\%$ can be observed when doubling the number of cores, implying the MGRIT algorithm is highly parallelizable.



(a) three cores

(b) six cores

Figure 4: CPU times for MGRIT using V-cycles and backward Euler on the unit square for a fixed problem size ($h = 2^{-6}$) adopting a different number of cores. The cores are evenly distributed over the nodes.

5 CONCLUSIONS

In this paper, we successfully combined Isogeometric Analysis with the Multigrid Reduced in Time (MGRIT) method to solve the time-dependent diffusion equation. Here, both (curved) single patch and multipatch geometries have been considered. Furthermore, different time integration methods and MGRIT hierarchies have been adopted. Numerical results show for all considered benchmarks that the MGRIT method converges independent of the considered mesh width h, spline degree p or time step size Δt. Furthermore, the use of an implicit time integration method has shown to be more robust compared to explicit time integration methods when applied within MGRIT. In general, a two-level hierarchy as well as the use of F-cycles leads to a slightly lower number of MGRIT cycles, but they are associated to higher costs per iteration. CPU timings show that the time needed to reach convergence does not only dependent on the number of time steps, but also on the spline degree of the B-spline basis functions when a standard iterative method is considered for the spatial solves. Future work will therefore focus on the use of state-of-the-art solvers for Isogeometric Analysis within MGRIT to mitigate this dependency. As increasing the number of CPUs significantly decreases the computational times, future research will focuss as well on the parallel performance of the MGRIT method and its comparison to traditional sequential time integration methods for large-scale simulations.

REFERENCES

[1] Hughes, T.J.R. and Cottrell, J.A. and Bazilevs, Y. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement *Computer Methods in Applied Mechanics and Engineering*, Vol. **194**, pp. 4135–4195, (2005).

[2] Falgout, R.D. and Friedhoff, S. and Kolev, Tz.V. and MacLachlan, S.P. and Schröder. J.B. Parallel Time Integration with Multigrid *SIAM Journal on Scientific Computing*, Vol. **36**, pp. C635–C661, (2014).

[3] Cottrell, J.A. and Reali, A. and Bazilevs, Y. and Hughes, T.J.R. Isogeometric Analysis of structural vibrations *Computer Methods in Applied Mechanics and Engineering*, Vol. **195**(41-43), pp. 5257–5296, (2006).

[4] Bazilevs, Y. and Calo, V.M. and Zhang, Y. and Hughes, T.J.R. Isogeometric Fluid-structure Interaction Analysis with Applications to Arterial Blood Flow *Computational Mechanics*, Vol. **38**, pp. 310–322, (2006).

[5] Wall, W.A. and Frenzel, M.A. and Cyron, C. Isogeometric structural shape optimization *Computer Methods in Applied Mechanics and Engineering*, Vol. **197**, pp. 2976–2988, (2008).

[6] Hughes, T.J.R. and Reali, A. and Sangalli, G. Duality and unified analysis of discrete approximations in structural dynamics and wave propagation: Comparison of p-method finite elements with k-method NURBS *Computer Methods in Applied Mechanics and Engineering*, Vol. **197**, pp. 4104–4124, (2008).

[7] Ries, M. and Trottenberg, U. and Winter, G. A note on MGR methods *Linear Algebra and its Applications*, Vol. **49**, pp. 1–26, (1983).

[8] Langer, U. and Moore, S.E. and Neumüller, M. Space-time isogeometric analysis of parabolic evolution problems *Computer Methods in Applied Mechanics and Engineering*, Vol. **306**, pp. 342–363, (2016).

[9] Günther, S. and Gauger, N.R. and Schröder, J.B. A non-intrusive parallel-in-time adjoint solver with the XBraid library *Computing and Visualization in Science*, Vol. **19**, pp. 85–95, (2018).

[10] Lecouvez, M. and Falgout, R.D. and Woodward, C.S. and Top, P. A parallel multigrid reduction in time method for power systems *C2016 IEEE Power and Energy Society General Meeting (PESGM)*, pp. 1–5, (2016).

[11] De Boor, C. *A practical guide to splines*. Springer, 2006.

On Total Reuse of Krylov Subspaces for an iterative FETI-solver in multirate integration

Andreas S. Seibold*, Daniel J. Rixen*, Javier del Fresno Zarza*

* Chair of Applied Mechanics
Technical University of Munich
Munich, Germany
e-mail: {andreas.seibold, rixen, javier.fresno}@tum.de

Key words: FETI, Structural dynamics, Krylov-Subspaces, Recycling, Multirate Time-Integration

Abstract: *In this work, we adapt a Total Reuse of Krylov Subspaces for usage in a GMRES-solver and apply it to nonlinear structural-dynamics examples. These examples are then solved by a multirate FETI-method, the nonlinear BGC-macro method, which allows local subcycling in time within substructures, such that local time-stepping is performed between synchronization-time-steps. In these proposed examples, we show that the reuse-method reduces the total number of GMRES-iterations and shifts the eigenvalue-spectrum of the global system towards smaller eigenvalues.*

1 INTRODUCTION

Substructuring methods are widely valued for parallelizing large structural mechanics problems and a popular non-overlapping dual domain decomposition method is the *Finite Elements Tearing and Interconnecting* (FETI) method [4]. In cases of local computationally expensive dynamics in a substructure, e.g. due to local damage or contact, it might be favorable to adjust the time-step-sizes locally. For such an asynchronous or multirate time-integration, domain-decomposition-based methods have been developed, such as the linear subcycling-based GC-method by Gravouil and Combescure [10]. However, this method suffers from energy-dissipation and therefore the non-dissipative linear and nonlinear PH-methods by Prakash and Hjelmstad [15] and the linear BGC-macro [3] have been developed. Recently a nonlinear version of the BGC-macro method has been proposed [18] and applied to an iterative FETI-solver equipped with a Dirichlet-like preconditioner [19]. Hence, the next natural step is to further improve solver-efficiency by applying recycling-techniques to this new problem. In this work, we adapt a *Total Reuse of Krylov Subspaces* (TRKS) approach [9], successfully applied to linear and nonlinear structural dynamics in [12, 17], to a GMRES-solver and investigate its applicability to the nonlinear BGC-macro method.

In Section 2.1, we introduce the applied multirate-method nonlinear BGC-macro and, in Section 2.2, the TRKS and its application in a GMRES is described. Finally, we show in Section 3 numerical examples with the described methods and conclusions in Section 4.

2 FETI for nonlinear structural dynamics

For the parallelization of a Finite Elements discretized structural dynamics problem, we divide the structure spacially along the element's edges in non-overlapping substructures $\Omega^{(s)}$. These substructures are connected with Lagrange-multipliers $\vec{\lambda}$, that can be viewed as interface-forces, as stated in the FETI method [4]. In Section 2.1, we give a brief introduction to the governing equations and the multirate BGC-macro method and in Section 2.2, we describe the application of a TRKS method to a GMRES solver.

https://doi.org/10.4995/YIC2021.2021.12306

2.1 Multirate with nonlinear BGC-macro method

Throughout this work, we consider different time-step-sizes in each substructure, which is referred to as multirate or asynchronous time-integration. As depicted in Figure 1 for two substructures A and B, the global time-integration with time-steps n is sub-cycled with N_j smaller time-steps with size $\Delta t^{(B)} = \frac{\Delta t^{(A)}}{N_j}$ on a micro-substructure.

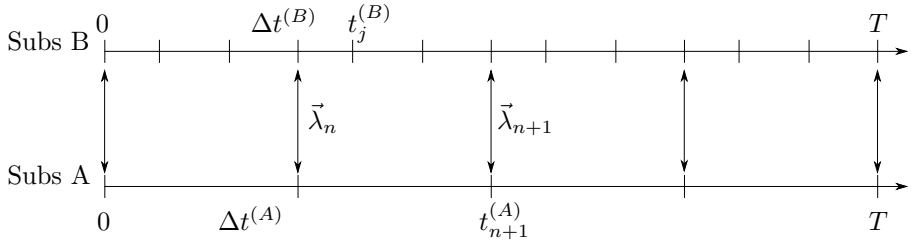


Figure 1: Multirate time-discretization.

The global time-steps are also referred to as macro-time-steps and the Lagrange-multipliers are interpolated linearly onto the local micro-timesteps

$$\vec{\lambda}_j = \left(1 - \frac{j}{N_j}\right) \vec{\lambda}_{n-1} + \frac{j}{N_j} \vec{\lambda}_n \tag{1}$$

resulting in the local differential equations of motion, written as a force-residual ${}_f\vec{res}_j^{(s)}$ here

$${}_f\vec{res}_j^{(s)} = \mathbf{M}^{(s)} \ddot{\vec{q}}_j^{(s)} + \vec{f}_{int}(\vec{q}_j^{(s)}) + \mathbf{B}^{(s)^T} \vec{\lambda}_j - \vec{f}_{ext}^{(s)}(t_j) = \vec{0} \tag{2}$$

at a discrete time-step j with a mass-matrix $\mathbf{M}^{(s)}$, nonlinear internal forces $\vec{f}_{int}$ and external forces $\vec{f}_{ext}^{(s)}$, as well as displacements $\vec{q}^{(s)}$, velocities $\dot{\vec{q}}^{(s)}$ and accelerations $\ddot{\vec{q}}^{(s)}$. The Lagrange-multipliers $\vec{\lambda}$ are applied to the local *degrees of freedom* (dof) by a signed Boolean matrix $\mathbf{B}^{(s)}$. The local solutions are then synchronized at the macro-time-scale, which is formulated by requiring the interface-velocities to coincide at the macro-timestep n in the interface-residual ${}_I\vec{res}_n$

$${}_I\vec{res}_n = \sum_{s=1}^{N_s} \mathbf{B}^{(s)} \dot{\vec{q}}_n^{(s)} = \vec{0}. \tag{3}$$

This approach is known from the linear BGC-macro method [3], which has been recently extended to nonlinear models [18]. To solve both equations (2) and (3), we choose as time-integration scheme one of the most popular ones, namely the Newmark-β scheme [14]

$$\begin{aligned}
{}_a\vec{res}_j^{(s)} &= -\frac{1}{\gamma \Delta t} \dot{\vec{q}}_{j-1}^{(s)} - \frac{1-\gamma}{\gamma} \ddot{\vec{q}}_{j-1} + \frac{1}{\gamma \Delta t} \dot{\vec{q}}_j^{(s)} - \ddot{\vec{q}}_j^{(s)} = \vec{0} \\
{}_d\vec{res}_j^{(s)} &= \vec{q}_{j-1}^{(s)} + \left(1 - \frac{\beta}{\gamma}\right) \Delta t \dot{\vec{q}}_{j-1}^{(s)} + \left(\frac{1}{2} - \frac{\beta}{\gamma}\right) \Delta t^2 \ddot{\vec{q}}_{j-1}^{(s)} + \frac{\beta}{\gamma} \Delta t \dot{\vec{q}}_j^{(s)} - \vec{q}_j^{(s)} = \vec{0},
\end{aligned}$$

with $\beta \in [0, 1/4]$, $\gamma \in [0, 1/2]$. The equations have been reformulated in residual-form ${}_a\vec{res}_j^{(s)}$ and ${}_d\vec{res}_j^{(s)}$ here. Analogously to the classical single-rate FETI in Farhat e.a. [6] and the

PH-method [15], all these equations are linearized for $\ddot{\tilde{q}}_j^{(s)}$, $\dot{\tilde{q}}_j^{(s)}$, $\tilde{q}_j^{(s)}$ and $\vec{\lambda}_n$, resulting in

$$\tilde{\mathbf{M}}^{(s)} = \begin{bmatrix} \mathbf{M}^{(s)} & \mathbf{0} & \mathbf{K}^{(s)} \\ -\gamma\Delta t^{(s)}\mathbf{I} & \mathbf{I} & \mathbf{0} \\ -\beta\Delta t^{(s)^2}\mathbf{I} & \mathbf{0} & \mathbf{I} \end{bmatrix} \qquad \tilde{\mathbf{C}}^{(s)} = \begin{bmatrix} \mathbf{B}^{(s)^T} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \qquad \tilde{\vec{r}}_j^{(s)} = \begin{bmatrix} {}_f\vec{r}\vec{e}s_j^{(s)} \\ {}_a\vec{r}\vec{e}s_j^{(s)} \\ {}_d\vec{r}\vec{e}s_j^{(s)} \end{bmatrix}$$

$$\mathbf{N}^{(s)} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\Delta t^{(s)}(1-\gamma)\mathbf{I} & -\mathbf{I} & \mathbf{0} \\ -\Delta t^{(s)}(1/2-\beta)\mathbf{I} & -\Delta t^{(s)}\mathbf{I} & -\mathbf{I} \end{bmatrix} \qquad \tilde{\mathbf{B}}^{(s)^T} = \begin{bmatrix} \mathbf{0} \\ \mathbf{B}^{(s)^T} \\ \mathbf{0} \end{bmatrix} \qquad \tilde{\vec{q}}_j^{(s)} = \begin{bmatrix} \ddot{\tilde{q}}_j^{(s)} \\ \dot{\tilde{q}}_j^{(s)} \\ \tilde{q}_j^{(s)} \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \tilde{\mathbf{M}}_1^{(s)} & & & \\ \mathbf{N}^{(s)} & \tilde{\mathbf{M}}_2^{(s)} & & \\ & \ddots & \ddots & \\ & & \mathbf{N}^{(s)} & \tilde{\mathbf{M}}_{N_j}^{(s)} \end{bmatrix}}_{\mathbf{A}^{(s)}} \underbrace{\begin{bmatrix} \Delta\tilde{\vec{q}}_1^{(s)} \\ \Delta\tilde{\vec{q}}_2^{(s)} \\ \vdots \\ \Delta\tilde{\vec{q}}_{N_j}^{(s)} \end{bmatrix}}_{\Delta\bar{\tilde{q}}^{(s)}} + \underbrace{\begin{bmatrix} \frac{1}{N_j}\tilde{\mathbf{C}}^{(s)} \\ \frac{2}{N_j}\tilde{\mathbf{C}}^{(s)} \\ \vdots \\ \frac{N_j}{N_j}\tilde{\mathbf{C}}^{(s)} \end{bmatrix}}_{\bar{\mathbf{C}}^{(s)}} \Delta\vec{\lambda}_n = \underbrace{\begin{bmatrix} -\tilde{\vec{r}}_1^{(s)}(\tilde{\vec{q}}_0^{(s)}, \tilde{\vec{q}}_1^{(s)}, \vec{\lambda}_1) \\ -\tilde{\vec{r}}_2^{(s)}(\tilde{\vec{q}}_1^{(s)}, \tilde{\vec{q}}_2^{(s)}, \vec{\lambda}_2) \\ \vdots \\ -\tilde{\vec{r}}_{N_j}^{(s)}(\tilde{\vec{q}}_{N_j-1}^{(s)}, \tilde{\vec{q}}_{N_j}^{(s)}, \vec{\lambda}_n) \end{bmatrix}}_{\bar{\vec{r}}^{(s)}}$$

This local problem is solved for the local states $\tilde{\vec{q}}_j^{(s)}$ and inserted into the compatibility condition

$$\bar{\mathbf{B}}^{(s)}\Delta\bar{\vec{q}}^{(s)} = \begin{bmatrix} \mathbf{0} & \dots & \mathbf{0} & \tilde{\mathbf{B}}^{(s)} \end{bmatrix} \Delta\bar{\vec{q}}^{(s)} = -{}_I\vec{r}\vec{e}s_n(\tilde{\vec{q}}_{N_j}^{(s)})$$

$$\vec{r}_k = \underbrace{\sum_{s=1}^{N_s} \bar{\mathbf{B}}^{(s)}\mathbf{A}^{(s)^{-1}}\bar{\mathbf{C}}^{(s)}}_{\mathbf{F}} \Delta\vec{\lambda}_n - \underbrace{\left(\sum_{s=1}^{N_s} \bar{\mathbf{B}}^{(s)}\mathbf{A}^{(s)^{-1}}\bar{\vec{r}}^{(s)} + {}_I\vec{r}\vec{e}s_n(\tilde{\vec{q}}_{N_j}^{(s)})\right)}_{\vec{d}} = \vec{0},$$

where $\mathbf{A}^{(s)}$ is invertible due to the regularizing nature of the mass-matrix. This is the so-called interface-problem and its only unknown are the global interface-forces. In contrast to the classical FETI-method, the interface-operator $\mathbf{F}$ is non-symmetric, which implies that we have to use a *Generalized Minimal Residual* (GMRES) method [16] here, as a Conjugate Gradient requires the problem to be symmetric.

2.2 TRKS for GMRES

The general idea of recycling relies on constructing an auxiliary coarse-space $\mathbf{C}$ similarly to the natural or kernel coarse space in FETI for statical problems. Hence, this is usually referred to as two-level FETI [7]. This auxiliary coarse-space can be built on FETI-search-directions from earlier solver-runs in which case these search-directions are projected out from the overall interface-problem, resulting in the TRKS [9]. This would lead to a reduced solution space and the iterative solver will not have to find the full set of search-directions every time anew. The auxiliary coarse-space adds another constraint

$$\mathbf{C}^T\mathbf{F}^T\vec{r}_k = \vec{0} \tag{4}$$

to the interface-problem with search-space $\mathbf{C}$ and constraint-space $\mathbf{FC}$ according to Gaul [8], where k is the GMRES-iteration counter. Here, $\mathbf{C}$ contains $l2$-orthonormal search-directions from previous GMRES-solver-runs. This coarse-space $\mathbf{C}$ is filled up until a predefined coarse-space-size N_C is reached. To fulfill constraint (4) in each iteration, we construct an auxiliary

coarse-grid projector $\mathbf{P}_C$. In the original TRKS for FETI, the projector was described for symmetric systems and a Conjugate Gradient method [9]. In our case, the projector

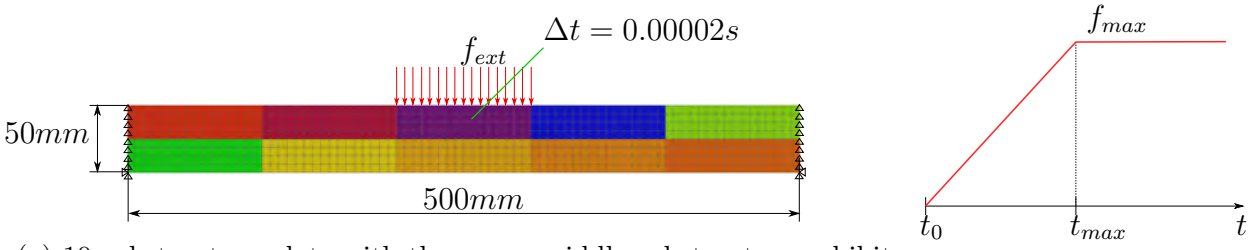
$$\mathbf{P}_C = \mathbf{I} - \mathbf{FC}\left(\mathbf{C}^T\mathbf{F}^T\mathbf{FC}\right)^{-1}\mathbf{C}^T\mathbf{F}^T \qquad \tilde{\mathbf{P}}_C = \mathbf{I} - \mathbf{C}\left(\mathbf{C}^T\mathbf{F}^T\mathbf{FC}\right)^{-1}\mathbf{C}^T\mathbf{F}^T\mathbf{F}$$

required some modifications for general matrices $\mathbf{F}$, as it is described in [8]. The projector $\tilde{\mathbf{P}}_C$ is required here for correcting the deflated solution.

This projector is then incorporated into the non-preconditioned GMRES algorithm 1. So, those search-directions, which are stored in $\mathbf{C}$, are the first N_C search-directions generated in the first Newton-Raphson- and GMRES-iterations and reused in all subsequent Newton-Raphson-iterations. From TRKS for the PCPG-algorithm it is known, that search-directions corresponding to high convergence-inhibiting eigenmodes are usually generated in the first iterations, which creates a suitable coarse-space [12]. A similar behavior is expected for the GMRES algorithm.

Algorithm 1: Two-level GMRES

$\Delta\vec{\lambda}_0 = \vec{0}$, $\Delta\hat{\lambda}_0 = \vec{0}$
$\Delta\vec{\lambda}_C = \mathbf{C}\left(\mathbf{C}^T\mathbf{F}^T\mathbf{FC}\right)^{-1}\mathbf{C}^T\mathbf{F}^T\left(\vec{d} - \mathbf{F}\Delta\vec{\lambda}_0\right)$
$\vec{r}_0 = \vec{d} - \mathbf{F}\Delta\vec{\lambda}_C$
$\vec{w}_0 = \mathbf{P}_C\vec{r}_0$
$\beta = \|\vec{z}_0\|$
$\mathbf{V}_0 = \vec{w}_0/\|\vec{w}_0\|$
while $\|\vec{r}_k\| > \varepsilon_{F,abs}$ ***and*** $\|\vec{r}_k\|/\|\vec{r}_0\| > \varepsilon_{F,rel}$ ***and*** $k \leftarrow 0$ **to** k_{end} **do**
  $\vec{q}_k = \mathbf{F}\mathbf{V}_k$
  $\vec{w}_k = \mathbf{P}_C\vec{q}_k$
  for $l \leftarrow 0$ **to** k **do**
    $\mathbf{H}_{l,k} = \vec{w}_k^T\mathbf{V}_l$
    $\vec{w}_k = \vec{w}_k - \mathbf{H}_{l,k}\mathbf{V}_l$
  $\mathbf{H}_{k+1,k} = \|\vec{w}_k\|$, $\vec{e}_1 = \begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}$
  $\vec{u}_k = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T(\beta\vec{e}_1)^T$
  $\vec{v}_k = \tilde{\mathbf{P}}_C\mathbf{V}\vec{u}_k$
  $\vec{r}_k = \vec{r}_0 - \mathbf{F}\vec{v}_k$
  $\mathbf{V}_{k+1} = \vec{w}_k/\|\vec{w}_k\|$
  $k \leftarrow k + 1$
$\Delta\vec{\lambda} = \Delta\vec{\lambda}_0 + \Delta\vec{\lambda}_C + \vec{v}_k$
$\mathbf{C} = \begin{bmatrix}\mathbf{C} & \mathbf{V}\end{bmatrix}$



(a) 10-substructure plate with the upper middle substructure exhibiting a micro-time-stepsize with time-step-ratio 10. All surrounding substructures exhibit $\Delta t = 0.0002s$.

(b) Applied load over time.

Figure 2: 2D benchmark example with multirate time-integration.

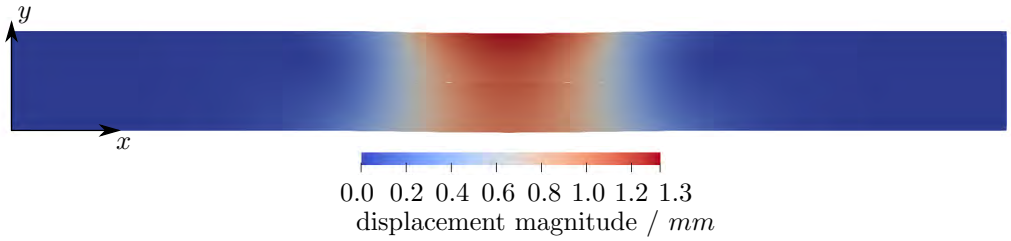


Figure 3: Displacements of converged solution at time $0.001s$

3 Numerical Experiments

Here, we provide our numerical results. Section 3.1 describes the setup of the presented bending-plate example. In Section 3.2, we investigate the solver's convergence behavior and the captured eigenmodes and in Section 3.3 the influence of the micro-time-scale.

3.1 Model setup

Throughout the following experiments, we used a 2D plate under impact-load as benchmark-example, which is depicted in Fig. 2. This example is composed of 2D rectangular substructures with Quad-4 elements and a geometrically nonlinear St. Venant-Kirchhoff material representing an Aluminium beam (Young's modulus $E = 70 \cdot 10^3 N/mm^2$, Poisson's ratio $\nu = 0.34$, density $\rho = 2.7 \cdot 10^{-6} kg/mm^3$, thickness $h = 5.0mm$). The external load is applied as a ramped up impact-like pressure applied on the middle substructure's top edge, as shown in Fig. 2b with $f_{max} = 5.0 \cdot 10^3 N/mm$, $t_{max} = 0.001s$. This model is created in our in-house Open-Source Python-Fortran FE-code AMfe [1] and solved with our Python FETI-library AMfeti [2]. The solvers were set up with absolute tolerances $\varepsilon_{N,abs} = 1.0 \cdot 10^{-6}$ and $\varepsilon_{F,abs} = 1.0 \cdot 10^{-7}$ and relative tolerances $\varepsilon_{N,rel} = 1.0 \cdot 10^{-10}$ and $\varepsilon_{F,rel} = 1.0 \cdot 10^{-10}$, such that the Newton-solver is considered converged if either $max(\|\vec{r}_i^{(s)}\|) < \varepsilon_{N,abs}$ or $max(\|\vec{r}_i^{(s)}\|)/max(\|\vec{r}_0^{(s)}\|) < \varepsilon_{N,rel}$ and the GMRES is converged if $\|\vec{r}_k\| < \varepsilon_{F,abs}$ or $\|\vec{r}_k\|/\|\vec{r}_0\| < \varepsilon_{F,rel}$. The resulting displacements of the solution at time $0.001s$ are shown in Figure 3. There are some small incompatibilities visible on the interfaces between the micro- and macro-substructures, resulting from intermediate oscillations in the velocities [18, 15].

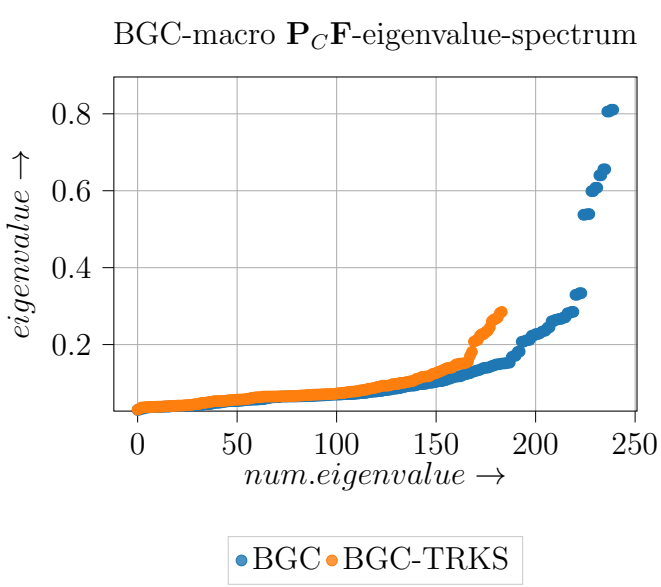
3.2 Convergence behavior and capturing eigenmodes

In singlerate dynamics, the PCPG solver's convergence behavior is bounded by the condition-number of the projected preconditioned interface-operator [12, 9]. A GMRES-solver's conver-

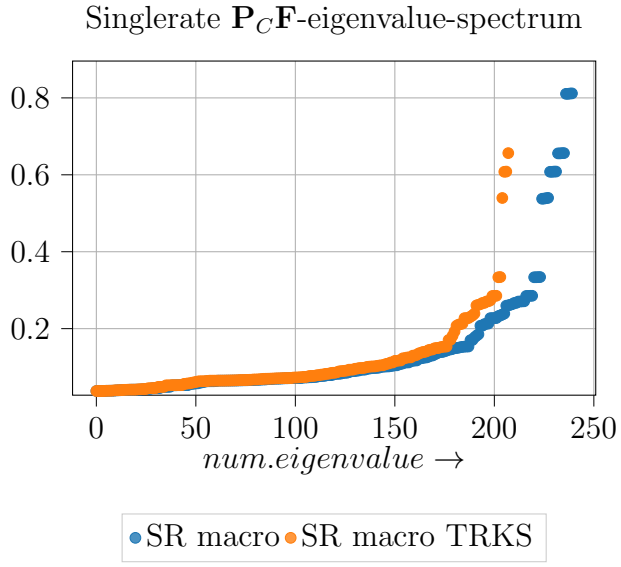
rank($\mathbf{F}$)	$\mathbf{F}$ size	$\mathbf{F}$ symmetry	$\|\mathbf{F}^T\mathbf{F} - \mathbf{F}\mathbf{F}^T\|$	condition number $\mathbf{A}^{(s)}$	$\mathbf{P}_C\mathbf{F}$ symmetry
240	264 x 264	$1.19 \cdot 10^{-3}$	$4.59 \cdot 10^{-4}$	$2.56 \cdot 10^{19}$	$6.83 \cdot 10^{-2}$

$\|\mathbf{F}^T\mathbf{P}_C^T\mathbf{P}_C\mathbf{F} - \mathbf{P}_C\mathbf{F}\mathbf{F}^T\mathbf{P}_C^T\|$	condition number coarse problem
$8,51 \cdot 10^{-3}$	$7.33 \cdot 10^2$

Table 1: System's average characteristic numbers for BGC-macro case with TRKS-projection. Matrix symmetries are checked with $\|\mathbf{A}^T - \mathbf{A}\|$ for some square matrix $\mathbf{A}$.



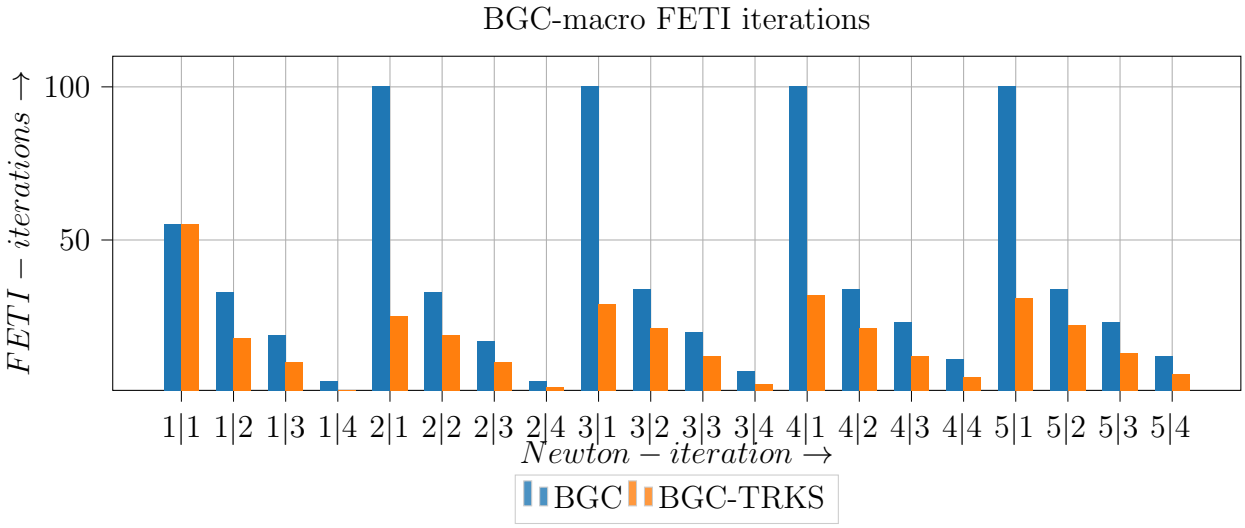
(a) Multirate case solved with GMRES



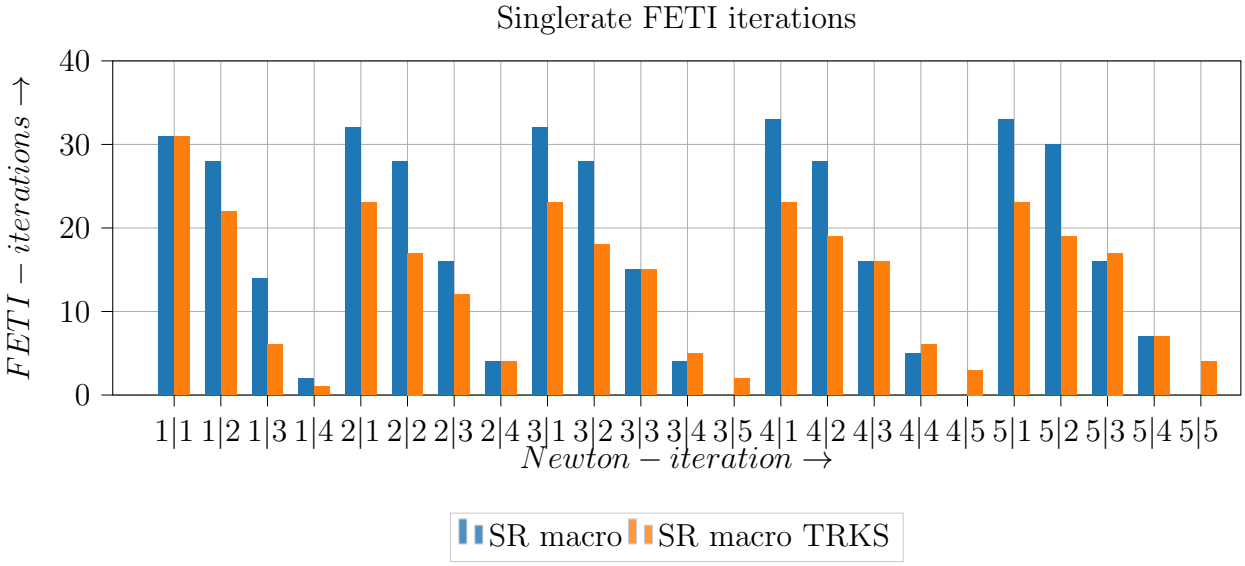
(b) Singlerate case solved with GMRES for comparison

Figure 4: Eigenvalue-spectrum of the eigenvector-matrix of $\mathbf{P}_C\mathbf{F}$ for multirate and singlerate cases in macro-timestep 1 and Newton-iteration 2.

gence behavior is only determined by this condition-number in case of normal matrices and not necessarily for nonnormal matrices, as pointed out by Greenbaum e.a. [11]. As shown in Table 1, the $\mathbf{F}$-operator is indeed a non-normal matrix, which is checked by evaluating $\|\mathbf{F}^T\mathbf{F} - \mathbf{F}\mathbf{F}^T\|$: if this norm is close to 0, $\mathbf{F}$ is considered normal. This also applies for the deflated case with $\|\mathbf{F}^T\mathbf{P}_C^T\mathbf{P}_C\mathbf{F} - \mathbf{P}_C\mathbf{F}\mathbf{F}^T\mathbf{P}_C^T\|$. However, one can still formulate an upper bound for the residuals by the condition-number of the eigenvector-matrix of $\mathbf{P}_C\mathbf{F}$, as proposed by Gaul [8]. In Figure 4, the eigenvalue-spectra for the BGC-macro case and for the single-rate case (macro-time-step in all substructures) are shown. Note that zero-eigenvalues have been removed in these plots and the eigenvalues are sorted in ascending order. Two aspects arise from these spectra: the eigenspectrum is very similar for both, the BGC-macro and the single-rate case. This implies, that the convergence-behavior is not as much governed by the micro-time-step, but by the macro-time-step. And the other aspect concerns differences in the captured eigenmodes associated to the removed eigenvalues by TRKS. In both cases, the coarse-space-size is limited to 50 and while in the single-rate case some high eigenvalues are kept, they are removed in the BGC-macro case. Krylov solvers capture high eigenmodes first, which lets the TRKS gather these high convergence inhibiting modes early [12]. Hence, a reason for this different behavior might be the initial number of FETI-iterations, as depicted in Figure 5. While in the single-rate case the GMRES-solver requires only 31 iterations in the first timestep and Newton-iteration, 55 are required to solve the interface-problem in the BGC-macro case. This difference with respect



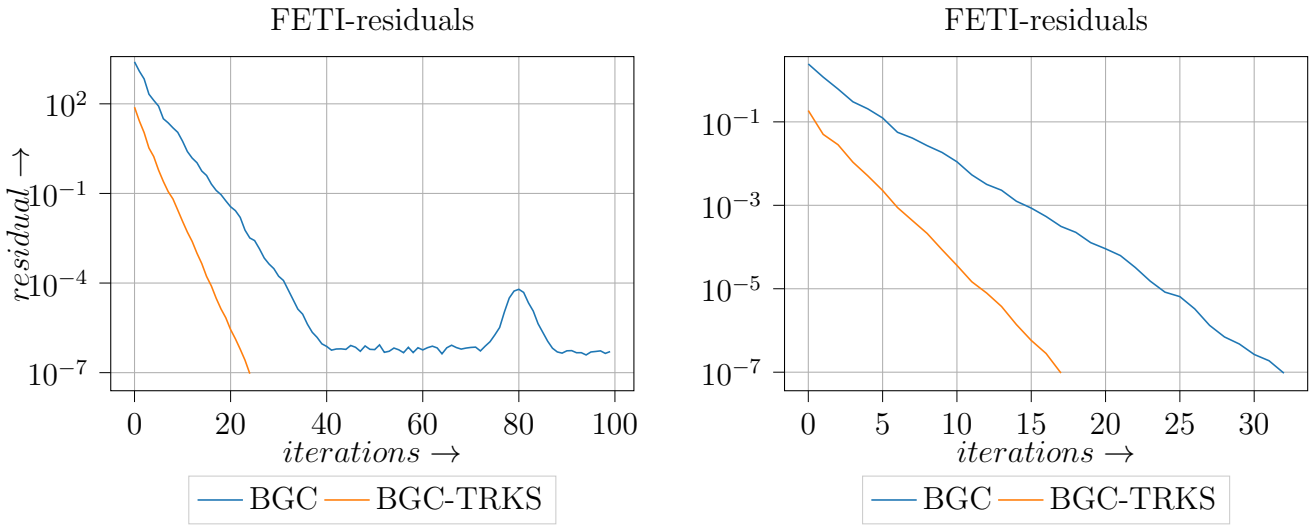
(a) Multirate case solved with GMRES



(b) Singlerate case solved with GMRES for comparison

Figure 5: FETI iterations required to reach convergence for consecutive macro-timesteps and Newton-iterations (see labels *time-step*|*Newton-iteration*).

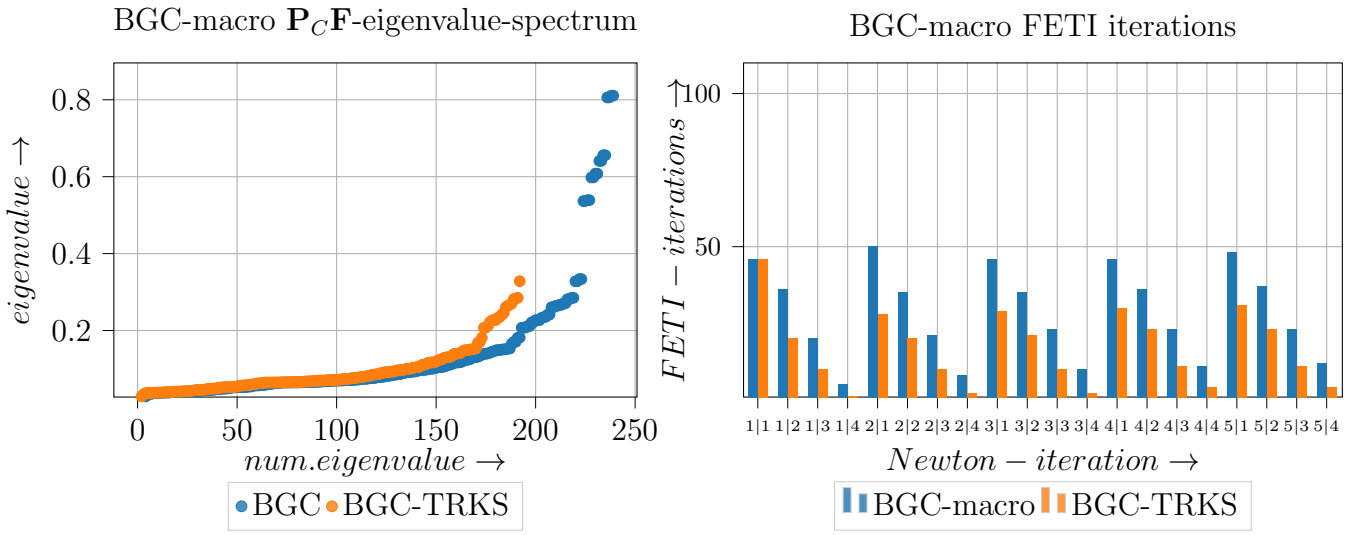
to the similar eigenvalue-spectra again emphasizes the fact that the condition number alone does not define the convergence behavior, but it provides a good estimate and describes the behavior of deflation well. That also means that only 31 search-directions are available for the coarse-space and, besides the large eigenvalues, smaller ones are captured in the singlerate-case earlier as well and therefore the coarse-space is enriched with less effective modes. During the subsequent Newton-iterations and time-steps the coarse-space is further filled up. In the BGC-macro case, the coarse-space is completely filled up in the first Newton-iteration. This also improves the relative reduction in the first Newton-iterations compared to the singlerate case. We have to point out, that the GMRES-solver didn't reach convergence in some nondeflated BGC-macro cases, though. However, that is likely a numerical issue related to the bad local conditioning, as the residuals stagnated at a low level, as depicted in Figure 6. Here, deflation



(a) Timestep 2, Newton-iteration 1

(b) Timestep 1, Newton-iteration 2

Figure 6: FETI residuals during FETI-iterations of BGC-macro case.



(a) Multirate case with micro-time-step-size 0.00001s at timestep 1 and Newton-iteration 2.

(b) FETI iterations required to reach convergence for multirate case with micro-time-step-size 0.00001s.

Figure 7: Eigenvalue-spectrum of the eigenvector-matrix of $\mathbf{P}_C\mathbf{F}$ and required iterations for multirate case with finer micro-time-scale.

also improved this stability.

3.3 Influence of the micro-time-scale

Finally, we further reduce the micro-time-step-size to 0.00001s, resulting in a time-step-ratio of 20. The resulting eigenvalue-spectrum and required FETI-iterations are depicted in Figure 7a and Figure 7b. The eigenvalue spectrum remains very similar to the one with a coarse micro-time-scale in Figure 4a. Hence, the micro-time-scale has limited influence on capturizing

the high eigenvalues. The required iterations are slightly reduced, but show similar convergence behavior as in Figure 5a apart from better stability in the nondeflated case.

4 CONCLUSIONS

The TRKS has been successfully applied to a GMRES-method for our multirate nonlinear BGC-macro FETI-solver. In our examples, the total recycling approach selects the high convergence inhibiting eigenmodes and therefore improves convergence. Of course, this does not affect the FETI-solver in the first Newton-iteration in the first time-step, as the eigenmodes are to be gathered in this step. With this reducing behavior of the recycling technique and the characteristic deflation of the eigenvalue-spectrum, we can say that recycling is also well applicable to a GMRES-solver and the nonlinear BGC-macro method. Moreover, we found, that the choice of local time-step-sizes hardly affects the performance of the global iterative GMRES-solver. The global Newton- and GMRES-solvers' performances are more governed by the synchronisation- or macro-timestep-size. We are currently working on the application of more selective recycling approaches for multirate methods and the application of preconditioning. Our results in this work also imply, that due to the little interface-problem's dependency of the micro time-steps, reusing search-directions from the singlerate case might be beneficial for a time-adaptive approach and will be investigated further in the future.

Acknowledgement: *We thank the DFG for the funding of project RI2451/8-1, in which context this work has been done.*

REFERENCES

[1] AMfe, Applied Mechanics Finite Elements Python-Fortran-library, `https://github.com/AppliedMechanics/AMfe`.

[2] AMfeti, Applied Mechanics Finite Element Tearing and Interconnecting Python-MPI-library, `https://github.com/AppliedMechanics/AMfeti`.

[3] Brun, M and Gravouil, A. and Combescure, A. and Limam, A. Two FETI-based heterogeneous time step coupling methods for Newmark and α-schemes derived from the energy method. *Computer Methods in Applied Mechanics and Engineering*, Vol. **283**, pp. 130–176, (2015), doi: 10.1016/j.cma.2014.09.010.

[4] Farhat, C. and Roux, F.X. A method of finite element tearing and interconnecting and its parallel solution algorithm, *International Journal for Numerical Methods in Engineering*, Vol. **32**, pp. 1205–1227, (1991), doi: 10.1002/nme.1620320604.

[5] Farhat, C. and Crivelli, L. and Géradin, M. On the spectral stability of time integration algorithms for a class of constrained dynamics problems, *AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 34th and AIAA/ASME Adaptive Structures Forum, La Jolla, CA, Apr. 19-22, 1993, Technical Papers. Pt. 1 (A93-33876 13-39)*, pp. 80–97, (1993)

[6] Farhat, C. and Crivelli, L. and Roux, F.X. A transient FETI methodology for large-scale parallel implicit computations in structural mechanics, *International Journal for Numerical Methods in Engineering*, Vol. **37**, pp. 1945–1975, (1994), doi: 10.1002/nme.1620371111.

[7] Farhat, C. and Kendall, P. and Lesoinne, M. The second generation FETI methods and their application to the parallel solution of large-scale linear and geometrically non-linear

structural analysis problems. *Computer methods in applied mechanics and engineering*, Vol. **184**, pp. 333–374, (2000), doi: 10.1016/S0045-7825(99)00234-0

[8] Gaul, A. Recycling Krylov subspace methods for sequences of linear systems - Analysis and applications. *Technische Universität Berlin*, Dissertation, (2014), `https://depositonce.tu-berlin.de/bitstream/11303/4444/1/gaul_andre.pdf`.

[9] Gosselet, P. and Rey, C. and Pebrel, J. Total and selective reuse of Krylov subspaces for the resolution of sequences of nonlinear structural problems, *International Journal for Numerical Methods in Engineering*, Vol. **94**, pp. 60–83, (2013), doi: 10.1002/nme.4441.

[10] Gravouil, A. and Combescure, A. Multi-time-step explicit - Implicit method for non-linear structural dynamics, *International Journal for Numerical Methods in Engineering*, Vol. **50**, pp. 199–225, (2001), doi: 10.1002/1097-0207(20010110)50:1<199::AID-NME132>3.0.CO;2-A.

[11] Greenbaum, A. and Pták, V. and Strakoš, Z. Any Nonincreasing Convergence Curve is Possible for GMRES, *Society for Industrial and Applied Mathematics*, Vol. **17**, pp. 465–469, (1996)

[12] Leistner, M. C. and Gosselet, P. and Rixen, D. J. Recycling of solution spaces in multi-preconditioned FETI methods applied to structural dynamics, *International Journal for Numerical Methods in Engineering*, Vol. **116**, pp. 141–160, (2018), doi: 10.1002/nme.5918.

[13] Leyendecker, S. and Ober-Blöbaum, S. A variational approach to multirate integration for constrained systems, *Multibody Dynamics 2011, ECCOMAS Thematic Conference, Brussels, Belgium, 4-7 July 2011*, pp. 1–15, (2011).

[14] Newmark, N.M. Method of Computation for Structural Dynamics, *Journal of the Engineering Mechanics Division*, Vol. **85**, pp. 67–94, (1959).

[15] Prakash, A. and Taciroglu, E. and Hjelmstad, K. D. Computationally efficient multi-time-step method for partitioned time integration of highly nonlinear structural dynamics. *Computers and Structures*, Vol. **133**, pp. 51–63, (2014), doi: 10.1016/j.compstruc.2013.11.013.

[16] Saad, Y. and Schultz, M. H. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, Vol. **7**, pp. 856–869, (1986), doi: 10.1137/0907058

[17] Seibold, A. S. and Leistner, M. C. and Rixen, D. J. Localization of nonlinearities and recycling in dual domain decomposition. *Domain Decomposition Methods in Science and Engineering XXV*, pp. 474–482, (2018), doi: 10.1007/978-3-030-56750-7, isbn: 978-3-030-56749-1.

[18] Seibold, A. S. and Rixen, D. J. A variational approach to asynchronous time-integration of structural dynamics problems in the context of FETI and spurious oscillations on the interfaces. *EURODYN 2020, XI International Conference on Structural Dynamics*, pp. 26–43, (2020), doi: 10.47964/1120.9003.19111.

[19] Seibold, A. S. and Rixen, D. J. Preconditioning of a FETI-solver for a nonlinear asynchronous time-integrator applied to structural dynamics. *PAMM*, pp. 1–2, (2021), doi: 10.1002/pamm.202000213.

[20] Yeung, M.C. and Tang, J.M. and Vuik, C. On the Convergence of GMRES with Invariant-Subspace Deflation. *Reports of the Delft Institute of Applied Mathematics*, pp. 1–35, (2010), issn: 1389-6520.

Multilevel matrix-free preconditioner to solve linear systems associated with a time-dependent SP_N equations

A. Carreño*, A. Vidal-Ferràndiz†, D. Ginestar† and G. Verdú*

* Instituto Universitario de Seguridad Industrial, Radiofísica y Medioambiental (ISIRYM)
Universitat Politècnica de València
Valencia, Spain
e-mail: amcarsan@iqn.upv.es, gverdu@iqn.upv.es

† Instituto Universitario de Matemática Multidisciplinar (IMM)
Universitat Politècnica de València
Valencia, Spain
e-mail: anvifer2@imm.upv.es, dginesta@mat.upv.es

Key words: Simplified spherical harmonics equations, Finite element method, Multilevel preconditioner, Matrix-free implementation

Abstract: *The evolution of the neutronic power inside of a nuclear reactor core can be approximated by means of the diffusive time-dependent simplified spherical harmonics equations (SP_N). For the spatial discretization of these equations, a continuous Galerkin high order finite element method is applied to obtain a semi-discrete system of equations that is usually stiff. A semi-implicit time scheme is used for the time discretization and many linear systems are needed to be solved and previously, preconditioned. The aim of this work is to speed up the convergence of the linear systems solver with a multilevel preconditioner that uses different degrees of the polynomials used in the finite element method. Furthermore, as the matrices that appear in this type of system are very large and sparse, a matrix-free implementation of the preconditioner is developed to avoid the full assembly of the matrices. A benchmark transient tests this methodology. Numerical results show, in comparison with the block Gauss-Seidel preconditioner, an improvement in terms of number of iterations and the necessity of computational resources.*

1 INTRODUCTION

Inside the reactor core, the evolution of the neutronic power can be modelled by means of the multigroup simplified spherical harmonics equations (SP_N). Different time formulations for this approximation of the neutron transport equation can be developed [15]. This work uses a formulation where the partial derivative of the even moments of the flux are neglected such that it can be seen as a generalized diffusive system with derivatives of order two in the space.

Spatially, the problem is discretized by applying a continuous Galerkin high order finite element method. Two sets of time-dependent differential equations are obtained, that for usual reactor systems are stiff. One related to the neutron moments and other related to the delayed neutron precursor concentrations. Therefore, implicit time schemes must be used [7] that require to solve large linear systems at each time-step. Krylov solvers such as the Generalized Minimal Residual (GMRES) [13] have been shown to be very efficient to solve such large sparse linear systems, if they are applied with a reasonable preconditioner.

Different preconditioners can be applied to these linear systems. First, one can apply classical preconditioners based on an incomplete matrix factorization, such as the ILU decomposition or the ICC decomposition. The linear systems associated with the SP_N equations have a block structure that also permits to apply block preconditioners such as the block Jacobi or the block Gauss-Seidel preconditioner [13]. Although these types of preconditioners are efficient, its implementation implies to store the matrices, or one part of them, in memory, which is very demanding.

https://doi.org/10.4995/YIC2021.2021.12510

Recently, multilevel methods are successfully applied to a wide range of problems. The levels are obtained either from different finite element discretizations on the original grid [5], from a hierarchy of coarser meshes [14] or from several levels of energy groups [6]. In this type of problems, the multigrid preconditioner can be applied, but the initial spatial meshes used in the computation are taken as coarse meshes and homogenized cross-sections must be redefined on each cell. That leads to an expensive application of the preconditioner [4]. Using a preconditioner based on several levels of energy groups is good option to integrate problems with a high number of energy groups. In this work, a two-level preconditioner based on different degrees of the polynomials used in the finite element discretization is applied.

On the other hand, the spatial discretization of a realistic nuclear reactor system with a high-order FEM produces huge algebraic matrices that require high demands of computational memory. Thus, a matrix-free technique can be used where the matrices are not allocated in memory and matrix-vector products are computed on the fly by using a cell-based interface. This technique does not only reduce the computational memory, but also it can reduce the matrix-vector multiplication runtimes in some computer architectures [10]. The main inconvenience of this technique is that algorithms to solve linear systems only can use matrix-vector products, since it is very difficult to access to particular elements of the matrices. In this work, a matrix-free implementation of the multilevel preconditioner is provided.

The rest of the paper is organized as follows. Section 2 presents the simplified spherical harmonics equations. Section 3 briefly exposes the finite element method used for the spatial discretization and the backward method used for the time-discretization. Section 4 explains the multilevel preconditioner. Section 5 describes some details about the implementation, in particular, about the matrix-free technique. Section 6 contains the numerical results obtained to test the proposed methodology. Finally, Section 7 collects the main conclusions of this work.

2 SIMPLIFIED P_N EQUATIONS

The diffusive time-dependent simplified harmonics (SP_N) equations can be written as [15]

$$\begin{aligned}\boldsymbol{\mathcal{V}}\frac{\partial}{\partial t}\phi^n - \vec{\nabla}\cdot\left(\frac{n\,(\boldsymbol{\mathcal{S}}^{n-1})^{-1}}{(2n+1)(2n-1)}\vec{\nabla}\left((n-1)\phi^{n-2}+n\phi^n\right) + \frac{(n+1)(\boldsymbol{\mathcal{S}}^{n+1})^{-1}}{(2n+1)(2n+3)}\vec{\nabla}\left((n+1)\phi^n\right.\right.\\ \left.\left.+\,(n+2)\phi^{n+2}\right)\right) + \boldsymbol{\mathcal{S}}^n\phi^n = \delta_{n0}\boldsymbol{\mathcal{F}}\phi^n + \delta_{n0}\sum_{k=1}^{K}\boldsymbol{\mathcal{M}}_k\boldsymbol{\mathcal{C}}_k\,, \qquad n=0,2,\ldots,N-1\,,\end{aligned} \tag{1}$$

and the equations for delayed neutron precursor concentration are

$$\frac{\partial}{\partial t}\boldsymbol{\mathcal{C}}_k = -\lambda_k^d\boldsymbol{\mathcal{C}}_k + \boldsymbol{\mathcal{R}}_k\phi^0\,, \qquad k=1,\ldots,K, \tag{2}$$

where

$$\boldsymbol{\mathcal{S}}^n = \begin{pmatrix}\Sigma_{t1}-\Sigma_{s11}^n & \cdots & -\Sigma_{sG1}^n\\ \vdots & \ddots & \vdots\\ -\Sigma_{s1G}^n & \cdots & \Sigma_{tG}-\Sigma_{sGG}^n\end{pmatrix}, \quad \boldsymbol{\mathcal{F}} = \begin{pmatrix}\chi_1^p(1-\beta^1)\nu_1\Sigma_{f1} & \cdots & \chi_1^p(1-\beta^G)\nu_G\Sigma_{fG}\\ \vdots & \ddots & \vdots\\ \chi_G^p(1-\beta^1)\nu_1\Sigma_{f1} & \cdots & \chi_G^p(1-\beta^G)\nu_G\Sigma_{fG}\end{pmatrix}, \tag{3}$$

$$\boldsymbol{\mathcal{V}} = \begin{pmatrix}1/v_1 & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & 1/v_G\end{pmatrix}, \quad \boldsymbol{\mathcal{M}}_k = \begin{pmatrix}\lambda_k^d\chi_1^{d,k}\\ \vdots\\ \lambda_k^d\chi_G^{d,k}\end{pmatrix}, \quad \boldsymbol{\mathcal{R}}_k = \begin{pmatrix}\beta_k^1\nu_1\Sigma_{f1}\\ \vdots\\ \beta_k^G\nu_G\Sigma_{fG}\end{pmatrix}^T, \quad \phi^n = \begin{pmatrix}\phi_1^n\\ \vdots\\ \phi_G^n\end{pmatrix}.$$

The variable $\phi_g^n = \phi_g^n(x,t)$ denotes the nth-moment of the neutron flux for the energy group g ($g = 1, \ldots, G$). $\boldsymbol{\mathcal{C}}_k$ denotes the delayed neutron precursor concentration of group k ($k = 1, \ldots, K$). The value of Σ_t is the total cross-sections that is approximated by the transport cross-section Σ_{tr}. Σ_f is the fission cross-section. The value of Σ_s^n is the nth-component of the scattering cross section in the spherical harmonics expansion. In this work, $\Sigma_s^n = 0, \forall n > 0$ because it is assumed that the scattering is isotropic. ν denotes the mean number of neutrons produced by fission. v_g are the neutron velocities. The spectrum of the prompt and the delayed neutrons are denoted by χ_g^p and $\chi_g^{d,k}$. The fraction of the delayed neutrons is β_k^g such that $\beta^g = \sum_{k=1}^K \beta_k^g$. The neutron precursor delayed constants are λ_k^d.

The linear change of variables proposed in [9] is applied to the Equation (1). In the case of the SP_3 equations, the change of variables is

$$U^1 = \phi^0 + 2\phi^2\,, \quad U^2 = 3\phi^2\,, \tag{4}$$

to obtain a system of the form

$$\mathbb{V}\frac{\partial}{\partial t}U - \vec{\nabla}\cdot\left(\mathbb{D}\vec{\nabla}U\right) + \mathbb{S}U = \mathbb{F}U + \mathbb{C}\,, \qquad U = \left(U^1, U^2\right)^T\,, \tag{5}$$

where

$$\mathbb{D} = \begin{pmatrix} \frac{1}{3}(\boldsymbol{\mathcal{S}}^1)^{-1} & 0 \\ 0 & \frac{1}{7}(\boldsymbol{\mathcal{S}}^3)^{-1} \end{pmatrix}, \quad \mathbb{S}_{ij} = \sum_{m=1}^{2} \mathbf{c}_{ij}^{(\mathbf{m})}\boldsymbol{\mathcal{S}}^m, \tag{6}$$

$$\mathbb{V}_{ij} = \sum_{m=1}^{2} \mathbf{c}_{ij}^{(\mathbf{m})}\boldsymbol{\mathcal{V}}, \quad \mathbb{F}_{ij} = \mathbf{c}_{ij}^{(\mathbf{1})}\boldsymbol{\mathcal{F}}, \quad \mathbb{C}_i = \mathbf{d}_i \sum_{k=1}^{K} \boldsymbol{\mathcal{M}}_k\boldsymbol{\mathcal{C}}_k, \tag{7}$$

and the coefficients matrix, $\mathbf{c}^{(\mathbf{m})}$ and vector $\mathbf{d}$ are

$$\mathbf{c}^{(\mathbf{1})} = \begin{pmatrix} 1 & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{9} \end{pmatrix}, \quad \mathbf{c}^{(\mathbf{2})} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{5}{9} \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} 1 \\ -\frac{2}{3} \end{pmatrix}. \tag{8}$$

3 SPATIAL AND TIME DISCRETIZATIONS

A high-order continuous Galerkin Finite Element Method (FEM) for the spatial discretization of the problem (5) is used. The discretization yields an semi-discrete time-dependent problem of the form

$$\mathbf{V}\frac{d}{dt}\tilde{U} + \mathbf{T}\tilde{U} = \mathbf{F}\tilde{U} + \mathbf{d}\sum_{k=1}^{K}\mathbf{M}_k C_k\,, \tag{9}$$

$$\frac{d}{dt}\mathbf{P}\mathbf{C}_k = -\mathbf{L}_k\mathbf{C}_k + \mathbf{R}_k\tilde{\phi}^0, \tag{10}$$

where $\mathbf{V}$, $\mathbf{T}$, $\mathbf{F}$, $\mathbf{M}_k$, $\mathbf{L}_k$ and $\mathbf{R}_k$ are the discretized operator of the $\mathbb{V}$, $-\vec{\nabla}\cdot\mathbb{D}\vec{\nabla} + \mathbb{S}$, $\mathbb{F}$, $\boldsymbol{\mathcal{M}}_k$, $\boldsymbol{\mathcal{L}}_k$ and $\boldsymbol{\mathcal{R}}_k$, respectively. The form of these operators will depend on the formulation. The mass matrix P is not the identity matrix because the basis of the FEM, that is composed of Lagrange polynomials, is not orthonormal. Vectors $\tilde{U}$ and C_k contain the discrete version of the moments U and the delayed neutron precursor concentration $\mathbb{C}_k$. The finite element method is implemented using `deal.II` library [3] and its structures. More details about FEM can be found in [19, 18]. Generally, these matrices are not symmetric, but they have a block structure provided by the different energy groups and neutronic field moments.

Given a configuration of a nuclear reactor, the time-dependent semi-discrete system (9) is generally stiff. Thus, implicit methods are used for its time discretization. In this work, a semi-implicit scheme is used where each type of equation is integrated independently.

The time interval $[0, T]$ is divided into several subintervals $[t_h, t_{h+1}]$ where $\Delta t_h = t_{h+1} - t_h$. The equation for the moments at $t = t_{h+1}$ (Equation (9)) is integrated by applying a backward difference of first order to the time derivative. The other magnitudes are substituted by its value at time t_{h+1}, excluding the concentration of neutron precursors that is substituted by its value at t_h. In this way, the solution of the linear system

$$\left(\frac{1}{\Delta t_h}\mathbf{V}^{h+1} + \mathbf{T}^{h+1} - \mathbf{F}^{h+1}\right) U^{h+1} = \frac{1}{\Delta t_h}\mathbf{V}^h U^h + \mathbf{d}\sum_{k=1}^{K}\mathbf{M}_k^{h+1}\mathbf{C}_k^h\,, \tag{11}$$

gives an approximation of the moments at time t_{h+1}.

This linear system has a size of $(N+1)\times N_{dofs}\times G/2$, where N_{dofs} are the degrees of freedom of the FEM. It is solved with the GMRES method provided by the PETSc library [2]. This work is devoted to study a multilevel preconditioner (described in Section 4) to precondition this system.

The concentration of delayed precursors equation is also integrated by using a one-step implicit scheme. The rest of the magnitudes are substituted by its value at t_{h+1}. Thus, the concentration of precursors can be approximated by solving the linear system

$$\left(\frac{1}{\Delta t_h}\mathbf{P} + \mathbf{L}_k^h\right)\mathbf{C}^{h+1} = \frac{1}{\Delta t_h}\mathbf{P}\mathbf{C}^h + \mathbf{R}_k^h\tilde{\phi}^{0,h+1}\,. \tag{12}$$

The matrices of this system are much smaller than the previous ones with a size equal to N_{dofs}. Thus, it is simply solved with the GMRES method and the ILU(0) preconditioner from the PETSc library [2].

4 MULTILEVEL PRECONDITIONER

The linear systems of Equation (11) need to be preconditioned to integrate the SP$_N$ equations with a reasonable CPU demand. In this work, we study a multilevel preconditioner based on the finite element method. This multilevel preconditioner is based on the classical V-cycle multigrid method [8]. In general, multigrid methods are designed to accelerate the convergence of a simple iterative method (known as smoothing, which reduces the short frequencies error) by a correction obtained when a coarse problem is solved (which is cheaper to solve). To solve this coarse problem, also a smoother and a coarser problem can be used, obtaining in this way a hierarchy of problems, known as levels of the multigrid method [14]. For nuclear reactor computations, using different meshes is not feasible because it does not require very refine meshes, and constructing coarse meshes implies homogenizing the reactor materials at each level. To avoid this problem, smaller problems are defined by considering a degree of the polynomial in the FEM, p^*, smaller than the original value p. The same spatial mesh is considered, but the simplified problem associated has a smaller number of degrees of freedom. This method only makes sense if $p > 1$. For these applications, we use a two-level method with one problem associated with each level because a degree in the FEM equal to 3 is enough to obtain accurate results. However, in other applications where higher degrees in FEM would be necessary, a multilevel preconditioner with more than two levels can be applied following a similar process than the multigrid method.

It is assumed that we want to define a preconditioner to solve the discrete SP$_N$ problem using a degree p in the FEM (first level),

$$\mathbf{A}x = b\,, \tag{13}$$

and we define the smaller problem with degree p^* in the FEM (second level)

$$\mathbf{A}^* x^* = b^* . \tag{14}$$

The two-level preconditioner smooths the iteration error on the original level and correct the iterate by a second level correction in a two-level setting. Recurrent applications on a sequence of levels lead to a multilevel procedure. The multilevel preconditioner can be used without smoothing [12]. In that case, the coarse level solve damps the low-frequency error of the fine problem, but not high-frequency errors. In a more physically sense, the smoother is applied to approximate the finer details.

To apply the multilevel preconditioner, we need to define a restriction operator, $\boldsymbol{\mathcal{R}}$, that interpolates vectors defined on the problem of FEM with degree p into a smaller vector associated with the problem with degree p^*. The values at the nodes of the FEM with degree p (original problem) are interpolated into the nodes associated with FEM with low degree of polynomial. This interpolation is made element by element through a transfer matrix, t^{res}, between the nodes associated with degree p of one element and the nodes associated with degree p^* of such element. The elements of such transfer matrix are given by

$$t^{res}_{ij} = \boldsymbol{\mathcal{N}}^p_j(\eta^*_i), \quad j = 1, \ldots, p+1, \quad i = 1, \ldots, p^*+1, \tag{15}$$

where $\boldsymbol{\mathcal{N}}^p_j$ are the Lagrange shape functions of the expansion which characterize the finite element method of degree p such that $\boldsymbol{\mathcal{N}}^p_j(\eta_k) = \delta_{jk}, \quad k = 1, \ldots, p+1$, being δ_{jk} the Kronecker delta function and η_k the position of the k-th node in the element associated with a degree p. η^*_i denotes the position of the i-th node in the element of the FEM with degree p^*.

In the other way, one can define the prolongation operator, $\boldsymbol{\mathcal{P}}$, also through a interpolation process. However, in this case, the elements of the transfer matrix, t^{pr}, that interpolates the nodes of one element associated with degree p^* into the ones associated with degree p, are given by

$$t^{pr}_{ij} = \boldsymbol{\mathcal{N}}^{p^*}_j(\eta_i), \quad j = 1, \ldots, p^*+1, \quad i = 1, \ldots, p+1, \tag{16}$$

where $\boldsymbol{\mathcal{N}}^{p^*}_i$ are the Lagrange shape functions of the expansion which characterize the finite element method of degree p_* and η_i is the position of the i-th node in the element associated with a degree p. Figure 1 shows a scheme of these operators for a two-dimensional problem.

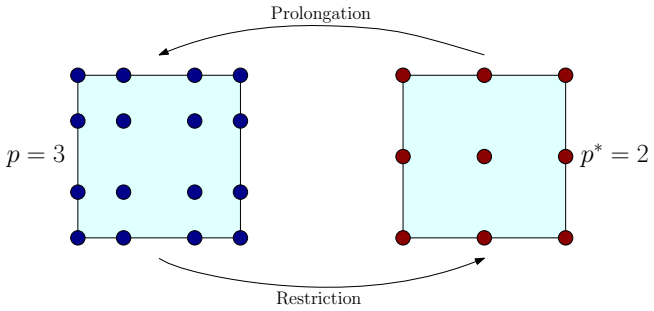


Figure 1: Restriction and prolongation operator of a element associated with degree $p = 3$ and $p^* = 2$ in the FEM.

The smoothing can be done with a Gauss-Seidel iterative method, but it needs to access to the matrix elements [4]. Therefore, Chebyshev polynomial smoothers are recommended for parallel computations and matrix-free implementations [1]. In particular, the implementation of the `deal.II` library is used [3]. The number of iterations for the smoother, its_{ch}, is set to $its_{ch} = 5$. To estimate the largest eigenvalue to apply the Chebyshev polynomial, ten iterations of the GMRES method preconditioned with the inverse diagonal of A are used.

To solve the small problem (14), the GMRES method is applied of the PETSc library [2]. To precondition the coarse problem, two type of strategies are tested. First, the coarse problem is assembled with a semi-matrix-free allocation and the block Gauss-Seidel preconditioner is used to solve this 'small' problem. Second, the matrix of the coarse problem is not assembled and the inverse of the diagonal elements is used to preconditioner the 'small' linear system. Note that to solve this small problem, if $p_* > 1$, we can also use a smaller degree to construct a smaller problem and repeat the process to obtain a three-level preconditioner.

The implementation of the particular case of two-level preconditioner is described in Algorithm 1.

Algorithm 1 Two-level preconditioner

Input: Vector x, matrices $\mathbf{A}$ and $\mathbf{A}^*$.

Output: Vector $y = \mathbf{P}x$ with $\mathbf{P} \approx \mathbf{A}^{-1}$ preconditioner of $\mathbf{A}$.

1: Pre-smooth $\mathbf{A}y = x$ with $its_{ch} = 5$ (Initialize the iterative method with $y_0 = 0$)
2: Restrict the residual $r = \mathbf{A}y - x$ to the second level by $r^* = \mathcal{R}(r)$
3: Solve $\mathbf{A}^* e^* = r^*$
4: Prolongate e^* by $e = \mathcal{P}(e^*)$
5: Correct $y = y + e$
6: Post-smooth $\mathbf{A}y = x$ with $its_{ch} = 5$ (Initialize the iterative method with $y^0 = y$)

The multilevel preconditioner is compared with the block Gauss-Seidel preconditioning (BGS). The number of moments and energy groups of the equations lead to a block structure of the matrices of the linear systems. In the block Gauss-Seidel preconditioner (BGS), each diagonal block is (approximately) solved with the conjugate gradient method and the incomplete Cholesky preconditioner. This preconditioner allows to save only the diagonal blocks and to use a semi-matrix-free implementation of the matrices [17].

5 MATRIX-FREE STRATEGY

A matrix-free strategy for the matrix $\mathbf{A}$ is applied to remove the computational cost of saving the matrices in memory. Nowadays, there are supercomputers without computational memory problems of capacity to solve this type of problem. In practice, they are available for some researching groups. In industrial sectors, such as nuclear engineering, there is a great interest in simulating the behaviour of the reactor without requiring high computational demands. On the other hand, this technique does not only reduce the computational memory, but also it can reduce the matrix-vector multiplication runtimes in some computer architectures [10]. Matrix-vector products are computed on the fly in a cell-based interface. For instance, we can consider a finite element Galerkin approximation, that leads to the matrix A, that takes a vector u as input and computes the integrals of the operator multiplied by trial functions to obtain the output vector is v. The operation can be expressed as a sum of N_c cell-based operations,

$$v = \mathbf{A}u = \sum_{c=1}^{N_c} \mathbf{P}_c^T \mathbf{A}^c \mathbf{P}_c u, \tag{17}$$

where $\mathbf{P}_c$ denotes the matrix that defines the location of cell-related degrees of freedom in the global vector and $\mathbf{A}^c$ denotes the submatrix of $\mathbf{A}$ on cell c. This sum is optimized through *sum-factorization*. Details about the implementation are explained in [10].

This type of implementation does not permit access to the matrix elements, which inabilities to use typical preconditioners such as ILU preconditioner. In this work, two types of allocations are used. First, the full matrix-free implementation (Full-MF) where any element of the

matrices are saved in memory. Second, a semi-matrix-free implementation (Semi-MF) where only the diagonal blocks of the matrices are assembled. The rest are implemented with the matrix-free technique.

6 NUMERICAL RESULTS

This Section tests the performance of the multilevel preconditioner in a transient defined from the movement of two banks of control rods in the tridimensional Langenbuch benchmark reactor [11]. The geometry of the reactor is modelled with 1170 cells. The transient is followed during 30 s.

The methodology has been implemented in C++ based on data structures provided by the libraries `deal.II` [3] and PETSc [2]. It has been incorporated to the open-source neutronic code `FEMFFUSION` [16]. For the computations, a computer with an Intel® Core™ i7-4790 @3.60GHz×8 processor with 32Gb of RAM running on Ubuntu 18.04 has been used.

Numerical results are presented to obtain the solution of the SP_1 and the SP_3 equations. Degree $p = 3$ in the polynomial of the FEM is used for the spatial discretization. Time-step for the backward method is set to $\Delta t_n = 0.1$ s to obtain a sequence of 300 linear systems. The two-level preconditioners are defined from small problems obtained by considering a degree in the FEM of the coarse problem (FEDC) equal to $p^* = 1$ and $p^* = 2$.

Table 1 shows the size of the matrices (Size($\mathbf{A}$)) for the different type of equations and degrees in the finite element method (FED). This Table shows as the full matrix-free implementation largely reduces the CPU memory required by the matrices. This property becomes more relevant as the problem is larger.

Table 1: Size of matrices associated with the linear systems and computational memory for the Full and Semi matrix-free implementations.

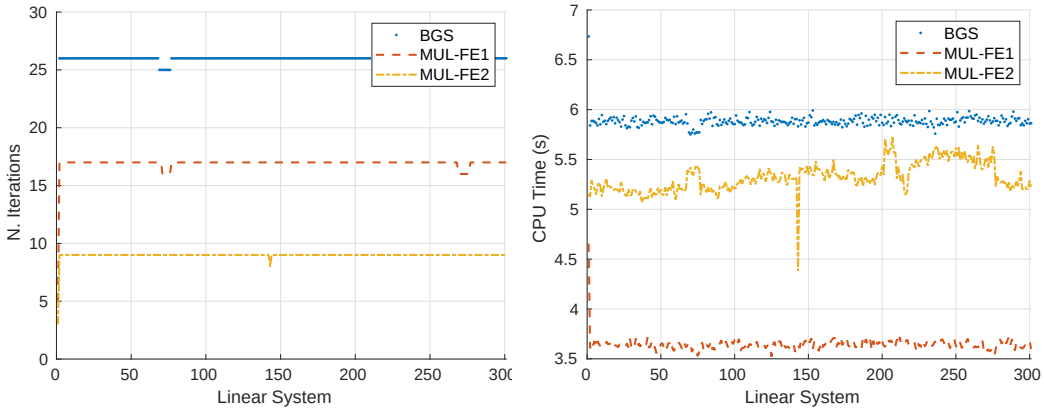
	FED=1			FED=2			FED=3		
Equations	Size (A)	CPU Memory		Size (A)	CPU Memory		Size (A)	CPU Memory	
		Full-MF	Semi-MF		Full-MF	Semi-MF		Full-MF	Semi-MF
SP_1	3080	0.07 MB	1 MB	21 546	0.16 MB	18 MB	69 440	0.34 MB	115 MB
SP_3	6160	0.14 MB	2 MB	43 092	0.23 MB	33 MB	138 880	0.40 MB	214 MB

First, we test the preconditioner used to solve the coarse problem in the multilevel preconditioner. Two types of implementation are compared: the BGS preconditioner with Semi-MF allocation and the inverse diagonal preconditioner with Full-MF allocation. Table 2 shows the mean number of iterations to solve the coarse problems with each preconditioner and the total CPU Time. The results are displayed to compute the sequence associated with the SP_1 equations if FEDC is equal to $p^* = 1$ and $p^* = 2$. Table 2 shows that the application of the BGS preconditioner solves the linear systems with much less iterations than using the inverse of the diagonal as preconditioner. However this preconditioner requires the assembly of the matrices and preconditioner. Numerical results shows that if FEDC is $p^* = 1$, the total CPU time of both implementations is similar. The CPU time of iterations is compensated by the assembling time. However, if FEDC is $p^* = 2$, the CPU time with the inverse diagonal is a bit higher, because in this last case the number of iterations needed by the inverse diagonal is 3 times higher than if BGS preconditioner is applied.

Table 2: Performance of the multilevel preconditioner to compute the sequence associated with the SP_1 equations.

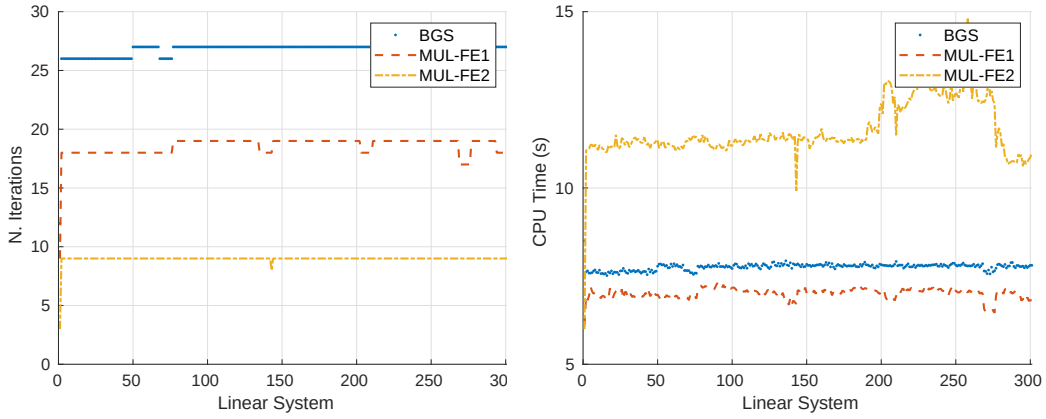
FEDC (p^*)	Type of preconditioner	Mean Number of Iterations	Total CPU time
1	BGS	8.12	1108 s
1	Inverse Diagonal	14.49	1096 s
2	BGS	19.31	1367 s
2	Inverse Diagonal	73.87	1602 s

Now, the multilevel preconditioner is compared with the BGS to solve the sequence of linear systems associated with the SP_1 and SP_3 equations. To apply the BGS and the multilevel preconditioner, the semi-matrix-free implementation and the full matrix-free implementation are used, respectively. The coarse problems in the multilevel preconditioner are solved with a full matrix-free implementation of the matrices and the inverse diagonal to have a full matrix-free implementation for the integration of the equations. Figure 2 shows the number of iterations (left) and the CPU time (right) needed by each type of preconditioner for every system in the sequence of the SP_1 equations. One can observe that the multilevel preconditioner reduces considerably the number of iterations required by the BGS, especially if the coarse problem is defined from a $p^* = 2$. However, in the CPU time, the most efficient preconditioner is the multilevel preconditioner with FEDC equal to $p^* = 1$, because the coarse problems are much smaller than the coarse problems with FEDC equal to $p^* = 2$ (Table 1). Figure 3 displays the results obtained in the sequence of systems associated with the SP_3 equations. Similar conclusions as the ones obtained for the SP_1 equations are obtained, even though the differences between the multilevel and FEDC equal to $p^* = 1$ and the BGS are not as high.



(a) N. iterations (b) CPU Time

Figure 2: Comparison of the multilevel preconditioner where the second level is obtained from $p_* = 1$ (MUL-FE1) and $p_* = 2$ (MUL-FE2), and the BGS preconditioner for the solution of the time-dependent SP_1 equations.



(a) N. iterations (b) CPU Time

Figure 3: Comparison of the multilevel preconditioner where the second level is obtained from $p_* = 1$ (MUL-FE1) and $p_* = 2$ (MUL-FE2), and the BGS preconditioner for the solution of the time-dependent SP_3 equations.

7 CONCLUSIONS

This work has proposed a two-level preconditioner based on different degrees in the polynomials of the finite element method to integrate the SP_N equations.

Numerical results have been shown the competitiveness of this multilevel preconditioner when it is compared with the block Gauss-Seidel preconditioner. Different coarse problems with degree $p^* = 1$ and $p^* = 2$ in the FEM are tested. Coarse problems with degree $p^* = 2$ are more convenient to reduce the number of iterations needed for the linear solver to converge. However, from a CPU time point of view, it is recommended to define the coarse problem from degree equal to $p^* = 1$. On the other hand, this two-level preconditioner can be applied by only using matrix-vector products allowing a full matrix-free implementation that removes the time to assembly the sparse matrices involved in the linear systems and reducing the necessary memory to allocate the matrices.

8 ACKNOWLEDGEMENTS

This work has been partially supported by Spanish Ministerio de Economía y Competitividad under projects ENE2017-89029-P and MTM2017-85669-P. Furthermore, this work has been financed by the Generalitat Valenciana under the project PROMETEO/2018/035.

REFERENCES

[1] M. Adams, M. Brezina, J. Hu, and R. Tuminaro. Parallel multigrid smoothing: polynomial versus Gauss–Seidel. *Journal of Computational Physics*, 188(2):593–610, 2003.

[2] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, et al. PETSc users manual. 2019.

[3] W. Bangerth, T. Heister, and Kanschat G. `deal.II` *Differential Equations Analysis Library*. `http://www.dealii.org`.

[4] A. Carreño, A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú. Block hybrid multilevel method to compute the dominant λ-modes of the neutron diffusion equation. *Annals of Nuclear Energy*, 121:513–524, 2018.

[5] B. Cockburn, O. Dubois, J. Gopalakrishnan, and S. Tan. Multigrid for an HDG method. *IMA Journal of Numerical Analysis*, 34(4):1386–1425, 2014.

[6] L.R. Cornejo, D.Y. Anistratov, and K. Smith. Iteration methods with multigrid in energy for eigenvalue neutron diffusion problems. *Nuclear Science and Engineering*, 2019.

[7] D. Ginestar, G. Verdú, V. Vidal, R. Bru, J. Marín, and J. L. Munoz-Cobo. High order backward discretization of the neutron diffusion equation. *Annals of Nuclear Energy*, 25(1-3):47–64, 1998.

[8] W. Hackbusch. *Multi-grid methods and applications*, volume 4. Springer Science & Business Media, 2013.

[9] S.P. Hamilton and Thomas M. Evans. Efficient solution of the simplified pn equations. *Journal of Computational Physics*, 284:155–170, 2015.

[10] M. Kronbichler and K. Kormann. A generic interface for parallel cell-based finite element operator application. *Computers & Fluids*, 63:135–147, 2012.

[11] S. Langenbuch, W. Maurer, and W. Werner. Coarse-mesh flux-expansion method for the analysis of space-time effects in large light water reactor cores. *Nuclear Science and Engineering*, 63(4):437–456, 1977.

[12] J.A. Roberts and B. Forget. Multigroup diffusion preconditioners for multiplying fixed-source transport problems. *Journal of Computational Physics*, 274:455–472, 2014.

[13] Y. Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[14] R.S. Sampath and G. Biros. A parallel geometric multigrid method for finite elements on octree meshes. *SIAM Journal on Scientific Computing*, 32(3):1361–1392, 2010.

[15] W.M. Stacey. *Nuclear reactor physics*, volume 2. Wiley Online Library, 2007.

[16] A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú. `FEMFFUSION`: A finite element method code for the neutron diffusion equation. `https://www.femffusion.imm.upv.es`, 2020.

[17] A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú. A block arnoldi method for the spn equations. *International Journal of Computer Mathematics*, 97(1-2):341–357, 2020.

[18] A. Vidal-Ferràndiz, R. Fayez, D. Ginestar, and G. Verdú. Solution of the lambda modes problem of a nuclear power reactor using an h–p finite element method. *Annals of Nuclear Energy*, 72:338–349, 2014.

[19] A. Vidal-Ferràndiz, R. Fayez, D. Ginestar, and G. Verdú. Moving meshes to solve the time-dependent neutron diffusion equation in hexagonal geometry. *Journal of computational and applied mathematics*, 291:197–208, 2016.

ISOGEOMETRIC AND NON-STANDARD DISCRETIZATION SCHEMES FOR COMPUTATIONAL STRUCTURAL AND SOLID MECHANICS

The effect of a consistent linearization on the numerical stability of hydrid-elements for quasi-incompressible hyperelastic solids

P. Schneider[1,2,*], J. A. Schönherr[3] and C. Mittelstedt[2]

[1] Deutsches Institut für Kautschuktechnologie e. V. (DIK)
Eupener Straße 33, D-30519 Hannover, Germany
e-mail: patrick.schneider@dikautschuk.de

[2] Institute for Lightweight Construction and Design (KLuB)
Technische Universität Darmstadt
Otto-Berndt-Straße 2, D-64287 Darmstadt, Germany
e-mail: patrick.schneider@klub.tu-darmstadt.de, christian.mittelstedt@klub.tu-darmstadt.de

[3] Center for Structural Materials, State Materials Testing Laboratory (MPA), Institute for Materials Technology (IfW)
Technische Universität Darmstadt
Grafenstraße 2, D-64283 Darmstadt, Germany
e-mail: josef.schoenherr@tu-darmstadt.de

Key words: hybrid finite element, mixed formulation, finite deformation, hyperelasticity, rubber-like material, quasi-incompressible material, discontinuous finite element

Abstract: *We revisit the well-known three-field formulation introduced by Simo and Taylor, [5]. However, while in [5] a semi-discretization is used to eliminate the additional primary unknowns before the problem is linearized in terms of the not yet discretized displacement field, we introduce hybrid/mixed elements based directly on the consistent linearization of the three-field formulation on the continuum-level. In the latter case, static condensation is used to eliminate the additional unknowns on the element-level after the linearization of the continuum formulation in order to derive discontinuous hybrid-elements.*

A family of Simo-Taylor-Pister (`STP`) elements, as well as a family of elements based on the continuum-level linearization (`CL3F`), designed to coincide in terms of the interpolation schemes, the number of assembled degrees of freedom and the number of integration points with the Abaqus hybrid-elements (`C3D8H,C3D20H,C3D10H`) are compared to those elements by benchmark tests. Material parameters were obtained by least-square fitting to experimental data of an industrial NR/IR-blend (natural rubber / isoprene rubber) used for damping applications.

All tested elements are free of volumetric locking. The `STP`-elements show severe stability issues. In general the maximum stable step-width of the Abaqus hybrid-elements is higher in comparison to the `STP`-elements. However, the `CL3F`-elements outperform the Abaqus elements in general without the usage of numerical stabilization. Especially in combination with strongly nonlinear compression models, the advantage of the `CL3F`-elements is huge – here the stable step-width is up to 22 times larger. Details can be found in a contribution which is currently under review, [7].

1 INTRODUCTION

Irreducible (purely displacement-based) finite element formulations are ill-posed for quasi-incompressible materials. In addition, these elements tend to be overly stiff which is an effect known as volumetric locking. To overcome these deficiencies, several so-called mixed and hybrid element formulations were developed. Simo-Taylor-Pister (`STP`) elements were originally introduced in [4, 5] and played a key role in the history of hybrid-element formulations. Usually this formulation is augmented in order to overcome numerical stability issues. Recent hybrid-element formulations often rely heavily on matching numerical stabilization techniques

https://doi.org/10.4995/YIC2021.2021.12174

as well.

2 STP-ELEMENTS

Simo-Taylor-Pister (STP) are based on the modified potential

$$\Pi_{\text{mod}}(\boldsymbol{\varphi}, \Theta, p) \coloneqq \int_\Omega \mathring{W}(\boldsymbol{\varphi}, \Theta)\,\mathrm{d}V + \int_\Omega p\,(J(\boldsymbol{\varphi}) - \Theta)\,\mathrm{d}V + \Pi_{\text{ext}}(\boldsymbol{\varphi}), \tag{1}$$

where the three (primary) unknown tensor fields are: the configuration $\boldsymbol{\varphi}$ (or alternatively the displacement $\boldsymbol{U}$), the dilation Θ and the hydrostatic pressure p. Here $J = \det(\boldsymbol{F})$ denotes the Jacobian and $\mathring{W}$ refers to the modified strain-energy density, where the displacement gradient $\boldsymbol{F}$ is replaced by its modified counterpart $\mathring{\boldsymbol{F}} \coloneqq (\Theta/J)^{1/3}\boldsymbol{F}$. The associated weak form, i.e. Gateaux derivative at the current state $(\hat{\boldsymbol{\varphi}}, \hat{\Theta}, \hat{p})$ in the direction $(\boldsymbol{\eta}, \psi, q)$ (virtual state) is

$$\begin{aligned} G\left(\left(\hat{\boldsymbol{\varphi}}, \hat{\Theta}, \hat{p}\right), (\boldsymbol{\eta}, \psi, q)\right) \coloneqq & \int_\Omega 2\frac{\partial \mathring{W}\left(\hat{\boldsymbol{\varphi}}, \hat{\Theta}\right)}{\partial \mathring{C}_{IJ}} \mathring{\hat{F}}_{iI}\,\mathring{\hat{F}}_{jJ}\,\operatorname{dev}\left(\eta^{\text{sym}}_{i,j}\right)\mathrm{d}V \\ + \int_\Omega \frac{2\psi}{3\hat{\Theta}}\frac{\partial \mathring{W}\left(\hat{\boldsymbol{\varphi}}, \hat{\Theta}\right)}{\partial \mathring{C}_{IJ}} \mathring{\hat{F}}_{iI}\,\mathring{\hat{F}}_{jJ}\,\delta_{ij}\,\mathrm{d}V & + \int_\Omega \hat{p}\,J\left(\hat{\boldsymbol{\varphi}}\right)\eta_{i,i} - \hat{p}\psi + q\left(J\left(\hat{\boldsymbol{\varphi}}\right) - \hat{\Theta}\right)\mathrm{d}V + \delta\Pi_{\text{ext}}(\hat{\boldsymbol{\varphi}}, \boldsymbol{\eta}). \end{aligned} \tag{2}$$

Here, the modified strain-energy density $\mathring{W}$ can still be an arbitrary function of the modified right Cauchy-Green tensor $\mathring{\boldsymbol{C}}$. After the treatment of the weak form, Simo and Taylor assume an additive split of $\mathring{W}$ into an isochoric and volumetric part, cf. [5], section 3.3. The implementation of STP-elements is based on the approach originally sketched in section 4.1 of [5], here referred to as "semi-discretization approach" and recapped below.

The same inter-element discontinuous interpolation for the dilation and the pressure and their virtual counterparts is utilized, using the same shape functions

$${}^{e}\Theta \approx \sum_{l=1}^{{}^{e}n_\Gamma} {}^{e,l}\Gamma(\boldsymbol{\xi})\,\widehat{{}^{e,l}\Theta}, \qquad {}^{e}p \approx \sum_{l=1}^{{}^{e}n_\Gamma} {}^{e,l}\Gamma(\boldsymbol{\xi})\,\widehat{{}^{e,l}p}. \tag{3}$$

Regarding just the additive term of the weak form (2) that incorporates the virtual pressure q and only the contribution of a single finite element ${}^{e}\Omega$, insertion of the interpolations (3) and factoring out the coefficients of the virtual pressure $\widehat{{}^{e,k}q}$ leads to a matrix-vector equation that has to equate to zero. By inversion of the matrix this equation can be solved for the nodal values of the dilation $\widehat{{}^{e,l}\Theta}$, leading finally to an intra-element interpolation for the dilation

$${}^{e}\hat{\Theta} \approx \sum_{k=1}^{{}^{e}n_\Gamma}\sum_{l=1}^{{}^{e}n_\Gamma} {}^{e}H^{-1}_{lk}\int_{{}^{e}\Omega} {}^{e,k}\Gamma(\boldsymbol{\xi})\,{}^{e}J\left(\hat{\boldsymbol{\varphi}}\right)\mathrm{d}V \quad \text{with} \quad {}^{e}H_{kl} \coloneqq \int_{{}^{e}\Omega} {}^{e,k}\Gamma(\boldsymbol{\xi})\,{}^{e,l}\Gamma(\boldsymbol{\xi})\,\mathrm{d}V, \tag{4}$$

that only depends on the current configuration $\hat{\boldsymbol{\varphi}}$. Hence, (4) can be used to eliminate the primary unknown Θ on element level. Applying the same procedure on the additive part of (2) incorporating the virtual dilation ψ, we also obtain an intra-element interpolation for the pressure

$${}^{e}\hat{p} \approx \sum_{k=1}^{{}^{e}n_\Gamma}\sum_{l=1}^{{}^{e}n_\Gamma} {}^{e}H^{-1}_{lk}\int_{{}^{e}\Omega} {}^{e,k}\Gamma(\boldsymbol{\xi})\,\frac{2}{3\hat{\Theta}}\frac{\partial \mathring{W}\left({}^{e}\hat{\boldsymbol{\varphi}}, {}^{e}\hat{\Theta}\right)}{\partial \mathring{C}_{IJ}}\,{}^{e}\mathring{\hat{F}}_{iI}\,{}^{e}\mathring{\hat{F}}_{jJ}\,\delta_{ij}\,\mathrm{d}V, \tag{5}$$

that only depends on the current configuration $\hat{\boldsymbol{\varphi}}$ by the use of (4). The elimination of the independent pressure and the independent dilation in the remainder of (2) by insertion of the intra-element interpolations (4) and (5) leads to a variant of the weak form depending solely on the current configuration $\hat{\boldsymbol{\varphi}}$

$$ {}^{e}g\left(\hat{\boldsymbol{\varphi}},\boldsymbol{\eta}\right) \approx \int_{\boldsymbol{\varphi}({}^{e}\Omega)} \left\{ {}^{e}\hat{\sigma}^{\mathrm{iso}}_{ij} + {}^{e}\hat{p}\,\delta_{ij} \right\} \eta^{\mathrm{sym}}_{i,j}\,\mathrm{d}v + \varphi_{*}\left\{\delta\Pi_{\mathrm{ext}}(\hat{\boldsymbol{\varphi}},\boldsymbol{\eta})\right\}. \tag{6} $$

(Here it is crucial to note that ${}^{e}\hat{p}$ is not the independent pressure anymore, but instead a shorthand notation for the insertion of (5), which also applies for ${}^{e}\hat{\Theta}$ below.) Since (6) only depends on the current configuration $\hat{\boldsymbol{\varphi}}$, i.e. not on the independent dilation and pressure anymore, it can be linearized by derivation of the Gateaux derivative in the direction of the displacement increment $\boldsymbol{u}$ only. The final spatial representation of the linearization reads

$$ \begin{aligned} \delta^{e}g\left(\left(\hat{\boldsymbol{\varphi}},\boldsymbol{\eta}\right),\boldsymbol{u}\right) = \int_{\boldsymbol{\varphi}({}^{e}\Omega)} & u_{i,j}\,{}^{e}\hat{\sigma}_{ij}\eta_{i,l} \\ & + u_{i,j}\eta_{k,l}\left[{}^{e}\hat{p}\left(\delta_{ij}\delta_{kl}-\delta_{il}\delta_{jk}-\delta_{ik}\delta_{jl}\right)+{}^{e}\hat{c}^{\mathrm{iso}}_{ijkl}\right] \\ & + \frac{\partial^{2}\mathring{W}\left({}^{e}\hat{\boldsymbol{\varphi}},{}^{e}\hat{\Theta}\right)}{\partial\Theta^{2}}\,\frac{{}^{e}\hat{\Theta}^{2}}{{}^{e}\hat{J}}\,\overline{\mathrm{div}}\,\boldsymbol{\eta}\,\overline{\mathrm{div}}\,\boldsymbol{u}\,\mathrm{d}v, \end{aligned} \tag{7} $$

where

$$ \overline{\mathrm{div}}\,\boldsymbol{u} = \frac{1}{{}^{e}\hat{\Theta}\left(\hat{\boldsymbol{\varphi}}\right)}\sum_{k=1}^{{}^{e}n_{\Gamma}}\sum_{l=1}^{{}^{e}n_{\Gamma}} {}^{e}H^{-1}_{lk}\int_{{}^{e}\Omega} {}^{e,k}\Gamma(\boldsymbol{\xi})\,{}^{e}J\left(\hat{\boldsymbol{\varphi}}\right)u_{i,i}\,\mathrm{d}V \tag{8} $$

is the so-called "discrete divergence operator", cf. [5]. Finally, the introduction of an inter-element continuous interpolation for the not yet discretized displacement field leads to `STP`-elements. Since only the remaining displacement degrees of freedom are assembled, the implementation is very similar to the procedure for purely displacement-based elements. Due to the used inter-element discontinuous pressure and dilation interpolation (3) `STP`-elements are classified as "discontinuous-type" hybrid elements.

3 CL3F-ELEMENTS

We propose hybrid/mixed elements – closely related to `STP`-elements – we call `CL3F`-elements, which can be of either the discontinuous or the continuous type. A detailed contribution including the full length derivation is currently under review, cf. [7]. The elements are based on the continuum-level linearization of the three-field potentials weak form (2) rather than the semi-discretization approach used for `STP`-elements recapped above.

We assume an isotropic, hyperelastic, quasi-incompressible, material of type

$$ \mathring{W} = \mathring{W}_{\mathrm{iso}}\left(\bar{\mathring{I}}_{1},\bar{\mathring{I}}_{2}\right) + \mathring{W}_{\mathrm{vol}}\left(\Theta\right), \tag{9} $$

where $\bar{\mathring{I}}_{1}$ and $\bar{\mathring{I}}_{2}$ refer to the first two isotropic invariants of $\mathring{\boldsymbol{C}}$. The isochoric and volumetric part of the stress are given by

$$ \sigma^{\mathrm{iso}}_{ij} = 2\,J^{-1}\,\frac{\partial\mathring{W}_{\mathrm{iso}}}{\partial\mathring{C}_{IJ}}\,\mathring{F}_{iI}\,\mathring{F}_{jJ}, \quad \sigma^{\mathrm{vol}}_{ij}(\Theta) = 2\,\Theta^{-1}\,\frac{\partial\mathring{W}_{\mathrm{vol}}}{\partial\mathring{C}_{IJ}}\,\mathring{F}_{iI}\,\mathring{F}_{jJ} = \frac{\partial\mathring{W}_{\mathrm{vol}}}{\partial\Theta}\,\delta_{ij}, \tag{10} $$

and the (isochoric) stiffness is given by

$$ c^{\mathrm{iso}}_{ijkl} = 4\,J^{-1}\,\frac{\partial^{2}\mathring{W}_{\mathrm{iso}}}{\partial\mathring{C}_{IJ}\,\partial\mathring{C}_{KL}}\,\mathring{F}_{iI}\,\mathring{F}_{jJ}\,\mathring{F}_{kK}\,\mathring{F}_{lL}. \tag{11} $$

With (10) and (11) the weak form (2) pushed to the spatial configuration reads

$$\begin{aligned} g\left(\left(\hat{\boldsymbol{\varphi}},\hat{\Theta},\hat{p}\right),(\boldsymbol{\eta},\psi,q)\right) = & \int_{\varphi(\Omega)} \left(\hat{\sigma}^{\text{iso}}_{ij}+\hat{p}\,\delta_{ij}\right)\,\eta^{\text{sym}}_{i,j}\,\mathrm{d}v \\ & + \int_{\varphi(\Omega)} \frac{\psi}{\hat{J}}\left(\frac{1}{3}\,\hat{\sigma}^{\text{vol}}_{ij}\left(\hat{\Theta}\right)\,\delta_{ij}-\hat{p}\right)\mathrm{d}v \\ & + \int_{\varphi(\Omega)} q\left(1-\frac{\hat{\Theta}}{\hat{J}}\right)\mathrm{d}v + \varphi_*\left\{\delta\Pi_{\text{ext}}(\hat{\boldsymbol{\varphi}},\boldsymbol{\eta})\right\}. \end{aligned} \tag{12}$$

Computing the Gateaux derivative of the weak form (2) for a material of type (9) at the current state $(\hat{\boldsymbol{\varphi}},\hat{\Theta},\hat{p})$ in the direction $(\boldsymbol{u},\omega,\gamma)$ (state increment), we obtain the consistent (continuum-level) linearization of the weak form, which reads pushed to the spatial configuration

$$\begin{aligned} & \delta g\Big(\Big((\hat{\boldsymbol{\varphi}},\hat{\Theta},\hat{p}),(\boldsymbol{\eta},\psi,q)\Big),(\boldsymbol{u},\omega,\gamma)\Big) \\ & \quad = \int_{\varphi(\Omega)} \Bigg\{ \hat{c}^{\text{iso}}_{ijkl}\,I^{\text{sym,dev}}_{ijab}\,I^{\text{sym,dev}}_{klcd}\,\eta_{a,b}\,u_{c,d} \\ & \qquad + \hat{\sigma}^{\text{iso}}_{jl}\left(2\,I^{\text{sym,dev}}_{ijab}\,I^{\text{dev}}_{ilcd}-I^{\text{sym,dev}}_{jlad}\delta_{bc}\right)\eta_{a,b}\,u_{c,d} \\ & \qquad + \hat{p}\,\left(\delta_{ab}\,\delta_{cd}-\delta_{cb}\delta_{ad}\right)\eta_{a,b}\,u_{c,d} \\ & \qquad + \frac{\omega\,\psi}{\hat{J}}\,\frac{\partial^2\hat{\mathring{W}}_{\text{vol}}}{\partial\Theta^2}+\gamma\,\eta_{i,i}-\frac{\gamma\,\psi}{\hat{J}}+q\left(u_{i,i}-\frac{\omega}{\hat{J}}\right)\Bigg\}\,\mathrm{d}v \\ & \quad + \varphi_*\left\{\delta\left(\delta\Pi_{\text{ext}}(\hat{\boldsymbol{\varphi}},\boldsymbol{\eta}),\boldsymbol{u}\right)\right\}, \end{aligned} \tag{13}$$

where we introduced

$$\begin{aligned} I^{\text{dev}}_{ijab} &:= \delta_{ia}\delta_{jb}-\frac{1}{3}\delta_{ij}\delta_{ab} && \Rightarrow \text{dev}(t_{ij}) = I^{\text{dev}}_{ijab}\,t_{ab}, \\ I^{\text{sym,dev}}_{ijab} &:= \frac{1}{2}\delta_{ia}\delta_{jb}+\frac{1}{2}\delta_{ib}\delta_{ja}-\frac{1}{3}\delta_{ij}\delta_{ab} && \Rightarrow \text{dev}(t^{\text{sym}}_{ij}) = I^{\text{sym,dev}}_{ijab}\,t_{ab}. \end{aligned}$$

Note that already the mathematical structure of (7) differs from (13): In the STP-linearization the dilation increment ω and pressure increment γ are missing by design, since the independent dilation and pressure variables were removed from the weak form before the linearization. The linearization (7) of the semi-discretization approach used for the STP-elements is not the consistent (continuum-level) linearization of the three-field potentials weak form (2) – it is the consistent linearization of (6).

The linearized problem of finding an incremented equilibrium state is

$$\begin{aligned} 0 &\overset{!}{=} g\left(\left(\hat{\boldsymbol{\varphi}}+\boldsymbol{u},\hat{\Theta}+\omega,\hat{p}+\gamma\right),(\boldsymbol{\eta},\psi,q)\right) \\ &\approx g\left(\left(\hat{\boldsymbol{\varphi}},\hat{\Theta},\hat{p}\right),(\boldsymbol{\eta},\psi,q)\right)+\delta g\Big(\Big((\hat{\boldsymbol{\varphi}},\hat{\Theta},\hat{p}),(\boldsymbol{\eta},\psi,q)\Big),(\boldsymbol{u},\omega,\gamma)\Big). \end{aligned} \tag{14}$$

Choosing the displacement field $\boldsymbol{U}$ as the first primary unknown (rather than the configuration $\boldsymbol{\varphi}$) and introducing a standard Lagrange interpolation

$${}^{e}U_i \approx \sum_{k=1}^{{}^{e}n_k} {}^{e,k}N^i(\boldsymbol{\xi})\,\widehat{{}^{e,k,i}U} \tag{15}$$

for $\boldsymbol{U}$ as well as different (potentially non Lagrange-type) interpolations for the dilation and pressure (3) with matching interpolations for all of the associated virtual quantities $\boldsymbol{\eta}, \psi, q$ and introducing a quadrature scheme for a specific element ${}^e\Omega$, allows us to factor out the nodal values of the virtual quantities as usual to obtain the linear equation system

$$\begin{array}{c} \\ \boldsymbol{\eta}\rightarrow \\ \psi\rightarrow \\ q\rightarrow \end{array} \begin{array}{c} \begin{array}{ccc} \overset{\boldsymbol{u}}{\downarrow} & \overset{\omega}{\downarrow} & \overset{\gamma}{\downarrow} \end{array} \\ \begin{bmatrix} {}^e\boldsymbol{K}_{uu} & \boldsymbol{0} & {}^e\boldsymbol{K}_{up} \\ \boldsymbol{0} & {}^e\boldsymbol{K}_{\Theta\Theta} & {}^e\boldsymbol{K}_{\Theta p} \\ {}^e\boldsymbol{K}_{pu} & {}^e\boldsymbol{K}_{p\Theta} & \boldsymbol{0} \end{bmatrix} \end{array} \cdot \begin{pmatrix} {}^e\boldsymbol{u} \\ {}^e\boldsymbol{\omega} \\ {}^e\boldsymbol{\gamma} \end{pmatrix} + \begin{pmatrix} {}^e\boldsymbol{R}_u - {}^e\boldsymbol{P}_u \\ {}^e\boldsymbol{R}_\Theta \\ {}^e\boldsymbol{R}_p \end{pmatrix} = \boldsymbol{0}, \tag{16}$$

on the element level from (14). By selecting (different) inter-element continuous interpolation schemes (3) and (15), we obtain continuous-type mixed elements by assembling all state variables (displacement, dilation and pressure). Although the element tangent stiffness matrix in (16) is singular (due to the diagonal element which is zero) the global tangent stiffness matrix assembled from (16) is regular.

To obtain discontinuous-type hybrid elements that are comparable to the original `STP`-elements, we choose an inter-element discontinuous interpolation scheme (3) for the dilation and pressure and an inter-element continuous one for the displacement (15). Eliminating the dilation and pressure increments from (16) (static condensation) leads to the equation system

$$\begin{aligned} &\left({}^e\boldsymbol{K}_{uu} + {}^e\boldsymbol{K}_{up}\left[{}^e\boldsymbol{K}_{\Theta p}\right]^{-1}{}^e\boldsymbol{K}_{\Theta\Theta}\left[{}^e\boldsymbol{K}_{p\Theta}\right]^{-1}{}^e\boldsymbol{K}_{pu}\right){}^e\boldsymbol{u} \\ &+{}^e\boldsymbol{R}_u - {}^e\boldsymbol{P}_u + {}^e\boldsymbol{K}_{up}\left[{}^e\boldsymbol{K}_{\Theta p}\right]^{-1}{}^e\boldsymbol{K}_{\Theta\Theta}\left[{}^e\boldsymbol{K}_{p\Theta}\right]^{-1}{}^e\boldsymbol{R}_p - {}^e\boldsymbol{K}_{up}\left[{}^e\boldsymbol{K}_{\Theta p}\right]^{-1}{}^e\boldsymbol{R}_\Theta = \boldsymbol{0}, \end{aligned} \tag{17}$$

$${}^e\boldsymbol{\omega} = -\left[{}^e\boldsymbol{K}_{p\Theta}\right]^{-1}\left({}^e\boldsymbol{K}_{pu}\,{}^e\boldsymbol{u} + {}^e\boldsymbol{R}_p\right), \tag{18}$$

$${}^e\boldsymbol{\gamma} = -\left[{}^e\boldsymbol{K}_{\Theta p}\right]^{-1}\left({}^e\boldsymbol{K}_{\Theta\Theta}\,{}^e\boldsymbol{\omega} + {}^e\boldsymbol{R}_\Theta\right). \tag{19}$$

Only the displacement degrees of freedom (DOFs) are assembled to the global system using the element tangent stiffness contributions and residual vector contributions according to (17). In contrast to the `STP`-approach, the current state values of the independent dilation $\hat{\Theta}$ and the pressure $\hat{p}$ are needed in order to compute the element stiffness and residual vector contributions. Hence, we need to keep track of these quantities, and it is mandatory that every instance of a finite element has internal state variables for the eliminated dilation and pressure DOFs. Once the global equation system – assembled from (17) – is solved, the nodal displacement increments for every element are known and the dilation and pressure increments are computed from (18) and (19) on element level. There is no need for such a secondary variable update scheme in the original `STP`-approach, since – in contrast to the increment-to-increment relations (18) and (19) used in the CL3F-approach – the current states dilation and pressure follow directly from the current displacement, cf. (4) and (5), and thus we have a purely displacement-based approach. The differences in the update schemes are illustrated by the algorithms 1 and 2.

repeat
  assemble $\boldsymbol{K}_T$ from ${}^e\boldsymbol{K}_T\left({}^e\hat{\boldsymbol{U}}, {}^e\hat{\Theta}, {}^e\hat{p}\right)$, cf. (13), (17)
  solve $\boldsymbol{K}_T\boldsymbol{u} = \lambda\boldsymbol{P}_u - \boldsymbol{R}\left(\boldsymbol{R}_u, \boldsymbol{R}_\Theta, \boldsymbol{R}_p\right)$ for $\boldsymbol{u}$, cf. (17)
  foreach element e **do**
    compute ${}^e\omega\left({}^e\boldsymbol{u}\right)$, ${}^e\gamma\left({}^e\boldsymbol{u}\right)$, cf. (18), (19)
    update ${}^e\hat{\Theta} := {}^e\hat{\Theta} + {}^e\omega$, ${}^e\hat{p} := {}^e\hat{p} + {}^e\gamma$
  end
  update $\hat{\boldsymbol{U}} := \hat{\boldsymbol{U}} + \boldsymbol{u}$
  assemble $\boldsymbol{R}$ from ${}^e\boldsymbol{R}_u\left({}^e\hat{\boldsymbol{U}}, {}^e\hat{p}\right)$, ${}^e\boldsymbol{R}_\Theta\left({}^e\hat{\boldsymbol{U}}, {}^e\hat{\Theta}, {}^e\hat{p}\right)$, ${}^e\boldsymbol{R}_p\left({}^e\hat{\boldsymbol{U}}, {}^e\hat{\Theta}\right)$, cf. (12), (17)
until $\|\lambda\boldsymbol{P}_u - \boldsymbol{R}\| \leq \text{tol}$

Algorithm 1: `CL3F` equilibrium iterations at load level $\lambda \in [0, 1]$

repeat
  assemble $\boldsymbol{K}_T$ from ${}^e\boldsymbol{K}_T\left({}^e\hat{\boldsymbol{U}}\right)$, cf. (4), (5), (7)
  solve $\boldsymbol{K}_T\boldsymbol{u} = \lambda\boldsymbol{P}_u - \boldsymbol{R}_u$ for $\boldsymbol{u}$
  update $\hat{\boldsymbol{U}} := \hat{\boldsymbol{U}} + \boldsymbol{u}$
  assemble $\boldsymbol{R}_u$ from ${}^e\boldsymbol{R}_u\left({}^e\hat{\boldsymbol{U}}\right)$, cf. (4), (5), (6)
until $\|\lambda\boldsymbol{P}_u - \boldsymbol{R}_u\| \leq \text{tol}$

Algorithm 2: `STP` equilibrium iterations at load level $\lambda \in [0, 1]$

Since the equation systems of both approaches differ from each other, the discontinuous type `CL3F`-elements differ from the original `STP`-elements, even if we choose the exact same interpolation schemes (3) and (15) in order to achieve a fair comparison. Comparing the equation systems on an element-level, it is interesting to observe that the element residual of the `STP`-approach equals the term ${}^e\boldsymbol{R}_u - {}^e\boldsymbol{P}_u$ in the `CL3F`-approach (compare (6) and the first line of (12)), if we ignore the differences between the pressures in both approaches stemming from the secondary variable update scheme. The additional summands of the vector residual of (17) stemming from ${}^e\boldsymbol{R}_\Theta$ and ${}^e\boldsymbol{R}_p$ are missing in the `STP`-approach due to the semi-discretization. On the other hand, the transformation from the system (17), (19), (18) used by the discontinuous `CL3F`-elements to the discretized consistent linearization (16) (used for the continuous type `CL3F`-elements) is an equivalence transformation on element level.

4 NUMERICAL BENCHMARK

We implemented several types of discontinuous hybrid elements in a finite element program build from scratch following the `CL3F` as well as the original `STP` approach. In order to compare the *formulations* all elements are implemented without extrapolation or any numerical stabilization. We always combine a Lagrange-type displacement interpolation with a polynomial-type dilation/pressure interpolation. The prefixes in the element names below indicate which approach was used. The interpolation schemes are selected to match the ones used for the Abaqus hybrid elements, cf. [6], whose names always start with `C3D` and end with `H`. In turn the number of assembled DOFs coincide. Also the quadrature schemes for `STP` and `CL3F`-elements are selected to match the number of integration points of the Abaqus elements. The Abaqus hybrid elements are used with default parameters, which includes the per default extrapolation.

In particular, we compare the following discontinuous hybrid elements:

- Hexahedral elements with linear Lagrange-type (8 node) displacement interpolation and constant dilation/pressure ansatz, 24 assembled DOFs: `C3D8H`, `STP-H1G8-P0`, `CL3F-H1G8-P0`

- Hexahedral elements with quadratic, serendipity-type Lagrange (20 node) displacement interpolation and linear dilation/pressure ansatz, 60 assembled DOFs: `C3D20H`, `STP-H2sG27-P1`, `CL3F-H2sG27-P1`

- Tetrahedral elements with quadratic Lagrange-type (10 node) displacement interpolation and constant dilation/pressure ansatz, 30 assembled DOFs: `C3D10H`, `STP-T2G4-P0`, `CL3F-T2G4-P0`

Also we implemented the following continuous type `CL3F`-elements for a comparison:

- `CL3F-H2sG27-L1`: A hexahedral element with quadratic, serendipity-type Lagrange (20 node) displacement interpolation and linear Lagrange-type (8 node) dilation/pressure ansatz, 76 assembled DOFs

- `CL3F-T2G4-L1`: A tetrahedral element with quadratic Lagrange-type (10 node) displacement interpolation and linear Lagrange-type (4 node) dilation/pressure ansatz, 38 assembled DOFs

To benchmark the influence of the strength of the nonlinearity in the material model, we combine the Neo-Hooke model for the isochoric part of the strain energy density ($\mu_0 = 1.0316\,\mathrm{MPa}$) with three different compression models. Sorted by the strength of the uplift from weakest to strongest nonlinearity, the compression models are:

- Ogden compression model, cf. [2], $\mathring{W}_{\mathrm{vol}} = \frac{K_0}{\beta^2}\left(\beta \ln \Theta + \Theta^{-\beta} - 1\right)$, $K_0 = 2781$ MPa, $\beta = -2$

- Standard compression model, $\mathring{W}_{\mathrm{vol}} = \frac{1}{2} K_0 \left(\Theta - 1\right)^2$, $K_0 = 2816$ MPa

- Hartmann-Neff model, cf. [1], $\mathring{W}_{\mathrm{vol}} = \frac{K_0}{2\beta^2}\left(\Theta^{\beta} + \Theta^{-\beta} - 2\right)$, $K_0 = 2290$ MPa, $\beta = 41$

All compression models are (least-square) fitted to the same experimental data obtained by confined axial compression testing of an industrial NR/IR-blend (natural rubber / isoprene rubber) with strongly nonlinear compression behavior used for damping applications, cf. [3].

The benchmark geometry, boundary conditions and the load-case are adapted from the well known block locking test, described in detail in [8, pp. 458].

Standard tests for finite element performance are usually mesh convergence studies: To asses an elements stiffness, the load-case is fixed and the mesh is refined in several steps towards the point where the solution does not change anymore within a certain tolerance. Usually one plots the maximum displacement vs. the mesh size for illustration. For each of the three groups of comparable discontinuous hybrid elements (and for each compression model), we see the same mesh convergence behavior. Here we only provide the plot for quadratic hexahedral elements for brevity, cf. Figure 1. The displacements fields coincide beside small round-off errors for different, comparable elements – even in the regime where the displacement is still mesh dependent. Especially all tested discontinuous elements are free of volumetric locking, like the original `STP`-elements, judged by the displacement-field. The continuous `CL3F`-elements have a different mesh convergence behavior, if compared to discontinuous elements with the same displacement ansatz. They need a slightly finer mesh to converge, but still the elements are free of volumetric locking. Their convergence is non-monotonic in contrast to the discontinuous elements. Also it is noteworthy that each equilibrium iteration is computationally more expensive than for comparable discontinuous elements due to the increased number of DOFs for a fixed mesh size. These disadvantages are however to some extent counterbalanced by the fact that

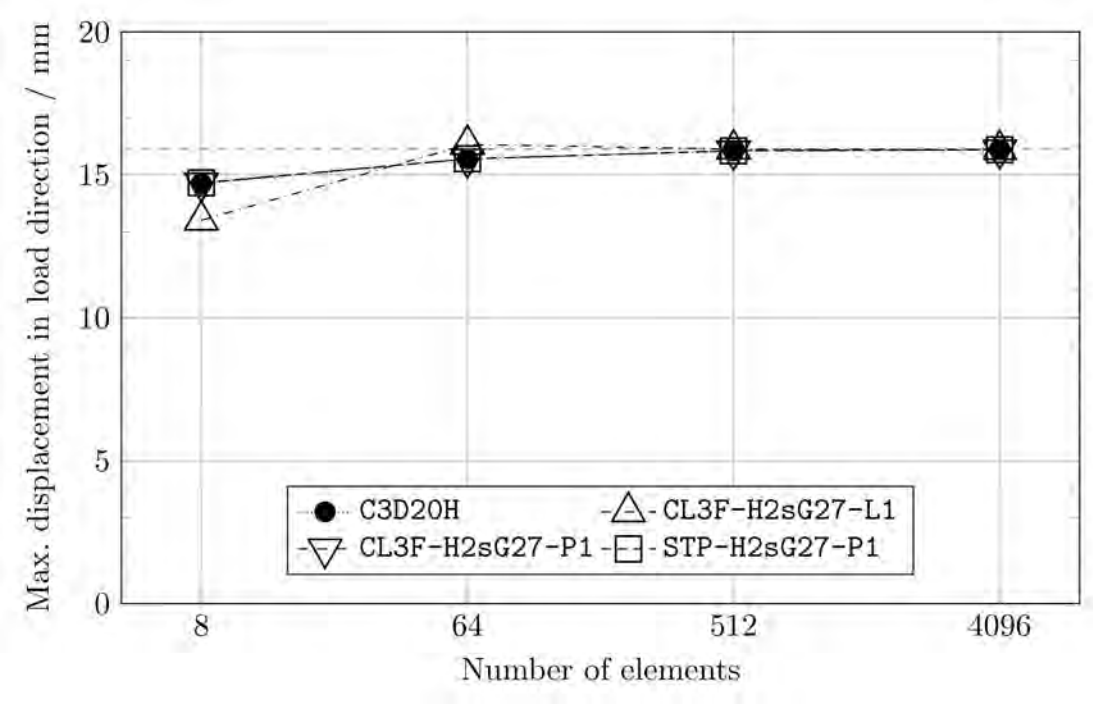


Figure 1: Mesh convergence study: quadratic hexahedral elements

the continuous elements need significantly less equilibrium iterations than the discontinuous elements in general.

The similarity in the mesh convergence behavior for the discontinuous `STP-` and `CL3F-` elements was expected since the semi-discretization approach eliminates variables from the weak form so that we did not expect pronounced differences for converged (i.e. equilibrium) states. However, due to the differences in the tangent stiffness matrix and right-hand side the elements differ in terms of the numerical stability. Under numerical stability (or robustness) we understand the maximum load-step size that can be applied (for a specific combination of hybrid-element and material model) so that the (only locally converging) Newton-Raphson scheme still converges. As a benchmark test we increase the total load, which is always applied in five equidistant steps, until for one of the five load steps the equilibrium iterations diverge. The maximum load for each combination of hybrid element and material model is displayed in Figure 2. It should be mentioned at this point, that a stability benchmark like this is in contrast to mesh convergence studies not at all a standard test in the literature, although the ability to apply large load-steps is obviously a desirable property.

In Abaqus only the standard compression model is readily available. Therefore, in combination with the standard compression model we tested the Abaqus elements (in red) twice: In combination with the internal material model and with the same material model implemented by a `UMAT`. (The nonstandard models are both implemented by `UMAT`s for the testing of the Abaqus hybrid elements.) Interestingly, the internal standard model performs slightly better than the `UMAT` implementation. Comparing only `UMAT` implementations, the Abaqus elements perform worse for compression models with stronger nonlinearity. Especially for the Hartmann-Neff model, the stable step-width of the Abaqus elements is significantly reduced. The original (not augmented) `STP`-elements (in gray) perform worse than the Abaqus elements and suffer from severe stability issues in general. In contrast, the (not augmented) `CL3F`-elements (in blue) perform better than the Abaqus hybrid elements in general. Furthermore, in contrast to the Abaqus elements, all `CL3F`-elements achieve the same stable step-width (within the margin of error of the test) independent of the used compression model. In turn, the advantage of

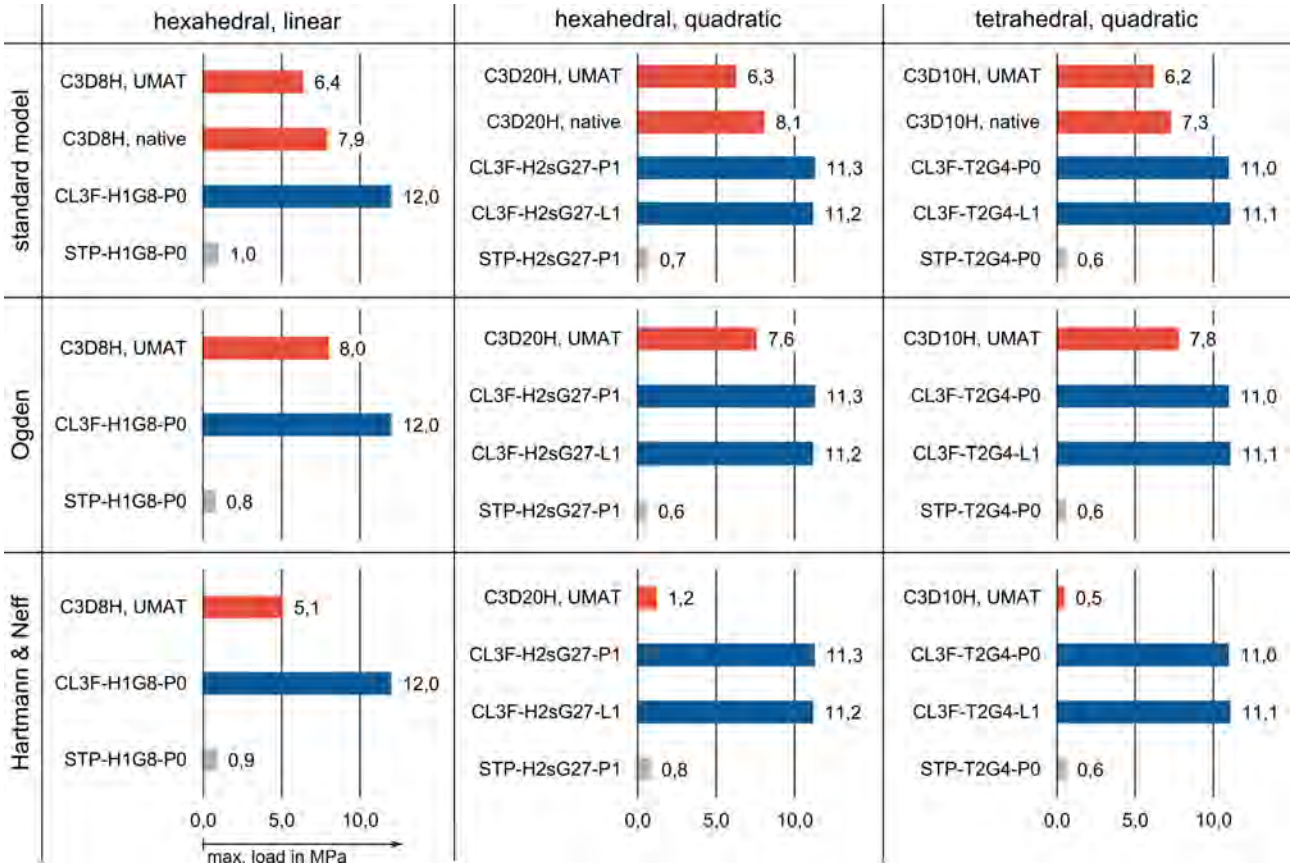


Figure 2: Numerical stability benchmark

the CL3F-elements is huge for the strongly nonlinear Hartmann-Neff model. (Here, the stable step-width is up to 22 times larger.) Another observation is that within the margin of error, we observe no difference in terms of the numerical stability between continuous CL3F-elements and comparable discontinuous CL3F-elements.

5 CONCLUSIONS

Comparing the Abaqus hybrid elements to the STP- and CL3F-elements it should be emphazized that the Abaqus/Standard default settings, especially the default extrapolation was used for the Abaqus hybrid elements, i.e. a method to determine the first guess to the incremental solution. In contrast the initial value of the incremental solution for the implementation of the STP- and CL3F-elements was simply zero. Also, no stabilization techniques were used for the STP- and CL3F-elements. Hence, the Abaqus hybrid elements (which are used as provided) have a huge advantage in the stability benchmark. Therefore, it is remarkable that the CL3F-elements overcome the stability issues of the STP-elements to an extend that they outperform the production stage Abaqus elements, especially since recent hybrid-element formulations often rely heavily on matching numerical stabilization techniques.

REFERENCES

[1] Hartmann, S. and Neff, P., Polyconvexity of generalized polynomial-type hyperelastic strain energy functions for near-incompressibility. *International Journal of Solids and Structures*, Vol. **40**, pp. 2767–2791, (2003).

[2] Ogden, R.W., Large deformation isotropic elasticity - on the correlation of theory and experiment for incompressible rubberlike solids. *Proceedings of the Royal Society A: Math-*

ematical, Physical and Engineering Sciences, Vol. **326**, pp. 565–584, (1972).

[3] Ricker, A., Fehse, A., Kröger, N. H. *Final report on IGF project No. 19916 N, characterization and modeling of compression moduli of technical rubber materials*, 2020. (in German)

[4] Simo, J.C., Taylor, R.L. and Pister, K.S., Variational and projection methods for the volume constraint in finite deformation elasto-plasticity. *Computer Methods in Applied Mechanics and Engineering*, Vol. **51**, pp. 177–208, (1985).

[5] Simo, J.C. and Taylor, R.L., Quasi-incompressible finite elasticity in principal stretches. continuum basis and numerical algorithms. *Computer Methods in Applied Mechanics and Engineering*, Vol. **85**(3), pp. 273–310, (1991).

[6] Simulia, *Abaqus 6.14 theory guide*, Dassault Systèmes Simulia Corp., section 3.2.3., (2014).

[7] Schönherr, J.A., Schneider P. and Mittelstedt, C., Robust hybrid/mixed finite elements for rubber-like materials under severe compression. Submitted to *Computational Mechanics*, under review, (2021).

[8] Wriggers, P. *Nonlinear finite element methods.* Springer, 2008.

Energy-momentum time integration of gradient-based models for fiber-bending stiffness in anisotropic thermo-viscoelastic continua

J. Dietzsch[1], M. Groß[2] and I. Kalaimani[3]

[1] Technische Universität Chemnitz, Professorship of applied mechanics and dynamics Reichenhainer Straße 70, D-09126 Chemnitz, julian.dietzsch@mb.tu-chemnitz.de

[2] michael.gross@mb.tu-chemnitz.de [3] iniyan.kalaimani@mb.tu-chemnitz.de

Key words: Fiber-bending stiffness, fiber-reinforced materials, locking behavior, mixed finite elements, mixed variational principle, gradient elasticity.

Abstract: *For our research, we are motivated by dynamic simulations of 3D fiber-reinforced materials in lightweight structures. In such materials, the material reinforcement is performed by fiber rovings with a separate bending stiffness, which can be modelled by a second-order gradient of the deformation mapping (see Reference [1]). Therefore, we extend a thermo-viscoelastic Cauchy continuum for fiber-matrix composites with single fibers by an independent field for the gradient of the right Cauchy-Green tensor. On the other hand, we focus on numerically stable dynamic long-time simulations with locking free meshes, and thus use higher-order accurate energy-momentum schemes emanating from mixed finite element methods. Hence, we adapt the variational-based space-time finite element method in Reference [2] to the new material formulation, and additionally include independent fields to obtain well-known mixed finite elements [3, 4, 5]. As representative numerical example, Cook's cantilever beam is considered. We primarily analyze the influence of the fiber bending stiffness, as well as the spatial and time convergence up to cubic order. Furthermore, we look at the influence of the physical dissipation in the material.*

1 INTRODUCTION

We consider an anisotropic material with the fiber roving direction $\boldsymbol{a}_0$, moving in the Euclidean space $\mathbb{R}^{n_{\text{dim}}}$ with the constant ambient temperature Θ_∞. The strain energy function of the material with a thermo-viscoelastic matrix and a thermoelastic fiber roving is given by

$$\Psi(\boldsymbol{C},\boldsymbol{C}_v,\Theta,\boldsymbol{a}_0) = \Psi_{\text{M}}(\boldsymbol{C},\boldsymbol{C}_v,\Theta) + \Psi_{\text{F}}(\boldsymbol{C},\Theta,\boldsymbol{a}_0) + \Psi^{\text{X}}_{\text{HOG}}(\dots,\boldsymbol{a}_0), \tag{1}$$

which is split into a matrix part Ψ_{M} a fiber roving part Ψ_{F} and a higher-order gradient part $\Psi^{\text{X}}_{\text{HOG}}$. Here $\boldsymbol{F} = \nabla \boldsymbol{q}$ define the deformation gradient by the position $\boldsymbol{q}$, $\boldsymbol{C} = \boldsymbol{F}^T\boldsymbol{F}$ define the right Cauchy-Green tensor, $\boldsymbol{C}_v$ define the viscous right Cauchy-Green tensor and Θ define the absolute temperature. The specific dependencies are given by

$$\Psi_{\text{M}}(\boldsymbol{C},\boldsymbol{C}_v,J,\Theta) = \Psi^{\text{iso}}_{\text{M}}(\boldsymbol{C},J) + \Psi^{\text{vol}}_{\text{M}}(J) + \Psi^{\text{cap}}_{\text{M}}(\Theta) + \Psi^{\text{coup}}_{\text{M}}(\Theta,J) + \Psi^{\text{vis}}_{\text{M}}(\boldsymbol{C}\boldsymbol{C}_v^{-1}) \tag{2}$$

$$\Psi_{\text{F}}(\boldsymbol{C},\Theta,\boldsymbol{a}_0,\dots) = \Psi^{\text{ela}}_{\text{F}}(\boldsymbol{C},\boldsymbol{a}_0) + \Psi^{\text{cap}}_{\text{F}}(\Theta) + \Psi^{\text{coup}}_{\text{F}}(\Theta,\boldsymbol{C}) \tag{3}$$

with the volume dilatation $J(\boldsymbol{C}) = \det[\boldsymbol{F}] = \sqrt{\det[\boldsymbol{C}]}$. The elastic part of the matrix function Ψ_{M} is split into an isochoric part $\Psi^{\text{iso}}_{\text{M}}$, a volumetric part $\Psi^{\text{vol}}_{\text{M}}$, a heat capacity part $\Psi^{\text{cap}}_{\text{M}}$, a part of the thermo-mechanical coupling effect $\Psi^{\text{coup}}_{\text{M}}$ and the viscoelastic free energy function of the matrix $\Psi^{\text{vis}}_{\text{M}}$. The parts of the fiber free energy is separated in the same manner. It is split into an elastic part $\Psi^{\text{ela}}_{\text{F}}$, a heat capacity part $\Psi^{\text{cap}}_{\text{F}}$ and a part of the thermo-mechanical coupling effect $\Psi^{\text{coup}}_{\text{F}}$. The functions of the

https://doi.org/10.4995/YIC2021.2021.12358

thermo-mechanical coupling $\Psi_{\mathrm{X}}^{\mathrm{coup}}$ with the coefficients of linear thermal expansion β_{X}, the structural tensor $\boldsymbol{M} = \boldsymbol{a}_0 \otimes \boldsymbol{a}_0$ and the fourth invariant $I_4 = \mathrm{tr}[\boldsymbol{C}\boldsymbol{M}]$ are given by

$$\Psi_{\mathrm{M}}^{\mathrm{coup}} = -2n_{\mathrm{dim}}\beta_{\mathrm{M}}(\Theta - \Theta_\infty)J\frac{\partial \Psi_{\mathrm{M}}^{\mathrm{vol}}(J)}{\partial J} \quad \Psi_{\mathrm{F}}^{\mathrm{coup}} = -2\beta_{\mathrm{F}}(\Theta - \Theta_\infty)\sqrt{I_4}\frac{\partial \Psi_{\mathrm{F}}^{\mathrm{ela}}(I_4,\dots)}{\partial I_4} \tag{4}$$

We distinguish between two different variants for the higher-order gradient part $\Psi_{\mathrm{HOG}}^{\mathrm{X}}$. One concerning the gradient of the deformation gradient $\boldsymbol{F}$ and one concerning the gradient of the right Cauchy-Green tensor $\boldsymbol{C}$. In comparison with $\Psi_{\mathrm{F}}^{\mathrm{ela}}$ which considers the fiber roving stretch, this part capture the bending of the fiber roving. The formulation regarding $\boldsymbol{F}$ is shown in Reference [1]. Here the sixth and seventh invariants are given by

$$I_6^{\mathrm{F}}(\boldsymbol{F},\nabla\boldsymbol{F}) = \boldsymbol{\kappa}_0^{\mathrm{F}} \cdot \boldsymbol{\kappa}_0^{\mathrm{F}} \qquad I_7^{\mathrm{F}}(\boldsymbol{F},\nabla\boldsymbol{F},\boldsymbol{C}) = \boldsymbol{\kappa}_0^{\mathrm{F}} \cdot \boldsymbol{C} \cdot \boldsymbol{\kappa}_0^{\mathrm{F}} \qquad \boldsymbol{\kappa}_0^{\mathrm{F}} = \boldsymbol{\Lambda}^{\mathrm{F}} \cdot \boldsymbol{a}_0 \tag{5}$$

with the referential representation

$$\boldsymbol{\Lambda}^{\mathrm{F}}(\boldsymbol{F},\nabla\boldsymbol{F}) = \boldsymbol{F}^T \cdot \boldsymbol{a}_0 \cdot \nabla\boldsymbol{F}^T \tag{6}$$

It is important to note here, that I_7 is depend on $\boldsymbol{C}$ as well as $\boldsymbol{\Lambda}$. Thus, for the strain energy function of the higher-order gradient, the dependencies are

$$\Psi_{\mathrm{HOG}}^{\mathrm{F}}(\boldsymbol{\Lambda}^{\mathrm{F}},\boldsymbol{C},\boldsymbol{a}_0) \quad = \quad \hat{f}(I_6^{\mathrm{F}}(\boldsymbol{\Lambda}^{\mathrm{F}}), I_7^{\mathrm{F}}(\boldsymbol{\Lambda}^{\mathrm{F}},\boldsymbol{C})) \tag{7}$$

A variant of the higher-order gradient formulation in $\boldsymbol{C}$ is shown in Reference [6]. From this we derive the following formula for the sixth invariant

$$I_6^{C}(\nabla\boldsymbol{C}) = (\boldsymbol{a}_0 \cdot \nabla\boldsymbol{C} \cdot \boldsymbol{a}_0) \cdot (\boldsymbol{a}_0 \cdot \nabla\boldsymbol{C} \cdot \boldsymbol{a}_0) \tag{8}$$

If we now set

$$\boldsymbol{\Lambda}^{\mathrm{C}}(\nabla\boldsymbol{C}) = \boldsymbol{a}_0 \cdot \nabla\boldsymbol{C} \tag{9}$$

we get the same expressions for the invariants as for $\boldsymbol{F}$, given by

$$I_6^{\mathrm{C}}(\nabla\boldsymbol{C}) = \boldsymbol{\kappa}_0^{\mathrm{C}} \cdot \boldsymbol{\kappa}_0^{\mathrm{C}} \qquad I_7^{\mathrm{C}}(\boldsymbol{C},\nabla\boldsymbol{C}) = \boldsymbol{\kappa}_0^{\mathrm{C}} \cdot \boldsymbol{C} \cdot \boldsymbol{\kappa}_0^{\mathrm{C}} \qquad \boldsymbol{\kappa}_0^{\mathrm{C}} = \boldsymbol{\Lambda}^{\mathrm{C}} \cdot \boldsymbol{a}_0 \tag{10}$$

and the final dependencies read

$$\Psi_{\mathrm{HOG}}^{\mathrm{C}}(\nabla\boldsymbol{C},\boldsymbol{C},\boldsymbol{a}_0) = f(I_6^{\mathrm{C}}(\nabla\boldsymbol{C}), I_7^{\mathrm{C}}(\nabla\boldsymbol{C},\boldsymbol{C})) \tag{11}$$

2 FINITE ELEMENT FORMULATION

The finite element discretization follows from the mixed principle of virtual power (see Reference [5, 2]). Here, we need the complete internal energy, which consists of the assumed temperature field $\tilde{\Theta}$, the entropy density field η as the corresponding Lagrange multiplier, the superimposed stress tensor $\tilde{\boldsymbol{S}}$ to derive an energy–momentum scheme, an independent mixed field $\tilde{\boldsymbol{C}}$ and the corresponding Lagrange multiplier $\boldsymbol{S}$. The internal energy functional reads

$$\begin{aligned}\Pi^{\mathrm{int}} = &\int_{\mathcal{B}_0} \Psi_{\mathrm{M}}(\tilde{\boldsymbol{C}},\tilde{J},\Theta)\mathrm{d}V + \int_{\mathcal{B}_0} \Psi_{\mathrm{F}}(\tilde{\boldsymbol{C}}_A,\Theta)\mathrm{d}V + \int_{\mathcal{B}_0} \frac{1}{2}\boldsymbol{S} : (\boldsymbol{C}(\boldsymbol{q}) - \tilde{\boldsymbol{C}})\mathrm{d}V + \int_{\mathcal{B}_0} \tilde{\boldsymbol{S}} : \tilde{\boldsymbol{C}}\mathrm{d}V \\ &+ \int_{\mathcal{B}_0} \eta\,(\Theta - \tilde{\Theta}) + \int_{\mathcal{B}_0} p\,(J(\tilde{\boldsymbol{C}}) - \tilde{J})\mathrm{d}V + \int_{\mathcal{B}_0} \tilde{p}\tilde{J}\mathrm{d}V + \int_{\mathcal{B}_0} \frac{1}{2}\boldsymbol{S}_A : (\tilde{\boldsymbol{C}} - \tilde{\boldsymbol{C}}_A)\mathrm{d}V \\ &+ \int_{\mathcal{B}_0} \tilde{\boldsymbol{S}}_A : \tilde{\boldsymbol{C}}_A\mathrm{d}V + \Pi_{\mathrm{HOG}}^{\mathrm{X}}\end{aligned} \tag{12}$$

We introduce an independent volume dilatation $\tilde{J}$ (see Reference [3]) and the field $\tilde{\boldsymbol{C}}_A$ (see Reference [4]) for the anisotropic part Ψ_{F} to avoid locking effects. Here, the Lagrange multiplier p plays the role of the hydrostatic pressure and the Lagrange multiplier $\boldsymbol{S}_A$ represents the stress tensor of the anisotropic part. To obtain an energy–momentum scheme, we also introduce the superimposed pressure $\tilde{p}$ and superimposed stress tensor $\tilde{\boldsymbol{S}}_A$. For the higher-order gradient fomulation with respect to $\boldsymbol{F}$ (HF), we introduce an independent field for $\boldsymbol{F}$, for $\nabla\boldsymbol{F}$ and for $\boldsymbol{\Lambda}^{\mathrm{F}}$

$$\begin{aligned}\Pi^{\mathrm{F}}_{\mathrm{HOG}} = &\int_{\mathcal{B}_0} \tilde{\boldsymbol{P}} : (\boldsymbol{F} - \tilde{\boldsymbol{F}})\mathrm{d}V + \int_{\mathcal{B}_0} \boldsymbol{B} \odot_3 (\nabla(\tilde{\boldsymbol{F}}) - \tilde{\boldsymbol{\Gamma}})\mathrm{d}V + \int_{\mathcal{B}_0} \boldsymbol{H} : (\boldsymbol{\Lambda}^{\mathrm{F}}(\tilde{\boldsymbol{F}}, \tilde{\boldsymbol{\Gamma}}) - \tilde{\boldsymbol{\Lambda}})\mathrm{d}V \\ &+ \int_{\mathcal{B}_0} \Psi^{\mathrm{F}}_{\mathrm{HOG}}(\tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{C}}_A, \boldsymbol{a}_0)\mathrm{d}V + \int_{\mathcal{B}_0} \tilde{\boldsymbol{H}} : \tilde{\boldsymbol{\Lambda}}\mathrm{d}V \end{aligned} \tag{13}$$

By the independent definition of $\tilde{\boldsymbol{F}}$ and $\tilde{\boldsymbol{\Gamma}}$ it is later in the discrete setting not necessary to construct a double gradient of the spatial shape functions. The introduction of $\tilde{\boldsymbol{\Lambda}}$ is necessary to have an objective quantity for the construction of an energy–momentum scheme with the superimposed field $\tilde{\boldsymbol{H}}$. For the higher-order gradient fomulation with respect to $\boldsymbol{C}$ (HC), we build the functional in a similar way

$$\begin{aligned}\Pi^{\mathrm{C}}_{\mathrm{HOG}} = &\int_{\mathcal{B}_0} \frac{1}{2}\boldsymbol{S}_G : (\boldsymbol{C} - \tilde{\boldsymbol{C}}_G) + \int_{\mathcal{B}_0} \boldsymbol{B} \odot_3 (\nabla(\tilde{\boldsymbol{C}}_G) - \tilde{\boldsymbol{\Gamma}})\mathrm{d}V + \int_{\mathcal{B}_0} \Psi^{\mathrm{C}}_{\mathrm{HOG}}(\tilde{\boldsymbol{\Gamma}}, \tilde{\boldsymbol{C}}_A, \boldsymbol{a}_0)\mathrm{d}V \\ &+ \int_{\mathcal{B}_0} \tilde{\boldsymbol{B}} \odot_3 \tilde{\boldsymbol{\Gamma}}\mathrm{d}V \end{aligned} \tag{14}$$

We introduce an independent field for $\boldsymbol{C}$ and $\nabla\boldsymbol{C}$. The further field with respect to $\boldsymbol{C}$ is introduced because $\boldsymbol{S}_G$ is assumed to be asymmetric, and therefore no symmetries in the Voigt notation are used later in the programming. Compared to the formulation in $\boldsymbol{F}$ (HF), we build the superimposed field based on $\tilde{\boldsymbol{\Gamma}}$. Furthermore, this leads to a less complex weak form. The superimposed fields (see Reference [2] and [5]), which have both variants in common, are given by

$$\tilde{\boldsymbol{S}} = \frac{\tilde{\Psi}(1) - \tilde{\Psi}(0) - \int \frac{\partial \Psi^{\mathrm{iso}}_{\mathrm{M}}}{\partial \tilde{\boldsymbol{C}}} : \dot{\tilde{\boldsymbol{C}}} - \int \frac{\partial (\Psi^{\mathrm{cap}}_{\mathrm{M}} + \Psi^{\mathrm{cap}}_{\mathrm{F}})}{\partial \Theta}\dot{\Theta} - \int \frac{\partial \Psi^{\mathrm{vis}}_{\mathrm{M}}}{\partial \boldsymbol{C}_v} : \dot{\boldsymbol{C}}_v}{\dot{\tilde{\boldsymbol{C}}} : \dot{\tilde{\boldsymbol{C}}}}\dot{\tilde{\boldsymbol{C}}} \tag{15}$$

$$\tilde{p} = \frac{\tilde{\Psi}(1) - \tilde{\Psi}(0) - \int \frac{\partial (\Psi^{\mathrm{iso}}_{\mathrm{M}} + \Psi^{\mathrm{vol}}_{\mathrm{M}})}{\partial \tilde{J}}\dot{\tilde{J}} - \int \frac{\partial \Psi^{\mathrm{coup}}_{\mathrm{M}}}{\partial \Theta}\dot{\Theta}}{\dot{\tilde{J}}\dot{\tilde{J}}}\dot{\tilde{J}} \tag{16}$$

$$\tilde{\boldsymbol{S}}_A = \frac{\tilde{\Psi}(1) - \tilde{\Psi}(0) - \int \frac{\partial \Psi^{\mathrm{ela}}_{\mathrm{F}}}{\partial \tilde{\boldsymbol{C}}_A} : \dot{\tilde{\boldsymbol{C}}}_A - \int \frac{\partial \Psi^{\mathrm{coup}}_{\mathrm{F}}}{\partial \Theta}\dot{\Theta}}{\dot{\tilde{\boldsymbol{C}}}_A : \dot{\tilde{\boldsymbol{C}}}_A}\dot{\tilde{\boldsymbol{C}}}_A \tag{17}$$

and the superimposed fields regarding the different higher-order gradient formulations read

$$\tilde{\boldsymbol{H}} = \frac{\tilde{\Psi}(1) - \tilde{\Psi}(0) - \int \frac{\partial \Psi^{\mathrm{F}}_{\mathrm{HOG}}}{\partial \tilde{\boldsymbol{\Lambda}}} : \dot{\tilde{\boldsymbol{\Lambda}}}}{\dot{\tilde{\boldsymbol{\Lambda}}} : \dot{\tilde{\boldsymbol{\Lambda}}}}\dot{\tilde{\boldsymbol{\Lambda}}} \qquad \tilde{\boldsymbol{B}} = \frac{\tilde{\Psi}(1) - \tilde{\Psi}(0) - \int \frac{\partial \Psi^{\mathrm{C}}_{\mathrm{HOG}}}{\partial \tilde{\boldsymbol{\Gamma}}} \odot_3 \dot{\tilde{\boldsymbol{\Gamma}}}}{\dot{\tilde{\boldsymbol{\Gamma}}} \odot_3 \dot{\tilde{\boldsymbol{\Gamma}}}}\dot{\tilde{\boldsymbol{\Gamma}}} \tag{18}$$

For the mixed principle of virtual power, we also need the kinetic power, given by

$$\dot{T} = \int_{\mathcal{B}_0} (\rho_0 \boldsymbol{v} - \boldsymbol{p}) \cdot \dot{\boldsymbol{v}}\mathrm{d}V + \int_{\mathcal{B}_0} \dot{\boldsymbol{p}} \cdot (\dot{\boldsymbol{q}} - \boldsymbol{v})\mathrm{d}V + \int_{\mathcal{B}_0} \boldsymbol{p} \cdot \ddot{\boldsymbol{q}}\mathrm{d}V \tag{19}$$

with the velocity $\boldsymbol{v}$, the linear momentum $\boldsymbol{p}$ and the mass density ρ_0. As external power, we assume

$$\begin{aligned}\dot{\Pi}^{\mathrm{ext}} = &-\int_{\mathcal{B}_0} \rho_0 \boldsymbol{g} \cdot \dot{\boldsymbol{q}}\mathrm{d}V - \int_{\partial\mathcal{B}_0} \boldsymbol{\lambda}_q \cdot (\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}^{\mathrm{ref}})\mathrm{d}A + \int_{\mathcal{B}_0} \nabla\left(\frac{\tilde{\Theta}}{\Theta}\right) \cdot \boldsymbol{Q}\mathrm{d}V + \int_{\mathcal{B}_0} \frac{\tilde{\Theta}}{\Theta} D^{\mathrm{int}}\mathrm{d}V \\ &+ \int_{\mathcal{B}_0} \dot{\boldsymbol{C}}_v : \mathbb{V}(\boldsymbol{C}_v) : \dot{\boldsymbol{C}}_v\mathrm{d}V \quad \boldsymbol{Q} = -\left[J(\tilde{\boldsymbol{C}}_A)\frac{k_{\mathrm{F}} - k_{\mathrm{M}}}{\tilde{\boldsymbol{C}}_A : \boldsymbol{M}}\boldsymbol{M} + kJ(\tilde{\boldsymbol{C}})\tilde{\boldsymbol{C}}^{-1}\right]\nabla\Theta \end{aligned} \tag{20}$$

Here, we have the Piola heat flux vector $\boldsymbol{Q}$ derived from Duhamel's law (see Reference [2]), where k_{M} and k_{F} denotes the material conductivity coefficients for matrix and fiber roving. The time evolution of a prescribed boundary displacement is given by $\dot{\boldsymbol{q}}^{\mathrm{ref}}$ with the Lagrange multiplier $\boldsymbol{\lambda}_q$. The vector $\boldsymbol{g}$ denotes the gravitational force. The non-negative internal viscous dissipation D^{int} is given by

$$D^{\mathrm{int}} = \dot{\boldsymbol{C}}_v : \mathbb{V}(\boldsymbol{C}_v) : \dot{\boldsymbol{C}}_v \qquad \mathbb{V}(\boldsymbol{C}_v) = \frac{1}{4}\left(V_{\mathrm{vol}} - \frac{V_{\mathrm{dev}}}{n_{\mathrm{dim}}}\right)\boldsymbol{C}_v^{-1}\otimes\boldsymbol{C}_v^{-1} + \frac{V_{\mathrm{dev}}}{4}\mathbb{I}_{\mathrm{s}} : \boldsymbol{C}_v^{-1}\otimes\boldsymbol{C}_v^{-1}, \tag{21}$$

with the viscosity constants V_{vol} and V_{dev}, which represent the volumetric and deviatoric viscosity constants and the fourth-order symmetric projection tensor $\mathbb{I}_{\mathrm{s}}$. The operator $\otimes$ represents the standard dyadic product.

The total energy balance $\dot{\mathcal{H}}$ thus reads

$$\dot{\mathcal{H}} = \dot{T}(\dot{\boldsymbol{q}},\dot{\boldsymbol{v}},\dot{\boldsymbol{p}}) + \dot{\Pi}^{\mathrm{ext}}(\dot{\boldsymbol{q}},\boldsymbol{\lambda}_q,\dot{\boldsymbol{C}}_v,\tilde{\Theta},\dot{\Theta}) + \dot{\Pi}^{\mathrm{int}}(\dot{\boldsymbol{q}},\tilde{\Theta},\dot{\eta},\dot{\boldsymbol{C}}_v,\dot{\tilde{\boldsymbol{C}}},\dot{\tilde{J}},\dot{\tilde{\boldsymbol{C}}}_A,\boldsymbol{S},p,\boldsymbol{S}_A,\dots) \tag{22}$$

Note, that we define the superimposed fields ($\tilde{\boldsymbol{S}}$,$\tilde{p}$,$\tilde{\boldsymbol{S}}_A$,$\tilde{\boldsymbol{H}}$,$\tilde{\boldsymbol{B}}$), the viscous dissipation D^{int} as well as the Piola heat flux vector $\boldsymbol{Q}$ as parameters not as arguments. We obtain the total weak forms by variation with respect to the variables in the argument of Eqn. (22). With $\int_T \delta_*\dot{\mathcal{H}}\mathrm{d}t \equiv \int_T[\delta_*\dot{T} + \delta_*\dot{\Pi}^{\mathrm{ext}} + \delta_*\dot{\Pi}^{\mathrm{int}}]\mathrm{d}t = 0$, the weak forms which occur in both variants of the higher-order gradient formulation read

$$\int_T\int_{\mathcal{B}_0}\left[\frac{1}{\rho_0}\boldsymbol{p} - \dot{\boldsymbol{q}}\right]\cdot\delta\dot{\boldsymbol{v}}\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\partial\mathcal{B}_0}\left[-\boldsymbol{\lambda}_q\right]\cdot\delta\dot{\boldsymbol{q}}\mathrm{d}A\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\eta + \frac{\partial\Psi}{\partial\Theta}\right]\delta\dot{\Theta}\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[\frac{\mathrm{Div}[\boldsymbol{Q}]}{\Theta} + \frac{D^{\mathrm{int}}}{\Theta} + \dot{\eta}\right]\delta\tilde{\Theta}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\frac{1}{2}\left[\dot{\tilde{\boldsymbol{C}}} - \dot{\boldsymbol{C}}\right] : \delta\boldsymbol{S}\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[\Theta - \tilde{\Theta}\right]\delta\dot{\eta}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\frac{\partial\Psi}{\partial\boldsymbol{C}_v} + \dot{\boldsymbol{C}}_v : \mathbb{V}(\boldsymbol{C}_v)\right] : \delta\dot{\boldsymbol{C}}_v\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\partial\mathcal{B}_0}\left[\dot{\tilde{\boldsymbol{q}}} - \dot{\boldsymbol{q}}^{\mathrm{ref}}(t)\right]\cdot\delta\boldsymbol{\lambda}_q\mathrm{d}A\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\dot{\tilde{J}} - \dot{J}\right]\delta p\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[p - \left[\frac{\partial\Psi}{\partial\tilde{J}} + \tilde{p}\right]\right]\delta\dot{\tilde{J}}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\frac{1}{2}\left[\dot{\tilde{\boldsymbol{C}}}_A - \dot{\tilde{\boldsymbol{C}}}\right] : \delta\boldsymbol{S}_A\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[\frac{1}{2}\boldsymbol{S}_A - \left[\frac{\partial\Psi}{\partial\tilde{\boldsymbol{C}}_A} + \tilde{\boldsymbol{S}}_A\right]\right] : \delta\dot{\tilde{\boldsymbol{C}}}_A\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\frac{1}{2}\boldsymbol{S} - \left(\frac{\partial\Psi}{\partial\tilde{\boldsymbol{C}}} + \frac{p}{2J(\tilde{\boldsymbol{C}})}\mathrm{cof}[\tilde{\boldsymbol{C}}] + \frac{1}{2}\boldsymbol{S}_A + \tilde{\boldsymbol{S}}\right)\right] : \delta\dot{\tilde{\boldsymbol{C}}}\mathrm{d}V\mathrm{d}t = 0$$

The weak forms associated with the higher-order gradient formulation in $\boldsymbol{F}$ (HF) are given by

$$\int_T\int_{\mathcal{B}_0}\left[\boldsymbol{S} : \frac{1}{2}\frac{\partial\dot{\boldsymbol{C}}}{\partial\dot{\boldsymbol{q}}} + \boldsymbol{P} : \frac{\partial\dot{\boldsymbol{F}}}{\partial\dot{\boldsymbol{q}}} - \dot{\boldsymbol{p}}\right]\cdot\delta_*\dot{\boldsymbol{q}}\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[\dot{\tilde{\boldsymbol{F}}} - \dot{\boldsymbol{F}}\right] : \delta_*\boldsymbol{P}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\boldsymbol{P} - \left(\boldsymbol{H} : \frac{\partial\boldsymbol{\Lambda}^{\mathrm{F}}}{\partial\dot{\tilde{\boldsymbol{F}}}} + \boldsymbol{B}\odot_3\frac{\partial\nabla\dot{\tilde{\boldsymbol{F}}}}{\partial\dot{\tilde{\boldsymbol{F}}}}\right)\right] : \delta_*\dot{\tilde{\boldsymbol{F}}}\mathrm{d}V\mathrm{d}t \qquad \int_T\int_{\mathcal{B}_0}\left[\nabla(\dot{\tilde{\boldsymbol{F}}}) - \dot{\tilde{\boldsymbol{\Gamma}}}\right]\odot_3\delta_*\boldsymbol{B}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\boldsymbol{\Lambda}^{\mathrm{F}} - \tilde{\boldsymbol{\Lambda}}\right] : \delta_*\boldsymbol{H}\mathrm{d}V\mathrm{d}t = 0 \qquad \int_T\int_{\mathcal{B}_0}\left[\boldsymbol{H} - \left[\frac{\partial\Psi}{\partial\tilde{\Lambda}} + \tilde{\boldsymbol{H}}\right]\right] : \delta_*\dot{\tilde{\boldsymbol{\Lambda}}}\mathrm{d}V\mathrm{d}t = 0$$

$$\int_T\int_{\mathcal{B}_0}\left[\boldsymbol{B} - \boldsymbol{H} : \frac{\partial\boldsymbol{\Lambda}^{\mathrm{F}}}{\partial\dot{\tilde{\boldsymbol{\Gamma}}}}\right]\odot_3\delta_*\dot{\tilde{\boldsymbol{\Gamma}}}\mathrm{d}V\mathrm{d}t = 0$$

and the weak forms associated with the higher-order gradient formulation in $\boldsymbol{C}$ (HC) take the form

$$\int_T \int_{\mathcal{B}_0} \left[\boldsymbol{S} : \frac{1}{2}\frac{\partial \dot{\boldsymbol{C}}}{\partial \dot{\boldsymbol{q}}} + \boldsymbol{S}_G : \frac{1}{2}\frac{\partial \dot{\boldsymbol{C}}}{\partial \dot{\boldsymbol{q}}} - \dot{\boldsymbol{p}} \right] \cdot \delta_* \dot{\boldsymbol{q}} \mathrm{d}V \mathrm{d}t = 0 \qquad \int_T \int_{\mathcal{B}_0} \left[\dot{\tilde{\boldsymbol{C}}} - \dot{\tilde{\boldsymbol{C}}}_G \right] : \delta_* \boldsymbol{S}_G \mathrm{d}V \mathrm{d}t = 0$$

$$\int_T \int_{\mathcal{B}_0} \left[\frac{1}{2}\boldsymbol{S}_G - \boldsymbol{B} \odot_3 \frac{\partial \nabla \dot{\tilde{\boldsymbol{C}}}_G}{\partial \dot{\tilde{\boldsymbol{C}}}_G} \right] : \delta_* \dot{\tilde{\boldsymbol{C}}}_G \mathrm{d}V \mathrm{d}t \qquad \int_T \int_{\mathcal{B}_0} \left[\nabla(\dot{\tilde{\boldsymbol{C}}}_G) - \dot{\tilde{\boldsymbol{\Gamma}}} \right] \odot_3 \delta_* \boldsymbol{B} \mathrm{d}V \mathrm{d}t = 0$$

$$\int_T \int_{\mathcal{B}_0} \left[\boldsymbol{B} - \left[\frac{\partial \Psi}{\partial \tilde{\Gamma}} + \tilde{\boldsymbol{B}} \right] \right] \odot_3 \delta_* \dot{\tilde{\boldsymbol{\Gamma}}} \mathrm{d}V \mathrm{d}t = 0$$

The operator $\odot_3$ represents the triple construction of two tensors. Obviously, for the higher-order gradient formulation in $\boldsymbol{C}$, we have less weak forms and thus the tangent becomes substantially simpler.

In the next step, we discretize all quantities over the elements in space and time and transform the integrals to reference elements. For the shape functions in space, N, we use Lagrangian shape functions (see Reference [7]) and approximate the different mixed fields independently. Also we use the same shape functions for the Lagrangian multipliers as for their corresponding mixed fields. We use Lagrangian shape functions in time as well (see Reference [2]), given by

$$M_i(\alpha) = \prod_{\substack{j=1 \\ j \neq i}}^{k+1} \frac{\alpha - \alpha_j}{\alpha_i - \alpha_j}, \; 1 \leq i \leq k+1 \qquad \tilde{M}_i(\alpha) = \prod_{\substack{j=1 \\ j \neq i}}^{k} \frac{\alpha - \alpha_j}{\alpha_i - \alpha_j}, 1 \leq i \leq k \tag{23}$$

The time rate variables and mixed fields ($\boldsymbol{q}$,$\boldsymbol{v}$,$\boldsymbol{p}$,$\tilde{\Theta}$,Θ,η,$\boldsymbol{C}_v$,$\tilde{\boldsymbol{C}}$,$\tilde{\boldsymbol{C}}_A$,$\tilde{J}$,$\tilde{\boldsymbol{\Gamma}}$,$\tilde{\boldsymbol{\Lambda}}$,$\tilde{\boldsymbol{F}}$,$\tilde{\boldsymbol{C}}_G$) are approximated by

$$(\bullet)^{e,h} = \sum_{I=1}^{k+1} \sum_{A=1}^{n_{no}} M_I(\alpha) \mathrm{N}^A(\boldsymbol{\xi}) (\bullet)_I^{eA} \tag{24}$$

and the approximation of Lagrangian multipliers and variation fields ($\boldsymbol{\lambda}_q$,$\boldsymbol{S}$,$\boldsymbol{S}_A$,p,$\boldsymbol{B}$,$\boldsymbol{H}$,$\boldsymbol{P}$,$\boldsymbol{S}_G$,$\delta_* \bullet$) takes the form

$$(\bullet)^{e,h} = \sum_{I=1}^{k} \sum_{A=1}^{n_{no}} \tilde{M}_I \mathrm{N}^A (\bullet)_I^{eA} \tag{25}$$

Here, k is the polynomial degree in time and n_{no} is the number of nodes of the spatial discretization. We approximate each integral with the corresponding Gaussian quadrature rule and condense out the resulting formulation at the element level to a displacement and temperature formulation (see Reference [4]), after eliminating $\boldsymbol{p}$ and η. Note, all mixed fields, except $\boldsymbol{q}$ and Θ, are discontinuous at the boundaries of spatial elements. The internal variable $\boldsymbol{C}_v$ is solved on the element level using the Newton-Raphson method, not at each spatial quadrature point. Since the higher-order gradient formulation results in internal torques, the conservation of angular momentum must be corrected. For the procedure which is described in Reference [11], we obtain for the formulation in $\boldsymbol{F}$

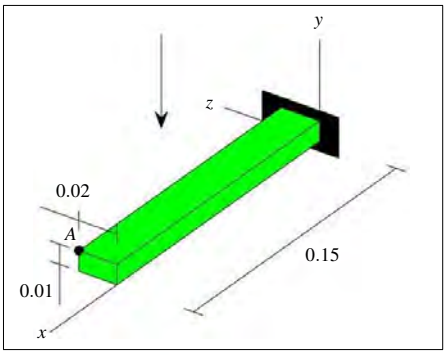
$$\mathcal{J}_{n+1} - \mathcal{J}_n = \int_{t_n}^{t_n+1} \int_{\mathcal{B}_0} \left[\left(\boldsymbol{H} : \frac{\partial \boldsymbol{\Lambda}^{\mathrm{F}}}{\partial \dot{\tilde{\boldsymbol{F}}}} + \boldsymbol{B} \odot_3 \frac{\partial \nabla \dot{\tilde{\boldsymbol{F}}}}{\partial \dot{\tilde{\boldsymbol{F}}}} \right) \times \tilde{\boldsymbol{F}} \right] \mathrm{d}V \mathrm{d}t + \int_{t_n}^{t_n+1} \int_{\partial \mathcal{B}_0} [\boldsymbol{q} \times \boldsymbol{\lambda}_q] \mathrm{d}A \mathrm{d}t + \int_{t_n}^{t_n+1} \int_{\mathcal{B}_0} [\boldsymbol{q} \times \rho_0 \boldsymbol{g}] \mathrm{d}V \mathrm{d}t \tag{26}$$

and for the formulation in $\boldsymbol{C}$

$$\mathcal{J}_{n+1} - \mathcal{J}_n = \int_{t_n}^{t_n+1} \int_{\mathcal{B}_0} \left[\boldsymbol{B} \odot_3 \frac{\partial \nabla \dot{\tilde{\boldsymbol{C}}}_G}{\partial \dot{\tilde{\boldsymbol{C}}}_G} \times \tilde{\boldsymbol{F}} \right] \mathrm{d}V \mathrm{d}t + \int_{t_n}^{t_n+1} \int_{\partial \mathcal{B}_0} [\boldsymbol{q} \times \boldsymbol{\lambda}_q] \mathrm{d}A \mathrm{d}t + \int_{t_n}^{t_n+1} \int_{\mathcal{B}_0} [\boldsymbol{q} \times \rho_0 \boldsymbol{g}] \mathrm{d}V \mathrm{d}t \tag{27}$$

We use our In-House Matlab code fEMcon based on the implementation and ideas shown in Reference [7]. To solve the linear systems of equations we use the Pardiso solver from Reference [8]. For the assembly procedure of all n_{el} finite elements, we use the fast sparse routine shown in Reference [9].

3 NUMERICAL EXAMPLES



$\varepsilon_1 = 0.1e6$	$\beta_F = 1e-6$
$\varepsilon_2 = 100e6$	$c_F^0 = 200$
$k_M = 0.1$	$c_F^1 = 0.0002$
$\beta_M = 1e-6$	$\rho_0 = 1000$
$c_M^0 = 800$	$\Theta_\infty = 300$
$c_M^1 = 0.0008$	$T = 1$
$\varepsilon_3 = 0$	$TOL = 1e-4$
$k_F = 100$	$g = [0\ -2\ 0]$

Figure 1: Geometry, configuration and simulation parameters of the cantilever beam for $n_{el} = 24$.

As numerical example serves a simple cantilever beam which oscillates in a gravitational field. The geometry, configuration and simulation parameters can be found in Figure 1. The corresponding strain energy functions are

$$\Psi_M^{\text{iso}} = \frac{\varepsilon_1}{2}(\text{tr}[\boldsymbol{C}] - 3 - 2\ln(J)) \quad \Psi_M^{\text{vol}} = \frac{\varepsilon_2}{2}\left(\ln(J)^2 + (J-1)^2\right)$$
$$\Psi_X^{\text{cap}} = c_X^0(1 - \Theta_\infty c_X^1)(\Theta - \Theta_\infty - \Theta\ln\frac{\Theta}{\Theta_\infty}) - \frac{1}{2}c_X^0 c_X^1(\Theta - \Theta_\infty)^2$$
$$\Psi_F^{\text{ela}} = \frac{\varepsilon_3}{2}(\text{tr}[\boldsymbol{CM}] - 1)^2 \quad \Psi_{\text{HOG}}^{\text{X}} = l^2\left(I_6^{\text{X}}\right)^2$$

The elastic part of the fiber roving Ψ_F^{ela} can be found in [10] and for the capacitive part the function Ψ_X^{cap} in Reference [2] . We use a quadratic serendipity mesh (20 nodes) with $n_{el} = 24$ and approximate $\tilde{J}$ linear and $\tilde{\boldsymbol{C}}_A$ constant to avoid potential locking effect. We introduce a length scale parameter l^2 with $c = \varepsilon_1 l^2$ for the material parameters of $\Psi_{\text{HOG}}^{\text{X}}$. Futhermore, the strain energy function of the viscous matrix part is given by $\Psi_M^{\text{vis}} = \Psi_M^{\text{iso}}(\boldsymbol{C}\boldsymbol{C}_v^{-1}) + \Psi_M^{\text{vol}}(\boldsymbol{C}\boldsymbol{C}_v^{-1})$.

First, we compare the stiffening behavior of the different higher-order gradient fomulations. In Figure 2 we can see that both formulations stiffen the bending behavior of the beam (HF and HC). However, it can also be done by the $\nabla\boldsymbol{C}$ formulation, although not to the same level (green). By adjusting the material parameters, we obtain a similar behavior here, too (blue). When we look at the angular momentum in Figure 3, we can see it is perfectly preserved for the different formulations. This also shows that the correction of the internal moments as a result of the gradient formulations. In Figure 4 we can see the increasing temperature by the viscous dissipation. As expected, the major

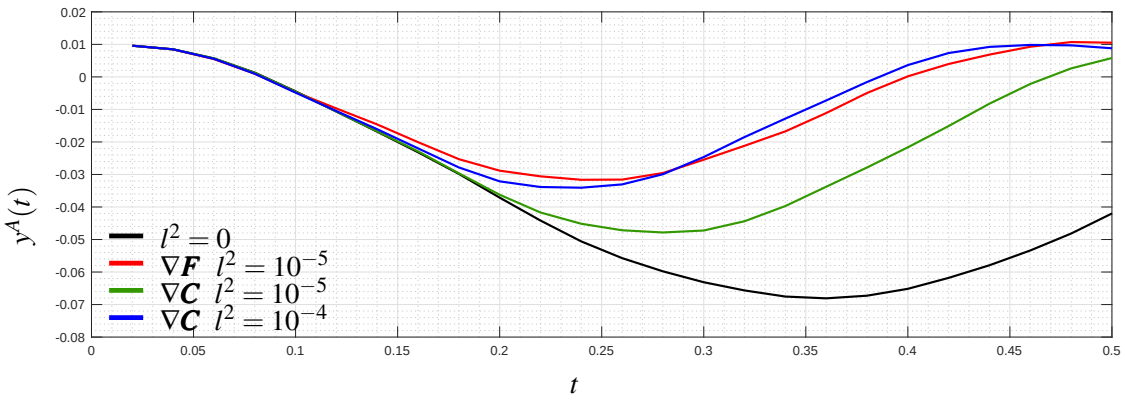


Figure 2: Trajectory of point A for the parameters shown in Figure 1 and the different formulations and $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$.

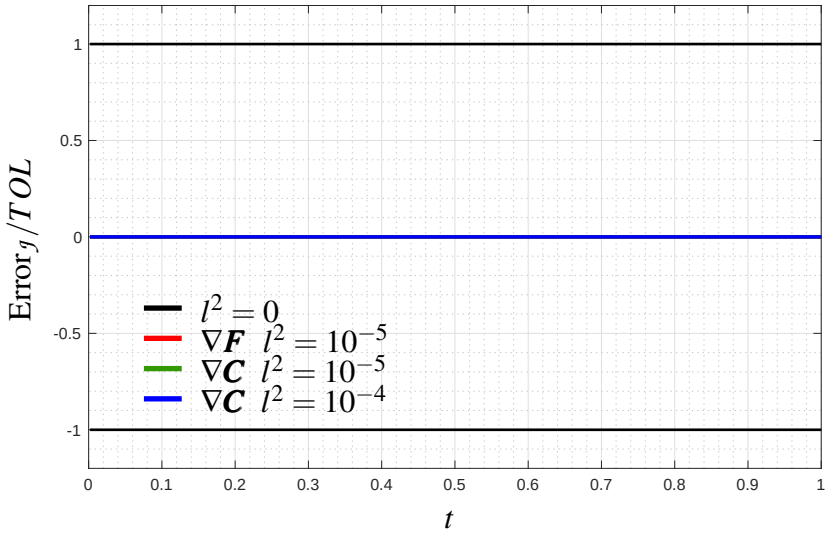


Figure 3: Error of angular momentum $\mathcal{J}$ for the parameters shown in Figure 1 and the different formulations and $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$.

increase in temperature is found at the mounting where the largest deformations occur. Next we check the objectivity of the new superimposed fields of the higher-order gradient formulation on the basis of a free-flying beam. Therefore, we set the initial rotational speed to $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$ and simulate until $T = 10$. In Figure 6-10 we can see that each higher-order gradient formulation and length scale parameter conserve the total energy. For the high l^2, a slightly higher energy error is observed, but this is also within the tolerance. For example, this can be explained by the fact that although the higher stiffness, we keep the time step size constant. In Figure 5 we show the current configuration and v. Mises equivalent stress σ_{VM} for $t = 10$. As expected, the beam is deformed by the rotation and shows the larger stresses at larger deformations.

4 CONCLUSIONS

We have shown that it is possible to formulate a higher-order gradient material formulation in terms of the right Cauchy-Green tensor. This is a remarkable result, because this formulation requires considerably less numerical effort and we can formulate the superimposed field directly in terms of $\nabla \boldsymbol{C}$ and thus achieve a roving direction independence. Also, both formulations work in a thermo-visoelastic context. And we have also shown that the higher-order energy-momentum time integrators

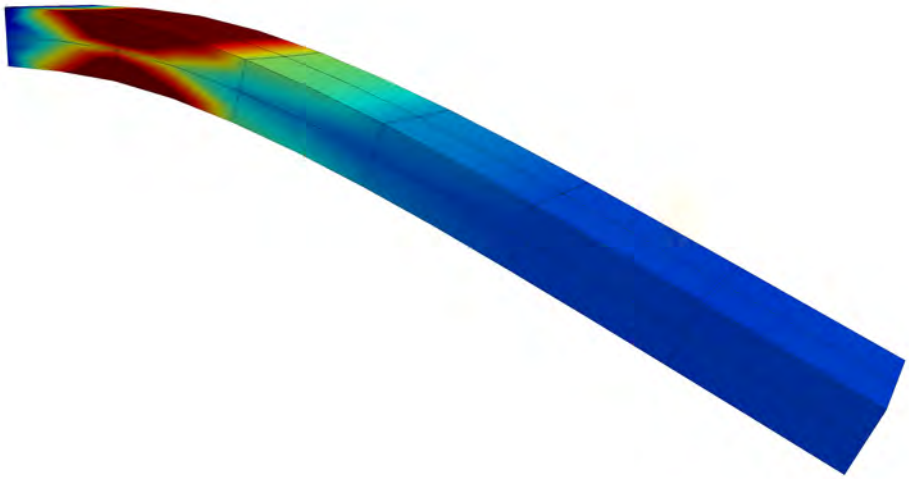


Figure 4: Configuration and temperature distribution Θ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $t = 0.24$, $\nabla \boldsymbol{C}$ and $l^2 = 10^{-4}$.

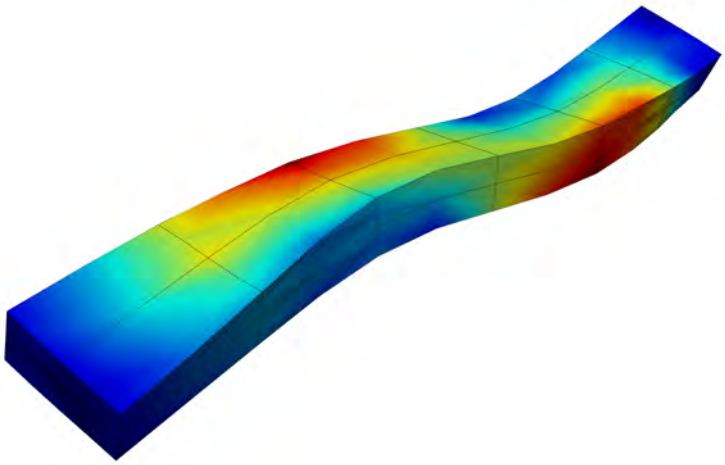


Figure 5: Configuration and v. Mises equivalent stress σ_{VM} for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $t = 10$, $\nabla \boldsymbol{C}$ and $l^2 = 10^{-4}$.

conserve energy in all cases. In the next step, we want to investigate other material formulation and will look on potential locking effects.

Acknowledgments

The authors thank the 'Deutsche Forschungsgemeinschaft (DFG)' for the financial support of this work under the grant GR3297/4-2 and GR3297/6-1 as well as Matthias Bartelt (GR 3297/2-2) for providing the programming basis for the current implementation.

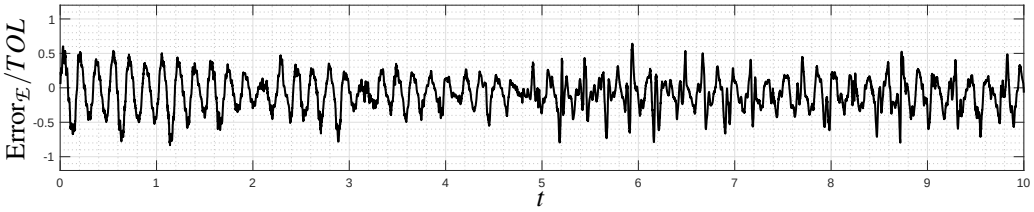


Figure 6: Error of energy $\mathcal{E}$ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $T = 10$ and $l^2 = 0$.

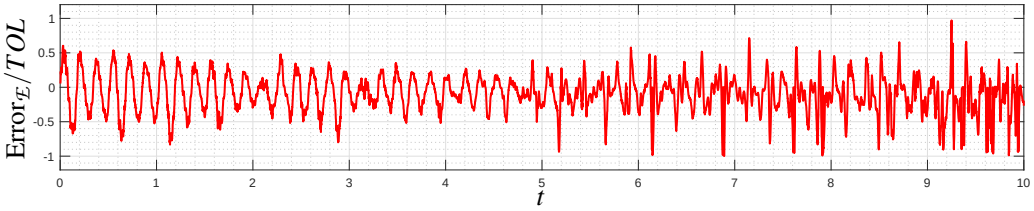


Figure 7: Error of energy $\mathcal{E}$ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $T = 10$, $\nabla \boldsymbol{F}$ (HF) and $l^2 = 10^{-5}$.

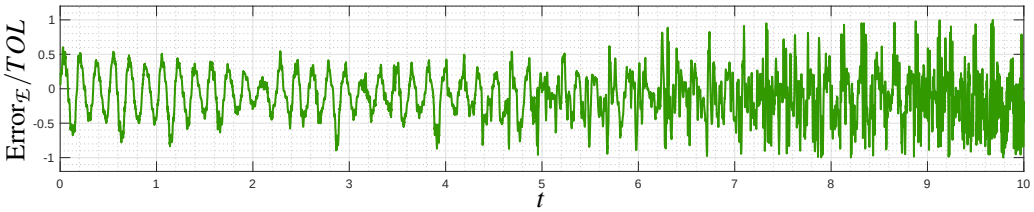


Figure 8: Error of energy $\mathcal{E}$ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $T = 10$, $\nabla \boldsymbol{F}$ (HF) and $l^2 = 10^{-4}$.

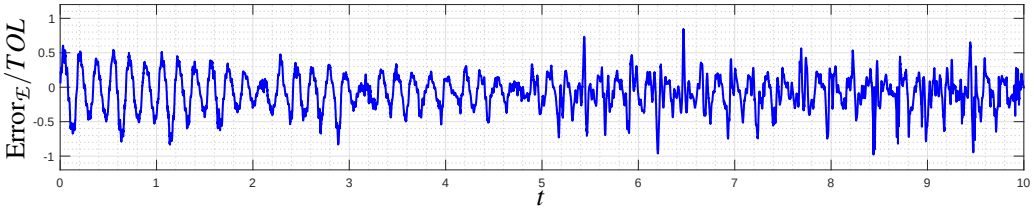


Figure 9: Error of energy $\mathcal{E}$ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $T = 10$, $\nabla \boldsymbol{C}$ (HC) and $l^2 = 10^{-4}$.

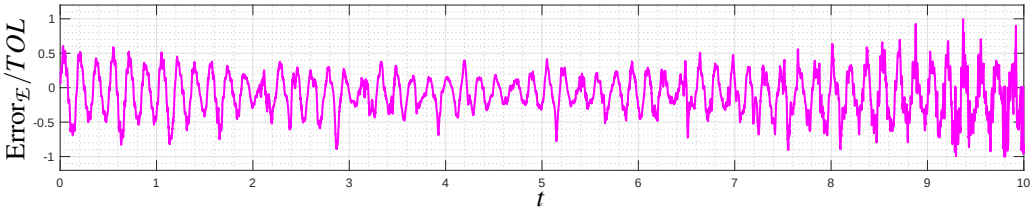


Figure 10: Error of energy $\mathcal{E}$ for the parameters shown in Figure 1, $(\boldsymbol{a}_0)^T = [1\ 0\ 0]$, $\boldsymbol{\omega}^T = [2\pi\ 2\pi\ 2\pi]$, $T = 10$, $\nabla \boldsymbol{C}$ (HC) and $l^2 = 10^{-3}$.

REFERENCES

[1] Asmanoglo, T., Menzel., A. (2017) A multi-field finite element approach for the modelling of fibre-reinforced composites with fibre-bending stiffness. *Comput. Methods Appl. Mech. Engrg.*, 317:1037–1067.

[2] Groß, M., Dietzsch, J., and Bartelt, M. (2018). Variational-based higher-order accurate energy–momentum schemes for thermo-viscoelastic fiber-reinforced continua. *Comput. Methods Appl. Mech. Engrg.*, 336, 353–418. https://doi.org/10.1016/j.cma.2018.03.019

[3] Simo, J. C., Taylor, R. L., and Pister, K. S. (1985). Variational and projection methods for the volume constraint in finite deformation elasto-plasticity. *Comput. Methods Appl. Mech. Engrg.*, 51(1–3), 177–208. https://doi.org/10.1016/0045-7825(85)90033-7

[4] Schröder, J., Viebahn, V., Wriggers, and P., Balzani, D. (2016). A novel mixed finite element for finite anisotropic elasticity; the SKA-element Simplified Kinematics for Anisotropy. *Comput. Methods Appl. Mech. Engrg.*, 310:475–494.

[5] J. Dietzsch and M. Groß, Mixed Finite Element Formulations for Polyconvex Anisotropic Material Formulations in WCCM-ECCOMAS2020.

[6] Ferretti, M., Madeo, A., dell'Isola, F., & Boisse, P. (2014). Modeling the onset of shear boundary layers in fibrous composite reinforcements by second-gradient theory. Zeitschrift Fur Angewandte Mathematik und Physik, 65(3), 587–612.

[7] Bartelt, M., Dietzsch, J., and Groß, M. (2018). Efficient implementation of energy conservation for higher order finite elements with variational integrators. *Math. Comput. Simulat.*, 150, 83–121. https://doi.org/10.1016/j.matcom.2018.03.002

[8] Alappat, C., Basermann, A., Bishop, A. R., Fehske, H., Hager, G., Schenk, O., Thies, J., and Wellein, G. (2020). A Recursive Algebraic Coloring Technique for Hardware-efficient Symmetric Sparse Matrix-vector Multiplication. *ACM Transactions on Parallel Computing*, 7(3). https://doi.org/10.1145/3399732

[9] Engblom, S., and Lukarski, D. (2016). Fast Matlab compatible sparse assembly on multicore computers. *Parallel Computing*, 56, 1–17. https://doi.org/10.1016/j.parco.2016.04.001

[10] Dal, H., Gültekin, O., Aksu Denli, F., and Holzapfel, G. A. (2017). Phase-Field Models for the Failure of Anisotropic Continua. *PAMM*, 17(1). https://doi.org/10.1002/pamm.201710027

[11] Groß, M., Dietzsch, J., and Röbiger, C. (2020). Non-isothermal energy–momentum time integrations with drilling degrees of freedom of composites with viscoelastic fiber bundles and curvature–twist stiffness. Computer Methods in Applied Mechanics and Engineering, 365, 112973. https://doi.org/10.1016/j.cma.2020.112973

Intrinsically Selective Mass Scaling with Hierarchic Structural Element Formulations

B. Oesterle*, J. Trippmacher†, A. Tkachuk‡, and M. Bischoff*

* University of Stuttgart, Institute for Structural Mechanics, Pfaffenwaldring 7, 70569 Stuttgart, Germany, e-mail: {oesterle,bischoff}@ibb.uni-stuttgart.de

† Ed. Züblin AG, Albstadtweg 3, 70567 Stuttgart, Germany, e-mail: jan.trippmacher@zueblin.de

‡ Department of Engineering and Physics, Karlstad University, 658 88 Karlstad, Sweden, e-mail: anton.tkachuk@kau.se

Key words: selective mass scaling, hierarchic formulations, frequency spectra, shear locking, isogeometric analysis

Abstract: *Hierarchic shear deformable structural element formulations possess the advantage of being intrinsically free from transverse shear locking, that is they avoid transverse shear locking a priori through reparametrization of the kinematic variables. This reparametrization results in shear deformable beam, plate and shell formulations with distinct transverse shear degrees of freedom. The basic idea of selective mass scaling within explicit dynamic analyses is to scale down the highest frequencies in order to increase the critical time step size, while keeping the low frequency modes mostly unaffected. In most concepts, this comes at the cost of non-diagonal mass matrices. In this contribution, we present first investigations on selective mass scaling for hierarchic formulations. Since hierarchic structural formulations possess distinct transverse shear degrees of freedom, they offer the intrinsic ability for selective scaling of the high frequency shear modes, while keeping the bending dominated low frequency modes mostly unaffected. The proposed instrinsically selective mass scaling concept achieves high accuracy, which is typical for selective mass scaling schemes, but in contrast to existing concepts it retains the simplicity of a conventianl mass scaling method and preserves the diagonal structure of a lumped mass matrix. As model problem, we study frequency spectra of different isogeometric Timoshenko beam formulations for a simply supported beam. We discuss the effects of transverse shear parametrization, locking and mass lumping on the accuracy of results.*

1 INTRODUCTION

Finite element solution schemes in the context of structural dynamics can be classified as explicit and implicit methods. Explicit algorithms are particularly popular for highly non-linear and non-smooth problems, since they do not require any iterative solution of the balance equations on the global level. In specific applications, like car crash or deep drawing simulations, they may be more robust than implicit methods. But due to the conditional stability of explicit methods, the admissible time step size is limited. The so-called critical time step Δt_{crit} crucially depends on the highest frequency ω_{max} of the discrete system

$$\Delta t_{\text{crit}} = \frac{2}{\omega_{\text{max}}}. \tag{1}$$

Several approaches to reduce computational cost are available and it is common to use a combination of various approaches simultaneously. First, locking-free and accurate finite element formulations can be used to achieve satisfactory results for coarse meshes, since mesh refinement indirectly increases computational cost for time integration. Second, adaptive mesh control on the basis of error estimators can be used, which is a standard approach for deep drawing simulations. Third, different time steps may be used in areas with different mesh density, which

https://doi.org/10.4995/YIC2021.2021.12418

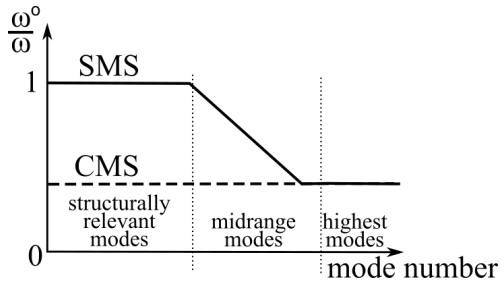


Figure 1: Schematic diagram of the ratio of the scaled eigenfrequencies to the original frequencies of a system for CMS and SMS.

is known as subcycling or asynchronous time integration. Fourth, reduced order modeling may increase efficiency. Fifth, mass scaling may increase the stable critical time step size and thus reduces the number of time steps and the total computational cost.

In the present contribution, we focus on innovative versions and a straight-forward combination of the first and the fifth approach. The outline is as follows. A short overview on established mass scaling concepts is presented in Section 2, before the novel, intrinsically selective mass scaling concept for hierarchic formulations is presented in Section 3. In Section 4, we study frequency spectra of different isogeometric Timoshenko beam formulations for a simply supported beam. We discuss effects of transverse shear parametrization, locking and mass lumping on the accuracy of results. Section 5 concludes our findings and provides an outlook on future work.

2 MASS SCALING

In the research field of mass scaling, it can be distinguished between conventional mass scaling (CMS) and selective mass scaling (SMS). All mass scaling techniques add artificial inertia to the global mass matrix. CMS adds inertia only on the diagonal entries, thus preserving the diagonal structure of the lumped mass matrix (LMM). When applied to translational inertia, as in case of continuum or solid shell element formulations with solely displacement degrees of freedom, translational inertia of the structure is increased. Uniform mass scaling for all elements significantly modifies the linear momentum of the entire structure and thus also affects the lowest, structurally relevant modes, see Figure 1. Therefore, application of CMS is usually limited to a small number of short and stiff elements that limit the critical time step size Δt_{crit}.

In the context of solid finite elements, the basic idea of selective mass scaling (SMS) is to add artificial contributions to both diagonal and off-diagonal entries of the mass matrix in order to preserve translational inertia. This results in a significant reduction of the highest eigenfrequencies, which are often irrelevant for structural response but limit the critical time step size. Manipulation of the low frequencies, which are essential for structural response, is reduced to a minimum. The qualitative picture of the desired ratio between scaled eigenfrequencies ω° and the unscaled eigenfrequencies ω, typically obtained with a LMM, is shown in Figure 1, comparing results obtained with SMS and CMS. The concept of SMS can provide a very good compromise between accuracy and critical time step size, but comes at the cost of non-diagonal mass matrices and thus the need to solve a linear system of equations at each time step. There exist a number of algebraic and variational SMS schemes, see for instance [1, 2, 3], but all of them are designed for continuum or solid shell elements with displacement degrees of freedom.

In case of structural element formulations with rotational degrees of freedom, the aforementioned SMS schemes are not extendable to rotational inertia in a straightforward manner. In fact, naively extending the SMS concept by Olovsson et al. [1] to the rotational part of the mass matrix is not capable of reducing the highest frequency and thus no benefit can be achieved

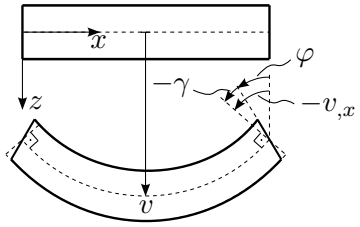


Figure 2: Geometrically linear kinematics of a planar, straight Timoshenko beam.

w.r.t. to increasing the critical time step size. But since in shear deformable structural element formulations the translational and the rotational part of the mass matrix can be computed separately, CMS of the rotational part may lead to some sort of semi-selective mass scaling. CMS of the rotational inertia is used in commercial explicit codes in dayly practice, as can be seen for instance in [4] in the context of isogeometric shell analysis in LS-DYNA. However, since the entire rotational inertia is increased by this concept, also rigid body rotations and the bending dominated modes are severely affected. In case of structural element formulations based on so-called first order shear defomation theory, that is Timoshenko or Reissner-Mindlin theory, the highest frequencies of the discretized system are typically related to transverse shear modes. The direct and isolated access of these high frequency shear modes by a simple and efficient concept is described in the following section.

3 HIERARCHIC REPARAMETRIZATION AND INTRINSICALLY SELECTIVE MASS SCALING

Hierarchic shear deformable structural element formulations are intrinsically free from transverse shear locking, that is they avoid transverse shear locking a priori through reparametrization of the kinematic variables. Although there already exist shear deformable beam, plate and shell formulations, we restrict ourselves to planar straight shear deformable Timoshenko beams with linearized kinematics within this study. In this chapter, different parametrizations of the Timoshenko beam model are briefly summarized and discussed. For further details we refer to [5] or [6]. Figure 2 shows the kinematics of a planar, straight Timoshenko beam, where v describes the mid-line displacement in z-direction, φ and γ are the total and the shear rotation of the beam's cross section. As a general rule in subsequent derivations, equal order interpolation of all involved primary fields is assumed.

The standard formulation of the Timoshenko model (T-st) introduces the total vertical displacement v and the total cross-sectional rotation φ as primary variables. The shear rotation γ and the curvature κ can be expressed as

$$\gamma = v_{,x} + \varphi \qquad \text{and} \qquad \kappa = \varphi_{,x}, \tag{2}$$

where $(\bullet)_{,x} = \frac{\mathrm{d}(\bullet)}{\mathrm{d}x}$ describes the derivative with respect to the spatial x-coordinate. When the discrete primary parameters v^h and φ^h are discretized via any equal order interpolation, pure bending with $\gamma^h = 0$, cannot be fulfilled. The imbalance of the shape functions $v^h_{,x}$ and φ^h leads to the well-known phenomenon of transverse shear locking.

Following the idea from [7], [8] and [9], the shear rotation γ may be introduced directly as primary variable instead of the total rotation φ. Accordingly, φ has to be expressed in terms of the two primary variables, that is

$$\varphi = -v_{,x} + \gamma. \tag{3}$$

formulation	displacement	total rotation	shear rotation	curvature
T-st (v-φ)	v	$\varphi = \varphi$	$\gamma = v_{,x} + \varphi$	$\kappa = \varphi_{,x}$
T-hr (v-γ)	v	$\varphi = -v_{,x} + \gamma$	$\gamma = \gamma$	$\kappa = -v_{,xx} + \gamma_{,x}$
T-hd (v-v_s)	v	$\varphi = -v_{,x} + v_{\mathrm{s},x}$	$\gamma = v_{\mathrm{s},x}$	$\kappa = -v_{,xx} + v_{\mathrm{s},xx}$

Table 1: Comparison of different Timoshenko beam formulations with different parametrizations of the kinematic variables.

The combination of Equations (2) and (3) yields the modified kinematics

$$\gamma = \gamma \qquad \text{and} \qquad \kappa = \varphi_{,x} = -v_{,xx} + \gamma_{,x}. \tag{4}$$

From Equation (4), the hierarchic structure of the kinematics is visible, since the formulation includes the Euler-Bernoulli beam model for vanishing shear strain γ. Since γ represents the shear rotation, which is superimposed on the rotated cross section according to the Euler-Bernoulli model, this formulation is denoted as Timoshenko beam formulation with hierarchic rotation (T-hr).

An alternative reparametrization is introduced in Timoshenko beam formulations with hierarchic displacements (T-hd). In contrast to the previously presented hierarchic split of the total rotation, the basic idea is the hierarchic split of the displacements into parts resulting from bending and shear, i. e.

$$v = v_\mathrm{b} + v_\mathrm{s}. \tag{5}$$

This idea is not new, in fact closed form solutions for static and dynamic problems can be found for instance in [10], finite element formulations to solve dynamic problems are presented in [11], among others. Based on Equation (5), a single-variable isogeometric formulation for shear deformable beams is presented in [12].

The following derivations are based on the notation of [5] and a practical modification of the initial concept presented in [13]. Starting from Equation (5), the reparametrized rotation can be written as

$$\varphi = -v_{,x} + \gamma = -(v_\mathrm{b} + v_\mathrm{s})_{,x} + v_{\mathrm{s},x} = -v_{\mathrm{b},x}. \tag{6}$$

In general, three different T-hd formulations can be derived by using two out of three displacment parameters v, v_b and v_s. The present study is restricted to the parametrization utilizing v and v_s as primary paramters, which is probably the most practical one, see also [13]. The Timoshenko beam formulations used herein are summarized in Table 1. For detailed interpretations and result w.r.t. locking in the context of static analyses, we refer to [5, 13, 6]. Some remarkable features of hierarchic formulations are:

1. The variational index is equal to two, thus consistency requires at least quadratic C^1-continuous shape functions. Thus, the smooth discretization schemes of isogeometric analysis [14] are well-suited for discretization.

2. Both T-hr and T-hd are free from transverse shear locking as the thin limit constraint $\gamma = 0$ can be trivially satisfied by the related degree of freedom being zero. T-hd has fully balanced kinematics, whereas in T-hr, the imbalance is shifted from γ to κ.

3. In the case of beams, shear is completely decoupled from bending.

4. Both formulations possess distinct shear degrees of freedom.

Starting from this point of departure, we introduce the novel idea of a selective and effective mass scaling strategy in the context of hierarchic structural element formulations. As introduced in Section 2, the idea of SMS adresses the effective reduction of the highest frequencies, while keeping the more relevant low frequency modes mostly unaffected. In case of shear deformable structural element formulations based on Timoshenko or Reissner-Mindlin theory, the highest frequencies of the discretized system are typically related to transverse shear modes. A deeper look at features 3 and 4 leads to the following hypotheses for a novel mass scaling strategy in the context of hierarchic formulations:

- The shear frequencies and corresponding modes can be directly accessed by the distinct shear degrees of freedom.
- Decoupling bending and shear facilitates selective scaling of high shear frequencies.
- Both aspects lead to a mass scaling strategy being as effective as a SMS strategy, while retaining the simplicity of a CMS scheme.

Starting point for the following derivations is d'Alembert's principle, specified for Timoshenko beam theory

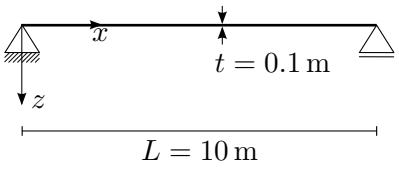
$$\delta W = \int_0^L (\delta\gamma GA\gamma + \delta\kappa EI\kappa)\ \mathrm{d}x + \int_0^L (\delta v\rho A\ddot{v} + \delta\varphi\rho I\ddot{\varphi})\ \mathrm{d}x - \delta W^{\mathrm{ext}} \quad = 0, \tag{7}$$

where E und G denote Young's modulus and shear modulus. The cross-sectional area, the second moment of inertia and the density are denoted by A, I und ρ, respectively. The consistent mass matrix (CMM) $\mathbf{M}_{\mathrm{C}}$ is computed as

$$\mathbf{M}_{\mathrm{C}} = \int_0^L \mathbf{N}^{\mathrm{T}} \begin{bmatrix} \rho A & 0 \\ 0 & \rho I \end{bmatrix} \mathbf{N}\, \mathrm{d}x, \tag{8}$$

with $\mathbf{N}$ being the matrix of shape functions w.r.t. displacement v and total rotation φ. The matrix $\mathbf{N}$ depends on the kinematic description of the chosen beam formulation, that is T-st, T-hr or T-hd, summarized in Table 1. Since for explicit dynamic simulations LMMs are desirable for efficiency reasons, they are of high interest. Cottrell et al. [15] showed for isogeometric analysis that a LMM obtained by row-sum lumping is only second order accurate, independent of the polynomial order. Nevertheless, row-sum lumping for isogeometric elements is still the state of the art in commercial software like LS-DYNA, see for instance [4]. Since the present contribution focuses on mass scaling and not mass lumping, we further consider traditional row-sum lumping and subsequent CMS of the rotational inertia by a scaling parameter α. In case of T-st the rotational part is related to the total rotation φ of the beam's cross section. Thus, any value $\alpha > 1$ for the scaling parameter leads to artificial rotational inertia and thus angular momentum is not preserved. How significantly the bending modes of the T-st formulation are influenced by $\alpha > 1$ is studied in the next section. For the T-hr formulation the rotational entries in the LMM are associated with the shear rotation γ, representing only the shear part of the total rotation φ, but not the rigid body part. Thus, it is expected that a scaling parameter $\alpha > 1$ mainly influences the shear modes, while keeping the bending modes significantly less affected.

For the second hierarchic Timoshenko beam formulation T-hd standard row-sum lumping leads to singular LMMs. In fact, the rotational entries vanish, since the parametrization of the rotation φ is fully balanced, as can be seen in Table 1. In the lumping process each contribution from v is canceled by the corresponding contribution from v_{s}. This issue is not addressed herein, but deserves further consideration in future work, for instance by a detailed study of alternative lumping schemes, as presented for instance in [16] or [17]. The development of accurate mass lumping schemes in general is still an open research topic in the context of isogeometric analysis.



$E = 210\,\mathrm{GPa}$
$\nu = 0.3$
$\rho = 7800\,\mathrm{kg/m^3}$

Figure 3: Simply supported beam, problem setup.

4 NUMERICAL EXAMPLE

As model problem, we study frequency spectra of various isogeometric Timoshenko beam formulations for the case of a simply supported beam, as shown in Figure 3. In all cases, the beam is discretized by 50 elements using quadratic, C^1-continuous B-splines, constructed from an open knot vector. The studied combinations of beam formulations and technologies to tackle transverse shear locking are listed as follows:

- **T-st**: standard Timoshenko beam formulation with displacment v and total rotation φ as primary variables.
- **T-st-low**: as T-st, but the shape functions for φ are one order lower to overcome transverse shear locking, as presented in [18] and [19] for isogeometric elements.
- **T-st-SRI**: As T-st, with selective reduced integration for the shear strain contributions to the stiffness matrix in order to remove transverse shear locking, as shown in [20].
- **T-hr**: Hierarchic Timoshenko beam formulation with hierarchic rotation with displacement v and shear rotation γ as primary fields.
- **T-hr-low**: As T-hr, but the shape functions for γ are one order lower to overcome the imbalance in the kinematic equations, which can be seen in Table 1.
- **T-hd**: Hierarchic Timoshenko beam formulation with hierarchic displacement, with total displacement v and shear displacement v_s as primary fields.

First, we study the accuracy of frequency spectra obtained by CMM and LMM with respect to analytical solutions from Cazzani et al. [21]. Figure 4 shows the results for the three T-st formulations, where in the top row of diagrams the frequency spectra are plotted and the bottom row of diagrams displays the ratio of numerical to analytical frequencies. As expected, the CMM provides more accurate results than the LMM in all cases. The T-st formulation suffers from shear locking, as can be seen from the very high frequencies in the bending dominated branch. The selective reduced integrated version T-st-SRI is locking-free, but exhibits low accuracy in the bending dominated branch of the spectrum. The most accurate results of all three standard formulations are obtained by T-st-low.

Figure 5 shows the results for the hierarchic formulations. The imbalance in the kinematic equations of the T-hr formulation leads to slightly increased frequencies in the right part of the bending dominated branch. This is surprising, since this element formulation has proven to be free from transverse shear locking for the static case. But the shifted imbalance from γ to κ theoretically leads to a locking effect in the very thick regime, although we have not observed any relevance of such a locking effect in static analyses. For this problem setup with this relatively fine discretization, the actual element slenderness ratio is $L_\mathrm{e}/t = 2$. For more slender elements, this effect vanishes also in the frequency spectra. Theoretically, T-hr-low and T-hd should achieve the same accuracy, which is clearly visible for the spectra obtained by a CMM. But, as already mentioned, the expression $\varphi = -v_{,x} + v_{\mathrm{s},x}$ is perfectly balanced in case of

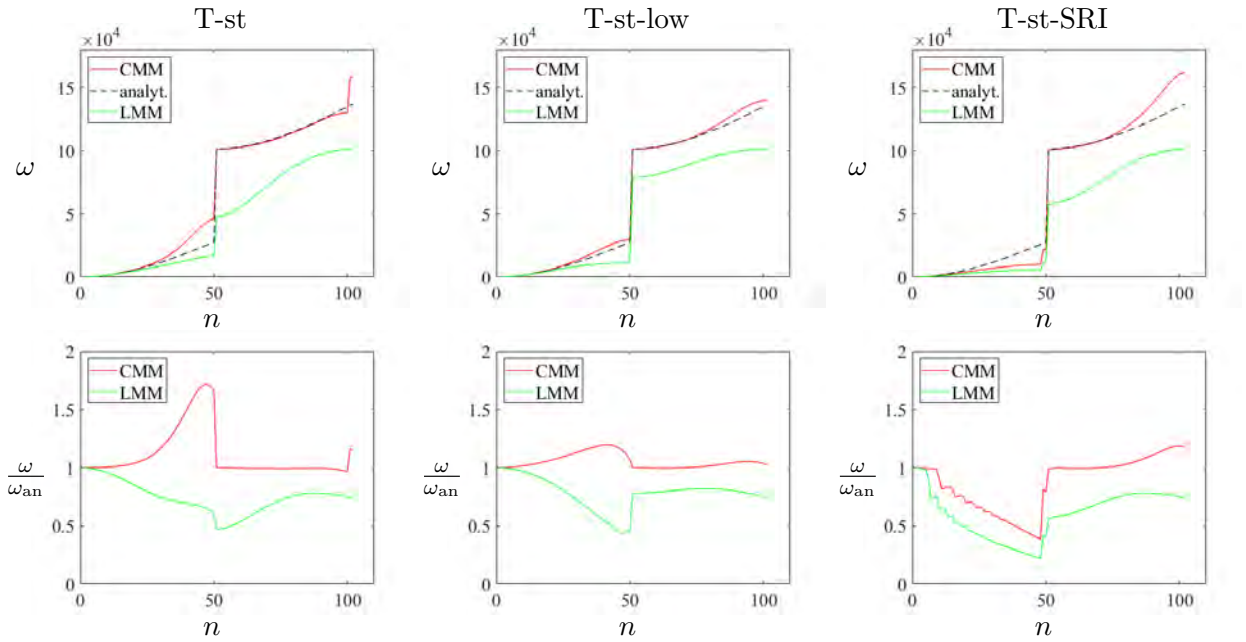


Figure 4: Simply supported beam: Standard formulations T-st, T-st-low and T-st-SRI, top: discrete spectra, bottom: ratio of numerical frequencies to analytical frequencies.

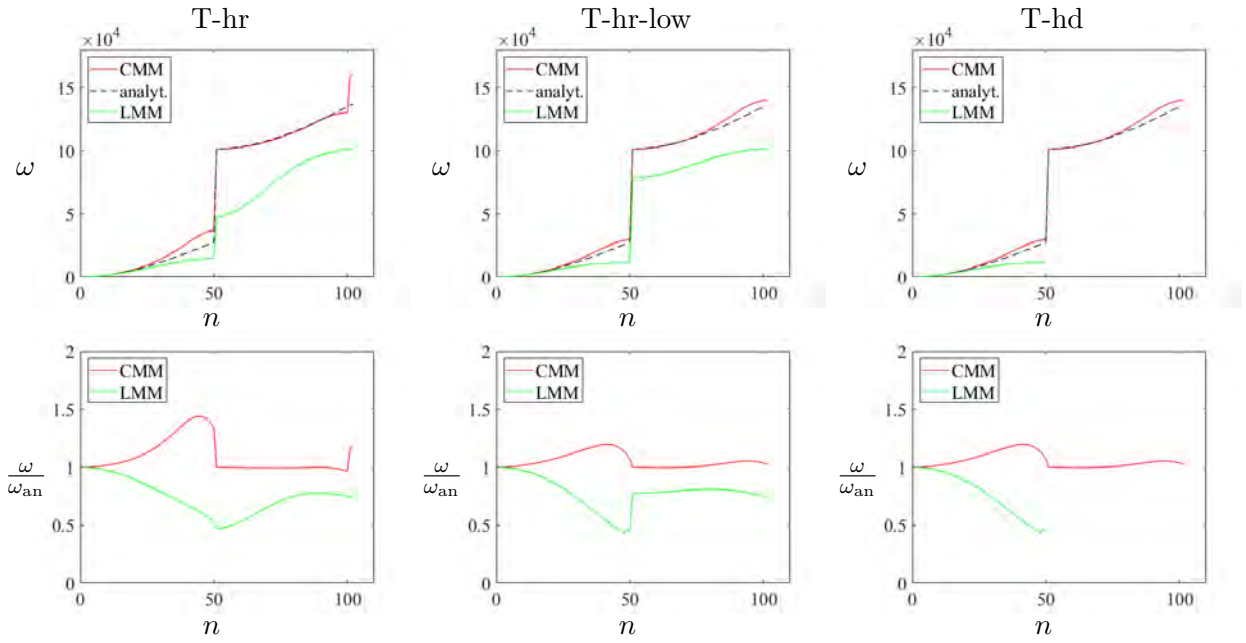


Figure 5: Simply supported beam: Hierarchic formulations T-hr, T-hr-low and T-hd, top: discrete spectra, bottom: ratio of numerical frequencies to analytical frequencies.

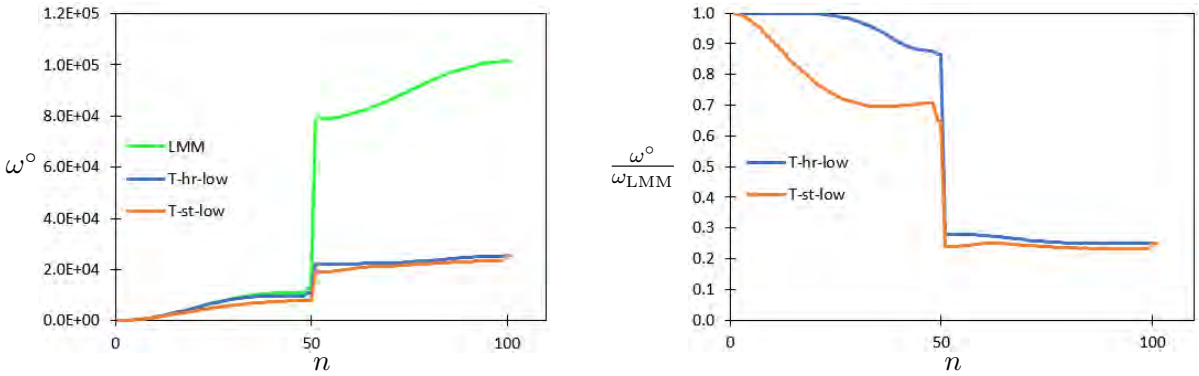


Figure 6: Simply supported beam: T-hr-low and T-st-low, left: scaled spectra in comparison to LMM (unscaled), right: ratio of scaled frequencies to unscaled frequencies for LMM.

T-hd, which leads to zero rotational mass in the LMM obtained by standard row-sum lumping. Alternative lumping strategies are subject of recent investigations, but not discussed herein. The results for T-hr-low are the most accurate ones from Figure 5 and they are practically identical to the results obtained by T-st-low. In fact, the maximum relative difference between both spectra obtained by LMM is 1.51%, while the highest frequency of $\omega_{\max} = 101435.07$ is identical.

Due to practically identical spectra, the two formulations T-st-low and T-hr-low are the optimal starting point for studying the accuracy of mass scaling and the corresponding accuracy of the scaled eigenfrequencies ω°. As stated in the previous section, in case of T-st or T-st-low the rotational part is related to the total rotation φ. Thus, any value for the scaling parameter $\alpha > 1$ leads to artificial rotational inertia and an influence of the bending dominated modes is expected. In contrast to T-st-low, for the T-hr-low formulation the rotational entries in the LMM are solely associated with the shear rotation γ.

Figure 6 compares the accuracy of the scaled eigenfrequencies ω° on the basis of LMMs and a simple CMS of the rotational masses by a scaling parameter α, as explained in Section 3. For both formulations α is chosen such that the maximum eigenfrequncy $\omega_{\max} = 101435.07$ is reduced by 75%. For T-st-low a scaling parameter of $\alpha = 28.52$ is needed to reduce the maximum frequency to $\omega^\circ_{\max} = 25358.84$, while the bending dominated frequencies are affected by up to 35% w.r.t. to the unscaled solution obtained by a LMM. In case of T-hr-low $\alpha = 16.0$ yields $\omega^\circ_{\max} = 25358.77$ and the highest deviation from LMM in the bending branch is only 14%. The largest deviation in the first quarter of the spectrum is only 1%. This highlights the significant influence of hierarchic parametrization on the selective scalability of the shear dominated frequencies and shows the high potential of the proposed *instrinsically selective mass scaling*, namely possessing high accuracy while preserving the diagonal structure of a lumped mass matrix.

5 CONCLUSIONS AND OUTLOOK

The concept of *instrinsically selective mass scaling* (ISMS) has been proposed. The key idea is to make use of the distinct shear degrees of freedom of shear deformable, hierarchic structural element formulations in the context of an efficient and effective mass scaling strategy. As a model problem, we studied frequency spectra of various isogeometric Timoshenko beam formulations. The results indicate that the proposed ISMS scheme is able to effectively reduce the highest shear frequencies, while keeping the low bending dominated frequencies mostly unaffected. This property is typical for SMS schemes, but, in contrast to standard SMS schemes,

the ISMS scheme proposed herein is as simple as a CMS scheme and preserves the diagonal structure of LMMs.

Further developments address the extension to other smooth discretization schemes, to hierarchic shell formulations and to nonlinear transient analyses. In addition, the studies on optimal scaling parameters and developments of time step estimates and more accurate mass lumping schemes are of high interest.

Acknowledgements

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) under grant OE 728/1-1. This support is gratefully acknowledged.

REFERENCES

[1] Olovsson, L., Simonsson, K. and Unosson, M. Selective mass scaling for explicit finite element analyses. *Int J Numer Methods Eng.*, Vol. **63**, pp. 1436–1445, (2005). https://doi.org/10.1002/nme.1293

[2] Cocchetti, G., Pagani, M. and Perego, U. Selective mass scaling and critical time-step estimate for explicit dynamics analyses with solid-shell elements. *Computers and Structures*, Vol. **27**, pp. 39-52, 2013. https://doi.org/10.1016/j.compstruc.2012.10.021

[3] Tkachuk, A. and Bischoff, M. Variational methods for selective mass scaling. *Comput Mech* Vol. **52**, pp. 563–570, (2013). https://doi.org/10.1007/s00466-013-0832-0

[4] Hartmann, S., Benson, D. J. Mass scaling and stable time step estimates for isogeometric analysis. *Int J Numer Methods Eng.* **102**, 671–687, (2015). https://doi.org/10.1002/nme.4719

[5] Oesterle, B., Ramm, E. and Bischoff, M. A shear deformable, rotation-free isogeometric shell formulation. *Comput. Methods Appl. Mech. Engrg.*, Vol. **307**, pp. 235–255, (2016). https://doi.org/10.1016/j.cma.2016.04.015

[6] Oesterle, B., Bieber, S., Sachse, R., Ramm, E., and Bischoff, M. Intrinsically locking-free formulations for isogeometric beam, plate and shell analysis. *Proc. Appl. Math. Mech.*, Vol. **18**, e201800399. https://doi.org/10.1002/pamm.201800399

[7] Başar, Y., Krätzig, W.B. Mechanik der Flächentragwerke, Vieweg, 1985.

[8] Long, Q., Bornemann, P.B. and Çirak, F. Shear-flexible subdivision shells. *Int J Numer Methods Eng.*, Vol. **90**, pp. 1549–1577, (2012). https://doi.org/10.1002/nme.3368

[9] Echter, R., Oesterle, B. and Bischoff, M. A hierarchic family of isogeometric shell finite elements. *Comput. Methods Appl. Mech. Engrg.*, Vol. **254**, 170–180, (2013). https://doi.org/10.1016/j.cma.2012.10.018

[10] Anderson, R. A. Transient response of uniform beams, Ph.d. thesis, California Institute of Technology, (1953).

[11] Marguerre, K., Wölfel, H. Mechanics of vibration, Sijthoff & Noordhoff [International Publishers], Alphen aan den Rijn, (1979).

[12] Kiendl, J., Auricchio, F., Hughes, T. J. R. and Reali, A. Single-variable formulations and isogeometric discretizations for shear deformable beams, *Comput. Methods Appl. Mech. Engrg.*, Vol. **284**, 988-1004, (2015). https://doi:10.1016/j.cma.2014.11.011.

[13] Oesterle, B., Sachse, R. Ramm, E. and Bischoff, M. Hierarchic isogeometric large rotation shell elements including linearized transverse shear parametrization, *Comput. Methods Appl. Mech. Engrg.*, Vol. **321**, pp. 383-405, (2017). https://doi:10.1016/j.cma.2017.03.031.

[14] Hughes, T.J.R., Cottrell, J.A. and Bazilevs, Y. Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. *Comput. Methods Appl. Mech. Engrg.*, Vol. **194**, 4135–4195, (2005). https://doi.org/10.1016/j.cma.2004.10.008

[15] Cottrell, J.A., Reali, A., Bazilevs, Y. and Hughes, T.J.R. Isogeometric analysis of structural vibrations. *Comput. Methods Appl. Mech. Engrg.*, Vol. **195**, 5257–5296, (2006). https://doi.org/10.1016/j.cma.2005.09.027

[16] Hinton, E., Rock, T., Zienkiewicz, O.C. A note on mass lumping and related processes in the finite element method. *Earthquake Engineering & Structural Dynamics*, Vol. **4**, 245–249, (1976). https://doi.org/10.1002/eqe.4290040305

[17] Anitescu, C., Nguyen, C., Rabczuk, T. and Zhuang, X. Isogeometric analysis for explicit elastodynamics using a dual-basis diagonal mass formulation. *Comput. Methods Appl. Mech. Engrg.*, Vol. **346**, 574–591, (2019). https://doi.org/10.1016/j.cma.2018.12.002

[18] Beirão da Veiga, L., Buffa, A., Lovadina, C., Martinelli, M., Sangalli, G. An isogeometric method for the Reissner-Mindlin plate bending problem. *Comput. Methods Appl. Mech. Engrg.*, Vol. **209–212**, pp. 45–53, (2012). https://doi.org/10.1016/j.cma.2011.10.009

[19] Kikis, G., Dornisch, W., Klinkel, S. Adjusted approximation spaces for the treatment of transverse shear locking in isogeometric ReissnerMindlin shell analysis. *Comput. Methods Appl. Mech. Engrg.*, Vol. **354**, pp. 850–870, (2019). https://doi.org/10.1016/j.cma.2019.05.037

[20] Adam, C., Bouabdallah, S., Zarroug, M. and Maitournam, H. Improved numerical integration for locking treatment in isogeometric structural elements, Part I: Beams. *Comput. Methods Appl. Mech. Engrg.*, Vol. **279**, pp. 1–28, (2014). https://doi.org/10.1016/j.cma.2014.06.023

[21] Cazzani, A., Stochino, F. and Turco, E. An analytical assessment of finite element and isogeometric analyses of the whole spectrum of Timoshenko beams. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, Vol. **96**, 1220–1244, (2016). https://doi.org/10.1002/zamm.201500280

A mixed isogeometric plane stress and plane strain formulation with different continuities for the alleviation of locking

L. Stammen* **and W. Dornisch**

Chair of Structural Analysis and Dynamics
Brandenburg University of Technology Cottbus-Senftenberg
Cottbus, Germany
e-mail: {lisa.stammen,wolfgang.dornisch}@b-tu.de

Key words: Isogeometric Analysis, Mixed Formulations, Spline Basis Functions, Continuity, Locking

Abstract: *Isogeometric analysis and mixed finite element methods offer promising opportunities to enhance analysis results for complex problems like incompressible elasticity and are able to cope with different locking phenomena. In this contribution, a mixed two-field isogeometric formulation with independent approximations for displacements and stresses is derived, and its ability to counteract different types of locking is investigated using two examples. Furthermore, the influence of the continuity of the stress shape functions on the accurancy of results and convergence behaviour is shown.*

1 INTRODUCTION

Whilst finite element methods have become a common analysis method in engineering, more recent approaches involve Isogeometric Analysis (IGA), which was founded by Hughes et al. [1] and tries to unify computer aided design (CAD) and finite element analysis (FEA) by using the same model for geometry representation and analysis. Therefore, in contrast to common finite element analysis, non-uniform rational B-splines (NURBS) and other kinds of splines are used as shape functions of the finite elements instead of the usual polynomials. Due to the exact representation of the geometry, analysis results can be improved [1, 2]. Furthermore, many fast and numerically stable algorithms have been developed that exhibit favourable mathematical properties [3]. Other investigations examine the use of different kinds of splines as well [2, 4, 5]. In linear elasticity, different locking phenomena can occur while solving incompressible elasticity problems or dealing with very slender structures for instance. Mixed formulations, where stresses and/or strains or pressures are approximated independently in addition to the usual displacement approximation, can counteract these effects and lead to more accurate results [6]. Recent investigations have already combined isogeometric analysis and mixed formulations in order to benefit from the advantages of both methods [7, 8, 9, 10, 11]. Furthermore, the continuity can have a decisive influence on the accuracy of results [12]. In this contribution, a mixed isogeometric method is proposed in order to improve the analysis results and to counteract locking. Therefore, spline basis functions are used and the displacement shape functions of a two-dimensional isogeometric plane stress and plane strain element are supplemented by independent stress shape functions. These additional stress shape functions are chosen to be of one order lower compared to the displacement shape functions, but with adapted continuity. Evaluating the errors for several examples, it is shown that the proposed mixed method can lead to improved accuracy of results compared to a standard isogeometric formulation and ensures convergence even for very slender geometries and very fine and distorted meshes. Furthermore, the influence of different continuities on the convergence behavior and the accuracy of the results is investigated.

https://doi.org/10.4995/YIC2021.2021.12554

2 MIXED FINITE ELEMENT METHODS

Mixed (or hybrid) formulations approximate primary and secondary variables independently and can be derived, e.g., from weak forms [13]. The resulting finite elements can be employed to reduce locking or cope with incompressible elasticity problems [6]. A starting point for the derivation of a mixed formulation is the following three-field functional

$$\Pi_{HW}(\boldsymbol{u},\boldsymbol{\epsilon},\boldsymbol{\sigma}) = \int_\Omega \frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{D}\boldsymbol{\epsilon} - \boldsymbol{\sigma}^T(\boldsymbol{\epsilon}-\boldsymbol{G}\boldsymbol{u}) - \boldsymbol{u}^T\boldsymbol{b}\ \mathrm{d}\Omega - \int_{\Gamma_t}\boldsymbol{u}^T\bar{\boldsymbol{t}}\ \mathrm{d}\Gamma\ - \int_{\Gamma_u}\boldsymbol{\sigma}^T(\boldsymbol{u}-\overline{\boldsymbol{u}})\ \mathrm{d}\Gamma\ , \quad (1)$$

also known as Hu-Washizu functional. In this equation, $\boldsymbol{u}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\sigma}$, $\boldsymbol{b}$, $\bar{\boldsymbol{t}}$ and $\overline{\boldsymbol{u}}$ represent the displacements, strains, stresses, body forces, boundary tractions and boundary displacements, respectively. $\boldsymbol{G}$ is a suitable differential operator. Using the constitutive equation

$$\boldsymbol{\epsilon} = \boldsymbol{D}^{-1}\boldsymbol{\sigma}\ , \quad (2)$$

we can derive the two-field functional

$$\Pi_{HR}(\boldsymbol{u},\boldsymbol{\sigma}) = \int_\Omega -\frac{1}{2}\boldsymbol{\sigma}^T\boldsymbol{D}^{-1}\boldsymbol{\sigma} + \boldsymbol{\sigma}^T\boldsymbol{G}\boldsymbol{u} - \boldsymbol{u}^T\boldsymbol{b}\ \mathrm{d}\Omega - \int_{\Gamma_t}\boldsymbol{u}^T\bar{\boldsymbol{t}}\ \mathrm{d}\Gamma\ - \int_{\Gamma_u}\boldsymbol{\sigma}^T(\boldsymbol{u}-\overline{\boldsymbol{u}})\ \mathrm{d}\Gamma\ , \quad (3)$$

which is known as the Hellinger-Reissner functional. The variation thereof reads

$$\delta\Pi_{HR}(\boldsymbol{u},\boldsymbol{\sigma}) = \int_\Omega -\delta\boldsymbol{\sigma}^T\boldsymbol{D}^{-1}\boldsymbol{\sigma} + \delta\boldsymbol{\sigma}^T\boldsymbol{G}\boldsymbol{u} + \boldsymbol{\sigma}^T\delta\boldsymbol{G}\boldsymbol{u} - \delta\boldsymbol{u}^T\boldsymbol{b}\ \mathrm{d}\Omega - \int_{\Gamma_t}\delta\boldsymbol{u}^T\bar{\boldsymbol{t}}\ \mathrm{d}\Gamma = 0 \quad (4)$$

for strongly fulfilled boundary conditions $\boldsymbol{u} = \overline{\boldsymbol{u}}$ on Γ_u and is known as the Hellinger-Reissner principle. In order to distinguish plane stress and plane strain formulations, the corresponding material matrix $\boldsymbol{D}$ is used. For more details, see [6, 14].

3 SPLINE BASIS FUNCTIONS

3.1 Construction

Basing on a non-decreasing knot vector

$$\boldsymbol{\Xi} = \{\xi_1,\dots,\xi_{n+p+1}\} \quad , \quad \xi_i \le \xi_{i+1} \quad , \quad i = 1,\dots,n+p \quad (5)$$

and a predefined degree p, the construction of spline basis functions follows the following recurrence algorithm taken from [3]:

$$N_{i,0}(\xi) = \begin{cases} 1 & \text{if } \xi_i \le \xi < \xi_{i+1} \\ 0 & \text{otherwise} \end{cases} \quad , \quad N_{i,p}(\xi) = \frac{\xi-\xi_i}{\xi_{i+p}-\xi_i}N_{i,p-1}(\xi) + \frac{\xi_{i+p+1}-\xi}{\xi_{i+p+1}-\xi_{i+1}}N_{i+1,p-1}(\xi) \quad (6)$$

The corresponding derivatives of the B-spline basis function are calculated by

$$N'_{i,p} = \frac{p}{\xi_{i+p}-\xi_i}N_{i,p-1}(\xi) - \frac{p}{\xi_{i+p+1}-\xi_{i+1}}N_{i+1,p-1}(\xi) \quad . \quad (7)$$

A NURBS surface $\boldsymbol{S}$ can then be represented by

$$\boldsymbol{S}(\xi,\eta) = \sum_{i=1}^{n}\sum_{j=1}^{m} R_{i,j}(\xi,\eta)\boldsymbol{P}_{i,j} \quad (8)$$

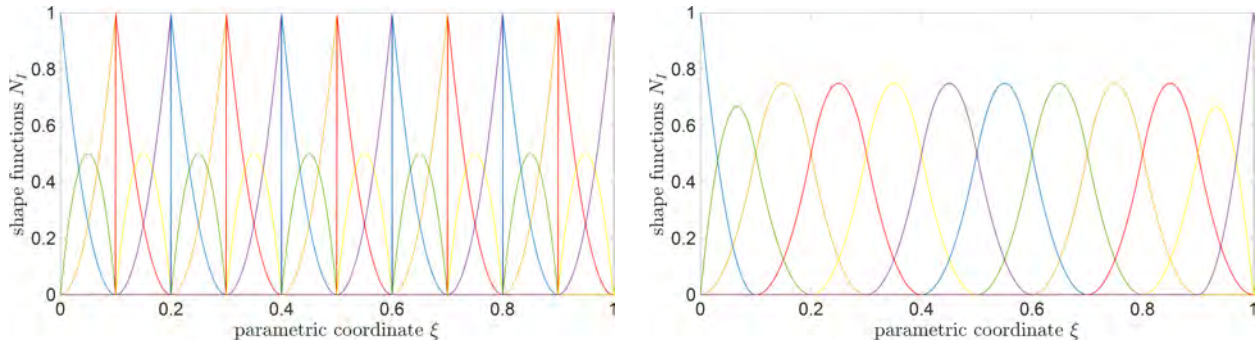
where $R_{i,j}$ are the piece-wise rational basis functions defined by

$$R_{i,j}(\xi,\eta) = \frac{N_{i,p}(\xi)N_{j,q}(\eta)\omega_{i,j}}{\sum_{k=1}^{n}\sum_{l=1}^{m}N_{k,p}(\xi)N_{l,q}(\eta)\omega_{k,l}} \quad . \quad (9)$$

In these formulas, $\boldsymbol{P}_{i,j}$ denotes the control points building up the control net of the surface and $\omega_{i,j}$ represents their corresponding weights. In the following chapters the number of local basis functions will be referred to by $n_{en} = (p+1)(q+1)$ and the total number of global basis functions is denoted by $n_{np} = n \cdot m$.

3.2 Refinement and continuity

There are two refinement methods for B-Splines, which are recalled according to [2]:
The first refinement method increases the number of basis functions by inserting additional knots (*knot insertion*). Thereby, the insertion of one knot leads to an increase in the number of basis functions by 1. As this is equal to dividing an element, this method is often called *h-refinement* in comparison to standard finite element methods. The second refinement method elevates the polynomial order by 1 (*order elevation*). As both methods are hierarchical refinement methods, each of the original basis function can be expressed as linear combination of the refined basis functions.
Based on these two refinement methods, two combined refinement procedures can be derived for isogeometric analysis. If knot insertion is performed before order elevation, this is called *p-refinement*. In this way, one basis function is added for each element. Using *k-refinement*, the spline order is elevated first and subsequently knot insertion is performed. In contrast to *p-refinement*, this method inserts less basis functions. Furthermore, maximal continuity is obtained, while *p-refinement* yields meshes with limited continuity. Figure 1 depicts the influence of the continuity on the shape functions for $p = 3$:



(a) discontinuity (C^{-1}) (b) C^1 - continuity

Figure 1: Influence of the continuity on the shape functions for $p = 3$

4 MIXED ISOGEOMETRIC ELEMENTS WITH TWO UNKNOWN FIELDS

Depending on which unknown fields are chosen, different mixed isogeometric finite elements can be developed from the corresponding variational principle. In this publication the $\boldsymbol{u}$-$\boldsymbol{\sigma}$-mixed formulation, that was derived in [15] for instance, shall be used and adapted to isogeometric analysis.
The chosen fields are approximated by independent shape functions as follows:

$$\boldsymbol{u}^h = \sum_{I=1}^{n_{en}^u} N_I^u \boldsymbol{u}_I \qquad \text{and} \qquad \boldsymbol{\sigma}^h = \sum_{I=1}^{n_{en}^\sigma} N_I^\sigma \boldsymbol{\sigma}_I \tag{10}$$

with

$$\boldsymbol{u}_I = \begin{pmatrix} u_{I1} \\ u_{I2} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\sigma}_I = \begin{pmatrix} \sigma_{I1} \\ \sigma_{I2} \\ \tau_{I12} \end{pmatrix} . \tag{11}$$

Thus, the strain-displacement relation becomes:

$$\boldsymbol{\epsilon}^h = \begin{pmatrix} \epsilon_{11}^h \\ \epsilon_{22}^h \\ 2\gamma_{12}^h \end{pmatrix} = \boldsymbol{G}\boldsymbol{u}^h = \sum_{I=1}^{n_{en}^u} \boldsymbol{B}_I^u \boldsymbol{u}_I \ , \tag{12}$$

where

$$\boldsymbol{B}_I^u = \begin{bmatrix} N_{I,1}^u & 0 \\ 0 & N_{I,2}^u \\ N_{I,2}^u & N_{I,1}^u \end{bmatrix} \tag{13}$$

contains the partial derivatives of the displacement shape functions in the first and second direction ($N_{I,1}^u$ and $N_{I,2}^u$). The interpolation of the variations of $\boldsymbol{u}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\sigma}$ read

$$\delta\boldsymbol{u}^h = \sum_{I=1}^{n_{en}^u} N_I^u \delta\boldsymbol{u}_I \,, \qquad \delta\boldsymbol{\epsilon}^h = \sum_{I=1}^{n_{en}^u} \boldsymbol{B}_I^u \delta\boldsymbol{u}_I \qquad \text{and} \qquad \delta\boldsymbol{\sigma}^h = \sum_{I=1}^{n_{en}^\sigma} N_I^\sigma \delta\boldsymbol{\sigma}_I \quad . \tag{14}$$

Inserting these relations into equation (4) leads to

$$\begin{aligned} \delta\Pi_{HR}^h(\boldsymbol{u}^h, \boldsymbol{\sigma}^h) = &\sum_{I=1}^{n_{en}^\sigma}\sum_{J=1}^{n_{en}^u} \delta\boldsymbol{\sigma}_I^T \int_\Omega N_I^\sigma \boldsymbol{B}_J^u \,\mathrm{d}\Omega\; \boldsymbol{u}_J \\ &+ \sum_{I=1}^{n_{en}^u}\sum_{J=1}^{n_{en}^\sigma} \delta\boldsymbol{u}_I^T \int_\Omega \boldsymbol{B}_I^{uT} N_J^\sigma \,\mathrm{d}\Omega\; \boldsymbol{\sigma}_J \\ &- \sum_{I=1}^{n_{en}^\sigma}\sum_{J=1}^{n_{en}^\sigma} \delta\boldsymbol{\sigma}_I^T \int_\Omega N_I^\sigma \boldsymbol{D}^{-1} N_J^\sigma \,\mathrm{d}\Omega\; \boldsymbol{\sigma}_J \\ &- \sum_{I=1}^{n_{en}^u} \delta\boldsymbol{u}_I^T \left[\int_\Omega N_I^u \boldsymbol{b} \,\mathrm{d}\Omega + \int_{\Gamma_t} N_I^u \overline{\boldsymbol{t}} \,\mathrm{d}\Gamma\right] = 0 \quad . \end{aligned} \tag{15}$$

The control point displacements $\boldsymbol{u}_I$ can now be assembled in the vector

$$\hat{\boldsymbol{u}} = \left(\boldsymbol{u}_1^T, \boldsymbol{u}_2^T, \ldots, \boldsymbol{u}_{n_{np}^u}^T\right)^T \,, \tag{16}$$

where n_{np}^u denotes the number of control points in the displacement mesh. The control point stresses are assembled analogously in

$$\hat{\boldsymbol{\sigma}} = \left(\boldsymbol{\sigma}_1^T, \boldsymbol{\sigma}_2^T, \ldots, \boldsymbol{\sigma}_{n_{np}^\sigma}^T\right)^T \,, \tag{17}$$

where n_{np}^σ denotes the number of control points in the stress mesh. The virtual displacements and virtual stresses $\delta\hat{\boldsymbol{u}}$ and $\delta\hat{\boldsymbol{\sigma}}$ are interpolated akin. Replacing the summations in equation (15) by matrix multiplications leads to:

$$\delta\Pi_{HR}^h(\boldsymbol{u}^h, \boldsymbol{\sigma}^h) = \delta\hat{\boldsymbol{\sigma}}^T \hat{\boldsymbol{C}} \hat{\boldsymbol{u}} + \delta\hat{\boldsymbol{u}}^T \hat{\boldsymbol{C}}^T \hat{\boldsymbol{\sigma}} + \delta\hat{\boldsymbol{\sigma}}^T \hat{\boldsymbol{A}} \hat{\boldsymbol{\sigma}} - \delta\hat{\boldsymbol{u}}^T \hat{\boldsymbol{f}}^u = 0 \tag{18}$$

This equation needs to be fulfilled for every arbitrary test function $\delta\hat{\boldsymbol{\sigma}}$ and $\delta\hat{\boldsymbol{u}}$ and can hence be splitted in two parts, which can be written in standard matrix form. This leads to the following global system of equations

$$\begin{bmatrix} \hat{\boldsymbol{A}} & \hat{\boldsymbol{C}} \\ \hat{\boldsymbol{C}}^T & \boldsymbol{0} \end{bmatrix} \begin{pmatrix} \hat{\boldsymbol{\sigma}} \\ \hat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{f}}^u \end{pmatrix} \,. \tag{19}$$

In a standard manner, these matrices are calculated at the element level and later assembled to the global system. Thus, using the n_{en}^u and n_{en}^σ shape functions which have influence in the respective element e, equation (15) results in:

$$\delta\Pi_{HR}(\boldsymbol{u}^h, \boldsymbol{\sigma}^h) = \bigcup_{e=1}^{n_{el}} \left[\delta\hat{\boldsymbol{\sigma}}^{eT} \hat{\boldsymbol{C}}^e \hat{\boldsymbol{u}}^e + \delta\hat{\boldsymbol{u}}^{eT} \hat{\boldsymbol{C}}^{eT} \hat{\boldsymbol{\sigma}}^e + \delta\hat{\boldsymbol{\sigma}}^{eT} \hat{\boldsymbol{A}}^e \hat{\boldsymbol{\sigma}}^e - \delta\hat{\boldsymbol{u}}^{eT} \hat{\boldsymbol{f}}^{ue}\right] = 0 \tag{20}$$

Defining

$$\hat{\boldsymbol{v}}^e = \begin{pmatrix} \hat{\boldsymbol{\sigma}}^e \\ \hat{\boldsymbol{u}}^e \end{pmatrix} \qquad \text{and} \qquad \delta\hat{\boldsymbol{v}}^e = \begin{pmatrix} \delta\hat{\boldsymbol{\sigma}}^e \\ \delta\hat{\boldsymbol{u}}^e \end{pmatrix} \tag{21}$$

results in the more comprehensive form

$$\delta\Pi_{HR}(\boldsymbol{u}^h, \boldsymbol{\sigma}^h) = \bigcup_{e=1}^{n_{el}} \left[\delta\hat{\boldsymbol{v}}^{eT} \hat{\boldsymbol{K}}^e \hat{\boldsymbol{v}}^e - \delta\hat{\boldsymbol{v}}^{eT} \hat{\boldsymbol{f}}^e \right] = 0 \quad , \tag{22}$$

where

$$\hat{\boldsymbol{K}}^e = \begin{bmatrix} \hat{\boldsymbol{A}}^e & \hat{\boldsymbol{C}}^e \\ \hat{\boldsymbol{C}}^{e^T} & \boldsymbol{0} \end{bmatrix} \tag{23}$$

is the system matrix at element level and

$$\hat{\boldsymbol{f}}^e = \begin{pmatrix} \boldsymbol{0} \\ \hat{\boldsymbol{f}}^{u^e} \end{pmatrix} \tag{24}$$

is the element load vector. The submatrices are computed by:

$$\begin{aligned} \hat{\boldsymbol{A}}^e &= -\int_{\Omega^e} \boldsymbol{N}^{\sigma^T} \boldsymbol{D}^{-1} \boldsymbol{N}^{\sigma} \, \mathrm{d}\Omega \\ \hat{\boldsymbol{C}}^e &= \int_{\Omega^e} \boldsymbol{N}^{\sigma^T} \boldsymbol{B}^{u} \, \mathrm{d}\Omega \\ \hat{\boldsymbol{f}}^{u^e} &= \int_{\Omega^e} \boldsymbol{N}^{u^T} \boldsymbol{b} \, \mathrm{d}\Omega + \int_{\Gamma_t^e} \boldsymbol{N}^{u^T} \bar{\boldsymbol{t}} \, \mathrm{d}\Gamma \quad . \end{aligned} \tag{25}$$

The displacement shape functions N_I^u, which are assembled in

$$\boldsymbol{N}^u = \begin{bmatrix} N_1^u \mathbf{1} & N_2^u \mathbf{1} & \cdots & N_{n_{en}^u}^u \mathbf{1} \end{bmatrix} \quad , \tag{26}$$

are determined as described in chapter 3.1. The stress shape functions N_I^σ are assembled analogously in $\boldsymbol{N}^\sigma$, where $\mathbf{1}$ is the identity matrix of the dimension 2 and 3, respectively. In standard finite element formulations, the number of necessary additional stress variables has to fulfill the stability condition

$$n_\sigma \geq n_u \tag{27}$$

for a two field approach, where $\boldsymbol{\sigma}$ is the primary variable and $\boldsymbol{u}$ is the constraint variable [6]. In this formula n denotes the number of degrees of freedom of the respective variable. Whether this condition is sufficient in Isogeometric Analysis as well has to be investigated in further studies. The additionally introduced stress variables can be condensed out, resulting in the final equation

$$\hat{\boldsymbol{C}}^T \hat{\boldsymbol{A}}^{-1} \hat{\boldsymbol{C}} \hat{\boldsymbol{u}} = -\hat{\boldsymbol{f}}^u \quad . \tag{28}$$

In the next section, the procedure leading to the shape functions for the two chosen fields is described.

5 DETERMINATION OF THE REQUIRED BASIS FUNCTIONS

For the presented mixed isogeometric method, displacements and stresses are chosen as two unknown fields and approximated independently. Thereby, the stress shape functions are chosen to be of one order lower than the displacement shape functions. Furthermore, the continuity of the stress shape functions can be adapted to study the effect on the analysis results. This is implemented in *MATLAB* [16] by using different refinement procedures on the original surface geometry, yielding two different meshes used for the displacements and the stresses, respectively. This procedure is depicted in Figure 2. The resulting meshes are exemplified in Figure 3. The corresponding shape functions can be seen in Figure 4.

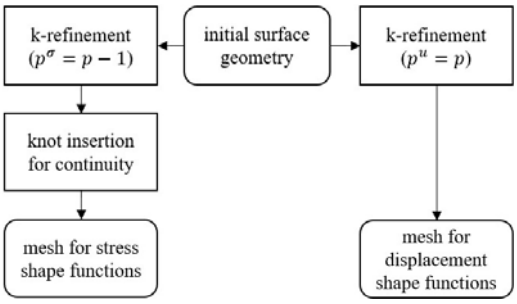
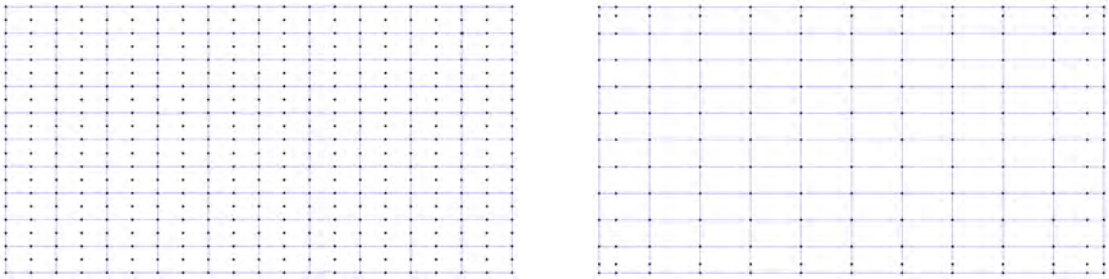


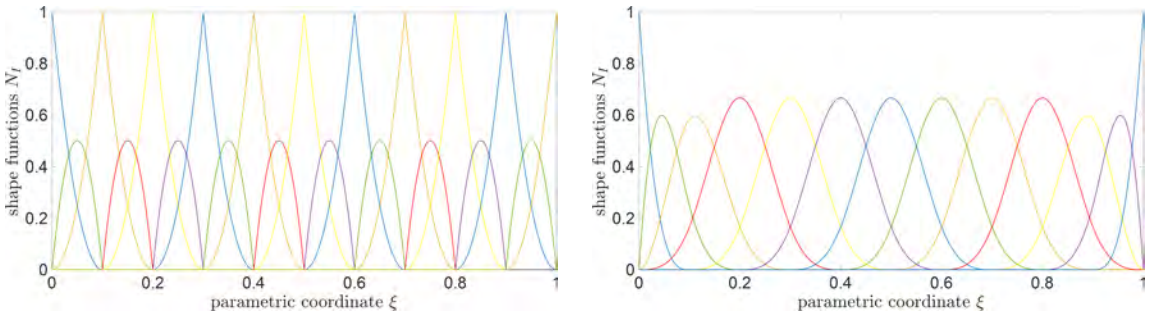
Figure 2: Refinement procedure for the generation of the two different meshes for the displacement and the stress shape functions using k-refinement for different degrees



(a) mesh for stress shape functions (C^0 - continuity)

(b) mesh for displacement shape functions (C^2 - continuity)

Figure 3: Resulting meshes for a rectangular domain divided into 10 elements per direction with their respective control points using $p^u = 3$ and $p^\sigma = 2$



(a) stress shape functions

(b) displacement shape functions

Figure 4: Resulting shape functions for the meshes depicted in Figure 3

6 NUMERICAL EXAMPLES

In this section, the ability of the proposed mixed formulation to counteract different types of locking and the influence of the continuity of the stress shape functions is investigated. Therefore, the results of the proposed mixed isogeometric formulation are compared to the results of a standard (pure displacement based) isogeometric formulation. Within these investigations, for the mixed approach, the continuity of the stress shape functions is varied, whereas the continuity of the displacement shape functions is set to maximal continuity C^{p^u-1} using k-refinement, according to the procedure depicted in Figure 2. The continuity of the shape functions for the standard formulation is varied between C^0 and maximal continuity C^{p-1} by using p- and k-refinement, respectively. Starting with the initial geometry, all meshes are refined regularly and equally in both directions using quadrilateral elements.

6.1 Beam subjected to pure bending

Firstly, a beam under pure bending is investigated. The initial distorted geometry including its control points (red) and relevant material and loading data is indicated in Figure 5.

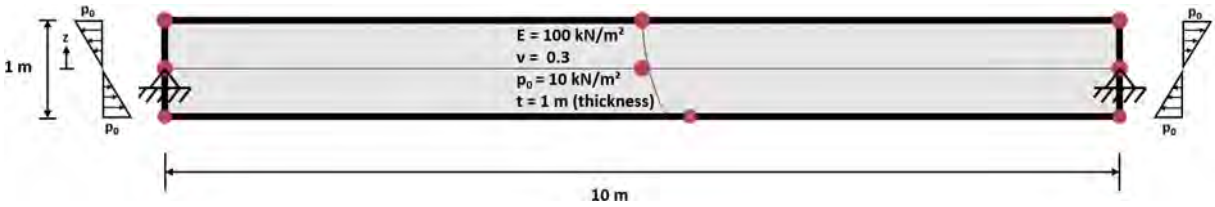
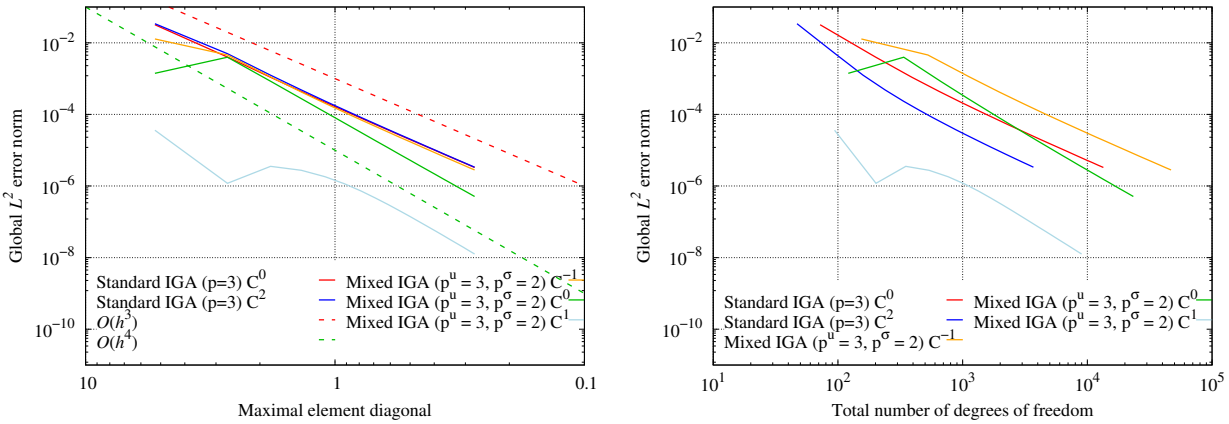


Figure 5: Initial geometry, material and loading data of beam subjected to pure bending

In order to investigate the ability of the formulation to counteract in-plane shear locking, a linearly varying load is applied on both vertical edges for the depicted distorted mesh and the results of the proposed mixed formulation are compared to the analytical solution given in equation (29) using the L^2-error norm of the stresses as indicated in equation (30).

$$\sigma_x(z) = 2 \cdot p_0 \cdot \frac{z}{h}\,, \qquad \sigma_z = 0 \text{ kN/m}\,, \qquad \tau_{xz} = 0 \text{ kN/m} \tag{29}$$

$$||\Delta\sigma||^2 = \sqrt{(\sigma_x - \sigma_x^h)^2 + (\sigma_y - \sigma_y^h)^2 + (\tau_{xy} - \tau_{xy}^h)^2} \tag{30}$$



(a) dependent on the maximal element diagonal (b) dependent on the number of degrees of freedom

Figure 6: Comparison of the L^2 error norm of stresses for the beam subjected to pure bending

As can be seen in Figure 6(a), the proposed mixed isogeometric formulation yields better results compared to a standard isogeometric formulation, for which only a minor difference between C^0- and C^2- continuity can be recognized for this example. Whereas no significant benefit can be achieved by the mixed formulation using discontinuous (C^{-1}) stress shape functions, the use of C^0- continuity offers a better convergence rate ($O(h^4)$) than the standard formulation ($O(h^3)$). Particularly interesting is the behavior for C^1- continuous stress shape functions, since proper convergence behavior begins later as in the other graphs, while constantly offering a much better result. This only holds if the L^2- error norm is calculated using the results of the introduced stress parameters (eq.(17)). If the stresses used for the calculation of the L^2-error norm are directly recalculated from the displacement parameters (eq.(16)), no benefit of the introduced mixed formulation can be observed compared to the standard formulation, and even partly worse results are achieved for C^1- continuity of the stress shape functions of

the mixed formulation. In order to maintain the benefits resulting from the introduced stress parameters if static condensation is used for the proposed mixed formulation, the stresses need to be recalculated using the equation

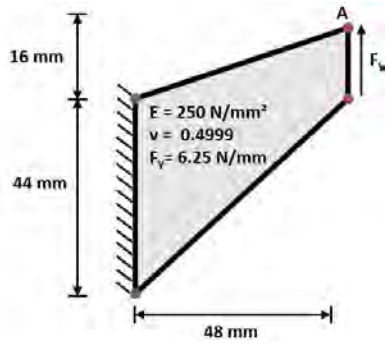
$$\hat{\boldsymbol{\sigma}} = -\hat{\boldsymbol{A}}^{-1}\hat{\boldsymbol{C}}\hat{\boldsymbol{u}} \quad . \tag{31}$$

Taking into account the number of degrees of freedom (cf. 6(b)), the benefits of the mixed formulation only hold for C^1- continuous stress shape functions. Lower continuities offer worse results for the same number of degrees of freedom compared to the standard formulation, for which the benefit of C^2- continuity is the lower number of degrees of freedom.

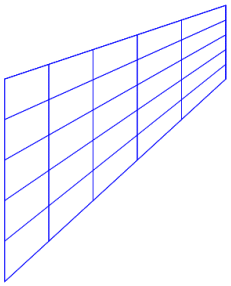
Figure 6 depicts the results for a slenderness ratio of 10. Varying the slenderness of the beam as the critical parameter (by reducing its height), the results of the mixed formulation are constantly better as those of the standard formulation. If the height is reduced to 0.001 m (increasing the slenderness ratio to 10000), the standard formulation diverges, whereas the proposed mixed formulation ensures convergence even for very slender structures.

6.2 Cook's Membrane

The Cook's Membrane is a standard problem to examine the robustness of finite element formulations. The initial geometry and a sample mesh are depicted in Figure 7.



(a) Geometry, material and loading

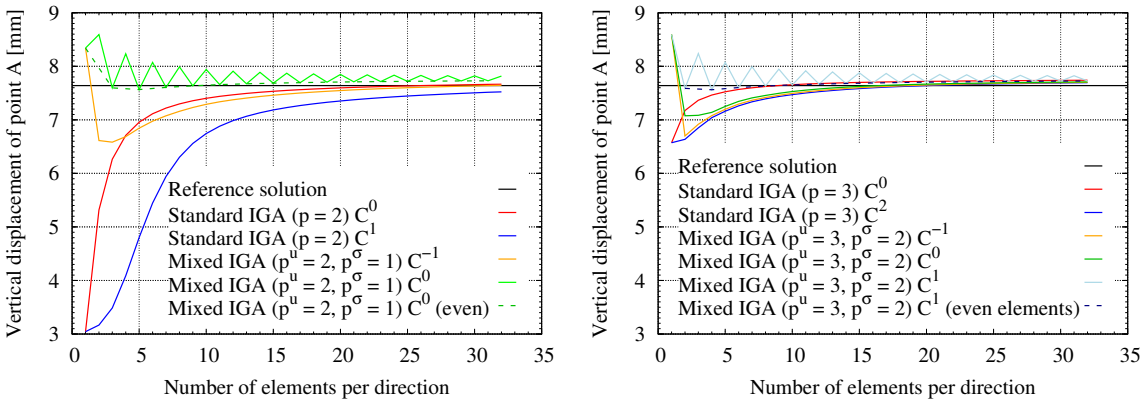


(b) Sample mesh with 5x5 elements

Figure 7: Data of Cook's Membrane problem

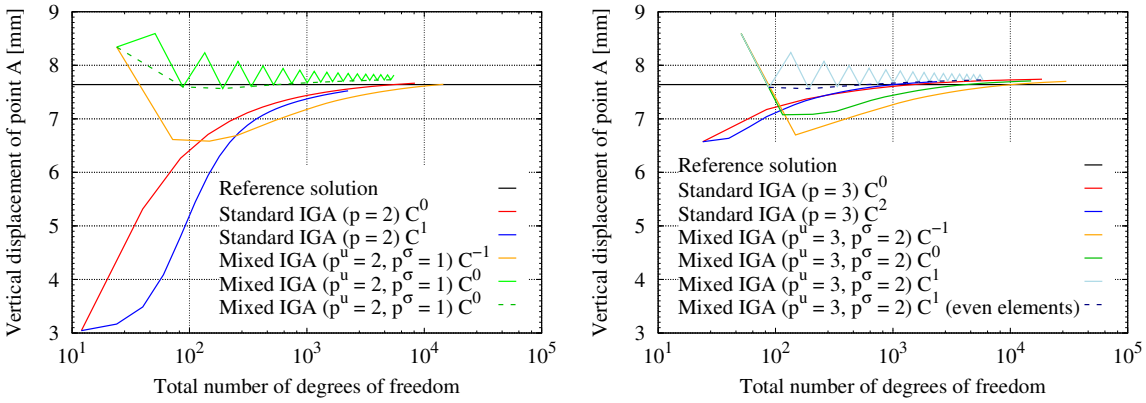
The relevant material and loading parameters were chosen to enable the comparison to the results presented in [17] and in order to investigate the ability of the presented formulation to cope with an incompressible elasticity problem and the resulting locking effects. Therefore, the vertical displacement of point A is compared to the reference solution (black lines in Figure 8 and Figure 9) taken from [17]. As can be observed in Figure 8, the mixed formulation offers better results in comparison to the standard formulation with maximal continuity. Using C^0-continuous shape functions, much better results can be obtained by the standard formulation for this example. Compared to this, the mixed formulation is only beneficial using stress shape functions with maximal continuity $C^{p^\sigma-1}$. However, in this case, the results oscillate between even and odd numbers of elements per direction. Hence, the stability of the mixed method obviously depends on the continuity of the stress shape functions. Examining the eigenvalues of the system matrix for this example, spurious zero eigenvalues occurred for maximal continuity of the stress shape functions, which may cause this instability. This issue will be investigated in detail in further studies. Considering only the results for even numbers of elements (dashed lines), the convergence behavior is superior compared to all graphs. The lower the degree of the shape functions is, the more locking occurs. Thus, the mixed formulation offers a higher benefit for the lowest possible degree of shape functions (cf. Figure 8 (a), (b)). Taking into account the

number of degrees of freedom resulting from the chosen continuity (cf. Figure 9), the benefit of the mixed formulation for maximal continuity of the stress shape functions becomes obvious (considering only even numbers of elements per direction). Especially for higher degrees, the use of C^0- continuous shape functions for the standard formulation looses its benefit in this context.



(a) $p^u = 2$ and $p^\sigma = 1$ (b) $p^u = 3$ and $p^\sigma = 2$

Figure 8: Comparison of the resulting vertical displacement of point A in dependence of the number of elements per direction



(a) $p^u = 2$ and $p^\sigma = 1$ (b) $p^u = 3$ and $p^\sigma = 2$

Figure 9: Comparison of the resulting vertical displacement of point A in dependence of the total number of degrees of freedom

7 CONCLUSIONS

In this contribution, a mixed isogeometric method is derived and its ability to counteract different locking effects is studied for a plane stress and a plane strain example. Furthermore, the influence of the continuity of the stress shape functions is investigated. It is shown that a mixed isogeometric formulation can yield results with a higher convergence rate compared to a standard formulation and is able to counteract different locking phenomena. Additionally, increasing the continuity of the stress shape functions yields better results of the proposed mixed formulation but can yield instabilities due to spurious zero eigenvalues for maximal continuity, despite offering the best results. Further research will focus on the stability for maximal continuity of the stress shape functions. Furthermore, different ansatz spaces for the stress shape functions should be investigated.

REFERENCES

[1] Cottrell, J.A., Hughes, T.J.R. and Bazilevs, Y. *Isogeometric Analysis: Towards integration of CAD and FEA*. John Wiley & Sons, 2009.

[2] de Borst, R., Crisfield, M.A., Remmers, J.J.C. and Verhoosel, C.V. *Nonlinear Finite Element Analysis of Solids and Structures*. John Wiley & Sons, 2012.

[3] Piegl, L. and Tiller, W. *The NURBS Book*. Springer, 1997.

[4] Hllig, K. *Finite Element Methods with B-Splines*. Society for Industrial and Applied Mathematics, 2003.

[5] Mner, B. *B-Splines als Finite Elemente*. Shaker, 2006.

[6] Zienkiewicz, O.C., Taylor, R.L. and Zhu, J.Z. *The finite element method: Its basis and fundamentals*. 6th Edition, Elsevier Butterworth-Heinemann, 2005.

[7] Echter, R., Oesterle, B. and Bischoff, M. A hierarchic family of isogeometric shell finite elements. *Comput. Methods Appl. Mech. Eng.*, Vol. **254**, pp. 170-180, (2013).

[8] Bouclier, R., Elguedj, T. and Combescure, A. Efficient isogeometric NURBS-based solid-shell elements: Mixed formulation and $\overline{B}$-method. *Comput. Methods Appl. Mech. Eng.*, Vol. **265**, pp. 86-110, (2013).

[9] Bouclier, R., Elguedj, T. and Combescure, A. An isogeometric locking-free NURBS-based solid-shell element for geometrically nonlinear analysis. *Int. J. Numer. Methods Eng.*, Vol. **101**, pp. 774-808, (2015).

[10] Bieber, S., Oesterle, B., Ramm, E. and Bischoff, M. A variational method to avoid locking - independent of the discretization scheme. *Int. J. Numer. Methods Eng.*, Vol. **114**, pp. 801-827, (2018).

[11] Zou, Z., Scott, M.A., Miao, D., Bischoff, M., Oesterle, B. and Dornisch, W. An isogeometric ReissnerMindlin shell element based on Bzier dual basis functions: Overcoming locking and improved coarse mesh accuracy. *Comput. Methods Appl. Mech. Eng.*, Vol. **370**, 113283, (2020).

[12] Adam, C., Bouabdallah, S., Zarroug, M and Maitournam, H. Improved numerical integration for locking treatment in isogeometric structural elements, Part I: Beams. *Comput. Methods Appl. Mech. Eng.*, Vol. **279**, pp. 1-28, (2014).

[13] Reddy, J.N. *An Introduction to the Finite Element Method*. McGraw-Hill, 2006.

[14] Knothe, K. and Wessels, H. *Finite Elemente: Eine Einfhrung fr Ingenieure*. Springer Berlin Heidelberg, 2008.

[15] Pian, T.H.H. and Sumihara, K. Rational approach for assumed stress finite elements. *Int. J. Num. Meth. Eng.*, Vol. **20**, pp. 1685-1695, (1984).

[16] The MathWorks, Inc. *MATLAB* (Version R2019b). Natick, Massachusetts, 2019.

[17] Bombarde, D.S., Nandy, A. and Gautam, S.S. A Two-Field Formulation in Isogeometric Analysis to Alleviate Locking. *Advances in Engineering Design: Select Proceedings of FLAME 2020.*, pp. 191-199, (2021).

Equilibrium-Based Finite Element Formulation for Timoshenko Curved Tapered Beams

Hugo A.F.A. Santos

Área Departamental de Engenharia Mecânica
Instituto Superior de Engenharia de Lisboa
Rua Conselheiro Emídio Navarro 1, 1959-007 Lisboa, Portugal
e-mail: hugo.santos@isel.pt

Key words: Finite Elements, Equilibrium Approach, Curved Beams, Tapered Cross-Sections, Timoshenko Theory

Abstract: *Curved tapered beams have been widely used in many engineering applications. Their complex geometries pose challenges to the development of robust approaches for the numerical modelling of their mechanical behaviour. The aim of the present contribution is to introduce a novel, simple and effective, finite element formulation for the quasi-static analysis of Timoshenko curved tapered beams. This formulation relies on a complementary variational approach based on a set of approximations that satisfy in strong form all equilibrium conditions of the boundary-value problem, resulting thus in a formulation that is free from both shear and membrane locking phenomena. The effectiveness of the formulation is numerically demonstrated through its application to different beam problems and the obtained results are analysed and discussed.*

1 INTRODUCTION

Due to their excellent mechanical performance and structural efficiency, curved beams have been widely used in many engineering applications, such as: bridge structures, piping systems, biomedical devices, aerospace and aeronautical structures, *etc.* Their complex geometries pose challenges to the development of robust approaches for the modelling of their mechanical behaviour.

Among the various approaches available in the literature for their analysis, those that are based on the finite element method have been the most successful. The simplest finite element modeling strategy for curved beams is an assembly of relatively short straight beam elements [1]. However, such an approach generally requires a large number of elements to obtain converged solutions. Furthermore, when applied to Timoshenko based structural models, some of these finite element approaches, in particular those that rely on the approximation of the displament fields based on lower-order shape functions, are prone to shear locking when the beam elements become slender and to membrane locking when their curvature increases [2]. Hybrid-mixed formulations, in which both displacement and force/bending moment fields are approximated, with the goal of avoiding the locking phenomena, were also explored [3]. However, most of the formulations that have been developed for the analysis of curved beams are limited to uniform cross-section cases. There are a few exceptions in which finite element formulations for curved beams with non-uniform cross-sections can be found in the literature, see *e.g.* [4, 5].

The aim of the present contribution is to introduce a novel, simple and effective, finite element formulation for the quasi-static analysis of Timoshenko curved tapered beams. Following the methodology adopted in [6, 7, 8], the proposed formulation relies on a complementary variational approach, only requiring the approximation of the internal force/bending moment fields. Such approximations are selected such that they satisfy in strong form all equilibrium conditions of the boundary-value problem. The formulation is naturally free from both shear and membrane locking phenomena. The effectiveness of the formulation is numerically demon-

https://doi.org/10.4995/YIC2021.2021.12567

strated through its application to different beam problems and the obtained results are analysed and discussed.

2 BOUNDARY-VALUE PROBLEM

Consider a two-dimensional curved beam whose geometry is described by its centroidal axis denoted by $\mathcal{C}$. The centroidal axis $\mathcal{C}$ is parameterized by $s \in [0, L]$, with L denoting the length of the beam in its reference configuration. $\mathcal{C}$ is decomposed into an internal part, represented by $\Omega =]0, L[$, and a boundary part, identified by $\Gamma = \Gamma_N \cup \Gamma_D = \{0, L\}$, where Γ_N and Γ_D correspond to the Neumann and Dirichlet boundaries, respectively, such that $\Gamma_N \cap \Gamma_D = \emptyset$.

Let the beam be subjected to: distributed loads defined per unit length, denoted by p and q, and bending moments, denoted by m, applied in Ω and assumed to depend on s, concentrated loads $\bar{N}$ and $\bar{V}$ and a concentrated moment $\bar{M}$ applied on Γ_N; prescribed displacements, $\bar{u}$ and $\bar{w}$, and a prescribed rotation $\bar{\phi}$ defined on Γ_D. While p, $\bar{N}$ and $\bar{u}$ represent axial quantities, q, $\bar{V}$ and $\bar{w}$ represent transverse quantities. m, $\bar{M}$ and $\bar{\phi}$ represent rotational quantities. The loads are assumed to act at the centroidal axis of the beam.

The kinematical differential equations of the beam model under consideration are given in Ω as

$$\varepsilon_{ss} = \frac{1}{1 + \frac{z}{R}}(\varepsilon + z\chi) \tag{1a}$$

$$\gamma_{sz} = \frac{1}{1 + \frac{z}{R}}\gamma \tag{1b}$$

in which

$$\varepsilon = u' - \frac{w}{R} \tag{2a}$$

$$\gamma = w' + \frac{u}{R} - \phi \tag{2b}$$

$$\chi = \phi' \tag{2c}$$

with ε being the axial deformation, γ the shear deformation and χ the bending curvature of the beam. R stands for the radius of curvature of the beam centroidal axis, which, in general, may depend on s. As the shear deformation γ is not disregarded, the adopted model is based on Timoshenko's beam theory.

The Dirichlet (kinematical) boundary conditions of the problem are given as follows

$$u - \bar{u} = 0, \text{ on } \Gamma_D \tag{3a}$$

$$w - \bar{w} = 0, \text{ on } \Gamma_D \tag{3b}$$

$$\phi - \bar{\phi} = 0, \text{ on } \Gamma_D \tag{3c}$$

where u and w are the axial and transverse displacements of the beam axis, respectively, and ϕ the rotation of the beam cross-section.

The equilibrium of an infinitesimal beam element can be expressed by the following set of differential equations in Ω

$$N' - \frac{V}{R(s)} + p(s) = 0 \tag{4a}$$

$$V' + \frac{N}{R(s)} + q(s) = 0 \tag{4b}$$

$$-M' + V + m(s) = 0 \tag{4c}$$

representing equilibrium of axial forces, shear forces and bending moments, respectively, where $(\cdot)'$ stands for the derivative of $(\cdot)$ with respect to s.

The Neumann (or static) boundary conditions of the problem are

$$nN - \bar{N} = 0, \text{ on } \Gamma_N \tag{5a}$$
$$nV - \bar{V} = 0, \text{ on } \Gamma_N \tag{5b}$$
$$nM + \bar{M} = 0, \text{ on } \Gamma_N \tag{5c}$$

with

$$n = \begin{cases} 1 & \text{if } x = L \\ -1 & \text{if } x = 0 \end{cases}$$

The constitutive equations are taken as the following relationships defined in Ω

$$\sigma_{ss} = E\varepsilon_{ss} \tag{6a}$$
$$\tau_{sz} = G\gamma_{sz} \tag{6b}$$

with E and G denoting Young's modulus and shear modulus, respectively, of the beam, such that

$$G = \frac{E}{2(1+\nu)} \tag{7}$$

with ν standing for Poisson's coefficient. A linear elastic material behavior is, thus, assumed in this study. The material properties E and ν are taken as constants in Ω.

The internal forces and bending moment fields correspond to the following stress resultants on a beam cross-section

$$N = \int_A \sigma_{ss} dA \tag{8a}$$
$$V = \int_A \tau_{sz} dA \tag{8b}$$
$$M = \int_A \sigma_{ss} z dA \tag{8c}$$

with dA an infinitesimal area element of the beam cross-section. Making use of these definitions, and upon substitution of (1) and (6), leads to

$$N = C_{11}\varepsilon + C_{12}\chi \tag{9a}$$
$$V = C_{33}\gamma \tag{9b}$$
$$M = C_{12}\varepsilon + C_{22}\chi \tag{9c}$$

with

$$C_{11} = \int_A \frac{E}{1+\frac{z}{R}} dA$$
$$C_{12} = \int_A \frac{Ez}{1+\frac{z}{R}} dA$$
$$C_{22} = \int_A \frac{Ez^2}{1+\frac{z}{R}} dA$$
$$C_{33} = \int_A \frac{f_s G}{1+\frac{z}{R}} dA$$

where f_s stands for the shear correction factor. It is worth noting the coupling between N and M. Notably, if $\frac{h}{R} \ll 1$, then this coupling disappears.

3 VARIATIONAL BASIS

The strain energy functional of a beam element assumes the following form

$$U = \frac{1}{2}\int_V \left(E\varepsilon_{ss}^2 + f_s G\gamma_{sz}^2\right)\, dV \tag{11}$$

with dV an infinitesimal volume element of the beam. Upon substitution of equations (2) and (6), the strain energy can be recast as

$$\begin{aligned} U &= \frac{1}{2}\int_V \left(\frac{E}{\left(1+\frac{z}{R}\right)^2}\left(\varepsilon + z\chi\right)^2 + \frac{f_s G}{\left(1+\frac{z}{R}\right)^2}\gamma^2\right)\, dV \\ &= \frac{1}{2}\int_V \left(\frac{E}{\left(1+\frac{z}{R}\right)^2}\left(\varepsilon^2 + 2z\chi\varepsilon + z^2\chi^2\right) + \frac{f_s G}{\left(1+\frac{z}{R}\right)^2}\gamma^2\right)\, dV \end{aligned}$$

Since $dV = \left(1+\frac{z}{R}\right) dA d\Omega$, with $1+\frac{z}{R}$ the Jacobian of the transformation, and since the deformations ε, χ and γ only depend on the curvilinear coordinate s, the strain energy can be rewritten as

$$U = \frac{1}{2}\int_\Omega \left(\left(C_{11}\varepsilon + C_{12}\chi\right)\varepsilon + \left(C_{12}\varepsilon + C_{22}\chi\right)\chi + \left(C_{33}\gamma\right)\gamma\right)\, d\Omega \tag{13}$$

or, making use of (9), as

$$U = \frac{1}{2}\int_\Omega \left(N\varepsilon + M\chi + V\gamma\right)\, d\Omega \tag{14}$$

The total potential energy of the boundary-value problem under study is, thus, the functional $\Pi_p : \mathcal{U}_k(\Omega) \to \mathcal{R}$ given by

$$\Pi_p(u, w, \phi) = U(\varepsilon(u, w), \chi(\phi), \gamma(u, w, \phi)) + F(u, w, \phi) \tag{15}$$

where U is the strain energy functional defined in (14) and F represents the external potential energy given by

$$F(u, w, \phi) = -\int_\Omega (pu + qw + m\phi)\, d\Omega - [\bar{N}u]_{\Gamma_N} - [\bar{V}w]_{\Gamma_N} - [\bar{M}\phi]_{\Gamma_N} \tag{16}$$

$\mathcal{U}_k$ is the kinematically admissible space defined as

$$\mathcal{U}_k = \{(u, w, \phi) \in \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) |\ u = \bar{u}, w = \bar{w}, \phi = \bar{\phi} \text{ on } \Gamma_D\} \tag{17}$$

where $\mathcal{H}^1(\Omega)$ represents a standard Sobolev space.

Inverting relations (9) gives

$$\varepsilon = \frac{C_{22}N - C_{12}M}{C_{11}C_{22} - C_{12}^2} \tag{18a}$$

$$\gamma = \frac{V}{C_{33}} \tag{18b}$$

$$\chi = \frac{C_{11}M - C_{12}N}{C_{11}C_{22} - C_{12}^2} \tag{18c}$$

On insertion of (18) into the strain energy functional (14) leads to the following complementary strain energy functional

$$U_c = \frac{1}{2}\int_\Omega \left(\frac{C_{22}N^2 - 2C_{12}NM + C_{11}M^2}{C_{11}C_{22} - C_{12}^2} + \frac{V^2}{C_{33}}\right)\, d\Omega \tag{19}$$

which only involves the internal forces/bending moment fields.

The associated total complementary energy $\Pi_c : \mathcal{U}_s(\Omega) \to \mathcal{R}$ comes out as

$$\Pi_c(N, V, M) = -U_c(N, V, M) + \Pi_{c,ext}(N, V, M) \tag{20}$$

in which $\Pi_{c,ext}$ represents the external complementary energy given as follows

$$\Pi_{c,ext}(N, V, M) = [nN\bar{u}]_{\Gamma_D} + [nV\bar{w}]_{\Gamma_D} + [nM\bar{\phi}]_{\Gamma_D} \tag{21}$$

and $\mathcal{U}_s$ stands for the statically admissible space defined as

$$\begin{aligned}\mathcal{U}_s = \{(N, V, M) \in \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) | \\ N' - \frac{V}{R(s)} + p(s) = 0,\ V' + \frac{N}{R(s)} + q(s) = 0,\ -M' + V + m(s) = 0 \text{ in } \Omega; \\ nN - \bar{N} = 0,\ nV - \bar{V} = 0,\ nM + \bar{M} = 0 \text{ on } \Gamma_N\}\end{aligned}$$

(N, V, M) is said to be a stationary point of Π_c if, and only if, the following condition holds

$$\delta\Pi_c = 0,\ \forall\ (\delta N, \delta V, \delta M) \in \mathcal{V}_s \tag{22}$$

where $\mathcal{V}_s$ represents the homogeneous statically admissible space defined as

$$\begin{aligned}\mathcal{V}_s = \{(\delta N, \delta V, \delta M) \in \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) |\ \delta N' - \frac{\delta V}{R(s)} = 0,\ \delta V' + \frac{\delta N}{R(s)} = 0, \\ -\delta M' + \delta V = 0,\ \text{in } \Omega; n\delta N = 0,\ n\delta V = 0,\ n\delta M = 0,\ \text{on } \Gamma_N\}\end{aligned}$$

A novel finite element formulation for the quasi-static analysis of Timoshenko curved tapered beams will be developed in the following on the basis of the complementary variational approach introduced above.

4 FINITE ELEMENT FORMULATION

As a starting point, let us define $\mathcal{H}^0_h$ and $\mathcal{H}^1_h$ as families of closed finite-dimensional subspaces of $\mathcal{H}^0$ and $\mathcal{H}^1$, respectively. A finite element approximation of (22) consists of seeking $(N^h, M^h, V^h) \in \mathcal{U}^h_s$ such that (22) holds for all $(\delta N^h, \delta M^h, \delta V^h) \in \mathcal{V}^h_s$, where $\mathcal{U}^h_s \subset \mathcal{U}_s$ and $\mathcal{V}^h_s \subset \mathcal{V}_s$ represent the discretized statically admissible spaces.

Let us assume that the entire domain Ω is partitioned in subdomains $\Omega_e \subset \Omega$, such that $\Omega = \cup_{e=1}^{n_e} \Omega_e$ in which n_e represents the number of beam elements. If the inter-element equilibrium conditions and Neumann boundary conditions are relaxed within the framework of the complementary energy principle, then, the following augmented Lagrangian, or hybrid complementary energy, has to be considered

$$L_c = \sum_{e=1}^{n_e} \Pi_{c,e} + \sum_{i=1}^{n_{int}} \left(\lambda_i^N [\![N]\!]_{\Gamma_i} + \lambda_i^V [\![V]\!]_{\Gamma_i} + \lambda_i^M [\![M]\!]_{\Gamma_i}\right) \tag{23}$$

where n_{int} represents the number of inter-element boundaries and Γ_i stands for the inter-element boundary i. $[\![(\cdot)]\!]$ stands for the jump of $(\cdot)$ on Γ_i. λ_i^N, λ_i^V and λ_i^M are the Lagrange multipliers, defined on Γ_i, that are energy-conjugate of N, V and M, respectively.

Without loss of generality, and only for the sake of simplicity, let us consider the case of beams with zero distributed loads, *i.e.*, $p(s) = q(s) = 0$ and $m(s) = 0$. Then, the solutions to

the equilibrium differential equations (4) are as follows

$$N(s) = c_2 \sin(k(s)) + c_1 \cos(k(s)) \tag{24a}$$
$$V(s) = c_2 \cos(k(s)) - c_1 \sin(k(s)) \tag{24b}$$
$$M(s) = c_3 + c_2 \int \cos(k(s))\ ds - c_1 \int \sin(k(s))\ ds \tag{24c}$$

where c_1, c_2 and c_3 are constants and $k(s)$ is defined as

$$k(s) = \int \frac{1}{R(s)}\ ds \tag{25}$$

It is worth noting that, if, additionally, the beam radius of curvature R is constant, then $k(s)$ results as

$$k(s) = \frac{s}{R} \tag{26}$$

and, therefore, the internal forces/bending moment functions (24) come out as

$$N(s) = c_2 \sin\left(\frac{s}{R}\right) + c_1 \cos\left(\frac{s}{R}\right) \tag{27a}$$
$$V(s) = c_2 \cos\left(\frac{s}{R}\right) - c_1 \sin\left(\frac{s}{R}\right) \tag{27b}$$
$$M(s) = c_3 + c_2 R \sin\left(\frac{s}{R}\right) + c_1 R \cos\left(\frac{s}{R}\right) \tag{27c}$$

These functions are taken as the trial finite element approximations and a Galerkin approach is adopted, *i.e.*, the problem is numerically approached assuming the same trial and test approximation function spaces within the framework of the augmented Lagrangian given by (23). The involved integrals are numerically computed using a 5-point Gaussian quadrature rule.

Differentiation of L_c^h with respect to all the unknown parameters gives rise to a governing system of linear equations that involve the element constants c_i and the Lagrange multipliers as fundamental unknowns.

5 NUMERICAL TESTS

5.1 Quarter-Circular Cantilver Uniform Beam Under Shear Force at its Free End - Shear and Membrane Locking Tests

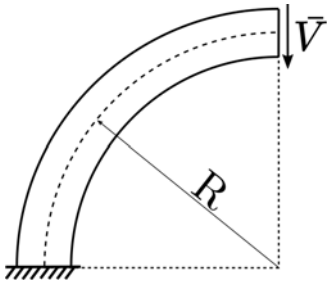


Figure 1: Quarter-Circular Cantilver Uniform Beam Under a Shear Force at its Free End

The classical problem of a quarter-circular uniform cantilever beam subjected to a shear force at its the free end as is illustrated in Figure 1 is analyzed first in order to test the capability of the proposed formulation to overcome the shear- and membrane-locking phenomena. The applied shear force, the cross-section width and the beam curvature radius were set to $\bar{V} = 1kN$,

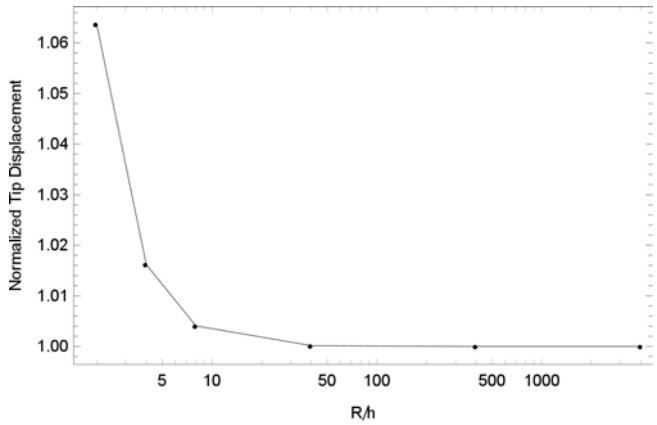


Figure 2: Quarter-Circular Cantilver Uniform Beam Under a Shear Force at its Free End - Shear and Membrane Locking Tests

$b = 0.4m$ and $R = 4m$. The analysis was carried out varying the slenderness ratio R/h of the beam. The beam was modeled using only one finite element. It is worth mentioning that finer finite element discretizations would exactly lead to the same results. The transverse displacements of the free end of the beam were computed and normalized with respect to their corresponding Euler-Bernoulli solutions, w_f^{EB}, for different values of the slenderness ratio, where

$$w_f^{EB} = \frac{\pi \bar{V} R^3}{4EI} + \frac{\pi \bar{V} R}{4EA} \tag{28}$$

The obtained results are shown in Figure 2. As it can be seen, as the slenderness ratio R/h increases, or, in other words, as the beam becomes thinner, the transverse tip displacements tend to the Euler-Bernoulli solutions. This shows that the proposed formulation does not suffer from either shear or membrane locking.

5.2 Clamped-Clamped Circular Beam with Tip Load

To validate and assess the accuracy and effectiveness of the proposed finite element formulation, a clamped-clamped circular beam with rectangular cross-section under tip loads as depicted in Figure 3 is herein analyzed. A uniform (constant h) beam is considered first and, afterwards, a tapered beam is studied. The following numerical parameters were considered for both situations: radius of curvature $R = 4m$, cross-section width $b = 0.4m$, shear correction factor $f_s = 5/6$ and opening angle $\theta_o = 2\pi/3$.

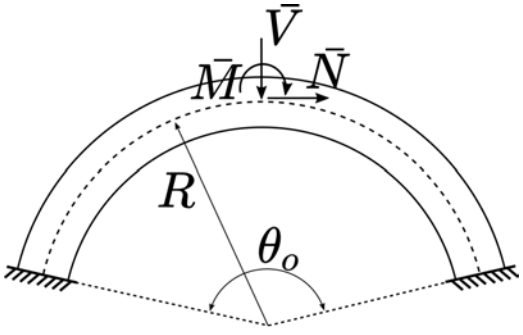


Figure 3: Clamped-Clamped Circular Beam

5.2.1 Clamped-Clamped Circular Uniform Beam with Tip Load - Accuracy test

In order to assess the accuracy of the proposed formulation, a uniform beam with a tip load is herein analysed. In this case, the cross-section height was set to $h = 0.6m$ and the Young's modulus and Poisson's ratio were set to $E = 30GPa$ and $\nu = 0.17$. The tip loads were assumed as $\bar{N} = \bar{V} = 1kN$ and $\bar{M} = 1kNm$. The beam was modeled using two finite elements. The accuracy of the proposed formulation is assessed by comparing the obtained results with the reference ones given in [4].

As it can be seen from the analysis of Table 1, the results produced by the proposed formulation are essentially the same as the reference ones.

Load Case	$\times 10^{-6}$	Ref. Sol. [4]	Present Study
$\bar{N} = 0,\ \bar{V} \neq 0,\ \bar{M} = 0$	$\frac{w}{R\theta_o}$	0.248781	0.248781
$\bar{N} \neq 0,\ \bar{V} = 0,\ \bar{M} = 0$	$\frac{u}{R\theta_o}$	0.12522	0.125221
	$\frac{\phi}{\theta_o}$	-0.379642	-0.379642
$\bar{N} = 0,\ \bar{V} = 0,\ \bar{M} \neq 0$	$\frac{u}{R\theta_o}$	-0.09491	-0.094910
	$\frac{\phi}{\theta_o}$	1.08224	1.082238

Table 1: Clamped-Clamped Circular Uniform Beam with Tip Load - Accuracy test

5.2.2 Clamped-Clamped Circular Tapered Beam with Tip Load

A tapered beam with a transverse tip load as illustrated in Figure 4 is now analysed. In this case, the initial cross-section height was set to $h_0 = h(0) = 0.6m$ and the Young's modulus and Poisson's ratio were set to $E = 70GPa$ and $\nu = 0.3$. The tip load was set to $\bar{V} = 1kN$. The beam was modeled firstly using two finite elements. The variation of the cross-section height was taken as $h(s) = h_0\left(1 - \frac{s}{1.1L}\right)$, which corresponds to a beam with a cross-section height at $s = L$ given by $h(L) = \frac{h_0}{11}$.

The obtained diagrams of axial force, shear force and bending moment are shown in Figures 5, 6 and 7, respectively. As expected, neither the axial force nor the bending moment diagrams exhibit symmetry with respect to a vertical axis crossing the mid-span of the beam. Likewise, the shear force diagram is not anti-symmetric with respect to the mentioned axis. This is in opposition to what would be obtained if a uniform beam would have been considered. It is also interesting to note that the bending moment at $s = L$ is considerably lower than that at $s = 0$. This is clearly a consequence of the lower bending stiffness of the beam at $s = L$ when compared to that at $s = 0$.

Finally, a mesh convergence study was performed, in which the tip transverse displacement was computed using 2, 4, 8 and 16 finite elements. The obtained results are provided in Table 2, showing that the finite element formulation converges to a solution.

6 CONCLUSIONS

- A novel finite element formulation for the quasi-static analysis of Timoshenko curved tapered beams was proposed.
- The formulation relies on a hybrid complementary energy variational principle leading to statically admissible solutions.
- The formulation proved to be effective and naturally insensitive to the shear and membrane locking phenomena.

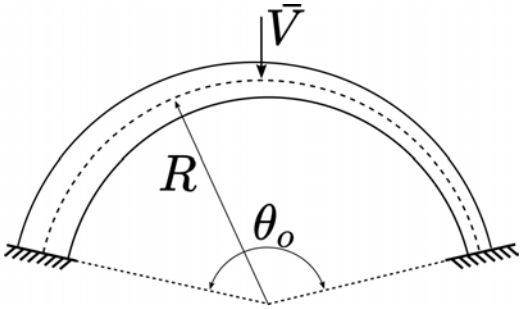


Figure 4: Clamped-Clamped Circular Tapered Beam with Tip Load

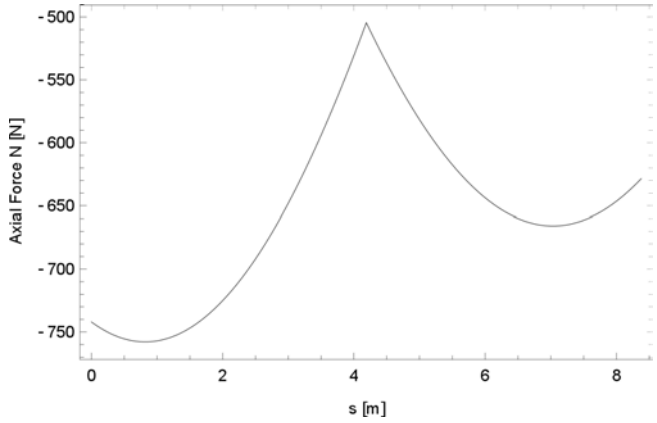


Figure 5: Clamped-Clamped Circular Tapered Beam with Tip Load - Axial Force Diagram

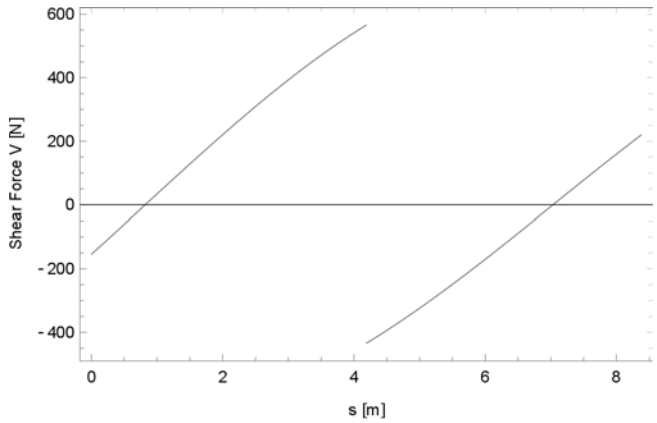


Figure 6: Clamped-Clamped Circular Tapered Beam with Tip Load - Shear Force Diagram

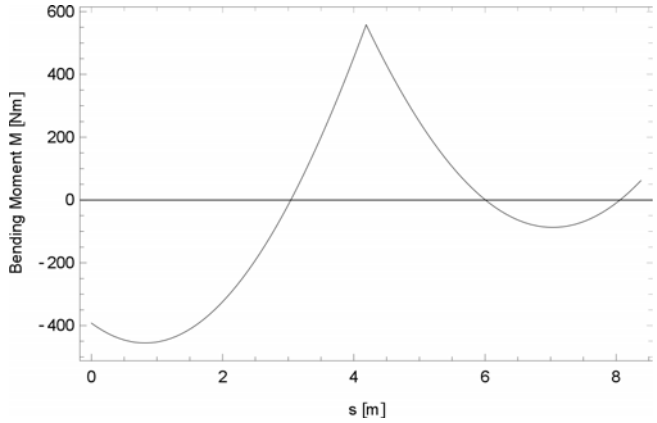


Figure 7: Clamped-Clamped Circular Tapered Beam with Tip Load - Bending Moment Diagram

n_e	$\frac{w\times10^{-6}}{R\theta_o}$
2	0.904064
4	0.938990
8	0.941257
16	0.941300

Table 2: Clamped-Clamped Circular Tapered Beam with Tip Load - Mesh Convergence Study

REFERENCES

[1] Kikuchi, F. On the validity of the finite element analysis of circular arches represented by an assemblage of beam elements, *Computer Methods in Applied Mechanics and Engineering*, Vol. **5**, pp. 253–276, (1975).

[2] Stolarski, H. and Belytschko, T. Shear and membrane locking in curved C0 elements, *Computer Methods in Applied Mechanics and Engineering*, Vol. **41**, pp. 279–296, (1983).

[3] Saleeb, A.F. and Chang, T.Y. On the hybrid-mixed formulations of C0 curved beam elements, *Computer Methods in Applied Mechanics and Engineering*, Vol. **60**, pp. 95–121, (1987).

[4] Tufekci, E. and Arpaci, A. Analytical solutions of in-plane static problems for non-uniform curved beams including axial and shear deformations, *Structural Engineering and Mechanics*, Vol. **22**, pp. 131–150, (2006).

[5] Tufekci, E., Eroglu, U. and Aya, S.A. A new two-noded curved beam finite element formulation based on exact solution, *Engineering with Computers*, Vol. **33**, pp. 261–273, (2017).

[6] Santos, H.A.F.A. Complementary-energy methods for geometrically non-linear structural models: an overview and recent developments in the analysis of frames, *Archives of Computational Methods in Engineering*, Vol. **18**, pp. 405–440, (2011).

[7] Santos, H.A.F.A. Variationally consistent force-based finite element method for the geometrically non-linear analysis of Euler-Bernoulli framed structures, *Finite Elements in Analysis and Design*, Vol. **53**, pp. 24–36, (2012).

[8] Santos, H.A.F.A. and Silberschmidt, V.V. Hybrid equilibrium finite element formulation for composite beams with partial interaction, *Composite Structures*, Vol. **108**, pp. 646–656, (2014).

An Isogeometric Element Formulation for Linear Two-Dimensional Elasticity Based on the Airy Equation

S. Held*, W. Dornisch and N. Azizi

Chair of Structural Analysis and Dynamics
Brandenburg University of Technology Cottbus-Senftenberg
Cottbus, Germany
e-mail:{susanne.held, wolfgang.dornisch, nima.azizi}@b-tu.de

Key words: Airy equation, isogeometric analysis, two-dimensional elasticity, NURBS, element formulation

Abstract: *The aim of this work is to derive a formulation for linear two-dimensional elasticity using just one degree of freedom. This degree of freedom is used to directly discretize the Airy bipotential equation, which requires higher order basis functions. Isogeometric structural analysis is based on shape functions of the geometry description in Computer-Aided design software. These shape functions can easily fulfill the continuity requirement of the bipotential equation. Thus, an Airy element formulation can be obtained through isogeometric methods. In this contribution Non-Uniform Rational B-splines are used to discretize the domain and to solve the occurring differential equations. Numerical examples demonstrate the accuracy of the evolved formulation for a quadratic plate under different load situations.*

1 INTRODUCTION

In 2005 Hughes et. al [1] introduced isogeometric analysis (IGA). The basic idea of IGA is to use one common geometry model for design in Computer-Aided design (CAD) software and analysis with the finite element method (FEM) to overcome model conversions between design and analysis. Therefore the basis functions, commonly Non-Uniform Rational B-splines (NURBS) basis functions, from CAD models are also utilized as the basis for the FEM. Besides the exact description of the geometry, NURBS can also provide high inter-element continuity. That is why results of equal accuracy to standard FEM can be achieved using less elements. That points out the huge potential of IGA also for complex geometries. For further information on basics of NURBS, see [2, 3].

Compared to classical FEM, the numerical effort in IGA is slightly shifted from solving to assembly. Thus, currently a strong focus in IGA research is set on efficient integration rules, since the total number of integration points scales very well with the assembly costs. Initially, the use of full and reduced Gauss integration was proposed in [1, 4]. Currently, integration rules which are computed from moment fitting equations for specific problems have shown highly improved efficiency [5, 6, 7]. A very different idea to lower the computational effort for two-dimensional linear elasticity might be to further make use of the high continuity and choose to discretize not the standard weak form of equilibrium, but rather the well-known Airy equation, which is able to describe two-dimensional linear elasticity with only one unknown. This partial differential equation (PDE) combines the set of PDEs of the classical two-dimensional linear elasticity formulation approach. Instead of the two unknown displacements per control point of a standard two-dimensional elasticity formulation, with the discretized Airy formula only one degree of freedom per control point is obtained.

2 BASIC NURBS TERMINOLOGY FOR 2D ELASTICITY

The considered parameter space is subdivided into knot spans, denoted as elements. The knots are arranged as a non-decreasing array, the knot vector $\boldsymbol{\Xi} = \{\xi_1, \xi_2, ..., \xi_{n+p+1}\}$. Here, n

https://doi.org/10.4995/YIC2021.2021.12598

is the number of basis functions of polynomial order p needed for the B-spline construction. A B-spline curve $\mathbf{C}(\xi)$ is constructed through control points $\mathbf{B}_i$ using a set of polynomial basis functions $N_i^p(\xi)$ by

$$\mathbf{C}(\xi) = \sum_{i=1}^{n} N_i^p(\xi)\ \mathbf{B}_i\,. \tag{1}$$

The underlying p-th order basis functions $N_i^p(\xi)$ are defined as follows.

$$N_i^0(\xi) = \begin{cases} 1 & \text{if} \quad \xi_i \leq \xi \leq \xi_{i+1} \\ 0 & \text{else} \end{cases} \tag{2a}$$

$$p > 0 : N_i^p(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_i^{p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1}^{p-1}(\xi) \tag{2b}$$

The derivatives of the basis functions are always a combination of lower order basis functions. The k-th derivative of the i-th basis function can be generalized to

$$\frac{d^k}{d^k\xi} N_i^p(\xi) = \frac{p!}{(p-k)!} \sum_{j=0}^{k} \alpha_{k,j}\ N_{i+j}^{p-k}(\xi) \tag{3}$$

with $\alpha_{0,0} = 1,\ \alpha_{k,0} = \frac{\alpha_{k-1,0}}{\xi_{i+p-k+1} - \xi_i},\ \alpha_{k,k} = \frac{-\alpha_{k-1,k-1}}{\xi_{i+p+1} - \xi_{i+k}}\ \alpha_{k,j} = \frac{\alpha_{k-1,j} - \alpha_{k-1,j-1}}{\xi_{i+p+j-k+1} - \xi_{i+j}},\ \ j = 1, ..., k-1.$
Only a small step is necessary to transform the B-splines N_i^p to NURBS R_i^p. Every control point $\mathbf{B}_i = \left[\mathbf{X}_i^T, w_i\right]^T$ contains in addition to its coordinates $\mathbf{X}_i$ also a weight factor w_i. The B-spline basis has to be divided through the weighting function $W(\xi) = \sum_{i=1}^{n} N_i^p(\xi)\ w_i$.

$$R_i^p(\xi) = \frac{N_i^p(\xi)\ w_i}{W(\xi)}\,. \tag{4}$$

For the two-dimensional case the parameter space is spanned by a tensor product of knot vectors $\mathbf{\Xi}_1$ and $\mathbf{\Xi}_2$ in two directions which leads to the shape functions

$$N_I(\xi^1, \xi^2) = R_{ij}(\xi^1, \xi^2) = \frac{N_i^{p_1}(\xi^1) N_j^{p_2}(\xi^2) w_{ij}}{\sum_{\hat{i}=1}^{n_1} \sum_{\hat{j}=1}^{n_2} N_{\hat{i}}^{p_1}(\xi^1) N_{\hat{j}}^{p_2}(\xi^2) w_{\hat{i}\hat{j}}}. \tag{5}$$

Thus, the parametric coordinates of a surface point can be interpolated as

$$\mathbf{X}(\xi^1, \xi^2) = \sum_{I=1}^{n_{en}} N_I(\xi^1, \xi^2) \mathbf{X}_I\,, \tag{6}$$

where $n_{en} = (p_1 + 1)(p_2 + 1)$ is the number of control points per element. To obtain the element formulation using the solution method of Airy, derivatives up to the 4$^{\text{th}}$ order are required. For shape functions of two-dimensional spaces, partial derivatives have to be computed according to [2, pp. 136-138].

3 CONTINUUM MECHANICAL FORMULATION

In a two-dimensional space only displacements in two directions exist. Let $u_1 = u(x, y)$ be the displacement in x-direction and $u_2 = v(x, y)$ the displacement in y-direction. The underlying mechanics for a two-dimensional body in this space are described through the three

main conditions of kinematics, material and equilibrium. The kinematic relations state normal strains and shear strains as derivatives of the displacements

$$\varepsilon_x = \frac{\partial u}{\partial x}, \quad \varepsilon_y = \frac{\partial v}{\partial y}, \quad \gamma_{xy} = \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}. \tag{7}$$

The two mentioned displacements u and v cause normal stresses σ_{xx} and σ_{yy} in the directions x and y and shear stresses τ_{xy} or τ_{yx}. Although we presupposed only displacements in two directions and assume plane stress ($\sigma_{zz}, \tau_{xz}, \tau_{zx} = 0$), normal strains do also appear in the third direction. These correlations between stresses and strains are covered by the following material law for linear elasticity of a two-dimensional body

$$\varepsilon_x = \frac{1}{E}(\sigma_x - \nu\sigma_y), \quad \varepsilon_y = \frac{1}{E}(\sigma_y - \nu\sigma_x), \quad \varepsilon_z = -\frac{\nu}{E}(\sigma_x + \sigma_y) \tag{8a}$$

$$\gamma_{xy} = \frac{2(1+\nu)}{E}\tau_{xy}, \quad \gamma_{xz} = \gamma_{yz} = 0\,, \tag{8b}$$

where Young's modulus E and Poisson ratio ν are the material parameters. For applied loads f_x and f_y, equilibrium is given in x-direction and in y-direction by

$$\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + f_x = 0 \text{ and } \frac{\partial \sigma_{yy}}{\partial y} + \frac{\partial \tau_{yx}}{\partial x} + f_y = 0\,, \tag{9}$$

respectively. In order to ensure a steady and cohesive displacement, also the compatibility condition

$$\frac{\partial^2 \varepsilon_x}{\partial y^2} + \frac{\partial^2 \varepsilon_y}{\partial x^2} - \frac{\partial^2 \gamma_{xy}}{\partial x \partial y} = 0 \tag{10}$$

between the single strain components has to be fulfilled. Combining Eqs. (7) to (10), a biharmonic equation is received. For $f_x = f_y = 0$, we obtain the Airy-equation

$$\Delta\Delta F = 0\,, \tag{11}$$

where F denotes the Airy stress function, a measure without a further physical meaning. However, its second derivatives directly deliver the stresses

$$\sigma_{xx} = \frac{\partial^2 F}{\partial y^2}, \quad \sigma_{yy} = \frac{\partial^2 F}{\partial x^2}, \quad \tau_{xy} = -\frac{\partial^2 F}{\partial x \partial y}. \tag{12}$$

Taking into account the definition of the Laplace operator Δ, Eq. (11) is extended to

$$F_{,xxxx} + 2F_{,xxyy} + F_{,yyyy} = 0 \tag{13}$$

with $F_{,ijkl} = \frac{\partial}{\partial i}\frac{\partial}{\partial j}\frac{\partial}{\partial k}\frac{\partial}{\partial l}F$. This equation marks the starting point for developing the element formulation in the subsequent section.

4 NURBS-BASED ISOGEOMETRIC DISCRETIZATION

We suppose F to be a scalar field $F \in S$ with $S = \{F \in H^4(\Omega) | F = 0 \text{ on } \delta\Omega_D\}$ in the area of Ω. The corresponding test function is chosen as $\delta F \in S$. Eq. (13) is multiplied with the test function which yields three similar integrals of the type $\int_\Omega F_{,ijkl}\ \delta F \mathrm{da}$. Partial integration is applied two times and leads to:

$$\int_\Omega F_{,ijkl}\ \delta F\ \mathrm{da} = \int_\Omega (F_{,ijk}\ \delta F)_{,l}\ \mathrm{da} - \int_\Omega (F_{,ij}\ \delta F_{,l})_{,k}\ \mathrm{da} + \int_\Omega F_{,ij}\ \delta F_{,kl}\ \mathrm{da} \tag{14}$$

Using the divergence theorem for scalar values $\int_\Omega f_{,i}\ \mathrm{da} = \int_{\partial\Omega} f \cdot n_i\ \mathrm{ds}$, the integration space is reduced to the boundaries $\partial\Omega$ and the associated normal vector n_i is multiplied to the integral. For a rectangular domain these boundary integrals are treated separately for each boundary. At x-boundaries $n_x = 1,\ n_y = 0$ is valid and prescribed stresses $\sigma_{xx} = \overline{\sigma}_{xx}$ and $\tau_{xy} = \overline{\tau}_{xy}$ are considered. Equally at y-boundaries $n_x = 0,\ n_y = 1$ is valid and $\sigma_{yy} = \overline{\sigma}_{yy}$ and $\tau_{yx} = \overline{\tau}_{yx}$ are possible prescribed stresses. The divergence theorem is applied to Eq. (14) and the boundary terms are evaluated as mentioned. Thus, the final continuous element formulation is received:

$$\begin{aligned}
&\int_\Omega F_{,xxxx}\ \delta F\ \mathrm{da} + \int_\Omega F_{,yyyy}\ \delta F\ \mathrm{da} + \frac{1}{2}\ \int_\Omega F_{,xx}\ \delta F_{,yy}\ \mathrm{da} \\
&+\frac{1}{2}\ \int_\Omega F_{,yy}\ \delta F_{,xx}\ \mathrm{da} + \frac{1}{2}\ \int_\Omega F_{,xy}\ \delta F_{,yx}\ \mathrm{da} + \frac{1}{2}\ \int_\Omega F_{,yx}\ \delta F_{,xy}\ \mathrm{da} \\
&= \\
&+\frac{1}{2}\ \int_{\partial\Omega_x} \overline{\sigma}_{xx}\ \delta F_{,x}\ \mathrm{ds} + \frac{1}{2}\ \int_{\partial\Omega_y} \overline{\sigma}_{yy}\ \delta F_{,y}\ \mathrm{ds} - \frac{1}{2}\ \int_{\partial\Omega_x} \overline{\tau}_{xy}\ \delta F_{,y}\ \mathrm{ds} \\
&-\frac{1}{2}\ \int_{\partial\Omega_y} \overline{\tau}_{yx}\ \delta F_{,x}\ \mathrm{ds} + \frac{1}{2}\ \int_{\partial\Omega_x} \overline{\tau}_{xy,y}\ \delta F\ \mathrm{ds} + \frac{1}{2}\ \int_{\partial\Omega_y} \overline{\tau}_{yx,x}\ \delta F\ \mathrm{ds}
\end{aligned} \tag{15}$$

4.1 Interpolation of geometry, unknowns and test functions

The geometry is interpolated as in Eq. (6). To solve the continuum formulation given in Eq. (15), next to geometry also all other expressions in the function space are discretized as an interpolation over all control points. The approximations F^h and δF^h of the unknown Airy function and the corresponding test function are interpolated as follows:

$$F^h = \sum_{I=1}^{n_{np}} N_I\ F_I, \quad \delta F^h = \sum_{J=1}^{n_{np}} N_J\ \delta F_J \tag{16}$$

Eq. (15) requires up to the fourth derivative of the Airy function and up to the second derivative of the test function. They are easily derived from Eq. (16) since F_I and ∂F_I are discrete values. This leads to the following general derivatives with respect to global coordinates (x, y):

$$F^h_{,i} = \sum_{I=1}^{n_{np}} N_{I,i}\ F_I, \quad F^h_{,ij} = \sum_{I=1}^{n_{np}} N_{I,ij}\ F_I, \quad F^h_{,ijk} = \sum_{I=1}^{n_{np}} N_{I,ijk}\ F_I, \quad F^h_{,ijkl} = \sum_{I=1}^{n_{np}} N_{I,ijkl}\ F_I \tag{17}$$

$$\delta F^h_{,i} = \sum_{J=1}^{n_{np}} N_{J,i}\ \delta F_J, \quad \delta F^h_{,ij} = \sum_{J=1}^{n_{np}} N_{J,ij}\ \delta F_J \tag{18}$$

Later on, integration will be performed as Gaussian quadrature on a bi-unit parent element using classical change of variables formulation taking into account the Jacobian determinant. Computation rules of NURBS shape functions and their derivatives in Sec. 2 are based on the (ξ^1, ξ^2) coordinate system as needed in the integration process. Thus a transformation rule for the first, second and fourth derivatives of shape functions appearing on the left side in Eq. (15) is required between (x, y) and (ξ^1, ξ^2) coordinate system. As shown in [3] it is useful to apply the chain rule

$$\frac{\partial N}{\partial x_i} = \frac{\partial N}{\partial \xi^\alpha}\frac{\partial \xi^\alpha}{\partial x_i}\,. \tag{19a}$$

The gradient of this mapping $\frac{\partial x_i}{\partial \xi^\alpha}$ is computed as part of the Jacobian. To keep it simple yet exact for our purposes which only include rectangular domains, higher order derivatives are calculated as follows:

$$\frac{\partial^2 N}{\partial x_i\ \partial x_j} = \frac{\partial^2 N}{\partial \xi^\alpha\ \partial \xi^\beta}\frac{\partial \xi^\alpha}{\partial x_i}\frac{\partial \xi^\beta}{\partial x_j} \tag{19b}$$

$$\frac{\partial^4 N}{\partial x_i\,\partial x_j\,\partial x_k\,\partial x_l} = \frac{\partial^4 N}{\partial \xi^\alpha\,\partial \xi^\beta\,\partial \xi^\gamma\,\partial \xi^\delta}\frac{\partial \xi^\alpha}{\partial x_i}\frac{\partial \xi^\beta}{\partial x_j}\frac{\partial \xi^\gamma}{\partial x_k}\frac{\partial \xi^\delta}{\partial x_l} \tag{19c}$$

4.2 Formulation of the condition matrix and final system of equations

Implementing the discretizations and moving the discrete values of Airy and test function out of the integrals in Eq. (15) leads to

$$\begin{aligned}
&\sum_I^{n_{np}}\sum_J^{n_{np}}\Bigg[\Bigg(\int_\Omega N_{I,xxxx}\;N_J\;\mathrm{da}+\int_\Omega N_{I,yyyy}\;N_J\;\mathrm{da}+\frac{1}{2}\int_\Omega N_{I,xx}\;N_{J,yy}\;\mathrm{da}\\
&+\frac{1}{2}\int_\Omega N_{I,yy}\;N_{J,xx}\;\mathrm{da}+\frac{1}{2}\int_\Omega N_{I,xy}\;N_{J,yx}\;\mathrm{da}+\frac{1}{2}\int_\Omega N_{I,yx}\;N_{J,xx}\;\mathrm{da}\Bigg)\;F_I\;\delta F_J\Bigg]\\
&=\sum_J^{n_{np}}\Bigg[\Bigg(\frac{1}{2}\int_{\partial\Omega_x}\overline{\sigma}_{xx}\;N_{J,x}\;\mathrm{ds}+\frac{1}{2}\int_{\partial\Omega_y}\overline{\sigma}_{yy}\;N_{J,y}\;\mathrm{ds}-\frac{1}{2}\int_{\partial\Omega_x}\overline{\tau}_{xy}\;N_{J,y}\;\mathrm{ds}-\frac{1}{2}\int_{\partial\Omega_y}\overline{\tau}_{yx}\;N_{J,x}\;\mathrm{ds}\\
&\qquad+\int_{\partial\Omega_x}\overline{\tau}_{xy,y}\;N_J\;\mathrm{ds}+\int_{\partial\Omega_y}\overline{\tau}_{yx,x}\;N_J\;\mathrm{ds}\Bigg)\;\delta F_J\Bigg].
\end{aligned} \tag{20}$$

The left side of Eq. (20) represents the system "stiffness" $\mathbf{B}$ and as in most cases the right side includes the loading. To avoid mistakes in the usage of F as Airy stress function, the loads are marked as $\mathbf{L}$. A loop over all n_{el} elements will sum up the individual parts B^e_{IJ} of the condition matrix and load components L^e_J. This changes Eq. (20) to

$$\bigcup_{e=1}^{n_{el}}\sum_{I=1}^{n_{nen}}\sum_{J=1}^{n_{nen}} B^e_{IJ}\;F_I\;\delta F_J=\bigcup_{e=1}^{n_{el}}\sum_{J=1}^{n_{nen}} L^e_J\;\delta F_J. \tag{21}$$

For an efficient computation of results, Eq. (21) has to be brought into a system of equations. Therefore, the element condition matrix and the load vector can be arranged as

$$\mathbf{B}^e=\begin{bmatrix} B^e_{11} & \dots & B^e_{1n_{en}}\\ \vdots & \ddots & \vdots\\ B^e_{n_{en}1} & \dots & B^e_{n_{en}n_{en}}\end{bmatrix},\quad \mathbf{L}^e=\begin{bmatrix} L^e_1\\ \vdots\\ L^e_{n_{en}}\end{bmatrix}. \tag{22}$$

To solve the global system of equations all element matrices and load vectors have to be mapped to a global condition matrix $\mathbf{B}$ and the global load vector $\mathbf{L}$, respectively. In addition to that, the discrete solution values are arranged in a vector $\hat{\mathbf{F}}=\left[F_1,\dots,F_{n_{np}}\right]^T$, where n_{np} is the number of global degrees of freedom. After converting Eq. (21) such that the test function is dropped, the final system of equations

$$\mathbf{B}\;\hat{\mathbf{F}}=\mathbf{L} \tag{23}$$

can be solved for $\hat{\mathbf{F}}$ with standard routines.

4.3 Treatment of boundary conditions

The right side of the derived element formulation in Eq. (20) involves already prescribed stresses. These stresses are handed in the process of computation as specified neumann boundary conditions. They can be treated easily as they are well known. But it is much more of interest how to prescribe dirichlet boundary conditions. They do not directly enter the formulatiom, but are necessary to receive full rank for the condition matrix. Without them a too wide

solution space of possible stress functions, leading all to the correct stress state, is spanned. As in all FEM formulations dirichlet boundary conditions are imposed for the unknown degrees of freedom. For the presented element formulation this means some values for F have to be prescribed without knowing the correct mechanical interpretation. Since within this work only stress boundary conditions are treated, any dirichlet conditions can be set which yields a stable computation. It is important that there are enough, yet not too many independent prescribed values in order to obtain a stable solution. As in standard two-dimensional plane stress FEM, it is required to set at least three nodal dirichlet conditions in order to prevent the three rigid body modes. More boundary conditions should not be set in order to not overconstrain the solution. As displacement boundary conditions are usually the dirichlet conditions in standard FEM formulations, it has to be stated out that displacements are not concerned and this procedure does only guarantee accuracy for the stress state.

5 NUMERICAL EXAMPLES

For the following numerical examples, the derived element formulation was embedded in a matlab working routine for IGA. The expressions $\overline{\tau}_{xy,y}$ and $\overline{\tau}_{yx,x}$ are neglected since the examples include only constant or even no shear. The formulation is tested on a quadratic plate with the dimensions 2×2. As dirichlet condition, F was fixed in both bottom corners and the upper left corner of the plate.

5.1 Quadratic plate under linearly varying uniaxial tension

A linearly varying tension along the vertical boundaries is applied in x-direction. At the upper boundary a positive and at the lower boundary a negative stress of 10 is prescribed. The exact solution for σ_{xx} is constant in x-direction and varies linearly in y-direction as the

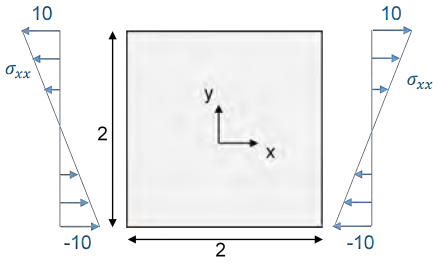


Figure 1: System sketch for quadratic plate under uniaxial tension

prescribed stresses. σ_{yy} and τ_{xy} are equal to zero. The stress results using the developed element formulation are provided in Figs. 2 to 4.

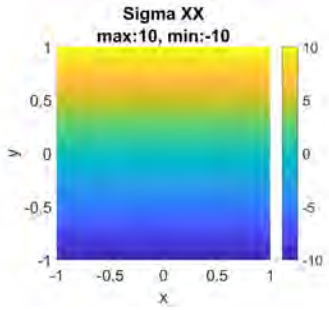


Figure 2: Stress σ_{xx} for quadratic plate under uniaxial tension

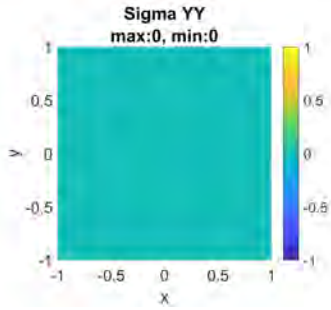


Figure 3: Stress σ_{yy} for quadratic plate under uniaxial tension

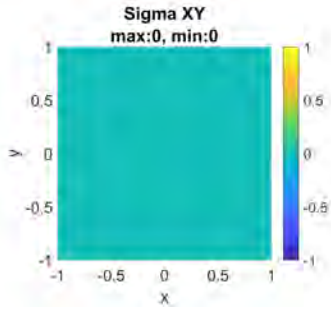


Figure 4: Stress τ_{xy} for quadratic plate under uniaxial tension

The mentioned exact analytic stress distribution can be found in all three stress plots. The values of τ_{xy} and σ_{yy} are actually numerical zero, only small computing errors in the range of the employed numerical precision appear. All discretizations yield the exact solution. This was also tested by the L_2-error norm. A negligible error in the range of 10^{-14} appeared for all refinement steps. To fulfill the condition of a full rank matrix three nodal dirichlet conditions have to be set.

5.2 Quadratic plate under pure shear

For the example of pure shear only shear stresses were prescribed at every boundary. As it is shown in the system sketch in Fig. 5, the unit square is constantly loaded with shear stresses of 10. The exact solution for τ_{xy} is constantly 10 over the whole plate while σ_{xx} and σ_{yy} are

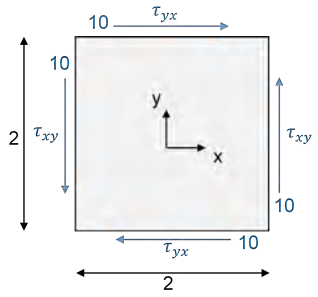


Figure 5: System sketch for quadratic plate under pure shear

equal to zero. For all three stresses the results are shown in Figs. 6 to 8. In the case of pure

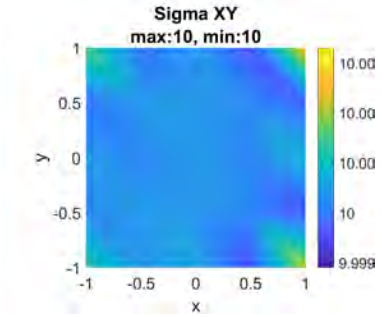


Figure 6: Stress τ_{xy} for quadratic plate under pure shear

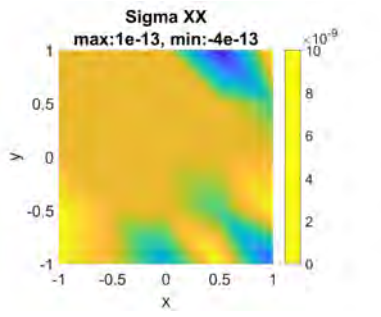


Figure 7: Stress σ_{xx} for quadratic plate under pure shear

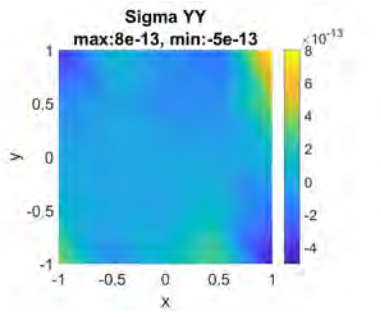


Figure 8: Stress σ_{yy} for quadratic plate under pure shear

shear the L_2-error norm is in the range of 10^{-13}, see Figs. 6 to 8. Under pure shear, as shown before under uniaxial tension, the exact solution can be obtained using just one element.

5.3 Quadratic plate under complex loads

A rectangular plate with the dimensions $-b \leq x \leq b$ and $-a \leq y \leq a$ is subjected to a complex load, which is derived from a chosen load function. This procedure is analogous to the method of manufactured solutions [8]. Here, a solution F is chosen, which is a simplified

version of the Fourier solution

$$\begin{aligned} F = &\sum_{n=1}^{\infty} \cos\beta_n x \left[B_n \cosh\beta_n y + C_n \beta_n y \sinh\beta_n y\right] \\ &+ \sum_{m=1}^{\infty} \cos\alpha_m y \left[D_m \cosh\alpha_m x + E_m \alpha_m x \sinh\alpha_m x\right] \end{aligned} \tag{24}$$

for uniaxial loading in y-direction with arbitrary shape but opposite sign on top and bottom, which is provided, e.g., in [9, pp. 175-178]. Since the complete Fourier solution is hard to compute with high precision, we simply use the solution with one Fourier term only. The stresses at the boundary are computed from this truncated solution and are applied as loading at the boundaries. Thus, we have loading functions at the boundary and at the same time we know the exact stress solution in every point within the domain. Since the occurring trigonometric functions and hyperbolic functions cannot be exactly described by NURBS basis functions, it is possible to study the convergence rates of the formulation.

The Airy function for this example is

$$\begin{aligned} F = &\cos\beta x \left[B \cosh\beta y + C\beta y \sinh\beta y\right] \\ &+ \cos\alpha y \left[D \cosh\alpha x + E\alpha x \sinh\alpha x\right] , \end{aligned} \tag{25}$$

where the constants are chosen in a way that the shear stresses τ_{xy} are zero along all boundaries. This yields $\alpha = \pi/a$, $\beta = \pi/b$,

$$D = -E(1 + \alpha b \coth\frac{\pi b}{a}) \qquad \text{and} \qquad B = -C(1 + \beta a \coth\frac{\pi a}{b}) . \tag{26}$$

Introducing Eq. (26) into Eq. (25) and using the required derivatives, the exact stresses are given by

$$\begin{aligned} \sigma_{xx} = &\beta^2 C \cos\beta x \left[\cosh\beta y + \beta y \sinh\beta y - \beta a \coth\frac{\pi a}{b} \cosh\beta y\right] \\ &- \alpha^2 E \cos\alpha y \left[\alpha x \sinh\alpha x - \cosh\alpha x - \alpha b \coth\frac{\pi b}{a} \cosh\alpha x\right] \\ \sigma_{yy} = &-\beta^2 C \cos\beta x \left[-\cosh\beta y + \beta y \sinh\beta y - \beta a \coth\frac{\pi a}{b} \cosh\beta y\right] \\ &+ \alpha^2 E \cos\alpha y \left[\alpha x \sinh\alpha x + \cosh\alpha x - \alpha b \coth\frac{\pi b}{a} \cosh\alpha x\right] \\ \tau_{xy} = &\beta^2 C \sin\beta x \left[\beta y \cosh\beta y - \beta a \coth\frac{\pi a}{b} \sinh\beta y\right] \\ &+ \alpha^2 E \sin\alpha y \left[\alpha x \cosh\alpha x - \alpha b \coth\frac{\pi b}{a} \sinh\alpha x\right] \end{aligned} \tag{27}$$

and can directly be used for the validation of the presented element formulation.

For the chosen domain it holds $a = 1$ and $b = 1$. The constants C and E, which basically govern the size of the load at the boundaries, are chosen to be $C = 10$ and $E = 1$. The resulting (Neumann) boundary condition are

$$\begin{aligned} \bar{\sigma}_{xx} = &\mp 10\beta^2 \left[\cosh\beta y + \beta y \sinh\beta y - \pi \coth\pi \cosh\beta y\right] \\ &\pm \alpha^2 \cos\alpha y \left[\pi \sinh\pi - \cosh\pi - \pi \coth\pi \cosh\pi\right] \end{aligned} \tag{28}$$

at $x = \pm b$ and

$$\begin{aligned} \bar{\sigma}_{yy} = &\pm 10\beta^2 \cos\beta x \left[-\cosh\pi + \pi \sinh\pi - \pi \coth\pi \cosh\pi\right] \\ &\pm \alpha^2 \left[\alpha x \sinh\alpha x + \cosh\alpha x - \pi \coth\pi \cosh\alpha x\right] \end{aligned} \tag{29}$$

at $y = \pm a$. At all boundaries $\bar{\tau}_{xy} = 0$ holds.

The following results in Figs. 9 to 11 are obtained from NURBS of order $p = 8$ using 50×50 elements within the patch. The error of the plotted stresses is in the range of 10^{-6}.

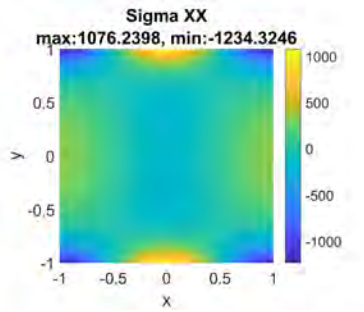


Figure 9: Stress σ_{xx} for quadratic plate under complex load

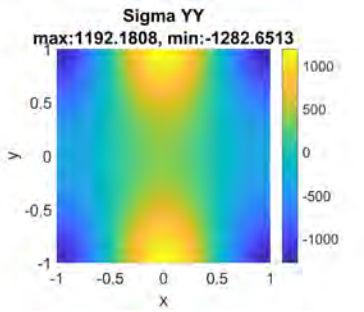


Figure 10: Stress σ_{yy} for quadratic plate under complex load

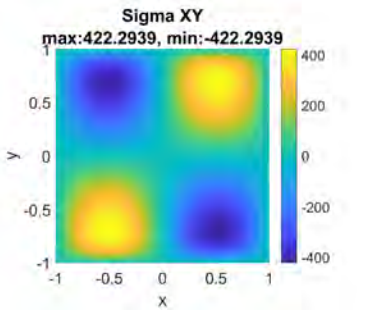


Figure 11: Stress τ_{xy} for quadratic plate under complex load

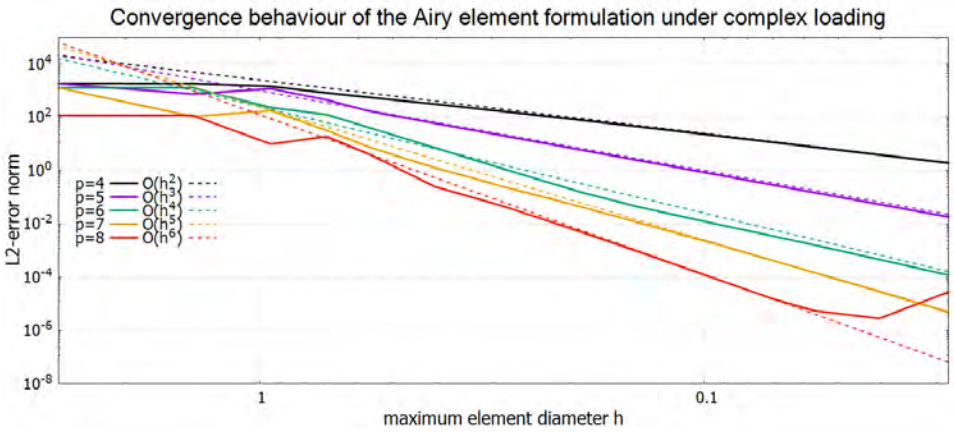


Figure 12: Convergence behaviour for the quadratic plate under complex load using orders $p = 4, .., 8$ for the discretization. The slopes can be seen to be around $p - 2$ for each curve.

Using the plot of the L_2-error norm in Fig. 12, the convergence behaviour of the proposed element formulation is assessed. To receive maximum C^{p-1} continuity, k-refinement is used. After some refinement steps a proper convergence behavior is obtained. All considered orders converge to the exact solution. The slope of the L_2 error norm of the stresses for a computation using basis functions of order p can be seen to be approximately equal to $p - 2$.

6 CONCLUSION

Within this work a one-degree of freedom formulation for two-dimensional linear elasticity problems has been obtained. It is based on the Airy equation and allows to compute stresses as direct solution of the underlying system of equations. Situations where only stress boundary conditions are relevant can be computed without defining suitable displacement boundary conditions. Thus, computations of that kind are strongly simplified. Possible applications could be the analysis on the micro level in an FE^2 formulation [10]. However, for that kind of applications it might be required to extend the formulation to handle non-rectangular meshes, which will be subject of future research. Within the described method the need for higher derivatives yields a more costly computation of the condition matrix, but the effort for the solution of the global system of equations will be significantly lower since just one degree of freedom per control point is required. Thus, the method can be competitive to a standard displacement-based

two-dimensional elasticity formulation. Future research will try to quantify this exactly. The presented formulation yields the exact solution in the first two simple examples as expected and shows proper convergence behavior in the third example with a very complex state of stress. Future work will focus on a mathematical proof of the convergence rates, on the imposition of displacement boundary conditions and the computation of arbitrarily shaped patches.

REFERENCES

[1] T. J. R. Hughes, J. A. Cottrell, and Y. Bazilevs, "Isogeometric analysis: Cad, finite elements, nurbs, exact geometry and mesh refinement," *Comput. Methods Appl. Mech. Engrg.*, vol. 194, pp. 4135–4195, 2005.

[2] L. Piegl and W. Tiller, *The NURBS book*, 2nd ed., ser. Monographs in visual communications. Berlin: Springer, 1997.

[3] J. A. Cottrell, T. J. R. Hughes, and Y. Bazilevs, *Isogeometric analysis: Toward integration of CAD and FEA*. Chichester: Wiley, 2009.

[4] T. J. R. Hughes, A. Reali, and G. Sangalli, "Efficient quadrature for nurbs-based isogeometric analysis," *Comput. Methods Appl. Mech. Engrg.*, vol. 199, pp. 301–313, 2010.

[5] R. R. Hiemstra, F. Calabrò, D. Schillinger, and T. J. Hughes, "Optimal and reduced quadrature rules for tensor product and hierarchically refined splines in isogeometric analysis," *Comput. Methods Appl. Mech. Engrg.*, vol. 316, pp. 966–1004, 2017.

[6] K. A. Johannessen, "Optimal quadrature for univariate and tensor product splines," *Comput. Methods Appl. Mech. Engrg.*, vol. 316, pp. 84–99, 2017.

[7] W. Dornisch and Y. Sikang, "Effiziente integrationsmethoden für isogeometrische schalenelemente," in *Berichte der Fachtagung Baustatik – Baupraxis 14*, M. Bischoff, M. von Scheven, and B. Oesterle, Eds. Stuttgart: Institut für Baustatik und Baudynamik, Universität Stuttgart, 2020, pp. 415–422.

[8] M. H. Gfrerer and M. Schanz, "Code verification examples based on the method of manufactured solutions for kirchhoff–love and reissner–mindlin shell analysis," *Eng. Comput.*, vol. 34, pp. 775–785, 2018.

[9] M. H. Sadd, *Elasticity: Theory, applications, and numerics*, 3rd ed. Amsterdam and Boston: Elsevier/Academic Press, 2014. [Online]. Available: http://www.sciencedirect.com/science/book/9780124081369

[10] S. Klarmann, F. Gruttmann, and S. Klinkel, "Homogenization assumptions for coupled multiscale analysis of structural elements: beam kinematics," *Comput. Mech.*, vol. 65, pp. 635–661, 2020.

Random vibration fatigue analysis with the method of Isogeometric Analyses (IGA)

Shubiao WANG, Leila KHALIJ, Renata TROIAN

Normandy universit, Laboratory of Mechanics of Normandy (LMN)
INSA Rouen Normandy, France
email: shubiao.wang@insa-rouen.fr

Key words: Isogeometric analysis; Finite element method; random vibration fatigue analysis

Abstract: *At present, Finite Element Analysis (FEA) is indispensable in the field of simulation technology, as this kind of numerical analysis method can help engineers to predict results difficult to obtain from experimental tests. However, the mesh generation process in FEA is time-consuming. It is estimated that about 80 percent of analysis time is devoted to mesh generation in some fields, such as automotive or shipbuilding industries. On the other hand, the imperfections of mesh models can lead to inaccurate results. In this study, we adopted a new numerical analysis method, Isogeometric Analysis (IGA) to develop a random vibration fatigue analysis on a wind turbine tower model. From the mesh generation process, it can be observed that the NURBS mesh creation is far more convenient and time-efficient than the finite element counterparts. From fatigue analysis results, we can conclude that IGA can predict fatigue damage using fewer mesh elements and integration points, corresponding very well with the finite element results.*

1 Introduction

It is necessary to predict the fatigue life of a structure during the design stage. In the numerical simulation, the fatigue analysis can be developed both in the time and frequency domain. However, compared with frequency domain fatigue analysis, the time domain fatigue analysis is computationally expensive. So, in this studying, we adopted the frequency domain fatigue analysis method to calculate the cumulative damage ratio based on Dirlik's approach, in which the input random vibration load and output stress are described by Power Spectrum Density functions (PSD).

At present, there are several disadvantages to FEA. The most significant one is to spend a long time in mesh generation. For example, it is estimated that about 80% of overall analysis time has been applied to the mesh creation process in automotive, aerospace, and shipbuilding industries [1]. In 2005, T.J.R. Hughes proposed a method, which is named Isogeometric analysis (IGA) to mainly solve the problems derived from the classical FEA.

IGA with NURBS basis function has been applied in various engineering problems, including contact mechanics [2, 3, 4], fluid mechanics [5, 6, 7], structural optimization [8, 9, 10, 11], shell analysis [12, 13, 14, 15], beam analysis [20, 16, 17] damage and fracture mechanics [18, 19], and structural vibration analysis [16, 20, 21], etc. In this paper, we mainly investigate the performance of the NURBS-based IGA LS-DYNA on a wind turbine tower model. Results are verified by classical FEA and matlab code.

The originality of this paper is that the isogeometric random vibration fatigue analysis is firstly employed on an industrial model. The structure of this article is as follows. In section 2, we briefly review some theoretical backgrounds. In section 3, isogeometric random vibration

https://doi.org/10.4995/YIC2021.2021.13262

fatigue analysis is applied on a wind turbine tower model and the results are verified by the FEA and own developed Matlab programming. In section 4, we conclude on the present studies.

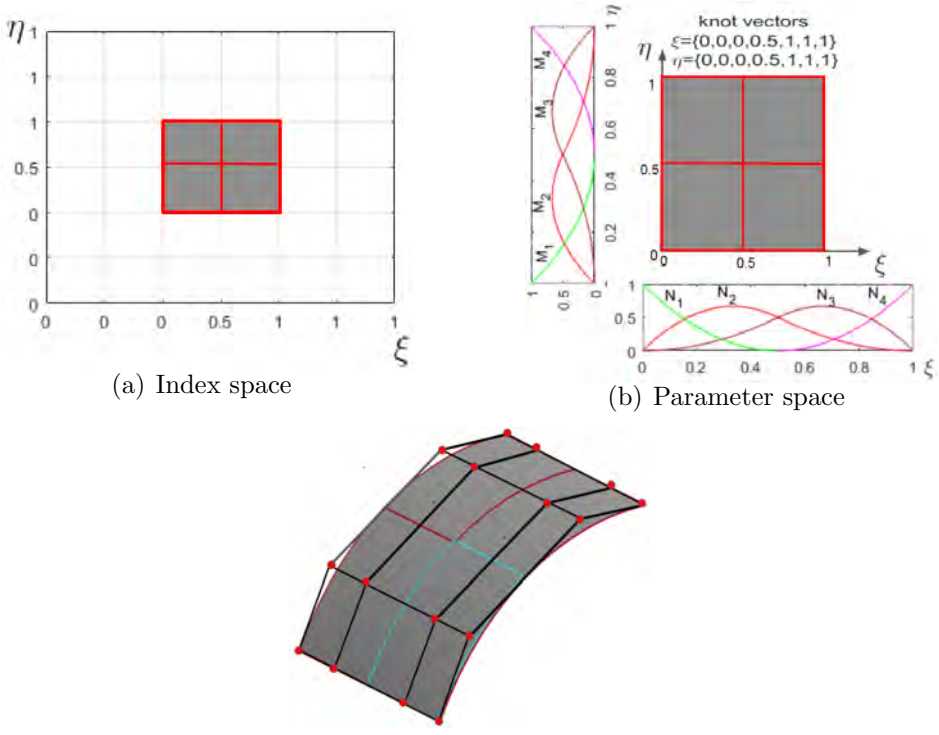
2 IGA modelling

At present, most IGA is developed, based on NURBS basis function, as it not only has a wide application in CAD systems but also possesses powerful capability in describing complex geometric models. The NURBS basis functions are defined by the B-spline basis function built from knot vectors. Details can be found in [1].

2.1 Some basic concepts of IGA

- I) Different spaces

The index space in two dimensions is an equally divided domain, no matter with the knot values of knot vectors. For example, in the case where the knot vectors are respectively $\Xi = \{0, 0, 0, 0.5, 1, 1, 1\}$ and $\eta = \{0, 0, 0, 0.5, 1, 1, 1\}$, the index space ranges from $[0, 1]$ (figure 1 (a)). The parameter space in two dimensions is the $[0, 1] \otimes [0, 1]$ domain where the NURBS basis functions are defined (figure 1 (b)). And the control points, physical mesh, and control mesh are defined in physical space (figure 1 (c)).



(a) Index space

(b) Parameter space

(c) Physical space

Figure 1: Schematic illustration of different spaces

- II) Knot vector

A knot vector in one dimension is defined as a series of non-decreasing coordinates in the parametric space, denoted by $\Xi = \{\xi_1, \xi_2, ..., \xi_{n+p+1}\}$, where $\xi_i \in R$ is the ith knot (or coordinate), and i is the knot index from $1, 2, \ldots, n + p + 1$, in which n is the number of B-spline

basis function along ξ parametric direction, and p is the polynomial order of B-spline basis function. In the construction of B-spline surface and solid, it is necessary to use 2 and 3-knot vectors, which are respectively directed along ξ and η directions. Each knot or coordinate of a knot vector is used to divide the parametric space of a geometrical model to obtain elements, meaning that all of the mesh elements can be selected by knot values of the knot vectors. In terms of the space between different knots, a knot vector can be referred to as a uniform or non-uniform knot vector. In a uniform knot vector, the knots are equally spaced in the parametric space, such as $\Xi = \{1, 2, 3, \ldots, \xi_{n+p+1}\}$. Similarly, in a non-uniform knot vector, the knots are unequally spaced in the parametric space, such as $\Xi = \{1, 1.5, 2.5, 3, \ldots, \xi_{n+p+1}\}$. In a knot vector, there can be repeated knots, and a knot vector is said to be open if its first and last knots repetition are equal to the $p + 1$, in which p is the polynomial order of the basis function. In one dimension, the basis functions constructed by an open knot vector interpolate the ends of parametric space.

- III) B-spline basis function and B-spline curve

The B-spline basis functions are defined by the following equation 1 and 2.
For $p = 0$, it is defined by:

$$N_{i,0}(\xi) = \begin{cases} 1 & if \quad \xi_i <\leqslant \xi < \xi_{i+1} \\ 0 & otherwise \end{cases} \tag{1}$$

For $p = 1$, 2, 3,..., they are defined by

$$N_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\xi) \tag{2}$$

B-spline curves are defined by the linear combination of B-spline basis functions and the corresponding control points, the vector - valued coefficients of the basis function $\mathbf{P}_i \in R$, $i = 1, 2, ..., n$, as in equation 3.

$$C(\xi) = \sum_{i=1}^{n} N_{i,p}(\xi)\mathbf{P}_i \tag{3}$$

- Non-Uniform Rational B-Spline (NURBS) basis function, NURBS curve, and NURBS surface

The univariate NURBS basis function is described by the rationale of weighted B-spline basis functions as:

$$R_{i,p}(\xi) = \frac{\omega_i N_{i,p}(\xi)}{W(\xi)} = \frac{\omega_i N_{i,p}(\xi)}{\sum_{i=1}^{n_{cp}} \omega_i N_{i,p}(\xi)} \quad 1 \leq i \leq p+1 \tag{4}$$

Where ω_i denotes the weight value of the control point $\mathbf{P}_i$, and $W(\xi)$ is the weighted linear combination of B-spline basis functions. Here, n denotes the total number of NURBS control points. The NURBS curve is defined by the linear combination of univariate NURBS basis function $R_{i,p}(\xi)$ and control point $\mathbf{P}_i$ by the following expression [1]:

$$C(\xi) = \sum_{i=1}^{n} R_{i,p}(\xi)\mathbf{P}_i \tag{5}$$

And the NURBS surafce is defined by:

$$C(\xi,\eta)=\sum_{i=1}^{n}\sum_{j=1}^{m}R_{i,j}^{p,q}(\xi,\eta)\mathbf{P}_{i,j} \tag{6}$$

where $R_{i,j}^{p,q}(\xi,\eta)$ is bivariate NURBS basis functions, which are defined by:

$$R_{i,j}^{p,q}(\xi,\eta)=\frac{N_{i,p}(\xi)M_{j,q}(\eta)w_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{m}N_{i,p}(\xi)M_{j,q}(\eta)w_{i,j}} \tag{7}$$

where $N_{i,p}(\xi)$ and $M_{j,q}(\xi)$ are pth and qth order B-spline basis function, which are defined in ξ and η parametric directions, respectively.

3 The IGA and FEA on a wind turbine tower model

In this section, isogeometric and finite element random vibration fatigue analysis are developed on a wind turbine tower model created based on the reference[22].

3.1 The analysis preparation

3.1.1 The geometric model, material properties

As shown in figure 2, the wind turbine tower model is assembled by a series of different thickness cylinders and conical shell sections, in which the geometry parameters like the height, thickness, etc are displayed in the form of mm. The tower model consists of 3 flange connections, whose base, middle and top flange thicknesses are respectively 300, 200 and 200 mm.

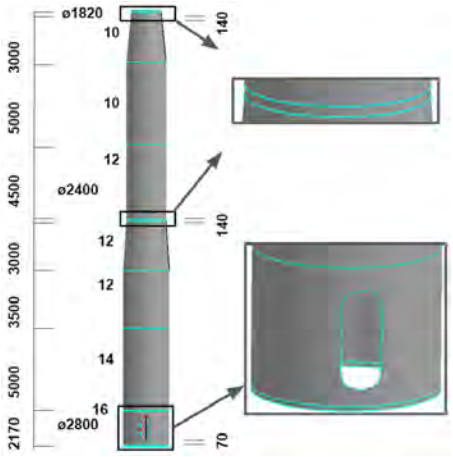


Figure 2: geometry model of the tower

The material properties are shown in table 1. And the material constants of the S-N curve are respectively $\beta=9.82$ and $C=4.0641\times10^{88}$ [23].

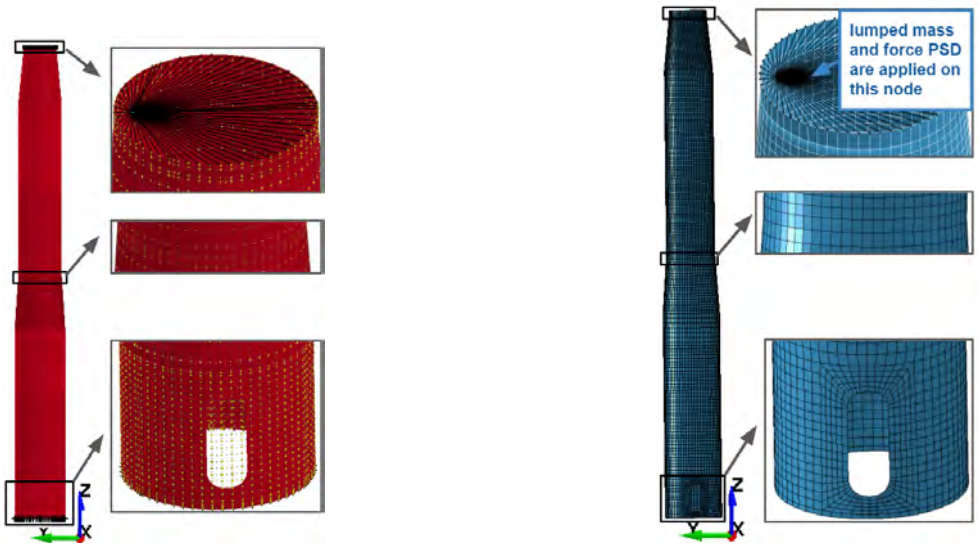
Table 1: Material properties

Mass density	Young's modulus	poisson's ratio
3.81e-3 g/mm^3	3.1e+11 Pa	0.33

3.1.2 Mesh models and boundary condition

The isogeometric and finite element mesh models are presented in figure 3, in which the number of control points and nodes are respectively 7639 and 12969. The finite element mesh model is created by quadrilateral 4 nodes mesh elements, and the shell element formulation of Belytschko-Tsay is chosen to develop fatigue analysis. For IGA, we used the isogeometric NURBS element, and adopted Hughes-Liu with rotational DOFs shell formulation; the polynomial order of univariate shape functions in s and r-directions in the parametric space are respectively 2, and in LS-DYNA, the mesh refinement method, SUBDIVISION, is used to create more isogeometric mesh elements. After mesh generation on each section, the keyword, NODE DUPLICATION, is used to merge duplicate control points (nodes for FEA) to assemble the different sections.

To simulate the weight effects of blades, turbines, and other parts on the top of the wind turbine tower, at the height of Z = 26460 and X= -750, Y = 0 mm, a node is created to substitute the concentrated mass element of 4.023e+7 g. Then the node is connected with all control points of the top flange edge, and the weight direction is set to in negative z-direction. During analysis, the base flange of the tower model is clamped in the translational and rotational local x, y, z-directions.



(a) (b)

Figure 3: Mesh models (a) IGA (b) FEA

3.2 Analysis results

3.2.1 Modal analysis results: the first five natural frequencies and vibration mode

Table 2, figure 4 and 5 respectively show the first five natural frequencies and corresponding vibration modes obtained from IGA and FEA, from which it can be observed that the frequencies and the vibration modes have a good agreement.

Table 2: The first five natural frequencies(Hz)

Method	1	2	3	4	5
IGA	4.47	4.55	26.83	27.26	30.47
FEA	4.47	4.54	27.15	27.21	30.48

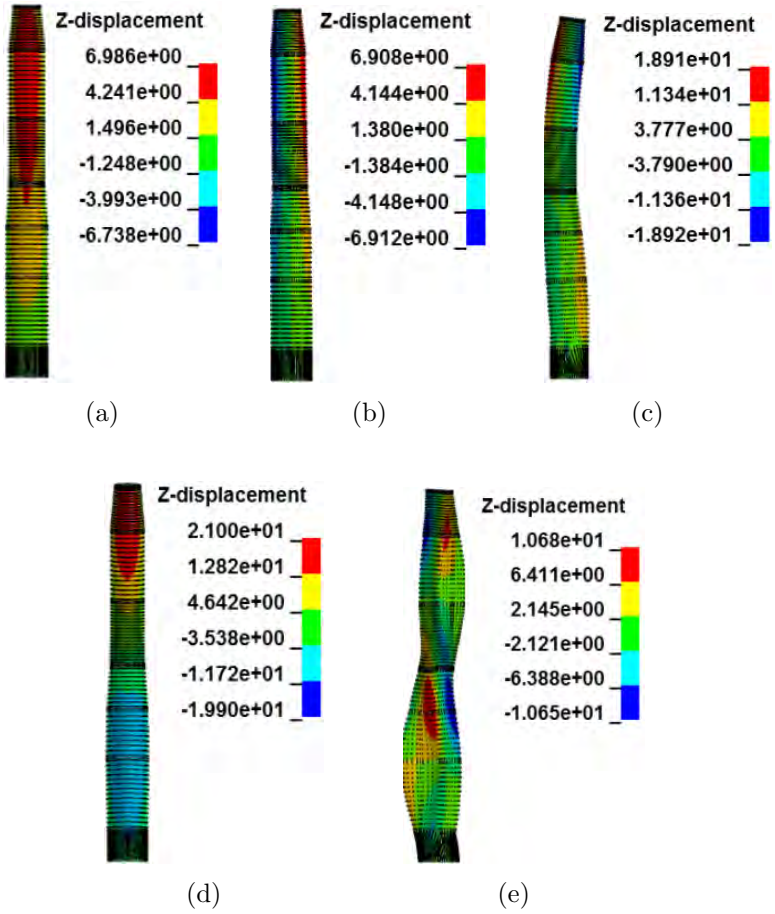
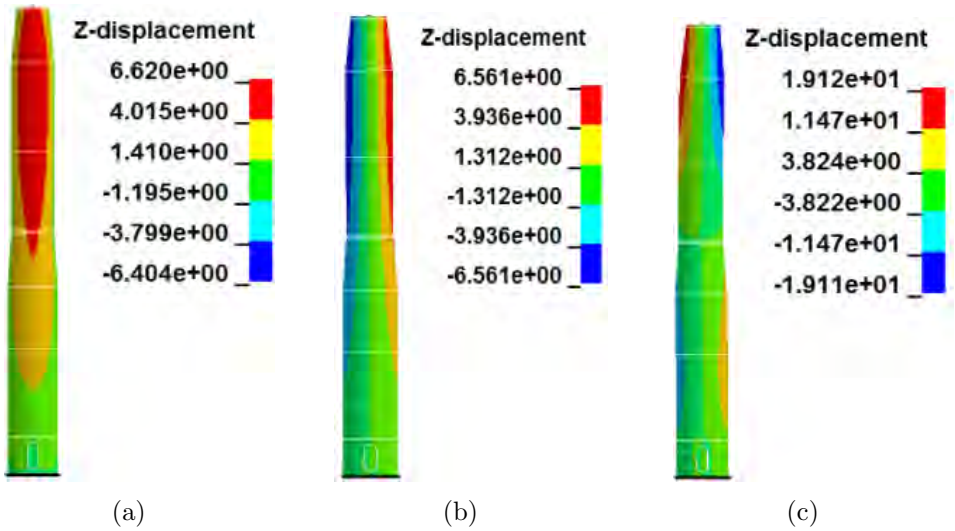
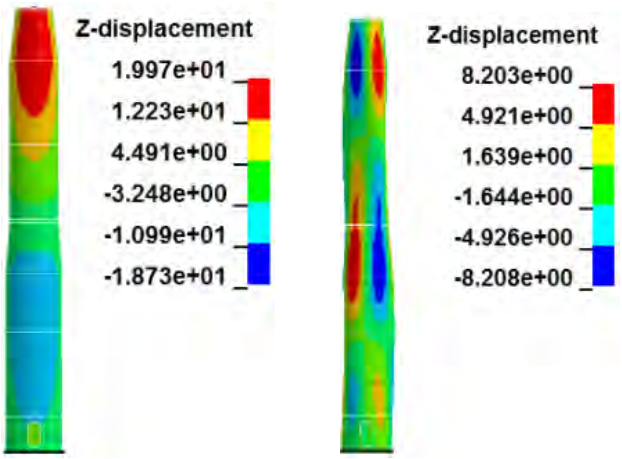


Figure 4: Isogeometric first five vibration mode





(d) (e)

Figure 5: Finite element first five vibration mode

3.2.2 Fatigue analysis results: effective stress PSD, RMS and cumulative damage ratio

The force PSD load, as shown in figure 6 is applied on the node substituting the element concentrated mass in the x-direction. The random vibration fatigue analysis of unit second, in which the damping ratio is set to 0.01, is developed to calculate the effective stress PSD, RMS, and cumulative damage ratio in Ls Dyna. Then based on obtained PSD, the cumulative damage ratio is validated in Matlab using Matlab program.

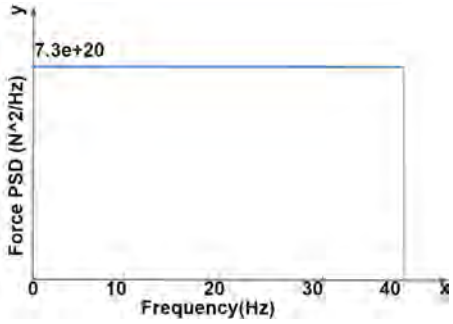


Figure 6: Applied load PSD

Figure 7, and 8 show the calculated isogeometric and finite element effective stress PSD and RMS, in which only the first natural frequency is excited by the applied force PSD. It is observed that isogeometric and finite element PSD and RMS display a good agreement, in which the maximum effective stress RMS from IGA and FEA is 3.151e+8 and 3.125e+8 pa respectively, leading to the relative error of 0.83%, based on the equation 8. From figure 9, it can be seen that the obtained isogeometric and finite element cumulative damage ratios are respectively 2.678e-4 and 2.638e-4, with a relative error of 1.52%, and the maximum damage ratios are located on similar elements close to the door edge. According to the equation 9, the expected isogeometric and finite element fatigue life $E[T_f]$ are 3.7341e+04, and 3.7908e+04 seconds respectively. Based on the Matlab program, the isogeometric and finite element damage ratios are respectively 2.6204e-04 and 2.6406e-04, which are in a good accordance with the damage ratios computed from Ls Dyna.

$$Relative\ \ error = \frac{IGAresult - FEAresult}{FEAresult} \tag{8}$$

$$E[T_f] = \frac{T}{E[D]} \tag{9}$$

Where T is the duration time (1 seconds in these analyses), $E[D]$ is the obtained cumulative damage ratio.

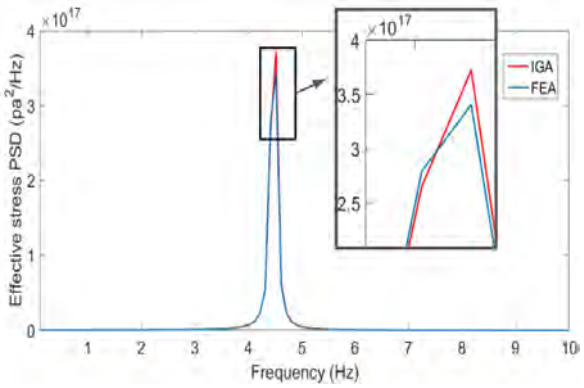
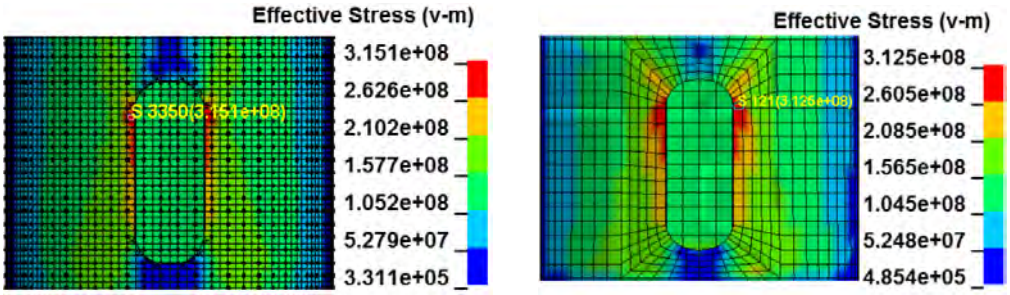


Figure 7: The effective stress PSD



(a) IGA (b) FEA

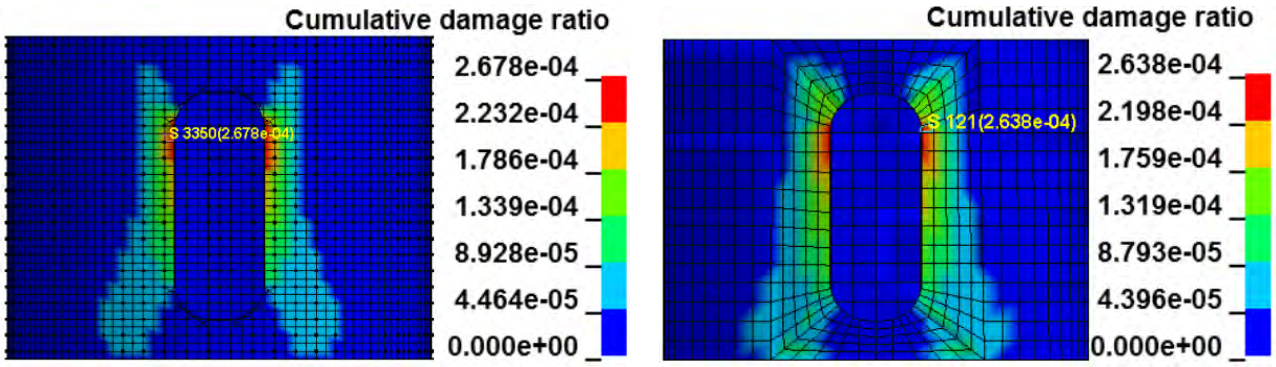
Figure 8: The effective stress RMS

4 Conclusion

In this studying, we considered random vibration fatigue analysis on a tower model using IGA and FEA, in which the isogeometric and finite element damage results are validated by the Matlab program.

During the analysis, the tower model is clamped on the base flange, and random force PSD in a vertical direction to the tower surface is applied to the concentrated mass element. From modal analysis, it can be found that the obtained first five natural frequencies and vibration modes from IGA and FEA have a good agreement. Fatigue analyses show that the obtained isogeometric and finite element maximum effective stress RMS are 3.151e+8 and 3.125e+8 *pa* with a relative error of 0.83%, and cumulative damage ratios are 2.678e-4 and 2.638e-4 with a relative error of 1.52%. Based on the Matlab program, the isogeometric and finite element damage are respectively 2.62e-4 and 2.64e-4, leading to the relative error of -0.76%.

On the other hand, in the aspect of the mesh refinement process, for IGA, it is not necessary to create mesh elements on the original geometry model. it is sufficient to develop mesh elements



(a) IGA (b) FEA

Figure 9: The cumulative damage ratio

on the previous mesh model, and so the mesh refinement time can be largely saved. However, for the FEA, the refinement process is mandatory to communicate with the original geometric model, and so this process is more time-consuming in LS Dyna software.

In addition, IGA can predict the fatigue life using fewer NURBS elements and integration points in the thickness direction, which correspond very well to the fatigue life computed by FEA, with the relative errors of 0.68% . Through the comparison of numerical analysis results, it can be observed that the obtained isogeometric, finite element PSD and RMS have a good agreement, leading to conclude that IGA is suitable for the random vibration fatigue analysis.

REFERENCES

[1] 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 (2005): 4135-4195.

[2] Lu J, *Isogeometric contact analysis: Geometric basis and formulation for frictionless contact* , Computer Methods in Applied Mechanics and Engineering, 200 (2011): 726-741.

[3] Temizer I, Wriggers P, Hughes T J R, *Contact treatment in isogeometric analysis with NURBS* . Computer Methods in Applied Mechanics and Engineering, 200 (2011): 1100-1112.

[4] Temizer I, Wriggers P, Hughes T J R, *Three-dimensional mortar-based frictional contact treatment in isogeometric analysis with NURBS* , Computer Methods in Applied Mechanics and Engineering, 200 (2012): 115-128.

[5] Y. Bazilevs, T.J.R. Hughes, *NURBS-based isogeometric analysis for the computation of flows about rotating components*, Comput Mech, 43 (2008): 143150.

[6] Y. Bazilevs,V.M. Calo, Y. Zhang, T.J.R. Hughes, *Isogeometric Fluidstructure Interaction Analysis with Applications to Arterial Blood Flow*, Comput Mech, 38 (2006): 310322.

[7] Y. Bazilevs,V.M. Calo, Y. Zhang, T.J.R. Hughes, *Isogeometric fluid-structure interaction: theory, algorithms, and computations*, Comput Mech, 43 (2008): 337.

[8] Xiaoping Qian, *Full analytical sensitivities in NURBS based isogeometric shape optimization*, Computer Methods in Applied Mechanics and Engineering, 199 (2010): 2059-2071.

[9] Wolfgang A. Wall, Moritz A. Frenzel, Christian Cyron, *Isogeometric structural shape optimization*, Computer Methods in Applied Mechanics and Engineering, 197 (2008): 2976-2988.

[10] B. Hassani, S.M. Tavakkoli, N.Z. Moghadam, *Application of isogeometric analysis in structural shape optimization*, Scientia Iranica, 18 (2011): 846-852.

[11] S. Shojaee and N. Valizadeh and M. Arjomand, *Isogeometric structural shape optimization using particle swarm algorithm*, Iran University of Science & Technology, 1 (2011): 633-645.

[12] J. Kiendl, K.-U. Bletzinger, J. Linhard, R. Wchner, *Isogeometric shell analysis with KirchhoffLove elements*, Computer Methods in Applied Mechanics and Engineering, 198 (2009): 3902-3914.

[13] D.J. Benson, Y. Bazilevs, M.C. Hsu, T.J.R. Hughes, *Isogeometric shell analysis: The ReissnerMindlin shell*, Computer Methods in Applied Mechanics and Engineering, 199 (2010): 276-289.

[14] D.J. Benson, Y. Bazilevs, M.C. Hsu, T.J.R. Hughes, *A large deformation, rotation-free, isogeometric shell*. Computer Methods in Applied Mechanics and Engineering, 200 (2011): 1367-1378.

[15] D.J. Benson, Y. Bazilevs, M.C. Hsu, T.J.R. Hughes, *T-spline finite element method for the analysis of shell structures*, International Journal for Numerical Methods in Engineering, 80 (2009): 507-536.

[16] Sang Jin Lee, Kyoung Sub Park,s, *Vibrations of Timoshenko beams with isogeometric approach*, Applied Mathematical Modelling, 37 (2013): 9174-9190.

[17] Weeger, O., Wever, U. & Simeon, B., *Isogeometric analysis of nonlinear EulerBernoulli beam vibrations*, Applied Mathematical Modelling, Nonlinear Dyn, 72 (2013): 813-835.

[18] Clemens V. Verhoosel Michael A. Scott Thomas J. R. Hughes, *An isogeometric analysis approach to gradient damage models*, International Journal for Numerical Methods in Engineering, 86 (2011): 115-134.

[19] Bazilevs, Y., Deng, X., Korobenko, A., Lanza di Scalea, F., Todd, M. D., and Taylor, S. G. *Isogeometric Fatigue Damage Prediction in Large-Scale Composite Structures Driven by Dynamic Sensor Data*, Journal of applied mechanics, 82 (2015): 115-134.

[20] J.A. Cottrell, A. Reali, Y. Bazilevs, T.J.R. Hughes, *Isogeometric analysis of structural vibrations*, Computer Methods in Applied Mechanics and Engineering, 195(2006): 5257-5296.

[21] Dongdong Wang, Wei Liu, Hanjie Zhang, *Novel higher order mass matrices for isogeometric structural vibration analysis*. Computer Methods in Applied Mechanics and Engineering, 260 (2013): 92-108.

[22] N Bazeos, G.D Hatzigeorgiou, I.D Hondros, H Karamaneas, D.L Karabalis, D.E Beskos, *Static, seismic and stability analyses of a prototype wind turbine steel tower*, Engineering Structures, 24 (2002): 1015 - 1025.

[23] X. Pitoiset, A. Preumont, *Spectral methods for multiaxial random fatigue analysis of metallic structures*, International Journal of Fatigue, 22 (2002): 541 - 550.

BONE TISSUE CHARACTERIZATION AND ITS STRUCTURAL SIMULATION

Explicit expressions for elastic constants of osteoporotic lamellar tissue and damage assessment using Hashin failure criterion

R. Megías[*], R. Belda[*], A. Vercher-Martínez[*,†] and E. Giner[*,†]

[*] Institute of Mechanical and Biomechanical Engineering (I2MB)
Universitat Politècnica de València
Valencia, Spain
[†] Department of Mechanical Engineering and Materials (DIMM)
Universitat Politècnica de València
Valencia, Spain
e-mail: ramedia@upv.es; ribelgon@upv.es; anvermar@dimm.upv.es; eginerm@mcm.upv.es

Key words: lamellar bone tissue, microporosity, bone mineral density, elastic constants, bone tissue damage, finite element method

Abstract: *In this work, we have derived explicit expressions to estimate the orthotropic elastic constants of lamellar tissue as a function of the porosity at tissue level (microporosity) and the bone mineral density. Our results reveal that the terms of the main diagonal of the stiffness matrix fit an exponential equation, while the cross terms of the matrix fit a polynomial expression. Regarding to bone damage, failure onset assessed by Hashin criterion is mainly due to matrix elements failure. Finally, a linear relationship was found between bone mineral density (BMD) and cancellous bone stiffness at the macro scale.*

1 INTRODUCTION

Bone is a biological material with a hierarchical structure that develops in an optimal condition, supporting the loads to which it is subjected using the minimum material. Specifically, cancellous bone is a highly porous and heterogeneous material whose structure is mainly struts and plates framework. This type of bone is laminated at the microscale and the tissue arranged at these layers is called lamellar bone tissue.

At tissue level, collagen fibrils are known to be oriented in the direction of the strut on the most external surface [1]. However, as we move deeper into the strut towards the inside, the collagen is more randomly distributed. For this reason, the need to orient the material properties arises because the behaviour will not be the same in each direction. This non-isotropic nature of lamellar tissue must be considered in the quantification of bone mechanical properties.

On the other hand, porosity at lamellar tissue (microporosity) and bone mineral content are two relevant parameters related to bone mechanics. It is known that an increase of bone mineral density (BMD) implies a stiffness rise, but an excessive increase will make the lamellar tissue more brittle. Microporosity contributes in the bone loss mechanical response. Porosity exerts strong influences on mechanical properties in structural materials [2, 3]. Similar dependencies exist for bone, its strength and stiffness vary inversely with increasing porosity [4, 5]. Bone porosity has two possible sources, natural porosity and pathological such as osteoporosis. Natural porosity is mainly due to canaliculi, lacunae and vascular porosity. Pathological porosity causes a widening of vascularisation channels, increase of empty lacunae due to death of osteocytes and degradation in bone architecture.

As regards bone damage assessment, some researchers have defined isotropic failure criteria for assessing bone failure. In line with the previous comments, lamellar tissue has a non-isotropic behaviour, so we may need a more complex failure criterion. In composite materials, interactive failure criteria are usually used to model damage initiation in a composite layer. In 1980 Hashin

https://doi.org/10.4995/YIC2021.2021.12442

[6] proposed two failure mechanisms based on the matrix and fiber failure respectively and distinguished between tension and compression for unidirectional fiber composites. Failure of these type of material fits with the lamellar tissue failure, so we will study bone damage by Hashin's orthotropic failure criterion.

Finally, failure analysis requires the strength limits of the material. Ascenzi and Bonucci conducted several studies of different osteon types in order to define their strength limits [7, 8, 9]. They carried out tensile tests and they concluded that the osteons with longitudinal arrangement in the consecutive lamellae are the stiffest ones [7]. Under compression, osteons with a transverse arrangement of the collagen fibrils are the stiffest [8]. Under shear loads [9], osteons with some transverse collagen fibril arrangement were found to be stiffer in relation to the other types, such as longitudinal osteons.

In this work, we have estimated expressions for the orthotropic elastic constants of lamellar tissue as a function of the porosity at the tissue level and the bone mineral density. For this task, we have developed finite element models in which porosity is explicitly modelled as ellipsoids and spheres. Moreover, bone failure onset has been modeled by Hashin criterion, while damage evolution has been assessed through the material property degradation method (MPD).

2 MATERIALS AND METHODS

2.1 Specimen description, scanning and numerical modelling

A swine lumbar trabecular bone sample will be modelled for assessing bone damage. The trabecular bone sample was prepared in Instituto de Biomecnica de Valencia (IBV) from a lumbar vertebrae of a mature skeletal swine. The specimen was cut in parallelepiped-shaped sample with 10 mm length side. The sample was scanned by μ-CT in Estacin de Bioloxa Maria from A Graa (Universidad de Santiago de Compostela), whose scanner is Skyscan1172 (Bruker, Kontig, Belgium) achieving images with an isotropic resolution of 13.58 μm (voltage 100 kV, intensity 100 μA, Al/Cu filter). The μ-CT images were segmented using a manual global thresholding procedure (ScanIP, Simpleware, UK). From the set of μ-CT images of the scanned sample, a 2 mm cube-shaped region of interest was digitally extracted for the subsequent numerical model generation.

Numerical simulations of tension and compression load cases will be conducted. We consider three values of bone mineral density (BMD) in this study, 0.653 g/cm^3, 0.85 g/cm^3 and 1.16 g/cm^3, in order to evaluate BMD influence in the mechanical response. BMD and porosity will be implicitly considered using explicit expressions for the stiffness matrix at bone tissue level, derived in this work. For damage bone assessment, the strength limits, shown in Table 1, were inferred from [8, 10]. In the simulations, the main growth direction of bone is defined as the longitudinal direction of the sample where plates structure dominate, while the remaining two directions are defined as transverse directions where struts prevail.

Table 1: Strength limit values (MPa) for lamellar tissue. The subscript t, c and s denote tension, compression and shear, respectively.

S_{1t}	S_{1c}	$S_{2t} = S_{3t}$	$S_{2c} = S_{3c}$	$S_{12s} = S_{13s}$	S_{23s}
120	-115	50	-160	46	38

2.2 Modelling porosity in lamellar tissue

We have modelled the two types of porosity: natural (due to lacunae) and pathological. In order to mimic natural porosity we have modelled ellipsoids, which represent the empty lacunae after osteocytes death. On the other hand, we have used spheres for mimicking lamellar tissue

holes due to osteoporosis. Natural porosity represents up to a 10% of the total bone porosity [11]. Therefore, we assume ellipsoid voids up to a 5% porosity and spherical voids for the whole range of porosities.

We have studied six percentages of porosity (1%, 5%, 10%, 15%, 20% and 25%) according to Martnez-Reina [11] and twelve levels of bone mineral density. The minimum value we consider for the BMD at tissue level is 0.653 g/cm^3 from Koller [12], while the maximum value is 1.50 g/cm^3 from [13].

Porosity does not appear with any pattern neither with a specific arrangement in bone tissue. For this reason, we have generated ten models with a random distribution of non-overlapping spheres to represent pores at tissue level while we used ellipsoids to mimic lacunae, see Figure 1. An average stiffness matrix is assessed for the ten random models of each pair of porosity and BMD values.

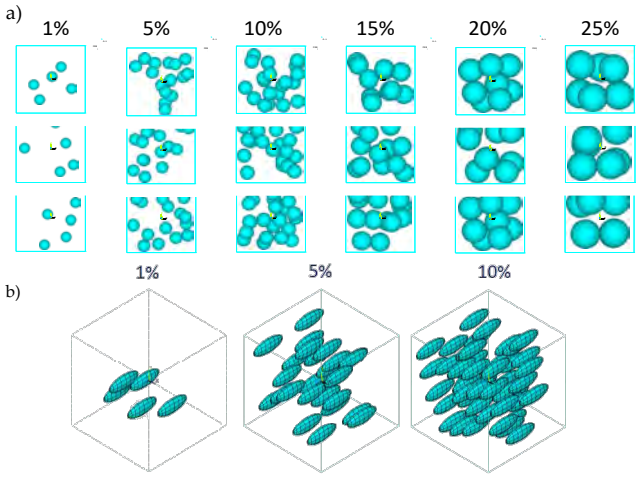


Figure 1: a) Random models with sphere-shaped pores and b) random models with ellipsoids

The reason why we have studied porosity with ellipsoids only until 5% is because they represent the lacunae which appear after osteocytes death. This type of porosity is due to natural bone porosity and it only represents at most the 5% of the total bone porosity. Regarding to sphere-shaped voids, we can model the whole range of porosity with them because with a little radius they represent natural bone porosity and with a larger radius they represent the holes that osteoporosis let at lamellar tissue.

The starting point of the current numerical model is considering the equations for estimating the elastic constants of a healthy bone as a function of the trabecular bone mineral density (BMD) in a multiscale approach [14].

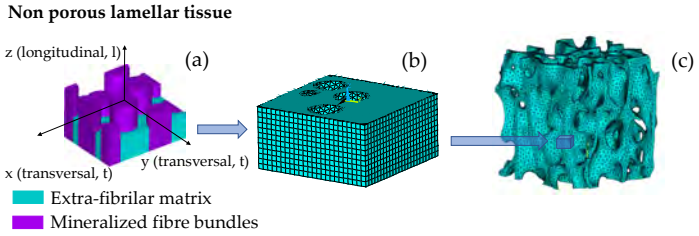


Figure 2: a) Transversely isotropic elastic properties of lamellar tissue as a function of BMD at tissue level [14]. b) Numerical model of the representative elementary volume of porous lamellar tissue. c) Trabecular bone numerical model at microscale.

Vercher et al. [14] assumed the lamellar tissue as a transversely isotropic material (Figure 2a). The elastic constants of the lamellar bone are given in [14].

The geometry of the numerical model is a region of interest of trabecular bone and is cube-shaped. We have modelled a representative periodically repeated volume called unit cell (Figure 2b). For this reason, we have applied periodic boundary conditions to guarantee that the analyzed hexahedron behaves as a continuous domain. Owing to the non-isotropic behaviour of lamellar tissue, the element coordinate system in the numerical model must be conveniently aligned with lamellar tissue properties. Regarding to the mesh, it is an important point of the numerical model due to the necessity of having a mirror mesh at each opposite faces of the model. Finally, a direct homogenization technique has been applied for estimating the average apparent stiffness of porous lamellar tissue.

2.3 Bone damage modeling

In this work, independent quasi-static load cases are numerically simulated. Bone failure onset has been modelled by Hashin's orthotropic failure criterion. On the other hand, bone damage evolution has been assessed through the material property degradation (MPD) method.

Hashin failure criterion is widely used to predict intralaminar failure in orthotropic materials. It assumes different failure mechanisms for tension and compression, both in the fiber and transverse directions. The formulation is the following [6]:

$$f_f = \left(\frac{\sigma_{11}}{X_t}\right)^2 + \frac{1}{S^2}\left(\tau_{12}^2 + \tau_{13}^2\right); \quad \sigma_{11} > 0 \tag{1}$$

$$f_f = \frac{\sigma_{11}}{X_c}; \quad \sigma_{11} < 0 \tag{2}$$

$$f_m = \frac{\left(\sigma_{22} + \sigma_{33}\right)^2}{Y_t^2} + \frac{\left(\tau_{23}^2 - \sigma_{22}\sigma_{33}\right)}{Q^2} + \frac{\left(\tau_{12}^2 + \tau_{13}^2\right)}{S^2}; \quad \sigma_{22} + \sigma_{33} > 0 \tag{3}$$

$$f_m = \frac{\left(\sigma_{22} + \sigma_{33}\right)}{Y_c}\left[\left(\frac{Y_c}{2Q}\right)^2 - 1\right] + \frac{\left(\sigma_{22} + \sigma_{33}\right)^2}{4Q^2} + \frac{\left(\tau_{23}^2 - \sigma_{22}\sigma_{33}\right)}{Q^2} + \frac{\left(\tau_{12}^2 + \tau_{13}^2\right)}{S^2}; \quad \sigma_{22} + \sigma_{33} < 0 \tag{4}$$

where X and Y are axial strength limits in longitudinal direction and transverse to the fiber, respectively ($X_t = S_{1t},\ X_c = S_{1c},\ Y_t = S_{2t},\ Y_c = S_{2c}$). Subscripts f and m denote fiber and matrix while subscripts t and c denote tension and compression. Furthermore, S and Q are shear strength limits in 12 and 23 planes respectively, being *1* the longitudinal direction of the fiber, normal direction to 23 plane ($S = Ss_{12},\ Q = Ss_{23}$).

We evaluate failure using the safety factor *f*, given by equation 5. Failure occurs for f values are greater than one.

$$f = max(f_f, f_m) \tag{5}$$

On the other hand, damage evolution is modelled through material property degradation. The load is progressively applied in quasi-static step increments until failure conditions are reached. Then the Young's modulus of the damaged elements is reduced.

In this work, the stiffness penalty for fiber and matrix failure is reduced differently. Fibers are stiffer and more resistant than matrix, so they transfer more load. If fibers fail, their stiffness is reduced in a 90%. On the other hand, if matrix fails, its stiffness is reduced to 50%.

3 RESULTS AND DISCUSSION

3.1 Expressions for the terms of the stiffness matrix as a function of porosity and bone mineral density

A non-linear multivariable regression by means of the least square fitting has been performed to adjust explicit expressions for the elastic constants of lamellar tissue as a function of the volumetric bone mineral density and porosity. For both porosity geometries, spheres and ellipsoids, the expressions estimated for the main diagonal of the stiffness matrix fit an exponential expression, equation 6.

$$y = ae^{b \cdot p} e^{c \cdot BMD} \tag{6}$$

Table 2: Values a, b and c for fitting the expression for the main diagonal for the stiffness matrix [C]. All the terms are expressed in GPa, BMD in g/cm^3 and porosity in %.

	Spheres			Ellipsoids		
	a	b	c	a	b	c
C_{11}	5.847	-0.02173	0.5817	5.757	-0.03656	0.5986
C_{22}	5.839	-0.02181	0.5830	5.784	-0.02004	0.6017
C_{33}	7.388	-0.02223	0.8213	7.119	-0.01527	0.8718
C_{44}	1.467	-0.02058	0.8123	1.336	-0.01318	0.8980
C_{55}	1.468	-0.02060	0.8119	1.334	-0.02199	0.8985
C_{66}	1.348	-0.02013	0.7298	1.254	-0.02277	0.7945

The exponential equation terms that fit equation 6 for each type of porosity are given in Table 2. The values which multiply the porosity are negative while those which multiply BMD are positive. Therefore, an increment of porosity in lamellar tissue causes a reduction of stiffness, while in bone mineral density leads to a stiffer material.

Figure 3 shows the terms of the main diagonal of the stiffness matrix for several bone mineral densities obtained with the estimated expressions for ellipsoids (cross markers) and sphere-shaped pores (circle markers). As can be seen, results for both expressions are really close between them, so we can use both indifferently. Moreover, the results for the term C_{33} are the highest in agreement with the most stiffest direction of the sample, the fibers direction. Besides, the terms C_{11} and C_{22} are almost identical according to the definition of the material which is transversely isotropic. Furthermore, it can be noticed that there is not a wide variation between the terms C_{44}, C_{55} and C_{66} of the main diagonal of the stiffness matrix.

The terms of the stiffness matrix [C] related to mutual influence and Chentsov coefficients are negligible in comparison with the rest of the terms and they only have a slightly variation with porosity and BMD. In order to complete the terms of the stiffness matrix we have to fit expressions for C_{12}, C_{13} and C_{23} terms. A polynomial function is the best fit for the numerical results obtained, but with slightly differences between spheres and ellipsoids holes. Equation 7 shows the fitting expression used for spheres-shaped pores, while equation 8 the corresponding for ellipsoids pores. Table 3 summarizes the values which correspond with each term of the fitting equation.

$$y = a + b \cdot p + c \cdot p^2 + d \cdot p^3 + e \cdot BMD + f \cdot BMD^2 + g \cdot BMD^3 \tag{7}$$

$$y = a + b \cdot p + c \cdot p^2 + d \cdot p \cdot BMD + e \cdot BMD + f \cdot BMD^2 \tag{8}$$

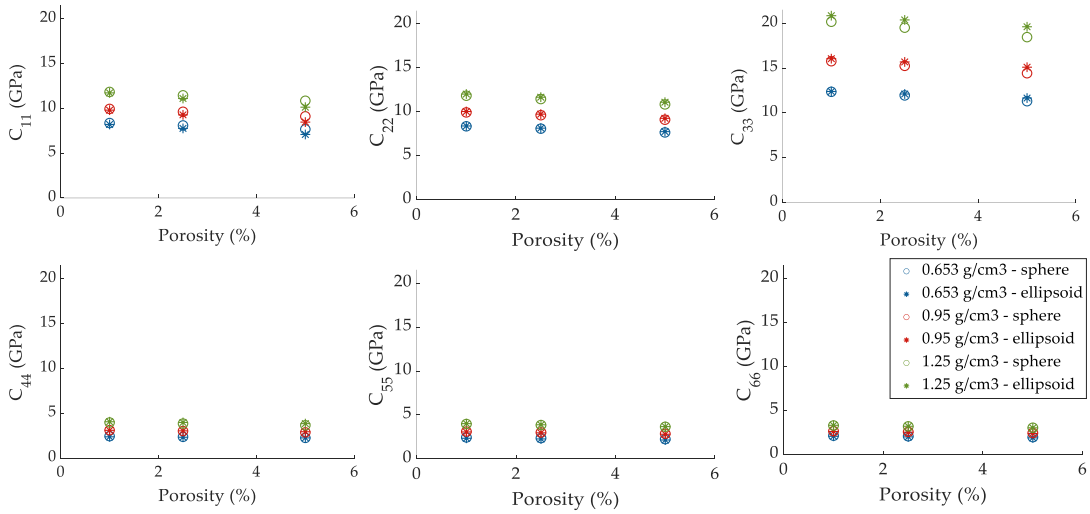


Figure 3: Comparison between the expressions estimated for the terms of the main diagonal of the stiffness matrix for mimicking porosity by ellipsoids and sphere-shaped holes.

Table 3: Values for the parameters a, b, c, d, e, f and g for fitting the polynomial expressions. All the terms are expressed in GPa, BMD in g/cm^3 and porosity in %. Subscript s denotes the spheres terms and subscript e denotes ellipsoids terms.

	a	b	c	d	e	f	g
$C_{12,s}$	2.1878	-0.12627	8.402210^{-4}	0	4.0292	-1.1405	0
$C_{13,s}$	-3.6721	-0.10889	-6.156610^{-4}	3.635010^{-5}	19.131	-10.812	0.58818
$C_{23,s}$	-6.6623	-0.11082	-3.934510^{-4}	3.022710^{-5}	30.459	-25.596	7.0279
$C_{12,e}$	0.7633	-0.1727	0.004519	-0.03203	6.949	-2.633	-
$C_{13,e}$	-1.261	-0.2131	0.004492	0.005465	12.67	-6.059	-
$C_{23,e}$	-1.486	-0.1581	0.002377	0.01366	13.15	-6.305	-

Figure 4 plots the results for the cross terms of the stiffness matrix. Both C_{13} and C_{23} terms show the same trend for the results. For these terms, the function has a maximum and then falls again, hence, the results for 0.95 g/cm^3 are greater than the ones for 1.25 g/cm^3. On the other hand, the equation followed for C_{13} is a polynomial that for all the studied values always grows. Therefore, as BMD increases the results increase as well.

3.2 Failure modeling results for tension and compression load cases

In this section, we have studied the failure behaviour of a vertebral trabecular swine bone numerical model. Figure 5 shows the results considering a 0.85 g/cm^3 bone mineral density and 5% of porosity. Tension load case is represented in blue whereas orange corresponds to the compression load case. Dashed lines represent the hypothetical linear behaviour of the sample in order to identify failure onset. The sample has a similar behaviour under tension and compression loads. However, the failure onset begins a bit earlier in tension than in compression. Moreover, the maximum stress is higher for compression than for tension.

The elements that fail under compression load at yielding and complete failure are represented in red in Figure 5. It can be noted that few elements failed at yielding, mainly due to matrix failure. After several load steps, the material collapses and several elements failed. Their stiffness has been reduced as compression has progressed. At this point, the component is not able to bear greater strains and collapses.

Three different values of BMD have been chosen for evaluating its influence on the failure

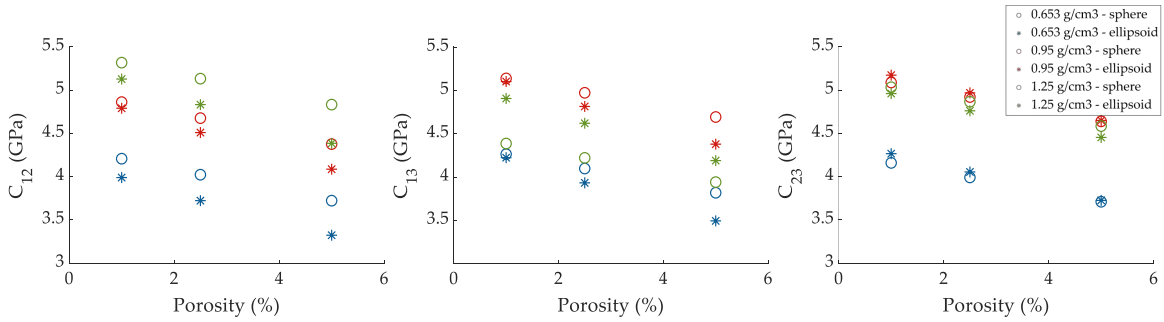


Figure 4: Comparison between the expressions estimated for the cross terms of the stiffness matrix for mimicking porosity by ellipsoids and sphere-shaped holes.

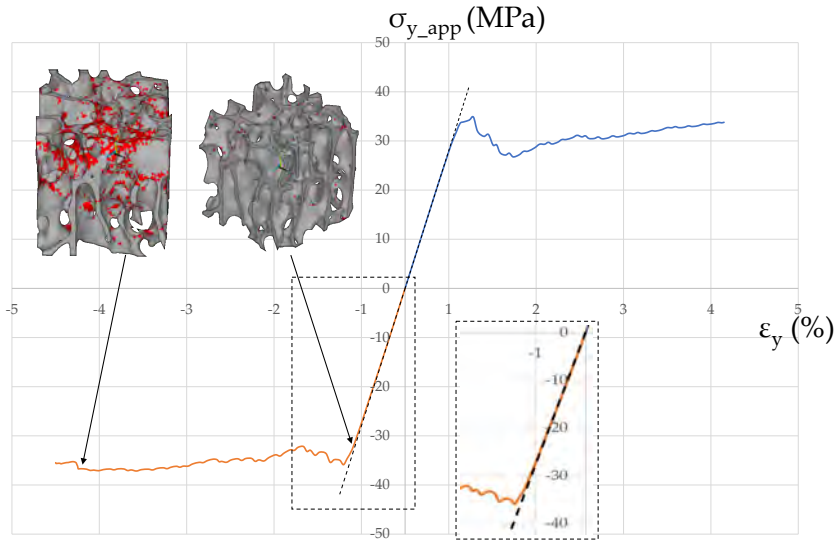


Figure 5: Tension and compression curves for the longitudinal direction (bone growth main direction) for a vertebral trabecular swine bone numerical model with 0.85 g/cm^3 bone mineral density and 5% of porosity.

of the vertebral trabecular swine bone. Porosity has been set for all cases in 5% because this percentage corresponds to a healthy bone, and there was no evidence about the presence of any disease or pathology in the animal.

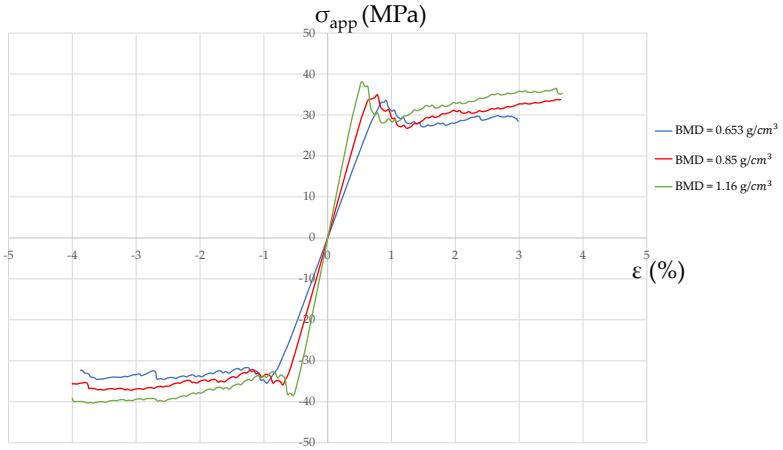


Figure 6: Tension and compression behaviour curves for three values of bone mineral density.

Figure 6 shows the tension-compression response of the cancellous bone specimen for three BMD values. As can be seen in the reported results, as BMD increases, the apparent stress that sample can withstand is greater. The reason is due to the fact that bone mineral density contributes to a higher stiffness in bone. It can be observed that as bone mineral density increases, yield strain is lower, but the apparent stress that each sample can withstand at the yield point is greater as stiffness increases.

4 CONCLUSIONS

In this work, we provide explicit expressions for the terms of the stiffness matrix as a function of porosity and bone mineral density. We have mimicked natural porosity at tissue level using ellipsoids due to the lacunae voids after osteocytes death and spheres for both sources of porosity, natural and pathological. The terms of the main diagonal of the stiffness matrix follow an exponential equation, whereas the cross terms fit a polynomial law. The results obtained indicate that an increment of porosity in lamellar tissue causes a reduction of stiffness, while in bone mineral density leads to a stiffer material.

We have detected that the importance of orientating the lamellar tissue in the numerical models is essential for obtaining results closely to an experimental test. The structure of the vertebral trabecular swine bone sample has dominant struts structure in the transverse directions, while plates prevail in the main growth direction of bone. For this reason, we have oriented the fibers vertically in plate directions and flat for transverse directions where struts exist.

Finally, we have proved that an orthotropic failure criterion can be used in order to analyse bone failure onset considering bone tissue as a composite material. Moreover, elastic property degradation method is an efficient procedure to analyse the failure propagation in a 3D numerical model due to its computational cost. Stress - strain curves for the trabecular bone numerical model show a similar behaviour both in tension and compression. Furthermore, the influence of BMD on the stiffness of the trabecular bone has been studied and we have seen that as BMD increase the apparent stiffness of the sample is greater.

ACKNOWLEDGEMENTS

The authors acknowledge the Generalitat Valenciana for the financial support received through Plan FDGENT 2018. The authors also ackowledge the Ministerio de Ciencia e Innovacin and the ERDF-FEDER programme through the project DPI2017-89197-C2-2-R.

REFERENCES

[1] Georgeadis, M. et al. Ultrastructure organization of human trabeculae assessed by 3D sSAXS and relation to bone microarchitecture. *Plos One*, 11 (8), e0159838, (2016).

[2] Hashin, Z. The elastic moduli of heterogeneus materials. *J. Appl. Mech.*, 29, 143-150, (1962).

[3] Brown, S., Biddulph, R.B. and Wilcox, P.D. A strenght - porosity relation involving different pore geometry and orientation. *Am. Ceram. Soc. J.*, 47, 320, (1964).

[4] Carter, D.R. and Hayes, W.C. The compressive behaviour of bone as a two - phase porus structure. *J. Bone Jt. Surg.*, 59A, 954-962, (1977).

[5] Martin, R.B. Porosity and specific surface of bone. *CRC Crit. Biomed. Engng.*, 10, 179-222, (1984).

[6] Hashin, Z. Failure criteria for unidirectional fiber composites. *J. Appl. Mech.*,Vol. 47, pp. 329-334, (1980).

[7] Ascenzi, A. and Bonucci, E. The tensile properties of single osteons. *The Anatomical Record*, 161(3):377-391, (1967).

[8] Ascenzi, A. and Bonucci, E. The compressive properties of single osteons. *The Anatomical Record*, 161(3):377-391, (1968).

[9] Ascenzi, A. and Bonucci, E. The shearing properties of single osteons. *The Anatomical Record*, 161(3):377-391, (1972).

[10] Giner, E., Arango, C., Vercher, A. and Fuenmayor, F.J. Numerical modelling of the mechanical behaviour of an osteon with microcracks. *Anat. Rec.*, 161:377-392, (2014).

[11] Martínez-Reina, J., Domínguez, J. and García-Aznar, J.M. Effect of porosity and mineral content on the elastic constants of cortical bone: a multiscale approach. *Biomech. Model. Mechanobiol.*, 10:309-322, (2011).

[12] Koller, B. and Laib, A. Calibration of micro-CT data for quantifying bone mineral and biomaterial density and microarchitecture. *Advance Bioimagin Technologies in assessment of the quality of bone and scaffold materials: Techniques and Applications*, DOI: 10.1007/978-3-540-45456-4-5, (2007).

[13] Yu, W. et al. Spinal bone mineral assessment in postmenopausal women: a comparison between dual X-ray absorptiometry and quantitative computed tomography. *Osteoporosis Int.*, 5:433-439, (1995).

[14] Vercher-Martínez A., Giner, E., Belda, R., Aigoun, A. and Fuenmayor, F.J. Explicit expressions for the estimation of the elastic constants of lamellar bone as a function of the volumetric mineral contents using a multi-scale approach. *Biomech. Model. Mechanobiol.*, 17:449-464, (2017).

ADVANCED DISCRETIZATIONS AND SOLVERS FOR COUPLED SYSTEMS OF PARTIAL DIFFERENTIAL EQUATIONS

Finite Element Simulation and Comparison of Piezoelectric Vibration-Based Energy Harvesters with Advanced Electric Circuits

A. Hegendörfer* **and J. Mergheim†**

* Institute of Applied Mechanics
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
e-mail: andi.hegendoerfer@fau.de

† Institute of Applied Mechanics
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
e-mail: julia.mergheim@fau.de

Key words: Smart materials, multiphysics simulation, piezoelectric energy harvesting, numerical simulation, finite element method, coupled problem

Abstract: *A system simulation method based on the Finite-Element Method (FEM) is applied to simulate a bimorph piezoelectric vibration-based energy harvester (PVEH) with different electric circuits: The standard circuit, the synchronized switch harvesting on inductor (SSHI) circuit and the synchronized electric charge extraction (SECE) circuit are considered. Moreover, nonlinear elasticity of the piezoelectric material is taken into account and different magnitudes of base excitations are applied. The holistic FEM-based system simulation approach allows the detailed evaluation of the influences of the considered electric circuits on the vibrational behavior of the PVEH. Furthermore, the harvested energy of the different applied electric circuits with respect to the magnitude of base excitation is compared and results from literature regarding the efficiency of electric circuits are confirmed.*

1 INTRODUCTION

A piezoelectric vibration-based energy harvester (PVEH) is composed of an electromechanical structure along with an electric circuit to extract the energy. The objective of such a device is converting otherwise unused mechanical energy to electrical energy to power e.g. wireless sensors. The piezoelectric effect, used as the energy conversion principle, describes the generation of electrical voltage when the piezoelectric material is mechanically deformed and vice versa.

The applied electric circuit significantly influences the energy output of a PVEH. The choice of the electric circuit depends on the electromechanical coupling of the harvesting structure, the excitation signal and the electric load among others. Therefore, numerous electric circuits have been introduced in the literature to improve the performance of PVEHs. The passive standard circuit is the simplest electric circuit and it was reported, that active circuits like the synchronized switch harvesting on inductor (SSHI) circuit and the synchronized electric charge extraction (SECE) circuit can increase the efficiency of a PVEH significantly compared to the standard circuit [1, 2]. The aim of this contribution is to simulate and compare the energy output of a PVEH with the standard circuit, the SSHI circuit and the SECE circuit for different magnitudes of base excitations. Nonlinear elasticity becomes important for large magnitudes of base excitation, thus, it has to be taken into account in the simulations [3]. Because the electric circuit and the electromechanical structure influence each other, an accurate simulation of a PVEH requires accurate modelling of both the electromechanical structure and the electric circuit. Analytical solutions are restricted to relatively simple geometries of the electromechanical structure and simple electrical loads. Equivalent circuit models describe only the electric behavior of PVEHs and important quantities like the stresses in the material

https://doi.org/10.4995/YIC2021.2021.12177

can not be evaluated. Simulations of PVEHs based on the Finite-Element Method (FEM) have been introduced and allow for arbitrarily shaped geometries and for modeling nonlinear material behavior. Recently, the authors presented a simulation method using only the FEM and allowing for nonlinear electromechanical structures and nonlinear electric circuits [4]. The influence of the electric circuit on the electromechanical structure is considered via the vector of external forces and an implicit time integration scheme is used. Because this method allows for efficient simulations of PVEHs considering nonlinearities of both the electric circuit and the structure, this simulation method is applied within this contribution.

2 GOVERNING EQUATIONS OF PIEZOELECTRICITY

Within this contribution index notation in accordance to [5] is applied. A piezoelectric body is characterized by the help of the linear strain tensor S_{ij} and the electric field E_i

$$S_{ij,i} = \frac{1}{2}\left[u_{i,j} + u_{j,i}\right] \text{ and } E_i = -\varphi_{,i} \tag{1}$$

Here, u_i is the mechanical displacement and φ is the electric voltage. The mechanical and the electric equations for a piezoelectric body Ω are given by the balance of linear momentum and Gauss' law considering that piezoelectric materials are insulating

$$T_{ij,i} = \rho \ddot{u}_j \text{ and } D_{i,i} = 0 \tag{2}$$

Here, ρ is the material density, D_i is the dielectric displacement and T_{ij} are the components of the mechancial stress. A constitutive law that specifies the material behavior must be introduced. In [3] nonlinear elasticity was identified as the primary source of nonlinearity and a nonlinear elastic constitutive law for a 1D setting was introduced, which is here extended to 3D

$$T_{ij} = c^E_{ijkl} S_{kl} - e_{kij} E_k + [c_4 S^3_{11} + c_6 S^5_{11}]\delta_{1i}\delta_{1j} \tag{3}$$

$$D_i = e_{ikl} S_{kl} + \varepsilon^S_{ij} E_j \tag{4}$$

Here, c^E_{ijkl} are the components of the elasticity tensor at constant electric field, e_{ikl} is the piezoelectric constant tensor, ε^S_{ij} is the dielectric constant at constant electric field and δ_{ij} is the Kronecker delta. Nonlinear elastic behavior is thus only introduced in 1-direction, which is a reasonable approach since this is the direction of the largest stresses and strains. The piezoelectric problem can be solved with appropriate boundary conditions

$$u_i = \bar{u}_i \text{ on } \partial\Omega_{Du} \quad \varphi = \bar{\varphi} \text{ on } \partial\Omega_{D\varphi} \quad T_{ij} n_j = \bar{t}_i \text{ on } \partial\Omega_{Nu} \quad D_i n_i = -\bar{Q} \text{ on } \partial\Omega_{N\varphi} \tag{5}$$

The prescribed surface traction $\bar{t}_i$ and the free surface charge density $\bar{Q}$ are introduced. The boundary $\partial\Omega$ of Ω consists of subsets that do not overlap, such that $\partial\Omega_D \cup \partial\Omega_N = \partial\Omega$ and $\partial\Omega_D \cap \partial\Omega_N = \emptyset$.

3 DISCRETIZATION OF THE EQUATIONS

A fundamental step of the FEM is to transform the partial differential equations from their strong formulation into their weak formulation. To obtain the weak formulation, equations (3) and (4) are essentially multiplied by test functions η_j and ξ and subsequently, integration by

parts is applied and the boundary conditions are introduced. The weak form results as

$$\underbrace{\int_\Omega \rho\eta_j\ddot{u}_j \; dV}_{\rightarrow \boldsymbol{f}^{dyn,u}} + \underbrace{\int_\Omega \eta_{j,i}T_{ij} \; dV}_{\rightarrow \boldsymbol{f}^{int,u}} - \underbrace{\int_{\partial\Omega_{Nu}} \eta_j\bar{t}_j \; dA}_{\rightarrow \boldsymbol{f}^{ext,u}} = 0 \tag{6}$$

$$\underbrace{\int_\Omega \xi_{,i}D_i \; dV}_{\rightarrow \boldsymbol{f}^{int,\varphi}} + \underbrace{\int_{\partial\Omega_{N\varphi}} \xi\bar{Q} \; dA}_{\rightarrow \boldsymbol{f}^{ext,\varphi}} = 0 \tag{7}$$

The particular integral terms of the weak form result after the FE discretization in the force terms specified under the brackets. Thereby, $\boldsymbol{f}^{dyn,u}$ is the mechanical inertial force, $\boldsymbol{f}^{int,u}$ is the internal mechanical force, $\boldsymbol{f}^{ext,u}$ is the mechanical external force, $\boldsymbol{f}^{int,\varphi}$ is the electric internal force and $\boldsymbol{f}^{ext,\varphi}$ is the electric external force. The FEM subdivides the domain Ω into small discrete elements and approximates the unknown solutions namely the mechanical displacement and the electric voltage elementwise by means of polynomial ansatz functions and nodal degrees of freedom. Thus, two coupled vector-valued equations for the unknown nodal displacements and electric voltage values result as

$$\boldsymbol{f}^{dyn,u} + \boldsymbol{f}^{int,u} - \boldsymbol{f}^{ext,u} = \boldsymbol{0} \tag{8}$$

$$\boldsymbol{f}^{int,\varphi} + \boldsymbol{f}^{ext,\varphi} = \boldsymbol{0} \tag{9}$$

The coupling to an electric circuit is easier if the electric surface current $-\dot{\bar{Q}}$ appears in the equation. Therefore, the time derivative of equation (9) is used for the system simulations. Furthermore, a mechanical damping force $\boldsymbol{f}^{damp,u}$ is introduced. The resulting system of equations thus becomes

$$\boldsymbol{f}^{dyn,u} + \boldsymbol{f}^{damp,u} + \boldsymbol{f}^{int,u} - \boldsymbol{f}^{ext,u} = \boldsymbol{0} \tag{10}$$

$$\dot{\boldsymbol{f}}^{int,\varphi} - \dot{\boldsymbol{f}}^{ext,\varphi} = \boldsymbol{0} \tag{11}$$

By means of $\boldsymbol{f}^{damp,u}$, a Rayleigh-type damping is modelled. For solving equations (10) and (11) the implicit Bossak-Newmark method is applied for direct time integration. In each time step the nonlinear system of equations is iteratively solved with Newton's method.

4 ELECTRIC CIRCUITS

In this contribution, the standard circuit, the SSHI and the SECE circuit are considered as electric circuits for a PVEH. In the following, the different electric circuits are presented along with a bimorph electromechanical structure introduced in [6].

4.1 Standard Circuit

The simplest way of an AC-DC converter is to use a diode bridge and supply the rectified voltage to an energy storage element. Figure 1 presents this standard circuit coupled with a bimorph electromechanical structure along with the typical waveforms of φ_{el} and $\dot{\bar{Q}}_{el}$ under a harmonic excitation. φ_{el} and $\dot{\bar{Q}}_{el}$ are the voltage and current values at the electrode of the electromechanical structure. They are available from the FEM simulation as decribed in section 5. If the absolute value of the piezoelectric voltage $|\varphi_{el}|$ is smaller than the conductive voltage V_C, the electromechanical structure is in open circuit mode and the current flowing out of the electrode $\dot{\bar{Q}}_{el}$ vanishes. The conductive voltage results from the drop voltage of the diodes V_D and the constant voltage V_{DC}. When $|\varphi_{el}|$ reaches V_C the diode bridge conducts and an electric current $\dot{\bar{Q}}_{el}$ charges the battery.

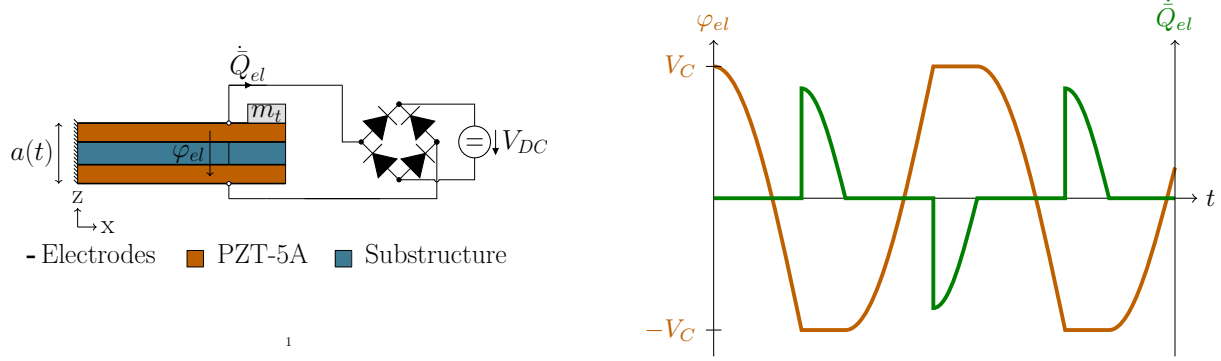


Figure 1: Bimorph electromechanical structure coupled to a standard circuit. The typical waveforms of φ_{el} and $\dot{Q}_{el}$ arising from the standard circuit under a harmonic base excitation of the electromechanical structure are provided.

4.2 SSHI Circuit

The SSHI circuit was introduced in [1] and adds a switch and an inductor (L) to the standard circuit. Figure 2 presents the SSHI circuit coupled to a bimorph electromechanical structure along with the typical waveforms of φ_{el} and $\dot{Q}_{el}$ under harmonic excitation of the structure. When $|\varphi_{el}|$ starts to decrease, the switch is closed, and only the inductor is connected to

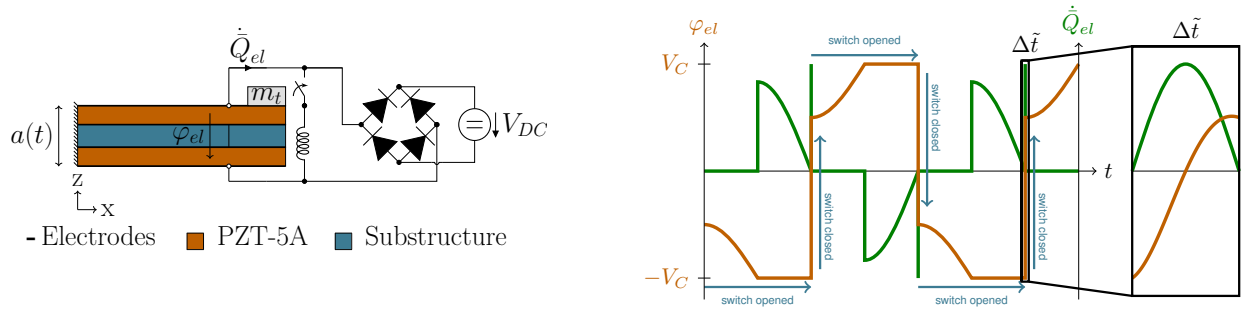


Figure 2: Bimorph electromechanical structure coupled to an SSHI circuit. The typical waveforms of φ_{el} and $\dot{Q}_{el}$ arising from the SSHI circuit under a harmonic base excitation of the electromechanical structure are provided.

the electromechanical structure. Since closing the switch creates an electric resonant circuit consisting of the piezoelectric capacitance and the inductor, the voltage is inverted. When the electric current in the inductor is zero-crossing, the switch is opened again.

4.3 SECE circuit

The SECE circuit was introduced in [2] and consists of a diode rectifier bridge and a buck-boost DCDC converter circuit composed of a switch, an inductor and a diode. To emulate an ideal energy storage device a constant voltage V_{DC} is assumed. During operation of the SECE circuit different phases can be distinguished [7]:

- *Extraction phase.* When the electric voltage φ_{el} reaches its maximum value, the switch is closed and thus only the inductance is connected to the electromechanical structure. Because the piezoelectric capacitance and the inductor form an electric resonance circuit, the electric voltage and the electric current start to oscillate as soon as the switch is closed. When the electric current in the inductance reaches its maximum value, the switch is opened again. When the switch is closed, energy stored in the piezoelectric capacitance is transferred to the inductor. During the extraction phase the absolute value of the

piezoelectric voltage $|\varphi_{el}|$ drops to $\pm 2V_D$. Because the electrical frequencies are usually by orders higher than the mechanical frequencies, the extraction process happens almost instantaneously.

- *Open circuit and freewheeling phase.* The energy stored in the inductor is transferred via the freewheeling diode to the battery. Because the switch is opened, the electromechanical structure is in open circuit mode and the electric current $\dot{\bar{Q}}_{el}$ vanishes.

The inductor is connected either to the electromechanical structure or to the battery, therefore, the electromechanical structure is decoupled from the battery. Figure 3 presents the SECE circuit coupled to a bimorph electromechanical structure along with the typical waveforms of φ_{el}, $\dot{\bar{Q}}_{el}$ and $\dot{Q}_c$ under harmonic excitation of the structure.

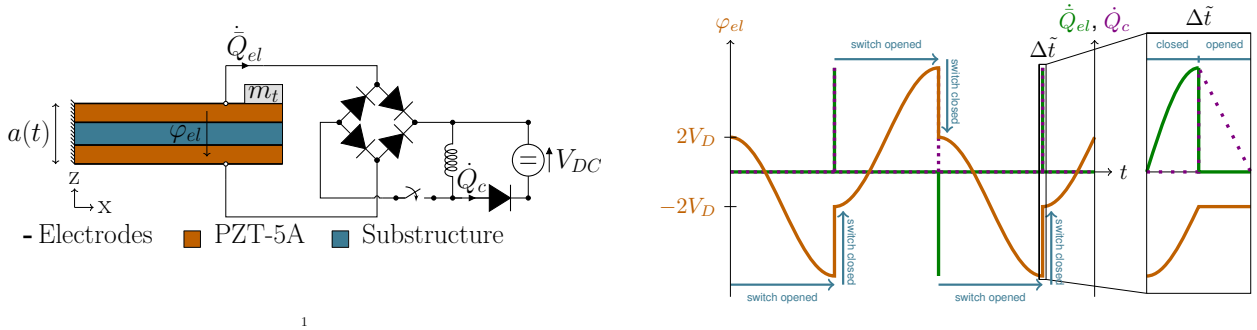


Figure 3: Bimorph electromechanical structure coupled to an SECE circuit. The typical waveforms of φ_{el}, $\dot{\bar{Q}}_{el}$ and $\dot{Q}_c$ arising from the SECE circuit under a harmonic base excitation of the electromechanical structure are provided.

5 FEM SYSTEM SIMULATION

The system simulation method introduced in [4] is applied to simulate a PVEH with a standard, an SSHI and an SECE circuit. Applying the FEM to model an electrode on a surface means to prescribe that all voltage degrees of freedom have the same value φ_{el} on this surface. One reference degree of freedom $\mathcal{F}$ is introduced for the voltage of the electrode. Furthermore, one electrode is assumed to be grounded. Hence, the voltage difference between the electrodes of a PVEH is just the electrical voltage φ_{el} of the non-grounded electrode with the corresponding degree of freedom $\mathcal{F}$. The electric boundary conditions for the grounded electrode is therefore a homogeneous Dirichlet boundary condition with $\bar{\varphi} = 0$. The electric circuit is coupled to the electromechanical structure via the boundary condition of the non-grounded electrode. To account for the influence of the electric circuit on the electromechanical structure, the surface current $\dot{\bar{Q}}_{el}$ leaving the electrode is prescribed. It can be constructed via an inhomogenous Neumann boundary condition

$$\dot{\bar{Q}}_{el} = \int_{\partial\Omega_{Electrode}} \dot{D}_i n_i \, \mathrm{dA} \tag{12}$$

Both, φ_{el} and $\dot{\bar{Q}}_{el}$ correspond to the same reference degree of freedom $\mathcal{F}$ of the non-grounded electrode. The electric current $\dot{\bar{Q}}_{el}$ appears in equation (11) and is introduced in $\dot{\boldsymbol{f}}^{ext,\varphi}$ in the entry for the reference degree of freedom $\mathcal{F}$ as

$$\dot{\boldsymbol{f}}^{ext,\varphi} = \begin{cases} \dot{\bar{Q}}_{el} & \text{for degree of freedom} = \mathcal{F} \\ 0 & \text{for degree of freedom} \neq \mathcal{F} \end{cases} \tag{13}$$

To consider the coupling of various electric circuits with the FE simulation, the relation between the electric current $\dot{\bar{Q}}_{el}$ and the voltage of the non-grounded electrode φ_{el} has to be specified. The implementation of the standard and SSHI circuits are described in detail in [4]. In the following, only the implementation of the SECE circuit is specified.
During operation of the SECE circuit two cases are possible:
Case 1: This is the open circuit case when the switch is opened, and the inductor is not connected to the electromechanical structure. Hence, the electric current $\dot{\bar{Q}}_{el}$ vanishes and thus can be modeled via a homogenous Neumann boundary condition in the FE simulation

$$\dot{\bar{Q}}_{el} = 0 \tag{14}$$

Case 2: In this case the switch is closed, and the inductor is connected to the electromechanical structure. The influence of the inductor on the electromechanical structure is considered via an inhomogeneous Neumann boundary condition. Depending on the sign of φ_{el} the relation is specified as

$$\ddot{\bar{Q}}_{el} = \begin{cases} \frac{\varphi_{el}-2V_D}{L} & \text{if } \varphi_{el} > 0 \\ \frac{\varphi_{el}+2V_D}{L} & \text{if } \varphi_{el} < 0 \end{cases} \tag{15}$$

In figure 4 the logic of the SECE circuit, i.e. the conditions how to switch between cases 1 and 2, is described.

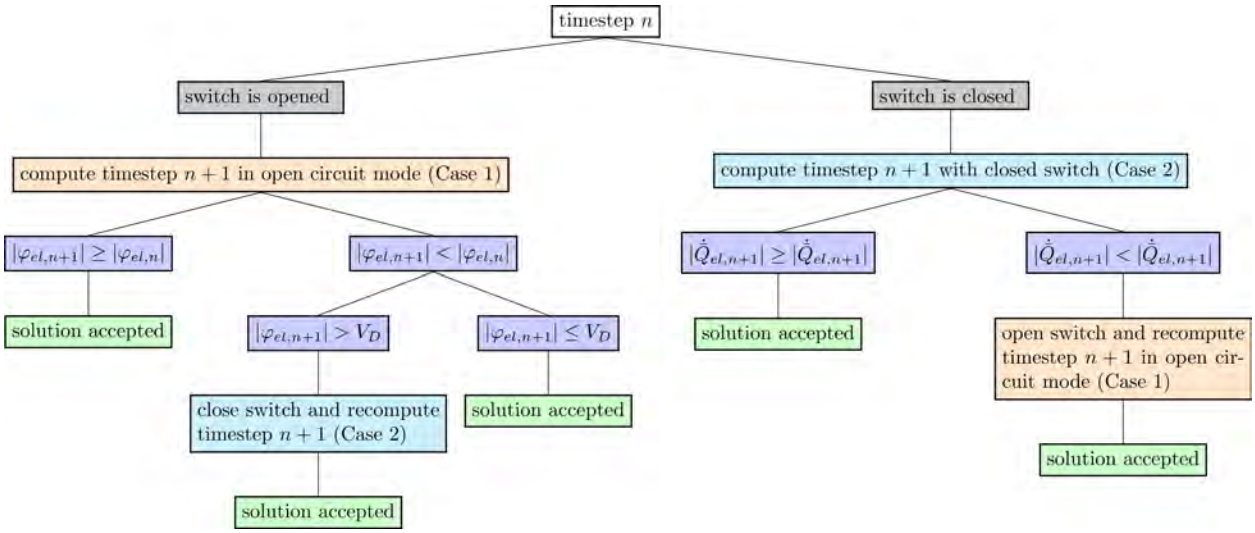


Figure 4: Logic for the definition of boundary conditions for the FE simulation of an electromechanical structure with an SECE circuit.

It has to be prevented in the simulation that the switch is triggered at inappropriate times. After closing the switch, φ_{el} decreases rapidly because the electrical oscillation frequency is usually by orders higher than the mechanical excitation frequency. The rapid decrease of φ_{el} after closing the switch acts like an actuation to the structure, which leads to oscillations of φ_{el} and the electromechanical structure. To prevent that the switch is triggered due to higher order oscillations of φ_{el} caused by its rapid decrease, a time span t_{ns} is introduced during which closing the switch is prohibited. Moreover, φ_{el} flutters during the instationary settling process when the mechanical excitation starts. Therefore, a time span t_{switch} is defined during which closing of the switch is prohibited at the beginning of the simulation.

The harvested energy $\mathcal{E}$ of the SECE circuit can be obtained as

$$\mathcal{E}(t) = \int_0^t V_{DC}\dot{Q}_c \, d\tau \tag{16}$$

whereby $\dot{Q}_c$ depends on the maximum of $\dot{\tilde{Q}}_{el}$ when the switch is opened. In the computation of the harvested energy the dissipation of the diodes is considered, for details please see [7].

6 APPLICATION EXAMPLE

As an application example the bimorph electromechanical structure with a tip mass m_t introduced in [6] coupled to the three different electric circuits is simulated. The bimorph cantilever has a mass mounted on its tip and consists of two layers of PZT-5A bracketing a layer of a passive substructure. The piezoceramic layers are poled in opposite directions. For the substructure, a linear elastic material behavior is assumed and for the PZT-5A the nonlinear piezoelectric constitutive law is applied. Figure 3 presents the considered PVEH and the appendix provides its parameters. A harmonic base acceleration $a(t)$ with a frequency of 47.8 Hz and with three different magnitudes of 0.5, 1 and 9.81 m/s^2 excites the PVEH. The harvested energy and the electric voltage are compared for the standard, the SSHI and the SECE circuit. Some of these results for the standard and SSHI circuits have already been discussed in [4], but here, for the first time, the results for all three circuits are carefully analyzed and compared for different excitation amplitudes.

To prohibit triggering the switch at inappropriate times $t_{ns} = 4$ ms and $t_{switch} = 10.8$ ms are chosen. The bimorph PVEH is discretized with 90 quadratic hexahedral elements. During the extraction phase when the switch is closed the time step size of the Bossak-Newmark time integration scheme is set to 10^{-5} ms since it happens almost instantaneously. After the switch is opened the time step size is set to 10^{-3} ms within a time span $t_{osc} = 2$ ms to precisely capture the higher order oscillation of the electromechanical structure. For the remaining time period a time step size of 10^{-1} ms is used.

Firstly, a magnitude of 0.5 m/s^2 of the harmonic base acceleration is applied. Figure 5 presents the results for φ_{el} and the harvested energy of the standard , the SSHI and the SECE circuit from [4]. The considered time period is 0 to 200 ms. The standard circuit harvests around 0.0014 mJ which is approximately 125% of the energy harvested by the SECE circuit and around 300% of the amount of energy harvested by the SSHI circuit. As shown in figure 5, left, the SSHI circuit rapidly changes the potential when the switch is closed by a much larger amount than the SECE circuit. Therefore, the excitation of high frequency vibration modes is more significant for the SSHI circuit and the energy dissipation is higher than for the SECE circuit. Moreover, the advantage of the SSHI circuit is usually to extend the time during which the diode rectifier is conducting. But in this case, due to the higher order vibration modes, the conduction time is actually shorter than for the standard circuit which can be seen in the zoom-in of the diagram in figure 5, left. Because of the higher dissipation due to the high frequency vibrations and the short conduction time, the SSHI circuit is the least effective here and harvests the least amount of energy. For a base acceleration of 0.5 m/s^2 the passive standard circuit, that does not excite high frequency vibration modes, is more efficient than the SSHI circuit and SECE circuit.

Figure 6 presents the results for φ_{el} and the harvested energy when a magnitude of 1 m/s^2 of harmonic base acceleration is applied. The considered electric circuits harvest nearly the same amount of energy within a time period of 200 ms namely around 0.006-0.007 mJ. The change of φ_{el} caused by the switching events of the SSHI and the SECE circuit are in the same order and therefore approximately an equal amount of energy is dissipated by the high frequency vibration modes. Furthermore, when a magnitude of 1 m/s^2 of harmonic base excitation is applied, the SSHI circuit extends the time fraction during which the diode bridge conducts compared to the standard circuit, as shown in the zoom-in in figure 6, left. The standard circuit does not excite high frequency vibration modes and has the same efficiency for the considered PVEH

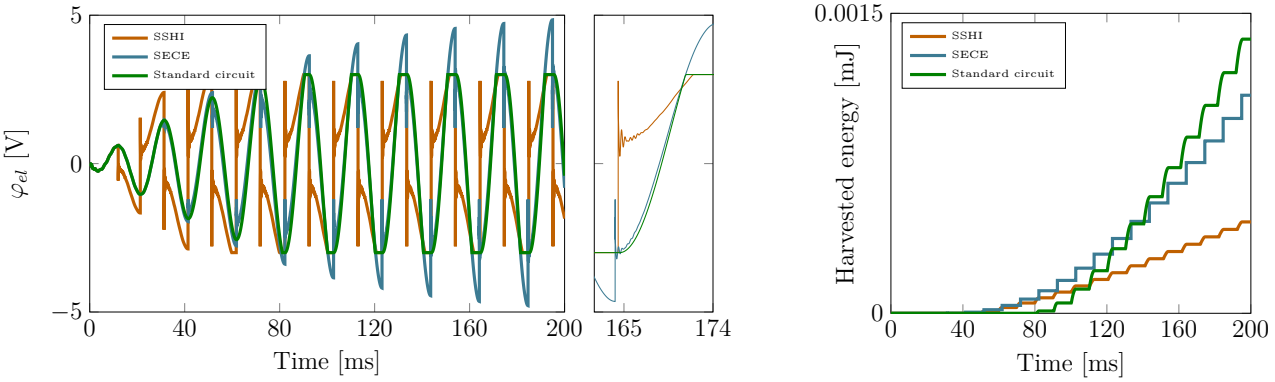


Figure 5: Electrode voltage φ_{el} and harvested energy $\mathcal{E}$ of the bimorph PVEH with standard, SSHI and SECE circuit under a harmonic base acceleration of 0.5 m/s^2 and a frequency of 48.7 Hz.

like the SSHI and SECE circuit for the considered time period. While the standard circuit and the SSHI circuit limit the piezoelectric voltage φ_{el} to $\pm V_C$, the SECE circuit does not limit φ_{el} and therefore the highest values of φ_{el} are reached.

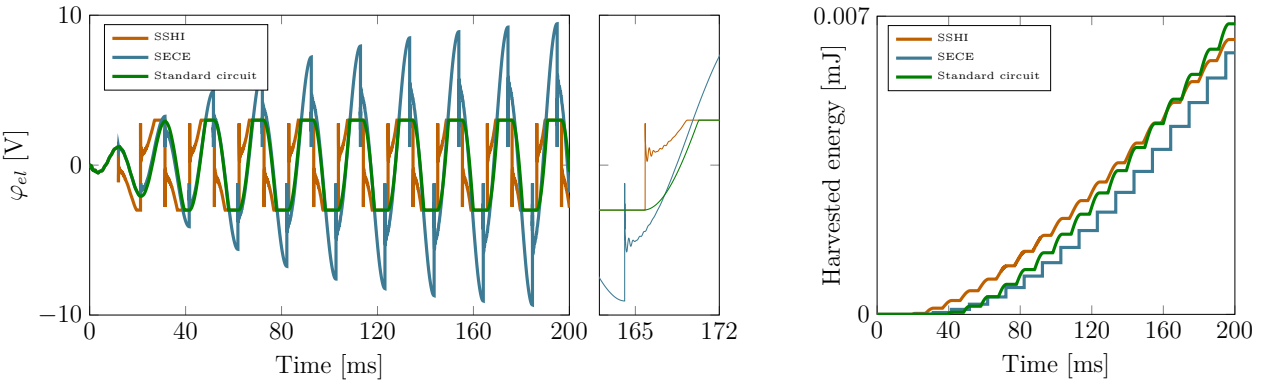


Figure 6: Electrode voltage φ_{el} and harvested energy $\mathcal{E}$ of the bimorph PVEH with standard, SSHI and SECE circuit under a harmonic base acceleration of 1 m/s^2 and a frequency of 48.7 Hz.

Figure 7 presents the results for φ_{el} and the harvested energy for the standard circuit, the SSHI circuit and the SECE circuit when a harmonic base acceleration with a magnitude of 9.81 m/s^2 is applied. The SECE circuit harvests approximately 0.3 mJ during 200 ms, which is around three times the amount of energy harvested by the standard circuit or the SSHI circuit. The high level of applied base acceleration would lead to high open circuit piezoelectric voltages compared to the conductive voltage V_C of the standard circuit and the SSHI circuit, which limit φ_{el} to $\pm V_C$. Therefore, the actuation of the electromechanical structure caused by voltage inversion for the SSHI circuit and, thus, the related dissipation of energy, are relatively small. However, because of the high level of mechanical base acceleration the advantage of the SSHI circuit, namely to extend the conduction time of the diode bridge, does not significantly improve the efficiency compared to the standard circuit, as is shown in the zoom-in in figure 7, left. Therefore, both the standard circuit and SSHI circuit harvest nearly the same amount of energy, namely around 0.09 mJ. In contrast, the harvested energy is independent of the applied electric load when the SECE circuit is used. Therefore, the SECE circuit is more efficient than the standard circuit and the SSHI circuit for this high level of base acceleration. To optimize the harvested energy of the standard circuit and the SSHI circuit for this setting a DC-DC converter must be applied after the diode bridge to regulate the conductive voltage V_C. This flexible adjustment of V_C to the current conditions would allow to harvest significantly more energy with the standard circuit and the SSHI circuit than for a constant conductive voltage

V_C [8].

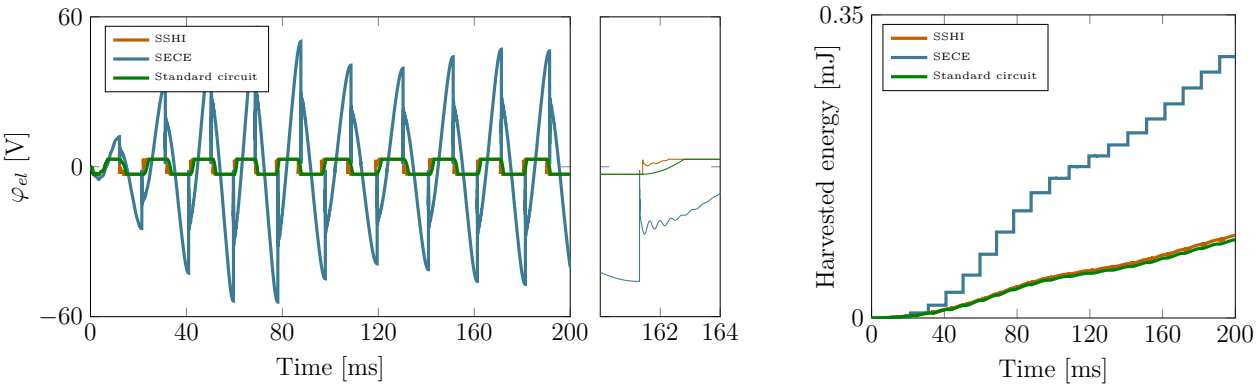


Figure 7: Electrode voltage φ_{el} and harvested energy $\mathcal{E}$ of the bimorph PVEH with standard, SSHI and SECE circuit under a harmonic base acceleration of 9.81 m/s^2 and a frequency of 48.7 Hz.

To sum up the simulation results, the drawback of the SSHI and the SECE circuit compared to the standard circuit is additional dissipation because of higher order vibration modes excited by switching events. This observation is consistent with the literature [9]. Furthermore, the simulation results confirm the advantage of the SECE circuit namely the independence of the harvested energy on the electric load. The efficiency of the standard and the SSHI circuit decreases compared to the SECE circuit for high levels of base acceleration. To improve the efficiency of the standard and the SSHI circuit a flexible adaption of the conductive voltage V_C to the current harvesting conditions would be necessary.

7 CONCLUSION

The FEM based system simulation method introduced in [4] is here applied to simulate a PVEH with three different circuits, a standard, an SSHI and an SECE circuit, and to compare their efficiency. Nonlinear elasticity of the electromechanical structure is taken into account and different magnitudes of a harmonic base acceleration are considered. Consistently with the literature, the SECE circuit and the SSHI circuit dissipate energy compared to the standard circuit through higher order vibration modes caused by switching events. This additional dissipated energy reduces the efficiency of the respective electric circuits. Because the harvested energy is independent of the electric load, the SECE circuit is more efficient than the considered standard circuit and the SSHI circuit without an additional DC voltage regulation stage for high levels of base accelerations. These results demonstrate the applicability of the system simulation method of [4] to develop or improve PVEHs.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge financial support for this work by the Deutsche Forschungsgemeinschaft under GRK2495/C.

REFERENCES

[1] Guyomar D., Badel A., Lefeuvre E., et al. Toward energy harvesting using active materials and conversion improvement by nonlinear processing. *IEEE Transactions of Ultrasonics, Ferroelectrics, and Frequency Control*, Vol. **52**, pp. 584-595, (2005).

[2] Lefeuvre, E., Badel, A., Richard, C., et. al. Piezoelectric Energy Harvesting Device Optimization by Synchronous Electric Charge Extraction. *Journal of Intelligent Material Systems and Structures*, Vol. **16**, pp. 865-876, (2005).

[3] Stanton S.C., Erturk A., Mann B.P., et al. Nonlinear nonconservative behavior and modeling of piezoelectric energy harvesters including proofmass effects. *Journal of Intelligent Material Systems and Structures*, Vol. **23**, pp. 685-704, (2012).

[4] Hegendörfer, A., Steinmann P. and Mergheim, J. Nonlinear finite element system simulation of piezoelectric vibration based energy harvesters. *Journal of Intelligent Material Systems and Structures*, Accepted. DOI 10.1177/1045389X211048222.

[5] Institute of Electrical and Electronics Engineers (IEEE) IEEE standard on piezoelectricity, ANSI-IEEE Std. 176-1987 (1988). Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=26560 (accessed 02 December 2020).

[6] Erturk A. and Inman D.J. An experimentally validated bimorph cantilever model for piezoelectric energy harvesting from base excitations. *Smart Materials and Structures*, Vol. **18**, 025009, (2009).

[7] Chen, C., Zhao, B. and Liang J. Revisit of synchronized electric charge extraction (SECE) in piezoelectric energy harvesting by using impedance modeling. *Smart Materials and Structures*, Vol. **28**, 105053, (2019).

[8] Ottman G.K., Hofmann H.F., Bhatt A.C., et al. Adaptive piezoelectric energy harvesting circuit for wireless remote power supply. *IEEE Transactions on Power Electronics*, Vol. **17**, pp. 669-676, (2002).

[9] Dorsch P., Bartsch T., Hubert F., et al. Implementation and Validation of a Two-Stage Energy Extraction Circuit for a Self Sustained Asset-Tracking System. *Sensors*, Vol. **19**, 1330, (2019).

APPENDIX

In the following, material parameters for PZT-5A used in this contribution are provided and both the parameters of the bimorph electromechanical structure from [6] and the parameters of the application example (Figure 3) are listed.

Width of the beam [mm]	31.8
Length of the beam [mm]	50.8
E modulus substructure [GPa]	105
Thickness substructure [mm]	0.14
Thickness PZT [mm]	0.26 (each)
Density substructure [kg/m^3]	9000
Tip mass m_t [kg]	0.012
Inductivity [mH]	0.1
V_D [V]	0.6
$V_{DC}[V]$	1.8

$$\boldsymbol{c}^E = \begin{bmatrix} 120.3 & 75.2 & 75.1 & 0 & 0 & 0 \\ 75.2 & 120.3 & 75.1 & 0 & 0 & 0 \\ 75.1 & 75.1 & 110.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 21.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 21.1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 22.6 \end{bmatrix} \text{GPa}$$

$$\boldsymbol{e} = \begin{bmatrix} 0 & 0 & 0 & 0 & 12.3 & 0 \\ 0 & 0 & 0 & 12.3 & 0 & 0 \\ -5.4 & -5.4 & 15.8 & 0 & 0 & 0 \end{bmatrix} \frac{\text{C}}{\text{m}^2}$$

$$\boldsymbol{\varepsilon}^S = \begin{bmatrix} 813.7 & 0 & 0 \\ 0 & 813.7 & 0 \\ 0 & 0 & 731.9 \end{bmatrix} \times 10^{-11} \frac{\text{F}}{\text{m}}$$

The nonlinear coefficients in equation (3) for the nonlinear piezoelectric constitutive law were identified in [3] as $c_4 = -9.7727 \times 10^{17}$ Pa and $c_6 = 1.4700 \times 10^{26}$ Pa for the considered PZT-5A material.

Monolithic Finite Element Method for the simulation of thixo-viscoplastic flows

N. Begum[*], **A. Ouazzi**[†], **S. Turek**[‡]

Institute for Applied Mathematics, LS III, TU Dortmund University, D-44227 Dortmund, Germany
[*]Naheed.Begum@math.tu-dortmund.de; [†]Abderrahim.Ouazzi@math.tu-dortmund.de; [‡]ture@featflow.de

Key words: Thixo-viscoplastic Flows, FEM, Newton-Multigrid, Monolithic, Generalized Navier-Stokes equations, Incompressible fluids

Abstract: This note is concerned with the application of Finite Element Method (FEM) and Newton-Multigrid solver to simulate thixo-viscoplastic flows. The thixo-viscoplastic stress dependent on material microstructure is incorporated via viscosity approach into generalized Navier-Stokes equations. The full system of equations is solved in a monolithic framework based on Newton-Multigrid FEM Solver. The developed solver is used to analyze the thixo-viscoplastic flow problem in a Lid-driven cavity configuration.

1 INTRODUCTION

The thixo-viscoplastic flows are introduced into yield stress flows by taking in consideration the internal material micro-structure using a structure parameter λ. Firstly, the viscoplastic stress is modified to include the thixotropic stress dependent on the structure parameter

$$\begin{cases} \sigma(\lambda) = 2\eta(\lambda)\mathbf{D}(u) + \tau(\lambda)\dfrac{\mathbf{D}(u)}{\|\mathbf{D}(u)\|}, & \text{if } \|\mathbf{D}(u)\| \neq 0, \\ \|\sigma(\lambda)\| \leq \tau(\lambda), & \text{if } \|\mathbf{D}(u)\| = 0, \end{cases} \tag{1}$$

where $\mathbf{D}(u)$ denotes the strain rate tensor. The norm for a tensor Λ is given by $\|\Lambda\| = \sqrt{\mathrm{Tr}(\Lambda^2)}$. We use $\|\mathbf{D}(u)\|$ and $\|\mathbf{D}\|$ alternately. η denotes plastic viscosity, and τ defines a yield stress that is a threshold parameter from which the material starts yielding. The shear stress has two contributions: a viscous part, and a strain rate independent part. Secondly, an evolution equation for the structure parameter is introduced to induce the time-dependent process of competition between the destruction (breakdown) and the construction (buildup) inhabited in the material

$$\left(\frac{\partial}{\partial t} + u \cdot \nabla\right)\lambda = \mathcal{F} - \mathcal{G}, \tag{2}$$

where, $\mathcal{F}$ and $\mathcal{G}$ are two nonlinear functions representing the buildup and breakdown of material micro-structure. A collection of thixotropic models with various choices of η, τ, $\mathcal{F}$ and $\mathcal{G}$ is given in Table 1;

Table 1: Thixotropic models

	η	τ	$\mathcal{F}$	$\mathcal{G}$
Worrall & Tulliani [16]	$\lambda\eta_0$	τ_0	$a(1-\lambda)\|\mathbf{D}\|$	$b\lambda\|\mathbf{D}\|$
Coussot et al.[4]	$\lambda^g\eta_0$		a	$b\lambda\|\mathbf{D}\|$
Houška [6]	$(\eta_0+\eta_1\lambda)\|\mathbf{D}\|^{n-1}$	$(\tau_0+\tau_1\lambda)$	$a(1-\lambda)$	$b\lambda^m\|\mathbf{D}\|$
Mujumbar et al. [9]	$(\eta_0+\eta_1\lambda)\|\mathbf{D}\|^{n-1}$	$\lambda^{g+1}G_0\Lambda_c$	$a(1-\lambda)$	$b\lambda\|\mathbf{D}\|$
Dullaert & Mewis [3]	$\lambda\eta_0$	$\lambda G_0(\lambda\|\mathbf{D}\|)\Lambda_c$	$(a_1+a_2\|\mathbf{D}\|)(1-\lambda)t^p$	$b\lambda\|\mathbf{D}\|t^{-p}$

where η_0 and τ_0 are initial plastic viscosity and yield stress resp. in the absence of any thixotropic phenomena, η_1 and τ_1 are thixotropic plastic viscosity and yield stress. Λ_c is the critical elastic strain,

https://doi.org/10.4995/YIC2021.2021.12250

and G_0 is the elastic modulus of unyielded material. a and b are buildup and breakage constants, and g, p, m, n are rate indices.

In quasi-Newtonian modeling approach for thixo-viscoplastic flows, an extended viscosity $\mu(\cdot,\cdot)$ is used for the generalized Navier-Stokes equations [10]. As for instance [13]:

$$\mu(D_{II},\lambda) = \eta(D_{II},\lambda) + \tau(D_{II},\lambda)\frac{\sqrt{2}}{2}\frac{1}{\sqrt{D_{II}}}\left(1 - e^{-k\sqrt{D_{II}}}\right), \tag{3}$$

where k is the regularization parameter. The generalized Navier-Stokes equations and the evolution equation for the structure parameter constitute the full set of modeling equations which is given as follows:

$$\begin{cases} \left(\dfrac{\partial}{\partial t} + u\cdot\nabla\right)u - \nabla\cdot\Big(2\mu(D_{II},\lambda)\mathbf{D}(u)\Big) + \nabla p = 0, & \text{in } \Omega, \\ \nabla\cdot u = 0, & \text{in } \Omega, \\ \left(\dfrac{\partial}{\partial t} + u\cdot\nabla\right)\lambda - \mathcal{F}(D_{II},\lambda) + \mathcal{G}(D_{II},\lambda) = 0, & \text{in } \Omega, \end{cases} \tag{4}$$

where u denotes velocity, p the pressure, λ the structure parameter, $\mathcal{F}$ and $\mathcal{G}$ the nonlinear functions for buildup and breakdown of material micro-structure. $D_{II} = \frac{1}{2}\big(\mathbf{D}(u) : \mathbf{D}(u)\big)$ is the second invariant of the strain rate tensor $\mathbf{D}(u)$.

2 FINITE ELEMENT DISCRETIZATION

To derive the variational form for thixo-viscoplastic flows, we consider the spaces $\mathbb{T} := L^2(\Omega), \mathbb{V} := (H_0^1(\Omega))^2$, and $\mathbb{Q} := L_0^2(\Omega)$ associated, respectively, with the corresponding L^2-norm, $\|\cdot\|_0$, H^1-norm, $\|\cdot\|_1$, and L^2-norm, $\|\cdot\|_0$. Let $\tilde{u} := (\lambda, u, p) \in \big(\mathbb{T}\cap H^1(\Omega)\big)\times\mathbb{V}\times\mathbb{Q}$, and $\tilde{v} := (\xi, v, q) \in \mathbb{T}\times\mathbb{V}\times\mathbb{Q}$ be a test function. The weak formulation for the thixo-viscoplastic flows reads: *Find* $\tilde{u} \in \big(\mathbb{T}\cap H^1(\Omega)\big)\times\mathbb{V}\times\mathbb{Q}$ *s. t.*

$$a_\lambda(\tilde{u})(\lambda,\xi) + a_u(\tilde{u})(u,v) + b(v,p) - b(u,q) = 0, \quad \forall\tilde{v}\in\mathbb{T}\times\mathbb{V}\times\mathbb{Q}, \tag{5}$$

where $a_\lambda(\tilde{u})(\cdot,\cdot)$, $a_u(\tilde{u})(\cdot,\cdot)$, and $b(\cdot,\cdot)$ are given as follows

$$a_\lambda(\tilde{u})(\lambda,\xi) = \int_\Omega \Big(-\mathcal{F}(D_{II},\lambda) + \mathcal{G}(D_{II},\lambda)\Big)\xi\, d\Omega + \int_\Omega u\cdot\nabla\lambda\,\xi\, d\Omega, \tag{6}$$

$$a_u(\tilde{u})(u,v) = \int_\Omega 2\mu(D_{II},\lambda)\mathbf{D}(u) : \mathbf{D}(v)\, d\Omega + \int_\Omega u\cdot\nabla u\, v\, d\Omega, \tag{7}$$

$$b(v,q) = -\int_\Omega \nabla\cdot v\, q\, d\Omega. \tag{8}$$

The finite element approximations of the problem (5) have to take care of its saddle point character, due to the bilinear form (8). Furthermore, since thixo-viscoplastic flows are usually slow, the only remaining issue is the control/continuity of the bilinear form (6) in the norm of space $\mathbb{T}$. We opt for higher order stable pair bi-quadratic for velocity and piece-wise linear discontinuous for the pressure, Q_2/P_1^{disc}, and higher order quadratic for structure parameter Q_2 with the appropriate stabilization terms [10, 15]. Indeed, let the domain Ω be partitioned by a grid $K \in \mathcal{T}_h$ which are assumed to be open quadrilaterals such that $\Omega = \text{int}\big(\bigcup_{k\in\mathcal{T}_h}\overline{K}\big)$. For an element $K \in \mathcal{T}_h$, we denote by $\mathcal{E}(K)$ the set of all 1-dimensional edges of K. Let $\mathcal{E}_i := \bigcup_{k\in\mathcal{T}_h}\mathcal{E}(K)$ be the set of all interior element edges of the grid $\mathcal{T}_h$.

We define the conforming finite element spaces $\mathbb{T}_h \subset \mathbb{T}$, $\mathbb{V}_h \subset \mathbb{V}$, and $\mathbb{Q} \subset \mathbb{Q}_h$ such that:

$$\mathbb{T}_h = \Big\{\xi_h \in \mathbb{T}, \xi_{h|K} \in\ Q_2(K)\,\forall\, K\in\mathcal{T}_h\Big\}, \tag{9}$$

$$\mathbb{V}_h = \big\{v_h \in \mathbb{V}, v_{h|K} \in\ (Q_2(K))^2\,\forall\, K\in\mathcal{T}_h,\, v_h = 0 \text{ on } \partial\Omega_h\big\}, \tag{10}$$

$$\mathbb{Q}_h = \Big\{q_h \in \mathbb{Q}, q_{h|K} \in\ P_1^{\text{disc}}(K)\,\forall\, K\in\mathcal{T}_h\Big\}. \tag{11}$$

The approximate problem reads: *Find* $\tilde{u} \in \mathbb{T}_h \times \mathbb{V}_h \times \mathbb{Q}_h$ *s. t.*

$$a_\lambda(\tilde{u})(\lambda,\xi) + j_{\tilde{u}}(\tilde{u},\tilde{v}) + a_u(\tilde{u})(u,v) + b(v,p) - b(u,q) = 0, \quad \forall \tilde{v} \in \mathbb{T}_h \times \mathbb{V}_h \times \mathbb{Q}_h. \tag{12}$$

The stabilization term $j_{\tilde{u}}(\cdot,\cdot)$ is given as follows

$$j(\cdot,\cdot) := j_u(\cdot,\cdot) + j_\lambda(\cdot,\cdot), \tag{13}$$

$$j_u(u,v) = \sum_{E\in\mathcal{E}_i} \gamma_u |E|^2 \int_E [\nabla u]\,[\nabla v]\, d\sigma, \tag{14}$$

$$j_\lambda(\lambda,\xi) = \sum_{E\in\mathcal{E}_i} \gamma_\lambda |E|^2 \int_E [\nabla \lambda]\,[\nabla \xi]\, d\sigma. \tag{15}$$

The stabilization (13) is consistent, control the convective terms and makes the coercivity and continuity match in $\mathbb{T}_h$ associated with the norm $|||\cdot|||$, where

$$|||\xi|||^2 = \|\xi\|_0^2 + j_\lambda(\xi,\xi). \tag{16}$$

3 GENERALIZED DISCRETE NEWTON

We use the Newton method to approximate the nonlinear residuals. Let $\mathcal{R}(\tilde{u}) = (\mathcal{R}_\lambda(\tilde{u}), \mathcal{R}_u(\tilde{u}), \mathcal{R}_p(\tilde{u})) = \big(\mathcal{R}_{(\lambda,u)}(\tilde{u}), \mathcal{R}_p(\tilde{u})\big)$ denote the residuals for the system (12). The nonlinear iteration is updated with the correction $\delta\tilde{u}$, $\tilde{u}^{k+1} = \tilde{u}^k + \delta\tilde{u}$. Then, the Newton linearization gives the following approximation for the residuals:

$$\mathcal{R}(\tilde{u}^{l+1}) = \mathcal{R}(\tilde{u}^l + \delta\tilde{u}) \simeq \mathcal{R}(\tilde{u}^l) + \left[\frac{\partial \mathcal{R}(\tilde{u}^l)}{\partial \tilde{u}}\right] \delta\tilde{u}. \tag{17}$$

The Newton's method iterations, assuming invertible Jacobians, are given as follows:

$$\tilde{u}^{l+1} = \tilde{u}^l - \omega_l \left[\frac{\partial \mathcal{R}(\tilde{u}^l)}{\partial \tilde{u}}\right]^{-1} \mathcal{R}(\tilde{u}^l). \tag{18}$$

The damping parameter $\omega_l \in (0,1)$ is chosen such that

$$\left\|\mathcal{R}(\tilde{u}^{l+1})\right\| \leq \left\|\mathcal{R}(\tilde{u}^l)\right\|. \tag{19}$$

The damping parameter is not sufficient for the convergence of this type of highly nonlinear problem, mainly due to the presence of Jacobian's singularities related to the problem or simply by being out of the domain of Newton's convergence [8, 10]. We use a generalized Newton's method which consists of using approximate Jacobians far away from the quadratic convergence range or close to singularities and accurate Jacobians in the quadratic region of convergence in an adaptive way [8]. Indeed, based on a priori analysis of Jacobians property. Let the Jacobian be written as follows:

$$\left(\frac{\partial \mathcal{R}(\tilde{u}^l)}{\partial \tilde{u}}\right) = \left(\frac{\partial \tilde{\mathcal{R}}(\tilde{u}^l)}{\partial \tilde{u}}\right) + \delta_l \left(\frac{\partial \hat{\mathcal{R}}(\tilde{u}^l)}{\partial \tilde{u}}\right). \tag{20}$$

The Jacobian (20) is splitted into a direct sum of corresponding operators with different properties. The parameter $\delta_l \in (0,1)$ is solely dependent on the rate of actual residual convergence [8]. It is worth mentioning that the operator-related damped Jacobian method (20) is related to the continuous Newton's method. The Jacobian approximation is only dependent on the rate of the actual residual convergence $\left(\left\|\mathcal{R}^l\right\| / \left\|\mathcal{R}^{l-1}\right\|\right)$. This generalized Newton's method assures a global nonlinear convergence [8].

4 MONOLITHIC MULTIGRID LINEAR SOLVER

To develop an appropriate linear solver, we segregate the Jacobian as follows

$$\left(\frac{\partial \mathcal{R}(\tilde{u})}{\partial \tilde{u}}\right) = \begin{pmatrix} \dfrac{\partial \mathcal{R}_{(\lambda,u)}(\tilde{u})}{\partial(\lambda,u)} & \dfrac{\partial \mathcal{R}_u(\tilde{u})}{\partial p} \\ \dfrac{\partial \mathcal{R}_p(\tilde{u})}{\partial u} & 0 \end{pmatrix}, \tag{21}$$

which is a saddle point problem. Then, the resulting linear system is treated with a Multilevel Pressure Schur Complement (MPSC) approach with Vanka-like smoother i.e.

$$\tilde{u}^{k+1} = \tilde{u}^k - \omega_k \sum_{K \in \mathcal{T}_h} \left(\left(\frac{\partial \mathcal{R}(\tilde{u}^l)}{\partial \tilde{u}} \right)_{|K} \right)^{-1} \mathcal{R}(\tilde{u}^l)_{|K}. \tag{22}$$

In (22), we solve exactly on real element, K, and perform an outer Gau-Seidel iteration [5]. We use standard geometric multigrid solver for linearized system with standard Q_2 and P_1^{disc} restriction and prolongation operators. The combination of a stable finite element approximations, Q_2/P_1^{disc}, for Stokes problem together with multigrid results in a high numerically accurate, flexible, and efficient FEM-multigrid solver.

5 THIXO-VISCOPLASTIC FLOW IN LID-DRIVEN CAVITY

Lid-driven cavity flows represent an academic common standard benchmark for incompressible CFD codes. Therefore, we present the corresponding results for Newtonian, viscoplastic, and thixo-viscoplastic flows. Furthermore, this problem is accepted as a test configuration to check points wise mesh convergences despite the lack of regularity due to the pressure singularity in the corners of upper-lid. From thixotropic collection models given in Table 1, we use Houška's material model ($m = 1$).

5.1 Newtonian flow in lid-driven cavity

The global accuracy of the approximation which consist of the L_2-norm of the velocity is investigated using the kinetic energy. On the other hand, the point wise accuracy is investigated using the velocity magnitude at vertical center-line beside the primary vortex and the lower left secondary vortex. In order to check the solver convergence, we list in Table 2 the kinetic energy, $\frac{1}{2}\int_\Omega \|u\|^2\, dx$, and Newton-multigrid iterations, the number of nonlinear iteration versus the average number of multigrid sweeps (N/M), w.r.t. mesh refinement for an increased Reynolds numbers $Re = 1000$, $Re = 5\,000$, and $Re = 10\,000$. The starting solution for any level is the interpolated one from one level coarser. Table 3 shows the primary vortex and lower left secondary vortex w.r.t. mesh refinement for Reynolds numbers $Re = 1000$, $Re = 5\,000$, and $Re = 10\,000$. Moreover, we provide in Figure 1 the stream function contours for the mesh refinement level 9 and the velocity magnitude at vertical center-line w.r.t. mesh refinement for Reynolds numbers $Re = 1000$, $Re = 5\,000$, and $Re = 10\,000$.
As can be seen in Table 2, grid independent results are achieved for the kinetic energy, as well as for Newton-multigrid solver. It is worth mentioning that for the coarser levels few extra nonlinear iterations are required in contrast to finer mesh due to the decrease of interpolation error in the starting solutions.

Table 2: **Newtonian flow in lid-driven cavity**: The kinetic energy and the number of Newton-multigrid iterations, nonlinear number of iterations and the average number of multigrid iterations (N/M), for different mesh refinement at Reynolds numbers $Re = 1\,000$, $Re = 5\,000$, and $Re = 10\,000$.

Re		1 000		5 000		10 000	
Level	cells	Energy$\times 10^2$	N/M	Energy$\times 10^2$	N/M	Energy$\times 10^2$	N/M
7	16384	4.452357	3/1	4.768669	4/1	4.868399	5/1
8	65536	4.451904	3/1	4.744815	3/2	4.783917	4/2
9	262144	4.451846	3/1	4.742921	3/1	4.773500	3/2
10	1048576	4.451834	2/1	4.742815	3/1	4.772692	3/1
Ref. values $\approx$		4.45		4.74		4.77	

Table 3: **Newtonian flow in lid-driven cavity**: The primary vortex and the lower left secondary vortex at $Re = 1\,000$, $Re = 5\,000$, and $Re = 10\,000$.

Re	1000		5000		10000	
Level	ψ_{max}	$\psi_{min} \times 10^3$	ψ_{max}	$\psi_{min} \times 10^3$	ψ_{max}	$\psi_{min} \times 10^3$
7	0.1189360	-1.72649	0.1225439	-3.077555	0.1236127	-3.2070181
8	0.1189361	-1.72851	0.1222499	-3.072411	0.1225210	-3.1831353
9	0.1189362	-1.72963	0.1222269	-3.073524	0.1224097	-3.1910101
10	0.1189366	-1.72965	0.1222259	-3.073589	0.1223892	-3.1797390
Ref.	0.1189[1]	-1.729[1]	0.1221[7]	-3.070[1]		

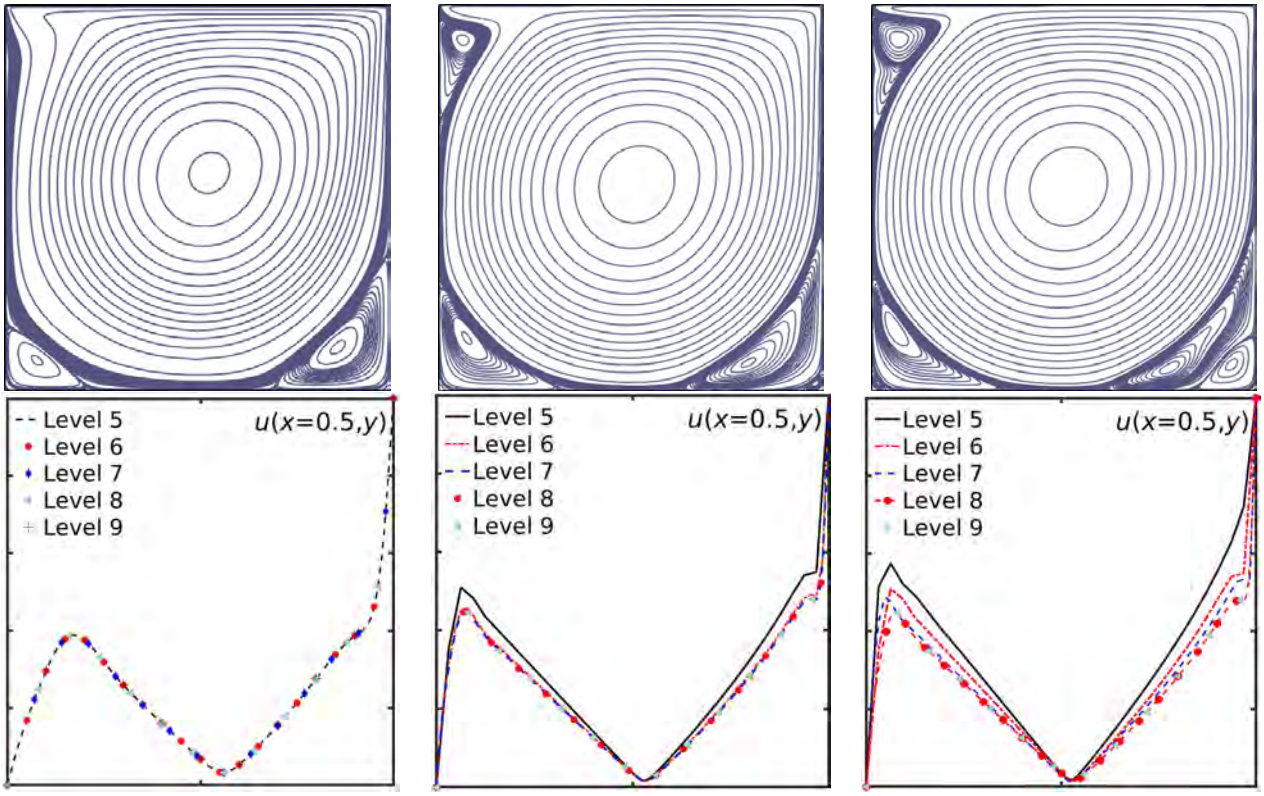


Figure 1: **Newtonian flow in lid-driven cavity:** The stream-function's contours at mesh refinement level 9 (TOP) and velocity magnitude at vertical centerline w.r.t. mesh refinement (BOTTOM) computed for Reynolds numbers $Re = 1\,000$, $Re = 5\,000$, and $Re = 10\,000$ resp. (LEFT to RIGHT).

5.2 Viscoplastic flow in lid-driven cavity flow

We acheived point wise convergence under mesh refinement for Newtonian flow. Moreover, no further improvement by increasing the resolution beyond mesh refinement level 8. Now, we investigate the impact of the regularization parameter in quasi-Newtonian modeling approach for viscoplastic flow. Figure 2 shows the boundary limit of the numerical approximation of the rigid zone w.r.t. regularization parameter k. Clearly, the relative convergence of the boundary limit of the rigid zone w.r.t. regularization parameter k is obtained. Furthermore, there is an optimal regularization $K_L \approx 2^{L8} \approx 10^3$ from from which there is no further accuracy improvement in capturing the rigid zone by increasing the regularization parameter k. In Figure 3, we use the optimal choice of the regularization parameter

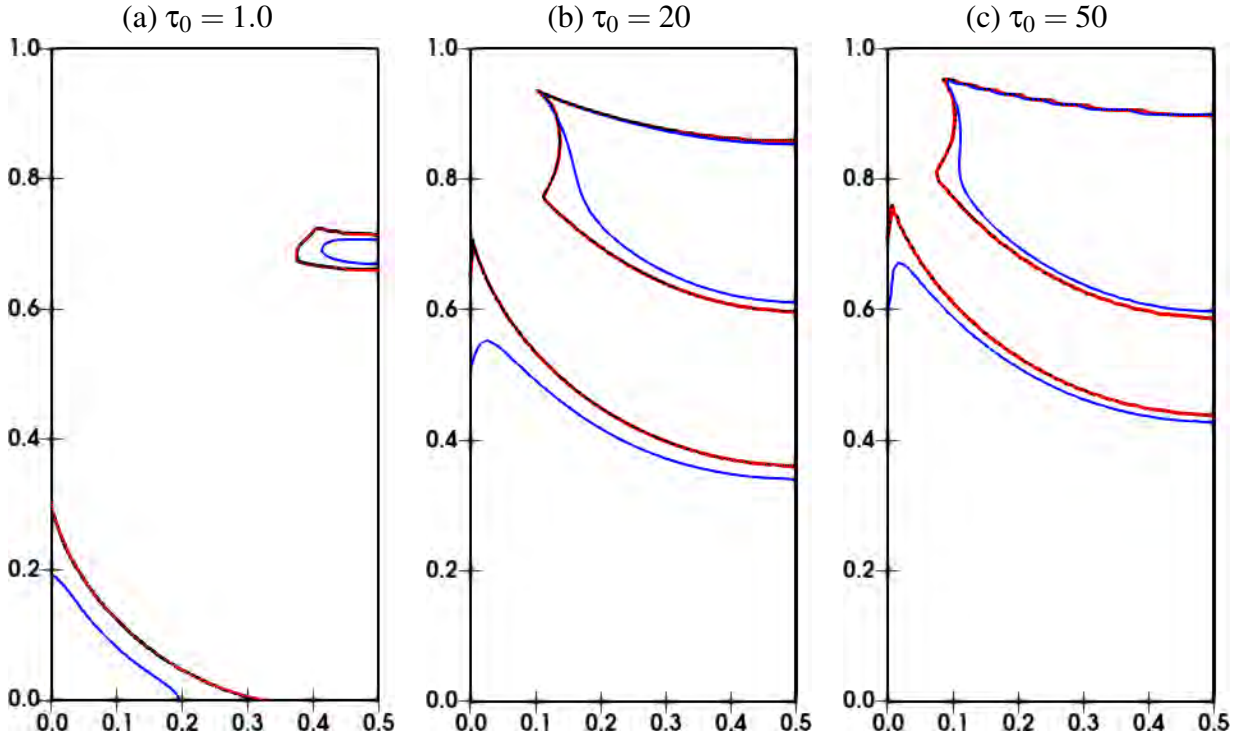


Figure 2: **Non-thixotropic (Bingham plastic) flow in lid-driven cavity:** The boundary limit of the numerical approximation of the plastic/rigid zone w.r.t. regularization parameter k, $k = 10^2$ (blue), $k = 10^3$ (red) and $k = 10^4$ (black), for different non-thixotropic yield stress parameter τ_0. The other parameters are set to $\eta_0 = 1.0$, $\eta_1 = 0.0$, and $\tau_1 = 0.0$. The solutions are calculated at mesh-refinement level 8.

and mesh refinement level to predict the relative position of the rigid zone to stream function contours for an increased non-thixotropic yield stress parameter $\tau_0 = 1$, $\tau_0 = 2$, $\tau_0 = 5$, $\tau_0 = 10$, $\tau_0 = 20$, and $\tau_0 = 50$. Furthermore, we provide the corresponding Newton-multigid data in Table 4 which depicts the number of Newton-multigrid iterations, i.e. the nonlinear number of iterations and the average number of multigrid iterations (N/M), w.r.t. different regularization parameters k and mesh refinement levels L. The solutions are calculated using the continuations process w.r.t. regularization k.
From Table 4, we conclude the Newton-multigrid solver is mesh refinement independent. Clearly, the nonlinearity of the problem is increased by increasing the non-thixotropic yield stress parameter τ_0, But, the slightly increases in the nonlinearity w.r.t. mesh refinement is due to the continuation process w.r.t. regularization parameter k used to obtain the solutions.

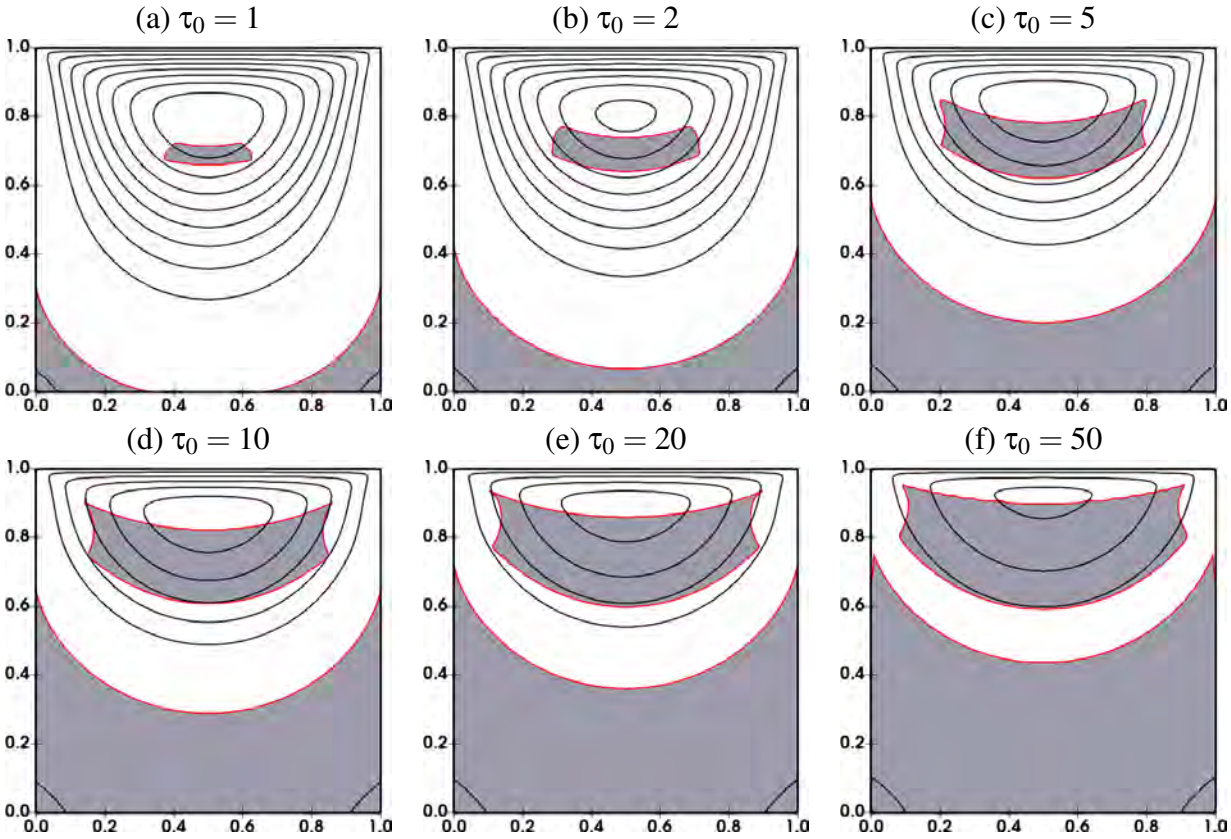


Figure 3: **Non-thixotropic (Bingham plastic) flow in lid-driven cavity:** The relative position of the plastic/rigid zone to streamline contours for an increased non-thixotropic yield stress parameter τ_0. The other parameters are set to $\eta_0 = 1.0$, $\eta_1 = 0.0$, and $\tau_1 = 0.0$. The Papanastasiou regularization parameter is set to $k = 10^4$. The solutions are calculated at mesh-refinement level 8.

Table 4: **Non-thixotropic (Bingham plastic) flow in lid-driven cavity**: The number of Newton-multigrid iterations, nonlinear number of iterations and the average number of multigrid iterations (N/M), w.r.t. different regularization parameters k and mesh refinement levels L for Bingham viscoplastic flow for different values of non-thixotropic yield stress parameters τ_0.

$k\backslash L$	5	6	7	5	6	7	5	6	7
	$\tau_0 = 1$			$\tau_0 = 2$			$\tau_0 = 5$		
1×10^1	3/1	3/1	3/1	3/1	3/1	3/1	4/1	4/1	4/1
1×10^2	3/1	3/1	3/1	3/1	3/1	3/1	4/1	4/1	4/1
1×10^3	2/2	3/2	3/1	3/1	3/1	4/1	4/1	5/2	5/2
1×10^4	2/1	2/2	5/1	3/1	3/1	6/1	4/1	5/4	6/3
	$\tau_0 = 10$			$\tau_0 = 20$			$\tau_0 = 50$		
1×10^1	5/1	5/1	5/1	6/1	6/1	6/1	5/1	7/1	7/1
1×10^2	5/2	4/1	4/1	5/2	5/2	5/1	6/5	5/4	5/1
1×10^3	5/2	7/4	9/1	5/5	7/2	8/1	5/5	9/2	9/2
1×10^4	6/1	7/2	8/3	6/3	5/5	7/3	6/3	7/3	8/2

5.3 Thixo-viscoplastic flow in lid-driven cavity flow

Armed with the knowledge of the point wise mesh convergence of viscoplastic driven cavity flow. Indeed, we obtained the point wise convergence of the boundary limit of the rigid zone w.r.t. the regularization parameter k. Furthermore, there is pair $(K,L) \approx (2^{L8},L8)$ regularization and mesh refinement level beyond which no further resolution improvement is possible. Now, we are ready the investigate thixo-viscoplastic driven cavity. Figure 4 sets out the relative position of the rigid zone to stream function contours for an increased thixotropic yield stress parameter τ_1. The solutions are calculated with the resolution barrier pair (K,L).

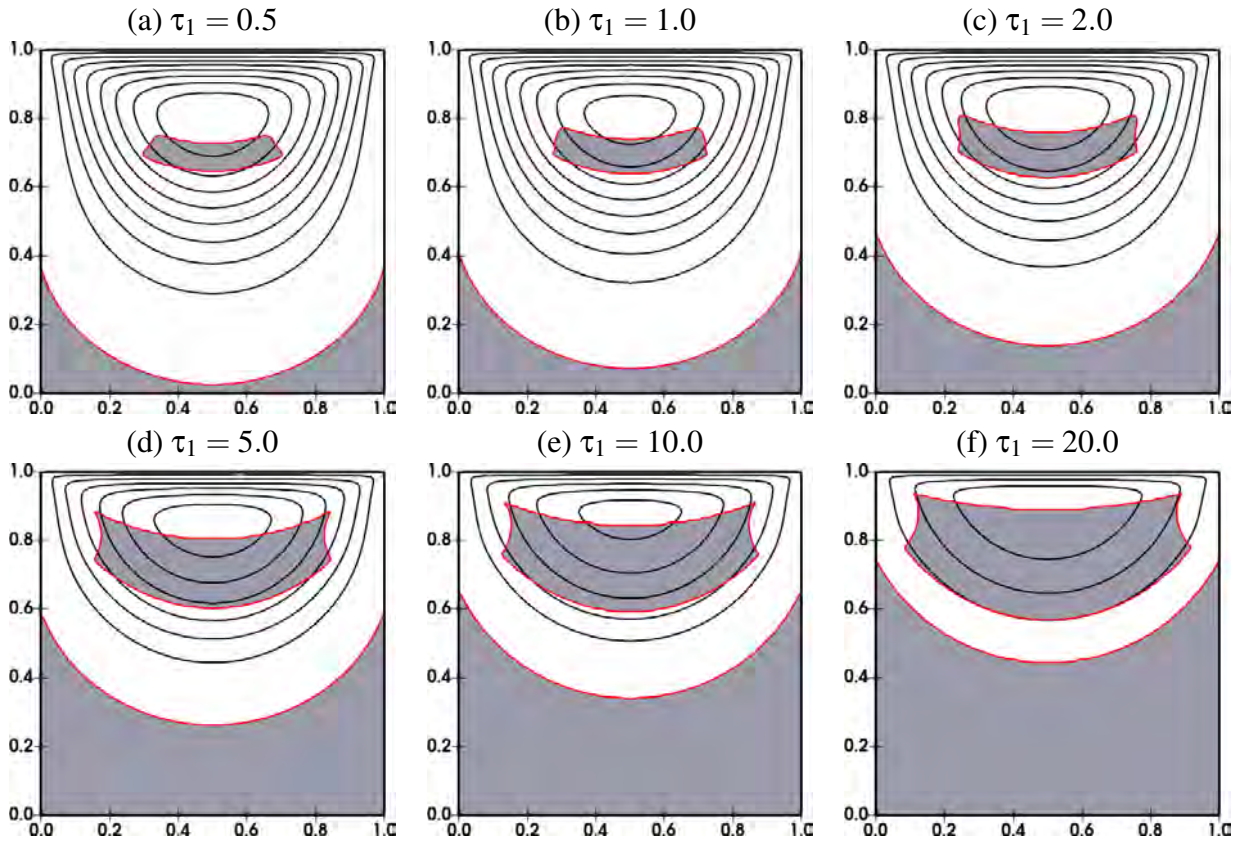


Figure 4: **Thixo-viscoplastic flow in lid-driven cavity:** The relative position of the plastic/rigid zone to streamline contours for an increased thixotropic yield stress parameter τ_1. The other parameters are set to $\eta_0 = 1.0$, $\eta_1 = 0.0$, $\tau_0 = 1.0$, $a = 1.0$ and $b = 0.1$. The Papanastasiou regularization parameter is set to $k = 10^4$. The solutions are calculated at mesh-refinement level 8.

In Table 5, we summarize the number of Newton-multigrid iterations, nonlinear number of iterations and the average number of multigrid iterations (N/M), w.r.t. different regularization parameters k and mesh refinement levels L for thixo-viscoplastic flow for different values of thixotropic yield stress parameters τ_1. The solutions are calculated using the continuations process w.r.t. regularization k.

Table 5 shows the mesh refinement independent of Newton-multigrid solver. Indeed, the nonlinearity of the problem is increased by increasing the thixotropic yield stress parameter τ_1, But, the slightly increases in the nonlinearity w.r.t. mesh refinement is due to the continuation process w.r.t. regularization parameter k used to obtain the solutions.

Table 5: **Thixo-viscoplastic flow in lid-driven cavity**: The number of Newton-multigrid iterations, nonlinear number of iterations and the average number of multigrid iterations (N/M), w.r.t. different regularization parameters k and mesh refinement levels L for thixo-viscoplastic flow for different values of thixotropic yield stress parameters τ_1. The solutions are calculated using the continuations process w.r.t. regularization k.

$k\backslash L$	5	6	7	5	6	7	5	6	7
	$\tau_1 = 0.5$			$\tau_1 = 1$			$\tau_1 = 2$		
1×10^1	5/2	5/3	6/2	5/2	5/2	9/1	5/2	5/2	9/1
1×10^2	4/1	4/2	5/1	4/1	4/2	7/1	4/2	4/2	8/1
1×10^3	4/1	4/1	4/1	4/2	4/2	8/1	4/4	6/1	7/1
1×10^4	4/1	4/2	4/2	5/1	7/1	4/1	7/1	10/1	8/2
	$\tau_1 = 5.0$			$\tau_1 = 10.0$			$\tau_1 = 20.0$		
1×10^1	6/2	6/2	10/1	11/1	8/2	11/1	10/1	9/2	11/1
1×10^2	4/2	5/2	11/1	10/1	5/3	8/1	12/1	6/3	10/1
1×10^3	5/2	9/1	10/1	10/1	9/1	7/1	8/2	9/1	9/2
1×10^4	5/1	5/2	5/1	8/3	7/1	5/1	8/2	7/1	9/1

6 SUMMARY

We presented a Newton-multigrid FEM solver for the quasi-Newtonian modeling approach for thixotropic flows. Based on a two-fields Stokes solver, we used higher order stable Q_2/P_1^{disc} FE approximations for velocity and pressure and higher order Q_2 FE approximation for the structure parameter field with appropriate stabilization term. The combination of a stable finite element approximations, Q_2/P_1^{disc}, for Stokes problem together with multigrid results in high numerically accurate, flexible and efficient FEM-multigrid solver. The nonlinearity is handled with generalized Newton's method w.r.t. the Jacobian's singularities having a global convergence property. For the numerical investigations; we used lid-driven cavity benchmark to find out the optimal setting, mesh refinement, and regularization. Indeed, we achieved a point-wise mesh convergence as well as a resolution barrier, (k,L) regularization mesh refinement level, beyond which no further resolution's improvement is possible. Furthermore, the solver shows a mesh refinement independency. For viscoplastic and thixo-viscoplastic solutions, we used the discrete continuation process w.r.t. regularization which might be integrated continuously within the solver in a black box manner.

Acknowledgments: The authors acknowledge the funding provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 446888252. Additionally, the authors acknowledge the financial grant provided by the Bundesministerium fr Wirtschaft und Energie aufgrund eines Beschlusses des Deutschen Bundestages through AiF-Forschungsvereinigung: Forschungs- Gesellschaft Verfahrens Technik e. V. - GVT under the IGF project number 20871 N. We would also like to gratefully acknowledge the support by LSIII and LiDO3 team at ITMC, TU Dortmund University, Germany.

REFERENCES

[1] Bruneau, C., Saad, M., The 2D lid-driven cavity problem revisited.*Computers & Fluids* (2006) **35**:326–348.

[2] Burgos, G. R., Alexandrou, A. N., Entov, V. Thixotropic rheology of semisolid metal suspensions. *J. Mater. Process. Technol.* (2001) **110**(2):164–176.

[3] Dullaert, K., Mewis, J. A Structural kinetics model for thixotropy. *J. Non-Newton. Fluid Mech.* (2006) **139**:21–30.

[4] Coussot, P., Nguyen, Q. D., Huynh, H. T., Bonn, D. Viscosity bifurcation in thixotropic, yielding fluids. *J. Rheol.* (2002) **46**(3):573–589.

[5] Damanik, H., Hron, J., Ouazzi, A., Turek, S. A monolithic FEM–multigrid solver for non–isothermal incompressible flow on general meshes. *J. Comp. Phys.* (2009) **228**:3869–3881.

[6] Houška, M. *Engineering aspects of the rheology of thixotropic liquids*. PhD thesis, Faculty of Mechanical Engineering, Czech Technical University of Prague, (1981).

[7] Kupperman, R. A central-difference scheme for a pure stream function formulation of incompressible viscous flow. *SIAM J. Sci. Comp.* (2001) **23**.

[8] Mandal, S., Ouazzi, A., Turek, S. Modified Newton solver for yield stress fluids, *Proceedings of ENUMATH 2015, the 11th European Conference on Numerical Mathematics and Advanced Applications*, Springer, (2016):481-490.

[9] Mujumdar, A., Beris, A. N., Metzner, A. B. Transient phenomena in thixotropic systems. *J. Nonnewton. Fluid Mech.* (2002) **102**(2):157–178.

[10] Ouazzi, A. *Finite Element Simulation of Nonlinear Fluids. Application to Granular Material and Powder*, Shaker Verlag, Achen (2006).

[11] Begum, N., Ouazzi, A., Turek, S. Monolithic Newton-multigrid FEM for the simulation of thixotropic flow problems.*Ergebnisberichte des Instituts fr Angewandte Mathematik Nummer 638, Fakultt fr Mathematik*, TU Dortmund University 634, (2021).

[12] Ouazzi, A., Begum, N., Turek, S. Newton-multigrid FEM solver for the simulation of quasi-newtonian modeling of thixotropic flows.*Ergebnisberichte des Instituts fr Angewandte Mathematik Nummer 638, Fakultt fr Mathematik*, TU Dortmund University 638, (2021).

[13] Papanastasiuo, T.C., Flow of materials with yield. *J. Rheol.*, (1987) **31**:385-404.

[14] Moller, P., Fall, A., Chikkadi, V. and Derks, D. and Bonn, D, An attempt to categorize yield stress fluid behaviour. *Phil Trans. R. Soc. A*, (2009) **367**:5139-5155.

[15] Turek, S., Ouazzi, A. Unified edge–oriented stabilization of nonconforming FEM for incompressible flow problems: Numerical investigations. *J. Numer. Math.* (2007) **15**(4):299–322.

[16] Worrall, W. E., Tuliani, S. Viscosity changes during the aging of clay-water suspensions. *Trans. Brit. Ceramic Soc.* (1964) **63**:167-185.

AN ADAPTIVE DISCRETE NEWTON METHOD FOR REGULARIZATION-FREE BINGHAM MODEL

A. Fatima*, S. Turek†, A. Ouazzi+ and M. A. Afaq‡

Institute for Applied Mathematics, LS III, TU Dortmund University,
Vogelpothsweg 87, 44227, Dortmund, Germany
e-mail: *arooj.fatima@math.tu-dortmund.de; †stefan.turek@math.tu-dortmund.de;
+abderrahim.ouazzi@math.tu-dortmund.de; ‡aaqib.afaq@math.tu-dortmund.de

Key words: Viscoplastic Fluids, Bingham Fluid, Divided Difference, FEM, Adaptive Newton Method, Regularization-Free

Abstract: *Developing a numerical and algorithmic tool which correctly identifies unyielded regions in yield stress fluid flow is a challenging task. Two approaches are commonly used to handle the singular behaviour at the yield surface, i.e. the Augmented Lagrangian approach and the regularization approach, respectively. Generally in the regularization approach, solvers do not perform efficiently when the regularization parameter gets very small. In this work, we use a formulation introducing a new auxiliary stress. The three field formulation of the yield stress fluid corresponds to a regularization-free Bingham formulation. The resulting set of equations arising from the three field formulation is solved efficiently and accurately by a monolithic finite element method. The velocity and pressure are discretized by the higher order stable FEM pair Q_2/P_1^{disc} and the auxiliary stress is discretized by the Q_2 element.*
Furthermore, this problem is highly nonlinear and presents a big challenge to any nonlinear solver. Therefore, we developed a new adaptive discrete Newton method, which evaluates the Jacobian with the divided difference approach. We relate the step length to the rate of the actual nonlinear reduction for achieving a robust adaptive Newton method. We analyse the solvability of the problem along with the adaptive Newton method for Bingham fluids by doing numerical studies for a prototypical configuration "viscoplastic fluid flow in a channel".

1 INTRODUCTION

A viscoplastic fluid is a viscous fluid with yield stress: a fluid that requires the applied stress above a certain non-zero limit of the yield stress to deform and to start flowing like a fluid. Below this non-zero limit of the yield stress the fluid behaves like a solid. The difference of this behaviour can be seen from the constitutive law of Bingham viscoplastic fluids.

$$\boldsymbol{\tau} = \begin{cases} 2\eta \mathbf{D}(\boldsymbol{u}) + \tau_s \dfrac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|} & \text{if } \|\mathbf{D}(\boldsymbol{u})\| \neq 0 \\ \|\boldsymbol{\tau}\| \leq \tau_s & \text{if } \|\mathbf{D}(\boldsymbol{u})\| = 0 \end{cases} \tag{1}$$

where $\mathbf{D}(\boldsymbol{u}) = \frac{1}{2}(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T)$ denotes the strain rate tensor, and τ_s denotes the yield stress. $\boldsymbol{\tau}$ is the stress tensor and η is the viscosity of the fluid. The Bingham model describes the nature of the viscoplastic fluids. These fluids are found in many practical applications, for example health/cosmetics (gels, creams, etc.), foods (yoghurt, butter, etc.), industrial (cement slurries, drilling mud, co-extrusion operations, etc.). One direct application is viscoplastic lubrication (hydraulic fracturing) and macro encapsulation [15]: heavy crude oil transportation along pipelines, coal-water slurry transportation and co-extrusion operations are examples of such lubrication. In this process, the stabilization of the interfaces in multi-layer shear flows [24] by means of viscoplastic fluids is the main interest. However, the accurate determination of yield surfaces is required. Developing a numerical and algorithmic tool which correctly

https://doi.org/10.4995/YIC2021.2021.12389

identifies unyielded regions in the flow is a challenging task. Indeed, to handle the singular behaviour at the yield surface leads researchers in the viscoplastic community to adopt two approaches. Firstly, the regularization approach [25, 11, 8] where the potentially "infinite" viscosity is replaced by a large finite effective viscosity making the yield surfaces dependent on the regularization. Secondly, the Augmented Lagrangian approach [12, 23] which is based on the exact yield stress model via a non-differential functional which is augmented with stabilization terms and typically solved iteratively using an Uzawa-type algorithm [7].

Generally in the regularization approach, solvers do not perform efficiently when the regularization parameter gets very small. In this work, we use a formulation introducing a new auxiliary stress [2]. The corresponding three-field formulation of yield stress fluids corresponds to a regularization-free Bingham model. The resulting saddle-point problem is solved efficiently and accurately by a monolithic finite element method.

2 GOVERNING EQUATIONS

It is difficult to model mathematically the Bingham constitutive law for viscoplastic fluids. The problem arises due to the non-differentiability of the viscosity in the constitutive law and needs to be treated in a special way. The Bingham constitutive law is given as follows

$$\boldsymbol{\tau} = \begin{cases} \left(2\eta + \dfrac{\tau_s}{\|\mathbf{D}(\boldsymbol{u})\|}\right)\mathbf{D}(\boldsymbol{u}) & \text{if } \|\mathbf{D}(\boldsymbol{u})\| \neq 0 \\ \|\boldsymbol{\tau}\| \leq \tau_s & \text{if } \|\mathbf{D}(\boldsymbol{u})\| = 0 \end{cases} \tag{2}$$

with non-linear viscosity:

$$\eta(\|\mathbf{D}(\boldsymbol{u})\|) = 2\eta + \frac{\tau_s}{\|\mathbf{D}(\boldsymbol{u})\|} \tag{3}$$

The problem of differentiability arises when the viscosity becomes infinite in the rigid zone, i.e. $\|\mathbf{D}(\boldsymbol{u})\| = 0$. Therefore, one approach is to use regularization to overcome this problem. The purpose is to make the viscosity smooth and differentiable over the whole domain. There are various regularization models in the literature. Allouche et al. [1] introduced a regularization parameter simply added in the denominator. Bercovier and Engelman [3] and Tanner et al. [19] proposed different regularization functions. Papanastasiou [21] introduced an exponential expression in the regularization model to hold for any shear rate by adding a small parameter. The corresponding Navier-Stokes equations for the steady incompressible flow reads

$$\begin{cases} -\nabla \cdot \boldsymbol{\tau} + \nabla p = 0 & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \\ \boldsymbol{u} = \boldsymbol{g}_D & \text{on } \Gamma_D \end{cases} \tag{4}$$

where $\boldsymbol{\tau}$ is stress tensor from (2) with regularized viscosity. We have already discussed above that the rigid zone produces a singularity and to overcome this problem, we use the Bercovier and Engelman regularization in this work. The real viscoplastic solution can only be achieved when the regularization parameter is very small ($\epsilon \to 0$) but this situation is difficult for the numerical solver. We proceed within the framework of a three-field Stokes problem, by introducing a new auxiliary stress [2] as follows:

$$\boldsymbol{\sigma} = \frac{\mathbf{D}(\boldsymbol{u})}{\|\mathbf{D}(\boldsymbol{u})\|_\epsilon} \tag{5}$$

Then, the three-field $(\boldsymbol{u}, \boldsymbol{\sigma}, p)$ system of Bingham fluid flow equations is given as follows:

$$\begin{cases} \|\mathbf{D}(\boldsymbol{u})\|_{\epsilon}\, \boldsymbol{\sigma} - \mathbf{D}(\boldsymbol{u}) = 0 & \text{in } \Omega \\ -\nabla \cdot (2\eta \mathbf{D}(\boldsymbol{u}) + \tau_s \boldsymbol{\sigma}) + \nabla p = 0 & \text{in } \Omega \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } \Omega \\ \boldsymbol{u} = \boldsymbol{g}_D & \text{on } \Gamma_D \end{cases} \tag{6}$$

System (6) represents the mixed formulation, which solves the regularized as well as the regularization-free Bingham problem, i.e. for $\epsilon = 0$. The numerical studies shown in the next sections describe the advantages of the formulation, particularly that we can achieve a true viscoplastic solution by solving a regularization-free Bingham model.

3 FINITE ELEMENT METHOD

The finite element method is chosen for the discretization in space. The strong form of the system of equations in (6) is converted into the weak formulation by multiplying it with the test functions and integrated over the whole domain. We consider three test functions $\boldsymbol{v}, q$ and $\boldsymbol{\tau}$, and multiply then with the system of equations (6). The resulting weak forms reads after partial integration:

$$\begin{aligned} \int_\Omega \Big(\|\mathbf{D}(\boldsymbol{u})\|_{\epsilon}\, \boldsymbol{\sigma} : \boldsymbol{\tau} \Big)\, dx - \int_\Omega \Big(\mathbf{D}(\boldsymbol{u}) : \boldsymbol{\tau} \Big)\, dx = 0 \quad \text{in } \Omega \\ \int_\Omega \Big(2\eta \mathbf{D}(\boldsymbol{u}) : \mathbf{D}(\boldsymbol{v}) \Big)\, dx + \int_\Omega \tau_s \Big(\boldsymbol{\sigma} : \mathbf{D}(\boldsymbol{v}) \Big)\, dx - \int_\Omega p\, \nabla \cdot \boldsymbol{v}\, dx = 0 \quad \text{in } \Omega \\ \int_\Omega q\, \nabla \cdot \boldsymbol{u}\, dx = 0 \quad \text{in } \Omega \end{aligned} \tag{7}$$

Let $\mathbb{V} = \boldsymbol{H}_0^1(\Omega) := (H_0^1(\Omega))^2$, $\mathbb{Q} = L_0^2(\Omega)$, and $\mathbb{M} = (L^2(\Omega))^{2\times 2}_{\text{sym}}$ be the spaces for the velocity, pressure and stress, respectively, associated with $\|.\|_{1,\Omega}$ and $\|.\|_{0,\Omega}$. Let $\mathbb{V}'$, $\mathbb{Q}'$, and $\mathbb{M}'$ be their corresponding dual spaces:

We introduce the approximation spaces:

$$\begin{aligned} \mathbb{V}^h &= \left\{ \boldsymbol{v}_h \in \mathbb{V}, \boldsymbol{v}_{h|K} \in (Q_2(K))^2 \right\} \\ \mathbb{M}^h &= \left\{ \boldsymbol{\tau}_h \in \mathbb{M}, \boldsymbol{\sigma}_{h|K} \in (Q_2(K))^{2\times 2} \right\} \\ \mathbb{Q}^h &= \left\{ q_h \in \mathbb{Q}, q_{h|K} \in P_1^{\text{disc}}(K) \right\} \end{aligned} \tag{8}$$

Velocity, stress and pressure are discretized using $Q_2, Q_2, P_1^{\text{disc}}$ finite elements [4], respectively, as shown in Figure 1.

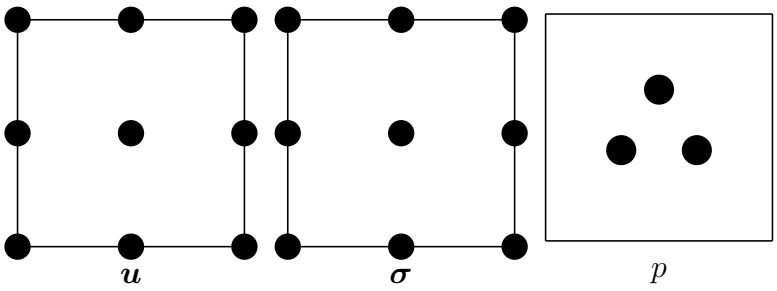


Figure 1: Finite elements Q_2, Q_2, P_1^{disc} for velocity, stress and pressure, respectively, on each quadrilateral

However, in the rigid zone $\|D\| = 0$, the finite element space $\mathbb{V}^h$ and $\mathbb{M}^h$ do not satisfy the LBB condition, the remedy is an apropriate stabilization technique. The following jump term might be added [20, 26]

$$j_{\boldsymbol{u}}(\boldsymbol{u}_h, \boldsymbol{v}_h) = \sum_{E \in \mathcal{E}_i} \gamma_u h \int_E [\nabla \boldsymbol{u}_h] : [\nabla \boldsymbol{v}_h] \; d\Omega \tag{9}$$

where γ_u is a constant parameter and h is the mesh size.

4 ADAPTIVE DISCRETE NEWTON

The problem (6) is highly nonlinear and presents a big challenge to any nonlinear solver. Iterative solvers, e.g. Newton and the fixed point iteration method, are used to solve such nonlinear problems in fluid dynamics. Since the Newton method usually has a faster convergence rate than the fixed point method, it is preferred in most of the cases but it is also very sensitive regarding the initial guess of the solution and depends strongly on the properties of the Jacobian matrices during the iterations. The Newton method solves the nonlinear steady system from (6) by the following steps:

Algorithm 1: Newton method solver

- Provide the input parameters, e.g. tolerance, parameters of the non linear solver, initial guess and the iteration number n
- Repeat until the tolerance is achieved
- Calculate the residual $\mathcal{R}(\mathcal{U}^n) = A\,\mathcal{U}^n - b$
- Build the Jacobian $J(\mathcal{U}^n) = \frac{\partial \mathcal{R}(\mathcal{U}^n)}{\partial \mathcal{U}^n}$
- Solve $J(\mathcal{U}^n)\;\delta\mathcal{U}^n = \mathcal{R}(\mathcal{U}^n)$
- Find the optimal value of the damping factor $\omega^n \in (-1, 0]$
- Approximate $\mathcal{U}^{n+1} = \mathcal{U}^n - \omega^n\;\delta\mathcal{U}^n$

The initial guess should be close to the final solution for achieving faster convergence. There are also some other factors in the Newton method which should be taken into account for the numerical stability, e.g. a damping factor when the solution is non-smooth. In our work, this factor is calculated by a root finding technique called line search method [9, 22]. First, the system of nonlinear equations is linearised using the Newton method, where $\mathcal{U} = (\boldsymbol{u}, \boldsymbol{\sigma}, p)$ and $\mathcal{R}_{\mathcal{U}}$ denote the discrete residuals. One Newton iteration reads:

$$\begin{bmatrix} \boldsymbol{u}^{n+1} \\ \boldsymbol{\sigma}^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}^{n} \\ \boldsymbol{\sigma}^{n} \\ p^{n} \end{bmatrix} - \omega_n \begin{bmatrix} \frac{\partial R_{\boldsymbol{u}}(\mathcal{U}^n)}{\partial \boldsymbol{u}} & \frac{\partial R_{\boldsymbol{u}}(\mathcal{U}^n)}{\partial \boldsymbol{\sigma}} & \frac{\partial R_{\boldsymbol{u}}(\mathcal{U}^n)}{\partial p} \\ \frac{\partial R_{\boldsymbol{\sigma}}(\mathcal{U}^n)}{\partial \boldsymbol{u}} & \frac{\partial R_{\boldsymbol{\sigma}}(\mathcal{U}^n)}{\partial \boldsymbol{\sigma}} & \frac{\partial R_{\boldsymbol{\sigma}}(\mathcal{U}^n)}{\partial p} \\ \frac{\partial R_{p}(\mathcal{U}^n)}{\partial \boldsymbol{u}} & \frac{\partial R_{p}(\mathcal{U}^n)}{\partial \boldsymbol{\sigma}} & \frac{\partial R_{p}(\mathcal{U}^n)}{\partial p} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{R}_{\boldsymbol{u}}(\mathcal{U}^n) \\ \mathcal{R}_{\boldsymbol{\sigma}}(\mathcal{U}^n) \\ \mathcal{R}_{p}(\mathcal{U}^n) \end{bmatrix} \tag{10}$$

In the Newton method, first derivatives of the residual are needed in every nonlinear iteration called Jacobian matrix. The Jacobian is either calculated analytically or approximated by the

divided difference method. The advantage of the approximation of the Jacobian is that this method acts in a black box manner so that it allows any nonlinear equations to be handled automatically without having to derive the corresponding calculations [5, 6]. In this work the Jacobian matrix is not computed exactly, instead its approximation is computed using divided differences and the corresponding j-th column is given as follows

$$\left[\frac{\partial \mathcal{R}(\mathcal{U}^n)}{\partial \mathcal{U}^n}\right]_j \approx \frac{\mathcal{R}(\mathcal{U}^n + \chi\delta_j) - \mathcal{R}(\mathcal{U}^n - \chi\delta_j)}{2\chi} \tag{11}$$

where δ_j is the vector with unit j-th component and zero otherwise. The parameter χ can be fixed or can be modified according to some norm of the solution $\|\mathcal{U}^n\|$ or the norm of the update in the previous step, i.e., $\|\delta\mathcal{U}^{n-1}\|$. The advantage of this approximation is that we don't need any knowledge of the Jacobian a priori. However, in this method, the step-length χ is a "free" parameter and the right choice might be a delicate task. Based on the perturbation analysis for the residum, it is often chosen according to the machine precision [14]. On the other hand, the sensitivity study of the nonlinear behavior of power law models w.r.t. the step-length parameter χ, the mesh width h and the strength of the nonlinearity suggest an adaptive choice [13, 18]. Indeed, choosing χ too big leads to the loss of the advantageous quasi-quadratic convergence behaviour, while very small parameter values for χ can lead to divergence, due to numerical instabilities. So, a process allowing for bigger step-length parameter χ is worthy for removing numerical instability. Loosely speaking, bigger step-length parameter χ increases the set of admissible Jacobian for nonregular solutions. As a result, there are thresholds of the residuum's norm which can be used for the choice of the step-length parameter χ as a step function. In order to relate continuously these thresholds of the residuum's norm to the successive nonlinear reduction

$$r_n = \frac{\|\mathcal{R}(\mathcal{U}^n)\|}{\|\mathcal{R}(\mathcal{U}^{n-1})\|} \tag{12}$$

we use the characteristic function introduced in [17]

$$f(r_n) = 0.2 + \frac{0.4}{0.7 + \exp\left(1.5 r_n\right)} \tag{13}$$

or the slightly modified ones introduced in [16]. A new adaptive step-length strategy is considered as follows

$$\chi_{n+1} = f^{-1}(r_n)\chi_n \tag{14}$$

5 NUMERICAL RESULTS

We analyse the solvability of the problem along with the adaptive Newton method for Bingham fluids by doing numerical studies for a prototypical configuration, i.e. "viscoplastic fluid flow in a channel".

5.1 Bingham viscoplastic fluid flow in channel

The two dimensional channel domain is considered as a domain between two parallel plates with h length apart and long. The problem is solved under the assumption of Dirichlet boundary conditions on the domain $\overline{\Omega} = [0, h]^2$ according to following analytical solution:

$$u_1 = \begin{cases} \frac{1}{8}\left[(h - 2\tau_s)^2 - (h - 2\tau_s - 2y)^2\right] & 0 \le y < \frac{h}{2} - \tau_s \\ \frac{1}{8}(h - 2\tau_s)^2 & \frac{h}{2} - \tau_s \le y \le \frac{h}{2} + \tau_s \\ \frac{1}{8}\left[(h - 2\tau_s)^2 - (2y - 2\tau_s - h)^2\right] & \frac{h}{2} + \tau_s < y \le h \end{cases} \tag{15}$$

$u_2 = 0$ and $p = -x + c$ [10]. The viscosity is set to be $\eta = 1$, the body force is $\boldsymbol{f} = 0$ and $h = 1$ is considered. The rigid zone is the region of constant velocity, i.e.

$$\frac{h}{2} - \tau_s \leq y \leq \frac{h}{2} + \tau_s \tag{16}$$

A comparison study is carried out between the new discrete adaptive Newton strategy and the classical Newton for the primitive variable formulation of a Bingham fluid in a channel flow. Applying both methods the number of nonlinear iterations is presented in Table 1. For the coarse refinement level (L=2 in the present case), starting with the zero solution as an initial guess, we perform Newton iterations until the tolerance is achieved. However, the next refinement level takes the solution from the previous refinement level as an initial solution. For the first test, we choose the yield stress value to be $\tau_s = 0.23$ because this value is aligned with the coarse mesh. It is observed that the primitive variable formulation along with the classical Newton method faces difficulties in convergence when the regularization parameter $\epsilon \to 0$. On the other hand, the adaptive Newton solver is able to converge even for very small values of ϵ, exhibiting the advantages of our newly developed solver. Moreover, it shows a good speed of convergence for all cases of regularized Bingham fluid. Testing the efficiency of the three-

Table 1: **Regularized viscosity approach in primitive variable $(\boldsymbol{u}, p)$:** Number of iterations of the nonlinear solver in a channel flow at yield stress $\tau_s = 0.23$ for the adaptive Newton and the classical Newton at different mesh refinement level L, the stopping criterion is 10^{-6}, "-" indicates that the simulation did not converged.

$\downarrow$ L/$\epsilon \rightarrow$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0
	Newton						**Adaptive Newton**					
3	2	3	-	-	-	-	4	4	5	5	9	-
4	2	3	-	-	-	-	4	4	5	5	9	-
5	2	3	-	-	-	-	4	4	6	5	9	-

Table 2: **Regularization-free three-field formulation:** Number of iterations of the nonlinear solver in a channel flow at yield stress $\tau_s = 0.23$ for the adaptive Newton and the classical Newton at different mesh refinement level L, the stopping criterion is 10^{-6}.

$\downarrow$ L/$\epsilon \rightarrow$	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0
	Newton						**Adaptive Newton**					
3	2	3	4	6	9	1	2	2	2	5	1	2
4	2	3	4	8	9	1	1	2	2	4	2	2
5	1	2	3	9	5	2	1	1	1	1	3	1

field formulation for the unregularized Bingham problem, a numerical study is carried out for both of the Newton strategies shown in Table 2. The efficiency of the three-field formulation and the robustness of the adaptive strategy for the discrete Newton is showcased successfully. The yield stress value is kept similar, i.e. $\tau_s = 0.23$ as in Table 1. Simulations are performed for different values of regularization parameter ϵ starting from 10^{-1} to 10^{-5} and then also for regularization-free Bingham $\epsilon = 0$. Figure 2 shows the velocity, pressure and norm of the strain rate tensor $\|\mathbf{D}(\boldsymbol{u})\|$ contours at refinement level L=5 ($h_x = 1/32, h_y = 1/96$) for regularization-free Bingham. The pressure distribution is different inside and outside of the rigid zone. It

shows a discontinuity near the interface and the distribution mainly depends on the yield stress value τ_s [10].

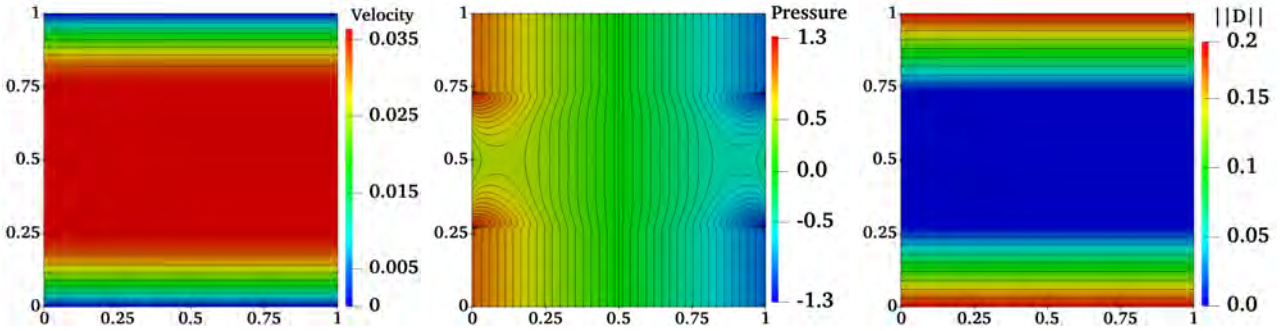


Figure 2: Visualization of the velocity contours, pressure and $\|\mathbf{D}(\boldsymbol{u})\|$ for the non-regularized Bingham fluid flow in a channel with $\tau_s = 0.23$ at refinement level L=5 ($h_x = 1/32, h_y = 1/96$).

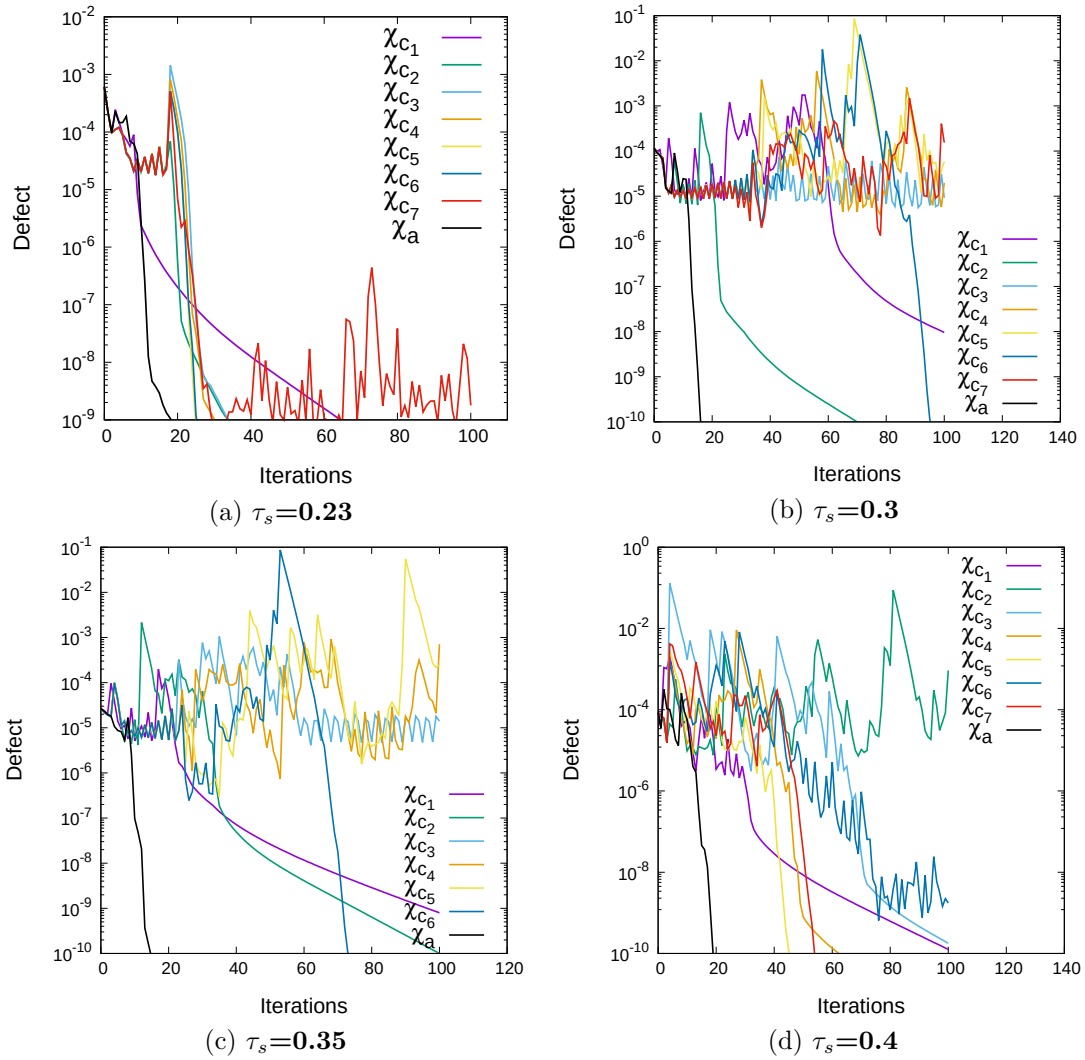
It can be seen from the figure that the three-field formulation accurately predicts the division of the rigid and fluid zone and provides the true solutions of the problem. Moreover, the formulation can be solved exactly irrespective of the Newton solver type (classical or adaptive). Figure 3 plots the comparison of the presented discrete adaptive Newton with the classical approach. When the length χ of the Jacobian approximation in the Newton method is chosen as constant the solver either converges very slowly or it starts to oscillate. In our adaptive Newton, χ changes dynamically between the iterations. Initially it is relaxed and once the solution enters the radius of convergence then χ gets smaller to achieve the accuracy of the solution. To highlight the efficiency and robustness of our newly developed solver, the yield stress value is increased from $\tau_s = 0.23$ to $0.3, 0.35$ and $\tau_s = 0.4$. All of these tests are carried out for the regularization-free Bingham case and the solver shows fast convergence by dynamically adapting the step-length during the iterations.

6 CONCLUSIONS

A new adaptive Newton and regularization-free solver for yield stress fluids is developed. Firstly, by introducing a new auxiliary stress in a three-field formulation. The resulting saddle-point problem is solved with a monolithic finite element method to simulate viscoplastic flows for the correct prediction of the yielded surfaces. The advantage of this formulation is achieving a true non-regularized viscoplastic solution, i.e. $\epsilon = 0$, efficiently and accurately. The method does not effect the shape of the yield surfaces. Secondly, a robust and accurate new adaptive discrete Newton method is developed, which evaluates the Jacobian matrix with the divided difference approach and converges faster as compared to classical Newton. We have carried out several numerical experiments for a benchmark problem. This experiment shows that the number of nonlinear iterations is significantly reduced for the three-field formulation with the combination of our newly developed adaptive discrete Newton method.

ACKNOWLEDGEMENTS

We would like to thank the Deutsche Forschungsgemeinschaft (DFG) for their financial support under the DFG Priority Program SPP 1962. The authors also acknowledge the support by LS3 and LiDO3 team at ITMC, TU Dortmund University.



(a) τ_s=**0.23** (b) τ_s=**0.3**

(c) τ_s=**0.35** (d) τ_s=**0.4**

Figure 3: **Nonlinear convergence w.r.t χ for adaptive-Newton method**: The norm of the residual versus number of iterations w.r.t two strategies (constant and adaptive) at refinement level L=2 ($h_x = 1/4, h_y = 1/12$) with the constant χ strategy (set as $\chi_{c_1} = 10^{-1}, \chi_{c_2} = 10^{-2}$,..., $\chi_{c_7} = 10^{-7}$) and the adaptive strategy (χ_a changing w.r.t non linear residuum reduction).

REFERENCES

[1] Allouche, M., Frigaard, I. A., and Sona, G. Static wall layers in the displacement of two visco-plastic fluids in a plane channel. *Journal of Fluid Mechanics 424* (2000), 243277.

[2] Aposporidis, A., Haber, E., Olshanskii, M., and Veneziani, A. A mixed formulation of the bingham fluid flow problem: Analysis and numerical solution. *Computer Methods in Applied Mechanics and Engineering 200* (2011), 2434–2446.

[3] Bercovier, M., and Engelman, M. A finite-element method for incompressible non-newtonian flows. *Journal of Computational Physics 36*, 3 (1980), 313–326.

[4] Boffi, D., and Gastaldi, L. On the quadrilateral q2p1 element for the stokes problem. *International Journal for Numerical Methods in Fluids 39*, 11 (2002), 1001–1011.

[5] Damanik, H. *FEM Simulation of Non-isothermal Viscoelastic Fluids*. TU Dortmund, Germany, 2011. PhD Thesis.

[6] Damanik, H., Hron, J., Ouazzi, A., and Turek, S. Monolithic Newton-multigrid solution techniques for incompressible nonlinear flow models. *International Journal for Numerical Methods in Fluids 71* (2012), 208–222.

[7] Dean, E., Glowinski, R., and Guidoboni, G. On the numerical simulation of bingham visco-plastic flow: Old and new results. *Journal of Non-Newtonian Fluid Mechanics 142* (2007), 36–62.

[8] Dean, E. J., and Glowinski, R. Operator-splitting methods for the simulation of bingham visco-plastic flow. *Chinese Annals of Mathematics 23* (2012).

[9] Dennis, J. E., and Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1996.

[10] El-Borhamy, M. *Numerical Simulation for viscoplastic fluids via finite element methods*. TU Dortmund, Germany, 2012. PhD Thesis.

[11] Frigaard, I. A., and Nouar, C. On the usage of viscosity regularisation methods for visco-plastic fluid flow computation. *Journal of Non-Newtonian Fluid Mechanics 127* (2005), 1–26.

[12] Glowinski, R., and Le Tallec, P. *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*. Society for Industrial and Applied Mathematics, 1989.

[13] Hron, J., Ouazzi, A., and Turek, S. A computational comparison of two FEM solvers for nonlinear incompressible flow. In *Lecture Notes in Computational Science and Engineering* (2003), vol. 35, Springer, pp. 87–109. New York.

[14] Kelley, C. T. *Iterative methods for linear and nonlinear equations*. SIAM, Philadelphia, 1995.

[15] Maleki, A., Hormozi, S., Roustaei, A., and Frigaard, I. A. Macro-size drop encapsulation. *Journal of Fluid Mechanics 769* (2015), 482521.

[16] Mandal, S. *Efficient FEM solver for quasi-Newtonian flow problems with application to granular material*. TU Dortmund, Germany, 2016. PhD Thesis.

[17] Mandal, S., Ouazzi, A., and Turek, S. Modified Newton solver for yield stress fluids. In *Proceedings of ENUMATH 2015, the 11th European Conference on Numerical Mathematics and Advanced Applications* (2016), Springer, pp. 481–490.

[18] Mandal, S., Turek, S., Schwarze, R., Haustein, M., Ouazzi, A., and Gladky, A. Numerical benchmarking of granular flow with shear dependent incompressible flow models. *Journal of Non-Newtonian Fluid Mechanics 262* (2018), 92–106.

[19] O'Donovan, E., and Tanner, R. Numerical study of the bingham squeeze film problem. *Journal of Non-Newtonian Fluid Mechanics 15*, 1 (1984), 75–83.

[20] Ouazzi, A. *Finite Element Simulation of Nonlinear Fluids: Application to Granular Material and Powder*. Industrial and applied mathematics. Shaker, 2005.

[21] Papanastasiou, T. C. Flows of materials with yield. *Journal of Rheology 31*, 5 (1987), 385–404.

[22] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[23] Roquet, N., and Saramito, P. An adaptive finite element method for bingham fluid flows around a cylinder. *Computer Methods in Applied Mechanics and Engineering 192*, 31 (2003), 3317–3341.

[24] Sarmadi, P., Mierka, O., Turek, S., Hormozi, S., and Frigaard, I. A. Three dimensional simulation of flow development of triple-layer lubricated pipeline transport. *Journal of Non-Newtonian Fluid Mechanics 274* (2019), 104201.

[25] Schmitt, H. Numerical simulation of bingham fluid flow using prox-regularization. *Journal of Optimization Theory and Applications 106* (2000), 603–626.

[26] Turek, S., and Ouazzi, A. Unified edge–oriented stabilization of nonconforming FEM for incompressible flow problems: Numerical investigations. *Journal of Numerical Mathematics 15*, 4 (2007), 299–322.

MONOLITHIC NEWTON-MULTIGRID SOLVER FOR MULTIPHASE FLOW PROBLEMS WITH SURFACE TENSION

M. A. Afaq*, S. Turek†, A. Ouazzi+ and A. Fatima‡

Institute for Applied Mathematics, LS III
TU Dortmund University,
Dortmund, Germany

e-mail: *aaqib.afaq@math.tu-dortmund.de; †stefan.turek@math.tu-dortmund.de;
+abderrahim.ouazzi@math.tu-dortmund.de; ‡arooj.fatima@math.tu-dortmund.de

Key words: Multiphase Flow, Curvature Free, Level Set Method, Cut-Off Material Function, Finite Element Method, Navier-Stokes Equations

Abstract: *We have developed a monolithic Newton-multigrid solver for multiphase flow problems which solves velocity, pressure and interface position simultaneously. The main idea of our work is based on the formulations discussed in [14], where it points out the feasibility of a fully implicit monolithic solver for multiphase flow problems via two formulations, a curvature free level set approach and a curvature free cut-off material function approach. Both formulations are fully implicit and have the advantages of requiring less regularity, since neither normals nor curvature are explicitly calculated, and no capillary time restriction has to be respected. Furthermore, standard Navier-Stokes solvers might be used, which do not have to take into account inhomogeneous force terms. The reinitialization issue is integrated within the formulations.*
The nonlinearity is treated with a Newton-type solver with divided difference evaluation of the Jacobian matrices. The resulting linearized system inside of the outer Newton solver is a typical saddle point problem which is solved using a geometrical multigrid method with Vanka-like smoother using higher order stable Q_2/P_1^{disc} FEM for velocity and pressure and Q_2 for all other variables. The method is implemented into an existing software package for the numerical simulation of multiphase flows (FeatFlow). The robustness and accuracy of this solver is tested for two different test cases, static bubble and oscillating bubble, respectively.

1 INTRODUCTION

Multiphase flows are of great interest in different industrial and engineering applications. The simplest example of multiphase flow is two-phase flow [6], two fluids/phases are separated by an interface, where surface tension forces are applied. If the fluids have different densities and viscosities, then a discontinuous pressure jump is observed near the interface. The interface moves/deforms due to the flow movement, and to capture this behaviour, an efficient tracking/capturing method should be applied. There are two main methods for interface modeling in the multiphase flow problems, i.e. Lagrangian and Eulerian methods. The volume of fluid (VOF) [8], phase field [1, 2] and level set [13, 15] are among the most famous Eulerian interface capturing methods, which are very favourable for the computational and implementation point of view.

In the present work, numerical methods based on the level set and material cut-off function for two-dimensional incompressible two phase flow are implemented. The approximation of the surface tension force does not require the calculation of the curvature, normals and the delta function. However, the level set method requires some sort of redistancing [17]. In our improved fully implicit level set method, the explicit redistancing is removed by integrating the reinitialization term within the formulations [14]. Moreover, the advantage of this approach is that there is no capillary time restriction and the standard Navier-Stokes solver can simulate

https://doi.org/10.4995/YIC2021.2021.12390

the multiphase problems with homogeneous force terms. The numerical studies are carried out for two different test cases, static bubble and oscillating bubble, which show the accuracy and robustness of these formulations in the context of FEM. The system of equations in each formulation is solved monolithically (in a fully coupled manner).

2 GOVERNING EQUATIONS

In the methodology of our work, first the Continuum Surface Force (CSF) [5] is introduced and then it is linked to the classical Continuum Surface Stress (CSS) [10]. In the CSF approach, the interface between the fluids is not considered as a sharp discontinuity but as a smooth transition. As a result, the surface tension is also assumed to be continuous everywhere in the transition regime. A detailed discussion of this approach can be found in [4]. In the CSS approach, a stress tensor is introduced and the surface force term is written as the divergence of the stress tensor [4]. The following formulations in section (2.1) and (2.2) are based on the CSS approach.

2.1 Curvature free level set approach

The curvature free level set formulation [14] is introduced by adding a tensor field in the Navier-Stokes equations. The system of equations is defined as:

$$\begin{cases} \rho(\psi)\Big(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\Big) - div\boldsymbol{\tau} + \nabla p = 0, & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{u} = 0, & \text{in } \Omega, \\ \dfrac{\partial \phi}{\partial t} + \boldsymbol{u}\cdot\nabla\phi = 0, & \text{in } \Omega, \\ \psi - \Big(\dfrac{-1}{1+exp(\frac{\phi}{\epsilon_\psi})} + 0.5\Big) = 0, & \text{in } \Omega. \end{cases} \tag{1}$$

Here, ρ is the density, $\boldsymbol{u}$ is the velocity, $\boldsymbol{\tau} = (\boldsymbol{\tau}_s + \boldsymbol{\tau}_m)$ is the full stress tensor, p is the pressure, ϕ is the level set function, ψ is the cut-off function and ϵ_ψ is the parameter for the interface thickness. The standard stress tensor $\boldsymbol{\tau}_s$ and the modified stress $\boldsymbol{\tau}_m$ (derived in [14]) are defined as

$$\boldsymbol{\tau}_s = 2\mu(\psi)\mathbf{D}(\boldsymbol{u}), \quad \boldsymbol{\tau}_m = -\sigma\Big(\frac{\nabla\psi\otimes\nabla\psi}{\|\nabla\psi\|}\Big), \tag{2}$$

where μ is the viscosity, $\mathbf{D}(\boldsymbol{u}) = \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T)$ is the deformation stress tensor and σ is the surface tension coefficient. The modified pressure p_m balances the pressure peaks at the interface. The mathematical expression is defined as

$$\nabla p_m = \nabla p - \nabla(\sigma\|\nabla\psi\|). \tag{3}$$

In order to circumvent the explicit reinitialization, the additional normal diffusion term can be integrated into the level set equation (1). The complete system of equations is defined as follows:

$$\begin{cases} \rho(\psi)\Big(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\Big) - \nabla\cdot\Big(2\mu(\psi)\mathbf{D}(\boldsymbol{u}) + \sigma\Big(\dfrac{\nabla\psi\otimes\nabla\psi}{\|\nabla\psi\|}\Big)\Big) + \nabla p = 0, & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{u} = 0, & \text{in } \Omega, \\ \dfrac{\partial \phi}{\partial t} + \boldsymbol{u}\cdot\nabla\phi - \nabla\cdot\Big(\gamma_{nd}\Big(\dfrac{\nabla\phi}{\|\nabla\phi\|}\cdot\nabla\phi - 1\Big)\dfrac{\nabla\phi}{\|\nabla\phi\|}\Big) = 0, & \text{in } \Omega, \\ \psi - \Big(\dfrac{-1}{1+exp(\frac{\phi}{\epsilon_\psi})} + 0.5\Big) = 0, & \text{in } \Omega. \end{cases} \tag{4}$$

Here, γ_{nd} is the relaxation parameter for normal diffusion and this term has the forward and backward diffusion property [11]. The system of equations (4) is a four field system with the unknowns $(\boldsymbol{u}, \phi, \psi, p)^T$. The two main advantages of this formulation are that neither normals nor curvature have to be explicitly computed, which are the sources of numerical errors.

2.2 Curvature free cut-off material function approach

In this approach, we are no longer in the need of the level set function, instead we use a new equation for the cut-off material function. Olsson and Kreiss [12] have introduced the equation for the material cut-off function with fictitious time as

$$\frac{\partial \psi}{\partial \tau} + \nabla \cdot \Big(\gamma_{nc}\psi(1-\psi)\boldsymbol{n}\Big) - \nabla \cdot \Big(\gamma_{nd}(\nabla\psi \cdot \boldsymbol{n})\boldsymbol{n}\Big) = 0, \tag{5}$$

where γ_{nc} and γ_{nd} are the relaxation parameters for the nonlinear convection in the normal direction and the normal diffusion, respectively. The nonlinear convection in the direction of the normal has the tendency to build the Heaviside step function, without depending on the convective parameter. Whereas, the sharpness of the interface is controlled by the normal diffusion. The full set of equations including the material cut-off function in physical time is defined as follows:

$$\begin{cases} \rho(\psi)\Big(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\Big) - \nabla\cdot\Big(2\mu(\psi)\mathbf{D}(\boldsymbol{u}) + \sigma\Big(\dfrac{\nabla\psi\otimes\nabla\psi}{\|\nabla\psi\|}\Big)\Big) + \nabla p = 0, \quad \text{in } \Omega, \\ \nabla\cdot\boldsymbol{u} = 0, \quad \text{in } \Omega, \\ \dfrac{\partial\psi}{\partial t} + \boldsymbol{u}\cdot\nabla\psi + \nabla\cdot\Big(\gamma_{nc}\psi(1-\psi)\dfrac{\nabla\psi}{\|\nabla\psi\|}\Big) - \nabla\cdot\Big(\gamma_{nd}\Big(\nabla\psi\cdot\dfrac{\nabla\psi}{\|\nabla\psi\|}\Big)\dfrac{\nabla\psi}{\|\nabla\psi\|}\Big) = 0, \quad \text{in } \Omega. \end{cases} \tag{6}$$

The system of equations (6) is a three field system with the unknowns $(\boldsymbol{u}, \psi, p)^T$. The momentum equation has homogeneous force terms. We are solving the systems in a fully coupled manner with our monolithic multiphase flow solver.

3 NUMERICAL METHOD

For solving the system of equations (4) and (6), first we discretize in time with a fully implicit 2nd order time stepping scheme, i.e. Crank Nicolson. For the space discretization, the velocity and pressure fields are discretized using higher order stable Q_2/P_1^{disc} FEM [3] and Q_2 for the level set function as well as for the cut-off material function, presented in Fig. 1. The nonlinearity in the system of equations (4) and (6) is treated by Newton solver and the resulting linear system is then solved using multigrid solver. Our Newton-multigrid solver is fully monolithic, which means it solves all the variables $(\boldsymbol{u}, \phi, \psi, p)$ simultaneously.

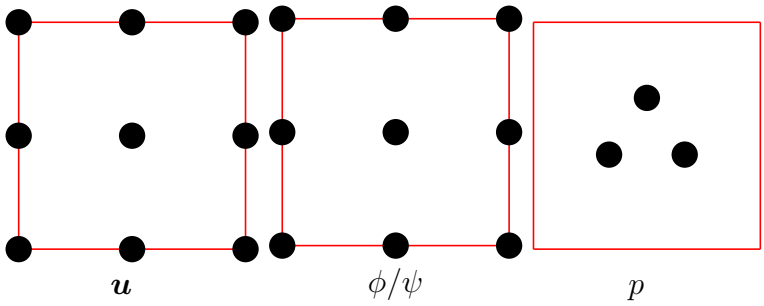


Figure 1: Higher order finite element $Q_2, Q_2, P_1^{\text{disc}}$ on quadrilaterals.

4 NUMERICAL RESULTS

Two numerical studies are performed to assess the accuracy and robustness of the solver. The static and the oscillating bubble are the prototypical configurations and present the behaviour of two phase flows.

4.1 Static bubble

A static bubble in a two-dimensional incompressible two phase flow is considered [18, 7, 16, 9]. This is a simple example for demonstrating the prototypical behaviour of multiphase flows. For simplicity, we consider a stationary bubble at equilibrium. Since the bubble is at rest inside the domain so it should show a zero velocity field but unfortunately spurious velocity/currents are observed near the interface [18, 7, 16, 9]. Moreover, flows involving interfaces lead to large pressure jumps. Approximation of the incompressibility constraints, the interface and the local external force are three different responsible sources for these phenomena [7]. The flow dynamics depends strongly on the magnitude of the spurious velocities. A non-physical movement of the interface might also be observed due to this spurious velocity.

4.1.1 Geometrical configuration

Both fluids are immiscible and separated by an interface Γ. The first fluid Ω_1 is completely inside of the second fluid Ω_2 as shown in Fig. 2 $(left)$. A circular static bubble of radius $r = 0.25$ is placed at the center $[0.5, 0.5]$ of a unit square $\overline{\Omega} = [0, 1]^2$. The surface tension coefficient, viscosities and densities are set to unity in the absence of the gravitational force. The relation between the pressure (inside and outside of the static bubble) should satisfies Laplace Young law:

$$p_i = p_o + \frac{\sigma}{r}. \tag{7}$$

Here, p_i is the pressure inside the bubble, p_o is the pressure outside the bubble and σ is the surface tension coefficient. The numerical studies for the systems of equations (4) and (6) are performed with a fixed time step $\Delta t = 10^{-2}$ until $t = 10$. The interface thickness is controlled by the parameter ϵ_ψ in both formulations. The spurious velocity/currents are observed and visually represented in Fig. 3 (a,b) and Fig. 4 (a,b). It can be seen in Fig. 3 (c,d) and Fig. 4 (c,d) that the pressure difference inside and outside of the bubble converges to the magnitude 4. Hence, the pressure jump across the interface successfully satisfies the Laplace Young law. By decreasing the value of interface thickness parameter ϵ_ψ, the surface tension force becomes sharp, resulting in a sharp pressure jump. The graphical representation of the pressure can be seen by a cross section through $y = 0$ in Fig. 3 (c,d) and Fig. 4 (c,d). Moreover, the cut-off function exhibits the same behaviour w.r.t. the interface thickness parameter ϵ_ψ in Fig. 3 (e,f) and 4 (e,f).

4.2 Oscillating bubble

A two-dimensional unsteady incompressible two phase flow is considered. These fluids are immiscible and separated by an interface Γ. The first fluid Ω_1 is completely inside the second fluid Ω_2 as shown in Fig. 2 $(right)$.

4.2.1 Geometrical configuration

The circle is initially perturbed to an elliptical shape with the semi-axis 0.25 in the x-direction and 0.125 in the y-direction. The ellipse is placed at the center [0.5, 0.5] of a unit

square $\overline{\Omega} = [0,1]^2$. The surface tension coefficient, viscosities and densities are set to unity. The gravitational force is neglected.

The numerical studies for the systems of equations (4) and (6) are performed with a fixed time step $\Delta t = 10^{-2}$ until $t = 100$. The transition of the radius (r_x, r_y) to a steady state with respect to time, is presented in Fig. 5. This study is performed for three different mesh refinement levels, i.e. $h = 1/16, 1/32$ and $1/64$. As the mesh is refined, the smooth transition from oscillating to steady state is illustrated in Fig. 5. It is observed that after certain oscillations the bubble reached the steady state, with no mass loss. It can be observed that the numerical instability arising from the interface capturing in the level set method vanishes in formulation (6), so the oscillating bubble reaches the steady state with much less oscillations. To analyse the temporal development of the interface, the bubble shapes are extracted at different time intervals, for mesh refinement level 6 ($h = 1/64$) in Fig. 6. At the final time, the ellipse is expected to reach an equilibrium state, that is a stable circular shape. As expected, the oscillating bubble is transformed into a stable circular shape.

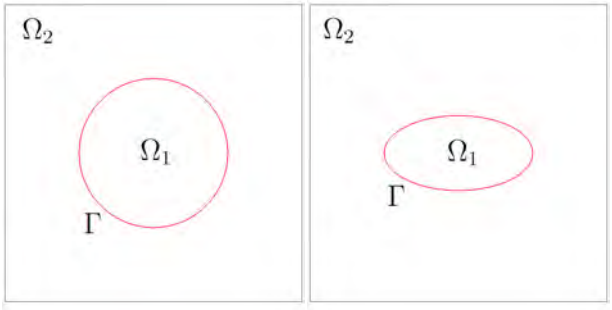
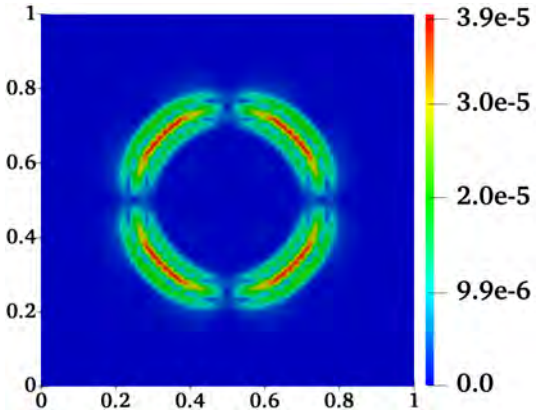
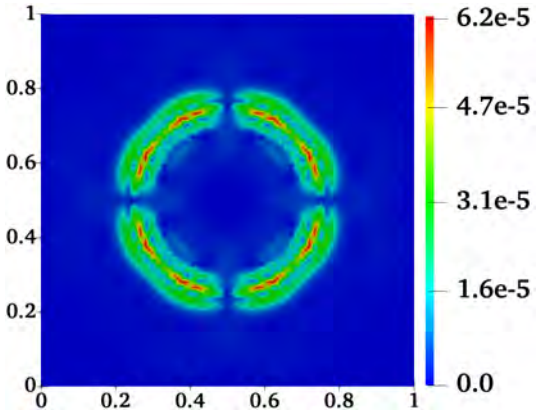


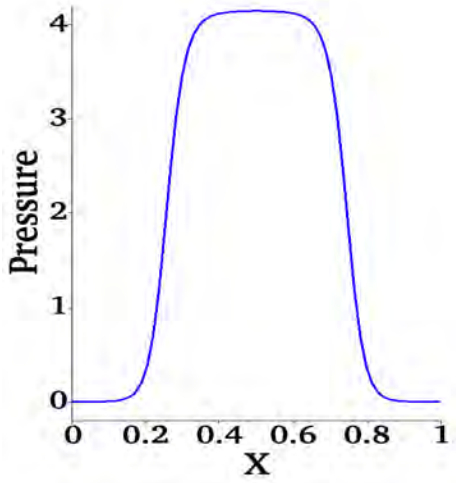
Figure 2: Computational domain of static bubble ($left$) and oscillating bubble ($right$).



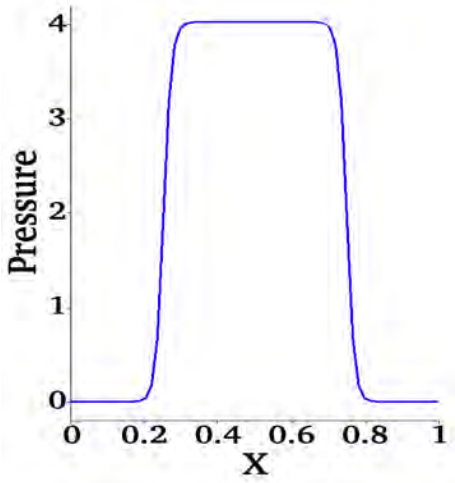
(a) Spurious velocity



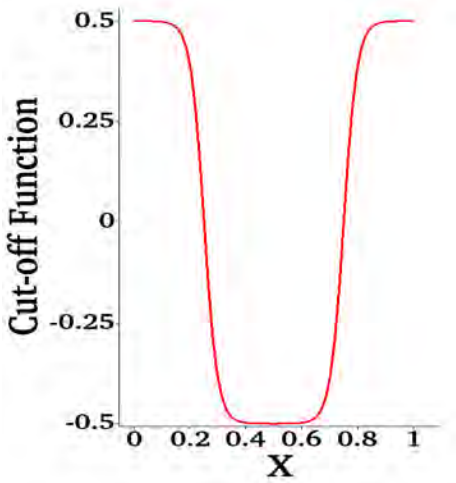
(b) Spurious velocity



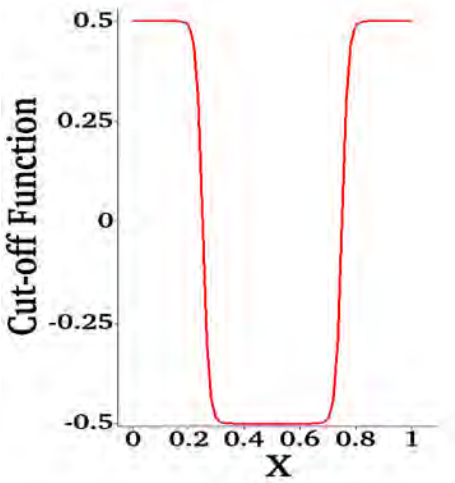
(c) Cross section through $y = 0$



(d) Cross section through $y = 0$

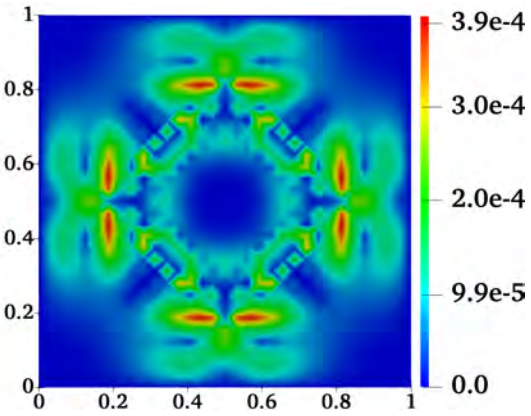


(e) Cross section through $y = 0$

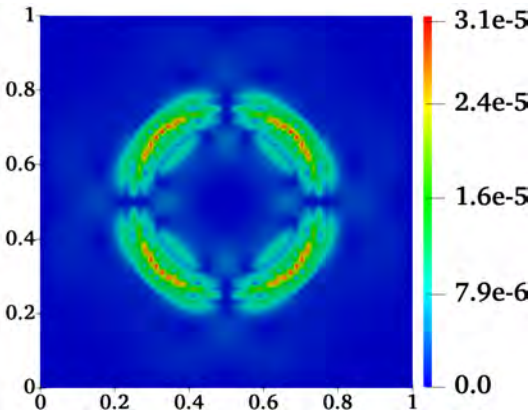


(f) Cross section through $y = 0$

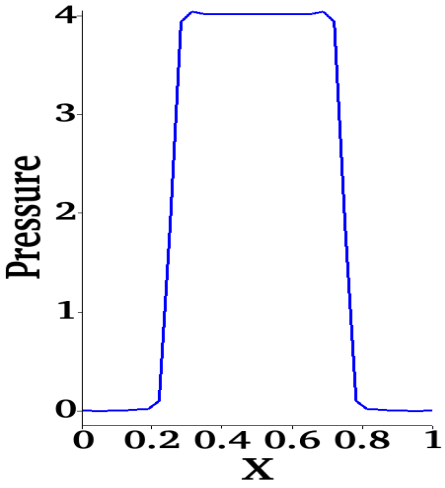
Figure 3: **Results for the system of equations (4):** The magnitude of the spurious currents (a, b), cross section of the pressure (c, d) through $y = 0$ and the cross section of the cut-off material function (e, f) through $y = 0$ for two different interface thickness $\epsilon_\psi = 1.5h$ $(Left)$ and $0.7h$ $(Right)$, with $h = 1/64$ at time $t = 10$, $\Delta t = 10^{-2}$.



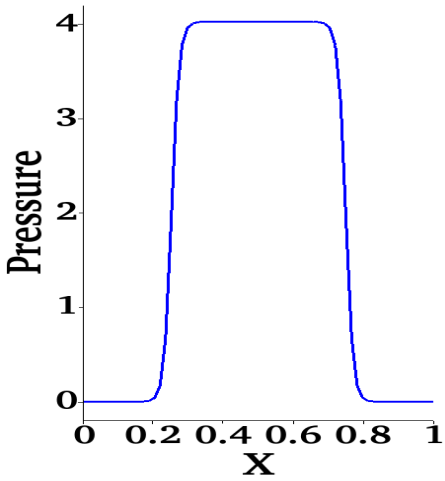
(a) Spurious velocity



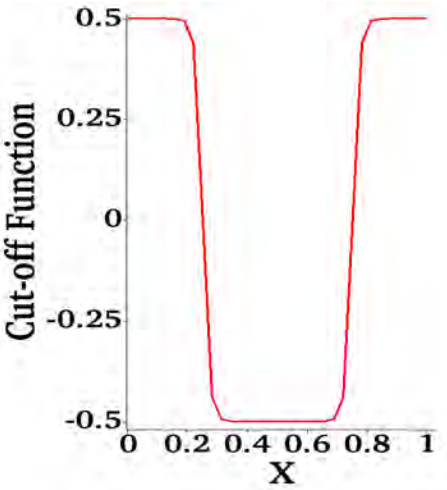
(b) Spurious velocity



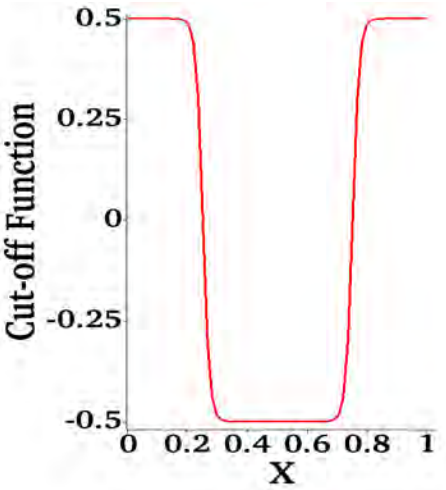
(c) Cross section through $y = 0$



(d) Cross section through $y = 0$

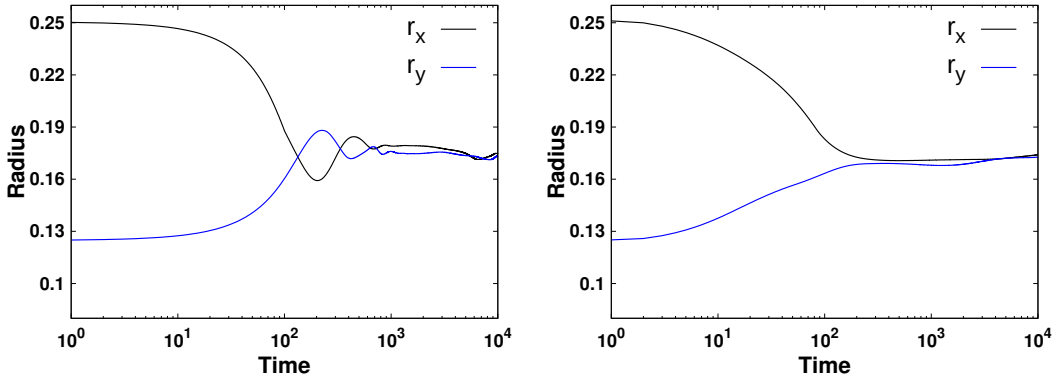


(e) Cross section through $y = 0$

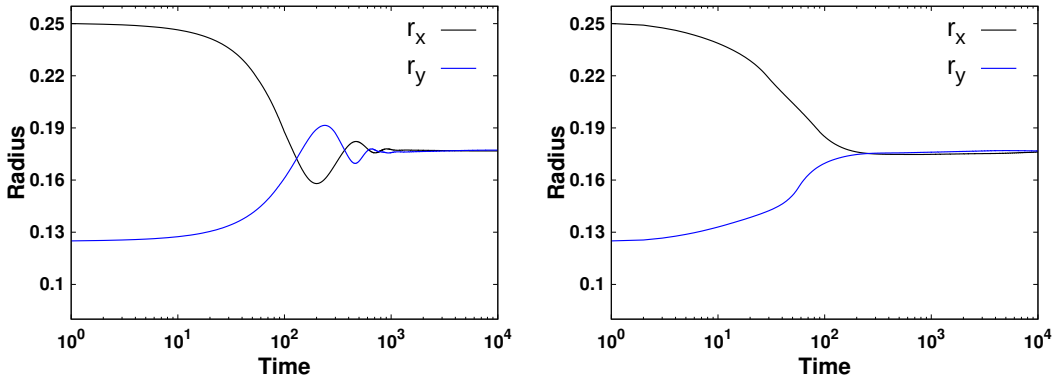


(f) Cross section through $y = 0$

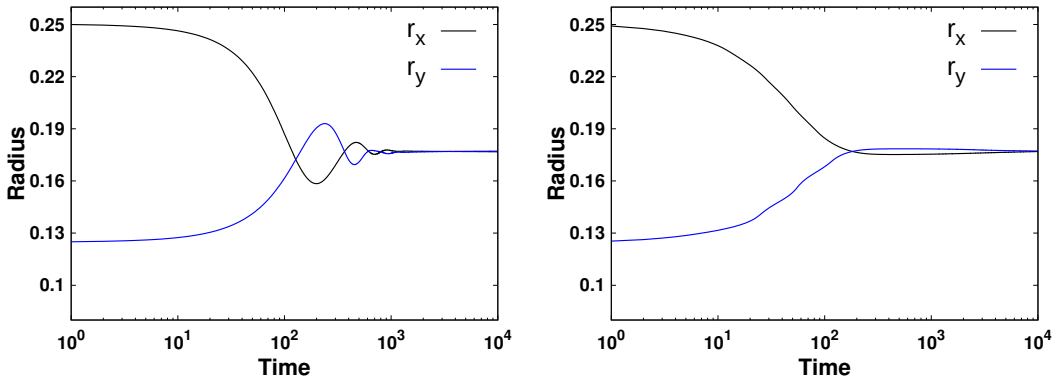
Figure 4: **Results for the system of equations (6):** The magnitude of the spurious currents (a, b), cross section of the pressure (c, d) through $y = 0$ and the cross section of the cut-off material function (e, f) through $y = 0$ for two different interface thickness $\epsilon_\psi = 1.5h$ $(Left)$ and $0.7h$ $(Right)$, with $h = 1/64$ at time $t = 10$, $\Delta t = 10^{-2}$.



(a) Refinement level 4 (h=1/16), **log scale** (b) Refinement level 4 (h=1/16), **log scale**



(c) Refinement level 5 (h=1/32), **log scale** (d) Refinement level 5 (h=1/32), **log scale**



(e) Refinement level 6 (h=1/64), **log scale** (f) Refinement level 6 (h=1/64), **log scale**

Figure 5: **Results for the system of equations (4**$(left)$**, 6**$(right)$**):** The radius (r_x, r_y) of oscillating bubble using Crank Nicolson time stepping scheme for three different mesh refinement levels, $\Delta t = 10^{-2}$.

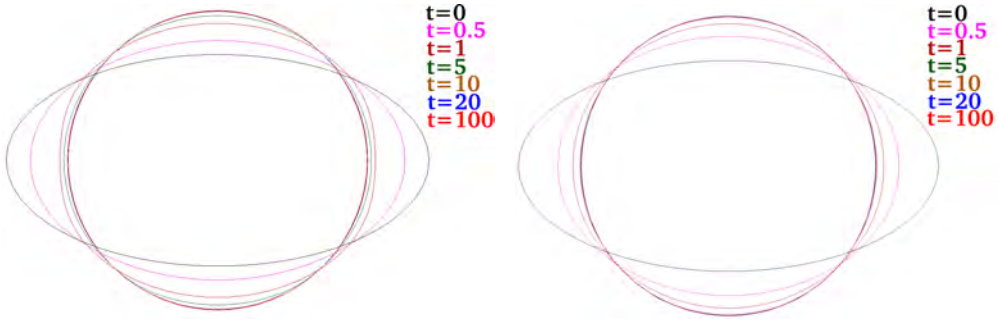


Figure 6: **Results for the system of equations (4**$(left)$**, 6**$(right)$**):** Oscillating bubble shape with respect to time t, $\Delta t = 10^{-2}$.

5 CONCLUSIONS

In this work, we have developed a monolithic Newton-multigrid solver for multiphase flow problems which solves velocity, pressure and interface position simultaneously. There are three main advantages of these formulations. Firstly, no explicit computation of the curvature and normals are required. Secondly, the explicit redistancing is removed by integrating the reinitialization term within the formulations. Thirdly, there is no capillary time restriction. In order to investigate the accuracy and robustness of the solver, numerical studies are performed for two-dimensional static and oscillating bubble. The results expectedly confirms the accuracy of the solution approximation as the magnitude of the pressure across the interface satisfy the Laplace Young's law. In both test cases, the bubble reached its steady state satisfying the theoretical predictions.

ACKNOWLEDGEMENTS

Muhammad Aaqib Afaq would like to thank Erasmus Mundus INTACT project, funded by the European Union as part of the Erasmus Mundus programme and the National University of Sciences and Technology (NUST) for their financial support. The authors also acknowledge the support by LS3 and LiDO3 team at ITMC, TU Dortmund University.

REFERENCES

[1] Anderson, D. M., McFadden, G. B., and Wheeler, A. A. Diffuse-interface methods in fluid mechanics. *Annual Review of Fluid Mechanics 30*, 1 (1998), 139–165.

[2] Badalassi, V. E., Ceniceros, H. D., and Banerjee, S. Computation of multiphase systems with phase field models. *Journal of Computational Physics 190*, 2 (2003), 371–397.

[3] Boffi, D., and Gastaldi, L. On the quadrilateral q2p1 element for the stokes problem. *International Journal for Numerical Methods in Fluids 39*, 11 (2002), 1001–1011.

[4] Boger, M. *Numerical Modeling of Compressible Two-Phase Flows with a Pressure-Based Method*. PhD thesis, University of Stuttgart, Institute of Aerodynamics and Gas Dynamics, Dec. 2013.

[5] Brackbill, J. U., Kothe, D. B., and Zemach, C. A continuum method for modeling surface tension. *Journal of Computational Physics 100*, 2 (1992), 335–354.

[6] Damanik, H., Ouazzi, A., and Turek, S. Numerical simulation of polymer film stretching. Tech. rep., Fakultät für Mathematik, TU Dortmund, 2013. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 485.

[7] Ganesan, S., Matthies, G., and Tobiska, L. On spurious velocities in incompressible flow problems with interfaces. *Computer Methods in Applied Mechanics and Engineering 196*, 7 (2007), 1193–1202.

[8] Hirt, C. W., and Nichols, B. D. Volume of fluid (VOF) method for the dynamics of free boundaries. *Journal of Computational Physics 39*, 1 (1981), 201–225.

[9] Hosseini, B. S., Turek, S., Mller, M., and Palmes, C. Isogeometric analysis of the Navier-Stokes-Cahn-Hilliard equations with application to incompressible two-phase flows. *Journal of Computational Physics 348* (2017), 171–194.

[10] Lafaurie, B., Nardone, C., Scardovelli, R., Zaleski, S., and Zanetti, G. Modelling merging and fragmentation in multiphase flows with SURFER. *Journal of Computational Physics 133*, 1 (1994), 134–147.

[11] Li, C., Gui, C., and Fox, M. D. Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on Image Processing 19*, 12 (2010), 3243–3254.

[12] Olsson, E., and Kreiss, G. A conservative level set method for two phase flow. *Journal of Computational Physics 210*, 1 (2005), 225–246.

[13] Osher, S., and Sethian, J. A. Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations. *Journal of Computational Physics 79*, 1 (1988), 12–49.

[14] Ouazzi, A., Turek, S., and Damanik, H. A curvature–free multiphase flow solver via surface stress–based formulation. *International Journal for Numerical Methods in Fluids 88*, 1 (2018), 18–31.

[15] Sethian, J. A. *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science.* Cambridge monographs on applied and computational mathematics. Cambridge University Press, 1999.

[16] Shirani, E., Ashgriz, N., and Mostaghimi, J. Interface pressure calculation based on conservation of momentum for front capturing methods. *Journal of Computational Physics 203*, 1 (2005), 154–175.

[17] Sussman, M., and Fatemi, E. An efficient, interface-preserving level set redistancing algorithm and its application to interfacial incompressible fluid flow. *SIAM Journal of Scientific Computing 20*, 4 (1999), 11651191.

[18] Turek, S., Ouazzi, A., and Hron, J. On pressure separation algorithms (PSEPA) for improving the accuracy of incompressible flow simulations. *International Journal for Numerical Methods in Fluids 59*, 4 (2008), 387–403.

Solution of heat transfer inverse problem in thin film irradiated by laser

A. Korczak* **and W. Mucha†**

Department of Computational Mechanics and Engineering
Silesian University of Technology
Gliwice, Poland
e-mail: *anna.korczak@polsl.pl; †waldemar.mucha@polsl.pl

Key words: heat transfer, Boltzmann transport equation, identification, evolutionary algorithm

Abstract: *The presented paper deals with an inverse problem in nanoscale heat transfer simulation. A thin metal film irradiated by the ultrashort laser pulse is modeled using the Boltzmann transport equation. Heat transfer parameters of the model arheat transfer, Boltzmann transport equation, identification, evolutionary algorithme identified using evolutionary algorithm an optimization algorithm inspired on biological evolution of species, where the difference between obtained and expected results is minimized.*

1 INTRODUCTION

In the presented research, identification of short-pulse laser parameters was carried out. In the discussed example thin metal film was influenced by a laser beam. The process was modelled numerically. In the identification, experimental data was used, in order to minimize the error between numerical and experimental results.

Heat flow in solids can be modelled using various models. When dealing with objects of small dimensions, of the order of nanometres, and with fast heating processes, comparable to relaxation times, then it is reasonable to use molecular dynamics or the Boltzmann transport equation (BTE). The presented coupled system of Boltzmann transport equations has the advantage over molecular dynamics that it has a less complicated mathematical apparatus and calculations proceed faster. This is an important advantage considering inverse problems, where computations are performed repeatedly for different possible combinations of identified parameters.

The goal of the identification presented in this paper is to obtain three parameters of the laser irradiation, such as the laser intensity, the optical penetration depth, the reflectivity and the laser pulse duration. The base of result evaluation is the outcome of the experiment described in [3] where experimental data are shown for electron temperatures in chosen node in a function of time. Proposed identification finds parameters of a numerical model that would recreate the real process flow as exactly as possible.

2 THE BOLTZMANN TRANSPORT EQUATION

In the presented problem as the governing equation is used the Boltzmann transport equation (BTE). According to the Debye simplifications the equivalent transformed form of energy density equation is analysed. This paper considers the one-dimensional heat transfer model in metals. As it is a coupled problem, then both types of energy carriers must be taken into account. The coupled system of equations can be written using the differential equation (subscript: e-electrons and ph-phonons) [4]

$$\frac{\partial e_e}{\partial t} + \mathbf{c}_e \frac{\partial e_e}{\partial x} = -\frac{e_e - e_e^0}{\tau_e} + Q_e \tag{1}$$

https://doi.org/10.4995/YIC2021.2021.12580

$$\frac{\partial e_{ph}}{\partial t} + \mathbf{c}_{ph}\frac{\partial e_{ph}}{\partial x} = -\frac{e_{ph} - e_{ph}^0}{\tau_{ph}} + Q_{ph} \tag{2}$$

where $e_e = e_e(t,x)$, $e_{ph} = e_{ph}(t,x)$ are the energy densities, $e_e^0 = \frac{e_e}{2}$, $e_{ph}^0 = \frac{e_{ph}}{2}$ are the equilibrium energies densities, $\mathbf{c}_e$, $\mathbf{c}_{ph}$ are the propagation speed, τ_e, τ_{ph} are the relaxation times, t is the time and $Q_e = Q_e(t,x)$, $Q_{ph} = Q_{ph}(t,x)$ are the energies sources for electrons and lattice respectively.

Since the system of governing equations is formulated for energy densities there is a need for formulas that allow to recalculate energy density to temperature and vice versa. Such conversion can be made using presented formulas [2]

$$e_e(T_e) = \left(n_e \frac{\pi^2}{2}\frac{k_b^2}{\varepsilon_F}\right) T_e^2 \tag{3}$$

$$e_{ph}(T_{ph}) = \left(\frac{9\eta_{ph}k_b}{\Theta_D^3}\int_0^{\frac{\Theta_D}{T_{ph}}} \frac{x^3}{exp(x)-1}dx\right) T_{ph}^4 \tag{4}$$

where k_b is the Boltzmann constant, Θ_D is the Debye temperature of the metal, T_e, T_{ph} are the temperatures for electrons and phonons respectively, while n_e and η_{ph} are densities of these carriers. The electrons and the phonons energy sources depend on temperature of both carriers, the electron-phonon coupling factor which vary for deferent materials and can be calculated using the following expressions [1, 5]

$$Q_e(t,x) = Q(t,x) - G(T_e(t,x) - T_{ph}(t,x)) \tag{5}$$

$$Q_{ph}(t,x) = G(T_e(t,x) - T_{ph}(t,x)) \tag{6}$$

The electron-phonon coupling factor is a coefficient which characterizes the energy exchange between both carriers. To make model complete the equations (1) and (2), should be supplemented by the boundary-initial conditions. In the paper are considered the 2nd type of the boundary conditions (BC) on both edges, particularly adiabatic condition, because the laser heating lasts for a short period and then the heat losses from the both surfaces of the thin film can be neglected. To solve direct problem based on presented system of equations (1) and (2) the lattice Boltzmann method (LBM) was applied. For D1Q2 model in the LBM the discrete set of two propagation directions with appropriate velocities for electrons and phonons (Figure 1) are defined.

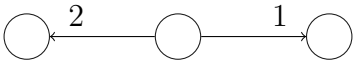


Figure 1: Directions of propagation energy carriers

Moreover, in the mathematical model the internal heat source $Q(t,x)$ is applied. It takes into account the temporal variation of the laser pulse approximated by a form of exponential function

$$Q(t,x) = \sqrt{\frac{\beta}{\pi}}\frac{1-R}{t_p\delta_s} I_0 e^{\frac{-x}{\delta_s} - \beta\frac{t-2t_p}{t_p}} \tag{7}$$

where I_0 is the laser intensity, R the reflectivity, t_p the laser pulse duration defined as full width at half maximum of the laser pulse, δ_s the optical penetration depth, x the depth measured from the front surface, $\beta = 4\ln(2)$.

3 INVERSE PROBLEM SOLUTION

The considered inverse problem consists of identification of four model parameters, describing the laser irradiation. The performed identification is defined as an optimization problem where the goal is to minimize the differences between the results obtained from model with given parameters, and the expected values. The problem is solved using an evolutionary algorithm.

3.1 Evolutionary algorithm

Evolutionary algorithm (Figure 2) is a metaheuristic optimization algorithm that is inspired by the mechanisms of biological evolution of species. It operates on a set (population) of potential solutions (individuals) to a given problem. The quality of each individual is evaluated by the minimized goal function, that determines the adaptation of the individual to the environment. The higher the adaptation is, the bigger are the chances of the individual to survive. Genetic operators such as crossover (mixing genes from more than one individual) or mutation (random changes in genes) are applied in order to create populations for subsequent generations [6][8].

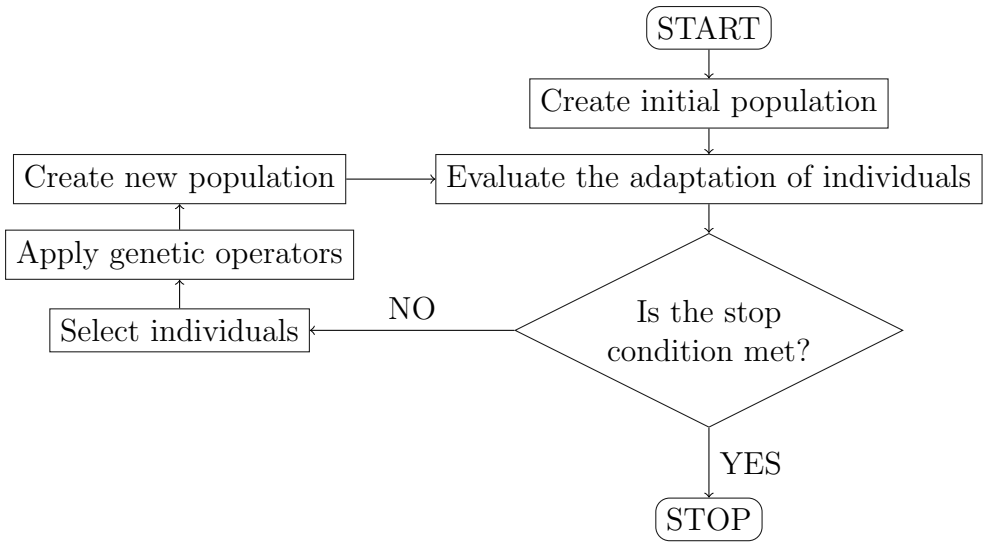


Figure 2: Evolutionary algorithm

3.2 Identification problem parameters

The goal is to adjust the numerical model, described using BTE, to fit experimental results published in [1], as accurate as possible. The considered results are electron temperatures in a chosen node in the function of time. The identified model parameters are: (a) laser intensity I_0, (b) optical penetration depth δ_s, (c) reflectivity R, and (d) laser pulse duration t_p. The boundaries of the parameters' values assumed in the identification are presented in Table 1.

Table 1: Boundaries of the identified parameters

Parameter	Lower bound	Upper bound	Unit
I_0	5	100	J/m^2
δ_s	10^{-11}	50×10^{-9}	m
R	0.01	1	-
t_p	10^{-14}	10^{-12}	ps

The goal (fitness) function F in the identification (optimization) problem was formulated as follows:

$$F(I_0, \delta_s, R, t_p) = \sum_{i=1}^{n} (T_i^{exp} - T_i^{num})^2 \tag{8}$$

where T_i^{exp} and T_i^{num} are the experimentally measured and numerically computed nodal electron temperatures at time sample i.

The parameters of the evolutionary algorithm were adapted as follows: population size 50, scattered crossover with probability of 0.8, and Gaussian mutation. The stop conditions were maximum number of generations 400 and 50 stall generations.

3.3 Obtained results

The convergence of the algorithm can be observed in Figure 3 as the mean fitness function value F of the population, over subsequent generations. The algorithm converged in 218 generation, after 50 stall generations. The fitness function value F for the best individual was 6.9964×10^4, while the mean value for whole final population was 7.0549×10^4.

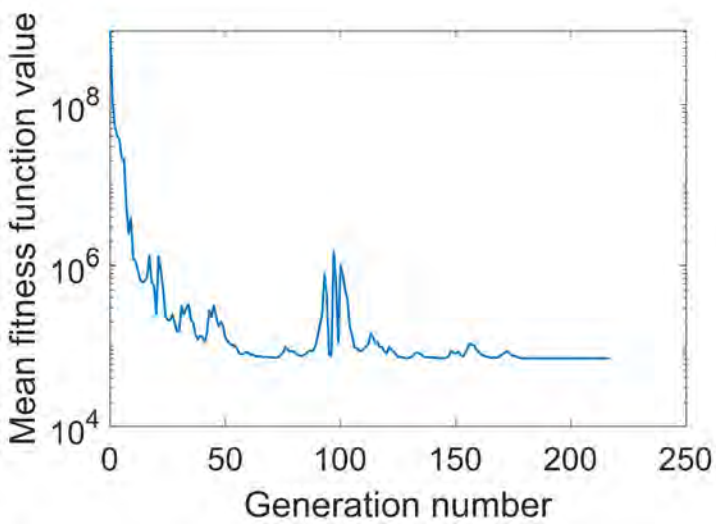


Figure 3: Convergence of the evolutionary algorithm

The identified parameters of the model are: $I_0 = 18.8843 J/m^2$, $\delta_s = 18.386 \times 10^{-9} m$, $R = 0.9619$ and $t_p = 0.0259 \times 10^{-12} s$. The time plot of the electron temperature of the source experimental data and that obtained from the identified numerical model are compared in Figure 4.

4 CONCLUSIONS

In the presented identification problem, the values of four model parameters (laser intensity I_0, optical penetration depth δ_s, reflectivity R, and laser pulse duration t_p) were searched. These parameter values introduced to the numerical model based on BTE were supposed to give electron temperature distribution that fit experimental results. Evolutionary algorithm was implemented to the identification problem. As indicated by Figure 3, the convergence process was successful, minimum value of fitness function was reached in 218 generations, after 50 stall generations. The accuracy of the BTE model with identified parameters values can be considered as satisfying, as can be observed in Figure 4 where comparison with experimental data is presented.

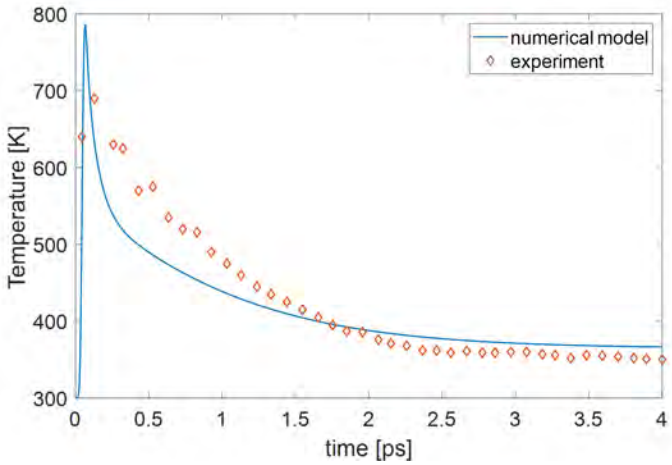


Figure 4: Comparison of experimental and numerical results

ACKNOWLEDGEMENTS

The research is funded from the projects Silesian University of Technology, Faculty of Mechanical Engineering, 2021.

REFERENCES

[1] Escobar, R. A., Ghai, S. S., Jhon, M. S., and Amon, C. H. Multi-length and time scale thermal transport using the lattice boltzmann method with application to electronics cooling. *International Journal of Heat and Mass Transfer 49*, 1 (2006), 97–107.

[2] Eshraghi, M., and Felicelli, S. D. An implicit lattice boltzmann model for heat conduction with phase change. *International Journal of Heat and Mass Transfer 55*, 9 (2012), 2420–2428.

[3] J. K. Chen, J. E. B. Numerical study of ultrashort laser pulse interactions with metal films. *Numerical Heat Transfer, Part A: Applications 40*, 1 (2001), 1–20.

[4] Joshi, A. A., and Majumdar, A. Transient ballistic and diffusive phonon heat transport in thin films. *Journal of Applied Physics 74*, 1 (1993), 31–39.

[5] Piasecka Belkhayat, A., and Korczak, A. Modelling of transient heat transport in metal films using the interval lattice boltzmann method. *Bulletin of the Polish Academy of Sciences: Technical Sciences 64*, No 3 (2016), 599–606.

RECENT ADVANCES IN SPACE-TIME METHODS

A Space-Time FE Level-set method for convection coupled phase-change processes

L. Boledi[1,*], B. Terschanski[1], S. Elgeti[2,3] and J. Kowalski[1,4]

[1] Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, 52056 Aachen, Germany
email: boledi@aices.rwth-aachen.de, benjamin.terschanski@rwth-aachen.de

[2] Institute of Lightweight Design and Structural Biomechanics (ILSB), TU Wien, 1040 Vienna, Austria
e-mail: elgeti@ilsb.tuwien.ac.at

[3] Chair for Computational Analysis of Technical Systems (CATS), RWTH Aachen University, 52056 Aachen, Germany

[4] Abteilung Computational Geoscience, Georg-August-Universität Göttingen, 37077 Göttingen, Germany
e-mail: julia.kowalski@geo.uni-goettingen.de

Key words: Ghost Cells, Phase Change, Stefan Problem, Space-Time Finite Elements

Abstract: *Phase-transition processes have great relevance for both engineering and scientific applications. In production engineering, for instance, metal welding and alloy solidification are topics of ongoing research, whereas understanding the melting of ice and permafrost is at the centre of many geoscience research questions. In this contribution we focus on one specific phase-change process, namely the convection-coupled solid-liquid phase change of a single species, e.g. water. The material is assumed to be incompressible within the two phases, but we account for density changes across the phase interface. To describe the process, we need to solve the incompressible Navier-Stokes equations and the heat equation for both phases over time. The position of the phase interface is tracked with a level-set method [1]. The level-set function is advected according to the phase interface's propagation speed. Such speed depends on local energy balance across the interface and it is determined through a heat-flux jump condition referred to as the Stefan condition [2]. One of the challenges of this method lies in the approximation of the heat-flux discontinuity at the interface based on the evolving temperature and velocity fields.*

To model the temperature and velocity fields within each phase, we employ the space-time finite element method. However, commonly used interpolation functions, such as piecewise-linear functions, fail to capture discontinuous derivatives over one element that are needed to assess the level-set's transport term. Available solutions to this matter, such as local enrichment with extended finite elements [3], are often not compatible with existing space-time finite element codes and require extensive implementation work. Instead, we consider a different method and we decide to extend the ghost-cell technique to finite element meshes [4]. The idea is that we can separate the two subdomains associated with each phase and solve two independent temperature problems. We prescribe the melting temperature at an additional node close to the interface and we retrieve the required heat flux on each side of the interface. This allows us to locally evaluate the heat-flux jump.

In this work we describe the ghost-cell method applied to our space-time finite element solver [5]. Then, we demonstrate test cases in 3D in view of future applications.

https://doi.org/10.4995/YIC2021.2021.12329

1 INTRODUCTION

In this paper we propose a numerical strategy to simulate convection-coupled solidification and melting processes. Many approaches to track the phase-change interface (PCI) are available [6], but we focus on the level-set method [1]. In our formulation we need to retrieve discontinuities in the first derivatives at the PCI. Extended finite elements solve the problem by locally enriching the basis functions [3], but this changes the number of degrees of freedom over time. Instead, we decide to build upon existing work from Gibou et Al. and extend the ghost-cell approach to our space-time finite element framework [4, 7, 8].

2 NUMERICAL MODELLING

We consider a domain of interest $\Omega \subset \mathbb{R}^d$, where d is the number of space dimensions, that consists of a solid region and a liquid region. The two phases are separated by a distinct PCI. The goal of the model is to determine the evolving velocity, pressure and temperature fields in both phases and over time.

2.1 Governing equations for flow and temperature

Let $t \in (0,T)$ be a time instant. We call $\Omega_1(t)$, $\Omega_2(t)$ the two time-dependent subdomains associated with the liquid region and the solid region, respectively, such that $\Omega_1(t) \cup \Omega_2(t) = \Omega$ for each t. We describe the flow problem with the incompressible Navier-Stokes equations for a Newtonian fluid

$$\rho_* \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} - \mathbf{f} \right) + \nabla p - \mu_* \Delta \mathbf{u} = 0 \quad \text{in } \Omega \times (0,T), \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \times (0,T), \tag{2}$$

where ρ_* and μ_* denote the density and the dynamic viscosity. To model the temperature field we consider the transient heat equation

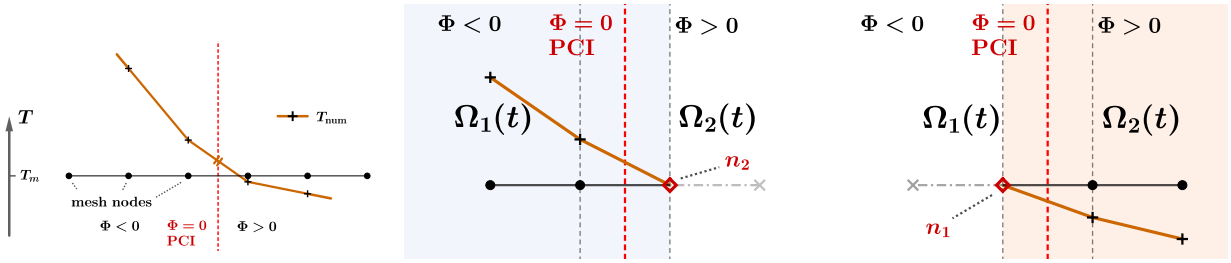
$$\rho_* (c_p)_* \left(\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T \right) = \kappa_* \Delta T \quad \text{in } \Omega \times (0,T), \tag{3}$$

where $(c_p)_*$ is the heat capacity and κ_* is the thermal conductivity. The subscript $*$ in Eqs. (1)-(3) indicates the phase-dependent material properties, such that $\rho_*(\mathbf{x},t) = \rho_1$ if $\mathbf{x} \in \Omega_1(t)$ and $\rho_*(\mathbf{x},t) = \rho_2$ if $\mathbf{x} \in \Omega_2(t)$. Note that such properties are phase-wise constant. The advective term $\mathbf{u}$ in Eq. (3) gives rise to a one-way coupling with the Navier-Stokes Equations (1), (2).

We solve both problems with our in-house space-time finite element solver [5, 9]. The stabilised space-time formulation can be found in [10], together with the values for the stabilisation terms.

2.2 Tracking the phase-change interface: Level-set method

We now introduce the level-set formulation to track the evolving PCI. Let $\Phi : \Omega \times (0,T) \to \mathbb{R}$ be a scalar, continuous function such that $\Phi(\mathbf{x},t) < 0$ in $\Omega_1(t)$ and $\Phi(\mathbf{x},t) > 0$ in $\Omega_2(t)$. The phase interface is the zero level set of Φ, that is every $\mathbf{x} \in \Omega : \Phi(\mathbf{x},t) = 0, \forall t \in (0,T)$ [11]. Thus, Φ gives information on which subdomain a point $\mathbf{x}$ is located. We then describe the material properties as function of Φ, e.g. $\rho_* = \rho_1 + (\rho_2 - \rho_1) H_\epsilon(\Phi)$. The function $H_\epsilon(\cdot)$ is the smoothed Heaviside function introduced in [12], which alleviates numerical difficulties.



(a) Complete domain (b) Left phase (c) Right phase

Figure 1: 1D example of ghost split. Fig. 1a shows a fictitious temperature field computed on the whole domain. Note that the solution is differentiable across one element and we do not retrieve the heat-flux jump at the PCI. Figs. 1b, 1c show the independent temperature problems solved for each phase after the ghost split. The melting temperature T_m is imposed at the ghost nodes n_1 and n_2.

The evolution of the PCI is described by the level-set equations

$$\begin{aligned} \frac{\partial \Phi}{\partial t} + \mathbf{v} \cdot \nabla \Phi = 0 \quad &\text{in } \Omega \times (0,T), \\ \Phi(\mathbf{x},0) = \Phi_0(\mathbf{x}) \quad &\text{in } \Omega, \end{aligned} \tag{4}$$

where $\mathbf{v}$ denotes the local propagation velocity of the interface. We select the initial condition $\Phi_0(\mathbf{x})$ such that $\Phi(\mathbf{x},t)$ is the signed distance function from the PCI. Problem (4) is a scalar advection problem that shares many similarities with Eq. (3). More details on its space-time formulation are available in Section 3.10 of [13].

Note that the transport term $\mathbf{v}$ in Eq. (4) is not known, so that we need an additional relation to close the problem. Localized at the zero set of the level-set function, the propagation velocity $\mathbf{v}(\mathbf{x},t)$ needs to match the local phase-change rate and can be modelled as the Stefan condition [2]. Thus, $\mathbf{v}$ is proportional to the heat-flux jump at the interface

$$\rho\, h_m \mathbf{v}(\mathbf{x},t) = -\kappa_L \nabla T\big|_{\mathbf{X}^-} + \kappa_S \nabla T\big|_{\mathbf{X}^+} = [\kappa \nabla T]_L^S = q_L - q_S, \quad \forall \mathbf{x} : \Phi(\mathbf{x},t) = 0, \tag{5}$$

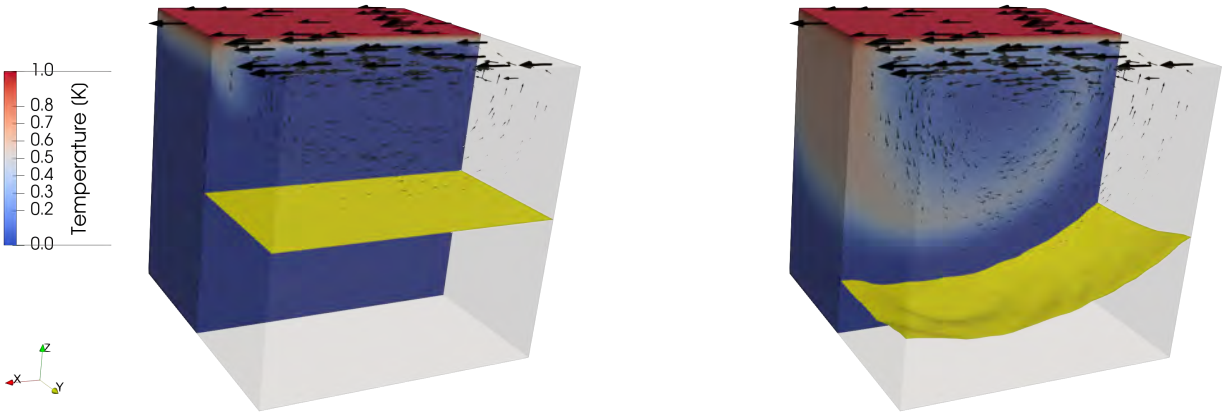
where h_m is the latent heat of melting, ρ denotes the material's density, κ the material's conductivity, $\mathbf{X}^{\pm}$ denote the limits taken from either side of the PCI and $[\cdot]_L^S$ refers to the liquid and solid regions. Note that we have closed the problem by coupling the level-set Eq. (4) with the temperature Eq. (3). However, we need to accurately retrieve the discontinuity of the heat flux within our numerical framework, which we will address in the next section.

3 HEAT-FLUX DISCONTINUITY AT THE PHASE INTERFACE

In the previous section we have described our numerical model for melting and solidification problems. What we need is a method to recover the heat-flux jump in Eq. (5) when using finite elements with element-wise continuously differentiable shape functions.

3.1 Evaluation points for the heat fluxes

First, we select evaluation points for the representative fluxes q_L, q_S in Eq. (5). The intuitive approach would be to choose points normal to the PCI, but this presents issues since the normal is not well-defined at the intersections with the mesh. Thus we adopt a different approach based on three assertions:



(a) $t = 1s$ (b) $t = 15s$

Figure 2: Phase-change coupled 3D lid-driven cavity. The temperature profile is shown at two different time instants. The yellow surface denotes the phase-change interface. The black arrows represent the velocity vectors at each point, their size is proportional to the velocity magnitude. The domain is transparent for $y > 0.5$.

1. If an element face is cut by the PCI, we consider the nodes that belong to the face as flux nodes. We use the numerical gradients at these nodes as representative fluxes in Eq. (5). The sign of the level-set function carries information on the corresponding phase q_L or q_S;

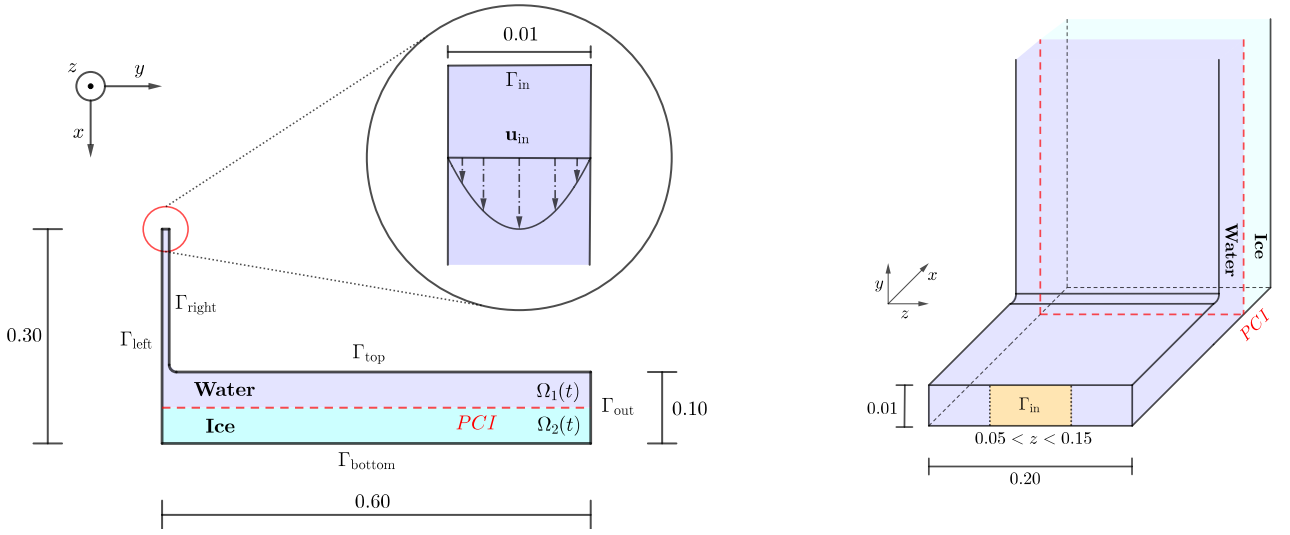
2. Each nodal gradient is computed as the arithmetic average of the gradients on all the elements that surround the node;

3. If the PCI intersects a mesh node, we consider the average of all the nodes in the adjacent faces.

3.2 The ghost-split method

The second issue comes from the computation of the numerical gradient, since its mathematical properties depend on the properties of the basis functions. In particular, we employ piecewise linear shape functions that can show discontinuities in the heat flux only at element nodes. This is where the ghost split comes into play. Since we know that the temperature solution at the PCI must equal the melting temperature T_m at each time instant, the PCI can be viewed as a Dirichlet type boundary for each phase. Then, we solve two independent temperature problems in each subdomain and retrieve the representative fluxes to compute the interface propagation velocity as in Eq. (5). However boundary conditions can be imposed only on mesh nodes, so we have to consider additional nodes for each subdomain to enforce the melting temperature at the approximate position of the PCI. These extra nodes are called ghost nodes. Figure 1 shows an example of ghost split for a 1D temperature profile. By solving two separate problems on each subdomain, we can retrieve the heat-flux jump at the PCI.

Now we can compute the interface velocity at the intersections with the mesh. As a last step we need to define the transport term $\mathbf{v}$ of Eq. (4) on all the mesh nodes. Given the Stefan velocity computed on a crossing, we extend such velocity to all the nodes that are closest to the crossing. Note that we do not need an additional search to find the nearest neighbours of



(a) Top view of the domain

(b) Close-up of the inflow

Figure 3: Phase-change coupled 3D corner flow. We show the computational domain of the second test case (not drawn to scale). On the left, a top view section is displayed. On the right, we show a close-up of the inflow to call attention to the boundary condition. Note that the parabolic velocity profile is imposed only on the part highlighted in orange.

the mesh nodes, as reinitialising the level-set function entails this information.

Recall that with the ghost split the melting temperature is assigned at an additional node close to the PCI. By doing so, we introduce an error in the interface location computed at the subsequent time step. Note, however, that the error depends on the mesh resolution and the position of the ghost node converges to the correct location of the PCI for finer grids [4]. Higher order schemes for the temperature extrapolation are available and can be investigated in the future [7].

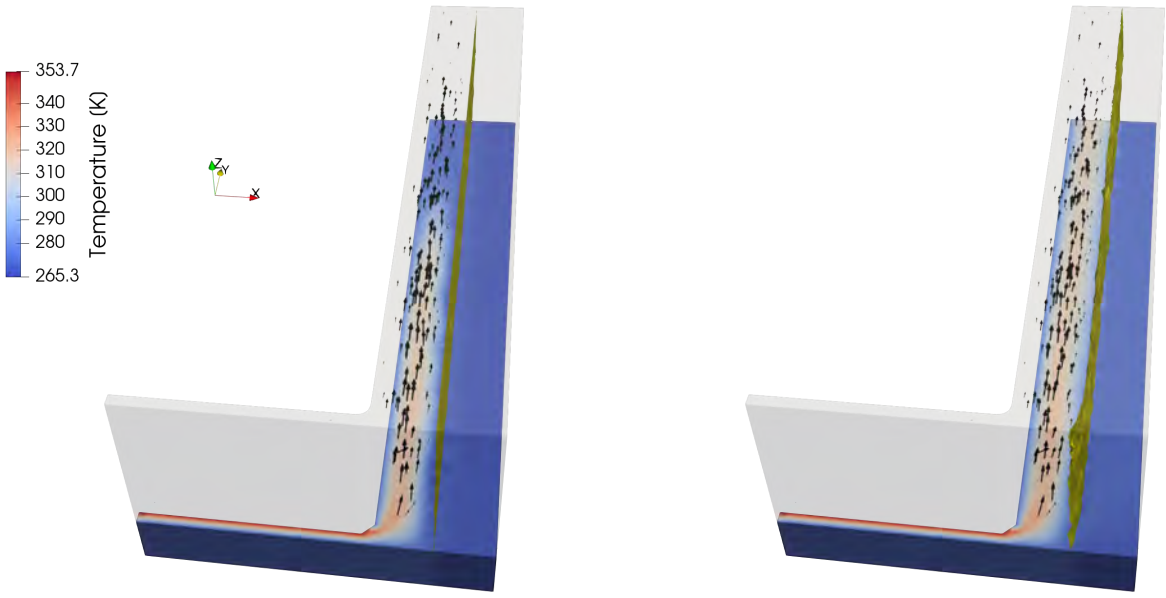
4 NUMERICAL EXAMPLES

In the last section we show two different numerical cases in 3D. A detailed verification of the numerical method against the one-phase Stefan problem can be found in [14], together with additional 2D examples. In this work we focus only on tridimensional problems in view of more complex applications.

Recall that the presented numerical method is not bound to the number of spatial dimensions. We have described the 1D ghost split in Fig. 1 for the sake of clarity, but a more detailed graphical description on a 2D mesh is available in [14]. Thus, the space-time finite element solver can handle 3D phase-change processes. Performance issues might arise with very fine meshes, for instance in the reinitialisation of the level-set function. However different strategies are available, e.g. the fast marching method [15], that can be investigated in the future.

4.1 Phase-change coupled 3D lid-driven cavity

We consider a $1 \times 1 \times 1$ domain, i.e. a unit cube, which is initially solid for $z < 0.5$ and liquid for $z > 0.5$. At the lateral and bottom boundaries we impose homogeneous Dirichlet boundary conditions for the velocity and homogeneous Neumann conditions for temperature. At the top edge we impose the temperature $T = 1$ and the velocity in x direction $\mathbf{u} = [1, 0, 0]^\intercal$. The initial temperature is $T_m = 0$. We select the parameters for the two phases ρ_1 =2, ρ_2 =1,



(a) $t = 50s$ (b) $t = 2500s$

Figure 4: Phase-change coupled 3D corner flow. The temperature profile is shown at two different time instants. The yellow surface denotes the phase-change interface. The black arrows represent the velocity vectors at each point, their size is proportional to the second component of velocity. The domain is transparent for $z > 0.05$.

$(c_p)_1$ =1e3, $(c_p)_2$ =1, κ_1 =1, κ_2 =1, μ_1=1, μ_2=1e4 and h_m =1, where all the values are in SI units. We simulate 500 time steps with $\Delta t = 0.1s$ on a uniform structured grid that comprises 35152 nodes.

Figure 2 shows the computed temperature profile at two time instants. After 10 time steps the PCI has not moved yet, but we retrieve the expected anti-clockwise circulation in the liquid region (2a). After 150 time steps the PCI has moved downwards (2b). Note that the left side of the domain melts faster, since the temperature propagation is driven by the convection of the flow field.

4.2 Phase-change coupled 3D corner flow

For the second example we consider a recent research topic, namely the flow that develops around a thermal melting cryorobot that descents into the ice [16]. Figure 3 shows the geometry of the test case, which resembles an idealised probe moving to the right. The inflow channel turns 90 degrees into a wider outflow channel. The latter contains two different phases that are separated by an evolving PCI. We impose a parabolic velocity profile at the inflow such that $\mathbf{u}_{\text{in}} = [5000y\,(0.01 - y), 0, 0]^\intercal$ if $0.05 < z < 0.15$. Furthermore, we impose no-slip conditions at each boundary except for the inflow and the outflow. We have Dirichlet temperature conditions on Γ_{right} and Γ_{top}, $T = 353\,\text{K}$ and $T = 278\,\text{K}$ respectively. On Γ_{left} we prescribe $T = 273\,\text{K}$ if $x < 0.25$, $T = 268\,\text{K}$ if $x > 0.25$. The initial conditions are $\mathbf{u}(\mathbf{x}, 0) = \mathbf{0}$, $T(\mathbf{x}, 0) = 273\,\text{K}$ in $\Omega_{1,0}$, $T(\mathbf{x}, 0) = 268\,\text{K}$ in $\Omega_{2,0}$. The material properties are selected according to water ice [17]. We simulate 500 time steps with $\Delta t = 5s$.

Figure 4 shows the computed temperature profile at two time instants. At the final time step a bulge is visible in the PCI. As expected, the ice melts as we introduce heat into the system and we can see the effect right after the 90 degree turn. We recall that this setup is not reproducible in 2D as heat and flow are applied only on a portion of the inflow, which

underlines the need for a numerical method that can represent 3D physical phenomena.

ACKNOWLEDGEMENTS

The authors were supported by the Helmholtz Graduate School for Data Science in Life, Earth and Energy (HDS-LEE). The work was furthermore supported by the Federal Ministry of Economic Affairs and Energy, on the basis of a decision by the German Bundestag (50 NA 1908). The authors gratefully acknowledge the computing time granted by the JARA Vergabegremium and provided on the JARA Partition part of the supercomputer JURECA at Forschungszentrum Jülich [18].

REFERENCES

[1] S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1):12 – 49, 1988.

[2] S. Chen, B. Merriman, S. Osher, and P. Smereka. A Simple Level Set Method for Solving Stefan Problems. *Journal of Computational Physics*, 135(1):8 – 29, 1997.

[3] M. Bernauer and R. Herzog. Motion Planning for the Two-Phase Stefan Problem in Level Set Formulation 1, 12 2010.

[4] F. Gibou, R. Fedkiw, L.-T. Cheng, and M. Kang. A Second-Order-Accurate Symmetric Discretization of the Poisson Equation on Irregular Domains. *Journal of Computational Physics*, 176:205–227, 02 2002.

[5] T. E. Tezduyar, M. Behr, S. Mittal, and A. A. Johnson. Computation of unsteady incompressible flows with the stabilized finite element methods: Space-time formulations, iterative strategies and massively parallel implementations. In *New Methods in Transient Analysis*, American Society of Mechanical Engineers, Pressure Vessels and Piping Division (Publication) PVP, pages 7–24. ASME, December 1992. Winter Annual Meeting of the American Society of Mechanical Engineers.

[6] S. Elgeti and H. Sauerland. Deforming fluid domains within the finite element method: Five mesh-based tracking methods in comparison. *Archives of Computational Methods in Engineering*, 23:323–361, 2016.

[7] F. Gibou and R. Fedkiw. A fourth order accurate discretization for the Laplace and heat equations on arbitrary domains, with applications to the Stefan problem. *Journal of Computational Physics*, 202:577–601, 01 2005.

[8] F. Gibou, L. Chen, D. Nguyen, and S. Banerjee. A level set based sharp interface method for the multiphase incompressible navier–stokes equations with phase change. *Journal of Computational Physics*, 222:536–555, 03 2007.

[9] T. E. Tezduyar, M. Behr, S. Mittal, and J. Liou. A new strategy for finite element computations involving moving boundaries and interfaces—the deforming-spatial-domain/space-time procedure: II. computation of free-surface flows, two-liquid flows, and flows with drifting cylinders. *Computer Methods in Applied Mechanics and Engineering*, 94(3):353–371, 1992.

[10] L. H. Pauli, S. T. Haßler, and M. Behr. Stabilized Finite Element Methods for Computational Design of Blood-Handling Devices. In *International Workshop on Modelling and Simulation in Biomedical Applications, 2017-10-24 - 2017-10-25, Mariatrost, Austria*, Oct 2017.

[11] A. Quarteroni. *Numerical Models for Differential Problems*. MS&A. Springer Milan, 2010.

[12] J. A. Sethian and P. Smereka. Level set methods for fluid interfaces. *Annual Review of Fluid Mechanics*, 35(1):341–372, 2003.

[13] J. Donea and A. Huerta. *Finite Element Methods for Flow Problems*, chapter 3, pages 79–146. John Wiley & Sons, Ltd, 2005.

[14] L. Boledi, B. Terschanski, S. Elgeti, and J. Kowalski. A level-set based space-time finite element approach to the modelling of solidification and melting processes. *Manuscript submitted for publication*, 2021. Preprint available at https://arxiv.org/abs/2105.09286.

[15] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, 95(15):8431—8435, July 1998.

[16] B. Dachwald, J. Mikucki, S. Tulaczyk, I. Digel, C. Espe, M. Feldmann, G. Francke, J. Kowalski, and C. Xu. Icemole: a maneuverable probe for clean in situ analysis and sampling of subsurface ice and subglacial aquatic ecosystems. *Annals of Glaciology*, 55(65):14–22, 2014.

[17] S. Ulamec, J. Biele, O. Funke, and M. Engelhardt. Access to glacial and subglacial environments in the Solar System by melting probe technology. *Environmental Science and Bio/Technology*, 6:71–94, 01 2007.

[18] Jülich Supercomputing Centre. JURECA: Modular supercomputer at Jülich Supercomputing Centre. *Journal of large-scale research facilities*, 4(A132), 2018.

Momentum conserving dynamic variational approach for the modeling of fiber-bending stiffness in fiber-reinforced composites

I. Kalaimani*, J. Dietzsch* and M. Groß*

* Technische Universität Chemnitz
Professorship of applied mechanics and dynamics
Reichenhainer Strasse 70, 09126 Chemnitz, Germany
e-mail: iniyan.kalaimani@mb.tu-chemnitz.de,
julian.dietzsch@mb.tu-chemnitz.de,
michael.gross@mb.tu-chemnitz.de

Key words: Fiber reinforced material, Fiber-bending stiffness, energy-momentum time integration, higher-order finite elements in space and time, Mixed variational principle

Abstract: *Rotor-dynamical systems made of 3D-fiber-reinforced composites which are subjected to dynamical loads exhibit an increased fiber bending stiffness in numerical simulations. We propose a numerical modeling approach of fiber-reinforced composites that treats this behavior accurately. Our model uses a multi-field mixed finite element formulation based on a dynamic variational approach, as demonstrated in [4], to perform long-term dynamic simulations that yield numerical solutions with increased accuracy in efficient CPU-time.*

We extend a Cauchy continuum with higher-order gradients of the deformation mapping as an independent field in the functional formulation, as suggested in [2], to model the bending stiffness of fibers accurately. This extended continuum also takes into account the higher-order energy contributions including the fiber curvature along with popular proven approaches that avoid the numerical locking effect of the fibers efficiently.

We apply the proposed approach on a cantilever beam with a hyperelastic, transversely isotropic, polyconvex material behavior in a transient dynamic analysis. The beam is subjected to a bending load with a strong dependence of the overall stiffness on the fiber orientation. The spatial and temporal convergence as well as the conservation properties are analyzed. It is observed that the model needs an improved numerical treatment to conserve total momenta as well as total energy.

1 INTRODUCTION

The finite element method for dynamical problems has received much attention over the last two decades, and approaches to solve them are still computationally demanding and time-consuming. The extensive development of new materials like fiber-reinforced composites constantly creates a need for more generalized algorithms for numerical simulations. The approach to avoid locking behavior in finite elements has significantly improved the accuracy and efficiency of almost any modern finite element code. Nevertheless, the usage of the same is not widely understood in the dynamic regime. Moreover, any modification of the standard continuum to better the accuracy of the numerical solution has to satisfy corresponding physical balance laws. Recent developments in the energy-momentum scheme provide a better opportunity to address these problems in a dynamic regime. The motivation of this work is to enhance the application of the lightweight design of rotating systems made of fiber-reinforced composites by taking advantage of the outcomes of the research mentioned above. In [1], the authors investigated the deformation patterns on mesoscopic level induced by the fibers in fiber-reinforced composites. Their results from the three-point bending test point out that these deformations eventually influence the bending stiffness of the composite material on the macroscopic level. Unfortunately, a standard Cauchy continuum is not well suited to capture

https://doi.org/10.4995/YIC2021.2021.12367

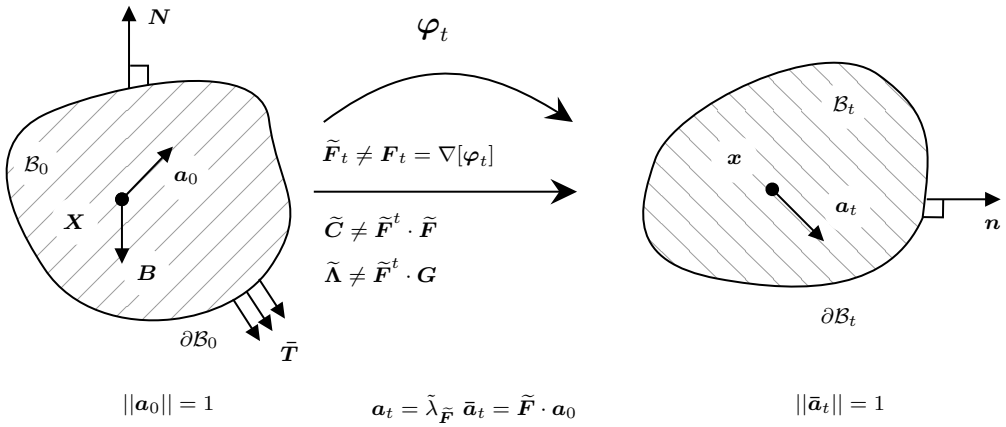


Figure 1: Transversely isotropic continuum with fibers oriented in direction of $\boldsymbol{a}_0$ and element-wise independent deformation gradient (cp. [3]).

these effects in fiber-reinforced models in numerical terms. Various approaches that have been proposed to solve this issue are limited to static problems [2, 8]. These drew our attention to capture the fiber bending stiffness in dynamical problems, which would help us reduce the unaccounted out-of-plane bending rigidity of an arbitrary geometry.

As a first step, to capture the fiber bending stiffness, we begin with assuming a constitutive model, where the strain energy function takes not only the strain and fiber direction vector into account, but also the information of fiber-curvature. A transversely isotropic continuum $\boldsymbol{\mathcal{B}}_0$ is considered with fibers at each point of the continuum oriented in direction of vector $\boldsymbol{a}_0$ in material configuration. In contrast to [2], we introduced a deformation gradient $\widetilde{\boldsymbol{F}}$ as an element-wise independent field in our Hu-Washizu based internal energy functional in [7]. Similarly, $\widetilde{\boldsymbol{\Gamma}}$ is introduced as an independent mixed field for $\nabla_{\boldsymbol{X}}[\widetilde{\boldsymbol{F}}]$ to capture the fiber curvature effects. In this work, we propose an additive split of strain energy function in terms of $\widetilde{\boldsymbol{C}}$ and $\widetilde{\boldsymbol{\Lambda}}$ as (see Figure 1),

$$\Psi^{\text{total}}(I_i(\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{\Lambda}})) \;=\; \Psi^{\text{iso}}(I_1(\widetilde{\boldsymbol{C}}), I_2(\widetilde{\boldsymbol{C}}), I_3(\widetilde{\boldsymbol{C}})) + \Psi^{\text{aniso}}(I_4, I_5) \;+\; \Psi^{\text{hg}}(I_6(\widetilde{\boldsymbol{\Lambda}})), \tag{1}$$

which is in line with the variation of [2], where $I_6 := \boldsymbol{k}_0 \cdot \boldsymbol{k}_0$, $\boldsymbol{k}_0 := (\widetilde{\boldsymbol{\Lambda}} \cdot \boldsymbol{a}_0)$ and $\widetilde{\boldsymbol{\Lambda}}$ is an independent mixed field for $\boldsymbol{\Lambda} := \widetilde{\boldsymbol{F}}^t \cdot \boldsymbol{G}$ which is the pure referential representation of $\boldsymbol{G}$. $\boldsymbol{G}$ is defined as the referential gradient of the spatial fiber direction vector $\boldsymbol{a}_t = \tilde{\lambda}_{\widetilde{\boldsymbol{F}}}\ \bar{\boldsymbol{a}}_t$ and $\tilde{\lambda}_{\widetilde{\boldsymbol{F}}}$ is the fiber stretch. Thus,

$$\boldsymbol{\Lambda} = \widetilde{\boldsymbol{F}}^t \cdot \left[\boldsymbol{a}_0 \cdot \nabla_{\boldsymbol{X}}[\widetilde{\boldsymbol{F}}^t] + \widetilde{\boldsymbol{F}} \cdot \nabla_{\boldsymbol{X}} \boldsymbol{a}_0\right]. \tag{2}$$

I_1, I_2, I_3, and I_4, I_5 are the usual isotropic and anisotropic principal invariants based on the right Cauchy green tensor $\widetilde{\boldsymbol{C}}$, which is a mixed field variable for $\boldsymbol{C} = [\widetilde{\boldsymbol{F}}]^t \cdot \widetilde{\boldsymbol{F}}$. For simplicity, anisotropic part of the strain energy due to I_4, I_5 is assumed to be constant and its effects are neglected within the framework of this work. In this way, we frame our new extended continuum.

2 PRINCIPLE OF VIRTUAL POWER

In the second step, we formulate the power functional for the extended continuum. The Equation (3) shows the internal power functional with new independent mixed field variables $\widetilde{\boldsymbol{F}}$ and $\widetilde{\boldsymbol{C}}$, energetically conjugated with independent first Piola-Kirchhoff stress tensor $\widetilde{\boldsymbol{P}}$ and second Piola-Kirchhoff stress tensor $\widetilde{\boldsymbol{S}}$, respectively. Similarly, $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\Lambda}}$ are energetically conjugated with independent $\widetilde{\boldsymbol{\mathfrak{B}}}$ and $\widetilde{\boldsymbol{\Lambda}}$, respectively.

$$\begin{aligned}\dot{\Pi}_{int} := & \int_{\boldsymbol{\mathcal{B}}_0} \left[\frac{\partial \Psi^{ela}(\widetilde{\boldsymbol{C}})}{\partial \widetilde{\boldsymbol{C}}} \dot{\widetilde{\boldsymbol{C}}}\right] \, dV + \int_{\boldsymbol{\mathcal{B}}_0} \left[\frac{\partial \Psi^{hg}(\widetilde{\boldsymbol{\Lambda}})}{\partial \widetilde{\boldsymbol{\Lambda}}} \dot{\widetilde{\boldsymbol{\Lambda}}}\right] \, dV - \int_{\boldsymbol{\mathcal{B}}_0} \widetilde{\boldsymbol{P}} : \left[\dot{\widetilde{\boldsymbol{F}}} - \nabla \dot{\boldsymbol{\varphi}}\right] \, dV \\ & - \int_{\boldsymbol{\mathcal{B}}_0} \widetilde{\boldsymbol{\mathfrak{B}}} \odot_3 \left[\dot{\widetilde{\boldsymbol{\Gamma}}} - \nabla \dot{\widetilde{\boldsymbol{F}}}\right] \, dV - \int_{\boldsymbol{\mathcal{B}}_0} \widetilde{\boldsymbol{A}} : \left[\dot{\widetilde{\boldsymbol{\Lambda}}} - \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{F}}} : \dot{\widetilde{\boldsymbol{F}}} - \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{\Gamma}}} \odot_3 \dot{\widetilde{\boldsymbol{\Gamma}}}\right] \, dV \\ & - \int_{\boldsymbol{\mathcal{B}}_0} \frac{1}{2} \widetilde{\boldsymbol{S}} : \left[\dot{\widetilde{\boldsymbol{C}}} - \dot{\widetilde{\boldsymbol{F}}^t \widetilde{\boldsymbol{F}}}\right] \, dV. \end{aligned} \tag{3}$$

Here we represent triple contraction of tensors by $\odot_3$.

The mass-specific body load $\boldsymbol{B}$ and a traction load $\bar{\boldsymbol{T}}$ on the Neumann boundary $\partial_T \boldsymbol{\mathcal{B}}_0$ are considered as external forces. Further, algorithmic stress tensors $\bar{\boldsymbol{S}}$ and $\bar{\boldsymbol{A}}$ are introduced in the external power functional to derive energy-momentum time integration. More details on this topic can be found in [5]. $\bar{\boldsymbol{\varphi}}$ denotes the prescribed boundary displacement with respect to the reaction force $\boldsymbol{R}$ as its associated Lagrange multiplier in the Dirichlet boundary $\partial_\varphi \boldsymbol{\mathcal{B}}_0$. These yield to the following external power functional,

$$\begin{aligned}\dot{\Pi}_{ext} := & - \int_{\boldsymbol{\mathcal{B}}_0} \rho_0 \boldsymbol{B} \cdot \dot{\boldsymbol{\varphi}} \, dV - \int_{\partial_T \boldsymbol{\mathcal{B}}_0} \bar{\boldsymbol{T}} \cdot \dot{\boldsymbol{\varphi}} \, dA - \int_{\partial_\varphi \boldsymbol{\mathcal{B}}_0} \boldsymbol{R} \cdot (\dot{\boldsymbol{\varphi}} - \dot{\bar{\boldsymbol{\varphi}}}) \, dA \\ & + \int_{\boldsymbol{\mathcal{B}}_0} \bar{\boldsymbol{A}} : \dot{\widetilde{\boldsymbol{\Lambda}}} \, dV + \int_{\boldsymbol{\mathcal{B}}_0} \frac{1}{2} \bar{\boldsymbol{S}} : \dot{\widetilde{\boldsymbol{C}}} \, dV. \end{aligned} \tag{4}$$

The algorithmic stress tensors are defined as,

$$\bar{\boldsymbol{A}} := \frac{\Psi(\widetilde{\boldsymbol{\Lambda}}_{n+1}) - \Psi(\widetilde{\boldsymbol{\Lambda}}_n) - \int_0^1 \frac{\partial \Psi(\widetilde{\boldsymbol{\Lambda}})}{\partial \widetilde{\boldsymbol{\Lambda}}} : \dot{\widetilde{\boldsymbol{\Lambda}}}}{\int_0^1 \dot{\widetilde{\boldsymbol{\Lambda}}} : \dot{\widetilde{\boldsymbol{\Lambda}}}} \quad \dot{\widetilde{\boldsymbol{\Lambda}}}, \tag{5}$$

$$\bar{\boldsymbol{S}} := \frac{\Psi(\widetilde{\boldsymbol{C}}_{n+1}) - \Psi(\widetilde{\boldsymbol{C}}_n) - \int_0^1 \frac{\partial \Psi(\widetilde{\boldsymbol{C}})}{\partial \widetilde{\boldsymbol{C}}} : \dot{\widetilde{\boldsymbol{C}}}}{\int_0^1 \dot{\widetilde{\boldsymbol{C}}} : \dot{\widetilde{\boldsymbol{C}}}} \quad \dot{\widetilde{\boldsymbol{C}}}. \tag{6}$$

And finally, the kinetic power functional with mass density ρ_0, velocity $\boldsymbol{v}$ and linear momentum $\boldsymbol{p}$ is defined by,

$$\dot{\mathcal{T}}_{kin} := \int_{\boldsymbol{\mathcal{B}}_0} \rho_0 \boldsymbol{v} \cdot \dot{\boldsymbol{v}} \, dV - \int_{\boldsymbol{\mathcal{B}}_0} \boldsymbol{p}(\dot{\boldsymbol{v}} - \ddot{\boldsymbol{\varphi}}) \, dV - \int_{\boldsymbol{\mathcal{B}}_0} \dot{\boldsymbol{p}}(\boldsymbol{v} - \dot{\boldsymbol{\varphi}}) \, dV. \tag{7}$$

3 WEAK FORMULATION

In the next step, to derive a weak formulation for extended continuum, we apply virtual power principle to the total energy balance of the system leading to the following equation,

$$\int_{t_n}^{t_{n+1}} \left[\delta_* \dot{\mathcal{T}}_{kin}\left(\dot{\boldsymbol{\varphi}}, \dot{\boldsymbol{v}}, \dot{\boldsymbol{p}}\right) + \delta_* \dot{\Pi}_{ext}\left(\dot{\boldsymbol{\varphi}}, \boldsymbol{R}\right) + \delta_* \dot{\Pi}_{int}\left(\dot{\boldsymbol{\varphi}}, \dot{\widetilde{\boldsymbol{F}}}, \dot{\widetilde{\boldsymbol{P}}}, \dot{\widetilde{\boldsymbol{C}}}, \widetilde{\boldsymbol{S}}, \dot{\widetilde{\boldsymbol{\Gamma}}}, \dot{\widetilde{\mathfrak{B}}}, \dot{\widetilde{\boldsymbol{\Lambda}}}, \widetilde{\boldsymbol{A}}\right)\right] dt = 0. \tag{8}$$

The variation of all power functionals is performed with respect to their dependencies to derive weak forms from the corresponding virtual powers. As in [4], the symbol δ_* is used in the sense of variation performed with respect to both temporally continuous time rate fields and temporally discontinuous Lagrange multiplier fields.

The resulting integrals of the weak forms of the extended Cauchy–Boltzmann continuum with fiber curvature is expressed in their continuous form in this paper for simplicity. The weak mechanical momentum equation is obtained as,

$$\begin{aligned}\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} [\dot{\boldsymbol{p}} - \rho_0 \boldsymbol{B}] \cdot \delta_* \dot{\boldsymbol{\varphi}}\, dV dt - \int_{t_n}^{t_{n+1}}\int_{\partial_T \mathcal{B}_0} \bar{\boldsymbol{T}} \cdot \delta_* \dot{\boldsymbol{\varphi}}\, dA dt \\ - \int_{t_n}^{t_{n+1}}\int_{\partial_\varphi \mathcal{B}_0} \boldsymbol{R} \cdot \delta_* \dot{\boldsymbol{\varphi}}\, dA dt + \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \widetilde{\boldsymbol{P}} : \nabla[\delta_* \dot{\boldsymbol{\varphi}}]\, dV dt \;=\; 0.\end{aligned} \tag{9}$$

To solve equation (9), the first Piola-Kirchhoff stress is required and determined from its weak form,

$$\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\left(\dot{\widetilde{\mathfrak{B}}} \odot_3 \frac{\partial(\nabla \dot{\widetilde{\boldsymbol{F}}})}{\partial \dot{\widetilde{\boldsymbol{F}}}}\right) + \left(\widetilde{\boldsymbol{A}} : \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{F}}}\right) + \frac{1}{2}\, \widetilde{\boldsymbol{S}} \cdot \left(\widetilde{\boldsymbol{F}}^t + \widetilde{\boldsymbol{F}}\right) - \widetilde{\boldsymbol{P}}\right] : \delta_* \dot{\widetilde{\boldsymbol{F}}}\, dV dt = \quad 0. \tag{10}$$

Similarly, to solve the above equation, independent strains tensors are obtained from their corresponding weak strain equations:

$$\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\dot{\widetilde{\boldsymbol{F}}} - \nabla \dot{\boldsymbol{\varphi}}\right] : \delta_* \widetilde{\boldsymbol{P}}\, dV dt = 0, \qquad \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\dot{\widetilde{\boldsymbol{C}}} - \overline{\widetilde{\boldsymbol{F}}^t \dot{\widetilde{\boldsymbol{F}}}}\right] : \delta_* \widetilde{\boldsymbol{S}}\, dV dt = 0, \tag{11}$$

and stress tensors from their corresponding weak stress equations:

$$\begin{aligned}\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[2\, \frac{\partial \Psi^{ela}}{\partial \widetilde{\boldsymbol{C}}} + \bar{\boldsymbol{S}} - \widetilde{\boldsymbol{S}}\right] : \delta_* \dot{\widetilde{\boldsymbol{C}}} dV dt = 0, \\ \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\frac{\partial \Psi^{hg}}{\partial \widetilde{\boldsymbol{\Lambda}}} + \bar{\boldsymbol{A}} - \widetilde{\boldsymbol{A}}\right] : \delta_* \dot{\widetilde{\boldsymbol{\Lambda}}} dV dt = 0.\end{aligned} \tag{12}$$

Weak curvature-strain equations are expressed as,

$$\begin{aligned}\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\dot{\widetilde{\boldsymbol{\Gamma}}} - \nabla \dot{\widetilde{\boldsymbol{F}}}\right] \odot_3 \delta_* \dot{\widetilde{\mathfrak{B}}} dV dt = 0, \\ \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \left[\dot{\widetilde{\boldsymbol{\Lambda}}} - \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{F}}} : \dot{\widetilde{\boldsymbol{F}}} - \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{\Gamma}}} \odot_3 \dot{\widetilde{\boldsymbol{\Gamma}}}\right] : \delta_* \widetilde{\boldsymbol{A}} dV dt = 0.\end{aligned} \tag{13}$$

Linear momentum needed in the weak mechanical momentum equation (9) can be obtained by dissolving weak velocity equation into weak momentum equation,

$$\int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} [\boldsymbol{v} - \dot{\boldsymbol{\varphi}}] \cdot \delta_* \dot{\boldsymbol{p}} \; dV dt = 0, \qquad \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} [\rho_0 \boldsymbol{v} - \boldsymbol{p}] \cdot \delta_* \dot{\boldsymbol{v}} \; dV dt = 0. \tag{14}$$

We discretize the weak forms spatially and temporally on the elemental level by Gaussian quadrature using Lagrangian ansatz funtions. The time rate variable fields $(\dot{\bullet})_i^{e,n}$ are approximated on the n-th time step by $k+1$-th order Lagrange polynomials corresponding to the normalized time $\alpha \in [0,1]$ on each time step $[t_n, t_{n+1}]$ by

$$(\bullet)^{h,n} = \sum_{i=1}^{k+1} M_i(\alpha)(\bullet)_i^{e,n}, \qquad (\dot{\bullet})^{h,n} = \sum_{i=1}^{k+1} \dot{M}_i(\alpha)(\bullet)_i^{e,n}, \tag{15}$$

and the stress fields as well as Lagrange multiplier fields are $(\tilde{\bullet})_i^{e,n}$ are approximated on the n-th time step by k-th order Lagrange polynomials by

$$(\tilde{\bullet})^{h,n} = \sum_{i=1}^{k} \tilde{M}_i(\alpha)(\tilde{\bullet})_i^{e,n}, \qquad M_i(\alpha) = \prod_{\substack{j=1 \\ j\neq i}}^{k+1} \frac{\alpha - \alpha_j}{\alpha_i - \alpha_j}, \; 1 \leq i \leq k+1. \tag{16}$$

Similarly e-th finite element are approximated in space using standard local shape functions $N^A(\xi), \; A = 1, \cdots, n_{node}$ defined on the reference domain. The resulting tangent matrix is condensated to pure displacement form by staggering the solution of globally discontinuous mixed fields on the elemental level. We implement this in our In-house finite element code 'fEMcon' and the resulting linear systems of equations are solved using PARDISO solver [6].

4 BALANCE LAWS

With the introduction of new independent field variables, the extended standard Cauchy continuum has to fulfill physical balance laws. To conserve total momentum and energy at every discrete time step entails doing a special numerical treatment.

Following the steps in [5], suitable test functions $\delta_* \dot{\boldsymbol{\varphi}}$, $\delta_* \dot{\widetilde{\boldsymbol{F}}}$ and $\delta_* \widetilde{\boldsymbol{P}}$ are employed in (9) for an arbitrary axial vector $\boldsymbol{c} =$ constant and eliminating the first Piola-Kirchhoff tensor yields a time-integrator that eventually conserves total angular momentum,

$$\begin{aligned} \boldsymbol{\mathcal{J}}_{n+1} \; - \; \boldsymbol{\mathcal{J}}_n \; = \; & \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} [\boldsymbol{\varphi} \times \rho_0 \boldsymbol{B}] \; dV dt + \int_{t_n}^{t_{n+1}}\int_{\partial_T \mathcal{B}_0} \left[\boldsymbol{\varphi} \times \bar{\boldsymbol{T}}\right] \; dA \; dt \\ & + \int_{t_n}^{t_{n+1}}\int_{\mathcal{B}_0} \boldsymbol{\epsilon} : \left[\left(\widetilde{\mathfrak{B}} \odot_3 \frac{\partial(\nabla \dot{\widetilde{\boldsymbol{F}}})}{\partial \dot{\widetilde{\boldsymbol{F}}}} \right) + \left(\widetilde{\boldsymbol{A}} : \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{F}}} \right) + \frac{1}{2} \, \widetilde{\boldsymbol{S}} \cdot \left(\widetilde{\boldsymbol{F}}^t + \widetilde{\boldsymbol{F}} \right) \right] \widetilde{\boldsymbol{F}}^t \; dV dt. \end{aligned} \tag{17}$$

Similarly, employing different choice of suitable test functions for $\delta_*\dot{\boldsymbol{\varphi}}$, $\delta_*\dot{\widetilde{\boldsymbol{F}}}$ and $\delta_*\widetilde{\boldsymbol{P}}$ and eliminating the first Piola-Kirchhoff tensor, we derive the total energy conserving time-integrator as below,

$$\boldsymbol{\mathcal{K}}_{n+1} - \boldsymbol{\mathcal{K}}_n = \int_{t_n}^{t_{n+1}}\int_{\boldsymbol{\mathcal{B}}_0} \dot{\boldsymbol{\varphi}} \cdot \rho_0 \boldsymbol{B}\, dV dt + \int_{t_n}^{t_{n+1}}\int_{\partial\boldsymbol{\mathcal{B}}_0} \dot{\boldsymbol{\varphi}} \cdot (\bar{\boldsymbol{T}} + \boldsymbol{R})\, dA\, dt$$
$$- \int_{t_n}^{t_{n+1}}\int_{\boldsymbol{\mathcal{B}}_0} \left[\left(\widetilde{\mathfrak{B}} \odot_3 \frac{\partial(\nabla \dot{\widetilde{\boldsymbol{F}}})}{\partial \dot{\widetilde{\boldsymbol{F}}}} \right) + \left(\widetilde{\boldsymbol{A}} : \frac{\partial \boldsymbol{\Lambda}}{\partial \widetilde{\boldsymbol{F}}} \right) + \frac{1}{2}\, \widetilde{\boldsymbol{S}} \cdot \left(\widetilde{\boldsymbol{F}}^t + \widetilde{\boldsymbol{F}} \right) \right] : \dot{\widetilde{\boldsymbol{F}}}^t\, dV dt. \tag{18}$$

5 NUMERICAL EXAMPLE

In order to understand the anisotropic behavior exhibited by the fibers, we apply our proposed approach on a simple cantilever beam of length 15cm, width 2cm, and height 1cm. The numerical model is assumed to be reinforced with a single family of extensible fibers submerged in the matrix material and exhibiting resistance to bending. 20-noded tri-quadratic serendipity elements are used to discretize the beam into 24 finite elements. A Gaussian quadrature scheme with 27 quadrature points is employed to evaluate the integrals numerically. The left end is chosen as the Dirichlet boundary, such that the displacement of nodes at this boundary are fixed in all three directions $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$. As a Neumann boundary condition, on the opposite

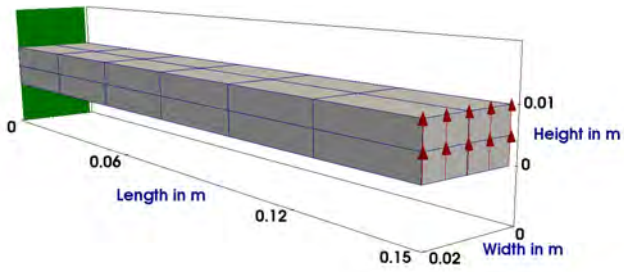


Figure 2: Dimensions and boundary conditions for the simply supported beam subject to bending load.

Figure 3: Assumed initial fiber orientation along the length of the beam ($\boldsymbol{a}_0 = \boldsymbol{e}_1$) shown in 2D.

free end of the beam, a deformation-dependent transient pressure load $\hat{p} = 200\hat{f}$ is prescribed, which always creates traction in the direction parallel to this surface. Standard Neo-Hookean type material ansatz is chosen for the isotropic part of the strain energy function and cI_6 for the higher-order gradient energy part as in [2],

$$\Psi^{total}(I_1, I_3, I_6; \lambda, \mu, c) = \lambda \frac{I_3 - 1}{4} - \left[\frac{\lambda}{2} + \mu \right] \ln\left(\sqrt{I_3}\right) + \frac{\mu}{2}\left[I_1 - 3\right] + c\, I_6. \tag{19}$$

For this setup, the simulation is performed for following test cases:

1. the fiber bending stiffness material parameter c is varied assuming the fibers are oriented with beam's axis, i.e. $\boldsymbol{a}_0 = \boldsymbol{e}_1$
2. the fiber orientation $\boldsymbol{a}_0$ is varied for a constant value of the fiber bending stiffness material parameter c

The material and simulation parameters chosen as per the table above are in SI units.

Table 1: Simulation parameters for the choice of initial fiber orientation and bending stiffness parameter

Ψ^{ela}		Ψ^{k_0}		Temporal parameters					
λ	μ	c	α	φ_0	$\boldsymbol{v}_0$	T	h_n	Tol	ρ_0
100×10^6	0.1×10^6	$[0.08, 12] \cdot 10^6$	$[0, \pi/2]$	0	0	[0, 1]	0.002	10^{-4}	10^3

Remark 1. *Note that in this article, we consider fiber bundles called rovings. Therefore the fiber bending stiffness parameter c with the unit $N = Pa \cdot m^2$ is calculated with respect to the average diameter $l \approx [0.9, 11] \cdot 10^{-3}$ of the fiber bundle using the relation $c = \mu l^2$. Consequently, the average diameter l represents the length scale of the material.*

5.1 Influence of the fiber bending stiffness

For $c = 0$, the numerical model exhibits the behavior of a non-reinforced beam. As expected, our results show that increasing the stiffness parameter value stiffens the overall response of the composite material. With the increasing values of c, the onset of the higher-gradient part of the energy function is more pronounced. However, fiber stiffness has no significant influence beyond a certain range of c for a chosen load. Figure 4 presents the trend of displacement of a point P at the top edge of the Neumann boundary in $\boldsymbol{e}_2$ direction.

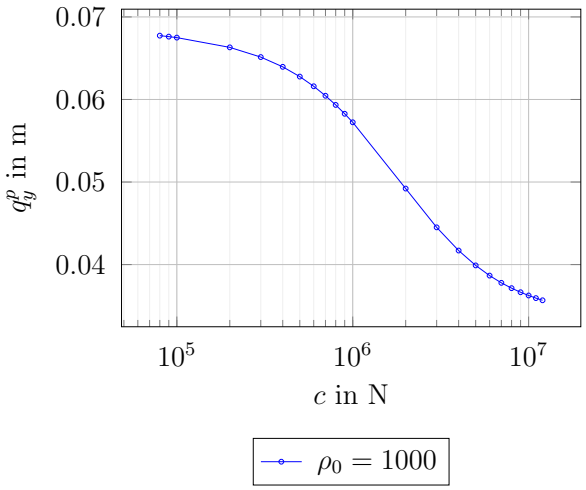


Figure 4: Displacement of a point P at the free end of the beam for varying fiber bending stiffness parameter c, with fibers oriented along $\boldsymbol{e}_1$.

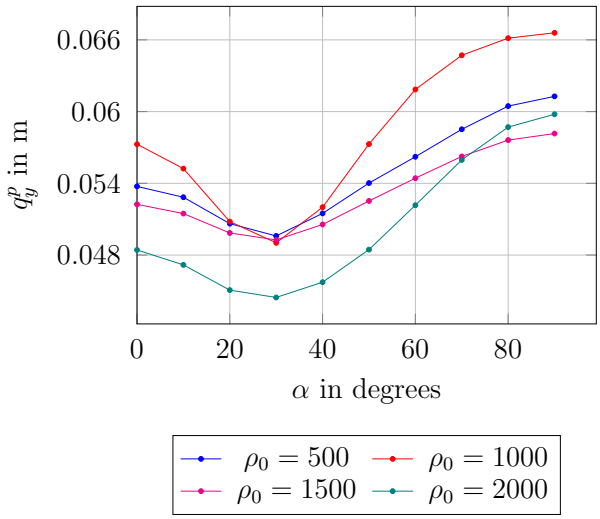


Figure 5: Displacement of a point P at the free end of the beam for varying fiber orientations α with fiber bending stiffness parameter $c = 10^6$.

5.2 Influence of the fiber orientation

For $c = 10^6$, the orientations of the fiber reinforcements are varied to understand the behavior of the beam. The fiber angle takes values between $\alpha = 0$ and $\alpha = \pi/2$. It is observed in Figure 5 that the cantilever is less stiffer for the fiber orientation $\boldsymbol{a}_0 = [1\ 1\ 1]^t$, and as expected, the composite is out of its longitudinal plane. What is surprising is the fact that the degree of stiffening achieved for $\alpha = \pi/6$ and $\alpha = \pi/4$ is more than for the orientation $\alpha = 0$, which is counter-intuitive. However, from Figure 5 we can understand that the trend of the plot is independent of inertia. Despite this, we can still state that our proposed time integrator conserves total momenta and total energy.

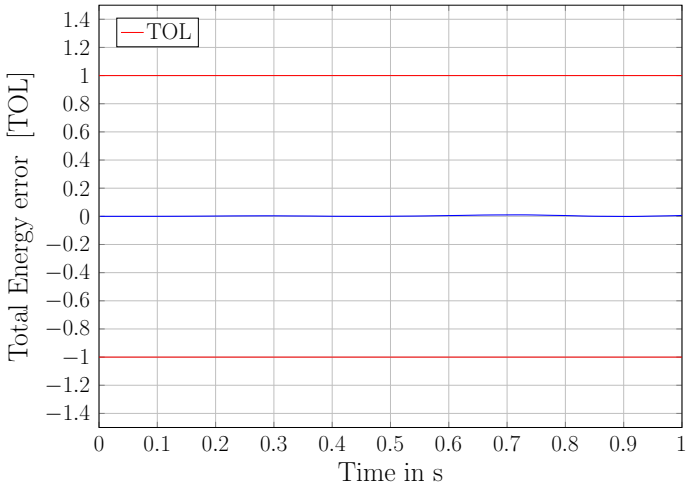


Figure 6: Total energy error plot for $c = 10^7$ with fibers oriented in the direction of $\boldsymbol{e}_1$ and a Newton-Raphson tolerance $TOL = 10^{-4}$.

5.3 Different type of energy function

Complementing the above fact on total energy conservation in previous test cases, we also studied the effect of the time integrator with a non-linear energy function based on the quadratic of curvature measure cI_6^2. It is evident from Figure 6 that for the chosen linear and non-linear type of anisotropic energy functions with respect to higher-order gradient of the deformation gradient, the total energy is conserved.

6 CONCLUSIONS

To sum up our work, we demonstrated the influence of fiber curvature on the bending stiffness of the cantilever beam as a numerical example. We introduced an independent field variable for spatial fiber direction vector in the continuum using Hu-Washizu's principle to capture the curvature effect. We have succeeded in combining an energy-momentum scheme with the principle of virtual power for the proposed mixed element formulation to preserve the time evolution of energy functions. In this way, the spurious errors arising from fibers are significantly reduced in numerical simulations. In addition to that, our energy-momentum scheme guarantees to obtain the desired accuracy with larger time steps and therefore reduced total CPU time.

The presented contribution has highlighted the importance of the curvature measure through the invariant I_6 in bending-dominated problems in dynamic scenarios. The maximum bending stiffness has been obtained with increased fiber bending stiffness parameter. For the simulated coarse mesh, it is observed that the bending stiffness is maximum when the fibers tend to align with the beam's axis except for smaller angles. To further our research, we intend to extend our formulations to thermomechanical problems.

Acknowledgments This research was made possible by the DFG under grants GR 3297/6-1 and GR 3297/4-2, which is gratefully acknowledged.

REFERENCES

[1] Madeo, A., Ferretti, M., dell'Isola, F. and Boisse, P. A Thick fibrous composite reinforcements behave as special second-gradient materials: three-point bending of 3D interlocks. *Z. Angew. Math. Phys.*, Vol. **66**, pp. 2041–2060, (2015).

[2] Asmanoglo, T. and Menzel, A. A multi-field finite element approach for the modelling of fibre-reinforced composites with fibre-bending stiffness. *Comput. Methods Appl.Mech. Engrg.*, Vol. **317**, pp. 1037–1067, (2017).

[3] Groß, M., Dietzsch, J. and Röbiger, C., Variational-based higher-order accurate energy–momentum schemes for thermo-viscoelastic fiber-reinforced continua. *Comput. Methods Appl.Mech. Engrg.*, Vol. **336**, pp. 353–418, (2018).

[4] Groß, M. and Dietzsch, J. Variational-based locking-free energy–momentum schemes of higher-order for thermo-viscoelastic fiber-reinforced continua. *Comput. Methods Appl.Mech. Engrg.*, Vol. **343**, pp. 631–671, (2019).

[5] Groß, M., Dietzsch, J. and Röbiger, C., Non-isothermal energy–momentum time integrations with drilling degrees of freedom of composites with viscoelastic fiber bundles and curvature–twist stiffness. *Comput. Methods Appl.Mech. Engrg.*, Vol. **365**, 112973, (2020).

[6] Alappat, C., Basermann, A., Bishop, A. R., Fehske, H., Hager, G., Schenk, O., Thies,J., and Wellein, G. A Recursive Algebraic Coloring Technique for Hardware-Efficient Symmetric Sparse Matrix-Vector Multiplication. *ACM Transactions on Parallel Computing*, Vol. **7**, No. 3(19),(2020).

[7] Kalaimani, I., Dietzsch, J. and Gross, M. Modeling of fiber-bending stiffness in fiber-reinforced composites with a dynamic mixed finite element method based on the principle of virtual power. *Proc. Appl. Math. Mech.*, (2021), Submitted.

[8] Witt, C., Kaiser, T. and Menzel, A. An isogeometric finite element approach to fibre-reinforced composites with fibre bending stiffness. *Archive of Applied Mechanics*, Vol. **91**, pp. 643–672, (2021).

Computing the jump-term in space-time FEM for arbitrary temporal interpolation

Eugen Salzmann*, Florian Zwicke† and Stefanie Elgeti†

* Chair for Computational Analysis of Technical Systems(CATS)
RWTH Aachen University
Schinkelstr. 2, 52062 Aachen, Germany
e-mail: Salzmann@cats.rwth-aachen.de

† Institute of Lightweight Design and Structural Biomechanics (ILSB)
Vienna University of Technology
Grumpendorferstr. 7, 1060 Vienna, Austria
e-mail: Zwicke@ilsb.tuwien.ac.at; Elgeti@ilsb.tuwien.ac.at

Key words: finite-elements, space-time, deforming domains, discontinuous galerkin

Abstract: *One approach with rising popularity in analyzing time-dependent problems in science and engineering is the so-called space-time finite element method that utilizes finite elements in both space and time. A common ansatz in this context is to divide the mesh in temporal direction into so-called space-time slabs, which are subsequently weakly connected in time with a discontinuous galerkin approach. The corresponding jump-term, which is responsible for imposing the weak continuity across space-time slabs, can be challenging to compute, in particular in the context of deforming domains. Ensuring a conforming discretization of the space-time slab at the top and bottom in time direction simplifies the handling of this term immensely. Otherwise, a computationally expensive and error prone projection of the solution from one time-level to another is necessary. However, when it comes to simulations with deformable domains, e.g. for free-surface flows, ensuring conforming meshes is quite laborious. A possible solution to this challenge is to extrude a spatial mesh in time at each time-step resulting in the so-called time-discontinuous prismatic space-time (D-PST) method [1]. However, this procedure is restricted to finite-elements of 1st order in time. We present a novel algorithmic approach for arbitrarily discretized meshes by flipping the mesh in time-direction for each time-step. This ansatz allows for a simple evaluation of the jump-term as the mesh is always conforming. The cost of flipping the mesh around its symmetry plane in time scales with the number of nodes, which makes it computationally cheaper than an additional update of the mesh to enforce conformity or the evaluation of a projection. We validate the approach on various physical problems with and without deforming domains.*

1 INTRODUCTION

Space-time is the extension of the finite element concept in time. It was first introduced in 1988 by Thomas J.R. Hughes for classical elastodynamics with a proven convergence theorem [2]. Nowadays, it is more commonly used in fluid problems, especially since the introduction of the deformable-spatial-domain/space-time (DSD/SST) method [1, 3]. DSD/SST is beneficial for free-surface-flows, where the computational domain is unknown as they allow for a convenient way to track the boundary [4]. The initial version of space-time, as well as many other adaptations, involves a discontinuous galerkin approach in time that leads to an additional jump term between so-called space-time slabs. The evaluation of this term can be challenging. Therefore we present an algorithmic approach for a more straightforward implementation of this term.

https://doi.org/10.4995/YIC2021.2021.12588

2 METHOD

In this section, we present our approach to the jump term in DG-Space-time methods. The section is structured as follows. First, we recap the basics of space-time methods, including the treatment of deforming domains. The next part focuses on possible treatments of the jump-term, including the flipping ansatz.

2.1 SPACE-TIME

Space-time methods utilize finite elements in space and time rather than finite differences as in semi-discrete settings, resulting in a finite element analysis of the full space-time domain. There are various space-time methods that can be categorized with respect to the employed element type as well as the time continuity. In this work, we choose prismatic elements and a discontinuous galerkin approach in time, resulting in an analysis of so-called space-time slabs illustrated in Fig. 2.1. These slabs can consist of one or multiple elements in time and can

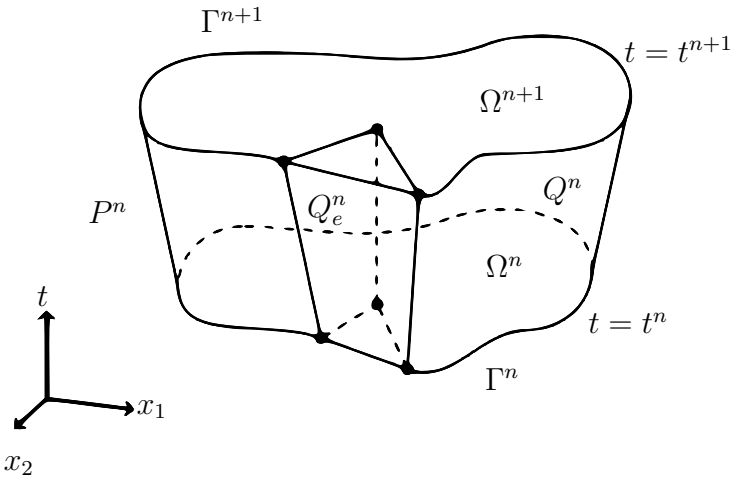


Figure 1: Illustration of a space-time-slab Q^n and an exemplary single element Q_e.

be considered an extension of a spatial mesh in time-direction. A jump-term, as typical for DG methods, weakly enforces continuity in time direction over multiple slabs. In this context, the weak form of a transient heat conduction equation reads: find $T \in S_a\left(Q^n\right)$ such that $\forall w \in S_t\left(Q^n\right)$:

$$\int_{Q^n} w \frac{\partial T}{\partial t} \,\mathrm{d}\mathbf{x} = \int_{Q^n} w \alpha \triangle T \,\mathrm{d}\mathbf{x} + \int_{\Omega_n} w \left(T|_{t_n}^{+} - T|_{t_n}^{-}\right) \mathrm{d}\mathbf{x}. \tag{1}$$

Where $Q^n = \Omega \times \left[t_n, t_{n+1}\right]$, $T|_{t_n}^{\pm} = \lim_{\epsilon \to 0} T(t_n \pm \epsilon)$, $\alpha \hat{=}$ thermal diffusivity and S_a, S_t are the appropriate ansatz and testing spaces on Q^n. Problems involving deforming domains can profit from space-time methods despite introducing additional complexity or restrictions. One benefit is incorporating deformation by formulating the weak form over the deforming or deformed domain. That way, the movement or deformation is considered in the solution procedure without modifying governing equations. When the movement is known, this can be particularly useful as the solution could be found over the entire time interval in one step. In problems where the deformation of a spatial domain is unknown beforehand, space-time methods with DG in time face a challenge concerning the jump term. The solution at the top of the slab from the previous time-step has to be projected to the bottom of the slab in the current time-step, which is easy when the mesh is conforming. In that case, there is a direct relation between the degrees of freedom. For non-conforming meshes, it is more challenging, and one possible

solution namely a projection introduces an additional error. Furthermore, the domain itself may deform, resulting in entirely non-matching domains. This can be avoided by (1) mesh update schemes that ensure mesh conformity or (2) a restriction to single-element layers in time.

2.2 MESH INVERSION/FLIPPING

We propose an alternative algorithmic approach, where the evaluation of the jump-term is easy to implement and works for arbitrary discretizations. Let us first focus on the implementation of the space-time method itself. Consider a 3D semi-discrete domain as compared to a 2D+time domain. Even though geometrically identical, algorithmically, one observes differences. These lie in the underlying operators in the PDE as well as the additional jump term. In terms of differential operators, in space-time approaches, the temporal derivative needs to be evaluated in a finite element sense whereas spatial derivatives need to be restricted to the spatial dimension only and can no longer be evaluated on the full domain. We adapt the FE mapping between the reference and physical space to consider the changes in the spatial differential operator and to scale the input mesh with the time-step size. As a result, within the computational mesh, the time coordinates are usually contained in $[0, 1]$ and the physical time is considered only through the mapping. Furthermore, for every space-time slab, the initial solution, from where the iterative solution scheme commences, is set to zero. Note that this entails that one can manipulate mesh coordinates without disturbing the solution process. We make use of this and invert the time coordinates. For every new time-step, before the iterative solution process starts, for every mesh node, we set the new t-coordinate t^* as

$$t^* = t_{max} - t + t_{min}. \tag{2}$$

Here, $t_{max/min}$ is the maximum/minimum value for t in the slab, which are commonly 1 and 0. As a result, the vertices and corresponding degrees of freedom move from the bottom to the top and vice versa without moving in space, ensuring conformity at the slab interface.

3 NUMERICAL STUDIES

This section aims to validate the presented approach. We focus on transient heat conduction problems and complement our test cases with a hyperplastic solid mechanical bending problem. For showing that the approach is valid for moving domains, we present a test case with a prescribed motion.

3.1 TRANSIENT HEAT ANALYSIS

In our first test case, we follow [6] and analyze heat conduction in a 2D rod. The rod is a rectangular domain and adiabatic everywhere except on the left side, where a fixed heat flux of 1 W is prescribed. Tab. 1 contains the geometry and simulation parameters for this test case.

Table 1: Parameters for transient heat conduction analysis of a 2D rod.

Parameter	Value	Unit
Length	20	m
Width	1	m
Thermal diffusivity ($\alpha = \frac{\kappa}{\rho c_p}$)	1	$\frac{m^2}{s}$
Reference time	1	s
Reference length	2	m

Reference time and length refer to the values at which we compare with the exact solution.

Despite the 2D geometry, the problem reduces to a 1D phenomenon as the heat propagates equally over the width. An analytical solution is given by

$$T(x,t) = 2\sqrt{\frac{t}{\pi}}\left[e^{\frac{-x^2}{4t}} - \frac{1}{2}x\sqrt{\frac{\pi}{t}}erfc\left(\frac{x}{2\sqrt{t}}\right)\right], \tag{3}$$

and shown in Fig. 2 together with a space-time and implicit Euler solution. The resulting solutions for the temperature distribution are depicted in Fig. 2. All solutions are visually identical. Therefore, we proceed to analyze the errors of the individual implementations. Our

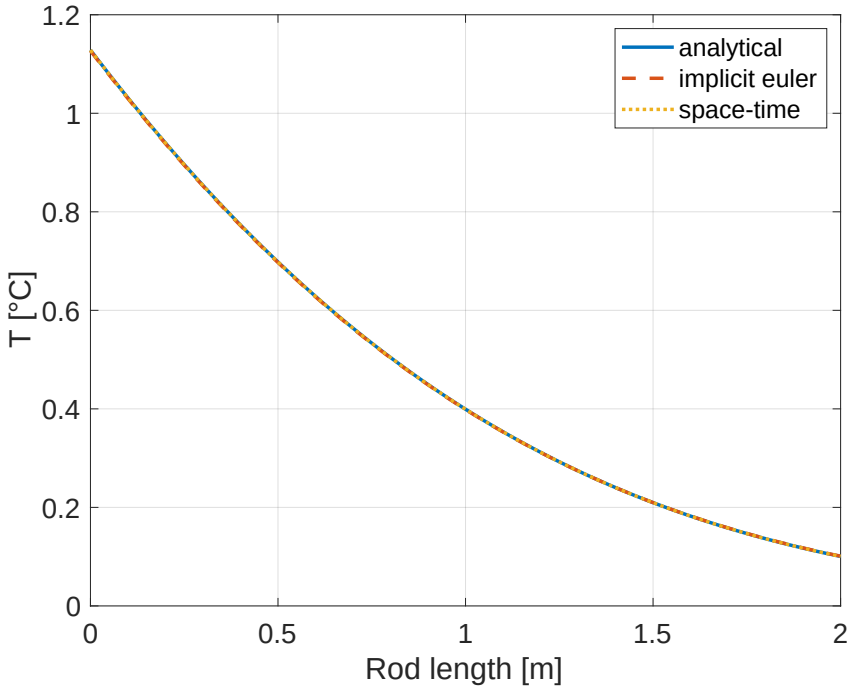


Figure 2: Exact temperature distribution on the most left 2 m of the rod after 1 sec.

first comparison is between the presented approach and a classical implementation of the space-time jump-term. Fig. 3 shows the error between these techniques and between each of them and the exact solution for two different discretizations. The curve index refers to the discretization while the "difference" curves show the error between the two implementations. The mesh details are the following:

Table 2: Mesh sizes for comparison between flipped and classical space-time implementations.

Curve index	elements in length	elements in width	time step size
1	1000	1	0.1
2	10000	1	0.05

From Fig. 3, it is evident that the effect of the mesh inversion is negligable in comparison to the overall discretization error. Fig. 4 shows the evolution of the error between the analytical solution and various space-time simulations as well as one implicit Euler simulation under temporal refinement. The error is evaluated as the L-2 norm over the first 2 m of the mesh normalized by the number of evaluation points (N). The analysis was performed on a quadrilateral 100.000x1 element spatial mesh. Note that the error of the implicit Euler scheme comes close to the theoretical convergence order of one. In comparison, the error of the space-time

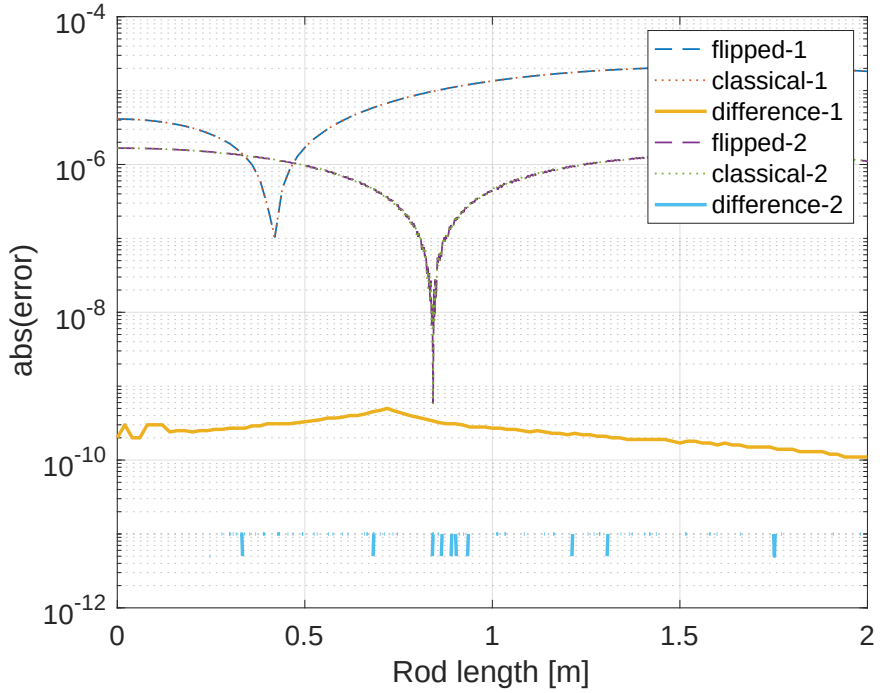


Figure 3: Error comparison of heat conduction analysis between classical and flipped space-time.

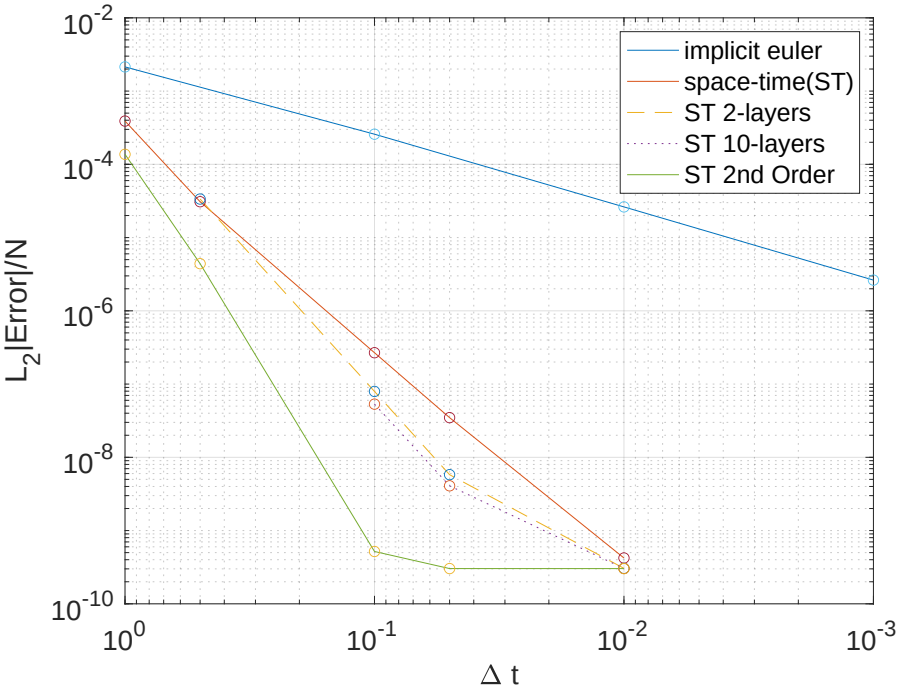


Figure 4: Relative error evolution of heat conduction analysis under temporal refinement.

ansatz decreases significantly faster when decreasing the time step size. However, not further than about $5.0e^{-10}$ where the spatial discretization error dominates and temporal refinement results in no improvement. For a fair comparison, the depicted time-step size Δt in the case of multiple time layer meshes corresponds to the layer thickness.

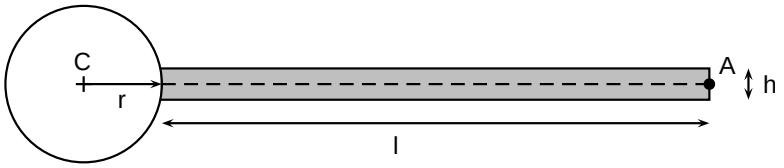


Figure 5: Illustration of the beam used for structural analysis [7].

3.2 STRUCTURAL ANALYSIS

The next test case we would like to present is the transient structural analysis of a beam that is fixed on the left side to a cylinder and bending under its weight. The clylinder is not part of the computational domain. The test case is taken from [7] where it is used to validate a structural solver before investigating fluid-structure-interaction. A hyperelastic material model, namely the St-Venant-Kirchhoff model, is employed. The problem is modeled from a classical lagrangian point of view, and we avoid 2nd order temporal derivatives by introducing a velocity field and solving it. The geometry is depicted in Fig. 5, and the following parameters were used:

Table 3: Parameters of structural analysis test-case.

Parameter	value	unit
length	35	cm
width	2	cm
cylinder radius r	5	cm
density	1000	$\frac{kg}{m^3}$
1st lamme parameter λ	2	$10^6 \frac{kg}{ms^2}$
2nd lamme prameter μ	0.5	$10^6 \frac{kg}{ms^2}$
gravity (y-direction)	-2	$\frac{m}{s^2}$

We compare the displacement of the reference point A at the tip of the beam shown in Fig. 5. Fig. 6 illustrates the displacement over 10 seconds in X and Y direction. The simulation details are given in Tab. 4. The results seem to be in good agreement. Nevertheless, there are minor

Table 4: Simulation parameters of structural analysis results.

run	elements	Δt
csm l4	5120	5ms
Crank-Nicolson (CN)	20x128	5ms
Space-time (ST)	20x128	5ms
CN dt=1ms	20x128	1ms

deviations. Fig. 7 zooms in on the last period of the oscillation between 9.2 and 10 seconds, and we focus on the displacement in x-direction as the differences are more visible in that direction. The result employing a Crank-Nicolson scheme and the same time-step is very close to the reference values, and different meshes can explain the occurring fine distinctions. CSM l4 uses an unstructured mesh with 5120 quadrilateral elements, while we choose a structured mesh of 20x128 quadrilateral elements. The space-time result exhibits slightly different behavior, which is close to the results of a Crank-Nicolson simulation with a refined time-step. However, as shown in the previous test case, space-time exhibits a superior accuracy with respect to the time-step size, so this result is to be expected.

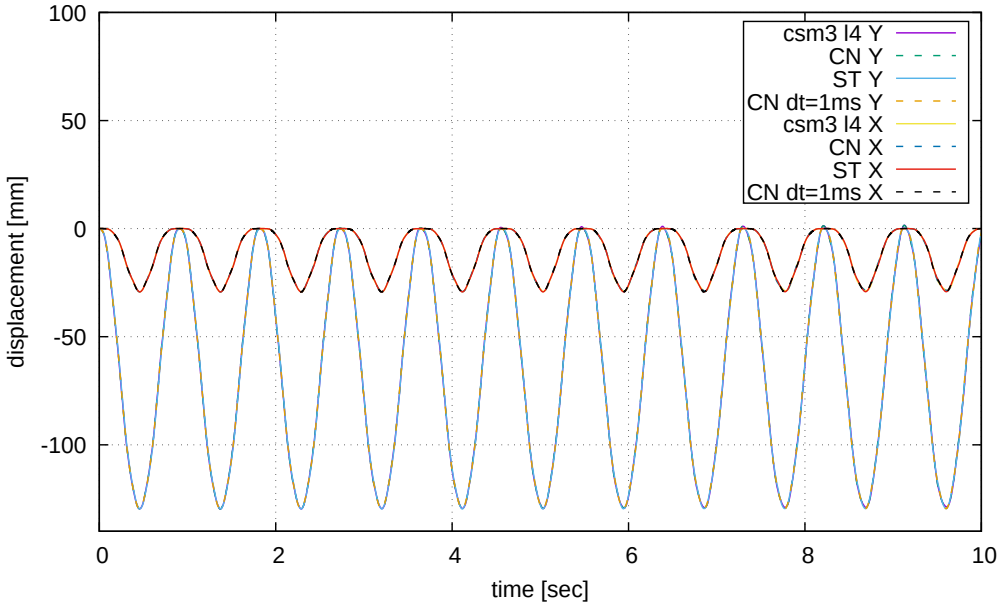


Figure 6: Displacement of reference point A over 10 seconds.

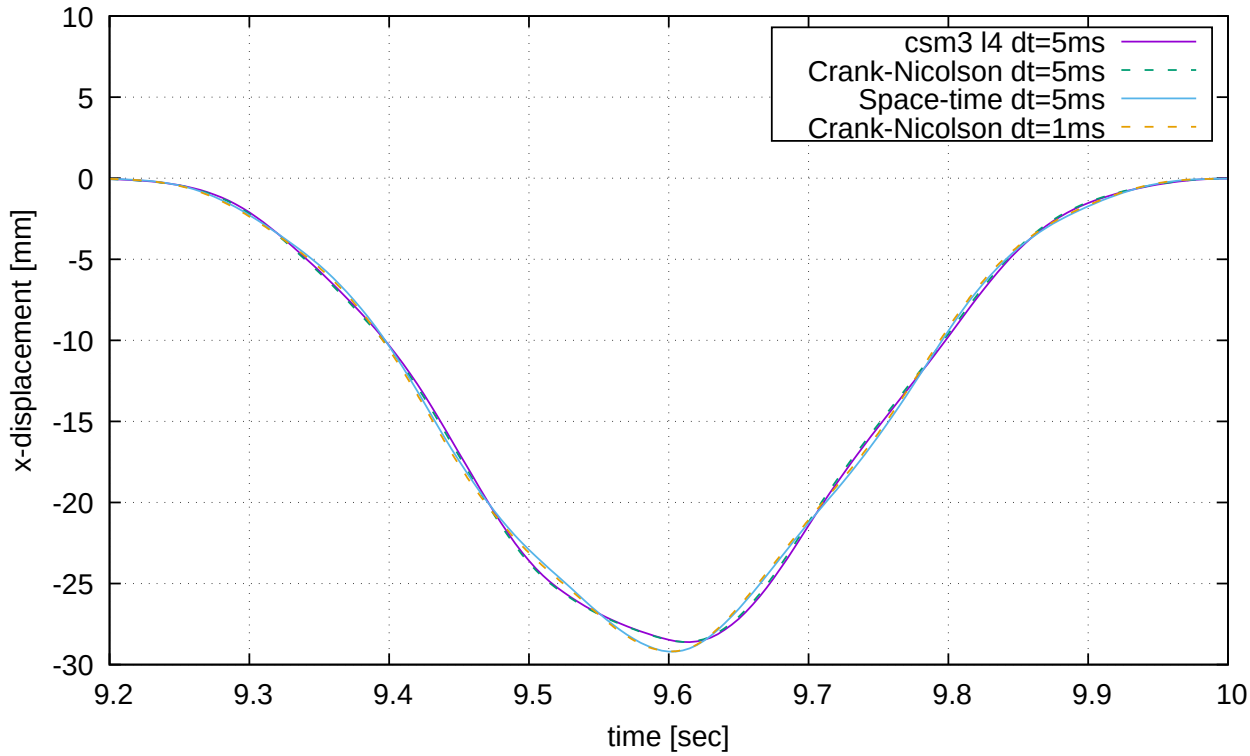


Figure 7: Displacement in x-direction within the last considered period.

3.3 RIGID-BODY MOVEMENT

The last test case we are discussing is the movement of a rigid body. This test case is somewhat artificial as we solve a continuum mechanical system of equation coupled with an elastic mesh-update problem and a free-surface approach, similar to what is described by Elgeti and Zwicke [4, 5]. However, as we are only interested in the deformation of the domain, we reduce the problem to a rigid body movement by imposing a fully developed velocity field of $v = (0.1, 0.1)^T$ as initial and boundary conditions. The test case is a 1x1 sized spatial domain

that then has to move $(1,1)^T$ in 10 seconds. Fig. 8 shows the result after one time-step with size 10. The t-axis is pointing out of the plane, and we view the x-y plane, meaning, at the bottom, the body is in its initial configuration, and the mesh connects it to its new position after 10 seconds at the top. The legend shows that the magnitude of the displacement is $\sqrt{2}$ after 10 seconds, which corresponds to the analytical solution.

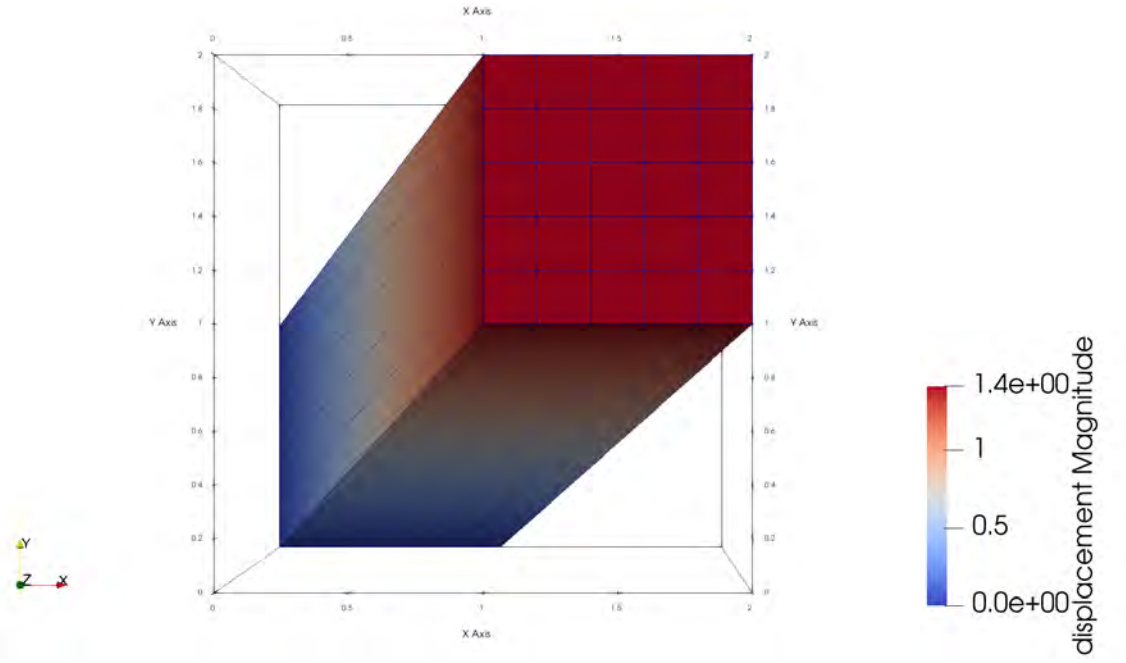


Figure 8: Rigid-body movement result after 10 seconds.

4 CONCLUSION

In this work, we presented an algorithmic approach to the treatment of jump terms in the context of space-time discontinuous finite element methods. We showed that inverting the space-time slab around its temporal axis leads to a one-to-one degree of freedom correspondence on the new bottom of the slab, allowing for easy evaluation of the jump term. Additionally, this alleviates the requirement of a conforming mesh on the top and bottom of the time-slab introducing new flexibility for the discretization. This ansatz is especially advantageous for problems involving moving domains, where the movement is not known apriori. Our numerical studies validated the approach in comparison to classical space-time and semi-discrete FEM solutions and analytical solutions for different physical problems.

REFERENCES

[1] Tezduyar, T.E. and Behr, M. A new stratgy for finite element computations involving moving boundaries and interfaces - The deforming-spatial-domain/ space-time procedure: I. The concept and the preliminary numerical tests* *Computer Methods in Applied Mechanics and Engineering*, Vol. 94, pp. 339-351, 1992

[2] Hughes, T.J.R. and Hulbert, G.M. Space-time finite element methods for elastodynamics: Formulations and error estimates* *Computer Methods in Applied Mechanics and Engineer-*

ing, Vol. 66, pp. 339-363, 1988

[3] Tezduyar, T.E. and Behr, M. A new stratgy for finite element computations involving moving boundaries and interfaces - The deforming-spatial-domain/ space-time procedure: II. Computation of free-surface flows, two-liquid flows, and flows with drifting cylinders* *Computer Methods in Applied Mechanics and Engineering*, Vol. 94, pp. 353-371, 1992

[4] Elgeti, S. and Sauerland, H. Deforming Fluid Domains Within the Finite Element Method: Five Mesh-Based Tracking Methods in Comparison *Archives of Computational Methods in Engineering*,Vol. 23, pp. 323361, 2016

[5] Zwicke, F. Eusterholz, S. and Elgeti, S. Boundary-Conforming Free-Surface Flow Computations: Interface Tracking for Linear, High-Order and Isogeometric Finite Elements *Computer methods in applied mechanics and engineering*, Vol 326, pp. 175-192, 2017

[6] Lewis, R.W. Nithiarasu,P. and Seetharamu, K.N. *Fundamentals of the Finite Element Method for Heat and Fluid Flow*. John Wiley & Sons, Ltd 2004

[7] Turek,S. and Hron, J. Proposal for Numerical Benchmarking of Fluid-Structure Interaction between an Elastic Object and Laminar Incompressible Flow *Fluid-Structure Interaction. Lecture Notes in Computational Science and Engineering*, Vol. 53 pp. 371-385, 2006

MATHEMATICAL MODELING OF COMPLEX SURFACE PROPERTIES

Inverse solution to the heat transfer coefficient for the oxidized ARMCO steel plate cooling by the air nozzle from high temperature

K. Jasiewicz*, Z. Malinowski† and A. Cebo-Rudnicka†

*Faculty of Metals Engineering and Industrial Computer Science
Department of Heat Engineering and Environment Protection
AGH University of Science and Technology
Cracow, Poland
e-mail: kjasiewi@agh.edu.pl

†Faculty of Metals Engineering and Industrial Computer Science
Department of Heat Engineering and Environment Protection
AGH University of Science and Technology
Cracow, Poland
e-mail: malinows@agh.edu.pl, cebo@agh.edu.pl

Key words: Inverse heat conduction problem, Finite Element Method, Air nozzle, Heat flux, Heat transfer coefficient

Abstract: *The inverse solution to the heat conduction equation for the heat transfer coefficient have been performed to the experimental data obtained during the oxidised Armco steel plate cooling by the air nozzle. A 3D numerical model of the heat transfer during the plate cooling has been considered. Steel products cooled in air from high temperatures are covered with the oxide layer having significantly lower conductivity, and a different surface structure comparing to the non-oxidised metal surface. The Armco steel has been selected as the experimental material because it oxidized in a similar way to carbon steels, but there is no microstructure evolution process in Armco steel below* $900°C$*. It eliminates in the inverse solutions serious problems caused by a latent heat of microstructure evolutions encountered during carbon steel cooling. In the present, study the steel plate has been heated to about* $900°C$ *and cooled by the air nozzle. The plate temperature has been measured by 36 thermocouples.*

1 INTRODUCTION

Air cooling is one of the most common methods of removing excess of heat from an object heated to high temperature. The heat transfer during this process consists of convection and radiation. During cooling with a stream of air, heat is mainly removed from the surface by forced convection. Under natural convection in air, the radiation part of heat transfer in the total heat flux increases. Due to the availability and high costs of obtaining other gaseous coolants, air is used most often. Cooling with a stream of air is used in the metallurgical industry, among others, during hot forging, rolling or heat treatment of metals. Air stream cooling is commonly used to cool turbines components [1]. To achieve the appropriate cooling parameters, and hence the appropriate properties of products, it is necessary to know a rate of heat removal from the surface during the cooling process. Due to the limited possibility of using other methods, especially in high-temperature processes, in order to determine the heat transfer between the coolant and the cooled surface, the inverse problem for the heat conduction equation is usually used [2, 3]. This method relies on temperature measurements at a few points inside the sample, which are then used in numerical calculations. The results of numerical calculations allow to determine the heat transfer on the cooled surface. To be able to perform numerical calculations, it is necessary to develop a heat transfer model. Mathematical and numerical models describing the heat transfer during cooling with an air stream were developed and improved by many scientists. One of the first was Beck [4]. Back presented an inverse method, by means of which, determined changes in the heat flux at the sample

https://doi.org/10.4995/YIC2021.2021.12344

surface from the temperature measurements inside a cooled copper sensor. Malinowski et al. [5] introduced an inverse method to determine three-dimensional heat transfer coefficient and heat flux as functions of time and location. Haw-Long et al. [6] presented an inverse algorithm for the solution of the inverse hyperbolic heat conduction problem. A significant problem during cooling of iron-containing products is the formation of a layer of scale on the surface. Scale formed on the cooled surface changes the heat transfer between the cooling medium and the cooled surface. Determining the thickness and properties of the scale layer formed on the surface, is important to obtain high accuracy numerical model. Li et al. [7] published the results of research concerning the thermal conductivity and diffusivity determination as temperature functions for FeO oxide. An example of a material where the problem of scale formation is significant is Armco steel. This material is used, inter alia, in the production of magnetically active parts of electrical devices in the petrochemical, energy, and shipbuilding industries. Maachou et al. [8] has tested an identification method using Volterra series to model a thermal diffusion in an Armco steel sample. The main purpose of the article is to identify the boundary conditions of heat transfer on the plate surface made of Armco steel during cooling with an air stream. The inverse method was used to determine the boundary conditions. Experimental tests were carried out, consisting of measuring the temperature inside the plate with 36 thermocouples. Temperature measurements taken during cooling were then implemented in a numerical program. The inverse solution for the heat conduction equation allowed to determine the heat transfer coefficient at the plate surfaces cooled by the air stream.

2 GENERAL SPECIFICATIONS

To identify the boundary conditions of heat transfer on the surface of the plate subjected to air cooling, it was necessary to measure the temperature change inside the plate during cooling. These measurements were carried out on an experimental stand, which consisted of three main parts: electric resistant furnace, a cooling chamber, and the temperature acquisition system. The experimental stand was equipped with a control system that allows to operate the feeder arm, furnace door, start cooling, and set furnace temperature (Fig. 1). The first stage of experimental measurements involved heating the plate in an electric furnace. The purpose of the heating was to obtain uniform temperature of about 900°C throughout the entire volume of the plate. After reaching the pre-set temperature, the plate was transported to the cooling chamber, where it was cooled by the MNM type air nozzle. The distance between the nozzle and the plate was 0,15 m. The furnace and the cooling chamber were separated by an automatic door. The plate was mounted vertically in the pneumatic feeder arm which was responsible for its transport between the furnace and the cooling chamber.

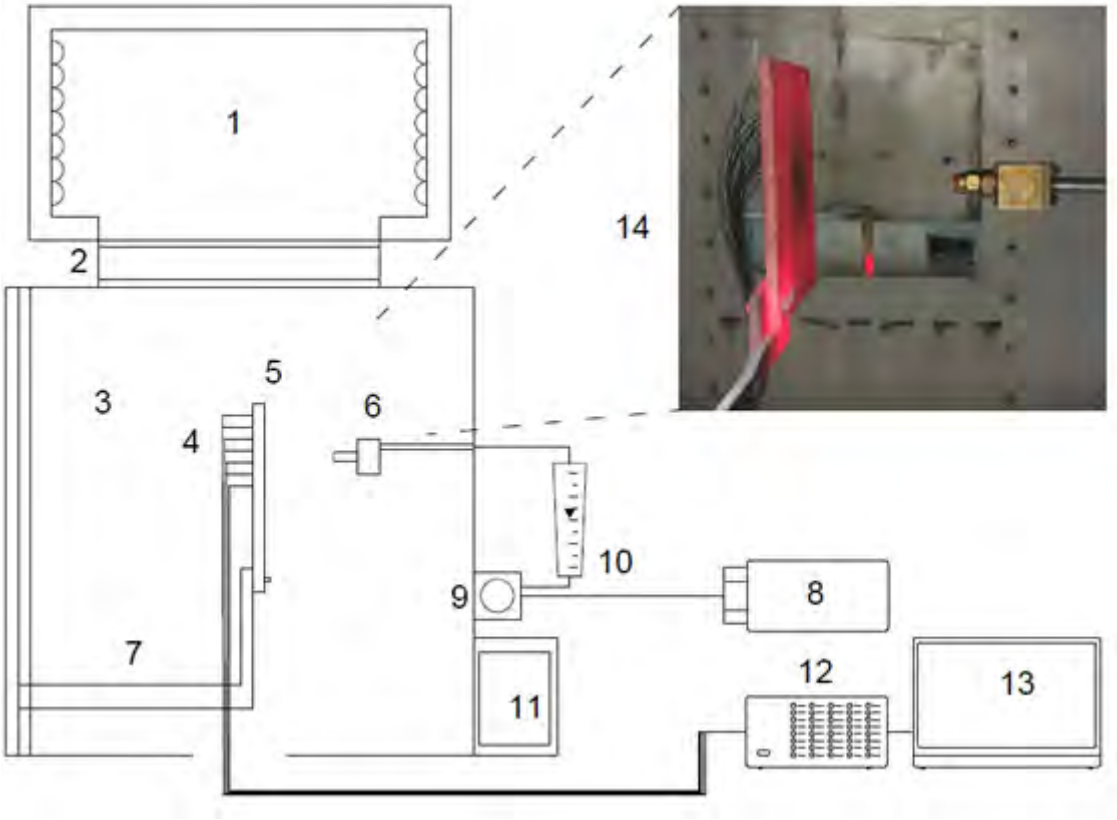


Figure 1: Experimental stand scheme. 1- electric furnace, 2- furnace door, 3- cooling chamber, 4- termocouples, 5- plate, 6- air nozzle, 7- feeder arm, 8- air compressor, 9- air pressure regulator, 10- rotameter, 11- control system, 12-temperature acquisition system, 13- laptop, 14- Armco steel plate cooled by air nozzle.

Experimental studies were carried out on a plate made of Armco steel Fig.1. The plate was $B = 10mm$ thick, $L = 245mm$ in length, and $H = 200mm$ in height. The temperature inside the plate was measured with 36 NiCr - NiAl (K type) thermocouples with a diameter of 1 mm. Thermocouples were numbered from P1 to P36. All thermocouples were placed 2 mm below the cooled surface, in holes 1 mm in diameter. The thermocouples were placed on a quarter of plate with a length of 90 mm. The arrangement of thermocouples is shown in (Fig. 4). The maximum temperature measurement error, related to the accuracy class of the thermocouple was 0.4% of the measured temperature [9]. The maximum temperature of the plate during the tests was $914°C$. It follows that the maximum temperature measurement error resulting from the accuracy class of the thermocouple was $3.66°C$. The temperature measured by the thermocouples was read with a data acquisition system [10]. The accuracy of the device was 0.2%, which means that the maximum reading error was $1.83°C$. These two sources of errors related to the temperature measurements gave the maximum temperature measurement error of about $5.5°C$. Additionally, three thermocouples were used to measure the temperature changes of the cooling chamber wall, and one thermocouple to measure the temperature of air supplied through the nozzle. Parameters of the cooling process has been presented in Table 1. The air flow during the cooling process was recorded by a rotameter.

Material	Air pressure	Distance between nozzle and plate	Cooling time	Average air temperature	Air flow
	[MPa]	[m]	[s]	[$^{\circ}C$]	[l/min]
Armco	0.1	0.15	2000	23.5	27.7

Table 1: Parameters of the cooling process.

3 THE INVERSE PROBLEM FORMULATION

The plate temperature $T(x_1, x_2, x_3, \tau)$ has been calculated from the finite element solution to the heat conduction equation:

$$\frac{\partial}{\partial x_1}\left(\lambda\frac{\partial T}{\partial x_1}\right)+\frac{\partial}{\partial x_2}\left(\lambda\frac{\partial T}{\partial x_2}\right)+\frac{\partial}{\partial x_3}\left(\lambda\frac{\partial T}{\partial x_3}\right)-\rho c\frac{\partial T}{\partial \tau}=0 \tag{1}$$

where:
c – Specific heat $[J/(kg\cdot K)]$,
T – Temperature $[^{\circ}C]$,
x_1, x_2, x_3 – Cartesian coordinates $[m]$,
λ – Thermal conductivity, $[W/(m\cdot K)]$,
ρ – Density $[kg/m3]$,
τ – Time $[s]$.

In the heat conduction model the thermal conductivity, and specific heat dependence on temperature has been considered for material selected for the experiments. The data given in [11] have been approximated with the polynomials (Fig. 2-3).

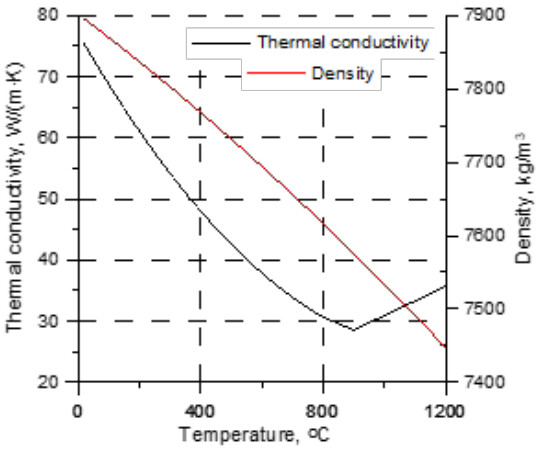


Figure 2: Thermal conductivity and density of Armco steel.

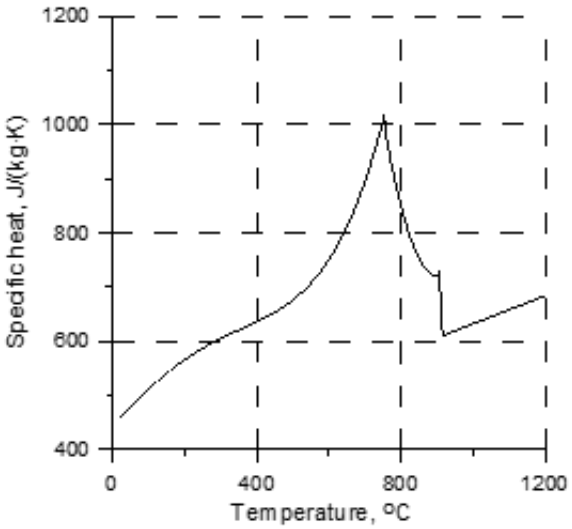


Figure 3: Specific heat of Armco steel.

A FEM algorithm and software developed by Malinowski et al. [5], has been employed to solve Eq. (1). In the FEM solution to Eq. (1) 600 elements with linear shape functions were employed. In the thickness of the plate 6 elements have been employed. The first layer of elements at the surface cooled by the nozzle had the thermophysical properties of iron oxide and a thickness of 0,097 mm. Thickness of the scale layer was determined based on measurements made after the end of the experiment. The properties of the scale implemented in numerical code were taken from Li et al. [7]. Since, the air nozzle was located in the center of the plate, and due to symmetry of the air flow, zero heat fluxes have been assumed at the two symmetry planes:

$$\dot{q}\left(x_1; x_2 = 0; x_3\right) = -\lambda \frac{\partial T}{\partial x_2} = 0 \tag{2}$$

$$\dot{q}\left(x_1; x_2; x_3 = 0\right) = -\lambda \frac{\partial T}{\partial x_3} = 0 \tag{3}$$

At the plate edges, and the vertical surfaces of the plate the boundary conditions have been approximated using the heat flux model:

$$\dot{q}_i = 5.67 \cdot 10^{-8} \frac{\left[T_s(x_1; x_2; x_3)\right]^4 - \left[T_c(\tau)\right]^4}{\frac{1}{\varepsilon_s(T)} + \frac{S_s}{S_c}\left(\frac{1}{\varepsilon_c} - 1\right)} + \propto_i \left[T_s(x_1; x_2; x_3) - T_a\right] \tag{4}$$

where:
$\dot{q}_i$ – Heat flux $[J/(kg \cdot K)]$,
S_c –Cooling chamber surface $[m^2]$,
S_s – Plate surface $[m^2]$,
T_a – Ambient temperature $[^\circ C]$,
T_c – Chamber temperature $[^\circ C]$,
T_s – Cooled surface temperature $[^\circ C]$,
ε_c – Emissivity of the cooling chamber surface,
ε_s – Emissivity of the plate surface,
$\propto_i$ – Heat transfer coefficient $[W/(m^2 \cdot K)]$.

The first term in Eq. (4) describes the radiation heat losses to the chamber walls. The cooling chamber was made of a stainless steel and had the surface $S_c = 4.33m^2$. Comparing the chamber temperature measurements given by the thermal camera with thermocouple's indications, the chamber emissivity $\varepsilon_c = 0.2$ was specified in the boundary condition model. The chamber surface temperature $T_c(\tau)$ was specified based on the thermocouple indications. The sample surface was $S_s = 0.107m^2$. The symbol i denotes a surface number at which a convection heat transfer coefficient α_i was calculated.

At the horizontal edge of the plate cooled from above the heat transfer coefficient (HTC) was calculated from the Nusselt number formula given by Lewandowski et al. [11].

$$Nu = 0.774 Ra^{\frac{1}{5}} \tag{5}$$

where:
Nu – Nusselt number,
Ra - Rayleigh number.

At the vertical edge of the plate, and at the vertical surface cooled under natural convection, the HTC was calculated from formula developed by Churchill and Chu [13].

$$Nu = \left\{ 0.825 + \frac{0.387 Ra^{\frac{1}{6}}}{\left[1 + \left(\frac{0.492}{Pr}\right)^{\frac{9}{16}}\right]^{\frac{8}{27}}} \right\} \tag{6}$$

where:
Pr – Prandtl number.

The boundary condition at the vertical plate surface cooled by the air nozzle has been approximated by the product of functions

$$\propto_{con} (\ \dot{w}, T_s, H_N, p) = \dot{w}\,(x_2, x_3, H_N, p)^{Awp} D(T_s, H_N, p) \tag{7}$$

where:
Awp – Parameter regulating the air flux distribution,
D – Thermal characteristic of air $[J/(kg \cdot K)]$,
H_N – Distance from nozzle to surface $[m]$,
p – Air pressure $[Pa]$,
$\dot{w}$ – Air flux $[kg/(s \cdot m^2)]$,
$\propto_{con}$ – Convection heat transfer coefficient $[W/(m^2 \cdot K)]$.

The function D depends on a local temperature T_s of the plate surface, air pressure p, and the nozzle distance to surface H_N. For a particular air pressure p and the nozzle distance to surface H_N the function D depends only on the plate surface temperature T_s. A scheme of the function D approximation has been shown in (Fig. 5). The plate surface temperature from the plate initial temperature T_0 to the air temperature T_a has been divided into 5 sections. The beginning, and the end of a particular section k is defined by the temperature T_k and T_{k+1}, respectively. The value of the function D at point T_k is defined by D_k parameter. However, for the plate surface temperature equal to the air temperature the convection HTC vanishes and therefore $D1 = 0$. The remaining D_k parameters for $k = 1$ to 5 must be determined from the minimum condition of the objective function (15).

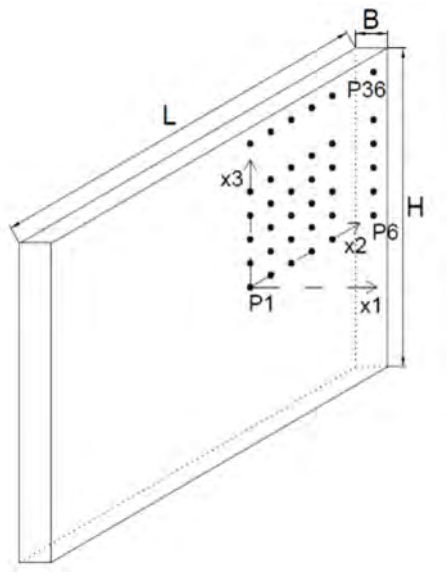


Figure 4: The arrangement of thermocouples

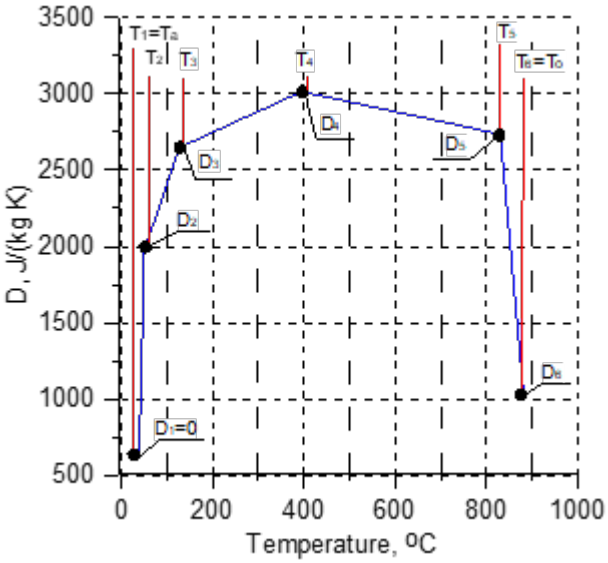


Figure 5: Scheme of thermal characteristic of air as a function of the plate surface temperature.

For surface temperature $T_a < T_s < T_2$ dimensionless temperature η is defined as

$$\eta = \frac{T_s - T_a}{T_2 - T_a} \tag{8}$$

and the function D is calculated from

$$D(\eta) = D_2\, \eta^2 \tag{9}$$

For surface temperature$T_s > T_2$ and $T_k < T_s < T_{k+1}$ dimensionless temperature η is defined as

$$\eta = \frac{T_s - T_k}{T_{k+1} - T_k} \tag{10}$$

and the function D is calculated from

$$D(\eta) = D_k\,(1-\eta) + D_{k+1}\,\eta \tag{11}$$

The function $\dot{w}\,(x_2, x_3, H_n, p)$ describes the rate of air flow over the cooled surface in $kg/(s \cdot m^2)$.

The local air flux has been determined on the basis of measurments that were done for water-air nozzle described in [14]. These measurments allowed to develop the hydraulic characteristic of the MNM nozzle. Distribution of air flux rate has been approximated in cylindrical coordinates using nondimensional distance from the stagnation point r_z:

$$r_z = c_1\, \frac{150}{H_N} \sqrt{x_2^2 + x_3^2} \tag{12}$$

The parameter r_z defines dimensionless radius at which air moving along the conical surface of the spray touches the plate regardless of the nozzle position HN. For the axially symmetrical approximation of the measured air flux two functions given by Eq. (13) and Eq. (14) have

been selected. The function defined by Eq. (13) approximates the air flux distribution from the stagnation point to a position $r_z = 1$:

$$\dot{w}\,(r, H_N, p) = 0,55 \left(\frac{150}{H_N}\right)^2 p^{c_5}\,(1 + c_2\ r_z^{c_3}) \exp\,(-r_z)\ \ for\ \ r_z\ \leq 1 \tag{13}$$

The function defined by Eq. (14) approximates the air flux distribution from a point $r_z = 1$ to infinity:

$$\dot{w}\,(r, H_N, p) = 0.55 \left(\frac{150}{H_N}\right)^2 p^{c_5}\,(1 + c_2\ r_z^{c_4})\, exp\,(-r_z^{c_4})\ \ for\ r_z \geq 1 \tag{14}$$

Nozzle Angle	Coefficient				
	c_1	c_2	c_3	c_4	c_5
45°	0.0512	0.1870	0.0000	1.0000	0.4800

Table 2: Coefficients employed in Eq. (12), Eq. (13) and Eq. (14) for a MNM nozzle angle of 45°.

It is important to notice that at point $r_z = 1$ the air flux calculated from Eq. (13), or Eq. (14) has the same value. The air flow rate calculated as the integral of Eq. (13) and Eq. (14) over the range from $r = 0$ to $r = 150mm$ was in a good agreement with the rotameter indication given in Table 1.

The parameters D_i and the Awp parameter defining the D function distribution have been obtained by minimizing the objective function:

$$E(D_i, Awp) = \frac{1}{NT \cdot NP} \sum_{m=1}^{NT} \sum_{n=1}^{NP} \left[\frac{1}{\sqrt{1 + \left(\frac{\Delta Te^{nm}}{\Delta \tau}\right)^2}} (Te^{nm} - T(D_i, Awp)^{nm}) \right]^2 \tag{15}$$

where:

Te^{nm} – Sample temperature measured by the sensor n at the time τ_m,

T^{nm} – Computed sample temperature at the location of the sensor n at the time τ_m,

NP – Number of temperature sensors,

NT – Number of temperature measurements performed by one sensor.

The objective function (15) defines the temperature difference between measured and computed temperatures along the normal to the measured temperature curve.

The radiation heat losses depended on the plate emissivity $\varepsilon_s(T)$ has been calculated from the emissivity model developed based on inverse solution to the Armco plate cooling under natural convection in air:

$$\varepsilon_s = 0.5 + 0.35\ \bar{t}_p^2 \tag{16}$$

4 IDENTIFICATION OF THE HEAT TRANSFER COEFFICIENT

The performed numerical calculations made it possible to determine the local convection HTC on the cooled surface of the Armco plate covered with scale. In Fig. 6 and Fig. 7, local HTC distributions in relation to the surface temperature excess, and time, respectively, calculated at 6 elements E1-E6 have been presented. The centers of the elements, in which the local HTC have been presented, corresponded to the positions of P1, P2, ..., P6 thermocouples inserted along x_2 axis (Fig. 4). In Fig. 8 the thermal characteristics of air versus the temperature excess has been presented.

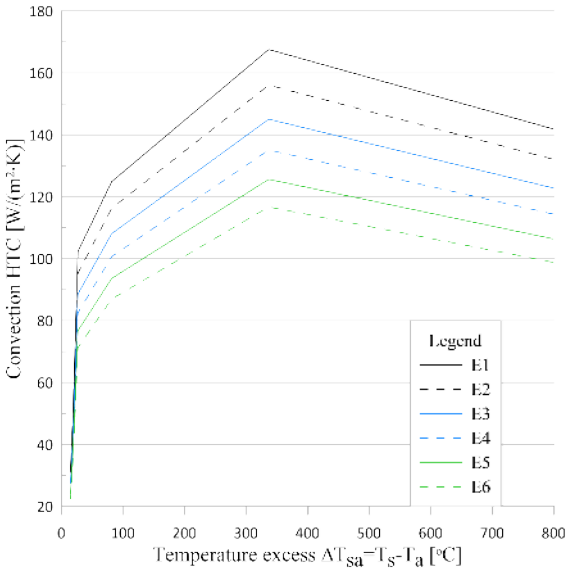


Figure 6: The local convection HTC distributins versus the temperature excess.

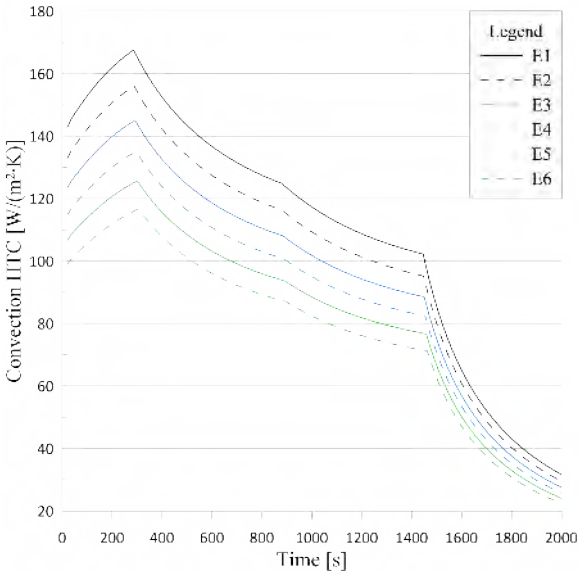


Figure 7: The local convection HTC distributions versus the time of cooling.

From the beginning of the air nozzle cooling, the convection HTC increase during about 300 s and reaches the maximum value depending on the location along x_2 axis (Fig. 7). The lowest value of a maximum convection HTC of $115W/(m^2 \cdot K)$ was reached at element E6, and a highest of $165W/(m^2 \cdot K)$ at element E1. It is related to the air velocity and mass flux. Near the center of the nozzle axis, the air velocity as well as the air mass flux are the highest. It increases the convection heat transfer process. As the distance from the nozzle axis increases, the air mass flux decreases. During the first 300 s of cooling process carried out with the air nozzle, the plate temperature decreases to about $350°C$ (Fig. 6). During that time the HTC has reached the maximum values, Next, the convection HTC decreases gradually, and after subsequent 1100 s has reached about $130W/(m^2 \cdot K)$ in E1, and $90W/(m^2 \cdot K)$ in E6 (Fig. 7). During this time, the plate temperature decreases slowly from $350°C$ to about $120°C$ (Fig. 6). In the last stage of cooling which lasted of about 600 s, a rapid drop in convection HTC values was observed.

In Fig. 8 the thermal characteristics of the air has been shown. This characteristic is presented as a function that determines the ability of a coolant to remove heat from the cooled surface. Such presentation of the results allows to eliminate the influence of the amount of air supplied on the efficiency of the cooling process. As shown in Fig. 8, the greatest ability of heat extracting from the cooled surface was obtained in element E1, which was located in the axis of the air stream. As the distance from the nozzle axis increases, the ability of heat removing decreases (Fig. 8). Such a behavior is related to the distribution of air velocity and the air mas flux distribution over the cooled surface.

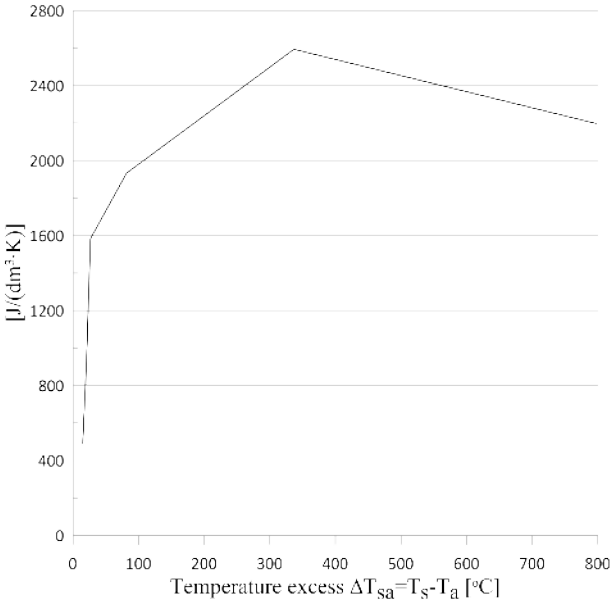


Figure 8: Function defining thermal characteristic of air versus the surface temperature excess.

The accuracy of the performed inverse solutions to the convection HTC identification has been presented in Table 3. The average temperature difference between measured and calculated temperature did not exceed 2K. This indicates a satisfactory accuracy of the numerical calculations.

Average deviation	Maximum negative deviation	Maximum positive deviation
[K]	[K]	[K]
1.667	-5.95	7.88

Table 3: Inverse solution accuracy

5 CONCLUSION

The inverse solution to the heat conduction equation for the heat transfer coefficient determination during the vertical plate cooling by the air nozzle has been obtained. Three-dimensional heat conduction problem has been solved using the finite element method. The thermophysical properties of the Armco steel have been considered as functions of temperature. The oxide layer on the cooled surface has been considered in the heat conduction model as well. The thickness of the oxide layer of about 0.1 mm has been determined based on the oxidation process kinetics. The boundary condition at the oxide layer has been defined. The boundary condition model has been specified as a product of two functions. The first function defined the air flux specific for a particular nozzle. The second function defined the air ability to extract heat from the cooled surface. The parameters of the heat conduction model were determined from the minimum condition of the object function. It has been found that the convection heat transfer coefficient increases rapidly as the plate temperature grows. However, the convection heat transfer coefficient has reached a maximum value at the plate temperature of $350°C$. For the plate temperature range from $350°C$ to $800°C$ a linear decrease in the HTC has been obtained. Air flux distribution over the cooled surface is a particularly important in the developed boundary condition model.

FOUNDING

Scientific study financed from the regular activity of the Faculty of Metals Engineering and Industrial Computer Science of AGH University of Science and Technology.

REFERENCES

[1] Yang L., Ren J., Jiang H., Ligrani P., Experimental and numerical investigation of unsteady impingement cooling within a blade leading edge passage, International Journal of Heat and Mass Transfer, Vol. 71, 2014, pp. 55-68

[2] Dou R., Wen Z., Zhou G., 2D axisymmetric transient inverse heat conduction analysis of air jet impinging on stainless steel plate with finite thickness, Applied Thermal Engineering, Vol 93, (2016), pp. 468-475

[3] Guo Q., Wen Z., Dou R., Experimental and numerical study on the transient heat-transfer characteristics of circular air-jet impingement on a flat plate, International Journal of Heat and Mass Transfer, Vol. 104, 2017, pp. 1177-1188

[4] Beck J.V, Nonlinear estimation applied to the nonlinear inverse heat conduction problem, International Journal of Heat and Mass Transfer 13, 1970, pp. 703-716

[5] Malinowski Z., Telejko T., Hadała B., Cebo-Rudnicka., Szajding A., Dedicated three dimensional numerical models for the inverse determination of the heat flux and heat transfer coefficient distributions over the metal plate surface cooled by water, International Journal of Heat and Mass Transfer, (2014), Vol. 75, pp. 347-361

[6] Lee H-L., Chang W-J., Wu S-Ch., Yang Y-Ch., An inverse problem in estimating the base heat flux of an annular fin based on the hyperbolic model of heat conduction, International Communication in heat and Mass Transfer, Vol 44, (2013), pp. 31-37

[7] Li M., Endo R., Akoshima M., Susa M., Temperature Dependence of Thermal Diffusivity and Conductivity of FeO Scale Produced on Iron by Thermal Oxidation, ISIJ International, Vol. 57, (2017), No. 12, pp. 2097-2106

[8] Maachou A., Malti R., Melchior P., Battaglia J-L., Oustaloup A., Hay B., Application of fractional Volterra series for the identification of thermal diffusion in an ARMCO iron sample subject to large temperature variations, IFAC Proceedings Volumes, Vol. 44, Issue 1, (2011), pp. 5621-5626

[9] Polska Norma PN-EN 60584-2: 1997 Termoelementy. Tolerancje

[10] Data acquisition system MGCplus datasheet, Hottinger Baldwin Messtechnik,

[11] Shanks H. R., Klein A., H., Danielson G. C., Thermal Properties of Armco Iron, Journal of Applied Physics, Vol. 38, (1967), pp. 2885-2892

[12] Lewandowski W. M., Radziemska E., Buzuk M., Bieszk H., Free Convection heat transfer and fluid flow above horizontal rectangular plates, Applied Energy, Vol. 15, (1972), pp. 2535-2549

[13] Churchill S.W., Chu H. H. S., Correlating equations for laminar and turbulent free convection from a vertical plate, Int. Journal of Heat and Mass Transfer, Vol. 18, (1975), pp. 1323–1329

[14] Cebo-Rudnicka A., Malinowski Z., Identification of heat flux and heat transfer coefficient during water spray cooling of horizontal copper plate, Int. Journal of Thermal Sciences, Vol.145 (2019), pp. 1-24

NOVEL COMPUTATIONAL METHODS FOR DESIGN, MODELLING, AND HOMOGENIZATION OF METAMATERIALS AND FUNCTIONAL SMART MATERIALS

Design and modelling of bioinspired 3D printed structures

C. Garrido[1] , E. Alabort[2] and D. Barba[1]

[1] Departamento de Materiales y Producción Aeroespacial
E.T.S de ingeniería Aeronáutica y del Espacio
Universidad Politécnica de Madrid
Madrid, Spain
e-mail: conrado.garrido@upm.es
daniel.barba@upm.es

[2] Alloyed Ltd
Oxford Industrial Park
Yarnton, United Kingdom
e-mail: enrique.alabort@alloyed.com

Key words: Lattice structure, bio-inspired materials, finite element methods models, metamaterials, titanium

Abstract: *Metamaterials are those that, through human engineering, have unusual properties that cannot be commonly found in nature. Recently, these metamaterials have been gaining importance due to the introduction of additive manufacturing technologies. Specifically, metamaterials known as lattice structures have advantages over bulk solid materials, such as increased strength and specific stiffness. However, in order to exploit these advantages of these exotic materials, we need robust and accurate tools to tailor and design their properties. The objective of this work is to present a complete systematic study of the different approaches for metamaterial computational design evaluating their advantages and drawbacks in terms of computational efficiency and accuracy in predicting the metamaterial mechanical properties.*

1 INTRODUCTION

Metamaterials are human designed materials which can acquire unforeseen properties not seen in nature [1]. These can be produced by different manufacturing methods [2]. There are different types of metamaterials, e.g., electromagnetic metamaterials modulating electromagnetic waves [3] or mechanical metamaterials with extraordinary specific mechanical behavior [4]. One type of mechanical metamaterials are lattice structures. Lattices have interesting properties like weight reduction compared to the solid structure, preserving other beneficial properties, for example the strength or biocompatibility [5]. Lattice structures are expected to revolutionize different fields [6],[7]. In the biomedical industry, these structures are selected as the perfect candidate for a new generation of biocompatible implants [6]. In the aerospace industry, the light-weighting potential of these structures can replace solid component with similar properties but higher weights [7]. In order to design these lattice structures, computational simulations are critical. Finite element modelling (FEM) stands as the preferred route to simulate lattice structures. The unresolved problem of using FEA for lattice design resides in the numerous variables that affect the results which can lead to a wrong design-optimization exercise. Among them are the type of mesh, the number of elements or heterogeneities in the material properties heritage from the additive manufacturing process. Furthermore, there is not a systematic study that shows the optimal form of modelling metamaterials [8].

To fill this gap, this work presents a systematic study of the effect of these variables on the mechanical simulations of latticed metamaterials. Three different methodologies are used: 3D explicit meshing, homogeneous beam models and heterogeneous beam models. The computational results are validated against experimental results of a lattice structure explicitly designed

https://doi.org/10.4995/YIC2021.2021.12529

and manufactured for this study.

2 METHODOLOGY OF THE STUDY

2.1 Lattice geometry

For this project, a typical lattice structure has been designed with a strut radius of 1.3 mm and target solid fraction of 24% following a Voronoi distribution, see Fig. 1. A summary of lattice structure features is presented in Table 1. The design has been additively manufactured (AM) by selective laser method in Renishaw AM250 using Ti6Al4V as base material. These are presented in Fig. 2.

Figure 1: Lattice structure used in this study.

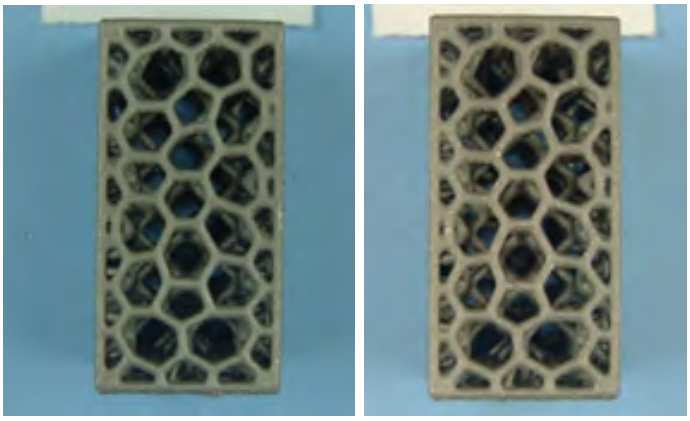


Figure 2: Additively manufactured Ti6Al4V lattice structures.

Structure	Strut radius	Solid fraction
Lattice	1.3 mm	23.78%

Table 1: Features of the lattice structure.

2.2 Mechanical Testing

The additively manufactured lattice samples have been subjected to compression testing using a servo-hydraulic MTS (model 810) universal testing machine equipped with a 100kN load cell. A strain rate of 10^{-3} s^{-1} was imposed during the test. Strain was monitored using digital image correlation (DIC) during the test. The DIC system have a camera (model BFS-U3-13Y3C-C) and also have a in-house code developed to group stress and strain points in real time. Two repeats were performed to address the consistency of the experimental results.

2.3 Computational study

The mechanical behaviour of the lattice structure has been addressed computationally. Abaqus (2018) FEM static analysis has been used for this purpose [9]. The lattice structure has been meshed with 2 different element types: (1) C3D10 volumetric quadratic elements and (2) quadratic B32 beam elements. Table 2 shows the number of elements for each type of mesh. Regarding the boundary conditions, the displacement of the nodes located at the lower face of the sample are restricted in all directions (U1 = U2 = U3 = 0). For the nodes at the upper face, a displacement equivalent to 5% of the total sample deformation is imposed. In terms of material model, material properties for the FE model are extracted from the experimental mechanical behaviour of AMTi6Al4V. A wire of AMTi6Al4V with the same thickness of the lattice struts (1.3 mm) was tested in a tension test and the experimental stress-strain curve was used as an Abaqus material database for the lattice structures [10].

Type of Mesh	Number of elements
Volumetric Mesh (C3D10 Abaqus Code)	462551
Beam Mesh (B32 Abaqus Code)	1051

Table 2: Number of elements for volumetric and beam meshes.

A typical feature of these lattice structures is the rounding of the struts at the lattice nodes to avoid stress concentrations promoting premature failures of the lattice, see Fig. 3. Due to this, the radius is not homogeneous along the axis of the struts. The strut radius is higher when approaching to nodes than in the centre of the lattice beams.

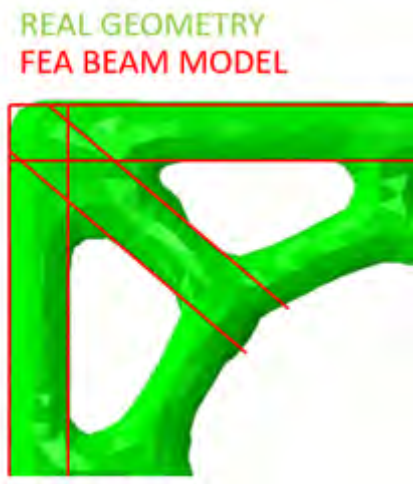


Figure 3: Representation of the curvature produced near the nodes.

The geometry at the nodes of the volumetric models is fully heritage from the explicit lattice structure model in the form of "*stl*" surface. However, for the beam models, the beam radius needs to be estimated in order to obtain a homogenised representative geometry of each strut in the lattice. In this work, two different approaches to target this problem have been proposed (see Fig. 4): (1) model A, assuming the lattice as a continuous beam network with a total volume equal to the experimental one or (2) model B, integrating the effect of the nodes idealising the lattice structure as a combination of beams (struts) and spheres (nodes). The details for each model are explained next. Beam model A: The first model assumes a network of beams with an idealised circular section of area πR_A^2, see Fig, 4. The total volume V_A of the beam network is calculated as:

$$V_A = \sum_i^N \pi \ R_A^2 \ L_A^i \tag{1}$$

Where L_A^i is the node-to-node length of the beam i and N is the total number of beams. By equaling this V_A to the real volume of the AM geometry V_{AM}, the R_A can be extracted as

$$R_A = \sqrt{\frac{V_{AM}}{\pi \ L_T}} \tag{2}$$

Where L_T is the total length of all beams of the structure. This radius R_A is used to define the section of the beams in model A. The calculated value is presented in Table 3. Beam model B idealises each node as a sphere of radius R_B and each beam as a cylinder of the same radius R_B, see Fig. 4. The total volume of the lattice VB can be calculated adding up the individual volume of all the nodes and beams in the lattice as:

$$R_B = \frac{4}{3} N_n \pi R_B^3 + \sum_i^N \left(L_B^i - 2R_B\right) \pi R_B^2 \tag{3}$$

where LB is the node-to-node length of the strut i and Nn is the total number of nodes in the lattice structure. By equaling this V_B to the real volume of the AM geometry V_{AM}, the R_B can be extracted solving the function:

$$R_B = \frac{4}{3} N_n \pi R_B^3 + L_T \pi R_B^2 - 2 R_B^3 N_B \tag{4}$$

Where N_B is the total number of beams in the structure. This radius R_B is used to define the section of the beams in model B. The calculated value is presented in Table 3.

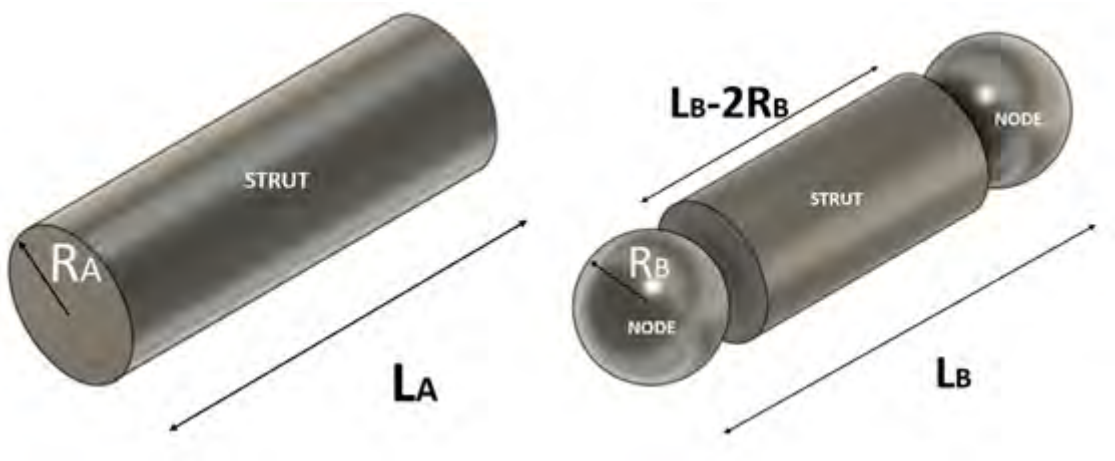


Figure 4: Beam model A (left) and beam model B (right).

Type of Models	Radius (mm)
Experimental at the centre of the struts	0.65
Model A (RA)	0.63
Model B (RB)	0.712

Table 3: Radius for each of the beam models used.

3 RESULTS AND DISCUSSION

In this section, the computational mechanical behaviour of the lattice structure is compared against the experiments. Next, the accuracy of each of the different modelling approaches is addressed. Finally, the advantage and disadvantages of each method are discussed.

3.1 Experimental behaviour

Experimental and FEM stress-strain curves of the lattice structure are presented in Fig. 5. The two repeats of the experimental tests present a good repeatability with less than 10% discrepancy between both curves. The lattice behaviour presents an initial elastic region with an elastic modulus proportional to the solid fraction of the lattice. After yielding, there is an initial hardening region followed by a plateau before failure (not studied in this work). The three different FEM approaches (Volumetric, Beam Model A and Beam Model B) present a similar qualitative behaviour. However, quantitively, the three models differ, with the volumetric model presenting the closest behaviour to the one experimentally observed.

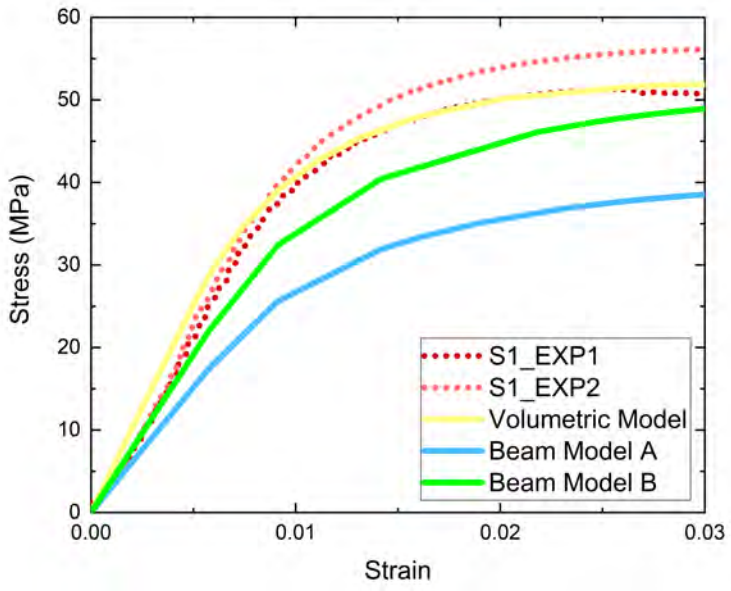


Figure 5: Experimental and FEM stress-strain curves. Red curves represent experimental lattice compression test, yellow curve represents the volumetric model and blue and green curves represent the beams models A and B, respectively.

The apparent elastic modulus and yield stress has been extracted from the stress-strain curves; they are shown in Fig. 6. The apparent elastic modulus of the volumetric model is higher than the experimental ones (∼10% higher) while the beam models present a substantially lower elastic modulus than the experiments (∼30% lower for Model A and ∼20% lower for Model B). Regarding the higher rigidity of the volumetry model when compared to the experiments, it is known that AM lattice present defects, especially for self-supported lattices like the ones in this work [11]. These defects can reduce rigidity of the lattice and might partially explain the small increase in the elastic modulus, which are not taking into account in the models. Another reason might be small deviations in the printed geometries from the ideal simulated ones [11]. In terms of the yield stress, all the models present lower values than the experimental ones. The volumetric and beam model B present the closest values to the experiments (<10% error) while the beam model A differs considerable (>20% lower yield strength). As a summary, the volumetric model shows superior accuracy when comparing with the experimental apparent elastic modulus and yield stress. On the other hand, beam model A presents the worst approximation for both, the apparent elastic modulus and the yield stress.

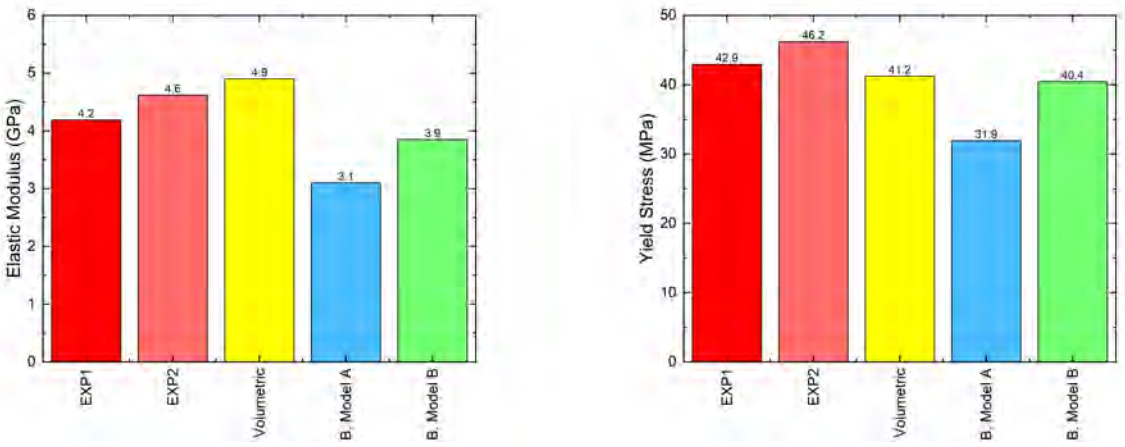


Figure 6: Experimental and FEM elastic modulus (left) and yield stress (right).

The mechanical behaviour of lattice structures is strongly influence by their solid fraction [12]. Small deviations of the solid fraction can produce significant variations in the mechanical behaviour of the lattice. Therefore, it is important to address this aspect between the experiments and the simulations. Solid fractions of experimental and FEM geometries are compared in Fig. 7. There are small deviations between the computational geometry (ideal design geometry) and the experimental one, arising from the imperfections in the additive manufacturing process [13]. Beam model A has the same solid fraction than the volumetric model because the radius is obtained equalling AM structure volume and beam model volume. On the other hand, beam model B has the bigger solid fraction due to the overlap of the beams produced at the nodes.

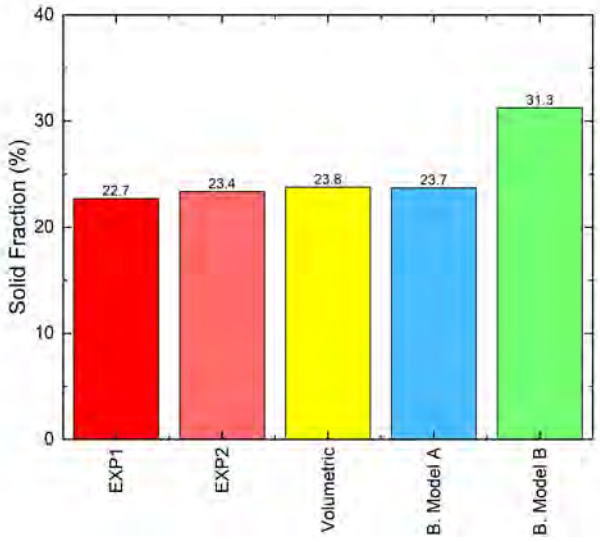


Figure 7: Solid fraction estimated for the samples tested and the numerical models, respectively.

3.2 Failure modes

In this section, focus is put on the failure mechanisms in the lattice structure. Fig. 8 shows the failure modes of each model compared to the experimental specimen. Volumetric failure modes correlate well with the ones observed experimentally. Both beam models (A and B) present the same failure modes, suggesting that the change in beam diameter between A and B did not affect the failure mechanism.

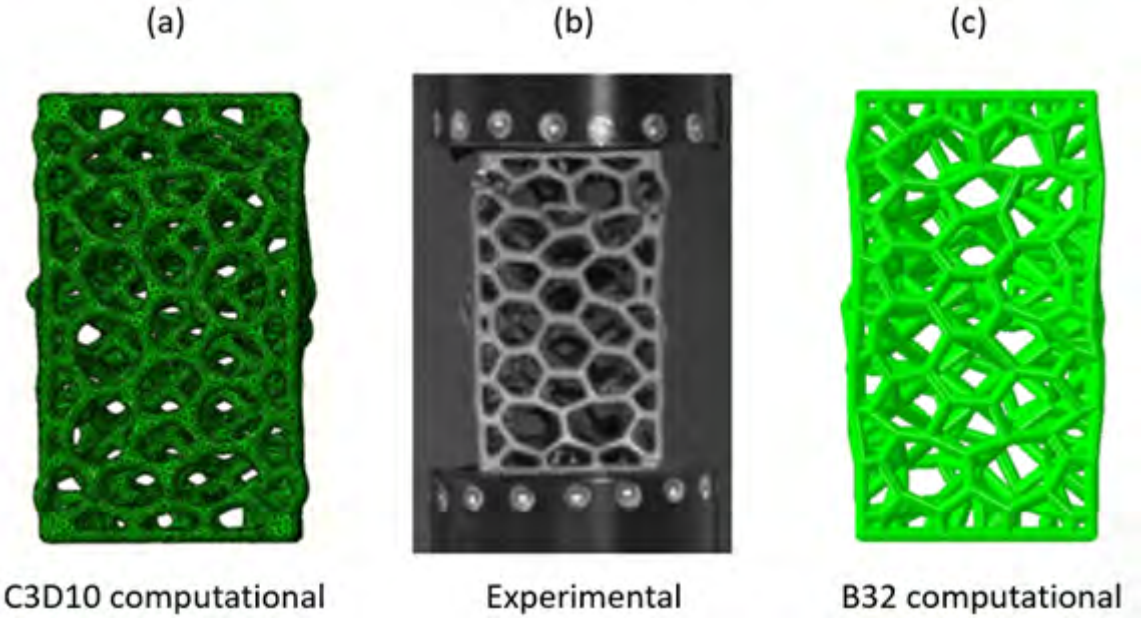


Figure 8: Comparison between failure modes in the volumetric model (a), experimental specimen (b) and the beam models (c).

3.3 Computational efficiency and discussion

Computational cost is a critical aspect in the design of lattice structures. In this regard, the computational time of each FEM approach is represented in Fig. 9. Beam models have considerably lower computational cost than the volumetric model. This supports the necessity of developing new beam theories adapted to AM lattice design capturing the peculiarities and defects in these structures.

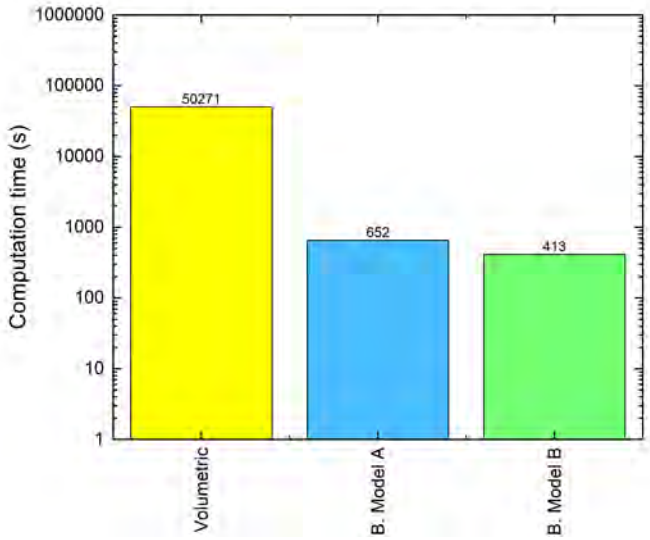


Figure 9: Computation time of FEM models.

4 CONCLUSIONS

In this work, the advantages and drawbacks of different FEM approaches in modelling and design AM metamaterial structures have been studied. The following conclusions can be drawn:

- Volumetric FE models with elasto-plastic material equations present the best accuracy when compared to experimental results. The minor difference in the mechanical response between experimental specimens and FE volumetric models are due to the defects present in the experimental specimens that have not been considered in the simulations. These models capture the plastic failure mechanics with great accuracy. However, these models present higher computational costs than beam models.

- Two criteria to establish FE radius in beam models have been presented: one based on the total experimental volume of the lattice (model A) and another simulating the lattice structure as combination of struts and spheres (model B). The beam model B presents a higher accuracy than model A. The beam model B has an error less than 10% compared to experiments in elastic modulus and yield stress.

- Beam models are computationally more efficient than volumetric models. However, the precision of these models is lower, and they do not correctly maintain the physics of the deformations. There is a need to develop new beam models that capture the same physics than volumetric models but with a lower computational cost.

REFERENCES

[1] T. J. Cui, D. R. Smith, and R. Liu, *Metamaterials*. Springer, 2010.

[2] Y. Ding, M. Akbari, X.-L. Gao, L. Ai, and R. Kovacevic, "Use of powder-feed metal additive manufacturing system for fabricating metallic metamaterials," in *Manufacturing techniques for materials: engineering and engineered*, pp. 51–65, CRC Press, 2018.

[3] R. M. Walser, "Electromagnetic metamaterials," in *Complex Mediums II: beyond linear isotropic dielectrics*, vol. 4467, pp. 1–15, International Society for Optics and Photonics, 2001.

[4] X. Zheng, H. Lee, T. H. Weisgraber, M. Shusteff, J. DeOtte, E. B. Duoss, J. D. Kuntz, M. M. Biener, Q. Ge, J. A. Jackson, *et al.*, "Ultralight, ultrastiff mechanical metamaterials," *Science*, vol. 344, no. 6190, pp. 1373–1377, 2014.

[5] X. Yan, Q. Li, S. Yin, Z. Chen, R. Jenkins, C. Chen, J. Wang, W. Ma, R. Bolot, R. Lupoi, *et al.*, "Mechanical and in vitro study of an isotropic ti6al4v lattice structure fabricated using selective laser melting," *Journal of Alloys and Compounds*, vol. 782, pp. 209–223, 2019.

[6] D. Mahmoud and M. A. Elbestawi, "Lattice structures and functionally graded materials applications in additive manufacturing of orthopedic implants: a review," *Journal of Manufacturing and Materials Processing*, vol. 1, no. 2, p. 13, 2017.

[7] G. Totaro and Z. Gürdal, "Optimal design of composite lattice shell structures for aerospace applications," *Aerospace Science and Technology*, vol. 13, no. 4-5, pp. 157–164, 2009.

[8] Z. Li, C. Wang, and X. Wang, "Modelling of elastic metamaterials with negative mass and modulus based on translational resonance," *International Journal of Solids and Structures*, vol. 162, pp. 271–284, 2019.

[9] E. Giner, N. Sukumar, J. Tarancón, and F. Fuenmayor, "An abaqus implementation of the extended finite element method," *Engineering fracture mechanics*, vol. 76, no. 3, pp. 347–368, 2009.

[10] G. F. de Vera Conrado and B. C. Daniel, "Smart modelling of additively manufactured metamaterials," in *2020 IEEE 10th International Conference Nanomaterials: Applications & Properties (NAP)*, pp. 02SAMA18–1, IEEE, 2020.

[11] D. Barba, C. Alabort, R. C. Reed, and E. Alabort, "On the size effects in additively manufactured titanium and the implications in am components," in *TMS 2020 149th Annual Meeting & Exhibition Supplemental Proceedings*, pp. 449–456, Springer, 2020.

[12] C. Chu, G. Graf, and D. W. Rosen, "Design for additive manufacturing of cellular structures," *Computer-Aided Design and Applications*, vol. 5, no. 5, pp. 686–696, 2008.

[13] B. Wu, Z. Pan, D. Ding, D. Cuiuri, H. Li, J. Xu, and J. Norrish, "A review of the wire arc additive manufacturing of metals: properties, defects and quality improvement," *Journal of Manufacturing Processes*, vol. 35, pp. 127–139, 2018.

NOVEL NUMERICAL METHODS FOR FLUID-STRUCTURE INTERACTION PROBLEMS

Adjoint-based methods for optimization and goal-oriented error control applied to fluid-structure interaction: implementation of a partition-of-unity dual-weighted residual estimator for stationary forward FSI problems in deal.II

T. Wick

Leibniz University Hannover
Institute of Applied Mathematics
Hannover, Germany
e-mail: thomas.wick@ifam.uni-hannover.de

Cluster of Excellence PhoenixD
(Photonics, Optics, and Engineering - Innovation Across Disciplines)
Leibniz University Hannover, Germany

Key words: goal-oriented error control, dual-weighted residuals, adjoint, mesh adaptivity, fluid-structure interaction, deal.II

Abstract: *In this work, we implement goal-oriented error control and spatial mesh adaptivity for stationary fluid-structure interaction (FSI). The a posteriori error estimator is accomplished using the dual-weighted residual method in which the adjoint equation arises. The fluid-structure interaction problem is formulated within a variational-monolithic framework using arbitrary Lagrangian-Eulerian coordinates. The overall problem is nonlinear and solved with Newton's method. We specifically consider the FSI-1 benchmark problem in which quantities of interest include the elastic beam displacements, drag, and lift. The implementation is based on the deal.II finite element library and provided open-source published on github* `https://github.com/tommeswick/goal-oriented-fsi`. *Possible extensions are discussed in the source code and in the conclusions of this paper.*

1 INTRODUCTION

Fluid-structure interaction (FSI) is well-known [11, 25, 27, 10, 5, 8, 46, 26, 59] and a prime example of a multiphysics problem. It combines several challenges such as different types of partial-differential equations (PDE), interface-coupling, nonlinearities in the equations and due to coupling, Lagrangian and Eulerian coordinates. These result into typical numerical challenges such as robust spatial discretization (in particular for the moving interface), robust time-stepping schemes, efficient and robust linear and nonlinear solution algorithms. Computational works include different coupling concepts [37, 30, 53, 42, 16], space-time multiscale [51], reduced order modeling [24, 40, 52, 33], optimal control, parameter estimation, uncertainty quantification [41, 7, 48, 39, 21, 62], and efficient solver developments [34, 4, 28, 43, 13, 45, 15, 38, 55].

In this work, the main objective is the application and open-source implementation of goal-oriented a posteriori error control using the dual-weighted residual (DWR) method [6, 3]. For applications in fluid-structure interaction, we refer to [32, 23, 54, 57, 44, 46, 20, 22]. A recent overview of our own work using the adjoint FSI equation in goal-oriented error estimation and optimization was done in [60]. In [49] a variational localization using a partition-of-unity (PU) was proposed, facilitating the application to coupled problems such as fluid-structure interaction. In view of increasing initiatives of open-source developments, another purpose of this work is to provide a documented open-source code. To this end, a stationary fluid-structure interaction problem is considered in order to explain the main steps of a PU-DWR estimator. The problem is formulated within a monolithic framework using arbitrary Lagrangian-Eulerian (ALE) coordinates. For some well-posedness results of such stationary FSI problems, we refer

https://doi.org/10.4995/YIC2021.2021.12332

to [31, 61]. Together with the goal functional under consideration, the FSI formulation serves as PDE constraint and the Lagrange formalism can be applied. Specifically, the monolithic formulation yields a consistent adjoint equation.

For the monolithic, stationary, FSI formulation we follow [47, 56] and for the PU-DWR error estimator, we follow [49]. The basis of our programming code is [58] (see also updates on github[1]) and we take some ideas from deal.II [1, 2] step-14[2]. Our resulting code can be found on github[3].

2 VARIATIONAL-MONOLITHIC ALE FLUID-STRUCTURE INTERACTION

2.1 Modeling

For the function spaces in the (fixed) reference domains $\widehat{\Omega}, \widehat{\Omega}_f, \widehat{\Omega}_s$, we define $\widehat{V} := H^1(\widehat{\Omega})^d$. In the fluid and solid domains, we define further:

$$\begin{aligned}
\widehat{L}_f &:= L^2(\widehat{\Omega}_f), \quad \widehat{L}_f^0 := L^2(\widehat{\Omega}_f)/\mathbb{R}, \quad \widehat{V}_f^0 := \{\widehat{v}_f \in H^1(\widehat{\Omega}_f)^d : \widehat{v}_f = 0 \text{ on } \widehat{\Gamma}_{\text{in}} \cup \widehat{\Gamma}_D\},\\
\widehat{V}_{f,\widehat{u}}^0 &:= \{\widehat{u}_f \in H^1(\widehat{\Omega}_f)^d : \widehat{u}_f = \widehat{u}_s \text{ on } \widehat{\Gamma}_i, \quad \widehat{u}_f = 0 \text{ on } \widehat{\Gamma}_{\text{in}} \cup \widehat{\Gamma}_D \cup \widehat{\Gamma}_{\text{out}}\},\\
\widehat{V}_{f,\widehat{u},\widehat{\Gamma}_i}^0 &:= \{\widehat{\psi}_f \in H^1(\widehat{\Omega}_f)^d : \widehat{\psi}_f = 0 \text{ on } \widehat{\Gamma}_i \cup \widehat{\Gamma}_{\text{in}} \cup \widehat{\Gamma}_D \cup \widehat{\Gamma}_{\text{out}}\},\\
\widehat{V}_s^0 &:= \{\widehat{u}_s \in H^1(\widehat{\Omega}_s)^d : \widehat{u}_s = 0 \text{ on } \widehat{\Gamma}_D\}.
\end{aligned}$$

As stationary FSI problem in variational-monolithic ALE form, we have [56][p. 29]:

Problem 2.1. *Find* $\{\widehat{v}_f, \widehat{u}_f, \widehat{u}_s, \widehat{p}_f\} \in \{\widehat{v}_f^D + \widehat{V}_{f,\widehat{v}}^0\} \times \{\widehat{u}_f^D + \widehat{V}_{f,\widehat{u}}^0\} \times \{\widehat{u}_s^D + \widehat{V}_s^0\} \times \widehat{L}_f^0$*, such that*

$$\begin{aligned}
&(\widehat{\rho}_f \widehat{J}(\widehat{F}^{-1}\widehat{v}_f \cdot \widehat{\nabla})\widehat{v}_f), \widehat{\psi}^v)_{\widehat{\Omega}_f}\\
+(\widehat{J}\widehat{\sigma}_f\widehat{F}^{-T}, \widehat{\nabla}\widehat{\psi}^v)_{\widehat{\Omega}_f} &- \langle \widehat{g}_f, \widehat{\psi}^v\rangle_{\widehat{\Gamma}_N} - (\widehat{\rho}_f\widehat{J}\widehat{f}_f, \widehat{\psi}^v)_{\widehat{\Omega}_f} = 0 \quad \forall \widehat{\psi}^v \in \widehat{V}_{f,\widehat{v}}^0,\\
(\widehat{F}\widehat{\Sigma}, \widehat{\nabla}\widehat{\psi}^v)_{\widehat{\Omega}_s} &- (\widehat{\rho}_s\widehat{f}_s, \widehat{\psi}^v)_{\widehat{\Omega}_s} = 0 \quad \forall \widehat{\psi}^v \in \widehat{V}_s^0,\\
(\widehat{\sigma}_{mesh}, \widehat{\nabla}\widehat{\psi}^u)_{\widehat{\Omega}_f} &+ (\widehat{v}_s, \widehat{\psi}^u)_{\widehat{\Omega}_s} = 0 \quad \forall \widehat{\psi}^u \in \widehat{V}_{f,\widehat{u},\widehat{\Gamma}_i}^0,\\
(\widehat{\text{div}}\,(\widehat{J}\widehat{F}^{-1}\widehat{v}_f), \widehat{\psi}^p)_{\widehat{\Omega}_f} &= 0 \quad \forall \widehat{\psi}^p \in \widehat{L}_f^0,
\end{aligned}$$

with $\widehat{F} = \widehat{I} + \widehat{\nabla}\widehat{u}$, $\widehat{J} = det(\widehat{F})$, $\widehat{\sigma}_f = -\widehat{p}_f\widehat{I} + \widehat{\rho}_f\nu_f(\widehat{\nabla}\widehat{v}_f\widehat{F}^{-1} + \widehat{F}^{-T}\widehat{\nabla}\widehat{v}_f)$, $\widehat{\Sigma} = 2\mu_s\widehat{E} + \lambda_s tr(\widehat{E})\widehat{I}$, $\widehat{E} = 0.5(\widehat{F}^T\widehat{F} - \widehat{I})$, $\widehat{\sigma}_{mesh} = \alpha_u\widehat{\nabla}\widehat{u}_f$*, volume forces* $\widehat{f}_f$ *and* $\widehat{f}_s$ *(both zero in this work), flow correction term* $\widehat{g}_f$ *(do-nothing [35]), densities* $\widehat{\rho}_s, \widehat{\rho}_f$*, kinematic viscosity* ν_f*, and the Lamé parameters* μ_s, λ_s*. All explanations are provided in [56][Chapter 3].*

2.2 Discretization and numerical solution

For spatial discretization, a conforming Galerkin finite element scheme on quadrilateral mesh elements is employed [12]. Specifically, we use Q_2^c elements for $\widehat{v}$ and $\widehat{u} := \widehat{u}_f + \widehat{u}_s$, and Q_1^c elements for $\widehat{p}$. For the flow problem $(\widehat{v}, \widehat{p})$, this is the well-known inf-sup stable Taylor-Hood element; see e.g., [29]. Due to variational-monolithic coupling and globally-defined finite elements, the fluid pressure must be extended to the solid domain, which is achieved via $\alpha_u[(\widehat{\nabla}\widehat{p}_s, \widehat{\nabla}\widehat{\psi}^p) + (\widehat{p}_s, \widehat{\psi}^p)]$, and α_u (as before) small, positive. This is only for convenience, an alternative is to work with the `FE_NOTHING`[4] element in deal.II. The nonlinear problem is solved

[1] `https://github.com/tommeswick/fsi`
[2] `https://www.dealii.org/current/doxygen/deal.II/step_14.html`
[3] `https://github.com/tommeswick/goal-oriented-fsi`
[4] `https://www.dealii.org/current/doxygen/deal.II/step_46.html`

with Newton's method. Therein, for simplicity in this work, we utilize a sparse direct solver [14]. For algorithmic descriptions of our implementation, we refer to [56].

3 PU-DWR GOAL-ORIENTED ERROR CONTROL

The Galerkin approximation reads: Find $\widehat{U}_h = \{\widehat{v}_{f,h}, \widehat{u}_{f,h}, \widehat{u}_{s,h}, \widehat{p}_{f,h}\} \in \widehat{X}^0_{h,D}$, where $\widehat{X}^0_{h,D} := \{\widehat{v}^D_{f,h} + \widehat{V}^0_{f,\widehat{v},h}\} \times \{\widehat{u}^D_{f,h} + \widehat{V}^0_{f,\widehat{u},h}\} \times \{\widehat{u}^D_{s,h} + \widehat{V}^0_{s,h}\} \times \widehat{L}^0_{f,h}$, such that

$$\hat{A}(\widehat{U}_h)(\widehat{\Psi}_h) = \hat{F}(\widehat{\Psi}_h) \quad \forall \widehat{\Psi}_h \in \widehat{X}_h, \tag{1}$$

where $\widehat{X}_h$ is the test space with homogeneous Dirichlet conditions.

3.1 Goal functional

The solution $\widehat{U}_h$ is used to calculate an approximation $J(\widehat{U}_h)$ of the goal-functional $J(\widehat{U}) : \widehat{X} \to \mathbb{R}$. This functional is assumed to be sufficiently differentiable. The drag value as goal functional reads

$$J(\widehat{U}) := \int_{\hat{S}} \widehat{J}\widehat{\sigma}_f \widehat{F}^{-T} \widehat{n}_f \, \widehat{d} \, \mathrm{d}\widehat{s},$$

where $\widehat{n}_f$ is the outward point normal vector of the cylinder boundary $\hat{S}$ [36] and the FSI interface $\widehat{\Gamma}_i$. Moreover, $\widehat{d}$ is a unit vector perpendicular to the mean flow direction. For the drag, we use $\widehat{d} = (1, 0)$.

3.2 Error representation

We use the (formal) Euler-Lagrange method, to derive a computable representation of the approximation error $J(\widehat{U}) - J(\widehat{U}_h)$. The task is: Find $\widehat{U} \in \widehat{X}^0_D$ such that

$$\min\{J(\widehat{U}) - J(\widehat{U}_h)\} \quad \text{s.t. } \widehat{A}(\widehat{U})(\widehat{\Psi}) = \hat{F}(\widehat{\Psi}) \quad \forall \widehat{\Psi} \in \widehat{X},$$

from which we obtain the optimality system

$$\mathcal{L}'_{\widehat{Z}}(\widehat{U}, \widehat{Z})(\delta\widehat{Z}) = \widehat{F}(\delta\widehat{Z}) - \widehat{A}(\widehat{U})(\delta\widehat{Z}) = 0 \quad \forall \delta\widehat{Z} \in \widehat{X}, \quad \text{(Primal problem)},$$
$$\mathcal{L}'_{\widehat{U}}(\widehat{U}, \widehat{Z})(\delta\widehat{U}) = J'(\widehat{U})(\delta\widehat{U}) - \widehat{A}'_{\widehat{U}}(\widehat{U})(\delta\widehat{U}, \widehat{Z}) = 0 \quad \forall \delta\widehat{U} \in \widehat{X}, \quad \text{(Adjoint problem)}.$$

Using the main theorem from [6], we obtain:

Theorem 3.1. *We have the error identity:*

$$J(\widehat{U}) - J(\widehat{U}_h) = \frac{1}{2}\rho(\widehat{U}_h)(\widehat{Z} - \widehat{\Phi}_h) + \frac{1}{2}\rho^*(\widehat{U}_h, \widehat{Z}_h)(\widehat{U} - \widehat{\Psi}_h) + \mathcal{R}^{(3)}_h, \tag{2}$$

for all $\{\widehat{\Psi}_h, \widehat{\Phi}_h\} \in \widehat{X}_h \times \widehat{X}_h$ *and with the primal and adjoint residuals:*

$$\rho(\widehat{U}_h)(\widehat{Z} - \widehat{\Phi}_h) := -A(\widehat{U}_h)(\cdot) + \hat{F}(\cdot),$$
$$\rho^*(\widehat{U}_h, \widehat{Z}_h)(\widehat{U} - \widehat{\Psi}_h) := J'(\widehat{U}_h)(\cdot) - A'(\widehat{U}_h)(\cdot, \widehat{Z}_h) + \hat{F}(\cdot).$$

The remainder term is $\mathcal{R}^{(3)}_h$ *is of cubic order. This error identity can be used to define the error estimator* η*, which can be further utilized to design adaptive schemes.*

Corollary 3.2 (Primal error)**.** *The primal error identity reads:*

$$J(\widehat{U}) - J(\widehat{U}_h) = \rho(\widehat{U}_h)(\widehat{Z} - \widehat{\Phi}_h) + \mathcal{R}_h^{(2)}. \tag{3}$$

3.3 Adjoint equation, discretization, and numerical solution

The adjoint equation reads: Find $\widehat{Z} = (\widehat{z}^v, \widehat{z}^u, \widehat{z}^p) \in \widehat{X}$ such that

$$J'(\widehat{U})(\widehat{\Phi}) = \widehat{A}'_{\widehat{U}}(\widehat{U})(\widehat{\Phi}, \widehat{Z}) \quad \forall \widehat{\Phi} \in \widehat{X},$$

and the explicit form can be found in [56, 60].

For the discretization, we briefly mention that higher-order information for the adjoint solution must be employed due to Galerkin orthogonality; in this work $\widehat{X}_h \subset \widehat{X}_h^{(2)} \subset \widehat{X}$. For simplicity, this is realized with global-higher order finite elements and in order to ensure again inf-sup stability, we use Q_4^c elements for $\widehat{z}^v$ and $\widehat{z}^u$, and Q_2^c elements for $\widehat{z}^p$. It is clear that this is an expensive choice. For the numerical solution, the same solvers as for the primal problem are taken (see Section 2.2), namely a Newton-type method and sparse direct solver. Since the adjoint problem is linear, Newton's method converges in one step. This is a trivial information, but for debugging reasons useful.

3.4 Localization

A PU localization [49] for stationary FSI reads:

Proposition 3.1. *We have for the primal error part $\rho(\widehat{U}_h)(\cdot)$ the a posteriori error estimate*

$$|J(\widehat{U}) - J(\widehat{U}_h)| \leq \eta := \big|\sum_{i=1}^{M} \eta_i\big| \leq \sum_{i=1}^{M} |\eta_i| \tag{4}$$

where M is the dimension of the PU finite element space $\widehat{V}_{PU}$ (composed of Q_1^c functions χ_i) and with the PU-DoF indicators

$$\begin{aligned}
\eta_i &= -A(\widehat{U}_h)((\widehat{Z}_h^{(2)} - i_h\widehat{Z}_h^{(2)})\widehat{\Psi}_i) + \hat{F}((\widehat{Z}_h^{(2)} - i_h\widehat{Z}_h^{(2)})\widehat{\Psi}_i) \\
&= -(\widehat{\rho}_f\widehat{J}(\widehat{F}^{-1}\widehat{v}_f \cdot \widehat{\nabla})\widehat{v}_f), \widehat{\psi}_i^v)_{\widehat{\Omega}_f} - (\widehat{J}\widehat{\sigma}_f\widehat{F}^{-T}, \widehat{\nabla}\widehat{\psi}_i^v)_{\widehat{\Omega}_f} + \langle \widehat{g}_f, \widehat{\psi}_i^v\rangle_{\widehat{\Gamma}_N} \\
&\quad - (\widehat{F}\widehat{\Sigma}, \widehat{\nabla}\widehat{\psi}_i^v)_{\widehat{\Omega}_s} - (\widehat{\sigma}_{mesh}, \widehat{\nabla}\widehat{\psi}_i^u)_{\widehat{\Omega}_f} - (\widehat{\mathrm{div}}\,(\widehat{J}\widehat{F}^{-1}\widehat{v}_f), \widehat{\psi}_i^p)_{\widehat{\Omega}_f} \\
&\quad + (\widehat{\rho}_f\widehat{J}\widehat{f}_f, \widehat{\psi}_i^v)_{\widehat{\Omega}_f} + (\widehat{\rho}_s\widehat{f}_s, \widehat{\psi}_i^v)_{\widehat{\Omega}_s}
\end{aligned}$$

with the interpolation $i_h : \widehat{X}_h^{(2)} \to \widehat{X}_h$ and the weighting functions are defined as

$$\widehat{\psi}_i^v := (\phi_{2h,v}^{(2)} - \phi_{h,v})\chi_i, \quad \widehat{\psi}_i^u := (\phi_{2h,u}^{(2)} - \phi_{h,u})\chi_i, \quad \widehat{\psi}_i^p := (\phi_{2h,p}^{(2)} - \phi_{h,p})\chi_i.$$

3.5 Adaptive algorithm

1. Compute the primal solution $\widehat{U}_h$ and the (higher-order) adjoint solution $\widehat{Z}_h^{(2)}$ on the present mesh $\mathcal{T}_h$.
2. Evaluate $|\eta| := |\sum_i \eta_i|$ in (4).
3. Check, if the stopping criterion is satisfied: $|J(\widehat{U}) - J(\widehat{U}_h)| \leq |\eta| \leq TOL$, then accept U_h within the tolerance TOL. Otherwise, proceed to the following step.

4. Mark all elements K_i for refinement that touch DoFs i with indicator η_i with $\eta_i \geq \frac{\alpha\eta}{M_{el}}$ (where M_{el} denotes the total number of elements of the mesh $\mathcal{T}_h$ and $\alpha \approx 1$).
   Alternatively, pure DoF-based refinement in i can be carried out.

4 NUMERICAL TESTS

In this section, we consider the FSI-1 benchmark [36] (see also the books [11, 10] and our own former results [47, 58]) and the 2D-1 benchmark [50]. The drag value is taken as goal functional. As previously mentioned, this paper is accompanied with a respective open-source implementation on github[5] based on the finite element library deal.II [1, 2] and our previous fluid-structure interaction code [58], which is also available on github[6].

4.1 FSI-1 benchmark

The configuration, all parameters, and reference values can be found in [36]. The reference value for computing the true error was computed on a five times refined mesh and is $1.5370185576528707e+01$ (see also in the provided github code). Our results from the file `dwr_results.txt` are:

Dofs	True err	Est err	Est ind	Eff	Ind
13310	2.58e-01	1.43e-01	4.37e-01	5.54e-01	1.69e+00
20921	9.00e-02	4.75e-02	1.60e-01	5.28e-01	1.77e+00
37874	3.20e-02	1.09e-02	5.96e-02	3.40e-01	1.86e+00
68754	1.84e-02	4.57e-03	2.77e-02	2.48e-01	1.51e+00

Furthermore, the terminal output yields

```
DisX  :     2.2656126465725842e-05
DisY  :     8.1965770448936843e-04
P-Diff:     1.4819455817646477e+02
P-front:    1.4819455817646477e+02
------------------
Face drag:      1.5351806985399641e+01
Face lift:      7.3933527637991259e-01
```

where `Face drag` represents the chosen goal functional. While the error reductions in the `True err` $J(\widehat{U}) - J(\widehat{U}_h)$ and the estimated error η are reasonable, the effectivity index `Eff` has room for improvement. The indicator index `Ind` (for the definition see [49]) performs quite well. The main reason for the intermediate effectivity indices might be the accuracy of the reference value. Second, we notice that only the primal error part ρ (Corollary 3.2) was used. As shown in our recent studies for quasi-linear problems, the adjoint error part ρ^* might play a crucial role in order to obtain nearly perfect effectivity indices for highly nonlinear problems [18]. Graphical solutions of the primal solution, including the adaptively refined mesh, and the adjoint solution are displayed in Figure 1.

4.2 Adaptation to flow benchmark 2D-1

The provided code can be adapted with minimal changes to the 1996 flow around cylinder benchmark 2D-1 [50]. In the `*.inp` file the material ids for solid must be set to 0 (flow), and

[5] `https://github.com/tommeswick/goal-oriented-fsi`
[6] `https://github.com/tommeswick/fsi`

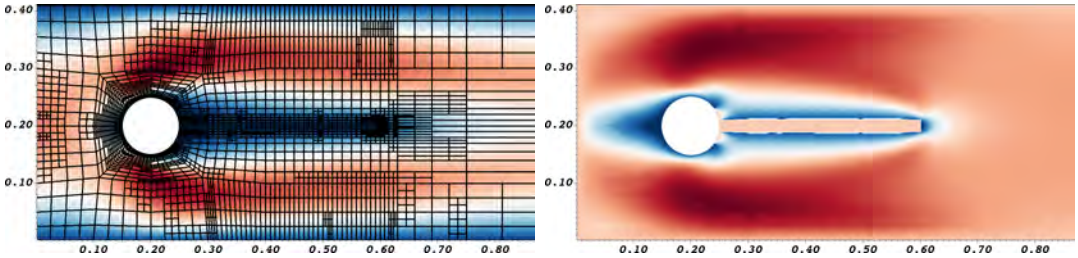


Figure 1: FSI-1 benchmark: Primal solution of $\widehat{v}_x$ and adjoint solution $\widehat{z}^{v_x}$. The adaptive mesh is displayed together with the primal solution (right).

the inflow and material parameters are adapted correspondingly. Of course, in this code, the displacement variables are still computed despite that they are zero everywhere, which increases the computational cost in comparison to a pure fluid flow code. Due to the zero displacements $\widehat{u} = 0$, the ALE mapping is the identity, yielding $\widehat{F} = \widehat{I}$ and $det(\widehat{F}) = 1$. Consequently, there is no mesh deformation and the Navier-Stokes equations fully remain in Eulerian coordinates. Here, extracting information from `dwr_results.txt`, the findings for the drag value as goal functional are:

Dofs	True err	Est err	Est ind	Eff	Ind
1610	3.51e-01	2.97e-01	6.20e-01	8.44e-01	1.76e+00
2586	8.80e-02	7.27e-02	2.21e-01	8.26e-01	2.51e+00
4764	1.89e-02	1.54e-02	7.11e-02	8.14e-01	3.75e+00
10830	3.23e-03	2.95e-03	1.82e-02	9.13e-01	5.62e+00

The pressure, drag (goal functional), and lift values are taken from the terminal output:

```
P-Diff:      1.1743527755157424e-01
P-front:     1.3213237901562136e-01
P-back:      1.4697101464047121e-02
------------------
Face drag:       5.5754969431700365e+00
Face lift:       1.0717678080199560e-02
```

These values fit well with the reference values given in [50]. Moreover, we observe very stable effectivity indices, which indicate that the primal error estimator ρ (Corollary 3.2) is for incompressible Navier-Stokes a sufficient choice. Indeed, using this part only, was already suggested in early work [6, 9]. Finally, we notice that extensions to multiple goal functionals for the 2D-1 benchmark were undertaken in [17, 19].

5 CONCLUSIONS

In this work, we developed and implemented PU-DWR goal-oriented error control and spatial mesh adaptivity for stationary fluid-structure interaction. An important part is the open-source programming code published on github. As numerical example, the FSI-1 benchmark is chosen. Therein, mesh adaptivity performs as expected and also the error reductions in the true error and estimated error are good. However, the effectivity index may be improved. Extensions of this work include inter alia the implementation of the adjoint error part ρ^*, local-higher order interpolations for the adjoint rather than using global-higher order finite elements, parallel iterative/multigrid linear solvers within Newton's method, and a 3D implementation. The latter is implementation-wise not difficult with deal.II's dimension-independent programming, but the linear solver becomes really important.

6 ACKNOWLEDGEMENTS

This work is supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the cluster of Excellence PhoenixD (EXC 2122, Project ID 390833453).

REFERENCES

[1] D. Arndt, W. Bangerth, T. C. Clevenger, D. Davydov, M. Fehling, D. Garcia-Sanchez, G. Harper, T. Heister, L. Heltai, M. Kronbichler, R. M. Kynch, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells. The `deal.II` library, version 9.1. *Journal of Numerical Mathematics*, 27:203–213, 2019.

[2] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells. The deal.ii finite element library: Design, features, and insights. *Computers & Mathematics with Applications*, 2020.

[3] W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Birkhäuser, Lectures in Mathematics, ETH Zürich, 2003.

[4] A. T. Barker and X.-C. Cai. Scalable parallel methods for monolithic coupling in fluid-structure interaction with application to blood flow modeling. *Journal of Computational Physics*, 229(3):642 – 659, 2010.

[5] Y. Bazilevs, K. Takizawa, and T. Tezduyar. *Computational Fluid-Structure Interaction: Methods and Applications*. Wiley, 2013.

[6] R. Becker and R. Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica, Cambridge University Press*, pages 1–102, 2001.

[7] C. Bertoglio, P. Moireau, and J. Gerbeau. Sequential parameter estimation in fluid-structure problems. application to hemodynamics. *Int. J. Numer. Meth. Biomed. Engrg.*, 28:434–455, 2012.

[8] T. Bodnár, G. Galdi, and Š. Nečasová. *Fluid-Structure Interaction and Biomedical Applications*. Advances in Mathematical Fluid Mechanics. Springer Basel, 2014.

[9] M. Braack and T. Richter. Solutions of 3d navier–stokes benchmark problems with adaptive finite elements. *Computers & Fluids*, 35(4):372 – 392, 2006.

[10] H.-J. Bungartz, M. Mehl, and M. Schäfer. *Fluid-Structure Interaction II: Modelling, Simulation, Optimization*. Lecture Notes in Computational Science and Engineering. Springer, 2010.

[11] H.-J. Bungartz and M. Schäfer. *Fluid-Structure Interaction: Modelling, Simulation, Optimization*, volume 53 of *Lecture Notes in Computational Science and Engineering*. Springer, 2006.

[12] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland, Amsterdam [u.a.], 2. pr. edition, 1987.

[13] P. Crosetto, S. Deparis, G. Fourestey, and A. Quarteroni. Parallel algorithms for fluid-structure interaction problems in haemodynamics. *SIAM J. Sci. Comp.*, 33(4):1598–1622, 2011.

[14] T. A. Davis and I. S. Duff. An unsymmetric-pattern multifrontal method for sparse LU factorization. *SIAM J. Matrix Anal. Appl.*, 18(1):140–158, 1997.

[15] S. Deparis, D. Forti, G. Grandperrin, and A. Quarteroni. Facsi: A block parallel preconditioner for fluid-structure interaction in hemodynamics. *Journal of Computational Physics*, 327:700–718, 2016.

[16] T. Dunne. An Eulerian approach to fluid-structure interaction and goal-oriented mesh adaption. *Int. J. Numer. Methods in Fluids*, 51:1017–1039, 2006.

[17] B. Endtmayer, U. Langer, J. P. Thiele, and T. Wick. Hierarchical DWR Error Estimates for the Navier Stokes Equation: h and p Enrichment, 2019.

[18] B. Endtmayer, U. Langer, and T. Wick. Multigoal-Oriented Error Estimates for Non-linear Problems. *Journal of Numerical Mathematics*, 27(4):215–236, 2019.

[19] B. Endtmayer, U. Langer, and T. Wick. Two-Side a Posteriori Error Estimates for the Dual-Weighted Residual Method. *SIAM J. Sci. Comput.*, 42(1):A371–A394, 2020.

[20] L. Failer. *Optimal Control of Time-Dependent Nonlinear Fluid-Structure Interaction*. PhD thesis, Technical University Munich, 2017.

[21] L. Failer and T. Richter. A newton multigrid framework for optimal control of fluid–structure interactions. *Optimization and Engineering*, 2020.

[22] L. Failer and T. Wick. Adaptive time-step control for nonlinear fluid-structure interaction. *Journal of Computational Physics*, 366:448 – 477, 2018.

[23] P. Fick, E. Brummelen, and K. Zee. On the adjoint-consistent formulation of interface conditions in goal-oriented error estimation and adaptivity for fluid-structure interaction. *Computer Methods in Applied Mechanics and Engineering*, 199:3369–3385, 2010.

[24] L. Formaggia, J.-F. Gerbeau, F. Nobile, and A. Quarteroni. On the coupling of 3d and 1d Navier-Stokes equations for flow problems in compliant vessels. *Comp. Methods Appl. Mech. Engig*, 191:561–582, 2001.

[25] L. Formaggia, A. Quarteroni, and A. Veneziani. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system*. Springer-Verlag, Italia, Milano, 2009.

[26] S. Frei, B. Holm, T. Richter, T. Wick, and H. Yang. *Fluid-structure interactions: Fluid-Structure Interaction: Modeling, Adaptive Discretisations and Solvers*. de Gruyter, 2017.

[27] G. Galdi and R. Rannacher. *Fundamental Trends in Fluid-Structure Interaction*. World Scientific, 2010.

[28] M. Gee, U. Küttler, and W. Wall. Truly monolithic algebraic multigrid for fluid–structure interaction. *Int. J. Numer. Meth. Engrg.*, 85(8):987–1016, 2011.

[29] V. Girault and P.-A. Raviart. *Finite Element method for the Navier-Stokes equations*. Number 5 in Computer Series in Computational Mathematics. Springer-Verlag, 1986.

[30] R. Glowinski, T. Pan, T. Hesla, D. Joseph, and J. Périaux. A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow. *Journal of Computational Physics*, 169(2):363 – 426, 2001.

[31] C. Grandmont. Existence for a three-dimensional steady state fluid-structure interaction problem. *Journal of Mathematical Fluid Mechanics*, 4:76–94, 2002.

[32] T. Grätsch and K.-J. Bathe. Goal-oriented error estimation in the analysis of fluid flows with structural interactions. *Comp. Methods Appl. Mech. Engrg.*, 195:5673–5684, 2006.

[33] N. Hagmeyer, M. Mayr, I. Steinbrecher, and A. Popp. Fluid-beam interaction: Capturing the effect of embedded slender bodies on global fluid flow and vice versa, 2021.

[34] M. Heil. An efficient solver for the fully coupled solution of large-displacement fluid-structure interaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1–23, 2004.

[35] J. G. Heywood, R. Rannacher, and S. Turek. Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations. *International Journal of Numerical Methods in Fluids*, 22:325–352, 1996.

[36] J. Hron and S. Turek. *Proposal for numerical benchmarking of fluid-structure interaction between an elastic object and laminar incompressible flow*, volume 53, pages 146 – 170. Springer-Verlag, 2006.

[37] T. Hughes, W. Liu, and T. Zimmermann. Lagrangian-Eulerian finite element formulation for incompressible viscous flows. *Comput. Methods Appl. Mech. Engrg.*, 29:329–349, 1981.

[38] D. Jodlbauer, U. Langer, and T. Wick. Parallel block-preconditioned monolithic solvers for fluid-structure interaction problems. *Int. J. Num. Meth. Eng.*, 117(6):623–643, 2019.

[39] J. Kratzke. *Uncertainty Quantification for Fluid-Structure Interaction: Application to Aortic Biomechanics*. PhD thesis, University of Heidelberg, 2018.

[40] T. Lassila, A. Manzoni, A. Quarteroni, and G. Rozza. A reduced computational and geometrical framework for inverse problems in hemodynamics. *Int. J. Numer. Meth. Biomed. Engrg.*, 29(7):741–776, 2013.

[41] M. Perego, A. Veneziani, and C. Vergara. A variational approach for estimating the compliance of the cardiovascular tissue: An inverse fluid-structure interaction problem. *SIAM Journal on Scientific Computing*, 33(3):1181–1211, 2011.

[42] C. Peskin. *The immersed boundary method*, pages 1–39. Acta Numerica 2002, Cambridge University Press, 2002.

[43] M. Razzaq, H. Damanik, J. Hron, A. Ouazzi, and S. Turek. FEM multigrid techniques for fluid-structure interaction with application to hemodynamics. *Appl. Numer. Math.*, 62(9):1156–1170, 2012.

[44] T. Richter. Goal-oriented error estimation for fluid–structure interaction problems. *Computer Methods in Applied Mechanics and Engineering*, 223-224:28 – 42, 2012.

[45] T. Richter. A monolithic geometric multigrid solver for fluid-structure interactions in ale formulation. *International Journal for Numerical Methods in Engineering*, pages 372–390, 2015.

[46] T. Richter. *Fluid-structure interactions: models, analysis, and finite elements*. Springer, 2017.

[47] T. Richter and T. Wick. Finite elements for fluid-structure interaction in ALE and fully Eulerian coordinates. *Comp. Methods Appl. Mech. Engrg.*, 199:2633–2642, 2010.

[48] T. Richter and T. Wick. Optimal control and parameter estimation for stationary fluid-structure interaction. *SIAM J. Sci. Comput.*, 35(5):B1085–B1104, 2013.

[49] T. Richter and T. Wick. Variational localizations of the dual weighted residual estimator. *Journal of Computational and Applied Mathematics*, 279(0):192 – 208, 2015.

[50] M. Schäfer and S. Turek. *Flow Simulation with High-Performance Computer II*, volume 52 of *Notes on Numerical Fluid Mechanics*, chapter Benchmark Computations of laminar flow around a cylinder. Vieweg, Braunschweig Wiesbaden, 1996.

[51] K. Takizawa and T. Tezduyar. Multiscale space–time fluid–structure interaction techniques. *Computational Mechanics*, 48:247–267, 2011.

[52] A. Tello, R. Codina, and J. Baiges. Fluid structure interaction by means of variational multiscale reduced order models. *International Journal for Numerical Methods in Engineering*, 121(12):2601–2625, 2020.

[53] T. Tezduyar, S. Sathe, R. Keedy, and K. Stein. Space-time finite element techniques for computation of fluid-structure interactions. *Comp. Meth. Appl. Mech. Engrg.*, 195:2002–2027, 2006.

[54] K. van der Zee, E. van Brummelen, I. Akkerman, and R. de Borst. Goal-oriented error estimation and adaptivity for fluid–structure interaction using exact linearized adjoints. *Computer Methods in Applied Mechanics and Engineering*, 200(37):2738–2757, 2011.

[55] M. Wichrowski. *Fluid-structure interaction problems: velocity-based formulation and monolithic computational methods*. PhD thesis, Polish Academy of Sciences, 2021.

[56] T. Wick. *Adaptive Finite Element Simulation of Fluid-Structure Interaction with Application to Heart-Valve Dynamics*. PhD thesis, University of Heidelberg, 2011.

[57] T. Wick. Goal-oriented mesh adaptivity for fluid-structure interaction with application to heart-valve settings. *Arch. Mech. Engrg.*, 59(6):73–99, 2012.

[58] T. Wick. Solving monolithic fluid-structure interaction problems in arbitrary Lagrangian Eulerian coordinates with the deal.II library. *Archive of Numerical Software*, 1:1–19, 2013.

[59] T. Wick. *Multiphysics Phase-Field Fracture: Modeling, Adaptive Discretizations, and Solvers*. De Gruyter, Berlin, Boston, 2020.

[60] T. Wick. On the Adjoint Equation in Fluid-Structure Interaction. WCCM-ECCOMAS2020, 2021.

[61] T. Wick and W. Wollner. On the differentiability of fluid–structure interaction problems with respect to the problem data. *Journal of Mathematical Fluid Mechanics*, 21(3):34, Jun 2019.

[62] T. Wick and W. Wollner. Optimization with nonstationary, nonlinear monolithic fluid-structure interaction. *International Journal for Numerical Methods in Engineering*, pages 1–20, 2020.

NUMERICAL METHODS FOR CHARACTERIZATION OF RAILWAY DYNAMICS AND VIBRO-ACOUSTICS

ANALYSIS OF THE INFLUENCE OF THE BALLAST TRACK IN THE DYNAMIC BEHAVIOUR OF SINGLE-TRACK RAILWAY BRIDGES OF DIFFERENT TYPOLOGIES

J. Chordà-Monsonís*†, M.D. Martínez-Rodrigo*, P. Galvín†, A. Romero† and E. Moliner*

* Departamento de Ingeniería Mecánica y Construcción
Universitat Jaume I
12006 Castelló, Spain

† Departamento de Mecánica de Medios Continuos y Teoría de Estructuras
Universidad de Sevilla
41092 Sevilla, Spain

e-mail: chordaj@uji.es, mrodrigo@uji.es, pedrogalvin@us.es, mrodrigo@uji.es, molinere@uji.es

Key words: Railway Induced Vibrations, Bridges, Track-Bridge Interaction, Resonance, Vertical Acceleration, Ballast Track.

Abstract: *Short-to-medium span simply-supported (SS) railway bridges are prone to experience high levels of vertical acceleration due to train passage. The necessity of predicting accurately their dynamic behaviour for design, safety and maintenance reasons, requires a deep understanding of the train induced vibrations in these structures. A key factor of this phenomenon is the influence exerted by the ballast track on their dynamic response. This paper provides a detailed sensitivity analysis over a single-track bridge catalogue covering lengths of interest from 10 to 25 m considering two different typologies, (i) girder-deck bridges and (ii) slab-deck bridges. The effect of the vertical flexibility of elastic bearings is also analysed. A 2D Finite-Element (FE) track-bridge interaction model is implemented with the aim to evaluate the influence of the track parameters on the modal properties of the bridges and the dynamic response under train passages. The results obtained reveal the influence of the ballast shear stiffness and damping in the dynamic behaviour of the structures, especially in the case of the shortest girder bridges.*

1 INTRODUCTION

In a context of an increasing demand of personal and freight mobility around the world, railway systems have experienced a sustained development that projects them as a reliable and sustainable way of transportation for the time to come. For this reason, dynamic effects on railway bridges are considered of major interest and concern for scientists and engineers, especially since the appearance of High Speed (HS) [1]. In this regard, short-to-medium span (10 – 25 m) SS railway bridges are particularly prone to experience an excessive level of vertical acceleration at the deck during train passage, due to its usually associated low mass and structural damping, especially at resonance [2]. This could cause discomfort for the passengers, flaws in the ballast layer, a rise in the maintenance service cost of the track and an increased risk of derailment in the worst-case scenario. Train induced vibrations in railway bridges is a rather complex interaction problem, which is affected by several factors. Apart from the mechanical and geometrical properties of the bridge and the characteristics of the train, interaction mechanisms regarding the vehicle, the track and the soil may also affect the response of the structure, which are currently under investigation [3]. In addition, the computational cost of including these mechanisms is considerable, thus, simplified models that usually disregard them are commonly used in engineering consultancies. This work is dedicated to the investigation of the effect exerted by the ballast track on the vertical dynamic response of SS

https://doi.org/10.4995/YIC2021.2021.12220

railway bridges. To this aim, the influence of the main track parameters on the bridge modal properties and on the dynamic response due to train passage is evaluated. With this purpose, a 2D FE track-bridge interaction model is implemented, where the track is represented using a three-layer discrete model, based on the work by Zhai et al. [4]. The model is employed to perform a sensitivity analysis over a bridge catalogue covering bridges of two different deck typologies and for a selected range of lengths of interest from 10 to 25 m. In sections 2 and 3, the bridge catalogue is presented, and the numerical model is described. In section 4, the results of dynamic analyses under train passage are included. Finally, in section 5, the main conclusions are summarized.

2 BRIDGE CATALOGUE

The catalogue contemplates single-track railway bridges of span lengths that range from 10 to 25 m in 5 m intervals. For each length, two common deck typologies are considered: (i) pre-stressed concrete girder decks; and (ii) voided or solid concrete slabs, or pre-stressed filler beams encased in a concrete pseudo-slab. As for the vertical support of the decks, infinitely rigid supports and elastic supports accounting for the vertical flexibility of neoprene bearings are differentiated.

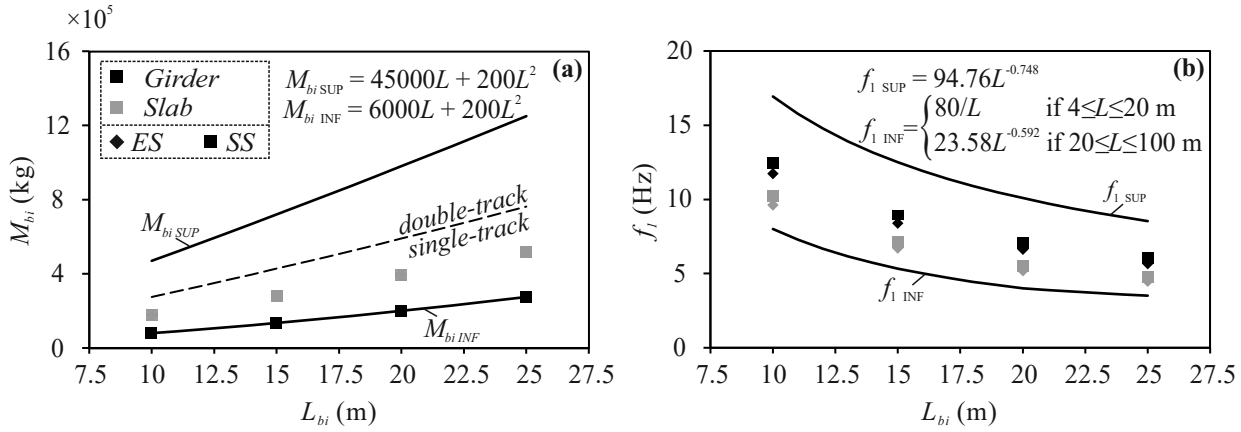


Figure 1: (a) Mass per span and (b) fundamental frequency of the bridges under study.

The main characteristics regarding the mass and the fundamental frequency of the bridges are calculated according to the work presented by Doménech et al. [5], where an ensemble of existing bridges of the considered typologies was studied. Fig. 1 shows for the 16 bridges of the catalogue the total mass per bridge span and the fundamental frequency. For the girder decks, the mass of the reported single-track existing bridges approaches the inferior limit. Additionally, this corresponds to the worst-case scenario for the vertical acceleration criterion. The fundamental frequency is selected as 50% of the difference between the Eurocode 1 (EC1) simplified method limits for each length [6]. For the slab decks, the mass value is selected as 25% of the difference between the upper and the lower limits for each length. This corresponds to an average value for the mass of existing single-track slab bridges. For the fundamental frequency, the same criterion is applied, and the frequency is calculated as 25% of the difference between the limits for each length. In addition, an elastically-supported (ES) version for each bridge is also defined admitting that the ratio κ between the bridge bending stiffness and the vertical stiffness of the bearings is approximately equal to 0.05, which leads to a reduction of the fundamental frequency of 3-4% with respect to the SS case [7], as indicated in Eq. 1. In this equation, $E_{bi}I_{ybi}$ stands for cross-section flexural stiffness of each section, $\bar{K}_{bi,dyn}$ for the

vertical dynamic stiffness of the elastic bearings and L_{bi} for the span length.

$$\kappa = \frac{E_{bi}\, I_{ybi}\, \pi^3}{\bar{K}^n_{bi,dyn}\, L^3_{bi}} \approx 0.05 \tag{1}$$

The mechanical properties of the bridges of the catalogue are shown in Tables 1 and 2, where the data is expressed per bridge span. From left to right, the columns show the information relative to the span length, L_{bi}, fundamental frequency, f_1, total mass, M_{bi}, cross-section flexural stiffness of the span section, $E_{bi}I_{ybi}$, and the vertical dynamic stiffness of the elastic bearings $\bar{K}_{bi,dyn}$, respectively. The last column stands for the identification code for each bridge, which contains the typology, the type of support and the span length (e.g. GD-ES-10 stands for girder-deck bridge, elastically-supported with 10 m of span length).

$L_{bi}[m]$	$f_1[Hz]$	$M_{bi}[t]$	$E_{bi}I_{ybi}[MN/m^2]$	$\bar{K}_{bi,dyn}[MN/m]$	ID
10	12.46	80.0	$3.56{\cdot}10^3$	∞	GD-SS-10
	11.72	80.0	$3.18{\cdot}10^3$	$3.12{\cdot}10^3$	GD-ES-10
15	8.92	135.0	$1.06{\cdot}10^4$	∞	GD-SS-15
	8.39	135.0	$9.63{\cdot}10^3$	$2.70{\cdot}10^3$	GD-ES-15
20	7.04	200.0	$2.41{\cdot}10^4$	∞	GD-SS-20
	6.62	200.0	$2.20{\cdot}10^4$	$2.49{\cdot}10^3$	GD-ES-20
25	6.02	275.0	$4.93{\cdot}10^4$	∞	GD-SS-25
	5.66	275.0	$4.51{\cdot}10^4$	$2.50{\cdot}10^3$	GD-ES-25

Table 1: Mechanical properties of the girder bridges.

$L_{bi}[m]$	$f_1[Hz]$	$M_{bi}[t]$	$E_{bi}I_{ybi}[MN/m^2]$	$\bar{K}_{bi,dyn}[MN/m]$	ID
10	10.22	177.5	$6.63{\cdot}10^3$	∞	SD-SS-10
	9.62	177.5	$6.06{\cdot}10^3$	$4.67{\cdot}10^3$	SD-ES-10
15	7.12	281.3	$1.66{\cdot}10^4$	∞	SD-SS-15
	6.70	281.3	$1.53{\cdot}10^4$	$3.59{\cdot}10^3$	SD-ES-15
20	5.52	395.0	$3.32{\cdot}10^4$	∞	SD-SS-20
	5.19	395.0	$3.05{\cdot}10^4$	$3.03{\cdot}10^3$	SD-ES-20
25	4.76	518.8	$6.41{\cdot}10^4$	∞	SD-SS-25
	4.48	518.8	$5.95{\cdot}10^4$	$2.96{\cdot}10^3$	SD-ES-25

Table 2: Mechanical properties of the slab bridges.

3 TRACK-BRIDGE INTERACTION MODEL

For the subsequent analysis, the discrete FE 2D track-bridge interaction model shown in Fig. 2 is implemented. A three-layer discrete model for the track is configured, based on that proposed by Zhai et al. [4], which couples a series of elastically or simply-supported bridge spans. The track admits Ahlbeck hypothesis, so it can be assumed that the load transmitted from each sleeper to the ballast has a cone distribution. In the proposed model, the rail is represented with a Bernoulli-Euler (B-E) beam, where E_r, I_{yr}, and m_r stand for the rail Young Modulus, cross-section moment of inertia with respect to the Y axis and linear mass, respectively. Below, the vertical damping and stiffness of the rail pads (C_p, K_p), of the mobilized ballast (C_b, K_b) and of the subgrade (C_f, K_f) are included at the sleepers locations. The

continuity and coupling effect of the interlocking ballast granules is also considered in the model by means of spring-damper elements (C_w, K_w) that link relative vertical displacements between adjacent ballast masses. Then, M_{sl} and M_b stand for the mass of each sleeper and the vibrating ballast mass under each support, respectively. Damping and stiffness on the bridge deck (C_f^b, K_f^b) are set to 0 and $100 \cdot K_f$, respectively, assuming that the ballast rests directly on the bridge deck. The longitudinal interaction between the rails and the deck through the ballast layer is disregarded in a first approach given the high flexural stiffness of the bridges. As shown in Fig. 2, rail and track parameters are multiplied by a factor of two, as only one rail is explicitly included in the model. The bridge is represented by means of N_{sp} simply or elastically-supported B-E beams representing each span of the bridge. In the present paper, N_{sp} is set to a value of 2, as two identical spans are considered for each bridge. The vertical stiffness of the neoprene bearings is introduced by the constant equivalent vertical stiffness $\bar{K}_{bi,dyn}$ at each end section of the i-th bridge span. The parameters L_{bi}, E_{bi}, I_{bi} and m_{bi} stand for the length, Young Modulus, cross-section moment of inertia with respect to the Y axis and linear mass of the i-th bridge span, respectively. Due to the presence of the continuous ballast track, a weak interaction takes place between successive spans. In the simulations, a track length of $L_{r,prev} = 20$ m is included before and after the bridge, which is considered sufficient according to previous publications [8], corresponding to 33.3 times the sleeper distances. The rail is discretized into two beam elements between consecutive sleepers, and so are the bridge beams.

The train excitation is represented by means of a constant moving load model, which implies that vehicle-structure interaction effects are neglected. In this sense, it is intended to isolate the effect of the track components affecting the dynamic behaviour of the bridges to investigate their influence separately. For the track parameters, an important dispersion has been found among different publications. Based on a review presented by the authors in [9], the values selected are shown in Table 3, expressed per rail seat. M_b, K_b and K_f are calculated with the equations given in [4]. Data from the European [10] and Spanish Standards [12], and from [11] are adopted for the rail, rail pads and sleepers properties. In the case of the ballast shear stiffness and damping, the authors have found that most of the times these parameters are not considered in track models. In the few cases where included, the majority of them adopted those proposed in [4]. For this reason, in this work, these same values are employed, and its influence is investigated. The model is implemented in ANSYS. For the computation of the bridges response under passing trains (see section 4), mass, stiffness and damping matrices are exported to MATLAB, and the equations of motion of the full model are integrated in the time domain applying the Newmark-beta constant acceleration algorithm. The time step for the numerical integration is set as the minimum between 1/50 times the smaller period of interest and 1/20 times the load travelling time between two consecutive sleepers.

4 SENSITIVITY ANALYSIS: MODAL PROPERTIES AND VERTICAL ACCELERATION

This section presents the results for the sensitivity analysis regarding the influence of the track properties on the dynamic behaviour of the bridges. The authors have found that the only parameters that affect significantly the modal properties of the bridges at low frequencies are the ballast shear stiffness and damping (K_w, C_w). In this sense, Zhai et al. [4] pointed out their great influence on the dynamic behaviour of the track too. Thus, in what follows, individual variations of these track parameters are considered to evaluate how this impacts the modal properties and the vertical acceleration on the bridge deck under train passages. It is also intended to determine what bridges are the most affected by these variations.

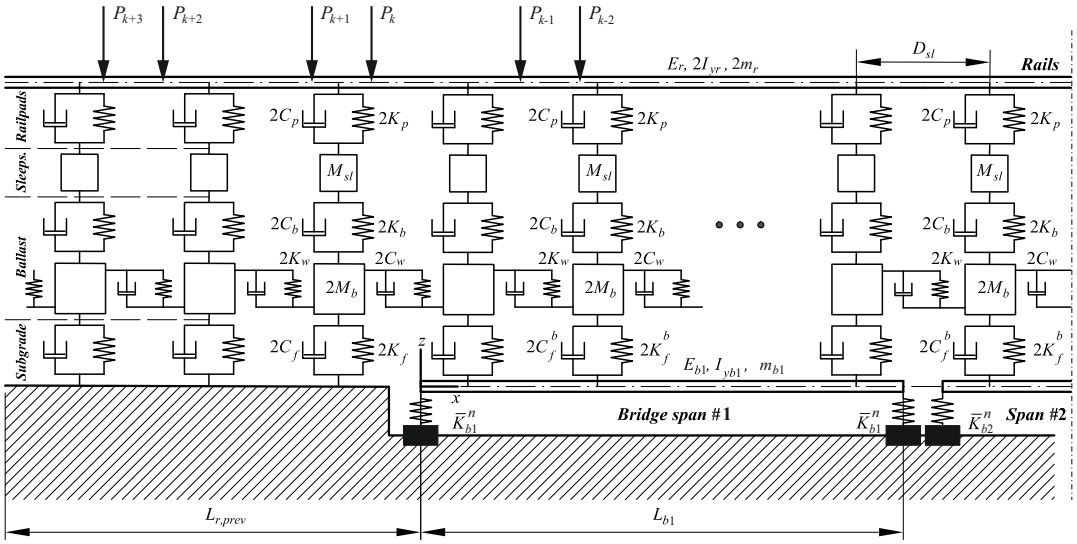


Figure 2: Track-bridge interaction model.

4.1 Influence of K_w on the bridge modal parameters

In this section the influence of the ballast shear stiffness on the bridge frequencies is evaluated. To this aim, the first, third and fifth longitudinal bending modal frequencies are calculated under individual variations of K_w. Fig. 3 shows the results for all the bridges in the catalogue, grouped per bridge length. Each plot shows the variation in the natural frequency f_i for $i =$ 1, 3, 5 when factors [0.0, 0.5, 1.0, 1.5, 2.0] multiply the nominal value of K_w (Table 3) with respect to the nominal case. From the results obtained, the following is observed:

- Natural frequencies increase with K_w. Bridges with shorter spans in a certain typology are more affected with the variation of this parameter.
- The fundamental frequency f_1 corresponding to the first longitudinal bending mode is significantly more affected than higher frequencies. The effect of K_w reduces with the frequency number.
- Regarding the typology, girder bridges, with lower longitudinal bending stiffness, are affected to a higher extent than slab bridges.
- As per the bridge supports, bridges on elastic supports are slightly more affected by K_w variations than rigidly supported bridges. Nevertheless, the difference is not significant, especially for modes higher than the fundamental one.

These results are consistent in all the considered bridges. From the sensitivity analysis it is concluded that regarding the modal parameters, short-span elastically-supported girder bridges are the most sensitive ones to the value of K_w. On this matter, the maximum variations for the frequency obtained for the first, third and fifth modes are 20%, 6% and 3%, respectively, for the shortest bridge considered (GD-ES-10), and 10%, 3% and 1.5% for the longest one (GD-ES-25).

4.2 Influence of K_w and C_w on the deck vertical acceleration due to train passage

The influence of K_w and C_w on the vertical acceleration at the bridge deck under train passages is investigated in this section. To this aim, several dynamic analyses are carried out on

Notation	Parameter	Value	Unit	Reference
E_r	Rail UIC 60 elastic modulus	$2.100 \cdot 10^{11}$	Pa	[10]
I_{yr}	Rail UIC 60 moment of inertia	$3038.3 \cdot 10^{-8}$	m^4	[10]
m_r	Rail UIC 60 mass per unit of length	60.21	kg/m	[10]
K_p	Rail pad vertical stiffness	$1.000 \cdot 10^8$	N/m	[11]
C_p	Rail pad damping	$7.500 \cdot 10^4$	Ns/m	[4]
M_{sl}	Sleeper mass	300	kg	[12]
D_{sl}	Sleeper distance	0.600	m	[12]
l_e	Half sleeper effective supporting length	0.950	m	[4]
l_b	Sleeper width	0.300	m	[12]
α	Ballast stress distribution angle	35	$\circ$	[4]
h_b	Ballast thickness	0.300	m	[12]
ρ_b	Ballast density	1800	kg/m^3	[4]
M_b	Ballast vibrating mass	317.910	kg	[4]
E_b	Ballast elastic modulus	$1.100 \cdot 10^8$	Pa	[4]
K_b	Ballast vertical stiffness	$1.933 \cdot 10^8$	N/m	[4]
C_b	Ballast damping	$5.880 \cdot 10^4$	Ns/m	[4]
E_f	Subgrade K_{30} modulus	$9.000 \cdot 10^7$	Pa/m	[4]
K_f	Subgrade vertical stiffness	$7.399 \cdot 10^7$	N/m	[4]
C_f	Subgrade damping	$3.115 \cdot 10^4$	Ns/m	[4]
K_w	Ballast shear stiffness	$7.840 \cdot 10^7$	N/m	[4]
C_w	Ballast shear damping	$8.000 \cdot 10^4$	Ns/m	[4]

Table 3: Bridge-track interaction model parameters, per rail seat.

the GD-ES-10 bridge under the circulation of HSLM-A1 Universal Train presented in the EC1. Only this bridge is selected for the sake of conciseness and for being the most influenced one by the ballast shear stiffness and damping properties. The acceleration response is calculated for the HSLM-A1 train in the range of velocities [40, 117] m/s (e.g. [144, 420] km/h) every 1 m/s at a quarter, mid-span and three quarters of both spans. A 3rd order Chebyshev filter is applied to the response in order to filter contributions below 1 Hz and above 60 Hz. Then, maximum response envelopes are obtained for each speed. The following individual variations of the track parameters are imposed: [0.0, 0.5, 1.0, 1.5, 2.0] $\cdot$ K_w and [0.5, 1.0, 1.5, 2.0] $\cdot$ C_w. Also, Rayleigh damping is assumed according to EC1 for pre-stressed concrete bridges as 1.7% for the GD-ES-10 bridge. This ratio is applied on the first and fifth natural frequencies.

In Fig. 4 (a-b), an envelope of the maximum acceleration response at the bridge deck is represented at the most critical section which corresponds to the center of the second span. The maximum acceleration level is not relevant as an unrealistically high design velocity is considered in order to capture low order and clear resonances of the bridge. Also, and in order to visualize how the variation of K_w and C_w affects the bridge response in different situations, the acceleration time-history at the same section is represented for three different velocities. In this way, the analysis is started with the second resonance speed of the first mode (e.g. $j = 2$, $n = 1$ in Eq. 2, according to [13]), which is equals to 380 km/h (see Fig. 4 (c-d)).

$$V_{nj}^r = \frac{d_k}{j\,T_n} = \frac{d_k\,\omega_n}{2\pi j} \tag{2}$$

In the previous equation, d_k stands for the characteristic distance of the HSLM-A1 train (18 m), T_n is the n-th natural period of the bridge and j the resonant order. Following that, the

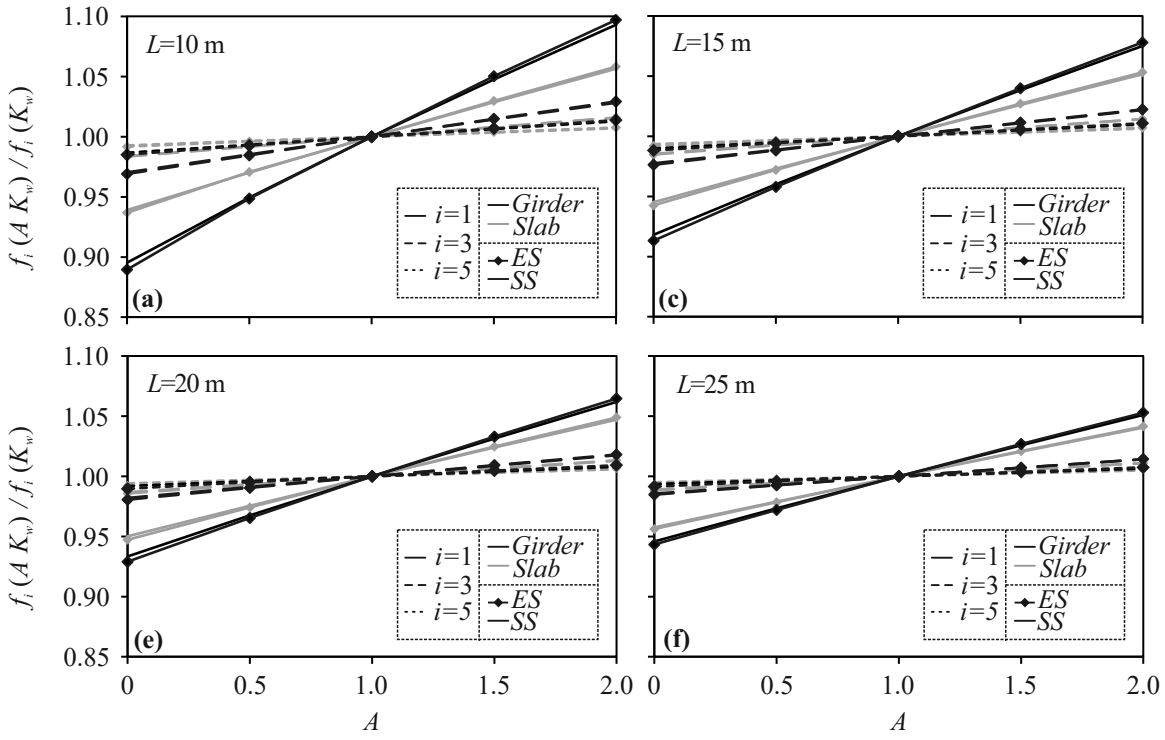


Figure 3: Influence of the variation of K_w in f_1, f_3 and f_5 with respect to the frequency in the nominal case.

response is computed at 324 km/h, far both from resonance and from cancellation of resonance (see Fig. 4 (e-f)). Finally, it is determined for a speed near a cancellation of resonance condition, given by Eq. 3, in agreement with [13]:

$$\left(\frac{L_{bi}}{d_k}\right)^c_{nji} = \left(\frac{\lambda_n}{n\pi}\right)^2 \frac{n}{2jK^c_{ni}}, \qquad n,j,i \geq 1 \tag{3}$$

In this way, when the relation L_{bi}/d_k between the length of each span and the characteristic distance of the train approaches the i-th cancellation ratio given by Eq. 3, the cancellation of the resonance is produced, and the vibration level gets significantly attenuated. For the case of the GD-ES-10 bridge associated to the circulation of the HSLM-A1 train, the third resonance speed of the first mode, equal to 253 km/h, approaches the first $(L_{bi}/d_k)^c_{nji}$ theoretical condition of cancellation for this resonance (e.g. $j = 3$, $n = 1$, $i = 1$, respectively), although it is not coincident (the difference is approximately 15%). Nevertheless, the phenomenon is visible, leading to a quite reduced resonant peak. These results are shown in Fig. 4 (g-h). In summary, the subsequent observations can be made:

- An increase in K_w leads to a rise in the resonant velocities, in the same proportion that the resonant frequency is modified by this parameter (in this particular case, neglecting or doubling K_w entails variations of -17.4% to +9.3% of the resonant velocity for the nominal case). This affects similarly different order resonances.
- For the range of K_w values considered, resonance at a certain speed may or may not take place depending on K_w (see Fig. 4(c-e)).
- Regarding the effect of the ballast shear damping, it is only relevant at resonance, leading to a pronounced reduction of the acceleration response. In this particular case, if C_w is

doubled with respect to its nominal value, the vertical acceleration reduces by a 26%. The effect of this parameter on the second resonant peak ($V = 380$ km/h) is much higher than the effect on the third one ($V = 253$ km/h). Nevertheless, this last peak is close to cancellation and no conclusions can be extracted in this regard.

- Finally, for the resonance speed approaching the cancellation conditions, a very significant attenuation of the acceleration level is observed with a small influence of the track parameters.

5 CONCLUSIONS

The longitudinal coupling effect exerted by the continuity of the ballasted track in single-track railway bridges composed by several isostatic consecutive spans is evaluated in this work. Specifically, the influence of the ballast shear stiffness and damping in the modal parameters and vertical acceleration under train passages is investigated. In the first place, a bridge catalogue considering short-to-medium span lengths and two common bridge deck typologies has been prepared. Then, a sensitivity analysis has been performed by means of a 2D FE track-bridge interaction model. Individual variations of the track parameters have been imposed in order to study their influence on the dynamic behaviour of the bridges. The main conclusions for this work are summarized as follows:

- In the discrete track model presented, the ballast shear stiffness and damping are the parameters that affect the most the bridge response in the frequency range of interest. The influence of the remaining parameters is negligible compared to these two.

- Regarding the modal parameters of the bridges, K_w exerts a notable influence on them, which is stronger in shorter bridges. When it comes to the typology, girder-deck bridges are the most affected due to their initially lower bending stiffness. The correlation with the flexibility of elastic supports is minor.

- With respect to the vertical acceleration level caused by the passage of a train, it is found that the effect of K_w and C_w is significant, especially at resonance. In particular, an increment of K_w leads to an important rise in resonant velocity, while an increment of C_w results into a reduction of the resonant acceleration amplitude. The effect of C_w far from resonance is negligible. These results are consistent, since, higher K_w values lead to an increase on the natural frequencies, especially of the fundamental one and in the case of short flexible structures.

- Future investigations are required in order to understand completely the influence of these shear parameters. It is also needed to find clear ways to determine their value, since their influence on the dynamic behaviour of railway bridges is significant and the information about it found in the literature is scarce. Experimentally appraised values for these parameters could be quite useful in the case of using discrete track models, which is a reasonable solution permitting solving the dynamic equations of motion in the time domain performing a full analysis in a reasonable amount of time.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support provided by the Junta de Andalucía and the European Social Fund through the contract USE-20617-D with the Universidad de Sevilla, and also the Spanish Ministry of Science and Innovation under research project PID2019-109622RB and Generalitat Valenciana under research project AICO2019/175.

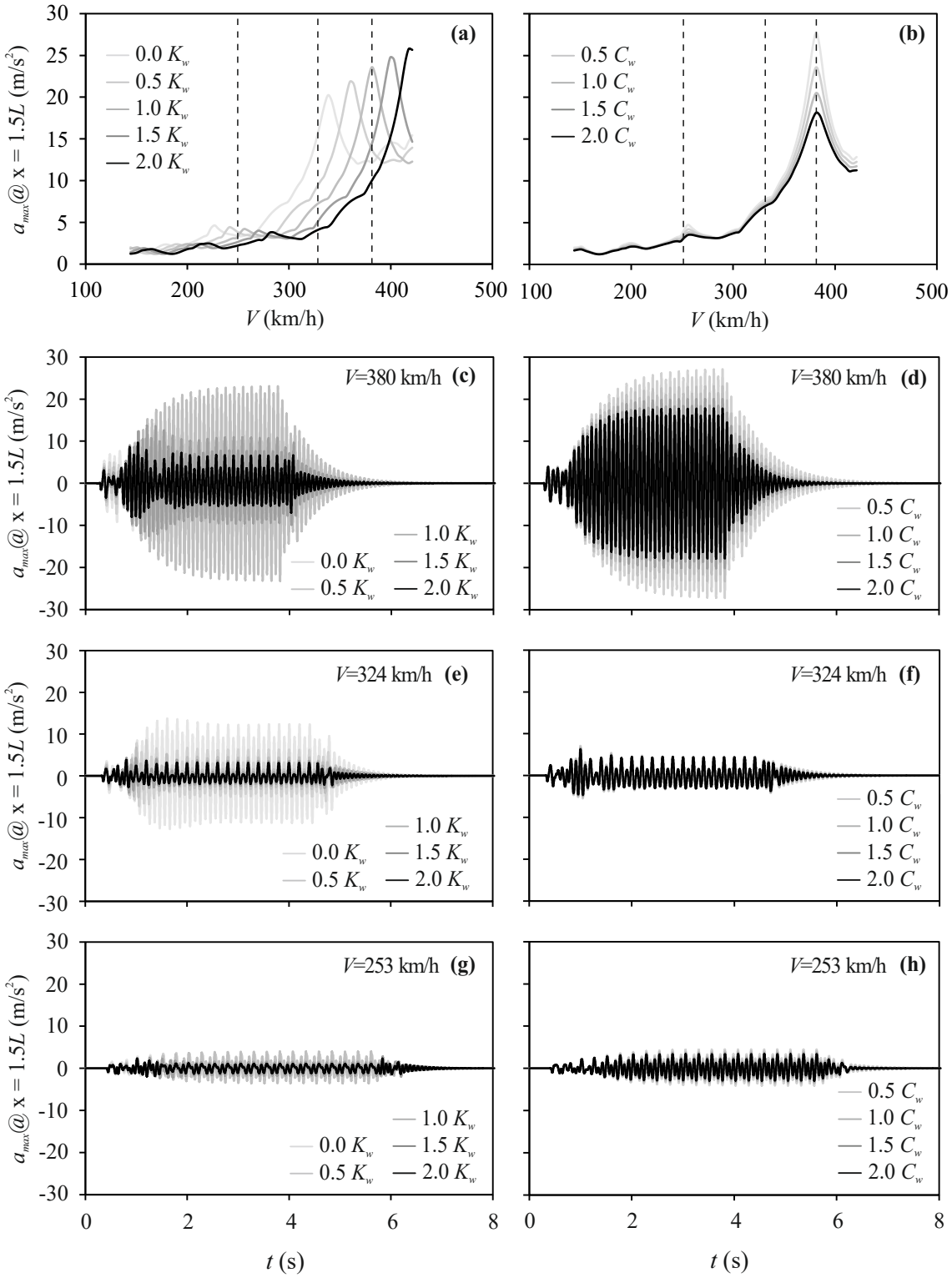


Figure 4: GD-ES-10 bridge. Maximum acceleration response for each velocity (a-b), and acceleration time-history at different speeds (c-h).

REFERENCES

[1] Frýba, L. Dynamic behaviour of bridges due to high-speed trains. *Bridges for High Speed Railways, CR, Press*, pp. 137-158, (2008).

[2] Zacher, M., Baeβler, M. Dynamic behaviour of ballast in railway bridges. *Dynamics of High-Speed Railway Bridges*, (2008).

[3] Rocha, J., Henriques, A., Calçada, R. Probabilistic safety assessment of a short span high-speed railway bridge. *Engineering structures*, Vol. **71**, pp. 99-11, (2014).

[4] Zhai, W., Wang, K., Lin, J. Modelling and experiment of railway ballast vibrations. *Journal of Sound and Vibration*, Vol. **270**, pp. 673-683, (2004).

[5] Doménech, A., Museros, P., Martínez-Rodrigo, M.D. Influence of the vehicle model on the prediction of the maximum bending response of simply-supported bridges under high-speed traffic. *Engineering Structures* Vol. **72**, pp. 123-139, (2014).

[6] CEN, EN 1991-2. Eurocode 1: Actions on Structures - Part 2: Traffic loads on bridges. *European Committee for Standardization*, Brussels, (2003).

[7] Martínez-Rodrigo, M.D., Moliner, E., Romero, A., Galvín, P. Maximum resonance and cancellation phenomena in orthotropic plates traversed by moving loads: Application to railway bridges. *International Journal of Mechanical Sciences* Vol. **169** 105316, (2020).

[8] Lou, P. A vehicle-track bridge interaction element considering vehicle's pitching effect. *Finite Elements in Analysis and Design* Vol. **41** 41, pp. 397-427, (2005).

[9] Galvín, P., Romero, A., Moliner, E., De Roeck, G., Martínez-Rodrigo, M.D. On the dynamic characterisation of railway bridges through experimental testing. *Engineering Structures* Vol. **226** (1) 111261, (2020).

[10] CEN/TC256, EN 13674-1:2011+A1:2017 Railway applications – Track – Rail – Part 1: Vignole railway rails 46 kg/m and above. *European Committee for Standardization*, Brussels, (2017).

[11] Nguyen, K., Goicolea, J., Gabaldón, F. Comparison of dynamic effects of high-speed traffic load on ballasted track using simplified two-dimensional and full three-dimensional model. *Journal of Rail and Rapid Transit* Vol. **228** (2), pp. 128-142, (2012).

[12] Ministerio de Fomento, Gobierno de España. *Instrucción de acciones a considerar en puentes de ferrocarril*, Actions in railway bridges (in spanish), (2010).

[13] Museros, P., Moliner, E., Martínez-Rodrigo, M.D. Free vibrations of simply- supported beam bridges under moving loads: Maximum resonance, cancellation and resonant vertical acceleration. *Journal of Sound and Vibration* Vol. **332**, pp. 326-345, (2013).

INFLUENCE OF TRACK MODELLING IN MODAL PARAMETERS OF RAILWAY BRIDGES COMPOSED BY SINGLE-TRACK ADJACENT DECKS

J.C. Sánchez-Quesada*†,E. Moliner†, A. Romero‡, P. Galvín‡ and M.D. Martínez-Rodrigo†

† Department of Mechanical Engineering and Construction
Universitat Jaume I
Castellón, Spain

‡ Escuela Técnica Superior de Ingeniería
Universidad de Sevilla
Sevilla, Spain

e-mail: jquesada@uji.es, molinere@uji.es, aro@us.es, pedrogalvin@us.es, mrodrigo@uji.es

Key words: railway bridges, vertical acceleration, track-bridge interaction, ballasted track, resonance

Abstract: *A significant number of railway bridges composed by simply-supported (SS) spans are present in existing railway lines. Special attention must be paid to short to medium span length structures, as they are prone to experience high vertical acceleration levels at the deck, due to their low weight and damping, compromising the travelling comfort and the structural integrity. The accurate prediction of the dynamic response of these bridges is a complex issue since it is affected by uncertain factors such as structural damping and complex interaction mechanisms such as vehicle-bridge, soil-structure or track-bridge interaction.*

Concerning track-bridge interaction, experimental evidences of a dynamic coupling exerted by the ballasted track between subsequent SS spans and also between structurally independent single-track twin adjacent decks have been reported in the literature. Nevertheless, this phenomenon is frequently disregarded due to the computational cost of models including the track and due to the uncertainties in the mechanical parameters that define the track system.

The present work contributes to the study of the coupling effect exerted by the ballasted track in railway bridges composed by SS adjacent decks. With this purpose a 3D finite element (FE) track-bridge interaction model is implemented with a continuous representation of the track components meshing the sleepers, ballast and sub-ballast with solid FE.

The numerical model is updated with experimental measurements performed on an existing railway bridge in a view to evaluate (i) the influence of the track continuity on the bridge modal parameters and (ii) the adequacy of the implemented numerical model.

1 INTRODUCTION

The ballasted track in railway bridges distributes the axle loads from the rails to the structure, acts as a high-frequency filter and introduces a restraining effect at the end sections [1]. In addition, experimental evidences of load and vibration transfer mechanisms between consecutive spans or adjacent SS decks sharing a continuous ballasted track have been reported over the last years [1, 2]. A vast description of different ballast models developed by researchers in the analysis of train-induced vibrations may be consulted in Reference [3]. These models fall into two main categories: discrete and continuous models. In discrete models the rail displacement is connected to the bridge deck through a set of spring, damper and lumped mass elements generally defined at the sleepers positions that represent the stiffness, damping and mass of the different track components (sleepers, railpads, ballast and sub-ballast), while the

https://doi.org/10.4995/YIC2021.2021.12283

rail is modelled as a continuous beam. From this basis, 2D and 3D representations of the track-bridge interaction have been reported in the scientific literature for different applications [4, 5]. Discrete models are conceptually simple and require less computational effort than continuous models. However they neglect potentially relevant aspects, such as the bending coupling between consecutive SS spans associated to the separation of the ballast and rails from the centre of gravity of the deck cross section and the contribution of the ballast bed to the global deck stiffness.

Continuous track models permit considering the composite action between the track and the bridge associated to the transmission of shear stress between the deck and the rails through the ballast. In these models, the ballast is generally considered as a continuum and is discretised into solid FE [6], admitting elastic and isotropic constant material properties. Additionally, in the ballast regions located at the joints between consecutive spans or decks, a few researchers propose the use of degraded material properties to take into account the possible loss of stiffness of the ballast due to the cyclic movement caused by passing trains [7]. In these works, the degradation is accomplished by reducing the elastic modulus of the general ballast. More refined techniques including the heterogeneous and granular nature of the ballast, such as the Discrete Element Method, are applied for the analysis of settlement and degradation under cyclic loading [8] but they require enormous computational resources which make them unfeasible for application.

The models for ballasted tracks require a significant number of parameters which are highly uncertain. Therefore, a better understanding of their influence is needed in bridge engineering to develop more realistic and adequate numerical tools that, at the same time, do not fall into inadmissible computational costs. In this regard the performance of experimental campaigns on bridges, the development of appropriate calibration methodologies and the experimental-numerical validation becomes crucial. However, the number of reported field measurements performed on multi-span SS viaducts or bridges composed by adjacent decks only coupled through the ballast is scarce to derive general conclusions. Rebelo et al. [1] performed experimental tests on some single-span ballasted railway bridges composed by two adjacent single-track slabs and pointed out the existence of a coupling effect exerted by the shared ballast, which was especially relevant in skewed decks.

In the present work, the authors analyse the coupling effect of the ballasted track taking as starting point Old Guadiana Bridge, a representative railway bridge from a conventional railway line in Spain. The bridge is composed by two identical SS spans and two structurally independent but adjacent single-track decks. A clear dynamic coupling between the spans attributable to the track continuity, and also between the adjacent decks through the shared ballast layer was detected during experimental tests [9]. This work aims to assess the extent of track-bridge interaction effects in such bridges and the key parameters affecting the dynamic coupling between structurally independent parts. With this purpose a 3D FE model is implemented. A degraded type of ballast with elastic anisotropic behaviour is assumed for the regions between subsequent spans or adjacent decks. Finally, the model is updated to reproduce the modal properties identified experimentally.

2 BRIDGE DESCRIPTION

The structure under study is a double-track bridge that belongs to the conventional railway line Madrid-Alcázar de San Juan-Jaén in Spain. It is composed by two identical SS spans of 11.93m length between supports centres. The horizontal structure is formed by two adjacent but structurally independent single-track decks. Each deck is made of a reinforced concrete slab of 0.25 m thickness resting on five pre-stressed concrete girders. The decks are weakly

connected along their longitudinal border through the ballast. Each track is conformed by Iberian gauge UIC60 rails and mono-block concrete sleepers separated 0.60 m [10]. A total ballast thickness "h_b" of 0.45 m is assumed. Underneath the sleepers, the ballast thickness is 0.34 m in accordance with current Spanish regulations [11]. The bridge substructure consists of two external abutments and one central support, and the girders rest on them through laminated rubber bearings.

On May 2019 the response of the bridge was measured to characterise the modal parameters and the dynamic response under operating conditions. Eighteen accelerometers were installed underneath the girders and the vertical response was measured under ambient vibration and several train passages. The accelerometers were installed at points 1 to 18 as indicated in Fig. 1. For details of the experimental campaign the reader is referred to Reference [9].

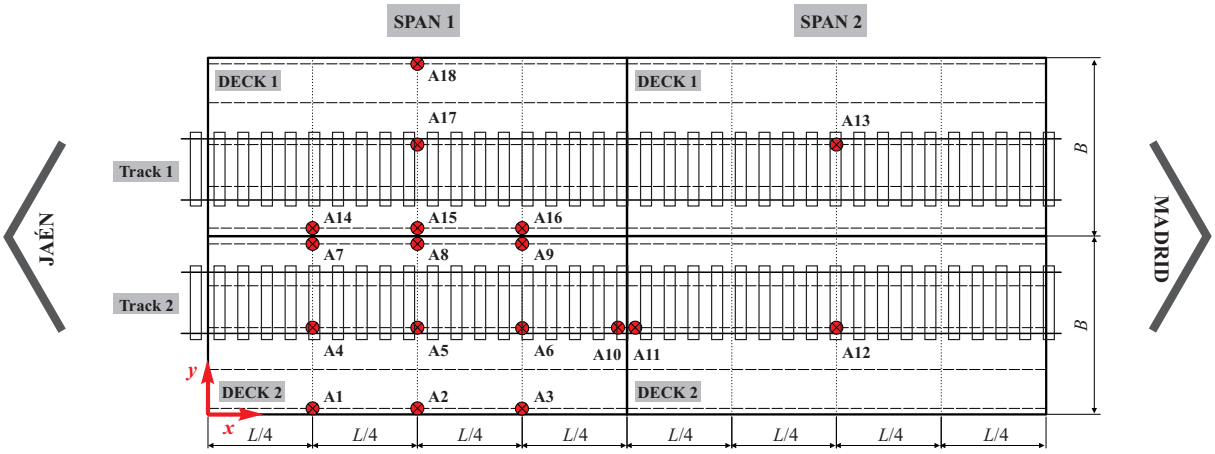


Figure 1: Sensors layout.

Notice that the number of sensors installed was limited, especially in the second span. As a result, the five modes shown in Fig. 3 in red trace (lines and crosses) were identified. The lowest one in frequency order corresponds to the first longitudinal bending mode of each span. The second mode was associated to the combined first torsion mode. In the third mode, the two adjacent decks conform a first transverse bending mode. The fourth and fifth modes were identified as in-phase torsional deformation mode and to the transverse bending mode of each deck, respectively. Fig 3 also provides their natural frequencies (f^{exp}) and modal damping ratios identified from ambient vibration (ξ^{exp}).

3 NUMERICAL MODEL

A 3D continuous track-bridge interaction model of the complete bridge is implemented in ANSYS. The model includes the structure and 15 m of track extension over the embankment before and after the bridge (Fig. 2). The slabs and girders of the bridge are discretised with shell FE, while the laminated rubber bearings with solid FE. The elastic modulus of the bearings was previously calibrated in order to reproduce the experimental static deflection measured during the load test proof of the bridge performed in 2005 [12]. Concerning the track plattform, the sleepers, ballast and subgrade layer are modelled with solid FE. For the rails, Timoshenko beam FE are used, which are connected to the sleepers thorugh the rail pads, considered as discrete spring-dashpot elements. Finally, the handrails are included as lumped masses uniformly distributed along the two external borders of the deck.

An optimisation iterative procedure implemented in ANSYS-MATLAB is performed to minimize an objective function which involves the differences in the predicted and measured natural frequencies and MAC values for the five modes identified from ambient vibration. Based on a

preliminary sensitivity analysis, the optimisation parameters were selected. Successive model samples are generated from variations of these parameters within reasonable limits with respect to nominal values extracted from the project information, scientific literature and current standards. Table 2 shows the main model parameters used in the numerical idealisation of Old Guadiana bridge. Among them, the selected optimisation parameters are those for which variation ranges are provided. In the cited table the following nomenclature is used: E, ν and ρ stand for the elastic modulus, Poisson's ratio, and mass density, respectively. Also, X, Y and Z refer to the longitudinal direction (parallel to the track), transverse and vertical directions, respectively. Concerning the track components, the spring-dashpot discrete properties of the rail pads (K_p and C_p) are provided. The main ballast presents isotropic elastic properties (E_b and ν_b identical in the three directions). The degraded ballast behaviour is considered as transversely isotropic material with elastic constants expressed as E_{bI}, G_{bIJ} and ν_{bIJ}, where I and J refer to the spatial directions X, Y and Z. This material is unequivocally defined by five independent constants:

$$E_{bX} = E_{bY} \qquad E_{bZ} \qquad G_{bXZ} = G_{bYZ} \qquad \nu_{bXY} \qquad \nu_{bXZ} = \nu_{bYZ} \tag{1}$$

In Eq. 1, $E_{bX} = E_{bY}$ are the in-plane elastic moduli, E_{bZ} and $G_{bXZ} = G_{bYZ}$ the out-of-plane elastic and shear moduli, respectively, and ν_{bXY} and $\nu_{bXZ} = \nu_{bYZ}$ the Poisson's ratios.

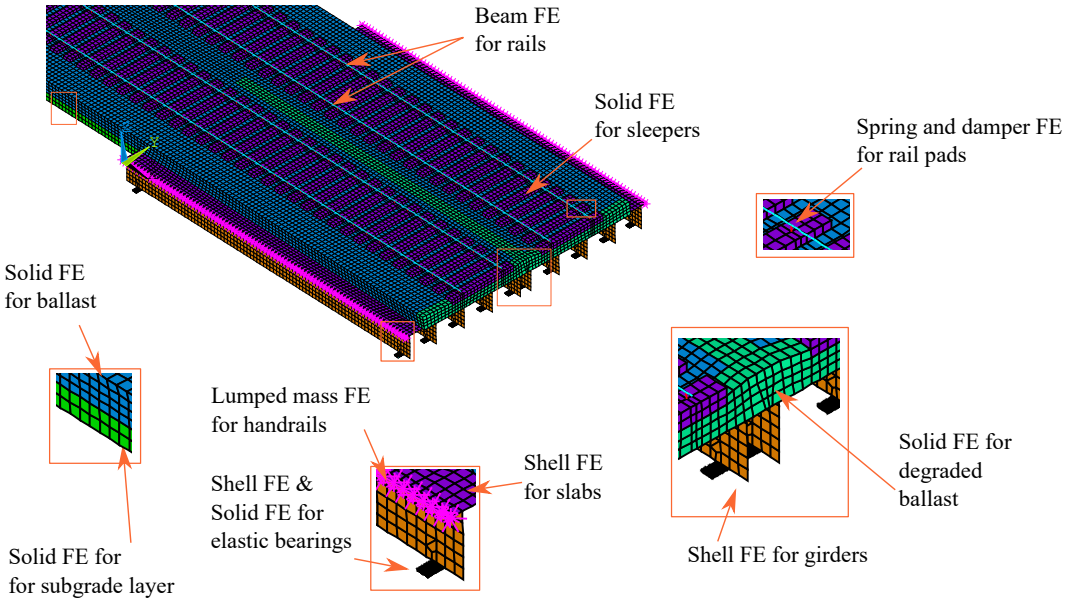


Figure 2: 3D numerical model: detail of one span and track extension.

The experimental and paired numerical mode shapes of the calibrated model are represented in Fig. 3, where the numerical frequency f^{num} is also provided for each paired mode. Table 1 shows the results of the model calibration in terms of frequency differences, calculated as $e_{100\%} = (f^{exp} - f^{num})/f^{exp} \times 100$ and MAC numbers.

Table 1: Frequency differences and MAC numbers of the paired modes after calibration.

Mode (i)	1	2	3	4	5
$e_{100\%}$ [-]	0.47	-3.17	5.37	-2.05	-9.75
MAC [-]	0.94	0.89	0.97	0.93	0.75

As can be seen in Fig. 3, the second and third modes, which are more affected by the continuity of the ballasted track between adjacent decks, are predicted with frequency differences

lower than 5.5%. Also their MAC numbers exhibit a satisfactory correlation with the measurements, with values close to 0.90 or even above. As it is shown, the correspondence of the fifth numerical mode with the experimental measurements is less accurate, even though very reasonable considering the limited number of sensors available in the experimental campaign for the identification of high frequency modes.

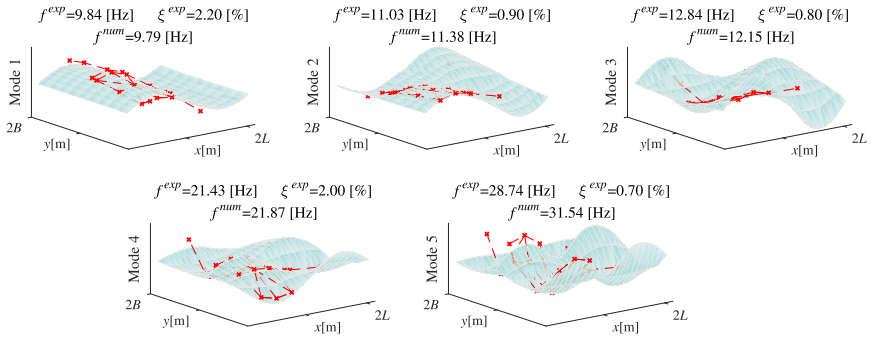


Figure 3: Experimental modes identified (lines path) and calibrated solution (surface)

4 SENSITIVITY ANALYSIS

4.1 Evolution of modal parameters with the thickness of the ballast layer

In this subsection the influence of the thickness of the ballast layer on the modal numerical-experimental correspondence is evaluated. Fig. 4 shows the MAC and $e_{100\%}$ values for the paired numerical modes under variations of the ballast thickness h_b in the range [0.3 – 0.7]m. Experimental frequencies and modal shapes (Fig. 3) are always used as reference values in what follows. The same thickness is assumed for both the main and the degraded ballast regions, based on in situ observations. The rest of the model parameters are kept unmodified and equal to their final updated values. In order to be able to separate the effect of the added mass and the added stiffness that the increase of h_b entails, two different results are provided. First, the total ballast mass is kept invariable, therefore, as h_b increases, the ballast density is modified accordingly and only the extra stiffness affects the results (dashed trace); Secondly, as h_b increases, the ballast density is kept unmodified and equal to its updated final value (and the ballast mass increases proportionally), i.e, both the ballast added mass and stiffness are taken into account (continuous trace). In the plots a black dashed horizontal line indicates a zero difference between the numerical and experimental natural frequencies and a black vertical dash-dot line points out the calibrated value of the elastic property.

From the analysis of the previous figures it can be observed that the fundamental frequency is the one most affected by h_b, leading to a reduction of the numerical frequency as a consequence of the added mass. Its mode shape alteration is negligible as it also is the contribution of the ballast added stiffness. The second (first torsion) mode evolution follows a similar trend, but the numerical frequency reduction is smaller and the effect of the added stiffness is higher when compared to the previous mode. It can also be observed that the natural frequency of the third (transverse bending) mode increases with the thickness of the ballast layer due to the predominant effect of the added stiffness, however its mode shape is only slightly modified. The influence of the ballast thickness on the fourth (second torsion) mode is negligible, but the MAC number reduces remarkably. Finally, the fifth (second transverse bending) mode frequency reduces slightly as the height of the ballast layer increases, and its mode shape remains unaltered.

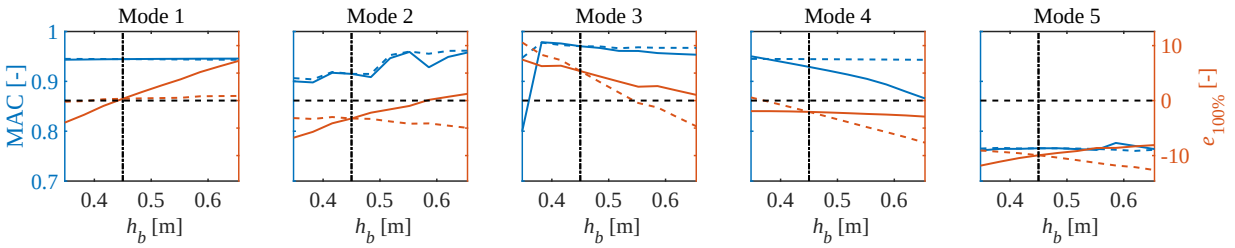


Figure 4: Modal parameters variation in terms of ballast thickness h_b. Dashed trace: constant ballast mass; Continuous trace: constant ballast density.

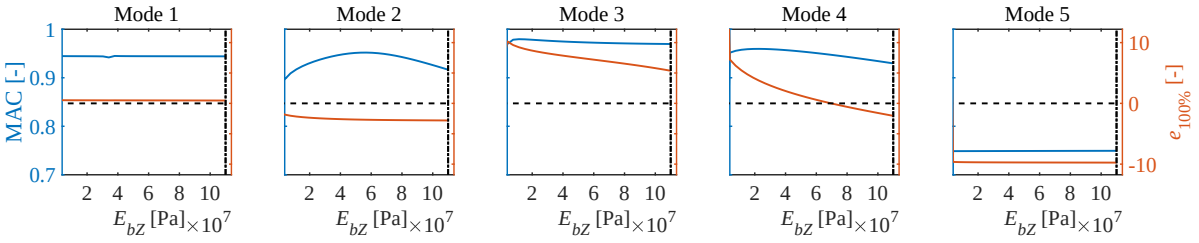
4.2 Evolution of modal parameters with degraded ballast elastic properties

Finally, a sensitivity analysis is performed in order to analyse the influence of the degraded ballast parameters on the frequency differences and MAC numbers of the paired modes. The parameters that are investigated are the elastic modulus in the vertical direction (Fig. 5), the elastic modulus in longitudinal and transverse directions (Fig. 6) and the independent shear moduli (Fig. 7). These parameters are modified and applied to the degraded ballast along the longitudinal shared border between the adjacent decks and along the transverse shared border between the two spans, separately. In all the cases, the model parameters are kept equal to their final updated value, except for the one that is modified.

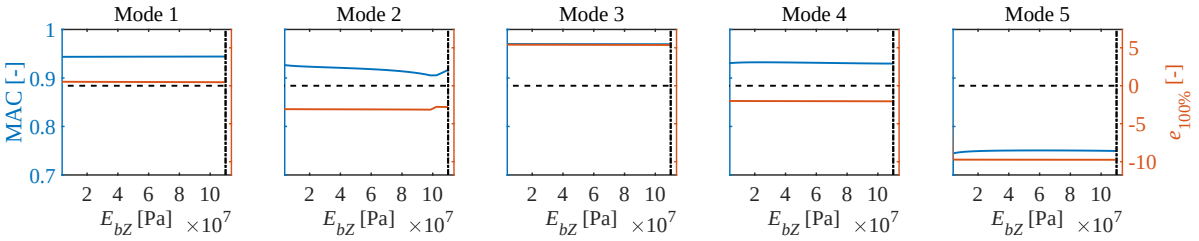
In Fig. 5a the MAC and frequency difference $e_{100\%}$ are represented versus E_{bZ} of the degraded ballast along the shared border between the adjacent decks that conform each bridge span, for the five paired modes. In Fig. 5b, the same quantities are represented but the degraded ballast property E_{bZ} is modified only along the transverse border between the spans. Similarly, in Figs. 6a and 6b and in Figs. 7a and 7b the same type of representations are included for the elastic modulus in the horizontal directions $E_{bX} = E_{bY}$ and for the shear modulus in the XZ and YZ planes $G_{bXZ} = G_{bYZ}$, respectively. As in Fig. 4, black dashed horizontal and vertical lines indicate, respectively, zero frequency difference and calibrated value of the model parameter.

From the analysis of the previous figures it can be concluded that the first longitudinal bending mode is the one least affected by the degraded ballast elastic properties. For an acceptable calibration of the frequency of the first torsion mode the elastic modulus in the horizontal directions $E_{bX} = E_{bY}$ must be substantially lower than the vertical elastic modulus E_{bZ} (no higher than 20% of E_{bZ}), both along the longitudinal and the transverse borders. The MAC of the torsion mode is the most affected by the value of the shear modulus in the XZ and YZ planes, $G_{bXZ} = G_{bYZ}$. Both the MAC number and frequency difference for this mode evolve in a similar way for variations of this parameter along both the longitudinal and the transverse borders. The third (first transverse bending) mode is the one most affected by the degraded ballast elastic properties. In this case the frequency difference increases with the reduction of E_{bZ} between the adjacent decks. This effect is also observed for the fourth mode. Nevertheless, the influence of this parameter is not very significant. A minimum value of the elastic modulus in the horizontal directions $E_{bX} = E_{bY}$, both between the adjacent decks and consecutive spans is needed to reproduce the experimental third mode, opposite to what happens with the torsion mode. For values higher than 4×10^7 Pa along the longitudinal border, the model becomes too rigid and the frequency difference is unacceptable. That is not the case for the degraded ballast between the spans. As for the shear modulus $G_{bXZ} = G_{bYZ}$, the value of this parameter does not affect much the MAC of the third mode but the frequency correspondence improves with the increase of this property. As for the fourth (second torsion) mode the influence of the degraded ballast elastic properties between the two spans is almost negligible. In this case the ballast zone affecting the most is $G_{bXZ} = G_{bYZ}$ of the degraded ballast between adjacent decks.

Both, the MAC and frequency difference reduce as this parameter increases. For an accurate prediction of the bridge behaviour a compromise should be found regarding the shear modulus between the second, third and fourth modes, as it affects in opposite ways the modal residuals. Finally, the fifth (second transverse bending) is not affected by either the vertical modulus of elasticity E_{bZ} or the horizontal one $E_{bX} = E_{bY}$ along either of the two edges of the deck. The only parameter affecting this mode is the shear modulus $G_{bXZ} = G_{bYZ}$ of the degraded ballast between spans, which increase leads to a slight improvement in the MAC number.

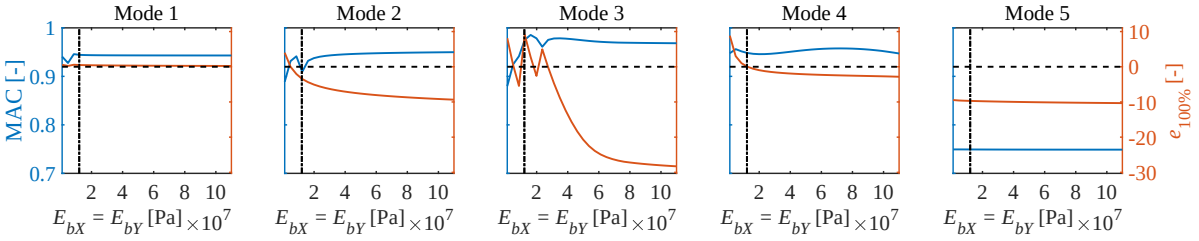


(a) E_{bZ} variations along the longitudinal border between adjacent decks.

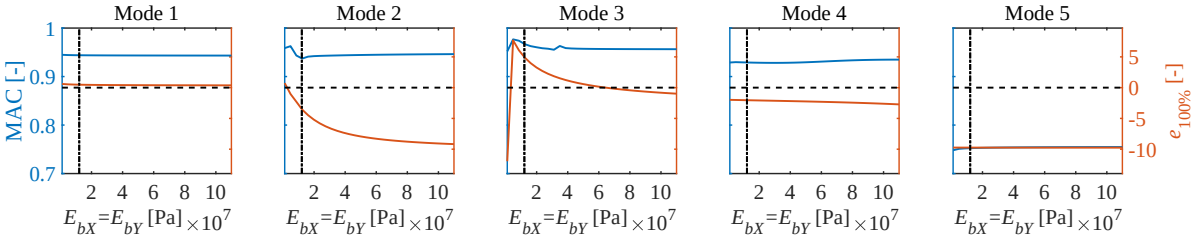


(b) E_{bZ} variations along the transverse border between spans.

Figure 5: Influence of E_{bZ} on the natural frequencies and MAC values.

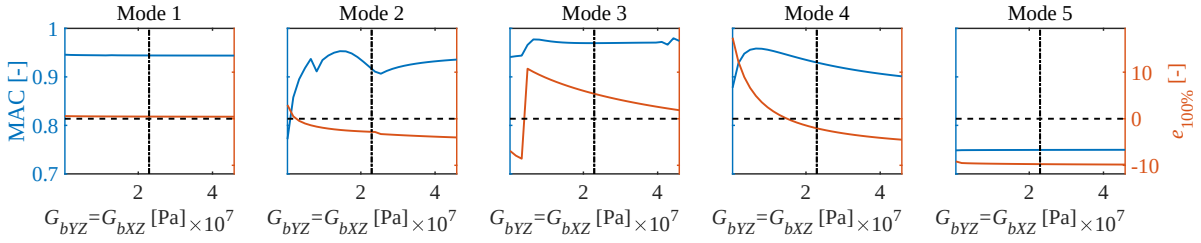


(a) $E_{bX} = E_{bY}$ variations along the longitudinal border between adjacent decks.

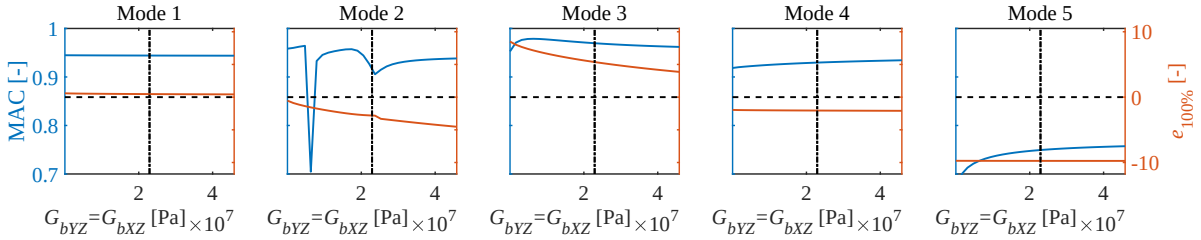


(b) $E_{bX} = E_{bY}$ variations along the transverse border between spans.

Figure 6: Influence of $E_{bX} = E_{bY}$ on the natural frequencies and MAC values.



(a) $G_{bXZ} = G_{bYZ}$ along the longitudinal border between adjacent decks.



(b) $G_{bXZ} = G_{bYZ}$ variations along the transverse border between spans.

Figure 7: Influence of $G_{bXZ} = G_{bYZ}$ on the natural frequencies and MAC values.

Table 2: Initial and final properties value for track and bridge superstructure.

Bridge component	Notation	Initial value	Calibration range	Calibrated value	Unit
Rail pad	K_p	$1.00 \cdot 10^8$	-	$1.00 \cdot 10^8$	N/m
	C_p	$7.50 \cdot 10^4$	-	$7.50 \cdot 10^4$	Ns/m
Ballast	h_b	0.34	-	0.34	m
	E_b	$1.10 \cdot 10^8$	-	$1.10 \cdot 10^8$	Pa
	ν_b	0.3	-	0.3	-
	ρ_b	1800	$[-30, 30]\%$	1584	kg/m^3
Degraded ballast	E_{bX}=E_{bY}	$1.10 \cdot 10^8$	$[-89, 0]\%$	$12.10 \cdot 10^6$	Pa
	E_{bZ}	$1.10 \cdot 10^8$	-	$1.10 \cdot 10^8$	Pa
	G_{bYZ}=G_{bXZ}	$4.58 \cdot 10^7$	$[-89, 0]\%$	$2.29 \cdot 10^7$	Pa
	ν_{bXY}=ν_{bYX}	0.2	-	0.2	-
	ν_{bXZ}=ν_{bYZ}	0.2	-	0.2	-
	ρ_b	1800	$[-30, 30]\%$	1584	kg/m^3
Handrail	m_b	50	-	50	kg/m
Girders	E_g	$3.60 \cdot 10^{10}$	$[-30, +45]\%$	$4.82 \cdot 10^{10}$	Pa
	ν_g	0.3	-	0.3	-
	ρ_g	2500	$[-30, +30]\%$	2504	kg/m^3
Slabs	E_s	$3.60 \cdot 10^{10}$	$[-30, +35]\%$	$3.10 \cdot 10^{10}$	Pa
	ν_s	0.3	-	0.3	-
	ρ_s	2500	$[-40, +40]\%$	2480	kg/m^3
Elastic bearings	E_{eb}	$2.39 \cdot 10^8$	-	$2.39 \cdot 10^8$	Pa
	ν_{eb}	0.2	-	0.2	-
	ρ_{eb}	1230	-	1230	kg/m^3

5 CONCLUSIONS

In this work the dynamic response of railway bridges composed by SS spans and adjacent single-track decks weakly connected through the ballasted track is investigated. The main aim is to assess the extent of track-bridge interaction effects in such bridges and the key parameters affecting the dynamic coupling between structurally independent parts. With this purpose a 3D FE model is implemented. A degraded type of ballast with elastic anisotropic behaviour is assumed at the regions between subsequent spans or adjacent decks to consider the potential degradation of the ballast due to the relative vertical movements under train passages. The model is updated with experimental results and the main ballast properties affecting the decks coupling are identified and evaluated by means of sensitivity analyses. The following conclusions are derived:

- The updated numerical model is able to reproduce the first five natural frequencies and mode shapes identified experimentally with an average error in the frequencies close to 4% and an average MAC of 0.9, and with a remarkably good correspondence in the particular case of the first longitudinal bending, third transverse bending and fourth second torsion modes.

- In order for the model to reproduce experimental modes higher than the second one, it is essential to consider the coupling effect of the ballast layer, especially between the adjacent decks.

- The predicted natural frequencies and mode shapes are not affected by the degraded ballast elastic modulus between spans in the vertical direction, E_{bz}.

- The first longitudinal bending mode is the one least affected by the degraded ballast elastic properties. In order to obtain a good prediction of the second (first torsion) mode natural frequency the ballast elastic moduli $E_{bX} = E_{bY}$ should be significantly smaller than the vertical elastic modulus E_{bz}.

- The third mode (first transverse bending mode) is the one most affected by the degraded ballast elastic properties. A minimum value of the elastic modulus in the horizontal directions $E_{bX} = E_{bY}$ is needed to reproduce the experimental third mode.

- As per the fourth mode, the most relevant parameter is the shear modulus in the vertical planes $G_{bXZ} = G_{bYZ}$ of the degraded ballast between the decks. Both, the MAC and frequency error reduce as this parameter increases.

- The fifth mode is only affected by the shear modulus $G_{bXZ} = G_{bYZ}$ between the two spans. The MAC number improves as it becomes stiffer.

- Regarding the thickness of the ballast layer, the added stiffness associated to a thicker ballast layer does not affect the fundamental mode. This effect is particularly relevant in the case of the third mode, leading to an important increase in its natural frequency.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support provided by the Spanish Ministry of Science and Innovation under research project PID2019-109622RB; FEDER Andalucía 2014-2020 Operational Program for project US-126491; Generalitat Valenciana and Universitat Jaume I under research projects AICO2019/175 and UJI/A2008/06; and the Andalusian Scientific Computing Centre (CICA).

REFERENCES

[1] Rebelo, C. and Simões da Silva, L. and Rigueiro, C. and Pircher, M. Dynamic behaviour of twin single-span ballasted railway viaducts - Field measurements and modal identification. *Engineering Structures*. Vol. **30**, (9), pp. 2460–2469, (2008).

[2] Galvín, P. and Romero, A. and Moliner, E. and De Roeck, G. and Martínez-Rodrigo, M.D. On the dynamic characterisation of railway bridges through experimental testing. *Engineering Structures*. Vol. **226** (111261).

[3] Zhai, W.M. and Han, Z.L. and Chen, Z.W. and Ling, L. and Zhu, S.Y. Train-track-bridge dynamic interaction: a state-of-the-art review. *Vehicle System Dynamics*. Vol. **57**, (7), pp. 984–1027, (2019).

[4] Zhu, Z.H. and Gong, W. and Wang, L.D. and Bai, Y. and Yu, Z.M. and Zhang, L. Efficient assessment of 3D train-track-bridge interaction combining multi-time-step method and moving track technique. *Engineering Structures*. Vol. **183**, pp. 290–302, (2019).

[5] Moliner, E. and Martínez-Rodrigo, M.D. and Galvín, P. and Romero, A. On the vertical coupling effect of ballasted tracks in multi-span simply-supported railway bridges under operating conditions. *International Journal of Mechanical Sciences*.

[6] Malveiro, J. and Ribeiro, D. and Sousa, C. and Calçada, R. Model updating of a dynamic model of a composite steel-concrete railway viaduct based on experimental tests. *Engineering Structures*. Vol. **164**, pp. 40–52, (2018).

[7] Bonifácio, C. and Ribeiro, D. and Calçada, R. and Delgado, R. Dynamic behaviour of a short span filler-beam railway bridge under high-speed traffic. *In: Proc. 2nd Int. Conf. on Railway Technology: Research, Development and Maintenance*, 2014.

[8] Guo, Y. and Zhao, C. and Markine, V. and Jing, G. and Zhai, W.M. Calibration for discrete element modelling of railway ballast: A review. *Transportation Geotechnics*. Vol. **23**, 100341, (2020).

[9] Galvín, P. and Romero, A. and Moliner, E. and De Roeck, G. and Martínez-Rodrigo, M.D. On the dynamic characterisation of railway bridges through experimental testing. *Engineering Structures.*, Vol. **226**, 111261, (2021).

[10] IAPF - Instrucción de acciones a considerar en puentes de ferrocarril, 2010. Ministerio de Fomento, Gobierno de España.

[11] IF3 - Instrucción para el proyecto y construcción de obras ferroviarias, 2015. Ministerio de Fomento, Gobierno de España.

[12] CITEF. *"Informe sobre resultado de registros en puente sobre río Guadiana p.k. 160.000"*. Línea Madrid-Cádiz, tramo Alcázar de San Juan-Manzanares.

On the calculation of the Kalker's creep coefficients for non-elliptical contact areas

J. Gómez-Bosch*, J. Giner-Navarro* and J. Carballeira*

*Instituto Universitario de Ingeniería Mecánica y Biomecánica
Universitat Politècnica de València
Valencia, Spain
e-mail: jorgobos@upvnet.upv.es, juanginer@upv.es and jacarmo@mcm.upv.es

Key words: Contact Mechanics, Railway Dynamics, Tangential Contact Problem, FastSim, Non-Hertzian Contact

Abstract: *FastSim is the most widely used tangential contact method due to its accuracy and computational efficiency. However, its use is limited to elliptic contact areas, as it needs results from Kalker's Linear Theory, a Hertzian contact theory, to obtain the so-called elastic parameters. This makes FastSim unable to face some of the current railway challenges, such as wear, corrugation, Rolling Contact Fatigue (RCF), wheel flats, etc. Taking this limitation into account, in the present work, an alternative methodology to Kalker's Linear Theory is proposed, which will enable FastSim to deal with non-Hertzian conditions.*

1 INTRODUCTION

Solving the tangential wheel-rail contact problem is always complex. Depending on the application, a trade-off between accuracy and computational cost is required. The most accurate tangential contact model existing is CONTACT [1, 2], but, because of its high computational cost, it is mainly used as a reference theory for validation. In railway dynamics simulation, simplified contact theories [3, 4, 5, 6] are usually required. These theories are much more computationally efficient, although they are less accurate. Among all the simplified theories, the most widely used is FastSim [6], due to its high-level performance and accuracy, and its ability to predict tangential stresses distribution and the stick-slip boundary [7]. FastSim is a contact theory that assumes that the surface displacements on a point are only related to the tangential stress on that point through the so-called elastic parameters [8]. To obtain these parameters, creep forces resulting from the full adhesion solution [8] (simplified contact theory which assumes adhesion over the entire contact area) are equalled to the ones obtained through Kalker's Linear Theory (KLT) [9] (exact contact theory but limited to Hertzian contact conditions). It is this limitation which has led various authors to find alternative methods to obtain these elastic parameters under non-Hertzian contact conditions [8, 10, 11], and so, to be able to extend FastSim validity to non-elliptic contact areas. The elastic parameter calculation under non-Hertzian condition has been carried out according to two different approaches [11]: a first approach, based on associating the contact area to one or several equivalent ellipses [12, 13]; and a second approach, in which, for each particular contact geometry, elastic parameters are obtained by solving the exact contact problem [8, 10]. Despite the methods based on the second approach are quite more accurate, their computational cost is much higher than the ones based on the first approach. That is the reason why equivalent ellipse based methods are used nowadays to study the influence of the non-Hertzian contact in actual railway vehicle dynamics [14, 15], as well as in complex tangential contact phenomena, such as wear [16, 17], Rolling Contact Fatigue (RCF) [18], wheel out-of-roundness [19, 20], etc.

In the present work, an alternative tangential contact model to KLT which allows the elastic parameters calculation under non-Hertzian hypothesis is proposed. This model derives from Kalker's Variational Theory [8], to which steady-state and full adhesion hypothesis are imposed.

https://doi.org/10.4995/YIC2021.2021.12313

Since the exact contact is solved to obtain the elastic parameters using this model, it is included within the second approach described above. Nevertheless, this model presents some advantages compared to existing alternatives [8, 10], as its resolution is not iterative, nor stress solutions are approximated to any polynomic function. These advantages make this model more suitable to face the new railway challenges, as the ones listed above.

The mathematical model of this work is developed in Section 2. In Section 3, the accuracy of the proposed contact model is analysed, when results are compared to the ones obtained with KLT on elliptic contact areas. Finally, in Section 4, according to presented results, the contribution of this work is concluded and justified.

2 MATHEMATICAL MODEL

In the present work, a mobile reference frame $\mathbf{X}_1\mathbf{X}_2\mathbf{X}_3$ is assumed, with origin at the theoretical contact point, and it moves with it as the vehicle travels along the track. $\mathbf{X}_1$ axis is parallel to the rolling direction, $\mathbf{X}_3$ axis is normal to the contact, being positive to the wheel, and $\mathbf{X}_2$ axis corresponds to the lateral direction in order to form a right-handed rectangular frame, as it is shown in Figure 1.

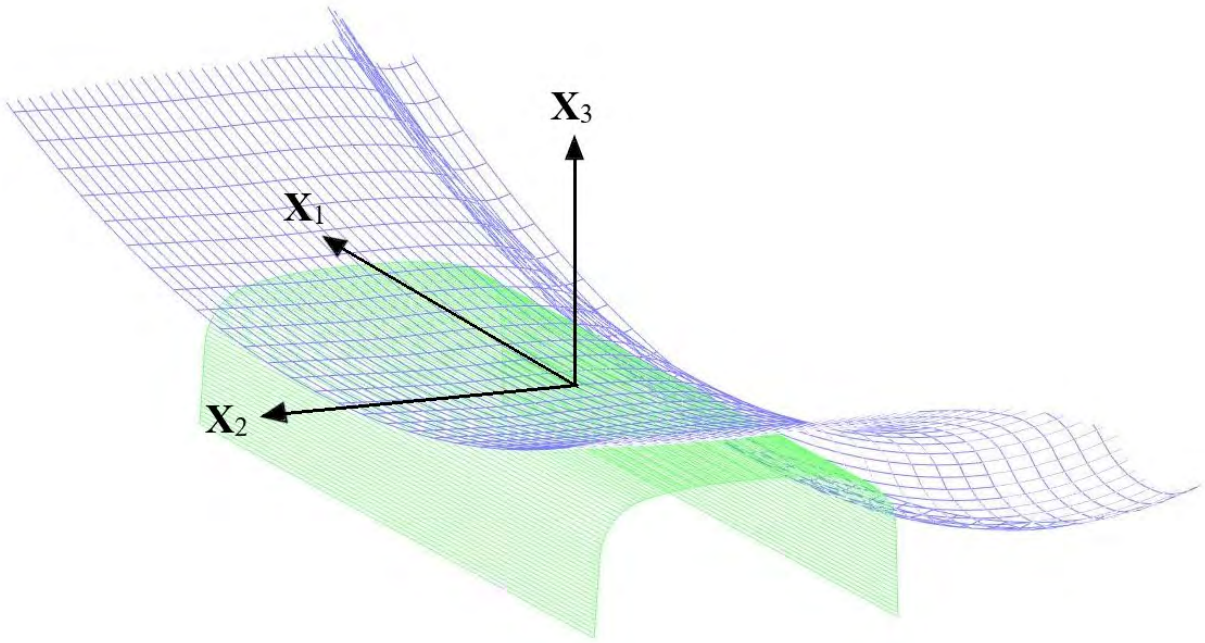


Figure 1: Mobile reference frame $\mathbf{X}_1\mathbf{X}_2\mathbf{X}_3$ at the theoretical contact point between rail (green) and wheel (blue).

As it is done in Kalker's Variational Theory [8], the kinematic equation that relates the rigid body displacements of the bodies in contact with the slip velocities and the elastic deformations can be written

$$\mathbf{s} = \mathbf{w} + 2\frac{D\mathbf{u}}{Dt} = \mathbf{w} + 2\frac{\partial \mathbf{u}}{\partial t} + 2V\frac{\partial \mathbf{u}}{\partial x_1}, \tag{1}$$

where $\mathbf{s}$ are the local slip velocities, $\mathbf{u}$ are the displacements related to the elastic deformation of the bodies in contact, V is the vehicle speed, and $\mathbf{w}$ are the velocities associated to the

undeformed configuration, which can be calculated from the creepages [8]. Assuming the steady-state hypothesis $\left(\frac{\partial}{\partial t}=0\right)$ and the full adhesion hypothesis ($\mathbf{s}=0$), Eq. (1) re-writes

$$\mathbf{w}=-2V\frac{\partial \mathbf{u}}{\partial x_1}. \tag{2}$$

Including the constitutive relationships in Eq. (2), it is possible to obtain an expression, which provides tangential stresses under steady-state and full adhesion conditions $\bar{p}_\tau$:

$$\mathbf{w}=-2V\left(\int_S \frac{\partial \mathbf{c}_1(\mathbf{x},\mathbf{y})}{\partial x_1}\bar{p}_1(\mathbf{y})+\frac{\partial \mathbf{c}_2(\mathbf{x},\mathbf{y})}{\partial x_1}\bar{p}_2(\mathbf{y})\right)ds(\mathbf{y}), \tag{3}$$

where $\mathbf{c}_1(\mathbf{x},\mathbf{y})$ and $\mathbf{c}_2(\mathbf{x},\mathbf{y})$ are two vectors that contain the elastic influence functions, and S is the contact surface. To solve Eq. (3), the contact area is discretised analogously as it is done in the TANG algorithm [8], assuming constant stresses on each element. For the j-th element of the mesh, Eq. (3) writes

$$\mathbf{w}^j=-2V\mathbf{C}^j\bar{\mathbf{p}}, \tag{4}$$

where $\mathbf{C}^j$ is the vector which contains the elastic influence coefficients derivatives, and $\bar{\mathbf{p}}$ is the column vector which contains tangential stresses under adhesion conditions of every element of the mesh. Figure 1 shows a scheme of the mesh used for solving Eq. (4), where a and b are half the size of the element on longitudinal and lateral directions, respectively. This equation is solved by a collocation method [1, 21], where the location of the collocation point can be controlled with a parameter α, which takes values in the range $[-1,1]$.

Once the tangential stresses have been obtained, the tangential contact forces can be obtained by summation of these stresses. As it is assumed that every element on the contact area is under adhesion, the tangential stresses and forces will be linear with creepages. By analogy with KLT, tangential forces under adhesion conditions $\bar{F}_\tau$ can be written as

$$\bar{F}_1=-f_{11}^*\xi \tag{5}$$

$$\bar{F}_2=-f_{22}^*\eta-\ f_{23}^*\phi, \tag{6}$$

where f_{11}^*, f_{22}^* and f_{23}^* are the analogous coefficients to the creep coefficients f_{11}, f_{22} and f_{23} defined by Kalker in Ref. [9]; and ξ, η and ϕ are longitudinal, lateral and spin creepages, respectively.

3 RESULTS

Using KLT creepage coefficients as a reference, it is possible to study the influence of the collocation point and the number of elements N on the accuracy of the results provided by the proposed method, for different ellipse axes ratio, r. Figure 2 shows the ratio between the creepage coefficients obtained by the proposed method and the ones obtained from KLT as a function of the location of the collocation point, for a mesh size of $N=6400$. Results for coefficients f_{11}^* and f_{22}^* are quite similar: the optimum collocation point is located at the centre of the element. Instead, to achieve higher accuracy on the f_{23}^* coefficient, it is convenient to move the collocation point forward to a value of parameter $\alpha=0.5$.

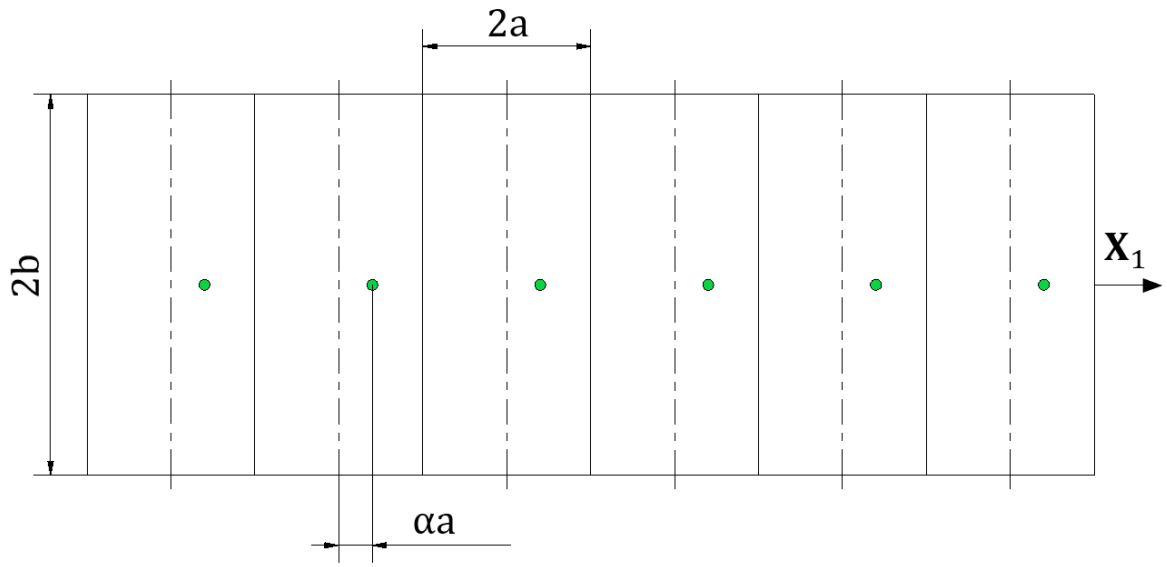


Figure 2: Mesh scheme with collocation points (green dots).

Figure 3 shows the evolution of the creepage coefficients ratio as a function of the number of elements of the mesh, N, for a collocation point at the centre of the element. According to these results, despite the ratio is close to 1 for a sufficient number of elements, thus being the error acceptable, the method does not present convergence. Assuming the full adhesion hypothesis leads to infinite tangential stress at the trailing edge of the contact area, which produces numerical errors, and the non-convergence of the method, as it is also concluded in Ref. [22].

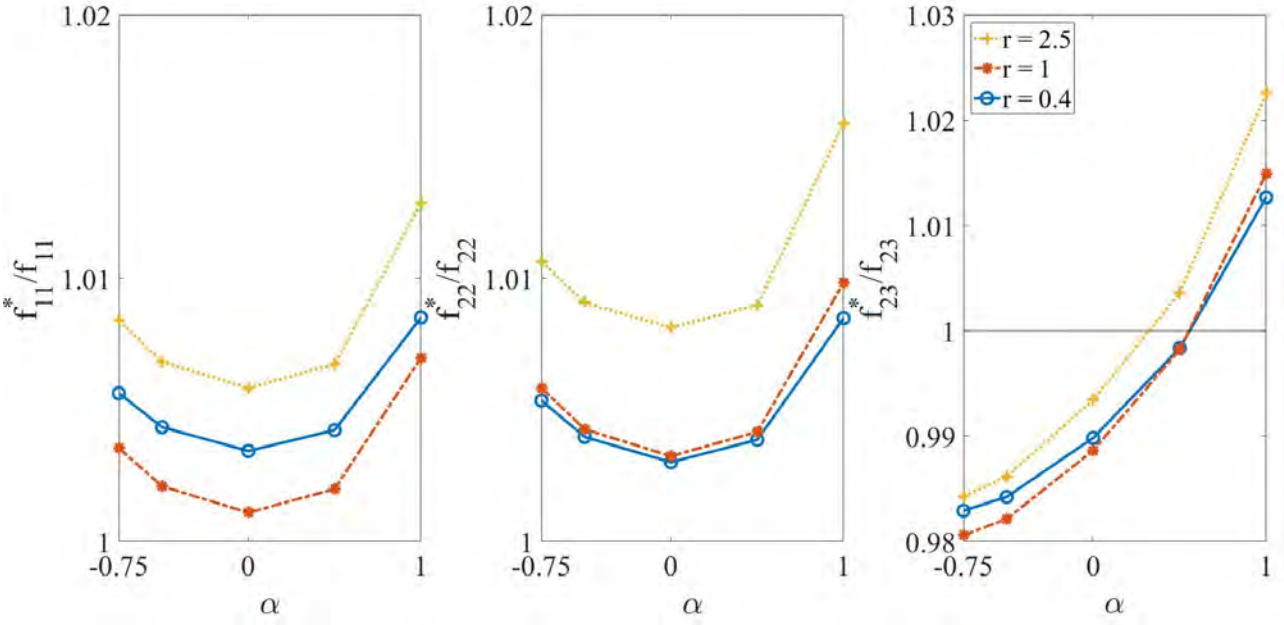


Figure 3: Creep coefficients ratio as a function of the collocation parameter α for three different ellipse ratios r. The number of elements of the mesh is $N = 6400$.

As coefficients f_{11}^* and f_{22}^* are the most relevant for tangential forces calculation, optimum collocation point is located at the centre of the element. The optimum mesh size is conditioned by the computational cost. To solve Eq. (4), it is needed to invert a $2N \times 2N$ matrix, so that, increasing the mesh size, exponentially increases the calculation time. Increasing mesh size from $N = 60 \times 60$ to$N = 80 \times 80$ elements implies four times more calculation time, but only a reduction of creep coefficients absolute error calculation of 0.2%. Therefore, as $N = 60 \times 60$ is the smallest mesh size with absolute errors on f_{11}^* and f_{22}^* calculations below 1%, authors propose an optimum mesh size of $N = 60 \times 60$ elements.

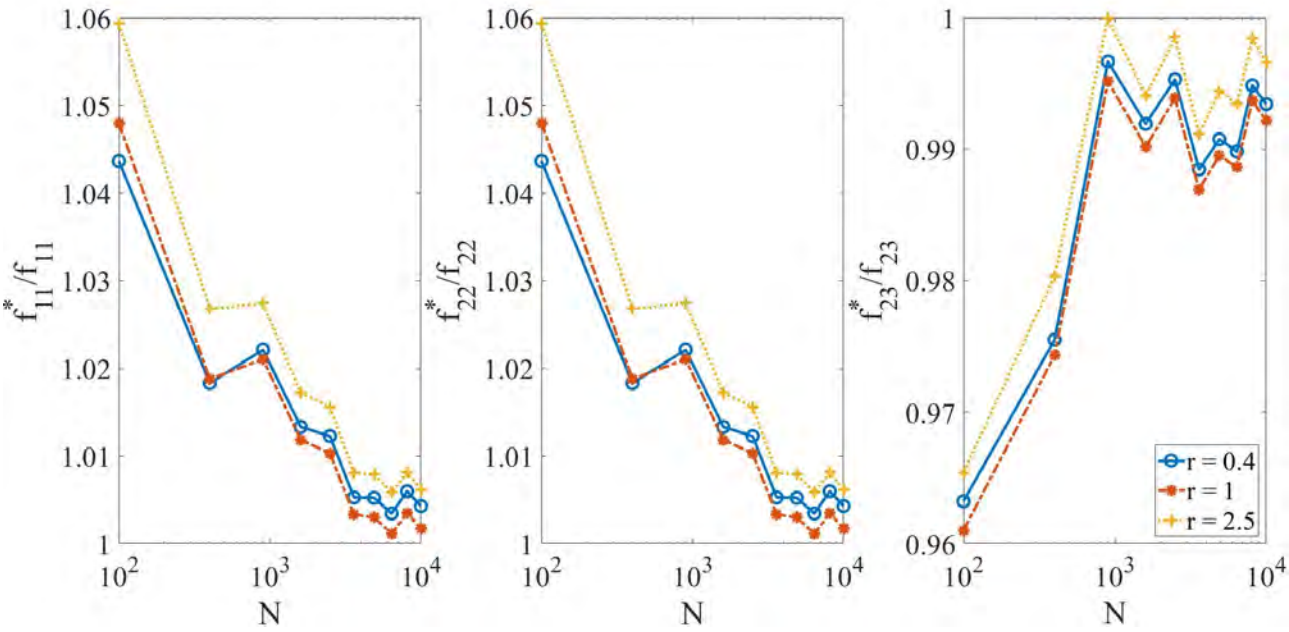


Figure 4: Creepage coefficients ratio as a function of the number of elements of the mesh N for three different ellipse ratios r. The collocation point is located at the centre of the element.

4 CONCLUSION

The FastSim algorithm is limited to elliptical contact areas because of the calculation of the elastic parameters based on Kalker's Linear Theory results. In this work, an alternative model has been proposed to deal with that restriction, allowing the calculation of the creep coefficients for non-Hertzian contact conditions, which can be used to obtain the elastic parameters according to the FastSim methodology. Based on results shown in this work, it has been proved that, combining both optimum collocation point ($\alpha = 0$) and mesh size ($N = 60 \times 60$ elements), sufficient to minimize the numerical error associated with the full adhesion hypothesis assumption, without considerably increasing the computational calculation time, the present model gives fairly accurate results on creep coefficients calculation for elliptic contact areas, without assuming Hertzian contact hypothesis. So, on future research, this method will be used together with FastSim to obtain results on tangential forces and stress distributions on non-Hertzian contact conditions.

AKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of Agencia Estatal de Investigación and European Regional Development Fund (grant PRE2018-084067 and project TRA2017-84701-R).

REFERENCES

[1] J. J. Kalker, "The computation of three-dimensional rolling contact with dry friction," International Journal for Numerical Methods in Engineering, vol. 14, no. 9, pp. 1293-1307, 1979.

[2] J. J. Kalker, Users Manual of the Fortran Program CONTACT, Delft: Delft University of Technology, Departament of Math. and Computer Science, 1986.

[3] K. L. Johnson, "The effect of spin upon the rolling motion of an elastic sphere on a plane," Journal of Applied Mechanics, vol. 25, pp. 332-338, 1958.

[4] P. J. Vermeulen and K. L. Johnson, "Contact of Nonspherical Elastic Bodies Transmitting Tangential Forces," Journal of Applied Mechanics, vol. 31, no. 2, pp. 338-340, 1964.

[5] O. Polach, "A Fast Wheel-Rail Forces Calculation Computer Code," Vehicle System Dynamics, vol. 33, no. 1, pp. 728-739, 1999.

[6] J. J. Kalker, "A fast algorithm for the simplified theory of rolling contact," Vehicle System Dynamics, vol. 11, no. 1, pp. 1-13, 1982.

[7] M. S. Sichani, R. Enblom and M. Berg, "An alternative to FASTSIM for tangential solution of the wheel-rail contact," Vehicle System Dynamics, vol. 54, no. 6, pp. 748 - 764, 2016.

[8] J. J. Kalker and K. L. Johnson, Three-Dimensional Elastic Bodies in Rolling Contact, Delft: ASME. J. Appl. Mech., 1993.

[9] J. J. Kalker, On the rolling contact of two elastic bodies in the presence of dry friction, T. H. Delft: Thesis, 1967.

[10] K. Knothe and L. T. Hung, "A method for the analysis of the tangential stresses and the wear distribution between two elastic bodies of revolution in rolling contact," Solids Structures, vol. 21, no. 8, pp. 889-906, 1985.

[11] M. S. Sichani, R. Enblom and M. Berg, "Comparison of non-elliptic contact models: Towards fast and accurate modelling of wheel-rail contact," Wear, vol. 314, no. Issues 1-2, pp. 111-117, 2014.

[12] J. Piotrowski and W. Kik, "A simplified model of wheel/rail contact mechanics for non-Hertzian problems and its application in rail vehicle dynamic simulations," Vehicle System Dynamics, vol. 46, no. 1-2, pp. 27-48, 2008.

[13] J. B. Ayasse and H. Chollet, "Determination of the wheel rail contact patch in semi-Hertzian conditions," Vehicle System Dynamics, vol. 43, no. 3, pp. 161-172, 2005.

[14] B. Liu and S. Bruni, "Comparison of wheel-rail contact models in the context of multibody system simulation: Hertzian versus non-Hertzian," Vehicle System Dynamics, pp. 1 - 21, 2020.

[15] Q. Guan, B. Liu and S. Bruni, "Effects of Non-Hertzian Contact Models on Derailment Simulation," in Proceedings of the 2020 Joint Rail Conference. 2020 Joint Rail Conference., St. Louis, Missouri, USA, 2020.

[16] G. Tao, Z. Wen, X. Zhao and X. Jin, "Effects of wheel-rail contact modelling on wheel wear simulation," Wear, Vols. 366 - 367, pp. 146 - 156, 2016.

[17] M. Meacci, Z. Shi, E. Butini, L. Marini, E. Meli and A. Rindi, "A railway local degraded adhesion model including variable friction, energy dissipation and adhesion recovery," Vehicle System Dynamics, pp. 1 - 22, 2020.

[18] S. Hossein-Nia, M. S. Sichani, S. Stichel and C. Casanueva, "Wheel life prediction model - an alternative to the FASTSIM algorithm for RCF," Vehicle System Dynamics, vol. 56, no. 7, pp. 1051 - 1071, 2018.

[19] G. Tao, Z. Wen, G. Chen, Y. Luo and X. Jin, "Locomotive wheel polygonisation due to discrete irregularities: simulation and mechanism," Vehicle System Dynamics, vol. 59, no. 6, pp. 872 - 889, 2021.

[20] U. Spangenberg, "Variable frequency drive harmonics and interharmonics exciting axle torsional vibration resulting in railway wheel polygonisation," Vehicle System Dynamics, vol. 58, no. 3, pp. 404 - 424, 2020.

[21] J. G. Giménez, A. Alonso and L. Baeza, "Precision analysis and dynamic stability in the numerical solution of the two-dimensional wheel/rail tangential contact problem," Vehicle System Dynamics, vol. 57, no. Issue 12, pp. 1822-1846, 2019.

[22] A. Alonso and J. G. Giménez, "Tangential problem solution for non-elliptical contact areas with the FastSim algorithm," Vehicle System Dynamics, vol. 45, no. 4, pp. 341-357, 2007.

SIMULATION OF THE CONTACT WIRE WEAR EVOLUTION IN HIGH SPEED OVERHEAD CONTACT LINES

S. Gregori*, J. Gil*, M. Tur*, A. Pedrosa* and F.J. Fuenmayor*

* Instituto de Ingeniería Mecánica y Biomecánica (I2MB)
Universitat Politècnica de València
Valencia, Spain
e-mail: sangreve@upv.es, jaigiro@upv.es, manuel.tur@mcm.upv.es, anpedsan@dimm.upv.es, ffuenmay@mcm.upv.es

Key words: Wear, Contact wire, Pantograph, Railway catenary.

Abstract: *The overhead contact line or catenary is the structure composed of support elements and wires responsible for the power supply of the locomotive through sliding contact with the pantograph. This contact causes wear not only on the pantograph contact strips but also in the contact wire, which produces a reduction on its effective section and eventually its replacement, resulting in the stoppage of the rolling stock with its associate economical and operational drawbacks. For this reason, it is important for catenary designers to count with appropriate tools able to predict the contact wire wear behaviour for extending the service life of the system. This work proposes a strategy to simulate the long-term contact wire wear evolution considering the mutual influence between the dynamic behaviour and wear of the system. The method is based on two pillars: the efficient simulation of the catenary-pantograph dynamic interaction and a heuristic wear model which considers mechanical wear due to friction and electrical wear produced by Joule effect and electric arcs. With the proposed simulation tool, we analyse the effect on the long-term contact wire worn height of the train speed.*

1 INTRODUCTION

Power supply in electric trains is usually carried out by the sliding contact between the overhead contact line or catenary and the contact strips of the pantograph. As shown in Figure 1a, the catenary is composed of the contact wire, which is held by droppers from the messenger wire. All the cabling is regularly supported by brackets attached to posts. By means of steady arms, the catenary is arranged in a zig-zag shape. The pantograph mechanism (see Figure **??**) is mounted on the roof of the locomotive. Powered by a pneumatic system, the mechanism unfolds and the contact strips push against the contact wire.

This sliding contact produces wear in both the contact wire and contact strips. While the latter are relatively easy and cheap to replace, the substitution of a worn contact wire requires a higher investment and the stoppage of the rolling stock. For this reason, it is important to establish a correct maintenance strategy that predicts when you will need to replace the contact wire. To this end, numerical simulations can be an interesting tool not only to predict wear but also to help catenary designers to develop catenaries with longer service life.

Some authors have proposed different models to compute the normal wear rate (NWR) of the contact wire, which is defined as the volume of material removed in a kilometre of the wire. Specifically, heuristic wear models fitted by experimental measurements [1, 2] and models based on the Lim-Ashby wear maps [3, 4] are the most representative. Other research is focused on the experimental measurement of the contact wire thickness variation along successive years [5, 6] and the simulation of the pantograph-catenary dynamic interaction with the worn contact wire height profile. The main conclusions reveal that the greater the wear, the greater the oscillations in the contact force.

In this work, we propose a simulation strategy to predict the long-term contact wire wear

https://doi.org/10.4995/YIC2021.2021.12566

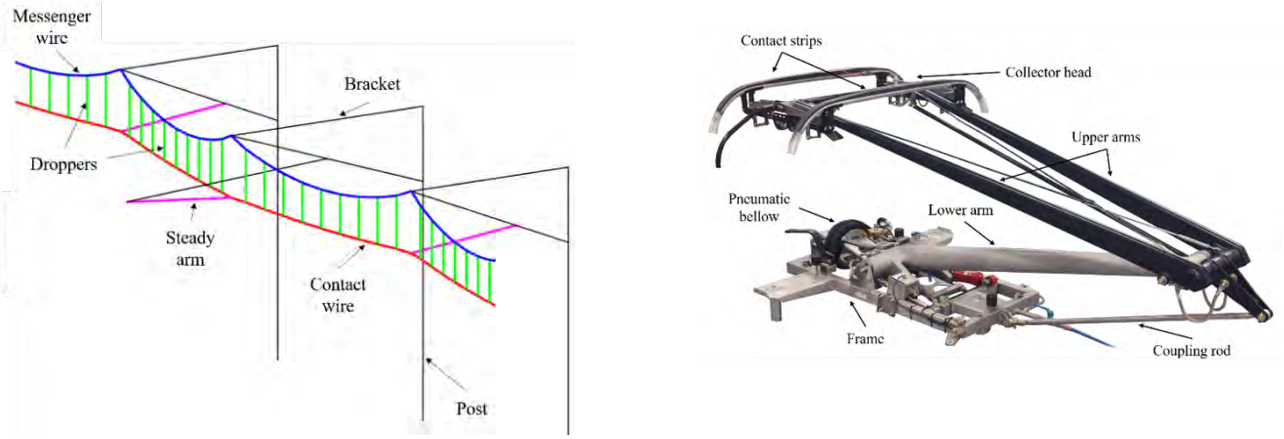


Figure 1: Main components a railway catenary (a) and pantograph (b)

evolvement considering the mutual influence between the dynamic behaviour of the system and the worn contact wire height profile. Starting from an unworn contact wire, this method allows to foresee when will be necessary to replace the contact wire and which are the most critical spots in which severe wear appears. The influence of the train velocity on the contact wire wear is also investigated.

2 LONG-TERM CONTACT WIRE WEAR SIMULATION

This section is devoted to give an overall view of the strategy proposed to compute the contact wire wear evolution which follows the flow diagram shown in Figure 2. Each of the steps involved in the procedure will be detailed in Section 3. The procedure starts by solving the initial configuration problem in which the nodal coordinates along with the element lengths are obtained for the Finite Element model of the nominal catenary with an unworn contact wire. From this point, a simulation loop is repeated until a given stopping criteria is reached, such as a certain percentage of the wire section is worn. The first step within this loop consists on solving the pantograph-catenary dynamic interaction. The main output obtained from this calculation is the contact force F_c between the pantograph and catenary contact wire. This force feeds the wear model to obtain the normal wear rate (NWR), which is the amount of area removed from each point of the contact wire due to wear. The removed section is then converted to an equivalent height following geometrical relations. At this step, as the contact wire section has changed, the total mass of the contact wire has decreased and therefore, it is needed to compute a static equilibrium problem to obtain the new position of the catenary and particularly the new contact wire height z_{cw}.

3 STAGES OF THE PROPOSED METHODOLOGY

This section is devoted to give a detailed insight of the models considered and the assumptions made in each of the stages that compose the proposed algorithm.

3.1 Initial configuration problem

In first place, the catenary model must be initialised. The Finite Element Method (FEM) with Absolute Nodal Coordinates Formulation (ANCF) elements is chosen to model the catenary cabling. The shape-finding problem consists of finding the nodal coordinates $\boldsymbol{q}$ and the undeformed element length $\boldsymbol{l_0}$ that fulfil both equilibrium equations and design constraints.

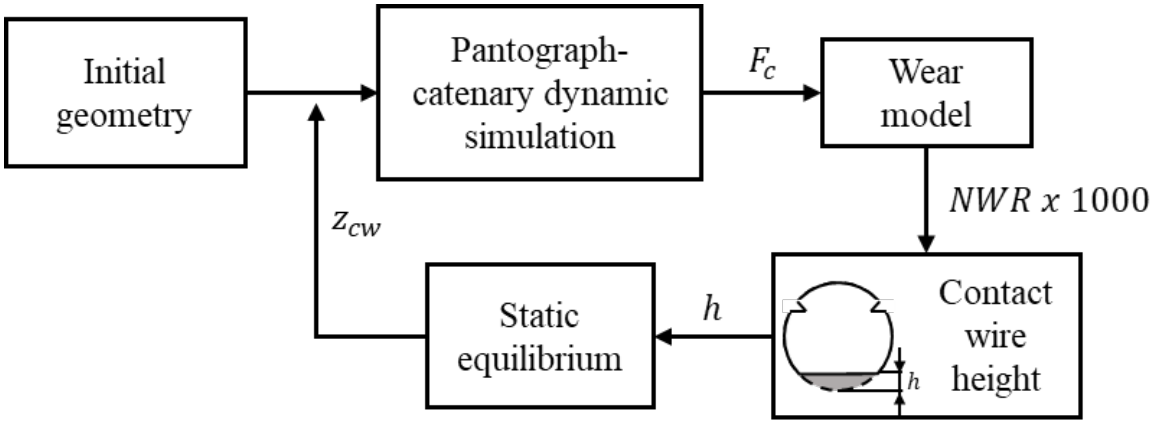


Figure 2: Flow diagram of the long-term contact wire wear simulation

The resultant non-linear problem is:

$$\begin{aligned} \boldsymbol{F_{int}}(\boldsymbol{q}, \boldsymbol{l_0}) + \boldsymbol{F_g}(\boldsymbol{l_0}) &= \boldsymbol{0} \\ \boldsymbol{C}(\boldsymbol{q}, \boldsymbol{l_0}) &= \boldsymbol{0} \end{aligned} \tag{1}$$

in which, $\boldsymbol{F_{int}}$ is the vector of internal forces, $\boldsymbol{F_g}$ is the vector of gravitational forces and $\boldsymbol{C}$ denotes for the design constraints such as tension of contact and messenger wires or position of dropper and stitch wire connection points. The interested reader is referred to [7] for a deeper explanation of the catenary initial configuration problem.

3.2 Pantograph-catenary dynamic interaction

The next step consists of solving the pantograph-catenary dynamic interaction problem. A lumped-mass model has been chosen to model the pantograph and the penalty method is used to consider the interaction between the pantograph and the contact wire.

This dynamic problem is governed by the following equation:

$$\boldsymbol{M}\ddot{\boldsymbol{u}} + \boldsymbol{C}\dot{\boldsymbol{u}} + \boldsymbol{K}\boldsymbol{u} = \boldsymbol{F} \tag{2}$$

in which, $\boldsymbol{M}$, $\boldsymbol{C}$ and $\boldsymbol{K}$ are the mass, Rayleigh damping and stiffness matrices respectively, $\boldsymbol{F}$ is the vector of external forces and $\boldsymbol{u}$, $\dot{\boldsymbol{u}}$ and $\ddot{\boldsymbol{u}}$ are the nodal displacement, velocity and acceleration vectors, respectively. This problem is also ruled by two nonlinearities, namely dropper slackening and pantograph contact losses.

For the simulation of the long-term evolvement of the contact wire wear, this dynamic problem must be solved hundreds of times. Thus, it is important to choose an efficient algorithm to perform the time integration of Eq. (2) with as low computation effort as possible. In this case, the fast algorithm proposed in [8] has been fully adopted.

3.3 Contact wire wear model

In this work we use the wear model of the copper contact wire proposed in [1]. This model differentiates three contributions on the total wear: (i) mechanical wear due to friction, (ii) electrical wear due to Joule effect of the current flow at the contact area and (iii) wear produced by electrical arcs when contact loss occurs. These three contributions to the contact wire wear are directly related to the three terms present in Eq. (3).

$$NWR = k_1 \left(\frac{1}{2}(1 + \frac{I_c}{I_0})\right)^{-\alpha} \left(\frac{F_c}{F_0}\right)^{\beta} \frac{F_c}{H} + k_2 \frac{R_c I_c^2}{Hv}(1-u) + k_3 \frac{V_a I_c}{v H_m \rho} u \tag{3}$$

The *NWR* represents the worn section and it is given in mm^2. The main factors that determine the wear rate are the electric current I_c, the sliding speed v and the contact force F_c. In this work we assume that F_c remains unaltered during a given number of pantograph passages in which it is not necessary to solve the dynamic interaction problem. Specifically, we have checked that 1000 passages gives a good balance between accuracy of the results and efficiency of the overall simulation.

A detailed description of all the parameters involved in the wear model is provided in [1] and unless otherwise indicated, we have kept the parameter values given in that reference. The only changes are the current intensity $I_c = 300$ A which is supposed constant, the contact force F_c and the appearance of contact loss u which come from the dynamic simulation and the electrical contact resistance R_c which, for a contact between a copper contact wire and a graphite contact strip, depends on the contact force as experimentally stablished in [9]:

$$R_c = 0.015 + 0.18e^{-(\frac{F_c-4}{7})} \tag{4}$$

3.4 Worn section height

The objective of this step of the proposed algorithm is to compute the total worn height h of the contact wire. For a contact wire section of radius R with an initial worn section A_0 (coloured region in Figure 3), the worn height due to the *NWR* produced by an additional thousand pantograph passages (grated area in Figure 2) is obtained from Eq. (5), in which the angle θ is first computed by solving the nonlinear Eq. (6), being $A = A_0 + 1000NWR$.

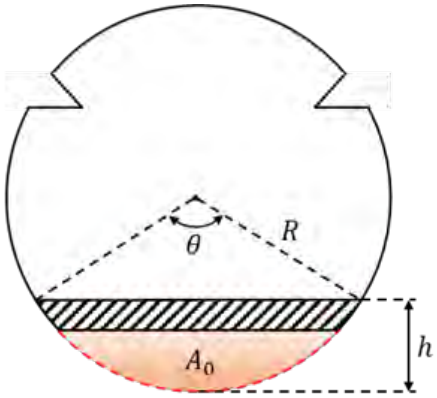


Figure 3: Worn section and worn height

$$h = R(1 - \cos\frac{\theta}{2}) \tag{5}$$

$$\frac{R^2}{2}(\theta - \sin\theta) - A = 0 \tag{6}$$

3.5 Static equilibrium

Once the contact wire height profile has been updated, it is important to note that the contact wire section has been reduced and therefore, the mass per unit length is modified. This weight loss modifies the force balance in the catenary model. That is why a static equilibrium problem is solved at this stage of the algorithm. This problem consists of solving Eq. (7) to obtain the new nodal coordinates that satisfy force equilibrium.

$$\boldsymbol{F_{int}}(\boldsymbol{q}) + \boldsymbol{F_g} = \boldsymbol{0} \tag{7}$$

Unless the initial configuration problem stated in Eq. (1), in this case, the element lengths are not set as unknowns.

4 NUMERICAL RESULTS

The numerical results presented in this work are obtained from the AC high speed contact line and the pantograph models provided in the standard EN-50318:2018 [10]. The initial contact wire section is 120 mm^2 and wear is only computed in a central region of the second catenary section, from kilometre point 400 to 800 m, to avoid boundary and transient effects. The stopping criteria for all the simulations performed is reaching a 20% of reduction on the contact wire section in a given kilometre point. This condition usually implies the replacement of the contact wire of the entire catenary section.

4.1 Nominal scenario

In this nominal case, the train speed is 300 km/h with an uplift force of 142.8 N acting on the pantograph mechanism. The main results obtained are shown in Figure 4.

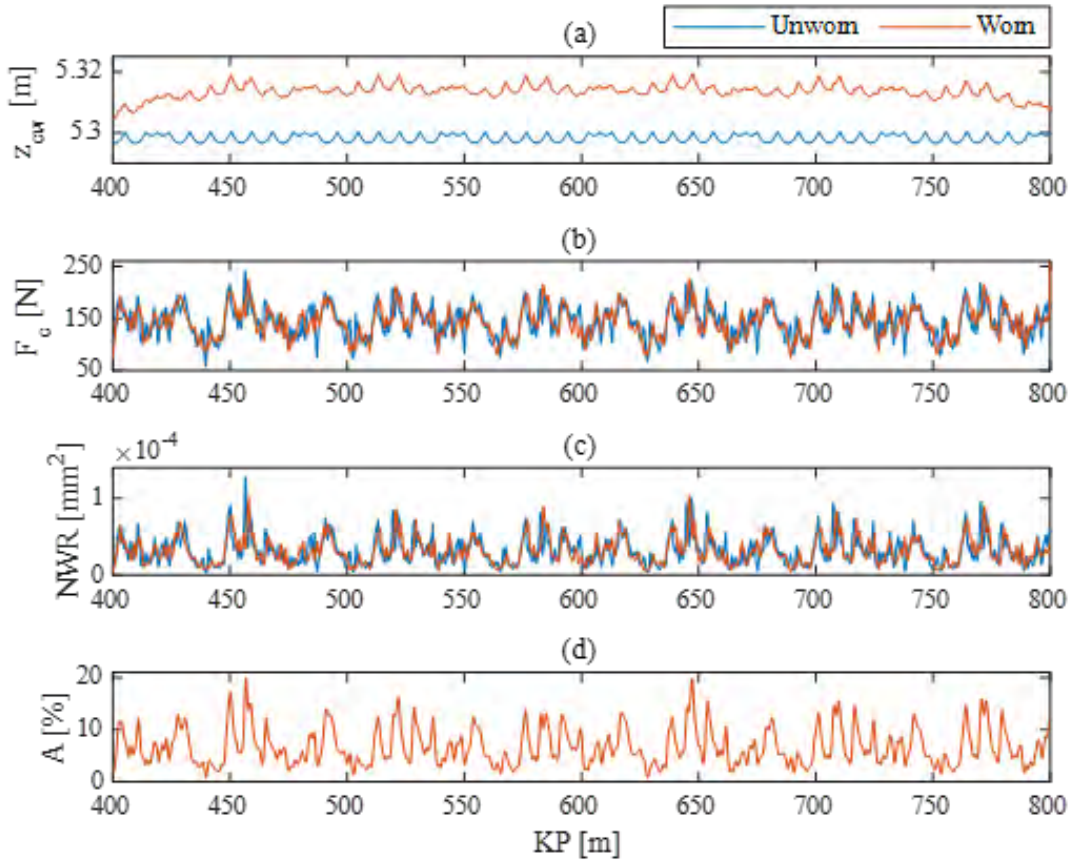


Figure 4: Comparison between the unworn and the worn catenaries in the nominal operating conditions: (a) contact wire height, (b) contact force, (c) normal wear ratio and (d) Percentage of worn area

The contact wire height is ploted in Figure 4a in which two effects can be distinguished. On the one hand, the main increase of the contact wire height (about 2 mm) is caused by the overall loss of weight. On the other hand, higher frequency irregularities are due to local wear effects. Figure 4b shows a comparison of contact force obtained from the unworn and the worn catenary. At first glance, a less oscilatory behaviour is observed in the contact force of the worn catenary, specially due to the disappearance of some local minima. The NWR obtained at the first and last pantograph passages is given in Figure 4c. An enlarged view of this plot is shown in Figure 5, in which a phase shift between the two wear rates is clearly observed.

This indicates that wear evolvement has a directional character which depends on the train travelling direction. This feature is also observed in the contact force. There is a clear direct relation between the contact force and the NWR since the main wear phenomena is mechanical friction in comparison to Joule wear.

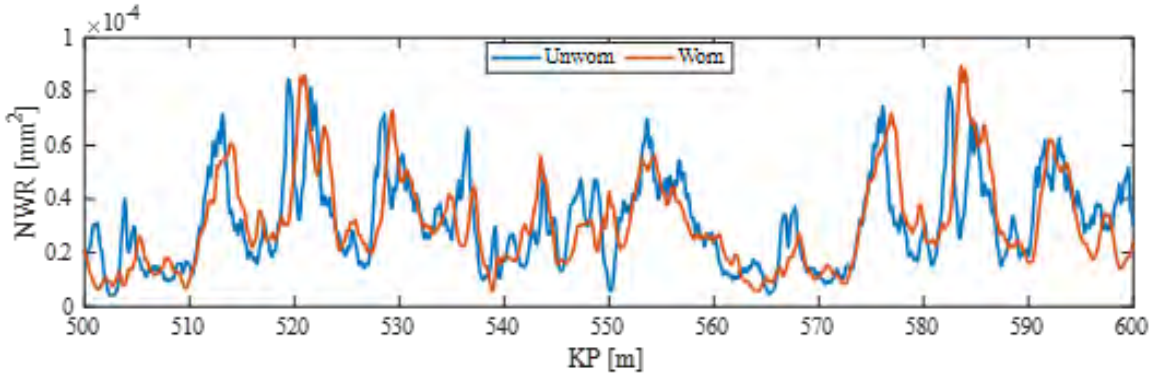


Figure 5: Enlarged view of the NWR obtained from the unworn and the worn catenary between kilometre points 500 and 600 m

Finally, in Figure 4d the percentage of worn section at the end of the simulation is shown. The 20% of worn area is reached at kilometre point 457 m after 250 dynamical simulations, which means 250.000 pantograph passages. Thus, this method provides a useful tool to foresee the life service period of a given catenary section.

4.2 Influence of train speed

It is well known that train speed has an important effect on the pantograph-catenary contact force and therefore, it is expected to also have it on the contact wire wear. In this section we compare the wear results obtained from simulations in which the pantograph travels at 200, 250, 300 and 350 km/h respectively. All the other parameters have been kept constant.

Figure 6 shows a comparison between the unworn and the worn catenary of the standard deviation σ, the maximum contact force F_c^{max} and the minimum contact force F_c^{min} for a different train speed. These results indicate that the worn catenary interacts with the pantograph producing a smother contact force than the unworn catenary as reflected by the lower value of *sigma* for all the studied velocities. This trend is confirmed by the lower values of the maximum force and the higher values of the minimum force found in the worn catenary. The relative variation of any of these values is also indicated in Figure 6. In general, such variations become more significative with the increase of train speed. This means that the contact wire wear evolvement tends to form a contact wire height profile that smoothes the contact force so that wear decelerates and the chance of electric arcs due to contact loss decreases.

The percentage of contact wire worn section is shown in Figure 7 for the four studied velocities. The limit of 20% of section reduction is reached at different kilometer points in each case (circles in Figure 7). The number of pantograph passages necessary to reach this value and replace the contact wire is 367.000, 350.000, 250.000 and 173.000 for the wear simulation with 200, 250, 300 and 350 km/h respectively.

The position of steady arms is marked with vertical dashed lines in Figure 7. For this catenary, the points that suffer the highest wear are located at midspan because the contact force presents higher values at this region.

It is important to mention that at low velocities the mean wear is higher, and there are several points with a worn section close to the 20% of the initial contact wire section. However, at high velocities most of the contact wire length suffers low wear while only a few local points

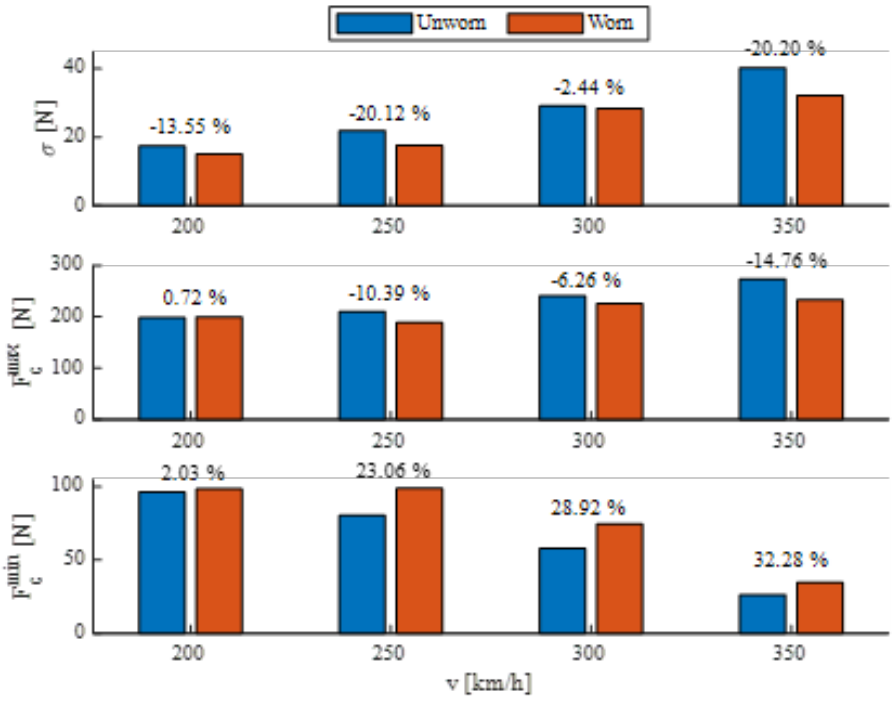


Figure 6: Statistics of the contact force under different train speeds for the unworn and the worn catenary

present severe wear. This implies to replace the whole contact wire of the catenary section because only in a few local spots the contact wire section has been reduced to its limit.

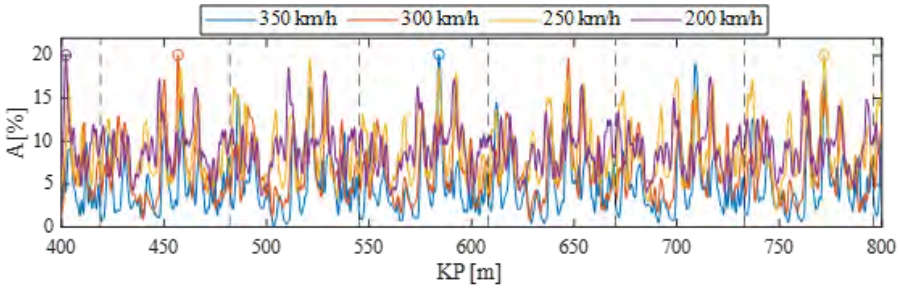


Figure 7: Percentage worn section along the contact wire for different train speeds

The higher overall amount of wear produced at low velocities is directly reflected in a higher contact wire height profile as shown in Figure 8. Specifically, the mean percentage of worn section is 9.11, 8.82, 6.76 and 4.85% for 200, 250, 300 and 350 km/h respectively.

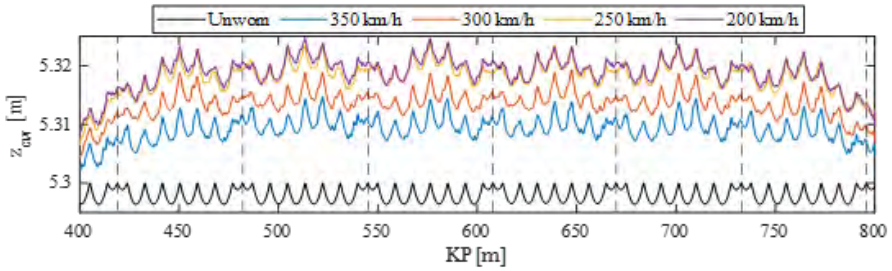


Figure 8: Contact wire height profile for the unworn catenary and the worn catenary when the pantograph travels at different velocities

5 CONCLUSIONS

This work proposes a simulation strategy to compute the long-term evolution of the contact wire wear for high-speed catenaries. The method uses an efficient dynamic solver of the pantograph-catenary interaction problem and a well-stablished contact wire wear model. As a novelty compared to other related works, the proposed algorithm considers the change in the catenary equilibrium position produced by the loss of material in the contact wire.

To exemplify the capabilities of the proposed method, it has been applied to a given catenary. Results such as the contact wire height profile, the contact force or the percentage of worn section were obtained for the nominal scenario concluding that wear evolution tends to provide a contact wire height that produces a smoother contact force so that, in a certain way, it seems to be beneficial for the current collection quality.

The effect of train speed on the contact wire wear has been also analysed. The main conclusion drawn from these simulations is that the increase of speed produces the localisation of wear on a few punctual regions of the contact wire, leading to its replacement even though on average it is little worn.

It is important to mention that this is an initial work on this field. Thus, the results and conclusions obtained cannot be extrapolated generally to other catenary-pantograph couples and they also need experimental measurements to be fully confirmed.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support received from the Spanish Ministry of Economy, Industry and Competitiveness (TRA2017-84736-R).

REFERENCES

[1] Bucca G. and Collina A. Electromechanical interaction between carbon-based pantograph strip and copper contact wire: A heuristic wear model. *Tribol. Int.* (2015) **92**:47-56.

[2] Derosa S., Nåvik P., Collina A., Bucca G and Rønquist A. A heuristic wear model for the contact strip and contact wire in pantograph-catenary interaction for railway operations under 15 kV 16.67 Hz AC systems. *Wear* (2020) **456**:203401.

[3] Bucca G. and Collina A. A procedure for the wear prediction of collector strip and contact wire in pantograph-catenary system. *Wear* (2009) **266**:46-59.

[4] Wei X.K., Meng H.F., He J.H. Jia L.M. and Li Z.G. Wear analysis and prediction of rigid catenary contact wire and pantograph strip for railway system. *Wear* (2020) **442**:203118.

[5] Wang H., Núñez A., Liu Z., Song Y., Duan F. and Dollevoet R. Analysis of the evolvement of contact wire wear irregularity in railway catenary based on historical data. *Veh. Syst. Dyn.* (2018) **56**(8):1207-1232.

[6] Song Y., Wang H and Liu Z. An investigation on the current collection quality of railway pantograph-catenary systems with contact wire wear degradations. *IEEE Trans. Instrum Meas.* (2021) **70**:9003311.

[7] Tur, M., García E., Baeza L. and Fuenmayor F.J. A 3D absolute nodal coordinate finite element model to compute the initial configuration of a railway catenary. *Eng. Struct.* (2014) **71**:234-243.

[8] Gregori S., Tur M., Nadal E. Aguado J.V., Fuenmayor F.J. and Chinesta F. Fast simulation of the pantograph-catenary dynamic interaction. *Finite Elem. Anal. Des.* (2017) **129**:1-13.

[9] Bucca G., Collina A., Manigrasso R., Mapelli F. and Tarsitano D. Analysis of electrical interferences related to the current collection quality in pantograph-catenary interaction. *Proc. Inst. Mech. Eng. Part F. J. Rail Rapid Transit.* (2011) **225**(5):483-500.

[10] EN-50318:2018. European Committee for Electrotechnical Standardization. Railway applications – Current collection systems – Validation of simulation of the dynamic interaction between pantograph and overhead contact line. (2018).

ROLLING NOISE REDUCTION THROUGH GA-BASED WHEEL SHAPE OPTIMIZATION TECHNIQUES

X. Garcia-Andrés*, J. Gutiérrez-Gil, V. T. Andrés, J. Martínez-Casas and F. D. Denia

Instituto Universitario de Ingeniería Mecánica y Biomecánica (I2MB)
Universidad Politècnica de València
Valencia, Spain

e-mail: xagaran@upv.es*, jorgugi1@upv.es, vicanrui@etsid.upv.es, jomarc12@mcm.upv.es, fdenia@mcm.upv.es

Key words: Railway wheel, Geometric optimization, Genetic Algorithms, Rolling noise

Abstract: *Railway rolling noise is nowadays a major source of acoustic pollution in urban areas, with nearly up to 12 million people daily affected in Europe by this phenomenon. Hence, the search for ways of decreasing such noise radiation has become a highly active and significant research field. Following this approach, a Genetic Algorithms-based shape optimization of the railway wheel is developed with the aim of minimizing rolling noise. Different approaches are considered to address the problem, such as directly minimizing radiated Sound poWer Level (SWL) or using the maximization of the natural frequencies if computational efficiency is especially critical. A parametric Finite Element model is implemented for the wheel based on the most relevant geometric parameters in rolling noise radiation. For the acoustic calculation, the sound radiation models used in the TWINS software are adopted, which also accounts for the whole dynamics of the wheel/rail system. Furthermore, for every candidate wheel proposed, a structural analysis is computed in order to check design feasibility in accordance with the corresponding standard. In all cases, new geometries for the wheel cross section are achieved that reduce the radiated rolling noise.*

1 INTRODUCTION

Railways are a highly efficient, cost-effective and low polluting transportation system. Unfortunately, there are also some drawbacks that need to be handled if the rail network is to be further expanded. Such issues are mainly related to acoustic pollution, what becomes especially important along urban environments, where it is estimated that about 12 million people are affected daily in Europe by the sound emitted by railway vehicles [1].

In that sense, one of the predominant types of noise emitted by railway vehicles is rolling noise [2], generated by the vibration of the wheel and track caused by the interaction force that emerges as a result of the irregularities present in their surfaces [3]. Through the different range of possible approaches that can be followed for rolling noise mitigation, those that consider its control at source are acknowledged as considerable cost-effective measures [4].

The present work therefore presents a procedure for the reduction of railway rolling noise by achieving optimal wheel geometries that minimize sound radiation. This is done by means of the global optimization technique known as Genetic Algorithm (GA) using two different methodologies, a first one based on the computed Sound Power Level (SWL), what includes solving the whole dynamic interaction, and another focused on the modal properties of the wheel. Moreover, the structural feasibility of each proposed candidate during the search is assured.

The document is structured as follows: firstly, the theoretical model used for the dynamic and acoustic calculations is introduced; Secondly, the optimization procedure is described along

https://doi.org/10.4995/YIC2021.2021.12577

with the defined objective functions and wheel shape parametrization; and, later, the results obtained are shown. Finally, a concluding discussion is presented.

2 THEORETICAL MODEL

For the present work, a methodology based on that of the commercial software TWINS is developed [5]. A system composed by a wheel and a continuously supported rail interacting at a contact point is considered. In order to derive the rolling noise radiation produced by the wheel, the whole coupled dynamic response of each of the components involved in the wheel/track interaction is solved through the use of linearised models in the frequency domain. Then, the wheel sound power is obtained with a semi-analytical formulation capable of computing the wheel acoustic efficiencies from the dynamic behaviour of its cross-section geometry.

2.1 Wheel response

The wheel response for the jth degree of freedom (d.o.f) is given by

$$u_{w,j} = -\sum_{i=1}^{3} H_{w,ji} F_{c,i}\,, \tag{1}$$

where $H_{w,ji}$ is the receptance of the wheel for the jth d.o.f. when the force is applied at the contact point in the ith direction and $F_{c,i}$ is the value of the contact force in the ith direction; x, y and z being represented by directions 1, 2 and 3, respectively.

The receptance of the wheel is given by modal superposition, with the associated modeshapes classified according to its number of nodal diameters n and nodal circumferences m, as [6]

$$H_{w,jk}\left(\omega\right) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{\Psi_{nm,j}\Psi_{nm,k}}{m_{nm}(\omega_{nm}^2 - \omega^2 + 2i\xi_{nm}\omega_{nm}\omega)}\,, \tag{2}$$

where, $\Psi_{nm,j}$ and $\Psi_{nm,k}$ are the modal amplitudes of the modeshape (n,m) for direction j and k, respectively, m_{nm} is the modal mass of the corresponding modeshape, ω_{nm} its natural frequency, ξ_{nm} the modal damping ratio and ω the angular frequency considered.

Regarding the derivation of the contact force $\bar{F}_c$, assuming that the excitation of the system is produced by the presence of a roughness amplitude r acting in the vertical direction, it can be stated that

$$\bar{r} = \boldsymbol{H}_{sys}\bar{F}_c\,, \tag{3}$$

with $\bar{r}$ being a vector with amplitude r in the vertical direction and $\boldsymbol{H}_{sys}$ the combined receptance of each component of the system defined as [2]

$$\boldsymbol{H}_{sys} = \boldsymbol{H}_w + \boldsymbol{H}_r + \boldsymbol{H}_c\,, \tag{4}$$

where $\boldsymbol{H}_w$, $\boldsymbol{H}_r$ and $\boldsymbol{H}_c$ are the receptances in matrix form of the wheel, rail and contact, respectively. In the present model, the rail receptance $\boldsymbol{H}_r$ is characterized considering the rail as a Timoshenko beam on a continuous foundation [2] and the contact receptance $\boldsymbol{H}_c$ describes the wheel/track interaction by means of a contact spring [7].

2.2 Wheel sound power

As a means to compute the radiated sound power of the wheel, the surface of this component is divided into six concentric rings and the tyre surface. Then, their velocity responses are calculated with the dynamic model introduced in the previous section and used for the

computation of the corresponding sound radiation. Therefore, it is possible to derive the wheel sound power W through [2]

$$W = \rho c_0 \sum_{l}^{N_m} \left(\sigma_l^a \sum_j (S_{a,j} \langle \overline{\tilde{v}_{a,jl}^2} \rangle) + \sigma_l^r S_r \langle \overline{\tilde{v}_{r,l}^2} \rangle \right) , \tag{5}$$

where index l refers to each the N_m modeshapes considered, ρ is the air density, c_0 the speed of sound, $\langle \overline{\tilde{v}_{a,jl}^2} \rangle$ and $\langle \overline{\tilde{v}_{r,l}^2} \rangle$ represent the mean squared vibration velocity averaged over time and surface area of the ring j and l^{th} modeshape for the axial and radial directions, respectively; $S_{a,j}$ refers to the axial area of the j^{th} ring, S_r to the surface used fo the radial radiation and σ_l^a and σ_l^r are the radiation efficiencies of the axial and radial contribution, respectively, for the l^{th} mode.

The radiation efficiencies, which are defined as the ratio of the amount of acoustic power radiated compared to that of a piston of the same area on an infinite wall when vibrating in the same manner [8], are obtained with a semi-analytical formulation detailed in [9].

3 OPTIMIZATION PROCEDURE

With the intention of minimizing the rolling noise radiated by the railway wheel, a Genetic Algorithms-based shape optimization procedure is developed and two different objective functions are studied: one based on the direct minimization of the radiated sound power ($L_{A,W}$-min), and another focused on the maximization of the natural frequencies of the wheel (NF-max).

Additionally, for the purpose of establishing a way of generating the different geometries propose for testing by the GA, a parametric FE model is defined using general axisymmetric elements [10]. The wheel cross section is set by the geometric characteristics found to be the most influential for the acoustic radiation [9, 11]: wheel radius x_1, fillet radius x_2, web thickness x_3 and web offset x_4. An overview of the described framework is presented in Fig. 1, while the design boundaries specified for this research are shown in Table 1. It should be noted that, as x_1, x_2 and x_4 are absolute parameters whose value directly correspond to the corresponding geometric property, x_3 is defined as a proportionality factor of the reference thickness along the web. Besides, due to constraints in the design process related with the modification of the wheel radius, two different optimizations are run for each procedure: one considering all the components in the parametrization and another in which the radius is kept as constant with value $x_1 = 0.45$ m.

The optimization algorithm proceeds as follows: the first step is to create a set of wheel candidates by using the defined parametrization, which conforms the generation $\boldsymbol{X_i}$; then, for every candidate $\bar{\boldsymbol{x}}_{\boldsymbol{j}}$ in $\boldsymbol{X_i}$, the structural feasibility of each proposed wheel is checked by following the standard EN13979-1 [12]. If the candidate is feasible, a modal analysis to obtain the N_m modeshapes $\boldsymbol{\Psi_{nm}}$ and natural frequencies ω_{nm}, needed for the calculation of the studied objective functions, is carried out by the FEM software ANSYS APDL. Afterwards, modeshapes are identified and classified in accordance to the number of nodal diameters and nodal circumferences (n, m) they present and the selected objective function Obj is evaluated.

Table 1: Design domain for the optimization methodologies.

	x_1 [m]	x_2 [m]	x_3	x_4 [m]
Reference	0.45	0.0427	0.0681	0.0300
Lower Boundary	0.40	0.0364	-0.1000	-0.2700
Upper Boundary	0.50	0.0484	0.1000	0.2700

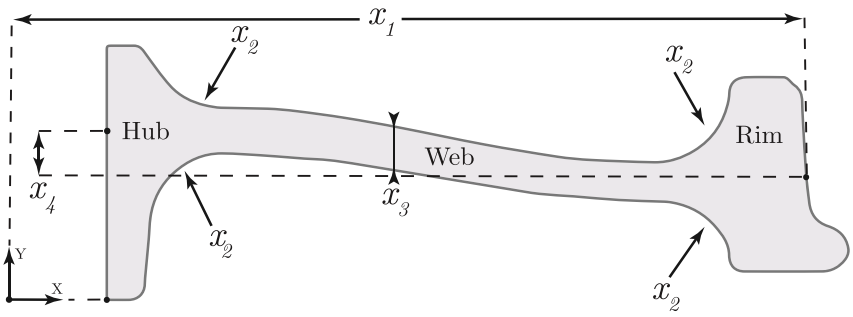


Figure 1: Design variables of the wheel cross section parametrization.

Once *Obj* has been computed for the whole generation, the stopping criteria is checked. If it is fulfilled, the candidate most suited for the objective function $\bar{\boldsymbol{x}}^*$ is selected as the Best Found Solution (BFS). Otherwise, a new generation is set that accounts for the geometrical information of the cross sections already analysed during each iteration and the described process is repeated. For further clarification, a flow diagram of the optimization algorithm is represented in Fig. 2.

3.1 Objective functions

As already mentioned in the prevous sections, two different objective functions are used along the optimization algorithm, $L_{A,W}$-min and NF-max. Below, their main features are explained with further detail.

3.1.1 $L_{A,W}$-min methodology

In the $L_{A,W}$-min methodology the goal is to directly minimize the radiated noise emitted by the wheel. With this intention, the SWL expressed in dB(A) is used, computed for every design as

$$\mathrm{SWL} = 10\log_{10}(\frac{W}{W_{ref}}) + A_{filter}\,, \tag{6}$$

where W is the sound power, $W_{ref} = 10^{-12}$ W and A_{filter} is the A-weighting filter for dB.

Next, *Obj* is defined as the summation in energy of the SWL in each frequency band.

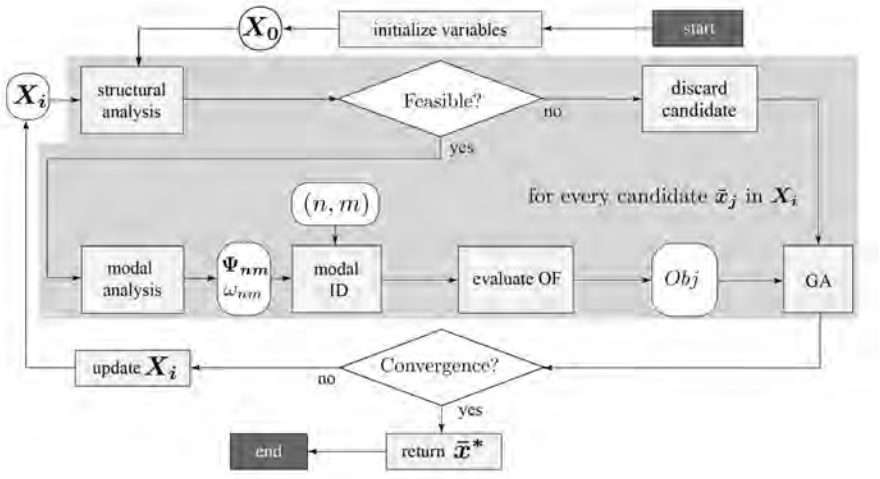


Figure 2: Flow diagram describing the optimization procedure.

Therefore,

$$Obj_{L_{A,W}} = 10\log_{10}(\sum_{i=1}^{n_{cf}} 10^{\frac{\mathrm{SWL}_i}{10}}), \tag{7}$$

where SWL_i is the SWL of the ith one-third octave band and n_{cf} the number of bands in the chosen frequency region.

3.1.2 NF-max methodology

In the NF-max methodology, the minimization of rolling noise is intended indirectly through the maximization of the natural frequencies of the wheel. The assumption made in this case is that, as the excitation of the system is dependent on the wheel-rail combined roughness and its content is lower in the high frequency region, shifting the natural frequencies to higher frequency regions should lead to wheel shapes whose vibration modes are less excited and, consequently, quieter designs. With this aim, the objective function Obj for the current methodology its defined as

$$Obj_{NF} = \frac{1}{\hat{\omega}_m}, \tag{8}$$

where $\hat{\omega}_m$ is the mean of all the N_m extracted natural frequencies of the wheel.

4 RESULTS

For the results presented in this section, the following specifications are used: an UIC54 rail with concrete bibloc sleepers separated 0.6 m, the parameters of which are shown in Table 2, and a standard roughness defined for a train speed of $V = 80$ km/h when a contact filter is applied [13]. In the dynamic calculations, the frequency range varies from 50 to 5000 Hz with a resolution of 1 Hz and the reference wheel, taken as a guideline to compare the changes observed for the wheel designs, is based on a simplified monobloc wheel with typical dimensions. As for the modal analysis made, a rigid constraint is applied at the nodes on the inner surface of the wheel hub, the maximum element size defined for the FE mesh is $h = 0.007$ m and a number of $N_m = 48$ modeshapes are considered. Additionally, in order to assure the correct development of the theoretical model, the combined SWL for all the components involved in the rolling noise radiation is compared to the results offered for the same case by the commercial package TWINS [5]. As it can be seen in Fig. 3, no significant discrepancies are observed, with a total variation in terms of energy of $\Delta L_{A,T} = 0.17$ dB(A).

The main results for both methodologies are presented in Table 3 and the wheel cross section geometries obtained for each procedure are shown in Fig. 4. The two approaches show a reduction in both the wheel SWL and total SWL for either the fixed radius case or the optimization with all the parameters. Thus, in the fixed radius case, it is clear that the obtained $L_{A,W}$ are lower than the reference wheel, with variations of $\Delta L_{A,W} = -3.94$ dB(A) and

Table 2: Track parameters used in SWL calculations.

Rail UIC54	*Vertical*	*Lateral*	**Foundation**	*Vertical*	*Lateral*
Bending stiffnes EI [Nm^2]	$4.93 \cdot 10^6$	$0.87 \cdot 10^6$	**Pad stiffness k'_p [N/m^2]**	$2.17 \cdot 10^9$	$1.17 \cdot 10^8$
Shear coefficient κ	0.4	0.4	**Pad loss factor η_p**	0.25	0.25
Loss factor η_r	0.02	0.02	**Ballast stiffness k'_b [N/m^2]**	$1.17 \cdot 10^8$	$5.83 \cdot 10^7$
Mass per length ρA [kg/m]	54		**Ballast loss factor η_b**	2	2
Cross receptance level	-15		**Sleeper mass per length m'_s [kg/m]**	203.33	

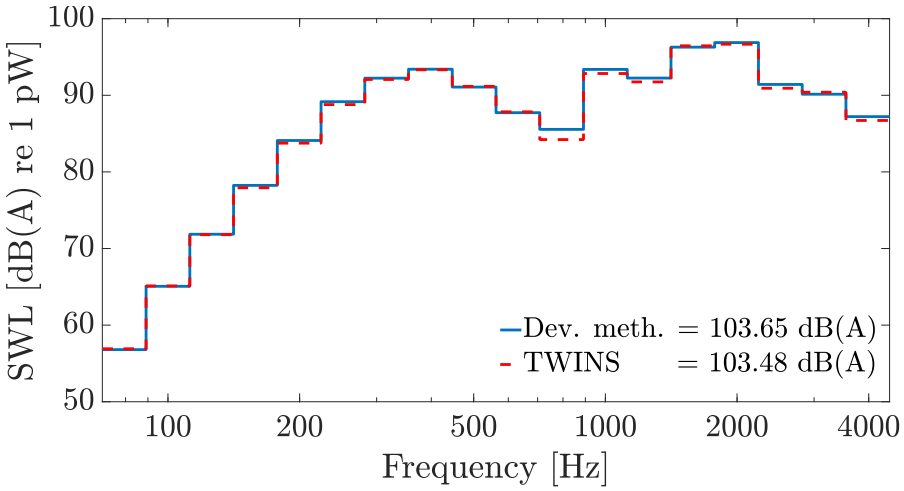


Figure 3: Total SWL produced by the commercial program TWINS (−−) and the methodology developed in the present work (—)

.

$\Delta L_{A,W} = -1.11$ dB(A) for the $L_{A,W}$-min and NF-max methodologies, respectively. Regarding the noise when considering all the components involved, reductions are kept with a change in $L_{A,T}$ of $\Delta L_{A,T} = -1.87$ dB(A) for $L_{A,W}$-min and $\Delta L_{A,T} = -0.43$ dB(A) NF-max. When all the geometric parameters are considered in the optimization, quieter wheel designs are achieved with improvements of up to $\Delta L_{A,W} = -4.96$ dB(A) and $\Delta L_{A,T} = -2.05$ dB(A) for the $L_{A,W}$-min approach. In all cases, an increase of the mean of the natural frequencies $\hat{\omega}_m$ is produced and the NF-max approach appears as computationally demanding methodology, requiring a lower number of generations n_{gen} to achieve convergence with a $\Delta n_{gen} = -24$ generations when compared with $L_{A,W}$-min in the optimization with fixed radius.

Concerning the evolution of the wheel shape along the optimization procedures, different Response Surfaces (RS) are generated in order to study it. For each RS, a pair of the defined geometric parameters are chosen and evaluated for each objective function in 676 different points along the solution space, allocated in the form of a 26×26 evenly distributed sampling grid. Some of the most relevant RSs generated are represented in Fig. 5. As exemplified in the results shown in Fig. 5a, the objective function defined for the NF-max approach presents mainly a planar form and the greatest maximization of natural frequencies is related with the decrease of radius x_1. Conversely, in the $L_{A,W}$-min methodology, the observed behaviour is in a more complex and variable manner. This can be seen in Fig. 5b, which also reveal the predominance of the web offset x_4 variable, followed by the radius x_1, in setting the value of the corresponding objective function. It should be noted that in all cases the minimum value obtained for the selected objective function were worse than that offered by the optimization.

Finally,for the purpose of further exploring the relation between Obj_{NF} and $Obj_{L_{A,W}}$, the

Table 3: BFS values for the optimization procedures. x_1, x_2 and x_4 are expressed in m. All L_A values are expressed in dB(A).

Methodology	x_1	x_2	x_3	x_4	$L_{A,W}$	$\Delta L_{A,W}$	$L_{A,T}$	$\Delta L_{A,T}$
$L_{A,W}$-min (Fixed rad.)	0.4500	0.0484	0.1000	-0.0128	97.35	-3.94	102.28	-1.87
NF-max (Fixed rad.)	0.4500	0.0484	0.1000	-0.0270	100.18	-1.11	103.73	-0.43
$L_{A,W}$-min (All param.)	0.4222	0.0483	0.0999	-0.0102	96.33	-4.96	102.10	-2.05
NF-max (All param.)	0.4000	0.0484	0.1000	-0.0167	99.26	-2.03	103.10	-1.05
Ref.	0.4500	0.0427	0.0700	0.0300	101.29	-	104.16	-

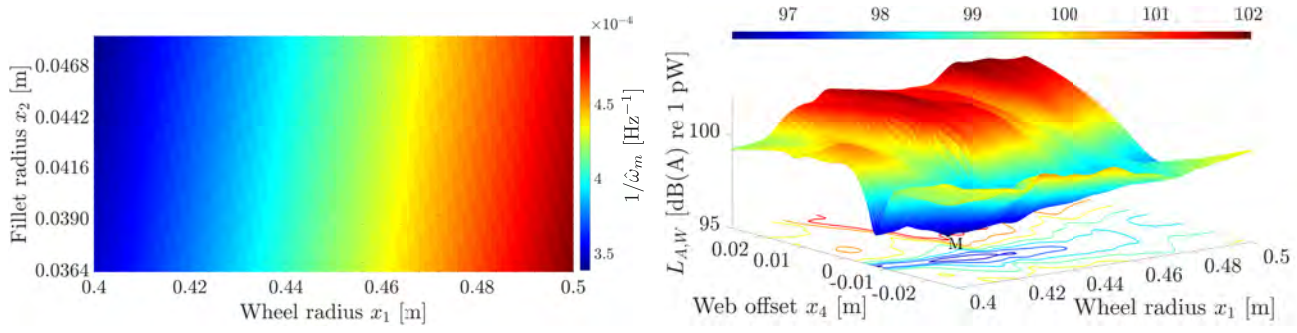
Figure 4: Wheel shapes comparison for the BFSs obtained by the optimization procedure with the fixed radius case (left) and considering all the geometric parameters (right). In both cases the results are shown for the NF-max (orange) and $L_{A,W}$-min (green) approaches together with the reference wheel (black).

objective function value $L_{A,W}$ for all candidates designs evaluated in the optimization runs is plotted against their natural frequencies mean in Fig. 6. Although some correlations can be found locally, there are not for the totality of sampled candidates: for the optimization with all the geometric parameters, the decreases of both objective functions value are coupled in the region where x_1 is above the optimum value ($x_1 = 0.42$ m), but below this point the trend shift; and in the fixed radius case, a wide range of emitted noise is present for the the candidates with minimum Obj_{NF} value. In both cases, the observable patterns are consistent with the existence of design variables with high influence on the acoustic behaviour but low on the fixing of the natural frequencies, as the web offset x_4.

5 CONCLUSIONS

With the goal of reducing acoustic radiation, a geometric optimization ot the railway wheel cross-section shape is performed by means of a GA-based optimizer. Two different methodologies are applied: the NF-max methodology, focused on the maximization of the natural frequencies, and the $L_{A,W}$-min methodology, based on the direct minimization of the wheel SWL. Furthermore, response surfaces for different combinations of geometric parameters are carried out in order to study their behaviour along the optimization process.

Results reflect that in all approaches a reduction is accomplished for both the wheel SWL, with improvements of up to 4.96 dB(A) in the $L_{A,W}$-min case, and the SWL when all components involved in the rolling noise radiation are considered. The differences in the evolution of



(a) (b)

Figure 5: Response surfaces for different combinations of geometric parameters: (a) with Obj_{NF} for x_1 and x_2; (b) with $Obj_{L_{A,W}}$ for x_1 and x_4. Fixed values correspond to those of the BFS for the corresponding optimization procedure and points M indicate the RS minima.

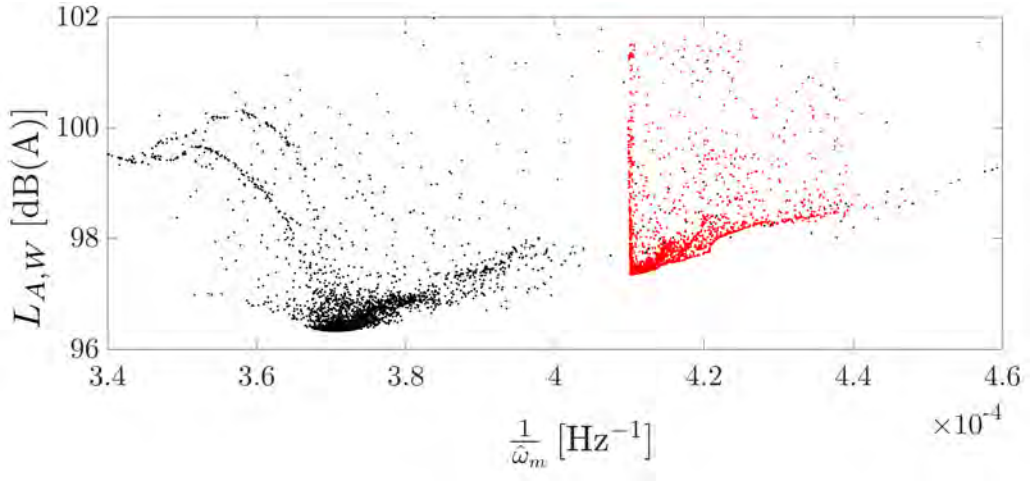


Figure 6: Obj_{NF} and $Obj_{L_{A,W}}$ values for candidate evaluations corresponding to the optimization runs using the $L_{A,W}$-min methodology. Black points are evaluations in a run with a design space consisting of all x_1 to x_4 variables. Red points are evaluations in a run considering a fixed wheel radius variable x_1.

each objective functions when modifying the wheel shape are established, identifying the radius x_1 and web offset x_4 as the principal factors in the changes observed. Finally, local correlations are found between the NF-max and $L_{A,W}$-min objective functions behaviour, although not for the totality of sampled wheel cross-sections. In all cases, the observed patterns are related with the existence of design variables with significant influence on the acoustic performance although not in a noticeable way on the fixation of the natural frequencies.

6 ACKNOWLEDGEMENTS

This study has been supported by the Agencia Estatal de Investigación and the European Regional Development Fund (project TRA2017-84701-R).

7 REFERENCES

[1] World Health Organization European Centre for Environment and Health, "Burden of disease from environmental noise," WHO, Tech. Rep., 2011.

[2] D. J. Thompson, *Railway Noise and Vibration. Mechanisms, modelling and means of control*, 1st ed. Elsevier, 2010. DOI: 10.1016/B978-0-08-045147-3.X0023-0.

[3] D. J. Thompson, "Wheel-rail noise generation, part I: Introduction and interaction model," *Journal of Sound and Vibration*, vol. 161, no. 3, pp. 387 –400, 1993. DOI: 10.1006/jsvi.1993.1082.

[4] J. Oertli, "The STAIRRS project, work package 1: A cost-effectiveness analysis of railway noise reduction on a European scale," *Journal of Sound and Vibration*, vol. 267, no. 3, pp. 431–437, 2003. DOI: 10.1016/S0022-460X(03)00705-3.

[5] D. J. Thompson, B. Hemsworth, and N. Vincent, "Experimental validation of the TWINS prediction program for rolling noise, part 1: Description of the model and method," *Journal of Sound and Vibration*, vol. 193, no. 1, pp. 123–135, 1996. DOI: 10.1006/jsvi.1996.0252.

[6] D. J. Thompson, "Wheel-rail noise generation, part II: Wheel vibration," *Journal of Sound and Vibration*, vol. 161, no. 3, pp. 401 –419, 1993. DOI: 10.1006/jsvi.1993.1083.

[7] D. J. Thompson, "Wheel-rail noise generation, part IV: Contact zone and results," *Journal of Sound and Vibration*, vol. 161, no. 3, pp. 447 –466, 1993. DOI: 10.1006/jsvi.1993.1085.

[8] F. Fahy and P. Gardonio, *Sound and Structural Vibration*, 2nd ed. Academic Press, 2007, ch. 3, pp. 135 –241. DOI: 10.1016/B978-012373633-8/50007-7.

[9] D. J. Thompson and C. J. C. Jones, "Sound radiation from a vibrating railway wheel," *Journal of Sound and Vibration*, vol. 253, no. 2, pp. 401–419, 2002.

[10] M. Petyt, *Vibration of Solids*, 2nd ed. Cambridge University Press, 2010. DOI: 10.1017/CBO9780511761195.007.

[11] J. C. O. Nielsen and C. R. Fredö, "Multi-disciplinary optimization of railway wheels," *Journal of Sound and Vibration*, vol. 293, no. 3-5, pp. 510–521, 2006. DOI: 10.1016/j.jsv.2005.08.063.

[12] UNE, "Railway applications. Wheelsets and bogies. Monobloc wheels. Technical approval procedure. Part 1: Forged and rolled wheels. UNE-EN-13979-1:2006," Asociación Española de Normalización (UNE), Technical Standard, 2011.

[13] X. Garcia-Andrés, J. Gutiérrez-Gil, J. Martínez-Casas, and F. D. Denia, "Wheel shape optimization approaches to reduce railway rolling noise," *Structural and Multidisciplinary Optimization*, 2020. DOI: 10.1007/s00158-020-02700-6.

RAILWAY ROLLING NOISE MITIGATION THROUGH OPTIMAL TRACK DESIGN

V. T. Andrés*, J. Martínez-Casas, J. Carballeira, F. D. Denia and D. J. Thompson†

* Instituto de Ingeniería Mecánica y Biomecánica (I2MB)
Universitat Politècnica de València
Valencia, Spain
e-mail: vicanrui@etsid.upv.es, jomarc12@mcm.upv.es, jacarmo@mcm.upv.es, fdenia@mcm.upv.es

† Institute of Sound and Vibration Research (ISVR)
University of Southampton
Southampton, United Kingdom
e-mail: djt@isvr.soton.ac.uk

Key words: Railway dynamics, rolling noise, sound radiation model, track design, design of experiments, noise mitigation.

Abstract: *The main goal of the present work lies in the identification of the railway track properties that influence acoustic radiation, as well as in the analysis of these properties for the reduction of sound levels. This is achieved through a dynamic model of the railway wheel and track that allows the study of rolling noise, produced as a result of the wheel/rail interaction. Once the vibrational response of the railway components is determined, the sound power radiated by them is evaluated. The influence of the track properties on the sound radiation is determined by analysing the acoustic power results of different track configurations. From the results obtained, a number of guidelines are presented for noise mitigation of the involved railway elements. Between the worst and the best track design, there are differences of approximately 7.4 dB(A) in the radiation considering the wheel, rail and sleeper noise.*

1 INTRODUCTION

Noise pollution due to transport is one of the most damaging environmental factors for humans, according to the World Health Organization [1]. The consequences of prolonged exposure to high noise levels include, in order of severity, hearing loss, hypertension, ischaemia, insomnia and even changes in the immune system [2]. Consequently, the development of tools for detection, analysis and mitigation of sound levels radiated from railway transport is of great importance. Among the sources of acoustic radiation of railway vehicles, rolling noise is considered one of the most relevant [3].

In this work, a dynamic model of the wheel and track is implemented, which allows calculating the rolling noise radiated by the different railway elements (wheel, rail and sleeper). With this approach, the geometric and viscoelastic parameters of the track that most influence sound radiation are identified. Also, the necessary changes in these factors to reduce railway noise levels are determined [4].

The vibroacoustic model and methodology for the analysis of track influence on sound radiation are presented in Section 2. Results of an optimal track design as well as some guidelines to achieve noise mitigation are given in Section 3. In Section 4 the conclusions of the work are summarised.

https://doi.org/10.4995/YIC2021.2021.12583

2 METHODOLOGY

2.1 Vibroacoustic model

To model the dynamic behaviour of the wheel, the Finite Element Method (FEM) is applied. Vibration modes of the wheel can be characterized according to the number of nodal lines (no vibration) that cross the wheel in a radial direction passing through its centre, known as nodal diameters [5]. This characterization allows grouping the contribution of the modes to the motion of the wheel and, consequently, to its acoustic radiation. By adopting a modal approach, the vibrational response of the wheel is evaluated.

After solving the dynamics of the railway wheel, its acoustic radiation is calculated as a postprocess of the vibrational field on its surface. The radiation model used in this work was

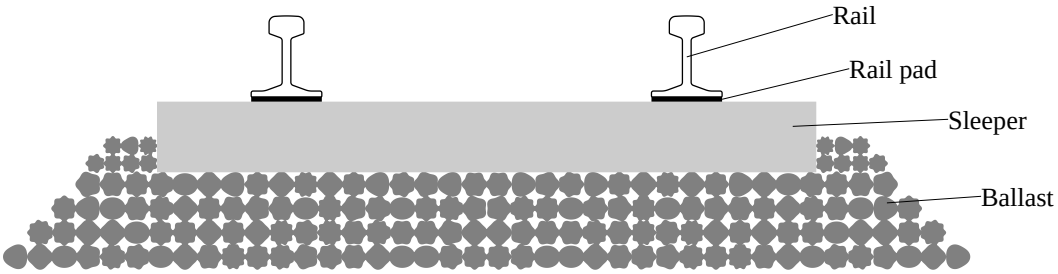


Figure 1: Railway track model configuration.

Due to the wheel/rail interaction force, structural waves propagate in the longitudinal direction through the infinite track. These propagating waves are evaluated applying the methodology proposed by Mead, whose formulation can be found in [9]. This consists of analysing a finite track segment using FE techniques. Using this approach, the displacement of any point on the track is obtained as a superposition of waves.

Regarding the sound radiation of the track, in this work it is assumed that there is a contribution from the rail and sleeper. The acoustic models of both components implemented in this work are described in [10] and it is assumed a two-dimensional radiation of each cross-section of the track, which is subsequently corrected to consider the three-dimensional nature of the sound radiation. Given the proportionality of the acoustic power and dynamic response of a component, the radiation from the rail and sleeper is also obtained as a superposition of the radiation associated with each wave.

The coupling between the wheel and track occurs through the wheel/rail interaction. The roughness present on the surfaces of both components is a source of excitation when the vehicle travels along the track. This excitation generates a vibrational field in the railway elements, producing rolling noise. A roughness spectrum defined in the standard EN13979-1 [11] is used. The contact model proposed by Thompson [12] is used in this work, which evaluates the interaction force from the wheel and rail combined roughness.

2.2 Influence analysis

This work aims to analyse the influence of the railway track design on the sound radiation of the wheel, the rail and the sleeper. In particular, the effect of the rail geometry and viscoelastic properties of the rail pad and ballast are studied. To do this, first, the rail profile is parameterised in six main variables (see Figure 2) and their limits are established; similarly, limits are set for the viscoelastic properties of the pad and ballast. Subsequently, a design of experiments and an ANOVA are carried out on the results, looking for a regression model that fits the calculated acoustic power. If the fit is good enough, the analysis of the regression coefficients allows knowing the influence of the different contributing variables on the sound radiation.

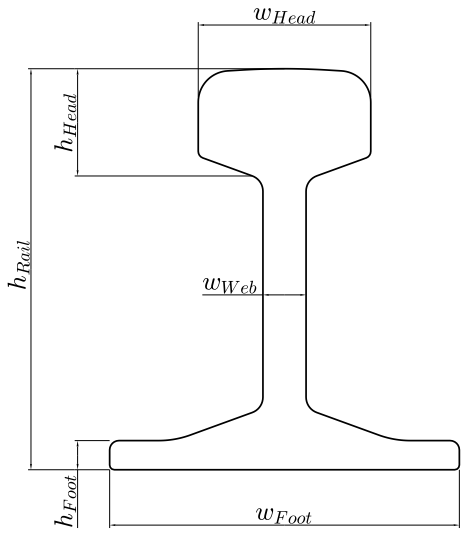


Figure 2: Rail profile parameterization.

In total, ten main parameters of the track are considered. Six of them describe the geometry of the rail (see Figure 2): w_{Head}, h_{Head}, w_{Foot}, h_{Foot}, w_{Web} and h_{Rail}; two represent the stiffness and damping of the rail pad, k_{Pad} and η_{Pad}, respectively; and the last two represent the stiffness and damping of the ballast, $k_{Ballast}$ and $\eta_{Ballast}$, respectively. In order to analyse the influence of these on sound radiation, a factorial design is proposed, covering all possible combinations of the variables. An ANOVA is performed on the result of the simulations, modifying the effects to ensure their statistical significance on radiation. The total acoustic power, which is the sum of the power of the rail, sleeper and wheel, is quantified by adding the energy contained in the frequency spectrum after including the A-weighting of the sound levels.

In this work both the influence of each parameter and its importance on the sound radiation are determined. For this, the technique developed by Pratt [13] is applied, by which the importance of each contributing variable is determined from the set of samples obtained from the factorial design calculation. For these samples, a polynomial regression is performed, given by:

$$\hat{\mathbf{y}} = \sum_j \beta_j \mathbf{x}_j, \tag{1}$$

where the response variable $\hat{\mathbf{y}}$ is the total radiation of each combination of the design of experiments previously standardised to unit variance and null mean, $\mathbf{x}_j$ is the standardised jth effect and β_j is the jth coefficient. An effect can be a simple parameter, an interaction

or a power. Note that the vector of the adjusted response variable $\hat{\mathbf{y}}$ is obtained as a linear combination of the standardised effects $\mathbf{x}_j$, which form the basis of the vectorial subspace of the model. For two effects, this concept can be visualized in Figure 3.

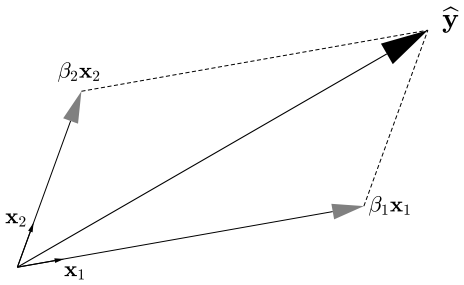


Figure 3: Vectorial subspace of the regression model.

The projection of $\beta_j\mathbf{x}_j$ onto $\hat{\mathbf{y}}$ represents the importance of the jth effect. In this work, the importance of the jth effect d_j is defined as follows:

$$d_j = \hat{\mathbf{y}} \cdot (\beta_j \mathbf{x}_j), \tag{2}$$

which represents the proportion of variance in the response variable that the jth effect explains. Consequently, the cumulative importance of all significant effects results in the coefficient of determination R^2 of the regression model.

3 RESULTS

The implemented vibroacoustic tool described in Section 2.1 has been verified with the commercial package TWINS [7, 8], which is considered as the reference program in railway rolling noise calculation.

To study the track influence on sound radiation, a design of experiments is carried out with five levels of each parameter. For each combination of them, the sound radiation of the railway components is calculated using the implemented tool. An ANOVA with the significant effects is performed on the results and the Pratt methodology, described in Section 2.2, is applied to determine the variability explanation of each effect. Using this technique, the variables influencing the total radiation are established, which are the width of the rail foot (w_{Foot}) and the four viscoelastic parameters of the track (k_{Pad}, η_{Pad}, $k_{Ballast}$ and $\eta_{Ballast}$). The polynomial regression model performed on the results of the factorial design has a coefficient of determination $R^2 = 99.43$ %. In Figure 4 the importance of each significant effect of the regression model as well as the cumulative importance are shown. The stiffness of the rail pad is the most important parameter, explaining 83.58 % of the sound radiation variability.

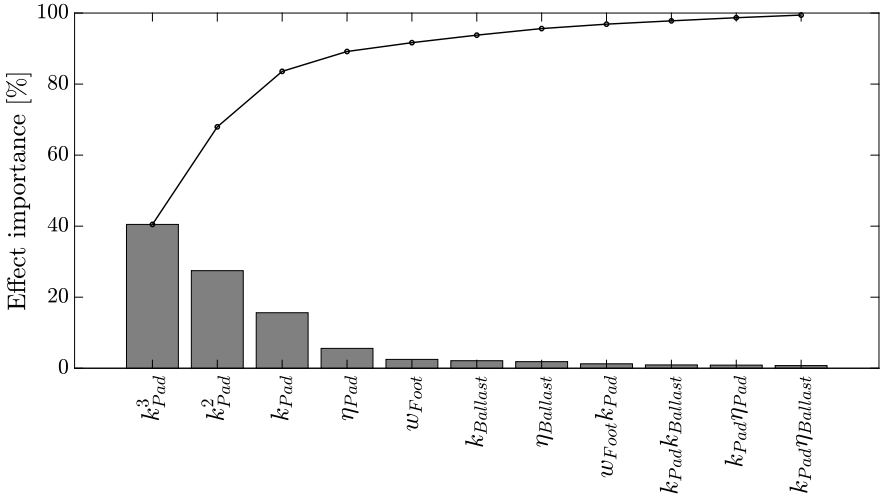


Figure 4: Importance and cumulative value of the significant effects.

An increase in the rail pad stiffness leads to a reduction in the rail and wheel noise and an increase in the sleeper radiation; for the total noise, at low stiffnesses the reduction of rail noise predominates and at high stiffnesses there is a balance between the three components. Regarding the rest of the parameters, a reduction in the rail foot width results in lower radiation levels of the rail as it reduces the radiation ratio and radiation area; the rail pad damping and ballast damping reduce the vibration amplitudes of the rail and sleeper, yielding a positive influence on their sound radiation; the ballast stiffness governs the vibrational response of the sleeper and, consequently, its acoustic power.

The optimal solution for the total sound power corresponds to a minimum value of w_{Foot} and maximum values of η_{Pad}, $k_{Ballast}$ and $\eta_{Ballast}$; regarding k_{Pad}, the minimum sound power levels are obtained with an intermediate/high stiffness, where the aforementioned balance is achieved. The regression model predicts that the optimal design is reached with the following parameters: $w_{Foot} = 120$ mm, $k_{Pad} = 780$ MN/m, $\eta_{Pad} = 0.5$, $k_{Ballast} = 100$ MN/m and $\eta_{Ballast} = 2$, with an acoustic power of 98.4 dB(A). In contrast, the worst design corresponds to the following parameters: $w_{Foot} = 150$ mm, $k_{Pad} = 130$ MN/m, $\eta_{Pad} = 0.25$, $k_{Ballast} = 40$ MN/m and $\eta_{Ballast} = 1$, with a power of 105.8 dB(A). Therefore, there is a difference between the best and the worst combination of 7.4 dB (A). Figure 5 shows the sound power levels of the track design with the worst combination of parameters and with the optimal combination.

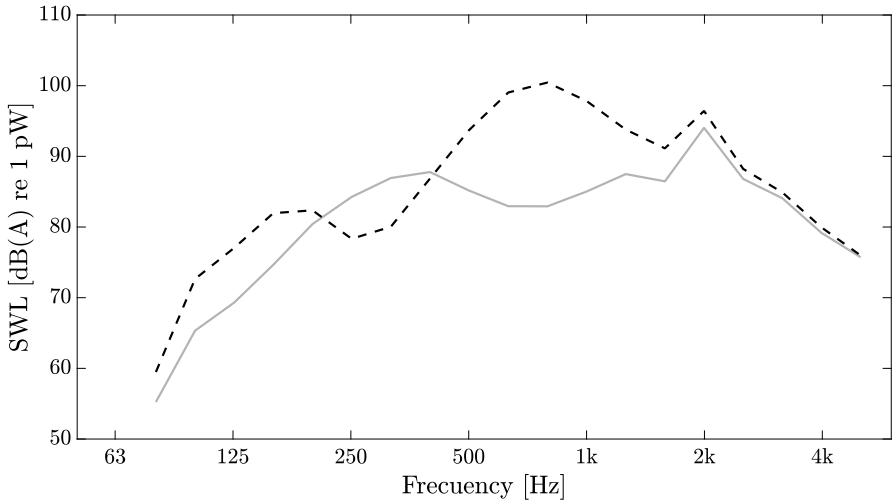


Figure 5: Sound power level for one wheel and associated track vibration. Best track design (——) and worst design (- - -).

4 CONCLUSIONS

A vibroacoustic model of the railway wheel and track has been implemented for the prediction of noise radiation from the wheel, rail and sleeper. A geometric parameterization of the rail profile has been carried out, which has allowed performing a design of experiments in order to analyse the influence of the track design on sound radiation. The geometry of the rail, represented by six variables, and the stiffness and damping of the rail pad and ballast, modelled through four additional variables, are studied.

The most important contributing variables are the viscoelastic properties of the rail pad and ballast and the width of the rail foot. The minimum sound power levels are found with minimum values of the rail foot width, maximum values of the rail pad and ballast damping, maximum values of the ballast stiffness and intermediate/high values of the rail pad stiffness. These values conform the optimal track design, which originates a total radiation 7.4 dB(A) lower than the worst track design found.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of Agencia Estatal de Investigación and European Regional Development Fund (grant FPU18/03999, project TRA2017-84701-R and project PID2020-112886RA-I00).

REFERENCES

[1] T. Münzel, S. Kröller-Schön, M. Oelze, T. Gori, F. P. Schmidt, S. Steven, O. Hahad, M. Röösli, J. M. Wunderli, A. Daiber and M. Sørensen. Adverse Cardiovascular Effects of Traffic Noise with a Focus on Nighttime Noise and the New WHO Noise Guidelines. *Annual Review of Public Health*, Vol. **41**(1), pp. 309–328, (2020). DOI: https://doi.org/10.1146/annurev-publhealth-081519-062400.

[2] W. Passchier-Vermeer and W. F. Passchier. Noise exposure and public health. *Environmental Health Perspectives*, Vol. **108**(1), pp. 123–131, (2000). DOI: https://doi.org/10.1289/ehp.00108s1123.

[3] X. Zhang, D. J. Thompson, E. Quaranta and G. Squicciarini. An engineering model for the prediction of the sound radiation from a railway track. *Journal of Sound and Vibration*, Vol. **461**, 114921, (2019). DOI: https://doi.org/10.1016/j.jsv.2019.114921.

[4] V. T. Andrés, J. Martínez-Casas, F. D. Denia and D. J. Thompson. Influence study of rail geometry and track properties on railway rolling noise. *Journal of Sound and Vibration*.

[5] D. J. Thompson. *Railway Noise and Vibration. Mechanisms, Modelling and Means of Control*. Elsevier, 2009. ISBN: 978-0-08-045147-3.

[6] D. J. Thompson and C. J. C. Jones. Sound radiation from a vibrating railway wheel. *Journal of Sound and Vibration*, Vol. **253**(2), pp. 401–419, (2002). DOI: https://doi.org/10.1006/jsvi.2001.4061.

[7] D. J. Thompson, B. Hemsworth and N. Vincent. Experimental validation of the TWINS prediction program for rolling noise, part 1: description of the model and method. *Journal of Sound and Vibration*, Vol. **193**(1), pp. 123–135, (1996). DOI: https://doi.org/10.1006/jsvi.1996.0252.

[8] D. J. Thompson, P. Fodiman and H. Mahé. Experimental validation of the TWINS prediction program for rolling noise, part 2: results. *Journal of Sound and Vibration*, Vol. **193**(1), pp. 137–147, (1996). DOI: https://doi.org/10.1006/jsvi.1996.0253.

[9] D. J. Mead. A general theory of harmonic wave propagation in linear periodic systems with multiple coupling. *Journal of Sound and Vibration*, Vol. **27**(2), pp. 235–260, (1973). DOI: https://doi.org/10.1016/0022-460X(73)90064-3.

[10] D. J. Thompson, M. H. A. Janssens and F. G. de Beer. Track Wheel Interaction Noise Software (TWINS) Theoretical Manual (version 3.4). TNO report, TNO Institute of Applied Physics, 2019.

[11] Railway applications - Wheelsets and bogies - Monobloc wheels - Technical approval procedure - Part 1: Forged and rolled wheels. EN 13979-1:2020. European Committee for Standardization, 2020.

[12] D. J. Thompson. Wheel-rail noise generation, part I: Introduction and interaction model. *Journal of Sound and Vibration*, Vol. **161**(3), pp. 387–400, (1993). DOI: https://doi.org/10.1006/jsvi.1993.1082.

[13] J. W. Pratt. Dividing the indivisible: Using simple symmetry to partition variance explained. *Proceedings of the Second International Conference in Statistics*, University of Tampere, Tampere, pp. 245–260, (1987).

A VIBROACOUSTIC MODEL OF THE STATIONARY RAILWAY WHEEL FOR SOUND RADIATION PREDICTION THROUGH AN AXISYMMETRIC APPROACH

V. T. Andrés*, J. Martínez-Casas, J. Carballeira and F. D. Denia

* Instituto de Ingeniería Mecánica y Biomecánica (I2MB)
Universitat Politècnica de València
Valencia, Spain
e-mail: vicanrui@etsid.upv.es, jomarc12@mcm.upv.es, jacarmo@mcm.upv.es, fdenia@mcm.upv.es

Key words: Stationary railway wheel dynamics, wheel axisymmetry, two-dimensional sound radiation model, computational efficiency.

Abstract: *In the literature, different dynamic models of the railway wheel have been developed to predict its sound radiation; however, there are still certain aspects that can be improved. Specifically, the high computational cost of these models, either because they solve the fluid-structure interaction or because they solve the dynamics and acoustics of the three-dimensional wheel, makes it difficult to carry out numerous simulations with the aim of achieving quieter designs. In the present work, a vibroacoustic model of the stationary wheel is developed through an axisymmetric approach, yielding an efficient and comprehensive acoustic prediction tool. The calculation methodology consists of, firstly, adopting an axisymmetric approach to solve the vibratory dynamics of the wheel from its cross-section, using finite element techniques; subsequently, the acoustic radiation of the three-dimensional wheel is calculated from the dynamics of the aforementioned section through an analytical formulation. Finally, the vibroacoustic model developed is validated via comparison with commercial software that solves the fluid-structure interaction, showing the aforementioned computational advantages that the former has over the latter.*

1 INTRODUCTION

Wheel/rail interaction generates a dynamic contact force due to the roughness of their surfaces. This excites the wheel causing a vibrational response which, in turn, leads to a sound radiation known as rolling noise. It is considered an important source of noise from railway activities [1], especially in urban areas where the vehicle velocity is relatively low [2]. The frequency range of interest for rolling noise radiation is approximately up to 6 kHz.

The interest in predicting the noise radiated by the railway wheel has resulted in the development of vibroacoustic models [3]; in general, the sound radiation is evaluated through the vibrational field of the wheel boundary. The wheel dynamic behaviour is commonly reproduced by the Finite Element Method (FEM) [4], which allows considering the flexibility of the body. Given the wheel geometry axisymmetry, a Fourier series expansion is feasible [5], solving analytically the vibrational response in the circumferential direction and therefore reducing the associated computational cost.

In this work, a vibroacoustic model of the axisymmetric wheel is presented. The description of the dynamic response of the wheel along the circumferential direction by means of Fourier series establishes a similar distribution of its modal properties. By adopting a modal approach, analytical relations between the vibrational field on the wheel boundary and on the wheel cross-section are found. This allows computing the acoustic problem also in a two-dimensional frame, with the computational advantages that it entails.

The mathematical formulation of the vibroacoustic model is presented in Section 2. The

https://doi.org/10.4995/YIC2021.2021.12584

results of this model are compared with results from a commercial software in Section 3. Finally, in Section 4 some conclusions are summarized.

2 VIBROACOUSTIC MODEL

2.1 Dynamics

Considering a cylindrical reference system, the displacement field of an axisymmetric wheel due to its flexible behaviour is expanded along the circumferential direction using Fourier series as follows [5]:

$$\begin{aligned} u_r =& u_{r,0} + \sum_{n>0} \left(u_{r,n}\cos(n\theta) - \overline{u}_{r,n}\sin(n\theta)\right), \\ u_\theta =& -\overline{u}_{\theta,0} + \sum_{n>0} \left(u_{\theta,n}\sin(n\theta) - \overline{u}_{\theta,n}\cos(n\theta)\right), \\ u_z =& u_{z,0} + \sum_{n>0} \left(u_{z,n}\cos(n\theta) - \overline{u}_{z,n}\sin(n\theta)\right), \end{aligned} \tag{1}$$

where subscripts r, θ and z indicate radial, tangential and axial direction, respectively. In this expansion, harmonic amplitudes without bar represent symmetric displacements about $\theta = 0$ and those with a bar represent antisymmetric displacements about $\theta = 0$, θ being the circumferential coordinate; all harmonic amplitudes are function of the coordinates r and z. Variable n symbolises each Fourier term. Similarly, the external forces applied on the flexible wheel can be expanded as Fourier series.

Making use of the expansion in Eq. (1), the kinetic energy of the flexible wheel E_k is analytically integrated over the circumferential direction. After that, it can be proved that the kinetic energy can be divided into the contribution of each motion associated with a Fourier harmonic and this, in turn, into the symmetric and antisymmetric displacements about $\theta = 0$. Thus, the kinetic energy can expressed as follows:

$$E_k = E_{k,0} + \overline{E}_{k,0} + \sum_{n>0} E_{k,n} + \sum_{n>0} \overline{E}_{k,n}. \tag{2}$$

Similarly, the strain energy of the wheel accomplishes the following expression:

$$E_p = E_{p,0} + \overline{E}_{p,0} + \sum_{n>0} E_{p,n} + \sum_{n>0} \overline{E}_{p,n}. \tag{3}$$

Applying the Lagrange Equations, a set of Equations of Motion (EoM) are obtained; each of these describes the motion associated with a Fourier term and a type of motion (symmetric or antisymmetric). The set of EoM for $n = 0$, considering a FE approach for the wheel cross-section, is given by:

$$\begin{aligned} \mathbf{M}_0\ddot{\mathbf{u}}_0 + \mathbf{K}_0\mathbf{u}_0 &= \mathbf{F}_0, \\ \overline{\mathbf{M}}_0\ddot{\overline{\mathbf{u}}}_0 + \overline{\mathbf{K}}_0\overline{\mathbf{u}}_0 &= \overline{\mathbf{F}}_0, \end{aligned} \tag{4}$$

where $\mathbf{u}_0$ contains the amplitudes $u_{r,0}$ and $u_{z,0}$ for each node of the wheel cross-section mesh while $\overline{\mathbf{u}}_0$ contains the amplitudes $\overline{u}_{\theta,0}$. The force vectors come from expanding the wheel/rail interaction force; $\mathbf{F}_0$ represents the even Fourier coefficients and $\overline{\mathbf{F}}_0$ the odd coefficients, both defined for $n = 0$. Similarly, the EoM for motion with $n > 0$ are given by:

$$\begin{aligned} \mathbf{M}_n\ddot{\mathbf{u}}_n + \mathbf{K}_n\mathbf{u}_n &= \mathbf{F}_n, \\ \overline{\mathbf{M}}_n\ddot{\overline{\mathbf{u}}}_n + \overline{\mathbf{K}}_n\overline{\mathbf{u}}_n &= \overline{\mathbf{F}}_n, \end{aligned} \tag{5}$$

where $\mathbf{u}_n$ contains the symmetric amplitudes $u_{r,n}$, $u_{\theta,n}$ and $u_{z,n}$ for each node of the wheel cross-section while $\overline{\mathbf{u}}_n$ contains the antisymmetric amplitudes $\overline{u}_{r,n}$, $\overline{u}_{\theta,n}$ and $\overline{u}_{z,n}$. Matrices accomplish the following relations:

$$\begin{aligned} \overline{\mathbf{M}}_n &= \mathbf{M}_n, \\ \overline{\mathbf{K}}_n &= \mathbf{K}_{-n}. \end{aligned} \tag{6}$$

Matrix $\mathbf{M}_n$ is indeed independent of n whereas $\mathbf{K}_n$ is defined for each n.

A modal approach is adopted in order to solve the dynamics of the flexible wheel. A set of modes coming from the EoM defined for a certain n is described in the literature as modes with n nodal diameters [2]. The eigenproblem of the EoM for $n = 0$ gives as a result a set of radial and axial modes with zero nodal diameters for the case of symmetric motion and a set of circumferential modes with zero nodal diameters for the case of antisymmetric motion. When considering the EoM for $n > 0$, for each vibration mode coming from the symmetric EoM, an analogous mode is obtained from the antisymmetric EoM, both being in quadrature of phase and with the same natural frequency. The wheel modeshapes can be also decomposed into harmonic functions with the angular coordinate similar to Eq. (1).

After solving the modal problem, the dynamic response of the wheel due to the contact force from the wheel/rail interaction is solved by modal superposition. Details of the interaction model can be found in [6]; in this work, the radial and axial directions are solved in the interaction problem. Although damping matrix is not considered in the EoM, spectral damping is introduced in the model as proposed by Thompson in [2], where it is suggested that modes with $n = 0$ have $\xi = 10^{-3}$, modes with $n = 1$ have $\xi = 10^{-2}$ and modes with $n \geq 2$ have $\xi = 10^{-4}$. Thus, the velocity of a point of the wheel with coordinates (r, θ, z) formulated in the frequency domain ω is given by:

$$v_j(r,\theta,z,\omega) = \sum_{p=1}^{m} A^p(\omega)\phi_j^p(r,\theta,z)\left(\sum_{k=r,z} F_k(\omega)\phi_{k,c}^p\right), \quad j = r,\theta,z, \tag{7}$$

where superscript p represents the pth vibration mode, m is the number of modes considered as a basis of the response, ϕ_j^p is the pth modeshape particularized in the jth Degree of Freedom (DoF) of the point, F_k is the kth component of the interaction force, $\phi_{k,c}^p$ is the pth modeshape particularized in the kth DoF of the wheel contact point and A^p is defined as:

$$A^p(\omega) = \frac{\mathrm{i}\omega}{\omega_p^2 - \omega^2 + 2\mathrm{i}\xi_p\omega_p\omega}, \tag{8}$$

with ω_p and ξ_p being the natural frequency and damping ratio, respectively, of the pth vibration mode.

2.2 Sound radiation

In this work, the acoustic model developed by Thompson [3] is employed. The sound radiation of the wheel is evaluated by postprocessing the vibrational field on its surface. Particularly, this model states that the acoustic power is the sum of the power associated with each set of modes with the same number of nodal diameters n. Thus, the sound power of the wheel W is given by:

$$W(\omega) = \rho c \sum_n \left(\sigma_{z,n}(\omega) S_z \widetilde{v_{z,n}^2}(\omega) + \sigma_{r,n}(\omega) S_r \widetilde{v_{r,n}^2}(\omega)\right), \tag{9}$$

where ρ is the density of air and c is the speed of sound. Functions σ are the radiation ratios and a set of fitting expressions for them is proposed in [3]. Surfaces S_z and S_r are the projected

surfaces of the wheel normal to the axial and radial direction, respectively. Squared velocities $v_{z,n}^2$ and $v_{r,n}^2$ are the projected velocities in the axial and radial direction, respectively, and they are averaged over time ($\widetilde{\ }$) and over the wheel surface ($\overline{\ }$). The former is evaluated in the frequency domain as the Root Mean Square (RMS) value of the velocity amplitude v whereas the latter is computed as an integral given by:

$$\overline{\widetilde{v}_{j,n}^2} = \frac{1}{S_j} \int_S \widetilde{v}_{j,n}^2 dS_j, \quad j = r, z, \tag{10}$$

where S is the wheel surface. Note that $\widetilde{v}_{j,n}$ is the contribution to the velocity of a set of modes with n nodal diameters, including both symmetric and antisymmetric ones, and it can be computed through the modal superposition approach presented in Eq. (7), where m is replaced by m_n, the last being the number of modes with n nodal diameters. The integral in Eq. (10) can be divided into an integral over the circumferential direction and an integral over the wheel cross-section boundary. The former can be evaluated analytically by means of Eq. (7) and the circumferential expansion of the displacements in Eq. (1). Furthermore, by developing this over the wheel modeshapes, some relations are found between the vibrational field of the three-dimensional wheel and the response of the wheel cross-section. Finally, a numerical approach based on the FEM for the cross-section boundary is performed to complete the evaluation of the integral in Eq. (10).

3 RESULTS

The vibroacoustic model presented in Section 2 is compared with the commercial software Ansys. To perform this comparison, the Frequency Response Function (FRF) of the contact point and the Sound poWer Level (SWL) of the wheel are evaluated with both approaches and the results are shown in this section. The fluid-structure interaction model available in Ansys computes the acoustic pressure field in the air surrounding the wheel, which requires a high computational cost as the number of DoF increases. In this work, a straight web wheel with a diameter of 900 mm and a S1002 profile [7] is considered, as well as a load per wheel of 50 kN. The receptances of the wheel contact point, computed with Ansys and with the proposed model, are shown in Figure 1.

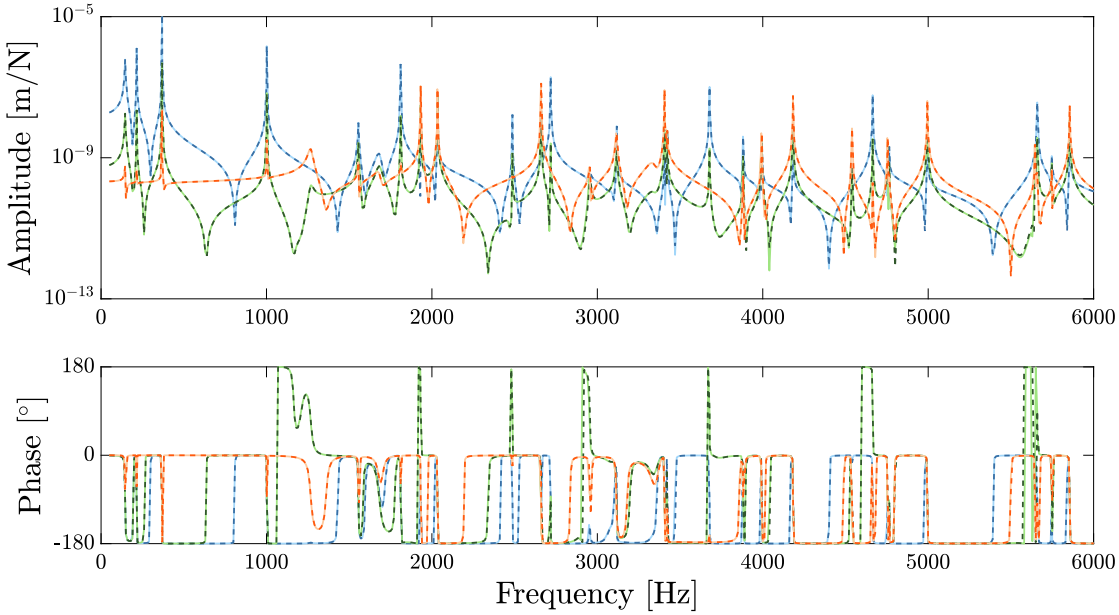


Figure 1: Receptances in the wheel contact point from Ansys (——) and axisymmetric approach (- - -): direct axial/axial (blue), cross axial/radial (green) and direct radial/radial (orange).

For the purpose of comparing the acoustic results from both approaches, the sound power radiated by the wheel is evaluated considering unit roughness excitation, the result being therefore a transfer function. The SWL of the wheel is shown in Figure 2. The greater differences between the proposed model and Ansys software appear at low and medium frequencies, where the sound power levels are low and the radiation ratios influence is important; at high frequencies, where the radiated levels are greater, the presented vibroacoustic model predicts the SWL accurately. The proposed model needs approximately 15 seconds for solving the vibroacoustic problem while Ansys software requires more than 24 hours, using a PC running with an ®Intel i7-9700 processor with 64 GB RAM.

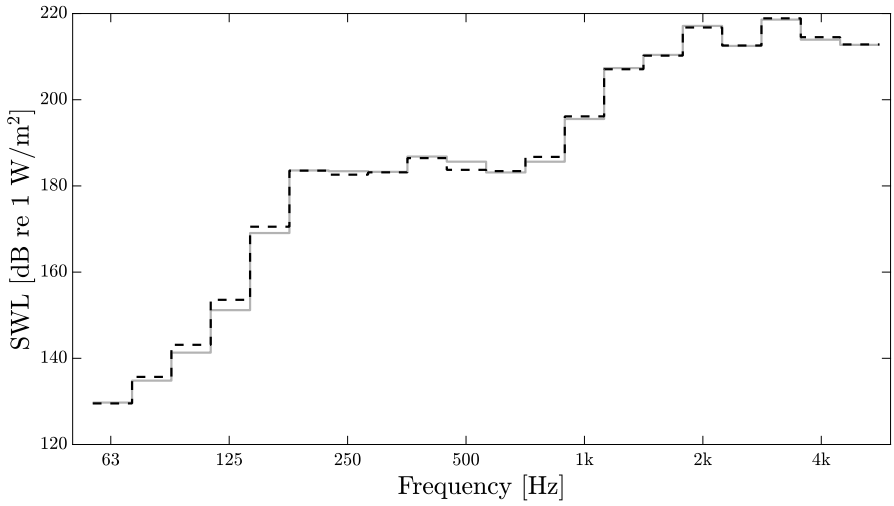


Figure 2: Wheel sound power level from Ansys (——) and axisymmetric approach (- - -).

4 CONCLUSIONS

A vibroacoustic model for a stationary and axisymmetric railway wheel is presented in this work, in which the displacement field is expanded using Fourier series. This allows solving analytically the dynamics and acoustics of the wheel along the circumferential direction, reducing the computational cost associated with numerical calculations. Also, the formulation developed leads to some relations between the dynamic response of the three-dimensional wheel and the response of the wheel cross-section, making it possible to employ the proposed acoustic methodology in combination with alternative three-dimensional dynamic models instead of that presented here. The vibroacoustic model is compared with the commercial FE package Ansys, which solves the fluid-structure interaction problem. The dynamic response and sound radiation results of both approaches show an excellent agreement, the computational cost of the presented methodology being much lower.

5 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of Agencia Estatal de Investigación and European Regional Development Fund (grant FPU18/03999, project TRA2017-84701-R and project PID2020-112886RA-I00).

REFERENCES

[1] X. Zhang, D. J. Thompson, E. Quaranta and G. Squicciarini. An engineering model for the prediction of the sound radiation from a railway track. *Journal of Sound and Vibration*, Vol. **461**, 114921, (2019). DOI: https://doi.org/10.1016/j.jsv.2019.114921.

[2] D. J. Thompson. *Railway Noise and Vibration. Mechanisms, Modelling and Means of Control*. Elsevier, 2009. ISBN: 978-0-08-045147-3.

[3] D. J. Thompson and C. J. C. Jones. Sound radiation from a vibrating railway wheel. *Journal of Sound and Vibration*, Vol. **253**(2), pp. 401-–419, (2002). DOI: https://doi.org/10.1006/jsvi.2001.4061.

[4] D. J. Thompson. Wheel-rail noise generation, part II: Wheel vibration. *Journal of Sound and Vibration*, Vol. **161**(3), pp. 401-–419, (1993). DOI: https://doi.org/10.1006/jsvi.1993.1083.

[5] M. Petyt. *Introduction to Finite Element Vibration Analysis*. Cambridge University Press, 2nd Edition, 2010. ISBN: 9780521191609.

[6] D. J. Thompson. Wheel-rail noise generation, part I: Introduction and interaction model. *Journal of Sound and Vibration*, Vol. **161**(3), pp. 387-–400, (1993). DOI: https://doi.org/10.1006/jsvi.1993.1082.

[7] Railway applications - Wheelsets and bogies - Wheels - Tread profile. EN 13715:2020. European Committee for Standardization, 2020.

Dynamic response of periodic infinite structure to arbitrary moving load based on the Finite Element Method

J. Gil*, S. Gregori, M. Tur and F.J. Fuenmayor

Instituto de Ingeniería Mecánica y Biomecánica (I2MB)
Universitat Politècnica de València
Valencia, Spain
e-mail: jaigiro@upv.es*, sangreve@upv.es, manuel.tur@mcm.upv.es and ffuenmay@mcm.upv.es

Key words: Infinite structure, Moving load, Periodic structure, Steady state

Abstract: *A common problem in railway engineering is the dynamic of repetitive structures subject to moving loads. Bridges, rails or catenaries are the most representative periodic structures, over which the train acts as a moving exciter. Usually, these structures are long enough to consider that their dynamic response is in permanent regime. To assume the steady-state regime some features have to be considered: infinite length structure, perfect periodicity and constant velocity of the moving load. This paper adopts these assumptions and provides the steady-state solution of a generic periodic structure subject to an arbitrary and also periodic moving load.*

The structure is divided into repetitive blocks modelled by the Finite Element Method. By applying the periodicity condition it is possible to consider the entire structure dynamics with only one block. The problem is stated in the frequency domain and moved back to time domain by means of Discrete Fourier Transform.

1 INTRODUCTION

The study of periodic structures subject to moving loads has a great relevance thanks to the wide use of high-speed trains. Rails, overhead contact lines or bridges are periodic structures whose dynamic response produced by the train has been studied under different approaches. The authors who consider an infinite periodic structure focus on the steady-state solution of the problem. The early analytic models found in the literature are based on an infinite continuous periodically supported string/beam [1, 2, 3]. In [1] an infinite periodic Euler-Bernoulli beam subject to a uniform moving harmonic pressure field is solved. The differential equation is solved in the domain between two supports and four boundary conditions allow to determine the coefficients of the solution. Boundary conditions are obtained from the periodicity condition of two consecutive supports and the momentum and shear equilibrium at these supports. In [2] a similar model subject to a constant moving load is solved using the modal method. A finite periodic supported beam is defined by N uncoupled differential equations based on a modal representation. The limit of the previous solution when $N \to \infty$ is computed for a moving constant load. The same problem is solved in [3], in which the Fourier Transform is used to shift to the frequency domain where the periodicity condition is easily formulated. The solution is obtained in the frequency domain and the Inverse Fourier Transform allows to obtain the response in the time domain. The presented approaches have in common the consideration of a periodic solution which allows considering only a single period or block of the string/beam between two consecutive supports.

The limitation of the previous references is their inability of modelling more complex structures. Some solutions have been found, for example in [4], in which an extension of the approach proposed in [3] is presented to solve a catenary model, including two strings and two spatial periods, one for supports and another for droppers. In [5], the beam is modelled by a two-and-a-half dimensional (2.5D) Finite Element model which allows to model any cross section of the

https://doi.org/10.4995/YIC2021.2021.12606

beam. The solution is divided into the response produced by the external load and the response produced by the reactions of the supports. Fourier Transform respect to position x and time t is performed to solve the differential equation and the periodicity condition is applied to the reactions of the supports in the frequency domain. The same authors presented an improved model in [6] in which the dynamic interaction of multiple wheels with the periodic model is computed by means of the Fourier Series decomposition of the contact force.

The Finite Element Method (FEM) can be used to model any periodic structure by means of the so-called Wave Finite Element Method (WFEM). This method allows to compute the frequency response of finite or infinite periodic structures [7, 8]. The frequency response of a periodic infinite structure obtained by WFEM can be used to compute the response under a moving load by means of the Fourier Transform [9]. WFEM makes possible to model finite-length structures and even structures with transition zones [8], but for periodic infinite structures we present an alternative in which some inconveniences of WFEM are avoided. For example, some slender structures (as catenaries) present ill conditioning behaviour in WFEM.

In this paper, the periodicity condition is applied on FEM models to obtain the frequency response of any generic periodic infinite structure. Then, the response to a temporal excitation is obtained by means of the Discrete Fourier Transform (DFT). Finally, the pantograph-catenary dynamic interaction is solved with this method.

2 HARMONIC RESPONSE

In this section we obtain the harmonic response of the model as a tool for the computation of the steady-state response. Let consider an infinite structure with a periodic pattern along the longitudinal axis as in Fig. 1. The repeated block is called substructure and it is modelled by the FEM. The dynamic equation of the substructure for a harmonic load can be written as:

$$\mathbf{D}(\omega)\mathbf{u} = \mathbf{F} \tag{1}$$

in which $\mathbf{u}$ is the nodal displacement vector, $\mathbf{D}(\omega) = \mathbf{K} + i\omega\mathbf{C} - \omega^2\mathbf{M}$ is the dynamic stiffness matrix of the substructure and $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ are the mass, damping and stiffness matrices, respectively. Note that the force vector $\mathbf{F}$ includes external forces and the reactions produced by the adjacent blocks. The nodes of the block can be divided into left (L) and right (R) boundary nodes and inner (I) nodes according to their positions. Thus, the previous equation can be split into:

$$\begin{bmatrix} \mathbf{D}_{\mathrm{LL}} & \mathbf{D}_{\mathrm{LI}} & \mathbf{D}_{\mathrm{LR}} \\ \mathbf{D}_{\mathrm{IL}} & \mathbf{D}_{\mathrm{II}} & \mathbf{D}_{\mathrm{IR}} \\ \mathbf{D}_{\mathrm{RL}} & \mathbf{D}_{\mathrm{RI}} & \mathbf{D}_{\mathrm{RR}} \end{bmatrix} \begin{Bmatrix} \mathbf{u}_{\mathrm{L}} \\ \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}_{\mathrm{L}} \\ \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\mathrm{R}} \end{Bmatrix} \tag{2}$$

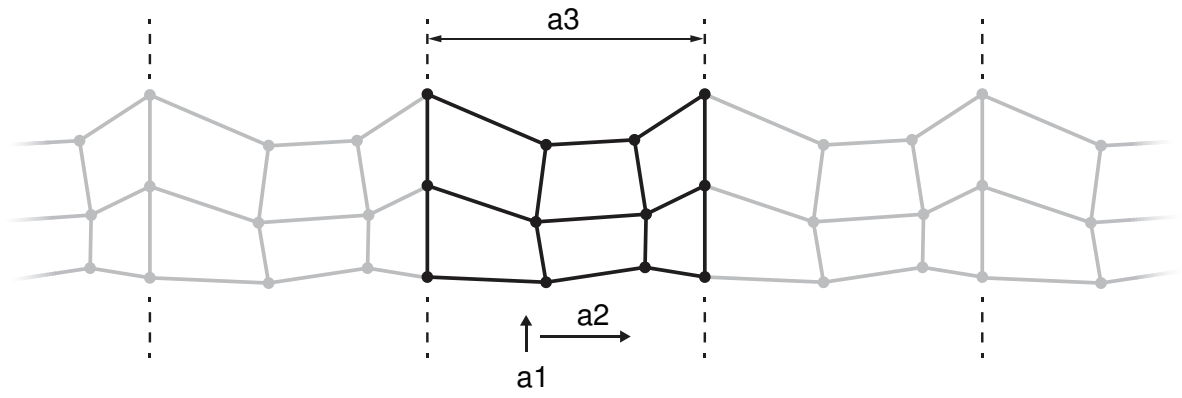


Figure 1: Periodic infinite FEM structure with moving load.

It is assumed that the load is repeated at every block so that the response of all blocks is identical but with a time lag that depends on the length L of the substructure and the velocity v of the moving load. This condition is called the periodicity condition and for the displacement u of any point it reads:

$$u(x,t) = u(x + nL, t + nL/v); \qquad n \in \mathbb{Z} \tag{3}$$

This condition allows to state the entire problem only in a single block of the structure, which is called the reference block. The periodicity condition can be moved to the frequency domain in which, the response of the next block to the reference one is:

$$\mathbf{u}^{next} = e^{-\frac{i\omega L}{v}}\mathbf{u} \tag{4}$$

Both blocks hold the following coupling condition in the common boundary:

$$\mathbf{u}_{\mathrm{R}} = \mathbf{u}_{\mathrm{L}}^{next} \tag{5}$$

so that the displacement of the left and right nodes of every substructure are related by:

$$\mathbf{u}_{\mathrm{L}} = e^{\frac{i\omega L}{v}}\mathbf{u}_{\mathrm{R}} \tag{6}$$

Applying this relation to Eq. (2):

$$\begin{bmatrix} \mathbf{D}_{\mathrm{LI}} & \mathbf{D}_{\mathrm{LR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{LL}} \\ \mathbf{D}_{\mathrm{II}} & \mathbf{D}_{\mathrm{IR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{IL}} \\ \mathbf{D}_{\mathrm{RI}} & \mathbf{D}_{\mathrm{RR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{RL}} \end{bmatrix} \begin{Bmatrix} \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}_{\mathrm{L}} \\ \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\mathrm{R}} \end{Bmatrix} \tag{7}$$

The same procedure can be considered for the nodal forces:

$$\mathbf{F}^{next} = e^{-\frac{i\omega L}{v}}\mathbf{F} \tag{8}$$

which must satisfy the action-reaction principle in the boundary:

$$\mathbf{F}_{\mathrm{R}} = \mathbf{F}_{\partial \mathrm{R}} - \mathbf{F}_{\mathrm{L}}^{next} \tag{9}$$

in which $\mathbf{F}_{\partial \mathrm{R}}$ is the external load at the right boundary. By combining Eqs. (8) and (9) the left and right nodal forces of every substructure can be related by:

$$\mathbf{F}_{\mathrm{L}} = e^{\frac{i\omega L}{v}}\left(\mathbf{F}_{\partial \mathrm{R}} - \mathbf{F}_{\mathrm{R}}\right) \tag{10}$$

Introducing this constraint in Eq. (7) it becomes into:

$$\begin{bmatrix} \mathbf{D}_{\mathrm{LI}} & \mathbf{D}_{\mathrm{LR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{LL}} \\ \mathbf{D}_{\mathrm{II}} & \mathbf{D}_{\mathrm{IR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{IL}} \\ \mathbf{D}_{\mathrm{RI}} & \mathbf{D}_{\mathrm{RR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{RL}} \end{bmatrix} \begin{Bmatrix} \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \end{Bmatrix} = \begin{bmatrix} \mathbf{0} & e^{\frac{i\omega L}{v}}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\partial \mathrm{R}} \end{Bmatrix} + \begin{bmatrix} -e^{\frac{i\omega L}{v}}\mathbf{I} \\ \mathbf{0} \\ \mathbf{I} \end{bmatrix} \mathbf{F}_{\mathrm{R}} \tag{11}$$

If all the unknowns are moved to the left-hand side,

$$\begin{bmatrix} \mathbf{D}_{\mathrm{LI}} & \mathbf{D}_{\mathrm{LR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{LL}} & e^{\frac{i\omega L}{v}}\mathbf{I} \\ \mathbf{D}_{\mathrm{II}} & \mathbf{D}_{\mathrm{IR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{IL}} & \mathbf{0} \\ \mathbf{D}_{\mathrm{RI}} & \mathbf{D}_{\mathrm{RR}} + e^{\frac{i\omega L}{v}}\mathbf{D}_{\mathrm{RL}} & -\mathbf{I} \end{bmatrix} \begin{Bmatrix} \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \\ \mathbf{F}_{\mathrm{R}} \end{Bmatrix} = \begin{bmatrix} \mathbf{0} & e^{\frac{i\omega L}{v}}\mathbf{I} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\partial \mathrm{R}} \end{Bmatrix} \tag{12}$$

the displacement and the nodal forces are given by:

$$\left\{ \begin{array}{c} \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \\ \mathbf{F}_{\mathrm{R}} \end{array} \right\} = \hat{\mathbf{H}}(\omega) \left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\partial \mathrm{R}} \end{array} \right\} \tag{13}$$

in which

$$\hat{\mathbf{H}}(\omega) = \left[\begin{array}{ccc} \mathbf{D}_{\mathrm{LI}} & \mathbf{D}_{\mathrm{LR}} + e^{\frac{i\omega L}{v}} \mathbf{D}_{\mathrm{LL}} & \mathbf{I} e^{\frac{i\omega L}{v}} \\ \mathbf{D}_{\mathrm{II}} & \mathbf{D}_{\mathrm{IR}} + e^{\frac{i\omega L}{v}} \mathbf{D}_{\mathrm{IL}} & \mathbf{0} \\ \mathbf{D}_{\mathrm{RI}} & \mathbf{D}_{\mathrm{RR}} + e^{\frac{i\omega L}{v}} \mathbf{D}_{\mathrm{RL}} & -\mathbf{I} \end{array} \right]^{-1} \left[\begin{array}{cc} \mathbf{0} & \mathbf{I} e^{\frac{i\omega L}{v}} \\ \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right] \tag{14}$$

Eq. (13) can be rewritten in terms of displacements of the substructure:

$$\left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}} \\ \mathbf{u}_{\mathrm{I}} \\ \mathbf{u}_{\mathrm{R}} \end{array} \right\} = \mathbf{H}(\omega) \left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}} \\ \mathbf{F}_{\partial \mathrm{R}} \end{array} \right\} \tag{15}$$

in which

$$\mathbf{H}(\omega) = \left[\begin{array}{c} e^{\frac{i\omega L}{v}} \hat{\mathbf{H}}_{\mathrm{R}}(\omega) \\ \hat{\mathbf{H}}_{\mathrm{I}}(\omega) \\ \hat{\mathbf{H}}_{\mathrm{R}}(\omega) \end{array} \right] \tag{16}$$

being $\hat{\mathbf{H}}_{\mathrm{I}}(\omega)$ and $\hat{\mathbf{H}}_{\mathrm{R}}(\omega)$ the two first rows of $\hat{\mathbf{H}}(\omega)$.

3 TIME-FREQUENCY ANALYSIS

To achieve the response to an arbitrary moving load, the excitation is moved to the frequency domain where the frequency response obtained in the previous section can be used. Then, the response in the frequency domain is moved back to the time domain.

The structure is excited by a moving load $\mathbf{f}(t)$ with period L/v. This moving load causes the nodal forces $\mathbf{F}_{\mathrm{I}}(t)$ and $\mathbf{F}_{\partial \mathrm{R}}(t)$ which are evaluated in N discrete times $t_n = n\Delta t$.

$$\left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}}(t_n) \\ \mathbf{F}_{\partial \mathrm{R}}(t_n) \end{array} \right\} = \left[\begin{array}{c} \mathbf{N}_{\mathrm{I}}^{\top}(t_n) \\ \mathbf{N}_{\mathrm{R}}^{\top}(t_n) \end{array} \right] \mathbf{f}(t_n) \tag{17}$$

$\mathbf{N}_{\mathrm{I}}(t_n)$ and $\mathbf{N}_{\mathrm{R}}(t_n)$ are the shape functions of the inner and right nodes evaluated at time t_n. These functions are used in FEM to transform nodal displacements into point displacements and it can also be used to transform point forces to nodal equivalent forces.

The Discrete Fourier Transform (DFT) is used to obtain the frequency representation of the nodal forces:

$$\left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}}(\omega_k) \\ \mathbf{F}_{\partial \mathrm{R}}(\omega_k) \end{array} \right\} = \sum_{n=0}^{N-1} \left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}}(t_n) \\ \mathbf{F}_{\partial \mathrm{R}}(t_n) \end{array} \right\} e^{\frac{-i2\pi kn}{N}} \tag{18}$$

in which

$$\omega_k = k\frac{2\pi}{N\Delta t}; \qquad k \in [0, N-1] \tag{19}$$

The DFT considers that the temporal function is N-periodic, thus a long enough sequence (high N) is necessary to ensure a negligible influence of other periods. Note that the block periodicity of the moving load is different from the periodicity of nodal forces, which is fictitious, created by the discrete analysis with Fourier.

As the moving load is repeated in every block, Eq. (18) can be used with times t_n in which the moving load is acting on the reference block, from $t_n = 0$ to $t_n = t_{M-1}$ with $M = L/(v\Delta t)$.

At the instants in which $n \geq M$, the load $\mathbf{F}_{\mathrm{I}}(t_n) = 0$ and $\mathbf{F}_{\partial \mathrm{R}}(t_n) = \mathbf{F}_{\partial \mathrm{L}}(t_{n-M})$, being $\mathbf{F}_{\partial \mathrm{L}}$ the external load at the left boundary. Then, Eq. (18) can be written as:

$$\left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}}(\omega_k) \\ \mathbf{F}_{\partial \mathrm{R}}(\omega_k) \end{array} \right\} = \sum_{n=0}^{M-1} \left\{ \begin{array}{c} \mathbf{F}_{\mathrm{I}}(t_n) \\ \mathbf{F}_{\partial \mathrm{R}}(t_n) + \mathbf{F}_{\partial \mathrm{L}}(t_n) e^{\frac{-i2\pi kM}{N}} \end{array} \right\} e^{\frac{-i2\pi kn}{N}} \tag{20}$$

Eq. (15) allows to obtain the displacements of the substructure in the frequency domain. Now, the Inverse Discrete Fourier Transform (IDFT) is used to return to time domain, resulting in:

$$\left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}}(t_n) \\ \mathbf{u}_{\mathrm{I}}(t_n) \\ \mathbf{u}_{\mathrm{R}}(t_n) \end{array} \right\} = \frac{1}{N} \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} \left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}}(\omega_k) \\ \mathbf{u}_{\mathrm{I}}(\omega_k) \\ \mathbf{u}_{\mathrm{R}}(\omega_k) \end{array} \right\} e^{\frac{i2\pi kn}{N}} \tag{21}$$

In addition, if $\mathbf{F}(t_n)$ is a real function, $\mathbf{F}(\omega_k) = conj\left(\mathbf{F}(\omega_{-k})\right)$ and the same applies for $\mathbf{u}(\omega_k) = conj\left(\mathbf{u}(\omega_{-k})\right)$ because $\mathbf{H}(\omega)$ exhibits Hermitian symmetry. Then, the IDFT can be computed as:

$$\left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}}(t_n) \\ \mathbf{u}_{\mathrm{I}}(t_n) \\ \mathbf{u}_{\mathrm{R}}(t_n) \end{array} \right\} = \frac{1}{N} \left(\left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}}(\omega_0) \\ \mathbf{u}_{\mathrm{I}}(\omega_0) \\ \mathbf{u}_{\mathrm{R}}(\omega_0) \end{array} \right\} + 2\mathrm{Re} \left(\sum_{k=1}^{\frac{N}{2}} \left\{ \begin{array}{c} \mathbf{u}_{\mathrm{L}}(\omega_k) \\ \mathbf{u}_{\mathrm{I}}(\omega_k) \\ \mathbf{u}_{\mathrm{R}}(\omega_k) \end{array} \right\} e^{\frac{i2\pi kn}{N}} \right) \right) \tag{22}$$

It is also possible to truncate and consider only the N_c first frequencies if the effect of higher frequencies is negligible.

If Eq. (17) is introduced in Eq. (20) and the result in Eq. (15), then in Eq. (21) and finally in Eq. (22), a condensed formulation is obtained.

$$\mathbf{u}(t_n) = \sum_{\hat{n}=0}^{M-1} \mathbb{I}(n, \hat{n}) \mathbf{f}(t_{\hat{n}}) \tag{23}$$

in which

$$\mathbb{I}(n, \hat{n}) = \frac{1}{N} \sum_{k=0}^{N_c-1} a_k \mathrm{Re} \left(\mathbf{H}(\omega_k) \left[\begin{array}{ccc} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ e^{\frac{-i2\pi kM}{N}} \mathbf{I} & \mathbf{0} & \mathbf{I} \end{array} \right] e^{\frac{i2\pi k(n-\hat{n})}{N}} \right) \left[\begin{array}{c} \mathbf{N}_{\mathrm{L}}^{\top}(t_{\hat{n}}) \\ \mathbf{N}_{\mathrm{I}}^{\top}(t_{\hat{n}}) \\ \mathbf{N}_{\mathrm{R}}^{\top}(t_{\hat{n}}) \end{array} \right] \tag{24}$$

being $a_k = 2$ if $k \neq 0$ or $a_k = 1$ if $k = 0$. Note that the variable $\hat{n}$ is used to distinguish the instant of application of the load from the instant of evaluation of the displacement n.

To reduce the computational cost, $\mathbb{I}(n, \hat{n})$ can be written as:

$$\mathbb{I}(n, \hat{n}) = \mathbb{J}(\lambda) \mathbf{N}(t_{\hat{n}}) \tag{25}$$

in which

$$\mathbb{J}(\lambda) = \frac{1}{N} \sum_{k=0}^{N_c-1} a_k \mathrm{Re} \left(\mathbf{H}(\omega_k) \left[\begin{array}{ccc} \mathbf{0} & \mathbf{I} & \mathbf{0} \\ e^{\frac{-i2\pi kM}{N}} \mathbf{I} & \mathbf{0} & \mathbf{I} \end{array} \right] e^{\frac{i2\pi k\lambda}{N}} \right) \tag{26}$$

and $\lambda = n - \hat{n}$.

4 NUMERICAL EXAMPLE

In this section, a numerical example of application of this method is analysed. The pantograph-catenary dynamic interaction is solved under the hypothesis of steady-state behaviour. The infinite catenary is composed of repetitive blocks as shown in Fig. 2 and the pantograph applies a vertical load f_c on the contact wire.

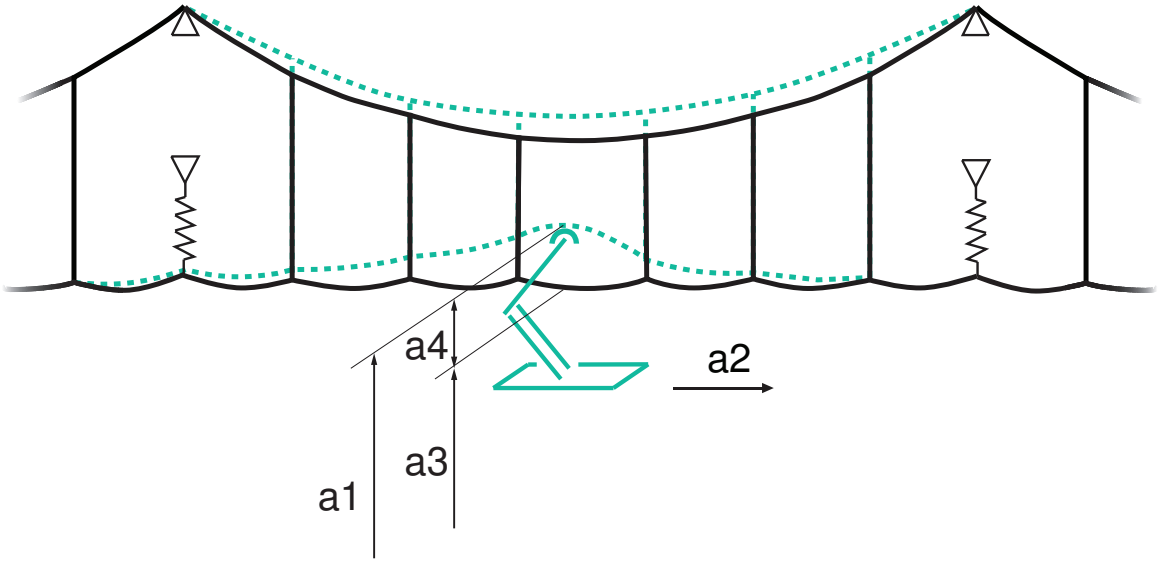


Figure 2: Pantograph interaction with the periodic catenary.

To solve the problem, Eq. (23) is only evaluated at the contact point instead of at all the degrees of freedom of the block. That is:

$$u_c(n) = \sum_{\hat{n}=0}^{M-1} \mathbb{I}_c(n, \hat{n}) f_c(t_{\hat{n}}) \tag{27}$$

in which $\mathbb{I}_c(n, \hat{n})$ is obtained from a simple transformation of $\mathbb{I}(n, \hat{n})$. This formulation provides a discrete matrix operator $\mathbb{I}_c(n, \hat{n})$ that relates the M values of the moving load with the M values of the contact point vertical displacement. The total height of the contact point is composed of the displacement produced by the load and the initial height profile of the contact wire:

$$z_c(n) = z_{cw}(n) + u_c(n) \tag{28}$$

A linear pantograph is considered whose dynamic response is defined by the frequency response function $H_p(\omega)$ of its contact point. This contact point is excited by an M-periodic force $-f_c(n)$ due to the action-reaction principle. The displacement $z_c(n)$ of the pantograph contact point can be also obtained by using the DFT:

$$z_c(n) = z_{ext} - \sum_{\hat{n}=0}^{M-1} \mathbb{I}_p(n, \hat{n}) f_c(t_{\hat{n}}) \tag{29}$$

in which z_{ext} is the displacement produced by a constant external load (produced by the bellow of the uplift mechanism) and:

$$\mathbb{I}_p(n, \hat{n}) = \frac{1}{M} \sum_{k=0}^{M/2} a_k \mathrm{Re}\left(H_p(\omega_k) e^{\frac{i2\pi k(n-\hat{n})}{M}} \right) \tag{30}$$

being:

$$\omega_k = k \frac{2\pi}{M\Delta t} \tag{31}$$

The contact force can be obtained imposing the same displacement $z_c(n)$ of the pantograph contact point (Eq. (29)) and the catenary contact point (Eq. (28)) at the M instants of time. That is:

$$z_{ext} - \sum_{\hat{n}=0}^{M-1} \mathbb{I}_p(n,\hat{n}) f_c(t_{\hat{n}}) = z_{cw}(n) + \sum_{\hat{n}=0}^{M-1} \mathbb{I}_c(n,\hat{n}) f_c(t_{\hat{n}}) \tag{32}$$

with $n = 0, .., M-1$. This system of M linear equations allows to compute the contact force $f_c(t_n)$.

As an example of a particular solution in which the dynamic interaction is produced at 300 km/h and the catenary model has 5 droppers per span, Fig. 3 gives a comparison between the solution obtained from the proposed method and the solution obtained from a FEM simulation performed with a long enough catenary to achieve the steady-state regime. The perfect coincidence shown gives validity to the proposed algorithm.

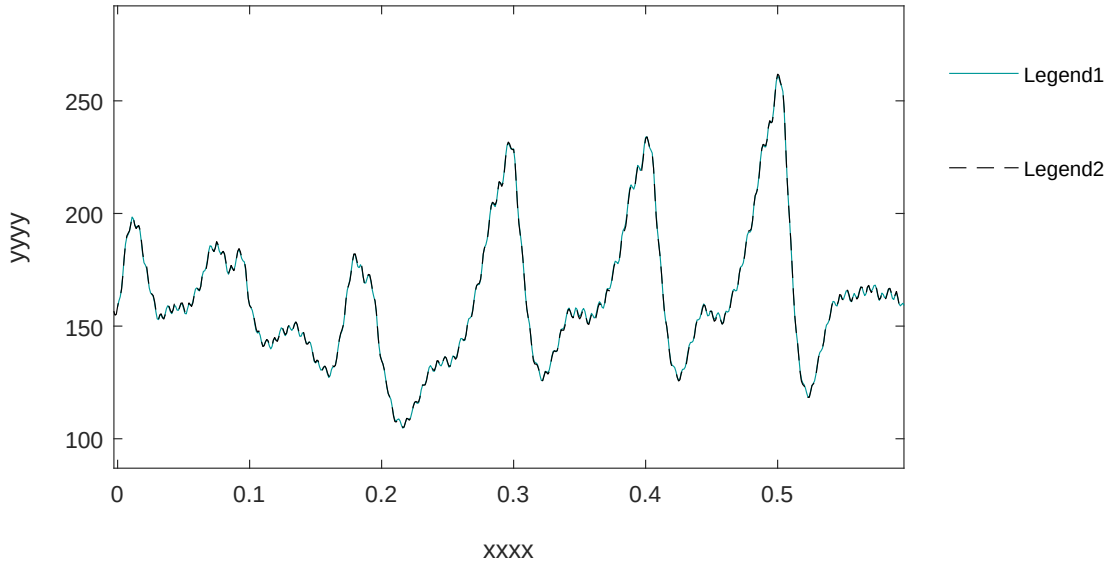


Figure 3: Solution of the infinite periodic model (PFEM) and a long conventional FEM model.

5 CONCLUSIONS

- The periodicity condition allows to compute the dynamic response of infinite periodic structures subject to a moving load combined with a FE model and discrete Fourier analysis.
- The repeated block is modelled by the Finite Element Method so that the proposed algorithm can be applied to any generic linear structure.
- The proposed method can be divided into two parts: the first devoted to compute the discrete operator that relates the load with the displacement of the structure and the second in which this operator is applied at different problems. The second part has a very low computational cost which makes it very suitable to perform (Hardware In the Loop) HIL tests or pantograph optimisation within the frame of pantograph-catenary dynamic interaction.

- A pantograph-catenary dynamic interaction problem is solved in this work to exemplify the proposed formulation. This has allowed us to validated the obtained results with a conventional FEM simulation, in which a long catenary must be used to ensure the steady-state interaction with the pantograph.

Acknowledgements

The authors would like to acknowledge the financial support received from the Spanish Ministry of Economy, Industry and Competitiveness (TRA2017-84736-R).

REFERENCES

[1] D. J. Mead, Vibration response and wave propagation in periodic structures, Journal of Engineering for Industry 93 (3) (1971) 783–792.

[2] C. Cai, Y. Cheung, H. Chan, Dynamic response of infinite continuous beams subjected to a moving force—an exact method, Journal of Sound and Vibration 123 (3) (1988) 461 – 472.

[3] P. Belotserkovskiy, On the oscillations of infinite periodic beams subjected to a moving concentrated force, Journal of Sound and Vibration 193 (3) (1996) 705 – 712.

[4] A. Metrikine, A. Bosch, Dynamic response of a two-level catenary to a moving load, Journal of Sound and Vibration 292 (3) (2006) 676 – 693.

[5] X. Sheng, C. Jones, D. Thompson, Responses of infinite periodic structures to moving or stationary harmonic loads, Journal of Sound and Vibration 282 (2005) 125–149.

[6] X. Sheng, C. Jones, D. Thompson, Using the fourier-series approach to study interactions between moving wheels and a periodically supported rail, Journal of Sound and Vibration 303 (2007) 873–894.

[7] J.-M. Mencik, On the low- and mid-frequency forced response of elastic structures using wave finite elements with one-dimensional propagation, Computers & Structures 88 (11) (2010) 674 – 689.

[8] B. Claudet, T. Hoang, D. Duhamel, G. Foret, J.-L. Pochet, F. Sabatier, Wave finite element method for computing the dynamic response of railway transition zones subjected to moving loads, 2019, pp. 4538–4547.

[9] T. Hoang, D. Duhamel, G. Forêt, J.-L. L. Pochet, F. Sabatier, Wave Finite Element Method and moving loads for the dynamic analysis of railway tracks, in: 13th World Congress on Computational Mechanics (WCCM XIII), New York, United States, 2018.

RECENT ADVANCES IN STABILISED METHODS FOR FLOW PROBLEMS

EFFICIENT AND HIGHER-ORDER ACCURATE SPLIT-STEP METHODS FOR GENERALISED NEWTONIAN FLUID FLOW

R. Schussnig[*,†,‡], D. R. Q. Pacheco[§,‡], M. Kaltenbacher[#,‡] and T.-P. Fries[†,‡]

[†] Institute of Structural Analysis, Graz University of Technology,
Lessingstraße 25/II, 8010 Graz, Austria

[§] Institute of Applied Mathematics, Graz University of Technology,
Steyrergasse 30/III, 8010 Graz, Austria

[#] Institute of Fundamentals and Theory in Electrical Engineering,
Graz University of Technology, Inffeldgasse 18, 8010 Graz, Austria

[‡] Graz Center of Computational Engineering, Krenngasse 37/I, 8010 Graz, Austria

e-mail: schussnig@tugraz.at, pacheco@math.tugraz.at, manfred.kaltenbacher@tugraz.at,
fries@tugraz.at

Key words: finite element method, generalised Newtonian fluid, incompressible flow, Navier-Stokes equations, split-step scheme, time-splitting method

Abstract: *In numerous engineering applications, such as polymer or blood flow, the dependence of fluid viscosity on the local shear rate plays an important role. Standard techniques using inf-sup stable finite elements lead to saddle-point systems posing a challenge even for state-of-the-art solvers and preconditioners. Alternatively, projection schemes or time-splitting methods decouple equations for velocity and pressure, resulting in easier to solve linear systems. Although pressure and velocity correction schemes of high-order accuracy are available for Newtonian fluids, the extension to generalised Newtonian fluids is not a trivial task. Herein, we present a split-step scheme based on an explicit-implicit treatment of pressure, viscosity and convection terms, combined with a pressure Poisson equation with fully consistent boundary conditions. Then, using standard equal-order finite elements becomes possible. Stability, flexibility and efficiency of the splitting scheme is showcased in two challenging applications involving aortic aneurysm flow and human phonation.*

1 INTRODUCTION

Various engineering and industrial applications such as automotive design, wind or hydraulic power production, medical devices or synthetics manufacturing share incompressible viscous flow as a central element. The modeling and simulation of fluids has thus been of great interest even before the beginning of computer aided design. More often than not, such fluids are modeled assuming a linear relationship between shear rate and viscous stress via constant viscosity. As it turns out, this modeling assumption may be invalid in various scenarios, blood and polymer flows being practically relevant examples. A vast majority of numerical schemes, however, focuses on Newtonian fluids, neglecting these effects. Depending on the problem and specific flow regime, non-Newtonian characteristics can heavily impact the results obtained and conclusions drawn from them [1, 2]. The most popular approach to incorporate phenomena such as plug flow or shear thinning/thickening is to consider the viscosity depending on the shear rate, leading to so-called generalised Newtonian or quasi-Newtonian assumptions.

Driven by the ever increasing demand, numerical treatment of the Navier-Stokes equations for incompressible flows have become a staple in modern day computational engineering. But despite the enormous efforts invested, large-scale flow problems still challenge state-of-the-art

https://doi.org/10.4995/YIC2021.2021.12217

high-performance computer architectures. The development of new algorithms and methods designed for such problems therefore remains an intense field of research. When employing finite elements, basis functions for velocity and pressure have to be chosen with caution, obeying the Ladyzhenskaya–Babuška–Brezzi (LBB) condition. Some extensions of classical workarounds for Newtonian fluids are readily available, ranging from penalty methods [3, 4] to pressure Poisson stabilisation [5] and local pressure projection [2]. Some residual-based stabilisations have been proposed [6, 7], and we recently presented a novel one [8, 9], eliminating spurious pressure boundary layers and poor conservation properties even in lowest-order discretisations. However, preconditioning the arising linear (block-) systems is a critical and often limiting task when developing numerical schemes, despite well-performing algorithms being available [10–12].

In view of these challenges, one might prefer projection or split-step schemes decoupling velocity and pressure [13, 14], thereby decomposing the system into convection-diffusion, Poisson and simple mass matrix problems. Nonetheless, projection methods suffer from artificial pressure boundary conditions (refer to Guermond et al. [15] for an excellent overview), which often call for corrective measures [16, 17]. As an alternative, Liu [18] combines explicit treatment of the convective velocity with a pressure Poisson equation (PPE) equipped with consistent boundary conditions. While schemes of similar kind have been applied to challenging incompressible flow problems [19–22], the extension to the non-Newtonian case is in many aspects challenging. Deteix and Yakoubi [14] proposed the so-called shear rate projection scheme which, despite being accurate and simple, requires LBB-stable velocity-pressure pairs and the solution of an advection-diffusion equation, two Poisson problems and more than ten scalar mass matrix problems per time step.

By contrast, we focus herein on the recent extension of the PPE scheme [18] to generalised Newtonian fluids [23, 24]. This new framework allows for continuous equal-order finite elements, is higher-order accurate, iteration-free, and consists of an advection-diffusion equation, a single PPE, and two mass matrix solves to recover pressure Dirichlet data and viscosity. We focus on the full-traction variant, additionally including Galerkin least-squares (GLS) stabilisation [25] to counteract dominant convective terms and the popular three-element Windkessel model together with backflow stabilisation.

2 PROBLEM STATEMENT

As a starting point, let us consider mass and momentum balance equations for an incompressible fluid in $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ and a time interval from $t = 0$ to T:

$$\rho\left[\partial_t \mathbf{u} + (\nabla \mathbf{u})\mathbf{u}\right] - \nabla \cdot \mathbb{S} + \nabla p = \mathbf{f} \qquad \text{in } \Omega \times (0, T], \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \qquad \text{in } \Omega \times (0, T], \tag{2}$$

with a constant density ρ, velocity $\mathbf{u}$, pressure p, volumetric body force $\mathbf{f}$ and viscous stress $\mathbb{S}$. For generalised Newtonian fluids the viscous stress $\mathbb{S}$ computes by

$$\mathbb{S} = 2\mu\nabla^s\mathbf{u}, \tag{3}$$

where $\mu(\mathbf{x}, t) \in \mathbb{R}^+$ denotes the variable dynamic viscosity and $\nabla^s\mathbf{u} := \frac{1}{2}[\nabla\mathbf{u} + (\nabla\mathbf{u})^\top]$ is the symmetric part of the velocity gradient. System (1)–(2) is supplemented by

$$\mathbf{u} = \mathbf{g} \qquad \text{on } \Gamma_D \times (0, T], \tag{4}$$

$$(\mathbb{S} - p\mathbb{I})\mathbf{n} = \mathbf{h} \qquad \text{on } \Gamma_N \times (0, T], \tag{5}$$

$$\mathbf{u} = \mathbf{u}_0 \qquad \text{at } t = 0, \tag{6}$$

given Dirichlet data $\mathbf{g}$ on Γ_D and Neumann data in terms of the full normal traction $\mathbf{h}$ on Γ_N, where $\Gamma_D \cup \Gamma_N = \partial\Omega$ and $\Gamma_D \cap \Gamma_N = \emptyset$. The rheological law describing the viscosity μ

depending on the shear rate $\dot{\gamma}$ is usually formulated in terms of a map $\eta : \mathbb{R}^+ \to \mathbb{R}^+ \setminus \{0\}$:

$$\mu = \eta(\dot{\gamma}), \quad \text{with } \dot{\gamma} := \sqrt{1/2 \nabla^s \mathbf{u} : \nabla^s \mathbf{u}}\,. \tag{7}$$

A popular choice in the context of shear-thinning haemodynamics or polymeric flows is the well-established Carreau model [26]

$$\eta\left(\dot{\gamma}\right) = \eta_\infty + \left(\eta_0 - \eta_\infty\right)\left[1 + \left(\lambda\dot{\gamma}\right)^2\right]^{\frac{n-1}{2}}, \tag{8}$$

with upper and lower limits η_0 and η_∞, respectively, and further fitting parameters λ and $n \leq 1$. Homogeneous Newtonian fluids are naturally included in this setting, e.g. for $n = 1$.

3 TIME-SPLITTING SCHEME

The time-splitting scheme is based on a consistently derived PPE equipped with suitable boundary conditions. So, let us start by taking minus the divergence of Eq. (1), to obtain

$$\begin{aligned} -\nabla \cdot (\nabla p) &= -\nabla \cdot \mathbf{f} + \rho \nabla \cdot [\partial_t \mathbf{u} + (\nabla \mathbf{u})\mathbf{u}] - \nabla \cdot (\nabla \cdot \mathbb{S}) = \\ -\Delta p &= -\nabla \cdot \mathbf{f} + \rho \partial_t \left(\nabla \cdot \mathbf{u}\right) + \rho \nabla \cdot [(\nabla \mathbf{u})\mathbf{u}] - \nabla \cdot (\nabla \cdot \mathbb{S})\,. \end{aligned}$$

We can further use $\nabla \cdot \mathbf{u} = 0$ and

$$\nabla \cdot \mathbb{S} = \nabla \cdot (2\mu \nabla^s \mathbf{u}) = \mu \nabla \left(\nabla \cdot \mathbf{u}\right) + \mu \Delta \mathbf{u} + 2\nabla^s \mathbf{u} \nabla \mu = \mu \Delta \mathbf{u} + 2\nabla^s \mathbf{u} \nabla \mu$$

to obtain

$$-\Delta p = -\nabla \cdot \mathbf{f} + \rho \nabla \cdot [(\nabla \mathbf{u})\mathbf{u}] - \nabla \cdot (2\nabla^s \mathbf{u} \nabla \mu) - \Delta \mathbf{u} \cdot \nabla \mu - \mu \Delta \left(\nabla \cdot \mathbf{u}\right),$$

which simplifies to

$$-\Delta p = \nabla \cdot [\rho (\nabla \mathbf{u})\mathbf{u} - 2\nabla^s \mathbf{u} \nabla \mu - \mathbf{f}] + [\nabla \times (\nabla \times \mathbf{u})] \cdot \nabla \mu \tag{9}$$

using

$$\Delta \mathbf{u} \equiv \nabla \left(\nabla \cdot \mathbf{u}\right) - \nabla \times (\nabla \times \mathbf{u}) = -\nabla \times (\nabla \times \mathbf{u})\,.$$

The Dirichlet condition for this auxiliary problem is obtained by dotting the traction boundary condition on Γ_N (5) with the unit outward normal vector $\mathbf{n}$:

$$\mathbf{n} \cdot [(\mathbb{S} - p\mathbb{I})\mathbf{n}] = \mathbf{n} \cdot \mathbf{h} \quad \therefore \quad \mathbf{n} \cdot [\mathbb{S}\mathbf{n} - \mathbf{h}] = p\mathbf{n} \cdot \mathbf{n} = p \tag{10}$$

and similarly, dotting the momentum balance equation (1) with $\mathbf{n}$

$$\mathbf{n} \cdot \nabla p = \mathbf{n} \cdot [\mathbf{f} - \rho \partial_t \mathbf{u} - \rho (\nabla \mathbf{u})\mathbf{u} + 2\nabla^s \mathbf{u} \nabla \mu - \mu \nabla \times (\nabla \times \mathbf{u})] \tag{11}$$

gives the Neumann condition for the PPE when restricted to Γ_D. For a detailed derivation, the interested reader is referred to our recent work [23], while we herein focus directly on a weak formulation of the split-step scheme. Let us denote the $L^2(\Omega)$ and $L^2(\Gamma_D)$ scalar products by $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle_{\Gamma_D}$, respectively, and start off by multiplying the PPE (9) with a test function $q \in H^1(\Omega)$, $q|_{\Gamma_N} = 0$, integrating by parts and inserting the Neumann boundary condition (11), thereby yielding

$$\begin{aligned} \langle \nabla q, \nabla p \rangle = &- \langle q, \mathbf{n} \cdot [\rho \partial_t \mathbf{u} + \mu \nabla \times (\nabla \times \mathbf{u})] \rangle_{\Gamma_D} \\ &+ \langle \nabla q, 2\nabla^s \mathbf{u} \nabla \mu + \mathbf{f} - \rho (\nabla \mathbf{u})\mathbf{u} \rangle + \langle q, [\nabla \times (\nabla \times \mathbf{u})] \cdot \nabla \mu \rangle, \end{aligned}$$

Table 1: Coefficients α_j^m and β_j^m of order $m = 2$ for BDF and extrapolation [27].

j	0	1	2
α_j^m	$\frac{2\Delta t^n+\Delta t^{n-1}}{\Delta t^n(\Delta t^n+\Delta t^{n-1})}$	$-\frac{\Delta t^n+\Delta t^{n-1}}{\Delta t^n\Delta t^{n-1}}$	$\frac{\Delta t^n}{\Delta t^{n-1}(\Delta t^n+\Delta t^{n-1})}$
β_j^m	$-$	$1+\frac{\Delta t^n}{\Delta t^{n-1}}$	$-\frac{\Delta t^n}{\Delta t^{n-1}}$

which is rewritten using integration by parts once again as

$$\langle \nabla q, \nabla p\rangle = \langle \nabla q, \mathbf{f} + 2\nabla^s \mathbf{u}\nabla\mu - \rho(\nabla\mathbf{u})\mathbf{u} - \mu\nabla\times(\nabla\times\mathbf{u})\rangle - \langle q\mathbf{n}, \rho\partial_t\mathbf{u}\rangle_{\Gamma_D},$$

Note however, that second-order derivatives are still present, which we mend via

$$\begin{aligned}\langle \mu\nabla q, \nabla\times(\nabla\times\mathbf{u})\rangle =& \langle \nabla q\times\mathbf{n}, \mu\nabla\times\mathbf{u}\rangle_\Gamma + \langle \nabla\times(\mu\nabla q), \nabla\times\mathbf{u}\rangle,\\ =& \langle \nabla q\times\mathbf{n}, \mu\nabla\times\mathbf{u}\rangle_{\Gamma_D} + \langle \nabla q, [\nabla\mathbf{u} - (\nabla\mathbf{u})^\top]\nabla\mu\rangle,\end{aligned}$$

omitting some of the details from [23] for brevity. Other vital ingredients of the split-step scheme are (i) the full decoupling of momentum balance and PPE through explicit treatment of the pressure gradient term in (1), (ii) projection of the PPE Dirichlet condition on Γ_N (10), (iii) recovering the viscosity μ via an L^2 projection and (iv) improving conservation of mass using divergence damping [20, 22]. For the time integration, we consider variable time steps $\Delta t^n = t^{n+1} - t^n$ in higher-order accurate backward differentiation (BDF) and extrapolation formulae (indicated by *) with coefficients α_j^m and β_j^m given in Tab. 1:

$$\partial_t\mathbf{u}(t^{n+1}) \approx \alpha_0^m\mathbf{u}^{n+1} + \sum_{j=1}^m \alpha_j^m\hat{\mathbf{u}}^{n+1-j}, \qquad \mathbf{u}^{n+1} \approx \mathbf{u}^\star := \sum_{j=1}^m \beta_j^m\mathbf{u}^{n+1-j}. \tag{12}$$

Then, given solutions from previous time steps, the split-step scheme reads

1. *Momentum balance*:
   Find $\mathbf{u}^{n+1} \in X_h \subset H^1(\Omega)$, such that $\mathbf{u}^{n+1}|_{\Gamma_D} = \mathbf{g}^{n+1}$ and

$$\begin{aligned}\left\langle \rho\mathbf{v}, \alpha_0^m\mathbf{u} + \sum_{j=1}^m\left[\alpha_j^m\left(\mathbf{u}^{n+1-j} - \nabla\varphi^{n+1-j}\right)\right] + (\nabla\mathbf{u})\mathbf{u}^\star\right\rangle + \langle\nabla\mathbf{v}, 2\mu^\star\nabla^s\mathbf{u}^{n+1} - p^\star\mathbb{I}\rangle\\ = \langle\mathbf{v}, \mathbf{f}^{n+1}\rangle + \langle\mathbf{v}, \mathbf{h}^{n+1}\rangle_{\Gamma_N} \qquad \forall\mathbf{v}\in X_h, \mathbf{v}|_{\Gamma_D} = \mathbf{0}.\end{aligned} \tag{13}$$

2. *Project viscosity*:
   Find $\mu^{n+1} \in Y_h \subset H^1\Omega$, such that

$$\langle v, \mu^{n+1}\rangle = \langle v, \eta(\mathbf{u}^{n+1})\rangle \qquad \forall v \in Y_h. \tag{14}$$

3. *PPE Dirichlet condition*:
   Recover the continuous $\zeta^{n+1} := \mathbf{n}\cdot\left[(2\mu^{n+1}\nabla^s\mathbf{u}^{n+1})\,\mathbf{n} - \mathbf{h}^{n+1}\right]$ on Γ_N via L^2 projection.

4. *Pressure Poisson step*:
   Find $p^{n+1} \in Z_h \subset H^1(\Omega)$, such that $p^{n+1}|_{\Gamma_N} = \zeta^{n+1}$ and

$$\begin{aligned}\langle\nabla q, \nabla p^{n+1}\rangle = \langle\nabla q, \mathbf{f}^{n+1} + 2\left(\nabla\mathbf{u}^{n+1}\right)^\top\nabla\mu^{n+1} - \rho(\nabla\mathbf{u}^{n+1})\mathbf{u}^{n+1}\rangle\\ -\langle q\mathbf{n}, \rho\sum_{j=0}^m\alpha_j^m\mathbf{u}^{n+1-j}\rangle_{\Gamma_D} + \langle\mathbf{n}\times\nabla q, \mu\nabla\times\mathbf{u}^{n+1}\rangle_{\Gamma_D} \qquad \forall q\in Z_h, q|_{\Gamma_N} = 0.\end{aligned} \tag{15}$$

5. *Divergence damping*:
   Find $\varphi^{n+1} \in Z_h$, such that $\varphi^{n+1}|_{\Gamma_N} = 0$ and

$$\langle \nabla\psi, \nabla\varphi^{n+1} \rangle = \langle \psi, \nabla \cdot \mathbf{u}^{n+1} \rangle \qquad \forall \psi \in Z_h, \psi|_{\Gamma_N} = 0, \tag{16}$$

   which will be used in the following time step.

Note here, that the Poisson problems (15)–(16) can be combined, solving only one Poisson problem per time step (*cf.* [18, 23]) and for Newtonian fluids, the viscosity projection step (14) is skipped. Also, the rheological law is easily replaced by swapping the right-hand side of (14) and was actually lumped in our experiments. Per time step, the scheme consists of solving a vector-valued advection-diffusion equation, an L^2 projection on Γ_N, a lumped mass matrix and one Poisson problem in the auxiliary variable $\hat{p} := p + \varphi$.

4 NUMERICAL EXAMPLES

The split-step algorithm (13)–(16) is implemented in the open-source finite element library `deal.II` [28], using parallel algebraic multigrid (AMG) methods provided by *Trilinos' ML* package [29] for preconditioning the FGMRES and BiCGStab methods used to solve linear systems corresponding to fluid momentum and PPE, respectively. The versatility and computational performance of the scheme is showcased in two fundamentally different applications in biomechanics, the first being flow through an abdominal aortic aneurysm and the second example considering human phonation.

4.1 Abdominal aortic aneurysm

Aneurysms are pathological vessel malformations giving rise to deformed, bulging lumina, altering flow fields and triggering various critical health conditions. A physiological setup is created similar to [30] based on flow data and geometry provided in [31, 32]. This prototypical segment of the abdominal aorta with length $L = 20$ cm and inlet/outlet radius $R = 1$ cm is subject to periodic inflow and outlet pressure $\bar{p}$ depicted in Fig. 1. Starting from the quiescent state, i.e., $\mathbf{u}_0 = \mathbf{0}$, we prescribe $\mathbf{u} = (u_1, 0, 0)^\top$ smoothly ramped by

$$\xi(t) = \begin{cases} \sin^2\left(\frac{\pi t}{2\tau}\right) & \text{for } t \leq \tau, \\ 1 & \text{otherwise,} \end{cases} \tag{17}$$

with $\tau = 0.2$ s and a quadratic velocity profile, matching the volumetric flow rate computed by the given mean velocity $\bar{\mathbf{u}}$. Concerning the fluid parameters, we set $\rho = 1060$ kg/m^3 and $\eta_0 = 56$ mPas, $\eta_\infty = 3.45$ mPas, $\lambda = 3.313$ s and $n = 0.3568$ in (8) according to [33]. Further modeling aspects such as three-element Windkessel models, backflow stabilisation and GLS stabilisation are included into the split-step scheme. These extensions, typical for haemodynamic applications, merely modify Neumann data $\mathbf{h}^{n+1}$ or add terms to the momentum equation, and a rigorous introduction is omitted for brevity. Moreover, we define the maximum element CFL and Reynolds numbers as

$$\mathrm{CFL}_e = \max_{e=1,\dots,N_e} \max_{i=1,\dots,d} \frac{|u_i^{n+1}|\Delta t^n}{h_i}, \quad \mathrm{Re}_e = \max_{e=1,\dots,N_e} \max_{i=1,\dots,d} \frac{\rho |u_i^{n+1}| h_i}{\mu}, \tag{18}$$

with the number of elements N_e and directional element size h_i taken as the maximum vertex distance in direction i. Based on (18), we aim for $\mathrm{CFL}_e \leq 0.5$, starting from an initial value of $\Delta t^0 = 10^{-3}$ s until five pulses are completed, i.e., $t \in (0, 5]$. The solution is periodic in time, spatially symmetric and characterised by strong recirculations during diastole, as exemplarily

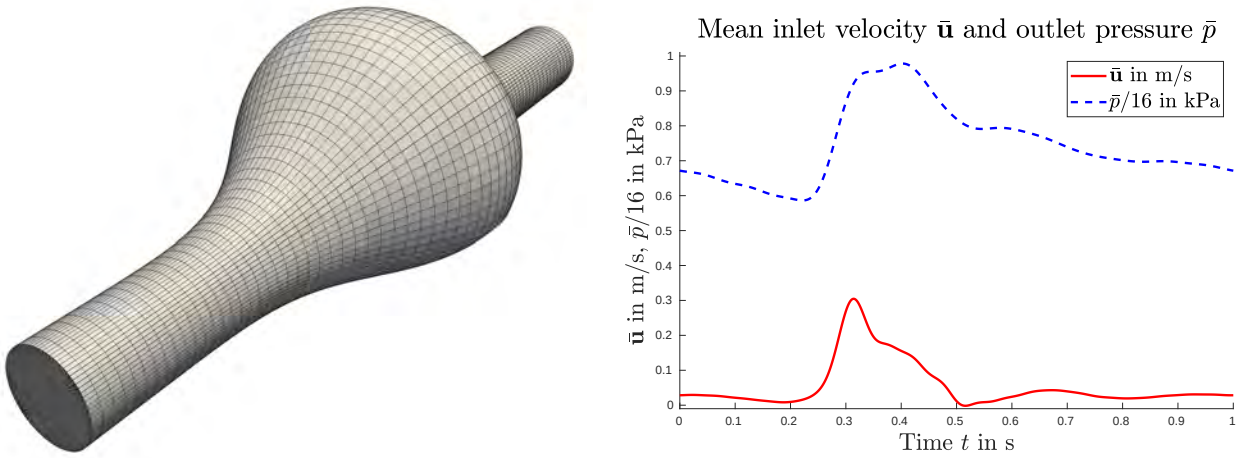


Figure 1: Abdominal aortic aneurysm: computational mesh (left) and boundary data (right).

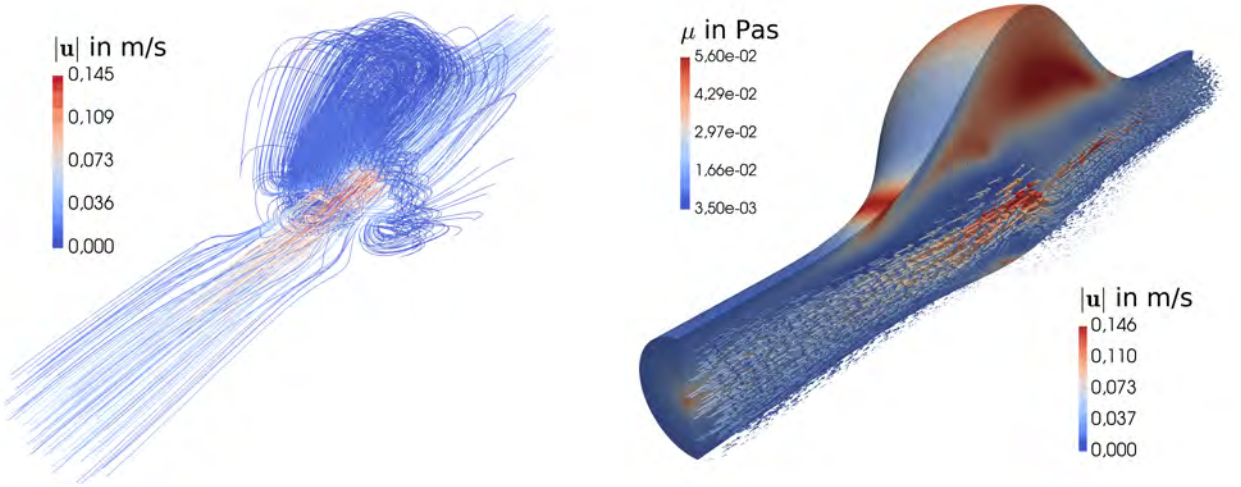


Figure 2: Strong recirculation and viscosity gradients in aneurysm at $t \approx 4.97$ s (diastole), selected streamlines (left) and viscosity in cut domain together with selected velocity vectors (right).

shown in Fig. 2 at $t \approx 4.97$ s. Consequently, viscosity spans the whole admissible spectrum $\eta_\infty \leq \mu \leq \eta_0$ due to large variations in the local shear rate. All linear systems are solved reducing the residual by a factor of 10^{-6}, taking the last timestep solution as the initial guess. Doing so, iteration counts for momentum balance ($N_{\mathbf{u}}$) and PPE (N_p) stay below 20, while the projection of pressure Dirichlet data on Γ_N needs a constant of 6 steps only for reaching convergence. Note here, how the former two mildly depend on the flow field as shown in Fig. 3, where we include the inlet velocity $\bar{\mathbf{u}}$ for reference. Fig. 3 also depicts the adapting time step size together with element CFL and Reynolds numbers, showing time steps decreasing from ≈ 0.015 to ≈ 0.002 shortly after peak inflow, yielding a maximum CFL_e of ≈ 0.85 without repeating time steps. $\mathrm{CFL}_e > 1$ is admissible in the split-step scheme (*cf.* Pacheco et al. [23]), but only at the cost of increasing iteration counts in the momentum balance solve.

4.2 Human phonation

In a second numerical test, we aim to simulate human phonation, which is the process of vocal folds interacting with air from the lungs, creating the human voice. However, in this preliminary two-dimensional study, the setup inspired by Kniesburges et al. [34] is limited to fixed vocal folds. Parameters representing air are selected as $\rho \approx 1.18$ kg/m^3 and $\mu = 0.0137$ mPas for a

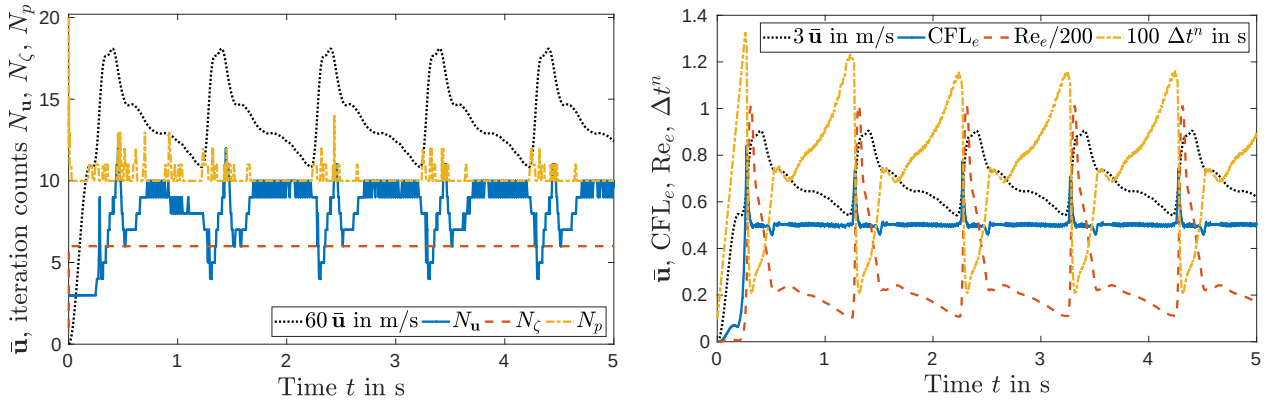


Figure 3: Iteration counts of momentum balance $N_\mathbf{u}$, pressure boundary projection N_ζ and PPE N_p (left); maximum element CFL and Reynolds numbers and Δt^n (right) with inlet velocity $\bar{\mathbf{u}}$ for reference.

Newtonian fluid. The glottis is modeled as a channel of total length of ≈ 50.4 mm and height of $H = 18$ mm, including the vocal folds (VFs) with a height of 8.9 mm as well as in a distance of 7.5 mm the false vocal folds with a height of 6.5 mm. The gap distance between the two VFs is $H_G = 0.2$ mm as depicted in Fig. 4.

Figure 4: Computational domain for the phonation example with vocal folds in dark grey.

Starting again from a quiescent state ($\mathbf{u}_0 = \mathbf{0}$) and ramping via (17) with $\tau = 0.01$ s, we enforce a quadratic inflow profile. The maximum inlet velocity is prescribed as 80 cm/s, yielding an intraglottal maximum velocity of $\bar{u}_G \approx 56$ m/s and $\mathrm{Re} = {}^{\rho \bar{u}_G H_G}/_{\mu} = \mathcal{O}(10^3)$ being in the physiological range [34]. On the outlet, a zero reference pressure is (approximately) set using $\mathbf{h} = \mathbf{0}$. Regarding the solver settings, we choose an initial $\Delta t^0 = 10^{-4}$ s, adapt the time step size such that CFL$_e \leq 0.8$ and reduce the residual by a factor of 10^{-8} with the last time step's solution as initial guess. The resulting velocity field is characterised by a strong jet, triggering vortices which in return influence the jet direction. Moreover, the pressure field features fluctuations in the vicinity of the jet as shown in Fig. 5. Low iteration counts result over

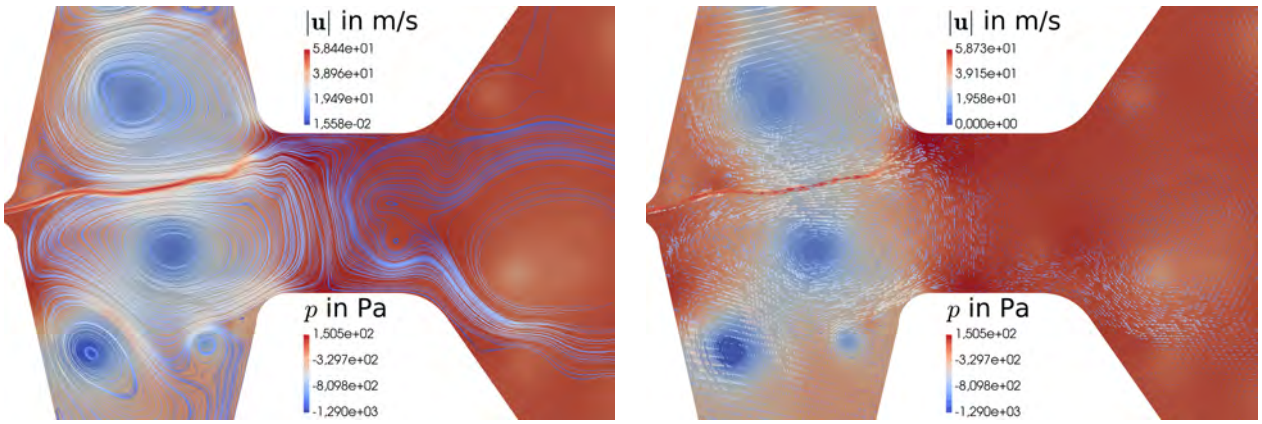


Figure 5: Snapshot of the solution at step 12000 ($t = 13.576$ ms) in the fold region: selected velocity streamlines (left) and vectors (right) colored by $|\mathbf{u}|$ over pressure p in the background.

the whole considered timespan and interestingly, almost constant iteration counts are observed

after the initial ramp-up phase. This is due to the time step size settling at $\Delta t \approx 10^{-6}$ s, giving an almost constant $\mathrm{Re}_e \approx 720$ and $\mathrm{CFL}_e \approx 0.8$. Therefore, we simply report mean iteration counts over the last 1000 time steps as $\bar{N}_\zeta = 3$, $\bar{N}_{\mathbf{u}} \approx 22.53$ and $N_p \approx 54.18$. Comparing to the previous aneurysm example, a slight increase is seen, which is due to a combination of worsened element aspect ratios and higher Reynolds number, but also depends on the more strict convergence criterion.

5 CONCLUSION

Within this work, a time-splitting scheme suitable for incompressible (generalised) Newtonian fluids has been presented. Momentum and mass balance equations are decoupled using an implicit-explicit treatment of the pressure, viscosity and convection terms. Thus, only an advection-diffusion equation for momentum balance and a PPE with fully consistent boundary conditions are computationally relevant steps. Lower equal-order interpolation of velocity and pressure is also found admissible, while temporal accuracy is determined by suitable BDF and extrapolation formulae. Two challenging examples in biomedical context were tackled, namely, flow through an abdominal aortic aneurysm and human phonation, demonstrating the effectiveness and versatility of the presented approach.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Graz University of Technology for the financial support of the Lead-project: Mechanics, Modeling and Simulation of Aortic Dissection.

REFERENCES

[1] G.P. Galdi, R. Rannacher, A.M. Robertson, and S. Turek. *Hemodynamical Flows*, volume 37 of *Oberwolfach Seminars*. Birkhäuser, Basel, 2008.

[2] L. John, P. Pustějovská, and O. Steinbach. On the influence of the wall shear stress vector form on hemodynamic indicators. *Comput. Vis. Sci.*, 18(4-5):113–122, 2017.

[3] G.F. Carey, K.C. Wang, and W.D. Joubert. Performance of iterative methods for Newtonian and generalized Newtonian flows. *Int. J. Numer. Meth. Fluids*, 9(2):127–150, 1989.

[4] M. Franta, J. Málek, and K.R. Rajagopal. On steady flows of fluids with pressure- and shear-dependent viscosities. *Proc. Math. Phys. Eng. Sci.*, 461(2055):651–670, 2005.

[5] S. Knauf, S. Frei, T. Richter, and R. Rannacher. Towards a complete numerical description of lubricant film dynamics in ball bearings. *Comput. Mech.*, 53(2):239–255, 2014.

[6] A. Masud and J. Kwack. A stabilized mixed finite element method for the incompressible shear-rate dependent non-Newtonian fluids: Variational Multiscale framework and consistent linearization. *Comput. Methods Appl. Mech. Eng.*, 200(5-8):577–596, 2011.

[7] V.L. Marrero, J.A. Tichy, O. Sahni, and K.E. Jansen. Numerical study of purely viscous non-Newtonian flow in an abdominal aortic aneurysm. *J. Biomech. Eng.*, 136(10), 2014.

[8] D.R.Q. Pacheco, R. Schussnig, O. Steinbach, and T.-P. Fries. A global residual-based stabilization for equal-order finite element approximations of incompressible flows. *Int. J. Numer. Methods Eng.*, 122(8):2075–2094, 2021.

[9] R. Schussnig, D.R.Q. Pacheco, and T.-P. Fries. Robust stabilised finite element solvers for generalised Newtonian fluid flows. *J. Comput. Phys. (in press)*, 2021.

[10] H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite Elements and Fast Iterative Solvers*. Oxford University Press, Oxford, 2014.

[11] S. Turek. *Efficient Solvers for Incompressible Flow Problems - An Algorithmic and Computational Approach*, volume 6 of *Lecture Notes in Computational Science and Engineering*. Springer, Berlin Heidelberg, 1999.

[12] T. Heister and G. Rapin. Efficient augmented Lagrangian-type preconditioning for the Oseen problem using Grad-Div stabilization. *Int. J. Numer. Meth. Fluids*, 71(1):118–134, 2013.

[13] A. Smith and D. Silvester. Implicit algorithms and their linearization for the transient incompressible Navier-Stokes equations. *IMA J. Numer. Anal.*, 17(4):527–545, 1997.

[14] J. Deteix and D. Yakoubi. Shear rate projection schemes for non-Newtonian fluids. *Comput. Methods Appl. Mech. Eng.*, 354:620–636, 2019.

[15] J.L. Guermond, P. Minev, and Jie Shen. An overview of projection methods for incompressible flows. *Comput. Methods Appl. Mech. Eng.*, 195(44-47):6011–6045, 2006.

[16] L.J.P. Timmermans, P.D. Minev, and F.N. Van de Vosse. An approximate projection scheme for incompressible flow using spectral elements. *Int. J. Numer. Meth. Fluids*, 22 (7):673–688, 1996.

[17] A. Poux, S. Glockner, and M. Azaïez. Improvements on open and traction boundary conditions for Navier–Stokes time-splitting methods. *J. Comput. Phys.*, 230(10):4011–4027, 2011.

[18] J. Liu. Open and traction boundary conditions for the incompressible Navier–Stokes equations. *J. Comput. Phys.*, 228(19):7250–7267, 2009.

[19] J.-G. Liu, J. Liu, and R.L. Pego. Stable and accurate pressure approximation for unsteady incompressible viscous flow. *J. Comput. Phys.*, 229(9):3428–3453, 2010.

[20] J. Jia and J. Liu. Stable and spectrally accurate schemes for the Navier–Stokes equations. *SIAM J. Sci. Comput.*, 33(5):2421–2439, 2011.

[21] Z. Sheng, M. Thiriet, and F. Hecht. A high-order scheme for the incompressible Navier-Stokes equations with open boundary condition. *Int. J. Numer. Meth. Fluids*, 73(1):58–73, 2013.

[22] L. Li. A split-step finite-element method for incompressible Navier-Stokes equations with high-order accuracy up-to the boundary. *J. Comput. Phys.*, 408:109274, 2020.

[23] D.R.Q. Pacheco, R. Schussnig, and T.-P. Fries. An efficient split-step framework for non-Newtonian incompressible flow problems with consistent pressure boundary conditions. *Comput. Methods Appl. Mech. Eng.*, 382:113888, 2021.

[24] D.R.Q. Pacheco and O. Steinbach. A continuous finite element framework for the pressure Poisson equation allowing non-Newtonian and compressible flow behavior. *Int. J. Numer. Meth. Fluids*, 93:1435–1445, 2021.

[25] T.J.R. Hughes and L.P. Franca. A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces. *Comput. Methods Appl. Mech. Eng.*, 65(1):85–96, 1987.

[26] Y.I. Cho and K.R. Kensey. Effects of the non-Newtonian viscosity of blood on flows in a diseased arterial vessel. Part 1: Steady flows. *Biorheology*, 28(3-4):241–62, 1991.

[27] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations 1 - Nonstiff problems*. Springer, 1993.

[28] D. Arndt, W. Bangerth, B. Blais, T.C. Clevenger, M. Fehling, A.V. Grayver, T. Heister, L. Heltai, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, R. Rastak, I. Thomas, B. Turcksin, Z. Wang, and D. Wells. The `deal.II` library, version 9.2. *J. Numer. Math.*, 28(3):131–146, 2020.

[29] M.A. Heroux and J.M. Willenbring. A new overview of the Trilinos project. *Sci. Program.*, 20(2):83–88, 2012.

[30] S. Lin, X. Han, Y. Bi, S. Ju, and L. Gu. Fluid-Structure Interaction in Abdominal Aortic Aneurysm: Effect of Modeling Techniques. *BioMed Res. Int.*, 2017:1–10, 2017.

[31] C.J. Mills, I.T. Gabe, J.H. Gault, D.T. Mason, J. Ross, E. Braunwald, and J.P. Shillingford. Pressure-flow relationships and vascular impedance in man. *Cardiovasc. Res.*, 4(4):405–417, 1970.

[32] C.A. Meyer, E. Bertrand, O. Boiron, and V. Deplano. Stereoscopically Observed Deformations of a Compliant Abdominal Aortic Aneurysm Model. *J. Biomech. Eng.*, 133(11), 2011.

[33] S. Kim, Y.I. Cho, A.H. Jeon, B. Hogenauer, and K.R. Kensey. A new method for blood viscosity measurement. *J Non-Newt. Fluid Mech.*, 94(1):47–56, 2000.

[34] S. Kniesburges, S.L. Thomson, A. Barney, M. Triep, P. Sidlof, J. Horacek, C. Brucker, and S. Becker. In Vitro Experimental Investigation of Voice Production. *Curr. Bioinform.*, 6 (3):305–322, 2011.

FLOW AND MECHANICS IN POROUS MEDIA

Numerical investigation on a block preconditioning strategy to improve the computational efficiency of DFN models

Laura Gazzola*, Massimiliano Ferronato*, Stefano Berrone†, Sandra Pieraccini† and Stefano Scialò†

* University of Padova
Padova, Italy
e-mail: laura.gazzola.1@phd.unipd.it
e-mail: massimiliano.ferronato@unipd.it

† Politecnico di Torino
Torino, Italy
e-mail: stefano.berrone@polito.it
e-mail: sandra.pieraccini@polito.it
e-mail: stefano.scialo@polito.it

Key words: Discrete Fracture Network, Preconditioning

Abstract: *The simulation of underground flow across intricate fracture networks can be addressed by means of discrete fracture network models. The combination of such models with an optimization formulation allows for the use of nonconforming and independent meshes for each fracture. The arising algebraic problem produces a symmetric saddle-point matrix with a rank-deficient leading block. In our work, we investigate the properties of the system to design a block preconditioning strategy to accelerate the iterative solution of the linearized algebraic problem. The matrix is first permuted and then projected in the symmetric positive-definite Schur-complement space. The proposed strategy is tested in applications of increasing size, in order to investigate its capabilities.*

1 INTRODUCTION

The simulation of the flow in highly fractured systems can be particularly demanding from a computational standpoint, because of the size and complexity of the domain and the uncertainty characterizing the rock properties and the fracture geometry.

In this context, discrete fracture network (DFN) models can be used, and are preferred particularly when the presence of fractures has a dominant impact on the fluid flow dynamics. DFN models represent only the fractures as intersecting planar polygons, neglecting the surrounding underground rock formation. Differently from homogenization-based techniques, DFN models provide an explicit representation of the fractures and their properties in a 3D structure, prescribing continuity constraints for the fluid flow along the linear intersections. The number of the fractures and their different size, that can change of orders of magnitude, entail a complex and multi-scale geometry, which is not trivial to address. The problem has been effectively reformulated as a PDE-constrained optimization problem in [1, 2]. The formulation relies on the use of non-conforming discretizations of the single fractures and on the minimization of a functional to couple intersecting planes. Thus, no match between the meshes of the fractures and the traces are required, simplifying the mesh generation process. Moreover, the problem on the entire DFN can be decoupled in several local problems on the fractures with a moderate exchange of data among fractures, being suitable for a massive parallel implementation [2].

The linearized algebraic problem that derives from such a formulation produces a large size symmetric saddle-point matrix with a rank-deficient leading block. In this work, we focus on accelerating the iterative solution of the linear system by introducing effective block preconditioning techniques. In particular, an appropriate permutation of the global matrix is first

https://doi.org/10.4995/YIC2021.2021.12234

performed, in order to avoid a singular leading block. Though the permuted matrix is no longer symmetric, this approach should be better suited for the solution with Krylov subspace methods. Then, the matrix is projected in the symmetric positive-definite Schur complement space of the fluxes along the intersection traces. The properties and the structure of matrix blocks are properly exploited in order to guarantee an efficient parallel implementation. The matrix properties are tested in applications of increasing size to verify pros and cons of the approach.

The manuscript is organized as follows. In section 2 the mathematical problem and the related discrete algebraic form are introduced. In section 3 the preconditioner framework is described. In section 4 numerical results for four problems of increasing size and complexity are analyzed and discussed.

2 PROBLEM STATEMENT

We consider a connected three-dimensional fracture network made by a system of intersected polygonal fractures surrounded by an impervious matrix. The flow occurs only along the fractures and their intersections, called traces. The flow along the fractures is modeled by means of Darcy's law with appropriate boundary conditions. Coupling conditions are imposed on the traces, in order to guarantee the continuity of the solution and the balance of the fluxes. The whole problem can be reformulated as PDE-constrained optimization problem [1]. Introducing an independent mesh on each fracture and trace, the Darcy equation, as well as the optimization problem, can be discretized following the standard finite element method. The result is the following algebraic problem [2]:

$$G^h\mathbf{h} - \alpha B\mathbf{u} + A^T\mathbf{p} = \mathbf{0}, \qquad \text{(energy minimization)} \tag{1a}$$

$$-\alpha B^T\mathbf{h} + G^u\mathbf{u} - C^T\mathbf{p} = \mathbf{0}, \qquad \text{(energy minimization)} \tag{1b}$$

$$A\mathbf{h} - C\mathbf{u} = \mathbf{q}, \qquad \text{(mass balance)} \tag{1c}$$

where $\mathbf{h} \in \mathbb{R}^{n^h}$ is the hydraulic head on the fractures, $\mathbf{u} \in \mathbb{R}^{n^u}$ is the flux on the traces, $\mathbf{p} \in \mathbb{R}^{n^p}$ are Lagrange multipliers and $\mathbf{q} \in \mathbb{R}^{n^p}$ derives from the boundary conditions and the forcing terms. Usually, $n^p = n^h$, while according to the problem n^u can be either larger or smaller than n^h. The coefficient $\alpha \in \mathbb{R}$ is a user-specified positive parameter, usually on the order of 1. The matrices $G^h \in \mathbb{R}^{n^h \times n^h}$, $A \in \mathbb{R}^{n^h \times n^h}$ and $C \in \mathbb{R}^{n^h \times n^u}$ are fracture-local, whereas $B \in \mathbb{R}^{n^h \times n^u}$ and $G^u \in \mathbb{R}^{n^u \times n^u}$ operate on degrees of freedom related to different fractures. Their properties can be summarized as follows:

- G^h and G^u are symmetric positive semi-definite (SPSD), usually rank-deficient;
- B and C are rectangular coupling blocks, whose entries are given by inner products between the basis functions of the main unknowns along the fracture traces;
- A is symmetric positive definite (SPD) with a block diagonal structure. Each diagonal block arises from the discretization of a $\nabla \cdot (\kappa\nabla)$ operator over a fracture, where κ is a proper diffusion tensor, hence inherits the usual structure of a 2-D discrete Laplacian. Block size depends on each fracture dimension and can be significantly different one from the other.

Equations (1) can be written in a compact form as:

$$\begin{bmatrix} G^h & -\alpha B & A^T \\ -\alpha B^T & G^u & -C^T \\ A & -C & 0 \end{bmatrix} \begin{pmatrix} \mathbf{h} \\ \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{q} \end{pmatrix} \quad \Rightarrow \quad \mathbf{K}\mathbf{x} = \mathbf{f} \tag{2}$$

where $\mathbf{K}$ is a symmetric saddle-point matrix with a rank-deficient leading block. Solution to such problems arise in several applications and is the object of a significant number of works. For a review on methods and ideas, see for instance [3]. With an SPD leading block, as it often arises in Navier-Stokes equations, mixed finite element formulations of flow in porous media, poroelasticity, etc., an optimal preconditioner exists based on the approximation of the Schur complement matrix [4]. However, if the leading block is singular the problem is generally more difficult and the only available result is for the case of maximal rank deficiency [5].

3 PRECONDITIONER FRAMEWORK

Matrix $\mathbf{K}$ in equation (2) is a classical example of the discretization of a coupled multi-physics problem. A general preconditioning framework for such problems can be developed following the results in [6], where the different unknown fields are approximately decoupled to obtain a block diagonal problem.

Theorem 1 of [6] holds true if the leading blocks of $\mathbf{K}$ are non singular. In order to satisfy this hypothesis, a proper row and column block permutation, $\mathbf{P}_r$ and $\mathbf{P}_c$, can be applied:

$$\widetilde{\mathbf{K}} = \mathbf{P}_r \mathbf{K} \mathbf{P}_c, \qquad \widetilde{\mathbf{x}} = \mathbf{P}_c^T \mathbf{x}, \qquad \widetilde{\mathbf{f}} = \mathbf{P}_r \mathbf{f}, \tag{3}$$

such that a decoupling operator can be computed for the equivalent system $\widetilde{\mathbf{K}}\widetilde{\mathbf{x}} = \widetilde{\mathbf{f}}$. A possible choice is:

$$\widetilde{\mathbf{K}} = \begin{bmatrix} A & 0 & -C \\ G^h & A^T & -\alpha B \\ -\alpha B^T & -C^T & G^u \end{bmatrix}, \qquad \widetilde{\mathbf{x}} = \begin{pmatrix} \mathbf{h} \\ \mathbf{p} \\ \mathbf{u} \end{pmatrix}, \qquad \widetilde{\mathbf{f}} = \begin{pmatrix} \mathbf{q} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{4}$$

Let us define the decoupling operator factors $\mathbf{G}, \mathbf{F} \in \mathbb{R}^{N \times N}$ of $\widetilde{\mathbf{K}}$, being $N = 2n^h + n^u$, as:

$$\mathbf{G} = \begin{bmatrix} I & 0 & 0 \\ G_{21} & I & 0 \\ G_{31} & G_{32} & I \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} I & F_{12} & F_{13} \\ 0 & I & F_{23} \\ 0 & 0 & I \end{bmatrix}, \tag{5}$$

with $G_{21}, F_{12} \in \mathbb{R}^{n^h \times n^h}$ and $G_{31}, G_{32}, F_{13}^T, F_{23}^T \in \mathbb{R}^{n^u \times n^h}$, and such that $\mathbf{G}\widetilde{\mathbf{K}}\mathbf{F} = \mathbf{S}$, with $\mathbf{S}$ a block diagonal matrix. Then, the off-diagonal blocks of $\mathbf{F}$ satisfy the relationships:

$$\begin{cases} AF_{12} = 0 \\ \begin{bmatrix} A & 0 \\ G^h & A^T \end{bmatrix} \begin{bmatrix} F_{13} \\ F_{23} \end{bmatrix} = \begin{bmatrix} C \\ \alpha B \end{bmatrix} \end{cases}. \tag{6}$$

from which we obtain:

$$F_{12} = 0, \qquad F_{13} = A^{-1}C, \qquad F_{23} = A^{-T}\left(\alpha B - G^h A^{-1} C\right). \tag{7}$$

Similarly, the off-diagonal blocks of $\mathbf{G}$ read:

$$\begin{cases} G_{21}A = -G^h \\ \begin{bmatrix} G_{31} & G_{32} \end{bmatrix} \begin{bmatrix} A & 0 \\ G^h & A^T \end{bmatrix} = \begin{bmatrix} \alpha B^T & C^T \end{bmatrix} \end{cases}, \tag{8}$$

which provides:

$$G_{21} = -G^h A^{-1}, \qquad G_{32} = C^T A^{-T}, \qquad G_{31} = \left(\alpha B^T - C^T A^{-T} G^h\right) A^{-1}. \tag{9}$$

It is easy to observe that $G_{32} = F_{13}^T$ and $G_{31} = F_{23}^T$, hence only three off-diagonal blocks, namely F_{13}, F_{23}, and G_{21}, are needed. Recalling that $\mathbf{GKF} = \mathbf{S}$, i.e.:

$$\begin{bmatrix} I & 0 & 0 \\ G_{21} & I & 0 \\ F_{23}^T & F_{13}^T & I \end{bmatrix} \begin{bmatrix} A & 0 & -C \\ G^h & A^T & -\alpha B \\ -\alpha B^T & -C^T & G^u \end{bmatrix} \begin{bmatrix} I & 0 & F_{13} \\ 0 & I & F_{23} \\ 0 & 0 & I \end{bmatrix} = \begin{bmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ 0 & 0 & S_3 \end{bmatrix} \tag{10}$$

we have:

$$S_1 = A, \qquad S_2 = A^T, \tag{11}$$

and

$$\begin{aligned} S_3 &= \left(F_{23}^T A + F_{13}^T G^h - \alpha B^T\right) F_{13} + \left(F_{13}^T A^T - C^T\right) F_{23} + G^u - F_{23}^T C - \alpha F_{13}^T B \\ &= G^u - F_{23}^T C - \alpha F_{13}^T B. \end{aligned} \tag{12}$$

Remark 1 *Using the definitions of F_{13} and F_{23}, it is easy to observe that the matrix S_3 of equation (12) is actually the Schur complement of $\widetilde{\mathbf{K}}$ computed with respect to the third block row:*

$$S_3 = G^u - \begin{bmatrix} \alpha B^T & C^T \end{bmatrix} \begin{bmatrix} A & 0 \\ G^h & A^T \end{bmatrix}^{-1} \begin{bmatrix} C \\ \alpha B \end{bmatrix}. \tag{13}$$

Similarly, S_1 and S_2 can be also regarded as the Schur complements computed with respect to the first and second block row of $\widetilde{\mathbf{K}}$, respectively.

Introducing the matrix $E = B - C$, the definition of the Schur complement (12) can be rewritten also as a function of F_{13} only:

$$S_3 = G^u + F_{13}^T \left(G^h - 2\alpha A\right) F_{13} - \alpha \left(E^T F_{13} + F_{13}^T E\right) \tag{14}$$

From equation (10) it follows immediately:

$$\widetilde{\mathbf{K}}^{-1} = \mathbf{F}\mathbf{S}^{-1}\mathbf{G}, \tag{15}$$

that is, the expression of the exact inverse of the block matrix $\widetilde{\mathbf{K}}$. Of course, equation (15) cannot be computed explicitly in large-size applications, because both the decoupling off-diagonal blocks in $\mathbf{F}, \mathbf{G}$ and the diagonal blocks in $\mathbf{S}^{-1}$ are dense. However, we can use the factorization (15) to build an inexact application of $\widetilde{\mathbf{K}}^{-1}$ that can be used as a *preconditioner* in a Krylov subspace method.

Since our aim is to compute the product of $\widetilde{\mathbf{K}}^{-1}$ by a vector $\mathbf{r} \in \mathbb{R}^N$, we do not necessarily need to form an explicit expression of $\mathbf{F}$ and $\mathbf{G}$, but just to define an algorithm to compute their products by portions of size n^h and n^u of a vector lying in $\mathbb{R}^N$. This can be done exactly and efficiently in a parallel computational environment by recalling the properties of matrix A (see section 2). Similarly, also S_1^{-1} and S_2^{-1} (equation (11)) can be exactly applied to a vector. Hence, the block preconditioner $\mathbf{M}^{-1}$ for $\widetilde{\mathbf{K}}$ can be defined as:

$$\mathbf{M}^{-1} = \mathbf{F}\hat{\mathbf{S}}^{-1}\mathbf{G}, \tag{16}$$

where $\hat{\mathbf{S}}^{-1}$ reads:

$$\hat{\mathbf{S}}^{-1} = \begin{bmatrix} A^{-1} & 0 & 0 \\ 0 & A^{-T} & 0 \\ 0 & 0 & \hat{S}_3^{-1} \end{bmatrix}, \tag{17}$$

$\hat{S}_3^{-1}$ being some approximation, either implicit or explicit, of S_3.

For the eigenspectrum of the preconditioned matrix $\mathbf{M}^{-1}\widetilde{\mathbf{K}}$, the following result holds true.

Lemma 1 *Let* $\widetilde{\mathbf{K}}, \mathbf{M}^{-1} \in \mathbb{R}^{N\times N}$ *be the matrices defined in* (3) *and* (16)*, respectively. Then, the eigenvalues* λ *of* $\mathbf{M}^{-1}\widetilde{\mathbf{K}}$ *are either 1, with multiplicity* $2n^h$*, or equal to those of the matrix* $\hat{S}_3^{-1}S_3$*.*

Proof 1 *By using equation* (16)*, the matrix* $\mathbf{M}^{-1}\widetilde{\mathbf{K}}$ *reads:*

$$\mathbf{M}^{-1}\widetilde{\mathbf{K}} = \mathbf{F}\hat{\mathbf{S}}^{-1}\mathbf{G}\widetilde{\mathbf{K}}, \tag{18}$$

which is similar to $\hat{\mathbf{S}}^{-1}\mathbf{G}\widetilde{\mathbf{K}}\mathbf{F}$*. Recalling* (10)*, we have:*

$$\begin{aligned} \hat{\mathbf{S}}^{-1}\mathbf{G}\widetilde{\mathbf{K}}\mathbf{F} &= \hat{\mathbf{S}}^{-1}\mathbf{S} \\ &= \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \hat{S}_3^{-1}S_3 \end{bmatrix}, \end{aligned} \tag{19}$$

which completes the proof.

The key for the effectiveness of $\mathbf{M}^{-1}$ as a preconditioner of $\widetilde{\mathbf{K}}$ is therefore the selection of $\hat{S}_3^{-1}$. In the next paragraph, we analyze the results from different choices for $\hat{S}_3^{-1}$.

4 NUMERICAL RESULTS

Since the effectiveness of $\mathbf{M}^{-1}$ depends on $\hat{S}_3^{-1}$ only, we reduce the system (4) on the flux space:

$$S_3\mathbf{u} = \mathbf{b} \quad \text{with} \quad \mathbf{b} = \left(\alpha B^T - F_{13}^T G^h\right) A^{-1}\mathbf{q} \tag{20}$$

Since S_3 is SPD, system (20) is solved by a preconditioned CG method, setting the maximum number of iterations to 1500 and the exit tolerance on the relative residual to 10^{-6}. Four problems of increasing size have been analyzed (Table 1). Figure 1 shows the mesh domain for the case $\mathrm{P_C}$.

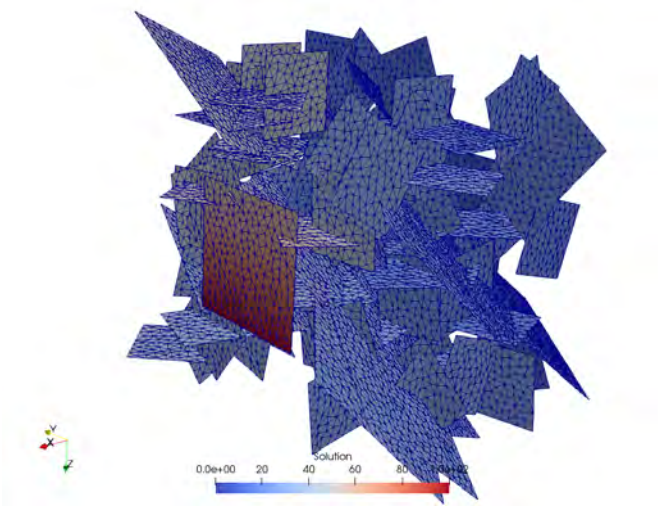


Figure 1: 3D mesh domain for the case $\mathrm{P_C}$.

Table 1: Problem size.

	$\mathrm{P_A}$	$\mathrm{P_B}$	$\mathrm{P_C}$	$\mathrm{P_D}$
n^h	787	13732	39288	93768
n^u	206	5085	8219	18276
N	1780	32549	86795	205812

The non-zero pattern of the matrices of the smallest problem is shown in figure 2. Matrices A, C and G^h are block diagonal. Being each block related to a fracture, these matrices are fracture-local. Instead, matrices B and G^u connect degrees of freedom related to different fractures. In particular, matrix B is made by the same diagonal blocks as C with additional extra-diagonal

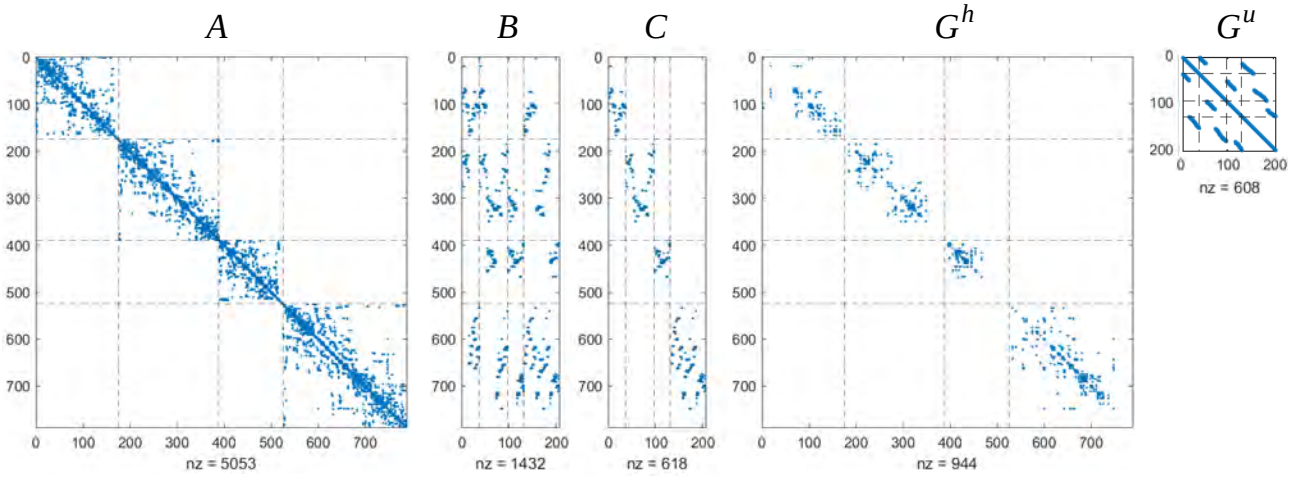


Figure 2: Structure and number of non-zeros of the matrices for case $\mathrm{P_A}$.

terms corresponding to intersections between fractures. Thus, the matrix $E = B - C$ is zero on the diagonal blocks and contains the same terms as B outside. Matrix F_{13}, being defined as $A^{-1}C$, is also block diagonal, with the same size and structure as C.

On the basis of these considerations, the Schur complement can be written as (see equation (14)):

$$S_3 = S_D - S_E \tag{21}$$

where $S_D = G^u + F_{13}^T \left(G^h - 2\alpha A\right) F_{13}$ and $S_E = \alpha \left(E^T F_{13} + F_{13}^T E\right)$. Matrix S_D contains the diagonal blocks of S_3 and S_E the off-diagonal part. Therefore, S_D is SPD, whereas S_E is indefinite.

A key property for $\hat{S}_3$ is being SPD. It is therefore natural to consider $\hat{S}_3 = S_D$, that is the block diagonal and positive definite part. The results in terms of number of iterations (iter), ratio between the non-zeros of the approximate Schur complement and the exact one (μ) and the conditioning number (ξ) are reported in Table 2. Despite the preconditioning, the number of iterations required to solve the system is still high and the conditioning number of the preconditioned matrix is not very different from the original.

Table 2: Results considering the approximation $\hat{S}_3 = S_D$. The $*$ indicates that the problem does not converge, with the residual stagnating around 10^{-5}.

Case	iter	μ	$\xi\left(\hat{S}_3^{-1} S_3\right)$	$\xi\left(S_3\right)$
$\mathrm{P_A}$	125	0.3921	3.10e+04	1.67e+04
$\mathrm{P_B}$	300	0.3958	2.08e+06	4.90e+05
$\mathrm{P_C}$	$*$	0.3619	1.40e+08	1.72e+09
$\mathrm{P_D}$	957	0.3594	7.39e+06	1.15e+09

Approximating S_3 with its diagonal blocks appears to be not enough for an efficient solution of the system. Thus, in the following also the off-diagonal part is taken into account. Aiming at understanding the importance of the single blocks of S_3 as a preconditioner, we filter the two contributions S_D and S_E separately, naming $\hat{S}_D$ and $\hat{S}_E$ their approximation. First, only the extra-diagonal part of S_3 is approximated:

$$\hat{S}_3 = S_D - \hat{S}_E \tag{22}$$

where $\hat{S}_E$ is obtained by filtering each column j of the product $E^T F_{13}$ neglecting the components

such that:

$$\left|\left(E^T F_{13}\right)_{ij}\right| < \tau \left\|\left(E^T F_{13}\right)_j\right\|_2 \tag{23}$$

Results for different values of τ are reported in Table 3.

Table 3: Results computing S_3 with the sparsified S_E. The * indicates the case when $\hat{S}_3$ becomes indefinite.

τ	case P_A			case P_B		
	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$
5×10^{-2}	*	0.8306	5.47e+03	*	0.4747	6.41e+08
10^{-2}	8	0.9398	1.79e+02	26	0.6171	5.07e+06

	case P_C			case P_D		
10^{-2}	*	0.9577	3.26e+09	*	0.8056	4.68e+07
10^{-3}	7	0.9950	2.28e+04	*	0.9742	3.44e+10

Finally, we consider the preconditioner $\hat{S}_3$:

$$\hat{S}_3 = \hat{S}_D - S_E \tag{24}$$

where the extra-diagonal blocks are computed exactly, while the diagonal ones are approximated neglecting the components s_{ij} of the product $F_{13}^T\left(G^h - 2\alpha A\right)F_{13}$ such that:

$$|s_{ij}| < \tau\sqrt{|s_{ii}\ s_{jj}|} \tag{25}$$

Results for the four matrices are reported in Table 4.

Table 4: Results computing S_3 after the sparsification of S_D. The * indicates the case when $\hat{S}_3$ becomes indefinite.

τ	case P_A			case P_B		
	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$
5×10^{-1}	*	0.6604	1.47e+04	*	0.6114	3.51e+08
10^{-1}	10	0.9217	1.99e+03	*	0.6351	4.62e+07
10^{-2}	3	0.9910	2.99e+00	*	0.8902	1.84e+05
10^{-3}	2	0.9987	1.06e+00	*	0.9925	2.33e+04

	case P_C			case P_D		
10^{-2}	*	0.9928	2.34e+04	*	0.9851	4.39e+05
10^{-3}	2	0.9993	2.47e+01	2	0.9987	6.88e+01

In both cases, i.e. when approximating only S_E or S_D, the level of fill-in of $\hat{S}_3$ required for the convergence is near to the one of the exact Schur complement ($\mu \simeq 1$). This is because $\hat{S}_3$ can easily become indefinite after the filtering. As an example, in figure 3 the ten smallest eigenvalues of the exact and the approximated (with τ equal to 5×10^{-1}) Schur complement for the case P_A are shown. While S_3 is positive definite, the eigenvalues of $\hat{S}_3$ are both positive and negative.

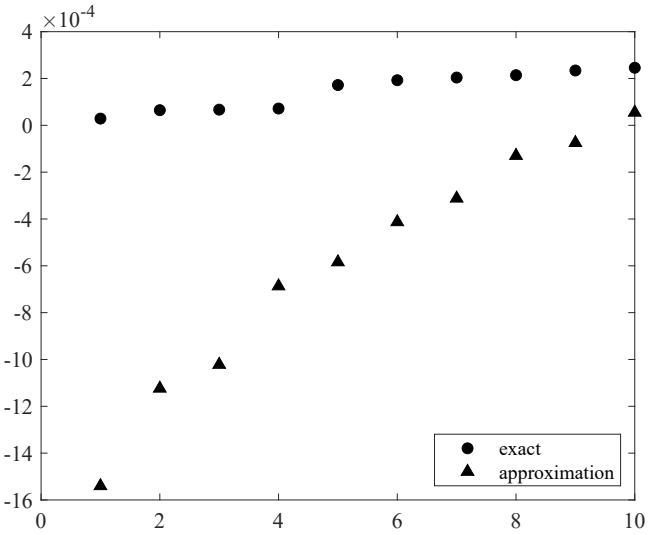


Figure 3: Case $\mathrm{P_A}$: ten smallest eigenvalues of S_3 and $\hat{S}_3$ computed through equation (24) with $\tau = 5 \times 10^{-1}$.

In the last test, the preconditioner is computed approximating both S_D and S_E:

$$\hat{S}_3 = \hat{S}_D - \hat{S}_E \tag{26}$$

To this aim, a sparsified F_{13} is computed by filtering the smallest components. Since F_{13} is block diagonal, it can be efficiently computed in a parallel computational environment exploiting a Cholesky factorization of the blocks of A. A relative drop tolerance is used, removing the components such that:

$$|F_{13,ij}| < \tau \left\|F_{13,j}\right\|_2 \tag{27}$$

Results are reported in Table 5. The iterations count can decrease significantly with respect to Table 2, with densities that are even smaller than those obtained keeping S_D only. However, in difficult problems, such as $\mathrm{P_C}$, quite a high fill-in can be required and the performance can be very sensitive to the τ selection.

Table 5: Results computing S_3 with the approximation of F_{13}.

τ	case $\mathrm{P_A}$			case $\mathrm{P_B}$		
	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$	iter	μ	$\xi\left(\hat{S}_3^{-1}S_3\right)$
10^{-1}	28	0.2697	3.75e+04	136	0.1226	1.88e+07
5×10^{-2}	19	0.8595	5.85e+04	57	0.2508	1.45e+07
10^{-2}	1	1.0000	1	17	0.6072	1.48e+06

	case $\mathrm{P_C}$			case $\mathrm{P_D}$		
5×10^{-2}	1483	0.5400	1.08e+11	445	0.3196	1.06e+09
2.5×10^{-2}	8	0.9952	1.64e+06	128	0.5663	8.89e+07
10^{-2}	4	0.9990	3.38e+04	41	0.8100	1.57e+07
10^{-3}	1	1.0000	1	5	0.9912	1.10e+05

Considering as a preconditioner an approximation $\hat{S}_3$ obtained by filtering S_3 or its components can be efficient (as results in Table 5 demonstrate), but also quite fragile because of the possible indefiniteness of the approximation (see Table 3 and Table 4).

5 CONCLUSIONS

A symmetric saddle-point matrix with a rank-deficient leading block arises from the combination of DFN models with an appropriate optimization formulation. Here, we focused on accelerating the iterative solution of this system with a block preconditioning technique. First, an appropriate permutation of the matrix is performed and then a projection on the Schur complement space of the flux is performed. The Schur complement proves to be the key for an effective preconditioner, therefore we investigate different approaches to approximate it. Both the diagonal and off-diagonal blocks of the Schur complement are fundamental for an efficient solution of the system. Independent filterings of such components reveal the fragility of the approximated Schur complement, that can easily become indefinite. When the filter step regards the matrix F_{13}, before the computation of the Schur complement, results are more promising. This suggests to investigate different other filtering approaches for F_{13}, aiming at finding a more robust and less τ dependent solution. Alternatively, a polynomial acceleration in a matrix-free implementation can help improving the performance when working in a parallel environment. Comparing the approaches we investigated, we noted that the conditioning number does not vary according to the number of iterations, as one can expect. This can be related to the distribution of the eigenvalues, that means that the eigenspectrum is mainly grouped, but there are a few outliers. In order to fix this problem, a deflation approach can be used to remove the eigenvectors related to the extreme eigenvalues. This technique requires the a priori knowledge of these eigenvalues, that is quite computational expensive, but reasonable in an iterative framework.

REFERENCES

[1] Berrone, S. and Pieraccini, S. and Scialò, S. A PDE-constrained optimization formulation for Discrete Fracture Network flows. *SIAM J. Sci. Comput.*, Vol. **35**, pp. B487–B510, (2013).

[2] Berrone, S. and Scialò, S. and Vicini, F. Parallel meshing, discretization and computation of flow in massive Discrete Fracture Networks. *SIAM J. Sci. Comput.*, Vol. **41**, pp. C317-C338, (2019).

[3] Benzi, M. and Golub, G.H. and Liesen, J. Numerical solution of saddle point problems. *Acta Numerica*, Vol. **14**, pp. 1–137, (2005).

[4] Elman, H.C. and Silvester, D.J. and Wathen, A.J. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. Oxford University Press, 2005.

[5] Estrin, R. and Greif, C. On nonsingular saddle-point systems with a maximally rank-deficient leading block. *SIAM Journal on Matrix Analysis and Applications*, Vol. **36**, pp. 367–384, (2015).

[6] Ferronato, M. and Franceschini, A. and Janna, C. and Castelletto, N. and Tchelepi, H.A. A general preconditioning framework for coupled multi-physics problems. *J. Comput. Phys.*, Vol. **398**, (2019).

ITERATIVE QUASI-NEWTON SOLVERS FOR POROMECHANICS APPLIED TO HEART PERFUSION

N. Barnafi*, J. W. Both†

* Dipartimento di Matematica "F. Enriques"
Università degli Studi di Milano
Milan, Italy
nicolas.barnafi@unimi.it

† Department of Mathematics
University of Bergen
Bergen, Norway
jakub.both@uib.no

Key words: Poromechanics, Iterative solvers, Fixed-stress split, Biomedical application

Abstract: *In this work, the efficient approximation of a nonlinear cardiac poromechanics model is investigated. Quasi-Newton solvers based on iterative two-way and three-way decoupling are proposed. For increased robustness and better performance, the iterative schemes are accelerated by additionally using Anderson acceleration. The solvers are tested for a numerical example simulating cardiac perfusion. The results obtained demonstrate a significant speed-up for the splitting approaches with respect to the standard monolithic Newton method.*

1 INTRODUCTION

Cardiac perfusion describes the fundamental process of blood supply of the heart muscle, but its importance at the outset of cardiac disease remains largely understudied. This motivates the use of mathematical models to deepen the understanding of this phenomenon. The complex network structure of the coronary vessels in the heart and tissue itself (myocardium) have been mainly addressed by the use of poroelastic models [14, 9, 11], which possess the advantage of greatly reducing the complexity of the vessels through formal averaging techniques [20].

Nonlinear poroelasticity consists in a complex multi-physics model, whose numerical approximation is still under active research. The linear case, i.e. Biot's equation, is instead better understood, with iterative coupling strategies presenting the most successful family of methods for this kind of problem. The main ones used in practice are the undrained [21] and fixed-stress [16] splitting schemes. These methods alternate between solving for flow and then solid variables until convergence, while keeping the others fixed. For guaranteed robustness, however, sufficient stabilization has to be used which can be obtained through analysis. Computational costs may be significantly reduced due to the decoupling, which relies on solving many times simpler sub-problems instead of solving once a difficult problem. The concept of decoupling can be extended to nonlinear problems as a quasi-Newton method [15, 4, 5], where the computational cost reduction can become even more relevant.

In this work, we study the nonlinear solution of a simplified nonlinear model for cardiac poromechanics. The model combines thermodynamically-consistent linearization [7] of a fully-nonlinear model [10], but with a nonlinear constitutive stress-strain relation [13]; the final model consists in a nonlinear coupled system of three physics. Quasi-Newton solvers are proposed integrating stabilized two-way and three-way decoupling, inspired by splitting schemes derived for the linearized model [3] which guarantee linear convergence. Our numerical results show that our iterative splitting quasi-Newton schemes outperform the widely used monolithic Newton method, with a reduction in computer times of up to a 50% for the three-way and an 85% for the two-way, making them an attractive choice for the fully-nonlinear models.

https://doi.org/10.4995/YIC2021.2021.12324

2 NONLINEAR POROELASTICITY MODEL FOR CARDIAC PERFUSION

The scope of the following model is twofold: on one hand, it captures the interaction between the deformation of the myocardium during a heartbeat and the myocardial coronary vessels, and on the other hand it provides a simple scenario in which numerical methods can be tested. We pose our problem on a prolate ellipsoid geometry Ω representing a left ventricle, cf. Figure 2.

The poromechanics model we consider is given by the following: Find a displacement $\boldsymbol{y}_s$, absolute fluid velocity $\boldsymbol{v}_f$ and pressure p such that

$$\begin{aligned} \mathcal{F}_s &:= \rho_s(1-\phi)\partial_{tt}\boldsymbol{y}_s - \operatorname{div}\,\mathbf{P}(\mathbf{F},t) + (1-\phi)\,\nabla p - \phi^2\boldsymbol{\kappa}_f^{-1}\,(\boldsymbol{v}_f - \partial_t\boldsymbol{y}_s) = \mathbf{0} \quad \text{in } \Omega, \\ \mathcal{F}_f &:= \rho_f\phi\partial_t\boldsymbol{v}_f - \operatorname{div}\,(\phi\boldsymbol{\sigma}_{\text{vis}}(\boldsymbol{v}_f)) + \phi\,\nabla p + \phi^2\boldsymbol{\kappa}_f^{-1}\,(\boldsymbol{v}_f - \partial_t\boldsymbol{y}_s) = \mathbf{0} \quad \text{in } \Omega, \\ \mathcal{F}_p &:= \frac{(1-\phi)^2}{\kappa_{\text{s}}}\partial_t p + \operatorname{div}\,(\phi\boldsymbol{v}_f) + \operatorname{div}\,((1-\phi)\,\partial_t\boldsymbol{y}_s) = 0 \quad \text{in } \Omega, \end{aligned} \tag{1}$$

where $\boldsymbol{\sigma}_{\text{vis}} := 2\mu_f\boldsymbol{\varepsilon}(\boldsymbol{v}_f)$ and $\boldsymbol{P}$ is the Piola stress tensor, given by $\boldsymbol{P}(\boldsymbol{F},t) := \frac{d\Psi}{d\boldsymbol{F}} + \boldsymbol{P}_a(\boldsymbol{F},t)$ for a Helmholtz potential Ψ and an active stress tensor $\boldsymbol{P}_a(\boldsymbol{F},t)$, specified further below. The remaining parameters are: solid density ρ_s, fluid density ρ_f, porosity ϕ, permeability tensor $\boldsymbol{\kappa}_f$ and bulk modulus κ_s.

To model the ventricle mechanics, a Guccione fiber oriented constitutive law [13] was used together with an artificial active contraction force. The constitutive law is given by

$$\Psi(\boldsymbol{F}) := C\exp\{Q(\boldsymbol{F})-1\} + \frac{\kappa}{2}(J-1)\log J, \tag{2}$$

$$Q := b_f E_{ff}^2 + b_s E_{ss}^2 + b_n E_{nn}^2 + 2(b_{fs}E_{fs}^2 + b_{fn}E_{fn}^2 + b_{sn}E_{sn}^2), \tag{3}$$

$$\boldsymbol{E} := \frac{1}{2}(\boldsymbol{F}^T\boldsymbol{F} - \boldsymbol{I}),\ \boldsymbol{F} = \nabla\,\boldsymbol{y}_s + \boldsymbol{I},\ J := \det(\boldsymbol{F}),\ E_{uv} := (\boldsymbol{E}\boldsymbol{v})\cdot\boldsymbol{u}, \tag{4}$$

where $\boldsymbol{f}, \boldsymbol{s}$ and $\boldsymbol{n}$ are a pointwise set of independent vectors directed towards the heart fibers, sheets and normal directions, and the active stress is given by

$$\boldsymbol{P}_a(\boldsymbol{F},t) := 3\cdot 10^4\,\sin(\pi t)\frac{(\boldsymbol{F}\boldsymbol{f})\otimes\boldsymbol{f}}{\|\boldsymbol{F}\boldsymbol{f}\|}. \tag{5}$$

We use the same parameters from [18]: $C = 0.88\cdot 10^3, b_f = 8, b_s = 6, b_n = 3, b_{fs} = 12, b_{fn} = 3, b_{sn} = 3, \kappa = 5\cdot 10^4$ for the nonlinear constitutive law, and the ones from [7] for the remaining parameters: $\rho_f = \rho_s = 10^3$, $\phi = 0.1$, $\boldsymbol{\kappa}_f = 10^{-7}$ and $\kappa_s = 10^8$. All parameters are considered within the SI unit system.

We note that this is a hybrid model, in the sense that it includes a nonlinear mechanics response but it does not account for large deformations in the fluid momentum and mass conservation. Still, it correctly captures the deformation pattern of a beating heart and, as our results show, provides an adequate framework for studying the interaction strength between the mechanics and the porous media flow. We thereby expect conclusions of this work to be also applicable to extended models.

2.1 Initial and boundary conditions

The initial conditions are simply given by

$$\boldsymbol{y}_s(0) = \boldsymbol{y}_{s0},\quad \partial_t\boldsymbol{y}_s(0) = \boldsymbol{v}_{s0},\quad \boldsymbol{v}_f(0) = \boldsymbol{v}_{f0},\quad p(0) = p_0. \tag{6}$$

The boundary conditions are defined as follows: the mechanics follow the Robin boundary conditions from [18] which model the interaction with the pericardium at both the epicardium

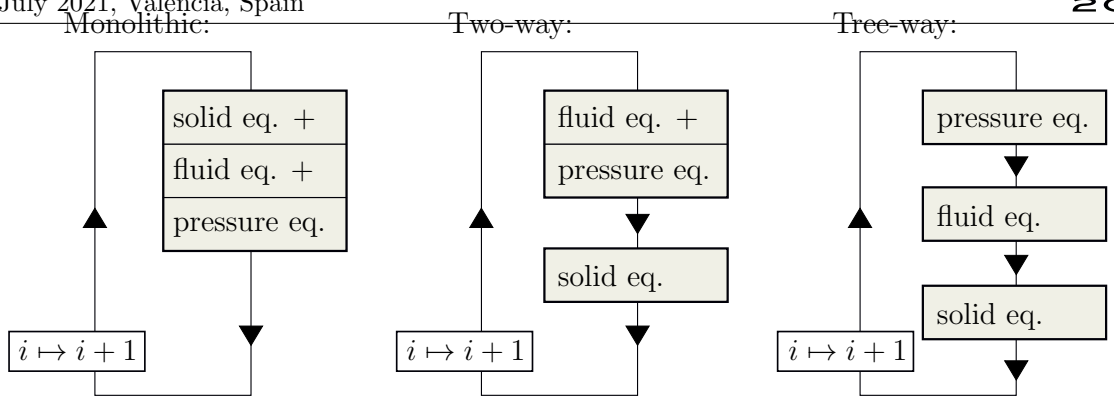


Figure 1: Summary of the monolithic, two-way, and three-way split solver strategies at an arbitrary iteration i, where connected blocks denote coupled physics.

(external surface) and the base (circular ring on top), whereas the endocardium (inner surface) uses a null Neumann condition. For the blood we consider a no-slip condition $\boldsymbol{v}_f = \partial_t \boldsymbol{y}_s$ on the endocardium and epicardium, and at the base of the heart we set a null Neumann condition which allows for the blood to freely leave the tissue.

2.2 Numerical discretization

We use the continuous Galerkin finite element method for solving this problem. For this, we consider the inf-sup stable family of generalized Taylor-Hood elements $\mathbb{P}_2 \times \mathbb{P}_2 \times \mathbb{P}_1$ [2] for the solid×fluid×pressure space, and we use an implicit Euler method with a fixed time-step Δt. We note that this approach is equally valid for the energy-consistent discretization shown in [6], as well as higher order methods. The no-slip condition is imposed weakly as in [6].

The use of higher order finite elements for the displacement, although less frequently used in the mechanics community, are common practice in the field of geomechanics, requiring an inf-sup stability condition for the displacement and pressure finite element spaces. Yet, we note that such discretization has also already been used in the context of cardiac poromechanics [11]. This relation is also true for the linear model we used as a base for our hybrid model, with the inf-sup constant being proportional to the solid porosity $1 - \phi$ for the case in which the displacement is approximated with the lower order $\mathbb{P}_1$ elements [2].

3 MONOLITHIC AND BLOCK-PARTITIONED NUMERICAL SOLVERS

We present three iterative solver strategies for solving the nonlinear problem from Section 2: a monolithic, a two-way splitting, and a three-way splitting approach, outlined in Fig. 1. The monolithic scheme is the standard Newton method, whereas the splitting schemes are formulated as quasi-Newton solvers, i.e. each iteration is a linearization iteration decoupling different physical sub-problems by a suitable choice of the inexact Jacobian. The decoupling strategies are closely related to previous developments for the corresponding linearized problem [3].

Through simultaneous linearization and decoupling, a significant reduction in overall computational cost can be expected as observed for other nonlinear poroelasticity problems [4]. Furthermore, we suggest employing Anderson acceleration to both improve the performance and slightly relax the need for well-chosen stabilization parameters.

All solvers are formulated in residual form, allowing in particular for a direct comparison. For this, we denote with $\mathcal{F}_s$, $\mathcal{F}_f$, and $\mathcal{F}_p$ the canonical residuals of the solid momentum, fluid momentum, and mass conservation/pressure equations, respectively. Throughout the remaining section, i will denote the current iteration index which decorates approximations, e.g. $\boldsymbol{y}_s{}^i$, as well as increments, e.g., $\delta \boldsymbol{y}_s{}^i$.

The resulting schemes consider at each iteration i some approximation $(\boldsymbol{y}_s{}^i, \boldsymbol{v}_f^i, p^i)$, for which

we compute an increment $(\delta \boldsymbol{y}_s{}^i, \delta \boldsymbol{v}_f{}^i, \delta p^i)$. The next iteration is then defined by $(\boldsymbol{y}_s{}^{i+1}, \boldsymbol{v}_f^{i+1}, p^{i+1}) := (\boldsymbol{y}_s{}^i, \boldsymbol{v}_f^i, p^i) + (\delta \boldsymbol{y}_s{}^i, \delta \boldsymbol{v}_f{}^i, \delta p^i)$. In the following, we specify the definition of the different linearization steps.

3.1 Monolithic Newton solver

The monolithic Newton solver is usually the first-choice linearization scheme for nonlinear problems (see [12] for the case of cardiac mechanics). The Jacobian is given by a full linearization of the governing equations (1) after discretization.

The linearization step at iteration $i \geq 0$ reads: Compute $(\delta \boldsymbol{y}_s{}^i, \delta \boldsymbol{v}_f{}^i, \delta p^i)$ such that

$$\frac{\rho_s(1-\phi)}{\Delta t^2}\delta \boldsymbol{y}_s{}^i - \operatorname{div}\, \partial_{\boldsymbol{y}_s} \boldsymbol{P}(\boldsymbol{F}^i) : \delta \boldsymbol{y}_s{}^i + (1-\phi)\,\nabla\, \delta p^i \\ -\phi^2 \boldsymbol{\kappa}_f^{-1}\left(\delta \boldsymbol{v}_f{}^i - \frac{\delta \boldsymbol{y}_s{}^i}{\Delta t}\right) = -\mathcal{F}_s(\boldsymbol{y}_s{}^i, \boldsymbol{v}_f^i, p^i), \tag{7}$$

$$\frac{\rho_f \phi}{\Delta t}\delta \boldsymbol{v}_f{}^i - \operatorname{div}\, \left(\phi \boldsymbol{\sigma}_{\text{vis}}(\delta \boldsymbol{v}_f{}^i)\right) + \phi\,\nabla\, \delta p^i \\ +\phi^2 \boldsymbol{\kappa}_f^{-1}\left(\delta \boldsymbol{v}_f{}^i - \frac{\delta \boldsymbol{y}_s{}^i}{\Delta t}\right) = -\mathcal{F}_f(\boldsymbol{y}_s{}^i, \boldsymbol{v}_f^i, p^i), \tag{8}$$

$$\frac{(1-\phi)^2}{\kappa_{\text{s}} \Delta t}\delta p^i + \operatorname{div}\, \left(\phi \delta \boldsymbol{v}_f{}^i\right) + \operatorname{div}\, \left((1-\phi)\,\frac{\delta \boldsymbol{y}_s{}^i}{\Delta t}\right) = -\mathcal{F}_p(\boldsymbol{y}_s{}^i, \boldsymbol{v}_f^i, p^i), \tag{9}$$

Algebraically, this can be written as

$$\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i & \boldsymbol{A}_{fs}^\top & -\boldsymbol{B}_s^T \\ \boldsymbol{A}_{fs} & \boldsymbol{A}_f & -\boldsymbol{B}_f^T \\ \boldsymbol{B}_s & \boldsymbol{B}_f & \boldsymbol{A}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta y_s}{}^i \\ \boldsymbol{\delta v_f}{}^i \\ \boldsymbol{\delta p}^i \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_s^i \\ \boldsymbol{R}_f^i \\ \boldsymbol{R}_p^i \end{bmatrix}, \tag{10}$$

with natural definitions of the block matrices $\boldsymbol{A}_{(\cdot)(\cdot)}, \boldsymbol{B}_{(\cdot)}$ and residual vectors $\boldsymbol{R}_{(\cdot)}$. The monolithic solver strategy does not utilize the fact that all blocks aside of the solid diagonal block $\boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i$ are constant. Splitting solvers are instead capable of making use of all constant blocks, i.e. $\boldsymbol{A}_{(\cdot)(\cdot)}$ and $\boldsymbol{B}_{(\cdot)}$.

3.2 Two-way splitting

We employ ideas previously developed for the linearized problem [3] justified by a similar coupling character of the exact Jacobian, cf. Sec. 3.1. For this, the mechanics equations are decoupled from the remaining two equations, and the mass conservation equation is stabilized with a weighted L^2-type term – essentially as in the fixed-stress split for Biot's equations.

Let β_p denote a user-defined stabilization parameter, and associate a weighted L^2-type bilinear form $(p, q) \mapsto \beta_p \langle p, q \rangle_{L^2}$ with a corresponding discretization matrix $\boldsymbol{S}_p$. Then the two-way split is given by a decoupled solver with diagonal L^2-type stabilization, which can be written algebraically at each iteration i as: Find the increment $(\boldsymbol{\delta y_s}{}^i, \boldsymbol{\delta v_f}{}^i, \boldsymbol{\delta p}^i)$ satisfying

$$\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i & \boldsymbol{A}_{fs}^\top & -\boldsymbol{B}_s^T \\ \boldsymbol{0} & \boldsymbol{A}_f & -\boldsymbol{B}_f^T \\ \boldsymbol{0} & \boldsymbol{B}_f & \boldsymbol{A}_p + \boldsymbol{S}_p \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta y_s}{}^i \\ \boldsymbol{\delta v_f}{}^i \\ \boldsymbol{\delta p}^i \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_s^i \\ \boldsymbol{R}_f^i \\ \boldsymbol{R}_p^i \end{bmatrix} \tag{11}$$

Equivalently, the two-way split can be performed in two separate steps. First the coupled fluid momentum and stabilized mass conservation equations are solved. Second, the solid

momentum equation is solved with updated fluid flow parameters. We highlight that the first step does not require any setup update over the course of iterations.

For the linearized problem, convergence can be showed for a range of stabilization values [3]. We expect similar robustness in the nonlinear case.

3.3 Three-way splitting

The two-way split still involves the solution of the coupled fluid momentum and mass conservation equations, which have the character of a time-dependent Stokes equations. Inspired by developments for the time-dependent Stokes equations [8], we apply additional decoupling with two sub-steps accounting for one of the two contributions (L^2-type and diffusion-type) in the fluid velocity diagonal block $\boldsymbol{A}_f$. Similarly to the fixed-stress split, the diffusion contribution in $\boldsymbol{A}_f$ suggests an L^2-type stabilization $\boldsymbol{S}_{CC,mass}$, associated to $(p,q) \mapsto \beta_{CC,mass}\langle p,q\rangle_{L^2(\Omega)}$, whereas the L^2-type contribution in $\boldsymbol{A}_f$ results in a Laplace-type stabilization $\boldsymbol{S}_{CC,diff}$, associated with $(p,q) \mapsto \beta_{CC,diff}\langle \nabla p, \nabla q\rangle_{L^2(\Omega)}$. Here, $\beta_{CC,mass}$ and $\beta_{CC,diff}$ denote two (additional) user-defined stabilization parameters.

The total increment is then obtained through mixing with parameter $\gamma \in [0,1]$

$$(\delta \boldsymbol{y_s}^i, \delta \boldsymbol{v_f}^i, \delta p^i) \coloneqq \gamma(\delta \boldsymbol{y_s}^i_{mass}, \delta \boldsymbol{v_f}^i_{mass}, \delta p^i_{mass}) + (1-\gamma)(\delta \boldsymbol{y_s}^i_{diff}, \delta \boldsymbol{v_f}^i_{diff}, \delta p^i_{diff}) \tag{12}$$

where the two increments are computed by solving the three-way splitting methods

$$\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i & \boldsymbol{A}_{fs}^\top & -\boldsymbol{B}_s^T \\ \boldsymbol{0} & \boldsymbol{A}_f & -\boldsymbol{B}_f^T \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_p + \boldsymbol{S}_p + \boldsymbol{S}_{CC,mass} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta y_s}^i_{mass} \\ \boldsymbol{\delta v_f}^i_{mass} \\ \boldsymbol{\delta p}^i_{mass} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_s^i \\ \boldsymbol{R}_f^i \\ \boldsymbol{R}_p^i \end{bmatrix}, \tag{13}$$

and

$$\begin{bmatrix} \boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i & \boldsymbol{A}_{fs}^\top & -\boldsymbol{B}_s^T \\ \boldsymbol{0} & \boldsymbol{A}_f & -\boldsymbol{B}_f^T \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_p + \boldsymbol{S}_p + \boldsymbol{S}_{CC,diff} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta y_s}^i_{diff} \\ \boldsymbol{\delta v_f}^i_{diff} \\ \boldsymbol{\delta p}^i_{diff} \end{bmatrix} = - \begin{bmatrix} \boldsymbol{R}_s^i \\ \boldsymbol{R}_f^i \\ \boldsymbol{R}_p^i \end{bmatrix}. \tag{14}$$

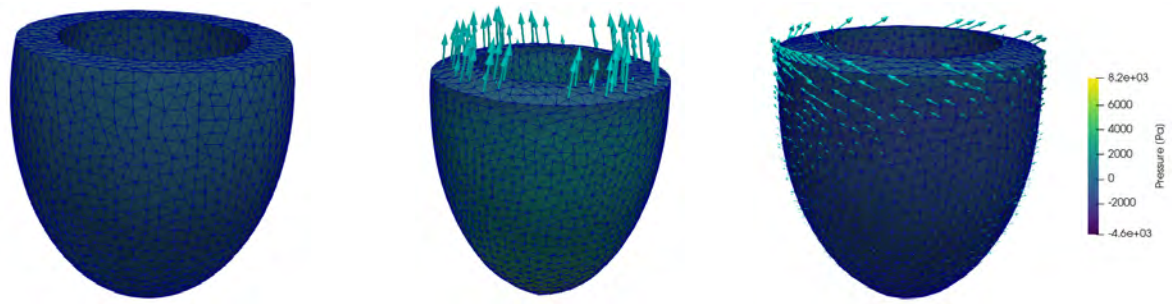
3.4 Choice of stabilization parameters and acceleration

The two-way and three-way splitting schemes involve the choice of user-defined stabilization parameters β_p, $\beta_{CC,mass}$, $\beta_{CC,diff}$, and a mixing parameter γ. In the numerical example in Section 4, inspired by the strategy in [17], we manually chose $\beta_p = 0.22$ as it resulted in the fewest iterations for the first 10 time steps. We keep the value fixed over the course of the entire simulation. The linear structure of the fluid-pressure coupling naturally suggests $\beta_{CC,mass} = 3\phi/(2\mu_f)$ and $\beta_{CC,diff} = \left(\rho_f(\Delta t)^{-1}\boldsymbol{I} + \boldsymbol{\kappa}_f^{-1}\right)^{-1}$. The mixing parameter is chosen as $\gamma = 0.9$ to favor the L^2-type stabilization.

The performance of the solvers depends on the choice of the parameters. However, the nonlinear character of the problem impedes optimization at each iteration; we note that the non-constant diagonal block $\boldsymbol{D}_{\boldsymbol{y}_s}\boldsymbol{R}_s^i$ in particular controls β_p. For remedy, we employ the multisecant and nonlinear GMRES method called Anderson acceleration [19]. As observed in [4], the need for optimized stabilization can be expected to be strongly relaxed, in addition to a generally improved performance.

4 NUMERICAL TESTS

In this section we present the performance of the quasi-Newton schemes with respect to the standard Newton method. We consider the solution of problem (1), whose solution is shown in Figure 2. It can be seen that (i) the base of the geometry allows for free flow of blood and (ii)



$t = 0.0$ $t = 0.1$ $t = 0.4$

Figure 2: Simulation of the poromechanics model at rest ($t = 0.0$), systole ($t = 0.1$) and dyastole ($t = 0.4$). Deformation is illustrated via the deformation of the geometry, and fluid velocity by the arrows.

it can be seen that the fluid is an absolute velocity, as it follows the heart's untwisting motion during diastole.

To study the splitting schemes we focus on the first 30 time-steps (up to $t = 3 \cdot 10^{-2}$), the mesh used yields around 130 000 degrees of freedom, and all sub-blocks being solved with the GMRES method. We use a right ILU preconditioner with 1 level of fill-in for the splitting methods, and readily highlight that this was not possible for the monolithic Newton solver, as it diverged. For convergence, we required 3 levels of fill-in for the monolithic case. The absolute and relative tolerances used for the linear solvers (GMRES) were 10^{-10} and 10^{-8} for the residual, and instead for the nonlinear solvers (Newton and quasi-Newton) we used 10^{-8} and 10^{-6}, computed through the residual as well. All tests were run in serial to avoid mixing the results with the parallel performance of the precondtioners[1]. The implementation was performed using the FEniCS library [1].

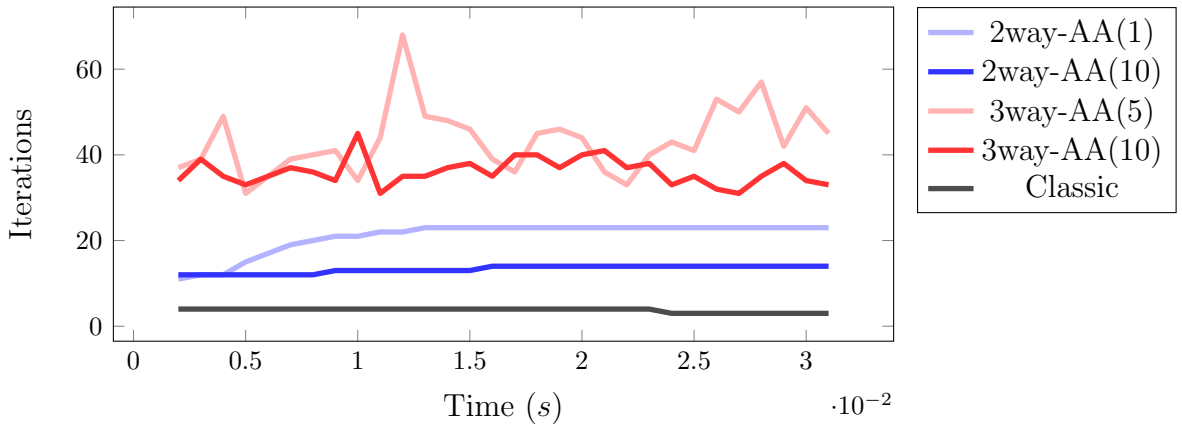
Results are shown in Figure 3, where we depict both the iterations of the nonlinear solvers and the wall-time. We note that Anderson acceleration was fundamental for the convergence of the quasi-Newton schemes, with different levels of depth being required; the depth denotes the amount of previous iterations utilized for determining the next approximation. The two-way split required a depth of at least one, whereas the three-way required a depth of at least 5. It must still be further studied whether the impact of Anderson acceleration is due to an improved robustness with respect to the stabilization parameters as observed in [4] or instead because it improves the convergence of the quasi-Newton itself.

For the iteration counts, cf. Figure 3a, we note that as expected the monolithic Newton method presents a much more robust behavior, with a maximum of three iterations per time-step. The accelerated two-way, albeit with more iterations, also does not present large oscillations in the iteration count, with a minor improvement obtained through further acceleration. The three-way instead varies from 33 to 68 iterations when using 5 levels of acceleration, this behavior being greatly reduced with an acceleration depth of 10, which presents iteration numbers between 31 and 43. This shows the effectiveness of Anderson acceleration in granting robustness to the iterative splitting schemes while the problem character changes over time due to the nonlinearities.

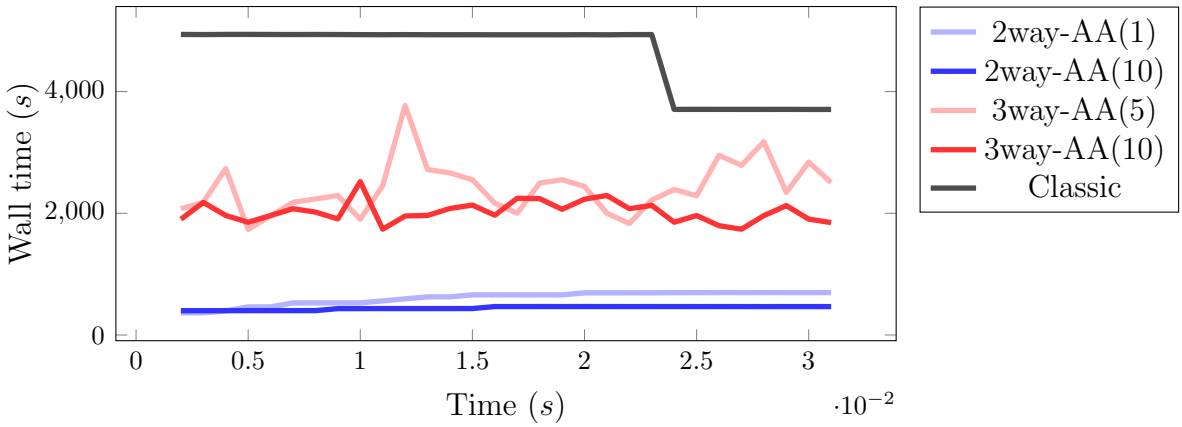
The solution times, cf. Figure 3b, show clearly the superiority of splitting schemes to the monolithic Newton method. The two-way split with one level of acceleration solves each time-step at roughly 15% of the time it takes the monolithic Newton method, whereas the three-way split with 5 levels of acceleration takes in average a 50% of the monolithic time. These times

[1]Simulations were performed in the Indaco cluster from the University of Milan.

can be further improved by means of acceleration, where using an acceleration depth of 10, we see that in average the solution time at each time-step is further reduced to roughly a 10% and a 43% for the two- and three-way respectively, both with respect to the Newton solver.



(a) Number of nonlinear solver iterations.



(b) Total solution time.

Figure 3: Iteration counts and solution time at each time step for different nonlinear solvers; depth of Anderson acceleration (AA) in parentheses.

5 CONCLUSIONS

In this work, we have proposed an extension of a two-way splitting scheme for a linearized model presented in [3], now applied as quasi-Newton methods for a nonlinear poroelasticity problem. Both two-way and three-way splitting schemes are considered. Both schemes have been numerically tested for a simplified, nonlinear model of cardiac poromechanics. Our results are very encouraging: the two-way splitting scheme presented an average reduction of the solution time of up to a 85% with respect to the classic Newton scheme, and the three-way a reduction of roughly 50%.

Anderson acceleration provided a crucial improvement to the quasi-Newton methods, without which they would have not converged. The amount of previous iterations required by Anderson depends on the scheme used, 1 being sufficient for the two-way, and instead 5 for the three-way. The three-way splitting scheme, although slightly less performant in this case, presents a more attractive alternative for high performance simulations as it does not require the preconditioning of a saddle point block.

The difference in performance between the monolithic and splitting schemes is mainly justified by the difficulty of devising an efficient preconditioner for the monolithic problem, as can already be seen by the requirement of using additional fill-in with the ILU (1 for the splitting schemes, 3 for the monolithic one). Splitting schemes leverage on the solution of the better understood sub-blocks to yield an overall more efficient solver with potentially better computational complexity.

Future work will be devoted to further investigate three-way decoupling techniques in the context of preconditioning. In addition, the application of quasi-Newton methods inspired by decoupling approaches for the fully nonlinear cardiac poromechanics will be further studied.

6 ACKNOWLEDGEMENTS

The development of this document has been supported by the following projects: "Modeling the heart across the scales: from cardiac cells to the whole organ" PRIN 2017AXL54F_003 P.I. S. Scacchi (NB), Project 250223 Research Council of Norway (JWB), and the FracFlow project funded by Equinor through Akademiaavtalen (JWB). The authors also thank Florin Radu, Paolo Zunino and Alfio Quarteroni for inspiring discussions.

REFERENCES

[1] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. Rognes, and G. Wells. The FEniCS project version 1.5. *Archive of Numerical Software*, 3(100), 2015.

[2] N. Barnafi, P. Zunino, L. Dedè, and A. Quarteroni. Mathematical analysis and numerical approximation of a general linearized poro-hyperelastic model. *Computers & Mathematics with Applications*, 2020.

[3] J. Both, N. Barnafi, F. Radu, P. Zunino, and A. Quarteroni. Iterative splitting schemes for a soft material poromechanics model. *Computer Methods in Applied Mechanics and Engineering*, 388:114183, 2022.

[4] J. Both, K. Kumar, J. Nordbotten, and F. Radu. Anderson accelerated fixed-stress splitting schemes for consolidation of unsaturated porous media. *Computers & Mathematics with Applications*, 77(6):1479 – 1502, 2019. 7th International Conference on Advanced Computational Methods in Engineering (ACOMEN 2017).

[5] J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu. Iterative methods for coupled flow and geomechanics in unsaturated porous media. In *Poromechanics VI*, pages 411–418. 2017.

[6] B. Burtschell, D. Chapelle, and P. Moireau. Effective and energy-preserving time discretization for a general nonlinear poromechanical formulation. *Computers and Structures*, 182:313–324, 2017.

[7] B. Burtschell, P. Moireau, and D. Chapelle. Numerical analysis for an energy-stable total discretization of a poromechanics model with inf-sup stability. *Acta Mathematicae Applicatae Sinica*, 35(1):28–53, 2019.

[8] J. Cahouet and J.-P. Chabard. Some fast 3d finite element solvers for the generalized stokes problem. *International Journal for Numerical Methods in Fluids*, 8(8):869–895, 1988.

[9] D. Chapelle, J. Gerbeau, J. Sainte-Marie, and I. Vignon-Clementel. A poroelastic model valid in large strains with applications to perfusion in cardiac modeling. *Computational Mechanics*, 46(1):91–101, 2010.

[10] D. Chapelle and P. Moireau. General coupling of porous flows and hyperelastic formulations - from thermodynamics principles to energy balance and compatible time schemes. *European Journal of Mechanics, B/Fluids*, 46:82–96, 2014.

[11] A. Cookson, J. Lee, C. Michler, R. Chabiniok, E. Hyde, D. Nordsletten, M. Sinclair, M. Siebes, and N. Smith. A novel porous mechanical framework for modelling the interaction between coronary perfusion and myocardial mechanics. *Journal of biomechanics*, 45(5):850–855, 2012.

[12] A. Gerbi, L. Dedè, and A. Quarteroni. A monolithic algorithm for the simulation of cardiac electromechanics in the human left ventricle. *Mathematics In Engineering*, 1(1):1–37, 2019.

[13] J. Guccione, A. McCulloch, and L. Waldman. Passive material properties of intact ventricular myocardium determined from a cylindrical model. *Journal of biomechanical engineering*, 113(1):42–55, 1991.

[14] J. Huyghe, T. Arts, D. van Campen, and R. Reneman. Porous medium finite element model of the beating left ventricle. *American Journal of Physiology-Heart and Circulatory Physiology*, 262(4):H1256–H1267, 1992.

[15] M. A. B. Reverón, K. Kumar, J. M. Nordbotten, and F. A. Radu. Iterative solvers for biot model under small and large deformations. *Computational Geosciences*, 25(2):687–699, 2021.

[16] A. Settari and F. Mourits. A coupled reservoir and geomechanical simulation system. *SPE Journal*, 3(3):219–226, 1998.

[17] E. Storvik, J. W. Both, K. Kumar, J. M. Nordbotten, and F. A. Radu. On the optimization of the fixed-stress splitting for biot's equations. *International Journal for Numerical Methods in Engineering*, 120(2):179–194, 2019.

[18] T. Usyk, I. LeGrice, and A. McCulloch. Computational model of three-dimensional cardiac electromechanics. *Computing and Visualization in Science*, 4(4):249–257, Jul 2002.

[19] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

[20] S. Whitaker. Flow in porous media i: A theoretical derivation of darcy's law. *Transport in porous media*, 1(1):3–25, 1986.

[21] O. Zienkiewicz, D. Paul, and A. Chan. Unconditionally stable staggered solution procedure for soil-pore fluid interaction problems. *International Journal for Numerical Methods in Engineering*, 26(5):1039–1055, 1988.

A numerical scheme for two-scale phase-field models in porous media

Manuela Bastidas*, Sohely Sharmin*, Carina Bringedal† and Sorin Pop*

* Faculty of Sciences
Hasselt University
Diepenbeek, Belgium
e-mail: manuela.bastidas@uhasselt.be, sohely.sharmin@uhasselt.be, sorin.pop@uhasselt.be

† Institute for Modelling Hydraulic and Environmental Systems,
University of Stuttgart
Stuttgart, Germany
e-mail: carina.bringedal@iws.uni-stuttgart.de

Key words: Multi-scale methods, phase-field models, two-phase flow

Abstract: *We consider the flow of two immiscible fluid phases in a porous medium. At the scale of pores, the two fluid phases are separated by interfaces that are transported by the flow. Furthermore, the surface tension at such interfaces depends on the concentration of a surfactant dissolved in one of the fluids. Here we discuss a two-scale model for two-phase porous-media flow, in which concentration-dependent surface tension effects are incorporated. The model is obtained by employing formal homogenization methods and relies on the phase-field approach, in which thin, diffuse interface regions approximate the interfaces. We propose a two-scale numerical scheme and present numerical results revealing the influence of various quantities on the averaged behaviour of the system.*

1 INTRODUCTION

Porous media are complex domains involving many alternating solid grains surrounded by void spaces (the pores). These form hierarchically organized structures in which various processes take place at different scales. Prominent examples in this sense are the fluid flow through the pores of the medium, the transport of chemically reactive substances, or mechanical deformation. In situations like the ones mentioned here, there are processes taking place at the scale of pores (from now on called the micro scale), whereas the main interest is in the averaged behavior of the system at a larger scale (the laboratory or even the field scale, from now on called the macro scale).

Two-phase flow in porous media are encountered in several real-life situations of practical relevance. Prominent examples in this sense are geological CO2 sequestration or oil recovery. Here we consider the flow of two immiscible fluid phases in a porous medium. At the micro scale, one encounters an interface separating the two fluids transported by the flow. Furthermore, we assume that the surface tension may change depending on the concentration of a surfactant dissolved in one of the fluid phases. Since the location of the interface is not known a-priori but depends on the (unknown) fluid velocities and the surfactant concentration, the resulting mathematical model involves free boundaries at the micro scale. Hence, the model equations are defined in time-dependent a-priori unknown micro-scale domains.

Two significant challenges can be identified in this context: the free boundaries at the micro scale and the complex structure of the micro-scale domain. To deal with the former, we consider a phase-field approach, in which the evolving interfaces are approximated by narrow diffuse-interface regions, which allows defining all model components on the entire micro-scale domain. For the latter, we recall that in practical applications, the main interest is in the system's behavior at the macro scale, not necessarily in the complex, micro-scale behavior.

https://doi.org/10.4995/YIC2021.2021.12571

Therefore, we apply formal homogenization techniques to derive a two-scale phase-field model, approximating the averaged, macro-scale behavior of the system. In the resulting two-scale model, the effective (macro-scale) parameters required at the macro scale are determined by solving micro-scale cell problems, which, in their turn, depend on the macro-scale quantities.

Similar situations are considered in [1,2], where macro-scale models are derived for two-phase porous-media flow, accounting for the evolving interfaces at the micro scale. Such results are extended in [3], where dynamic and hysteretic contact angles are incorporated in the micro-scale model before deriving macro-scale ones. Closest to the present contribution is the case involving a concentration-dependent surface tension studied in [4]. However, all these results are obtained for sharp-interface micro-scale models.

Phase-field models for two-phase porous-media flow, including the derivation of macro-scale models are discussed in [5–10]. More precisely, in [5, 6] phase-field pore-scale models are discussed, and the convergence to the corresponding sharp-interface model is proved when passing the diffuse-interface parameter to zero. A macro-scale model is derived in [9] under certain scaling assumptions, but without accounting for variable surface-tension effects. A macro-scale phase-field model for compressible fluids is derived in [8]. Here we consider a two-scale phase-field model derived by formal homogenization techniques [11]. The model includes variable surface-tension effects, depending on the concentration of a surfactant dissolved in one of the fluid phases. We propose an explicit numerical scheme, accounting for the coupling between the two scales.

The paper is organized as follows. In Section 2, the two-scale model is presented, and the interaction between the scales is highlighted. Then, in Section 3, an explicit numerical scheme is proposed for the numerical solution of the two-scale model. Finally, in Section 4, a numerical example is presented, for which the necessity of using adaptive meshes at the micro scale is discussed, and the influence of the macro-scale quantities on the micro-scale results is studied.

2 THE TWO-SCALE MODEL

At the macro scale, the porous medium is a bounded domain $\Omega \subset \mathbb{R}^2$, having Lipschitz-continuous boundary $\partial\Omega$. Let $\mathrm{T} \in (0, \infty)$ be the final time. To each macro-scale point $\mathbf{x} \in \Omega$, one micro-scale cell $Y = [0, 1]^2$ is associated. The micro-scale cell is divided into two subdomains: the inner grain G surrounded by the pore space P. We denote by ∂G the boundary of G and by $\mathbf{n}$ the unit normal to ∂G pointing into G. One has $Y = G \cup P \cup \partial G$, and we assume that the pore space P is filled by two fluids, "Fluid 1" and "Fluid 2". A sketch of the two-scale domain is shown in Figure 1.

Following the phase-field approach, the (micro-scale) sharp interface separating the two fluids is replaced by a narrow, diffuse interface. Consequently, the two fluids are identified at the micro scale through the phase field ϕ, ranging from 1 (corresponding to Fluid 1) to -1 (for Fluid 2). At the micro scale, this allows defining the velocity, pressure, and solute concentration for the mixture over the entire pore space P, without separating between the fluid phases. The corresponding macro-scale quantities are $\bar{\mathbf{v}}$, $\bar{\mathbf{v}}_\phi$, p and c. Also, S stands for the macro-scale saturation of Fluid 1. We consider the following macro-scale model for $\mathbf{x} \in \Omega$ and $t \in (0, T]$

$$(\mathbf{P}_p) \qquad \begin{cases} \bar{\mathbf{v}} = -\mathcal{K}\nabla p - \mathbf{M}\gamma(c), \\ \nabla \cdot \bar{\mathbf{v}} = 0, \end{cases}$$

$$(\mathbf{P}_S) \qquad \begin{cases} \bar{\mathbf{v}}_\phi = -\mathcal{K}_\phi \, \nabla p - \mathbf{M}_\phi \, \gamma(c), \\ \Phi \, \partial_t S + \dfrac{1}{2} \nabla \cdot \bar{\mathbf{v}}_\phi = 0, \end{cases}$$

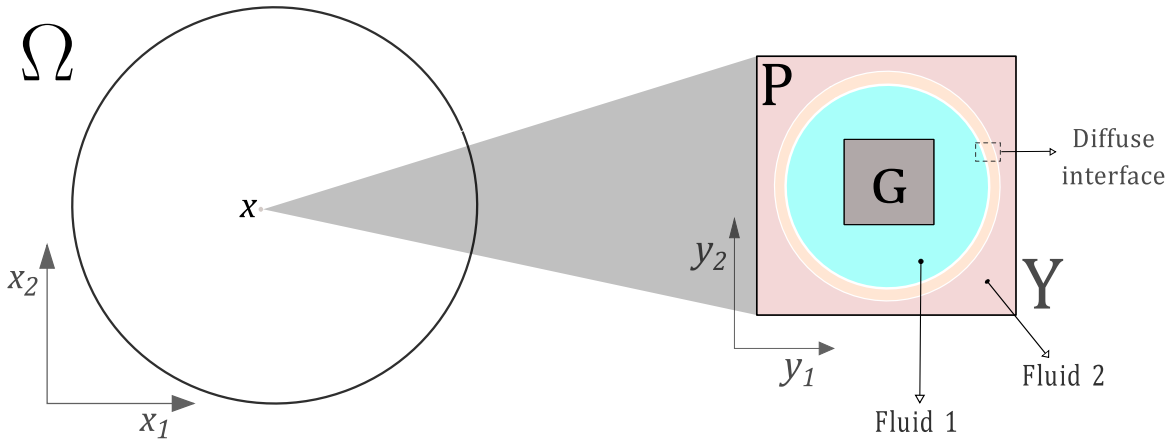


Figure 1: The porous medium: the macro-scale domain Ω (left) and the micro-scale domain Y (right) corresponding to a point $\mathbf{x} \in \Omega$.

$$(\mathbf{P}_c) \qquad \Phi\, \partial_t(Sc) + \frac{1}{2}\nabla \cdot (c(\bar{\mathbf{v}} + \bar{\mathbf{v}}_\phi)) = \frac{1}{\overline{\mathrm{Pe}_c}}\nabla \cdot (\mathcal{B}\nabla c + \mathbf{H}c).$$

Initial and boundary conditions complete the model. Here $\overline{\mathrm{Pe}_c}$ is the non-dimensional Péclet number and Φ denotes the medium porosity. The definition of the effective parameters $\mathcal{K}_\phi$, $\mathcal{K}$, $\mathbf{M}_\phi$, $\mathbf{M}$, $\mathcal{B}$, $\mathbf{H}$ depend on the micro-scale structure and on the evolution of the micro-scale phase field ϕ, as explained below. Note that $\bar{\mathbf{v}}$ is the velocity of the mixture of the two fluids, while $\bar{\mathbf{v}}_\phi$ accounts for the phase distribution. Hence $\frac{1}{2}(\bar{\mathbf{v}} + \bar{\mathbf{v}}_\phi)$ reflects the macro-scale velocity of Fluid 1.

At each micro-scale cell Y, the phase field ϕ and the potential ψ are computed by solving the following micro-scale cell problem

$$(\mathbf{P}_\phi) \qquad \begin{cases} \nabla \cdot (\mathbf{v}\phi) = \overline{\mathrm{A}_\phi}\lambda\Delta\psi, & \text{in } P, \\ \psi = \dfrac{\overline{\mathrm{A}_\psi}\gamma(c)}{\lambda}\left(\mathcal{C}P'(\phi) + I'(\phi) - \mathcal{C}\lambda^2\Delta\phi\right), & \text{in } P, \\ \nabla\phi \cdot \mathbf{n} = 0, & \text{on } \partial G, \\ \nabla\psi \cdot \mathbf{n} = 0, & \text{on } \partial G, \\ \phi,\ \psi \text{ are } Y\text{-periodic}, \\ \dfrac{1}{\Phi}\displaystyle\int_P \phi\, d\mathbf{y} = (2S - 1)\,. \end{cases}$$

Observe that t enters in $(\mathbf{P}_\phi)$ as a parameter, through the macro-scale saturation S and concentration c. Here $\overline{\mathrm{A}_\phi}, \overline{\mathrm{A}_\psi}$ are non-dimensional quantities and $\gamma(c)$ is the concentration-dependent surface tension, which introduces a coupling with the macro scale. The micro-scale velocity $\mathbf{v}$ is defined below and its average is by construction $\bar{\mathbf{v}}$. Moreover, we choose $P(\phi) = \frac{1}{4}(1-\phi^2)^2$ as the double-well type potential and $I(\phi) = \frac{1}{2}(1+\phi)$ as a characteristic function which is 1 in Fluid 1 and 0 in Fluid 2. The parameter λ is the diffuse interface thickness and $\mathcal{C} = \frac{3}{2\sqrt{2}}$ is a calibration constant.

The components of the effective matrices $\mathcal{K}$ and $\mathcal{K}_\phi$, appearing in the Darcy-type laws in $(\mathbf{P}_p)$ and in the evolution equation for the saturation $(\mathbf{P}_S)$, are found through

$$\mathcal{K}_{\mathtt{s},\mathtt{r}} = \int_P (\mathbf{w}_\mathtt{r})_\mathtt{s}\, d\mathbf{y} \quad \text{and} \quad (\mathcal{K}_\phi)_{\mathtt{s},\mathtt{r}} = \int_P (\mathbf{w}_\mathtt{r})_\mathtt{s}\phi\, d\mathbf{y}, \qquad \text{for } \mathtt{r},\mathtt{s} = 1,2. \tag{1}$$

Here $(\mathbf{w}_\mathtt{r})_\mathtt{s}$ are the components of $\mathbf{w}_\mathtt{r} = ((\mathbf{w}_\mathtt{r})_1, (\mathbf{w}_\mathtt{r})_2)^t$, where $(\mathbf{w}_\mathtt{r}, \Pi_\mathtt{r})$ solve the following

Stokes-type cell problems

$$(\mathbf{P}^{\mathtt{r}}_{\mathcal{K}}) \qquad \begin{cases} \overline{\mathrm{Eu}}\,(\mathbf{e}_{\mathtt{r}} + \nabla\Pi_{\mathtt{r}}) = -\dfrac{1}{\overline{\mathrm{Re}}}\nabla\cdot(2\mu(\phi)\boldsymbol{\varepsilon}(\mathbf{w}_{\mathtt{r}}))\,, & \text{in } P,\\ \nabla\cdot\mathbf{w}_{\mathtt{r}} = 0, & \text{in } P,\\ \mathbf{w}_{\mathtt{r}} = \mathbf{0}, & \text{on } \partial G,\\ \Pi_{\mathtt{r}}, \mathbf{w}_{\mathtt{r}} \text{ are } Y\text{-periodic} \quad \text{and} \quad \displaystyle\int_P \Pi_{\mathtt{r}}\, d\mathbf{y} = 0. \end{cases}$$

Here $\boldsymbol{\varepsilon}(\mathbf{w}_{\mathtt{r}}) = \frac{1}{2}\left((\nabla\mathbf{w}_{\mathtt{r}}) + (\nabla\mathbf{w}_{\mathtt{r}})^T\right)$ is the symmetric stress tensor and $\mathbf{e}_{\mathtt{r}}$ is the unit basis vector. The Euler and Reynolds numbers are denoted by $\overline{\mathrm{Eu}}$ and $\overline{\mathrm{Re}}$, respectively. Moreover, $\mu(\phi) = \frac{\mu^2\cdot(1+\phi)}{2} + \frac{\mu^1\cdot(1-\phi)}{2}$ is the viscosity of the mixture of the two fluids and μ^i with $i = 1, 2$ correspond to the viscosity of Fluid i. As before, t enters in $(\mathbf{P}^{\mathtt{r}}_{\mathcal{K}})$ as a parameter through ϕ.

Additionally, the components of the effective vectors $\mathbf{M}$ and $\mathbf{M}_\phi$, appearing in the Darcy-type law $(\mathbf{P}_p)$ and in the evolution equation for the saturation $(\mathbf{P}_S)$, are found through

$$\mathbf{M}_{\mathtt{s}} = \int_P (\mathbf{w}_0)_{\mathtt{s}}\, d\mathbf{y} \quad \text{and} \quad (\mathbf{M}_\phi)_{\mathtt{s}} = \int_P (\mathbf{w}_0)_{\mathtt{s}}\phi\, d\mathbf{y}, \qquad \text{for } \mathtt{r}, \mathtt{s} = 1, 2. \tag{2}$$

As before, $(\mathbf{w}_0)_{\mathtt{s}}$ are the components of $\mathbf{w}_0 = ((\mathbf{w}_0)_1, (\mathbf{w}_0)_2)^t$, where $(\mathbf{w}_0, \Pi_0)$ solve the following modified Stokes-type cell problem

$$(\mathbf{P}_{\mathbf{M}}) \qquad \begin{cases} \overline{\mathrm{Eu}}\nabla\Pi_0 = -\dfrac{1}{\overline{\mathrm{Re}}}\nabla\cdot(2\mu(\phi)\boldsymbol{\varepsilon}(\mathbf{w}_0)) + \dfrac{\mathcal{C}}{\overline{\mathrm{Re}}\;\overline{\mathrm{Ca}}}\left(\dfrac{1}{\lambda}P'(\phi) - \lambda\Delta\phi\right)\nabla\phi, & \text{in } P,\\ \nabla\cdot\mathbf{w}_0 = 0, & \text{in } P,\\ \mathbf{w}_0 = \mathbf{0}, & \text{on } \partial G,\\ \Pi_0, \mathbf{w}_0 \text{ are } Y\text{-periodic} \quad \text{and} \quad \displaystyle\int_P \Pi_0\, d\mathbf{y} = 0, \end{cases}$$

with $\overline{\mathrm{Ca}}$ being the capillary number. Observe that $(\mathbf{P}_{\mathbf{M}})$ is introduced to deal with the concentration-dependent surface tension.

The micro-scale cell velocities $\mathbf{w}_{\mathtt{r}}$ and $\mathbf{w}_0$ are also involved in the calculation of the micro-scale velocity $\mathbf{v}$, i.e.

$$\mathbf{v} = -\sum_{\mathtt{r}=1}^{2} \mathbf{w}_{\mathtt{r}}\, \partial_{x_{\mathtt{r}}} p - \mathbf{w}_0\gamma(c). \tag{3}$$

Notice that the macro-scale velocities $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}_\phi$ in $(\mathbf{P}_p)$ are related with the micro scale trough $\mathbf{v}$ and ϕ as follows

$$\bar{\mathbf{v}} = \int_P \mathbf{v}\, d\mathbf{y} \quad \text{and} \quad \bar{\mathbf{v}}_\phi = \int_P \mathbf{v}\phi\, d\mathbf{y}.$$

The components of the effective matrix $\mathcal{B}$ and the effective vector $\mathbf{H}$, appearing in the macro-scale equation for the solute concentration $(\mathbf{P}_c)$, are

$$\mathcal{B}_{\mathtt{s},\mathtt{r}} = \int_P I(\phi)\left(\delta_{\mathtt{s},\mathtt{r}} + \partial_{y_{\mathtt{s}}}\chi_{\mathtt{r}}\right)\, d\mathbf{y}, \quad \mathbf{H}_{\mathtt{s}} = \int_P I(\phi)\partial_{y_{\mathtt{s}}}\chi_0\, d\mathbf{y}, \qquad \text{for } \mathtt{r}, \mathtt{s} = 1, 2. \tag{4}$$

Here, $\chi_{\mathtt{r}}$ and χ_0 solve the following micro-scale cell problems

$$(\mathbf{P}^{\mathtt{r}}_{\mathcal{B}}) \qquad \begin{cases} \nabla\cdot[I(\phi)\,(\nabla\chi_{\mathtt{r}} + \mathbf{e}_{\mathtt{r}})] = 0, & \text{in } P,\\ I(\phi)\,(\nabla\chi_{\mathtt{r}} + \mathbf{e}_{\mathtt{r}})\cdot\mathbf{n} = 0, & \text{on } \partial G,\\ \chi_{\mathtt{r}} \text{ is } Y\text{-periodic} \quad \text{and} \quad \displaystyle\int_P \chi_{\mathtt{r}}\, d\mathbf{y} = 0. \end{cases}$$

$$(\mathbf{P_H}) \qquad \begin{cases} \nabla \cdot [I(\phi)\nabla\chi_0] = \nabla \cdot (I(\phi)\mathbf{v}), & \text{in } P, \\ I(\phi)\nabla\chi_0 \cdot \mathbf{n} = 0, & \text{on } \partial G, \\ \chi_0 \text{ is } Y\text{-periodic} \quad \text{and} \quad \displaystyle\int_P \chi_0 \, d\mathbf{y} = 0. \end{cases}$$

3 THE NUMERICAL SCHEME

We propose an explicit numerical scheme for solving the two-scale model for the two-phase flow porous-media problem presented in Section 2. With $N \in \mathbb{N}$, we let $\Delta t = \mathrm{T}/N$ be the time step size and define $t^n = n\Delta t$. The time-discrete solutions are denoted by $\phi^n := \phi(\cdot, \cdot, t^n)$ and $\nu^n := \nu(\cdot, t^n)$ where $\nu \in \{\mathcal{K}_\phi, \mathcal{K}, \mathbf{M}_\phi, \mathbf{M}, \mathcal{B}, \mathbf{H}, p, \bar{\mathbf{v}}, \bar{\mathbf{v}}_\phi, S, c\}$. For $n \geq 0$, assume S^n, c^n and ϕ^n given. The time stepping reads:

- For each $\mathbf{x} \in \Omega$, compute the solution of the time-discrete micro-scale cell problems corresponding to $(\mathbf{P}^{\mathrm{r}}_{\mathcal{K}})$ and $(\mathbf{P_M})$.
- Compute the first set of time-discrete effective parameters $\mathcal{K}^n_\phi$, $\mathcal{K}^n$, $\mathbf{M}^n_\phi$ and $\mathbf{M}^n$.
- Compute the macro-scale solution p^n and $\bar{\mathbf{v}}^n$ by solving the time-discrete macro-scale problems corresponding to $(\mathbf{P}_p)$.
- Compute the macro-scale solution $\bar{\mathbf{v}}^n_\phi$ and S^{n+1} by solving the time-discrete macro-scale problems corresponding to $(\mathbf{P}_S)$.
- For each $\mathbf{x} \in \Omega$, compute the micro-scale velocity $\mathbf{v}^n$ and the solution of the time-discrete micro-scale cell problems corresponding to $(\mathbf{P}^{\mathrm{r}}_{\mathcal{B}})$ and $(\mathbf{P_H})$.
- Compute the second set of time-discrete effective parameters $\mathcal{B}^n$ and $\mathbf{H}^n$.
- Compute the macro-scale solution c^{n+1} by solving the time discrete problem corresponding to $(\mathbf{P}_c)$.
- For each $\mathbf{x} \in \Omega$, compute the solution of the time-discrete phase-field problem corresponding to $(\mathbf{P}_\phi)$ to obtain ϕ^{n+1}.

The explicit scheme is sketched in Figure 2. We highlight that the two-scale problem itself is fully coupled, and an iterative structure could be considered here. We refer to [12,13] for similar approaches using iterations to handle the multi-scale interaction between the sub-problems.

Clearly, for the numerical simulations the explicit time stepping needs to be completed by the spatial discretization. More precisely, let $\mathfrak{T}_H$ be a triangular partition of the macro-scale domain Ω with elements T of diameter H_T and $H := \max_{T \in \mathfrak{T}_H} H_T$. For computing the micro-scale quantities, a micro-scale domain Y is assigned to each macro-scale element T. On each micro-scale domain Y we define another triangular partition $\mathfrak{T}_h$ with elements T_μ of diameter h_{T_μ} and $h := \max_{T_\mu \in \mathfrak{T}_h} h_{T_\mu}$. Finally, we use the mixed finite element method to calculate the numerical solution at both scales. For an effective computation we use adaptive mesh refinement on the micro scale (see [12, 14]).

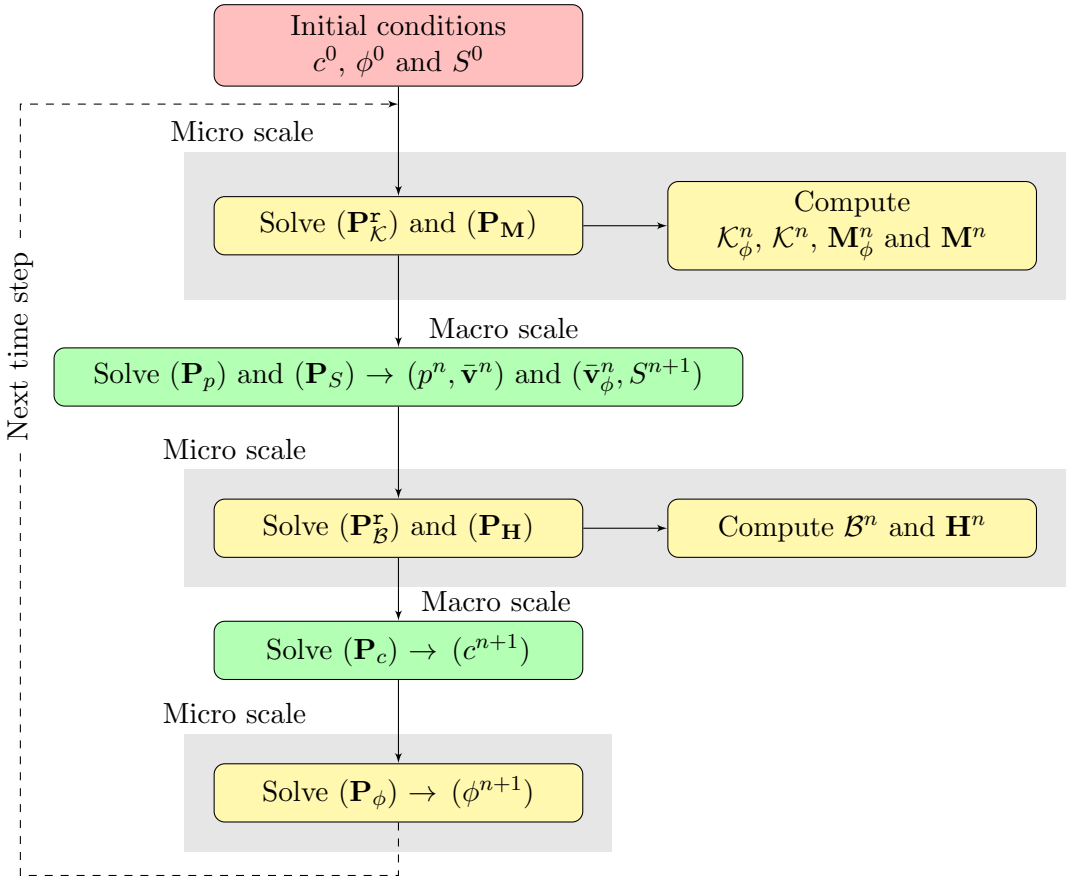


Figure 2: The explicit two-scale scheme.

4 NUMERICAL RESULTS

In this section, we present a micro-scale numerical experiment that highlights the relation between the macro-scale quantities and the micro-scale solutions. We restrict our implementations to the 2D case and all parameters remain non-dimensional. Here the micro-scale domain contains a centered square grain with side lengths 0.2 and we choose

$$\overline{\mathrm{Pe}_c} = \overline{\mathrm{Eu}} = \overline{\mathrm{Re}} = \overline{\mathrm{Ca}} = \overline{\mathrm{A}_\phi} = \overline{\mathrm{A}_\psi} = 1 \ \text{ and } \ \lambda = 0.08.$$

4.1 THE PHASE-FIELD AND THE MICRO-SCALE MESH

Figure 3 shows the initial phase field ϕ, corresponding to a saturation $S^0 = 0.639$, and the Laplacian of the initial phase field $\Delta\phi$, which is needed for computing the potential ψ in $(\mathbf{P}_\phi)$. The Laplacian is calculated numerically, and this calculation requires the construction of a very fine mesh around the transition zone to achieve sufficient accuracy. Close to the diffuse interface, the resolution of the micro-scale mesh $\mathfrak{T}_h$ is taken $h \ll \lambda$ to capture the diffuse transition zone and the variation in its derivatives. Following the ideas in [12], we refine the micro-scale mesh only close to the diffuse transition zone, making the computation of the phase field and the effective parameters accurate and efficient.

In Figure 3 we use an initially uniform mesh with 800 elements. Then, the mesh is refined around the transition zone such that the length of the smallest edge in the mesh is $\min_{T_\mu \in \mathfrak{T}_h} h_{T_\mu} = 1.25\text{E-}2 < \lambda$ and the length of the largest edge (located far from the transition zone) is $\max_{T_\mu \in \mathfrak{T}_h} h_{T_\mu} = 7.071\text{E-}2$.

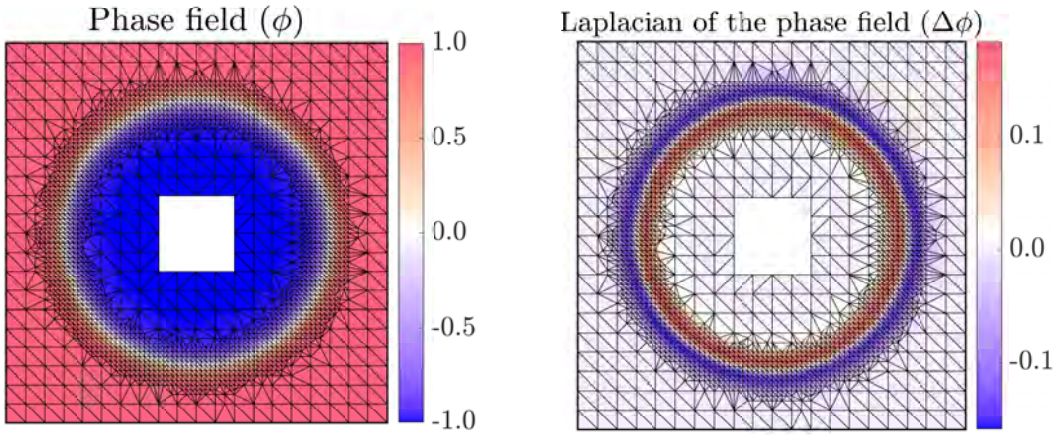


Figure 3: Phase-field initial condition (left) and the numerical calculation of the Laplacian of the phase field (right).

4.2 FIRST SET OF MICRO-SCALE PROBLEMS

Given the phase-field initial condition in Figure 3, we solve the micro-scale problems $(\mathbf{P}^{\mathbf{r}}_{\mathcal{K}})$ and $(\mathbf{P}_{\mathbf{M}})$ over the refined mesh. Figure 4 shows the scalar solutions Π_1, Π_2 and Π_0 of the problems $(\mathbf{P}^{\mathbf{r}}_{\mathcal{K}})$ and $(\mathbf{P}_{\mathbf{M}})$ in the simple case when the two fluids have the same viscosity, i.e. $\mu^1 = 1$ and $\mu^2 = 1$.

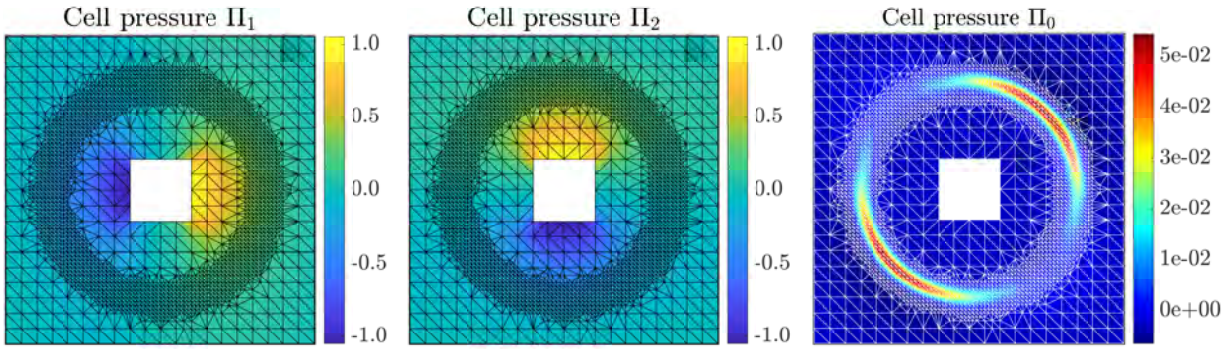


Figure 4: The solution of the first set of micro-scale problems $(\mathbf{P}^{\mathbf{r}}_{\mathcal{K}})$ (left and middle) and $(\mathbf{P}_{\mathbf{M}})$ (right).

Notice that for Π_0, the location of the changes in the solution coincides with the phase-field transition zone. This supports the requirement of a mesh refinement strategy to improve the accuracy and efficiency of further computations.

4.3 THE EFFECTIVE PARAMETERS

We show below the behavior of the effective parameters $\mathcal{K}_\phi$, $\mathcal{K}$, $\mathbf{M}_\phi$ and $\mathbf{M}$, depending on the saturation. Figure 5 displays the results for the effective tensors $\mathcal{K}_\phi$ and $\mathcal{K}$. We consider two cases: a simple case where the two fluids have same viscosity, i.e. $\mu^1 = \mu^2 = 1$, and a more complex case where the viscosities are $\mu^1 = 0.1$ and $\mu^2 = 1$.

The symmetry of the phase field at the micro scale implies that the effective tensors are isotropic. The non-diagonal components of $\mathcal{K}_\phi$ and $\mathcal{K}$ can be neglected, and in Figure 5 we only show the first component of the effective tensors.

Notice that when $\mu^1 = \mu^2 = 1$, the changes on the saturation do not affect the permeability $\mathcal{K}$. This is expected since the two fluids flow like one. In contrast, Figure 5 reflects that the changes in the saturation have an important effect if the two fluids have different viscosities.

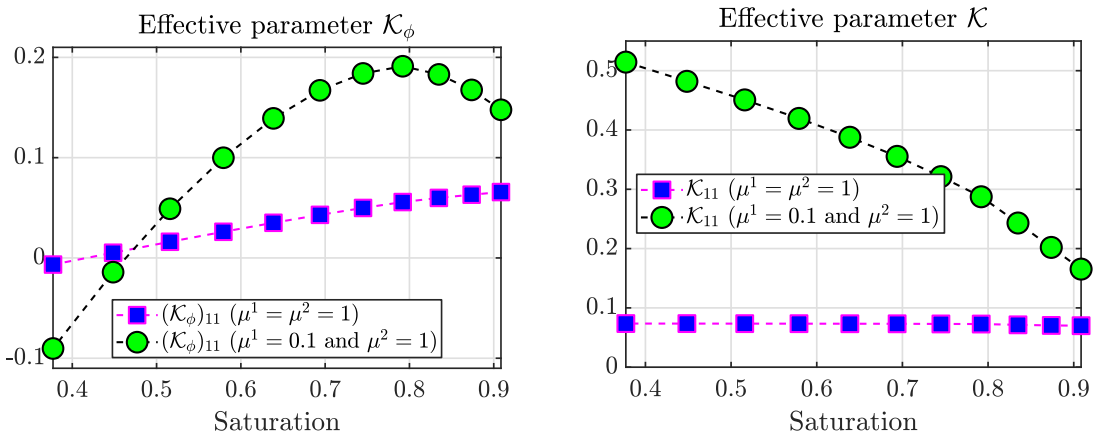


Figure 5: Dependence of the effective parameters $\mathcal{K}_\phi$ and $\mathcal{K}$ on the macro-scale saturation.

Commonly used two-phase porous-media flow models are relying on saturation-dependent quantities like relative permeability and capillary pressure. The situation here is similar, but capillary pressure is absent due to the assumed scaling of the capillary number [11]. Moreover, here $\mathcal{K}$ is not separated into absolute and relative permeability, and it reflects how the velocity of the mixture of the two fluids relate to the pressure gradient.

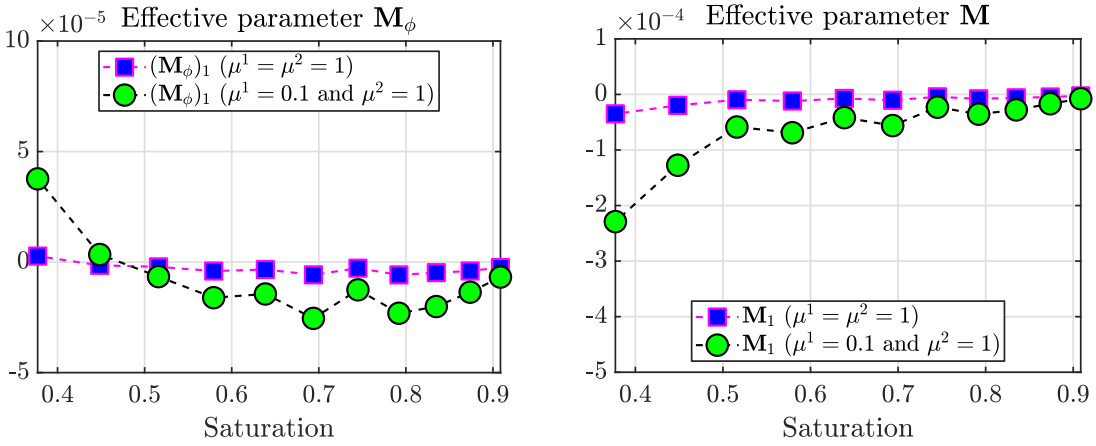


Figure 6: Dependence of the effective parameters $\mathbf{M}_\phi$ and $\mathbf{M}$ on the macro-scale saturation.

In Figure 6 we denote by $(\mathbf{M}_\phi)_1$ and $\mathbf{M}_1$ the first component of the effective vectors $\mathbf{M}_\phi$ and $\mathbf{M}$. Notice that for these macro-scale vectors, both components are equal due to the symmetry of the phase field. Moreover, Figure 6 shows that the variations in the effective parameters $\mathbf{M}_\phi$ and $\mathbf{M}$ are more relevant in the case of a large viscosity ratio.

5 SUMMARY AND OUTLOOK

We have considered a two-scale model for two-phase flow in a porous medium. The model describes the behavior of the mixture of two fluids and a surfactant dissolved into one of them. Here, the surface tension depends on the concentration of the solute. This model is the homogenized counterpart of a pore-scale phase-field model. For these phase-field formulations, a diffuse region approximates the moving interfaces separating the two fluids.

Based on the Euler explicit time stepping and the lowest order mixed finite element spatial discretization, we have proposed a two-scale numerical scheme. The scheme requires solving several micro-scale cell problems for each macro-scale point, depending on the macro-scale concentration and saturation. The solution of these micro-scale cell problems is used to deter-

mine the macro-scale parameters needed to compute the macro-scale model unknowns (velocity, pressure, saturation and concentration). For each micro-scale cell problem, the spatial mesh is refined or coarsened adaptively, improving the efficiency of the scheme.

We have presented numerical simulation results in two different situations, when the fluids have the same viscosities or when the viscosity ratio is large. Based on these results, we show the dependence of the macro-scale parameters depending on the saturation.

In the following research steps, we will analyze the possibility to compute the macro-scale parameters adaptively based on an active-passive node strategy. Furthermore, implicit or semi-implicit schemes will be considered, coupled with appropriate linearization approaches. Also, different regimes will be analyzed, possibly leading to models involving a capillary pressure.

ACKNOWLEDGEMENTS

This research is supported by the Research Foundation - Flanders (FWO) through the Odysseus programme (Project G0G1316N) and by the German Research Foundation (DFG) through the SFB 1313, Project Number 327154368.

REFERENCES

[1] A. Mikelic and L. Paoli, "On the derivation of the Buckley-Leverett model from the two fluid Navier-Stokes equations in a thin domain," *Computational Geosciences*, vol. 4, no. 1, pp. 99–101, 2000.

[2] D. Picchi and I. Battiato, "The impact of pore-scale flow regimes on upscaling of immiscible two-phase flow in porous media," *Water resources research*, vol. 54, no. 9, pp. 6683–6707, 2018.

[3] S. B. Lunowa, C. Bringedal, and I. S. Pop, "On an averaged model for immiscible two-phase flow with surface tension and dynamic contact angle in a thin strip," *Studies in Applied Mathematics*, vol. 1, p. 43, 2021.

[4] S. Sharmin, C. Bringedal, and I. S. Pop, "On upscaling pore-scale models for two-phase flow with evolving interfaces," *Advances in Water Resources*, vol. 142, p. 103646, 2020.

[5] H. Abels, H. Garcke, and G. Grün, "Thermodynamically consistent, frame indifferent diffuse interface models for incompressible two-phase flows with different densities," *Mathematical Models and Methods in Applied Sciences*, vol. 22, no. 03, pp. 1 150 013, 40, 2012.

[6] H. Garcke, K. F. Lam, and B. Stinner, "Diffuse interface modelling of soluble surfactants in two-phase flow," *Communications in Mathematical Sciences*, vol. 12, no. 8, pp. 1475–1522, 2014.

[7] O. R. A. Dunbar, K. F. Lam, and B. Stinner, "Phase field modelling of surfactants in multi-phase flow," *Interfaces and Free Boundaries*, vol. 21, no. 4, pp. 495–547, 2019.

[8] C. Rohde and L. von Wolff, "Homogenization of Nonlocal Navier–Stokes–Korteweg Equations for Compressible Liquid-Vapor Flow in Porous Media," *SIAM Journal on Mathematical Analysis*, vol. 52, no. 6, pp. 6155–6179, 2020.

[9] S. Metzger and P. Knabner, "Homogenization of two-phase flow in porous media from pore to Darcy scale: a phase-field approach," *Multiscale Modeling & Simulation. A SIAM Interdisciplinary Journal*, vol. 19, no. 1, pp. 320–343, 2021.

[10] Ľ. Baňas and H. S. Mahato, "Homogenization of evolutionary Stokes-Cahn-Hilliard equations for two-phase porous media flow," *Asymptotic Analysis*, vol. 105, no. 1-2, pp. 77–95, 2017.

[11] S. Sharmin, M. Bastidas, C. Bringedal, and I. S. Pop, "Upscaling of a Navier-Stokes-Cahn-Hilliard model for two-phase porous-media flow with solute-dependent surface-tension effects," in preparation.

[12] M. Bastidas, C. Bringedal, and I. S. Pop, "A two-scale iterative scheme for a phase-field model for precipitation and dissolution in porous media," *Applied Mathematics and Computation*, vol. 396, p. 125933, 2021.

[13] M. K. Brun, T. Wick, I. Berre, J. M. Nordbotten, and F. A. Radu, "An iterative staggered scheme for phase field brittle fracture propagation with stabilizing parameters," *Computer Methods in Applied Mechanics and Engineering*, vol. 361, p. 112752, 2020.

[14] M. Bastidas, C. Bringedal, I. S. Pop, and F. A. Radu, "Numerical homogenization of nonlinear parabolic problems on adaptive meshes," *Journal of Computational Physics*, vol. 425, p. 109903, 2020.

CHALLENGES IN SEA ICE MODELING, HIGH-RESOLUTION SIMULATION AND VALIDATION

Sea ice strength development from freezing to melting in the Antarctic marginal ice zone

F. Paul*, T. Mielke†, R. Audh# and D. C. Lupascu†

* Institute for Materials Science and Center for Nanointegration Duisburg-Essen (CENIDE)
University of Duisburg-Essen
Essen, Germany
e-mail: felix.paul@uni-due.de

Marine Research Institute
University of Cape Town
Cape Town, South Africa
e-mail: ADHRIE001@myuct.ac.za

† Institute for Materials Science and Center for Nanointegration Duisburg-Essen (CENIDE)
University of Duisburg-Essen
Essen, Germany
e-mail: doru.lupascu@uni-due.de, tommy.mielke@uni-due.de

Key words: Sea ice strength, compressive strength, Antarctic marginal ice zone

Abstract: *Sea ice growth in the Marginal Ice Zone of the Antarctic is one of the largest annual changes on earth with a huge impact on the global climate and ecology system [1]. The principles of sea ice growth and melting in the MIZ of the Antarctic are not yet as well researched as their polar counterparts in the north [2]. For this study, pancake ice, consolidated ice and floe ice were analyzed with a compression test in July, October and November 2019 in the marginal ice zone of the Antarctic. Newly formed pancake ice in July showed the highest compressive strength in the bottom layer (3 MPa), whereas consolidated ice was strongest at the top (5 MPa). Consolidated ice in October and November had the highest compressive strength in a middle layer with up to 13.5 MPa, the maximum strength at the top was 3 MPa. Floe ice, consisting of destroyed pack ice, did not show a clear strength development over sea ice depth.*

1 INTRODUCTION

Sea ice growth in the Marginal Ice Zone of the Antarctic is one of the largest annual changes on earth with a huge impact on the global climate and ecology system [1]. The principles of sea ice growth and melting in the MIZ of the Antarctic are not yet well researched. The annual freezing-thawing cycle can be divided into two parts. The first part is the pancake ice cycle, which describes the sea ice growth process in four steps [3]. The melting process is the second part and is dominated by the ice-ocean albedo feedback [4]. Both processes combined, as shown in Figure 1, can explain the full annual growth and melt process in the MIZ of the Antarctic.

https://doi.org/10.4995/YIC2021.2021.12249

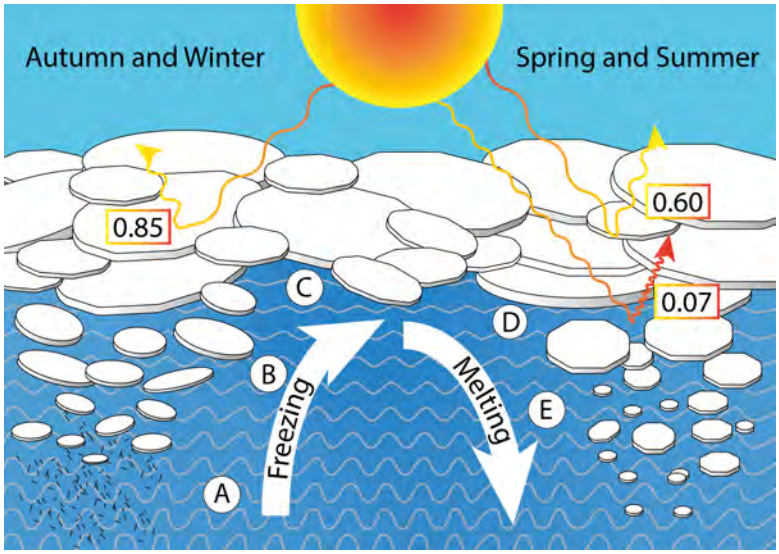


Figure 1: Full-year pancake cycle in the MIZ of the Antarctic. A) Freezing starts with the formation of frazil ice, which develops into grease ice. B) Grease ice grows thicker and starts to form ice floes named pancake ice due to their round appearance. Waves lead to pancake rafting and get attenuated by the ice cover. C) Closed ice cover overcast by snow with an albedo of 0.85. D) The ocean is still completely covered by ice. Solar radiation leads to the melting of the ice cover from the top (Albedo decreases to 0.60) and from the bottom. The bottom starts melting due to the ice-ocean albedo effect. E) Waves break up the ice and release ice floes.

Two prerequisites must be fulfilled for the beginning of sea ice growth: The water must be supercooled *and* under turbulent conditions. If both prerequisites are met, the growth of frazil ice takes place. Frazil ice, which appears as grease ice, a grey milky layer at the surface, is the first step in the annual freezing process. If a sufficient number of frazil ice crystals has formed, the crystals stick together and form flocs of ice. Flocs of frazil ice develop into larger agglomerations of crystals, forming first pans of ice. The size of pancakes varies between a few centimeters for the first pans up to 5 m in diameter for fully grown so called pancakes [5]. Pancakes get rafted by ocean waves and wind, starting to form larger ice floes by pancakes freezing together. The bonding process between the pancakes has not been observed in the laboratory or field yet and is therefore referred to as a welding mechanism [6]. A growing sea ice layer at the ocean surface attenuates the ocean waves, leading to a calmer ocean [7]. When the ocean waves are sufficiently damped, the ice cover freezes up completely. Snow at the sea ice surface increases the albedo of the ice cover to 0.85, preventing the ice and the ocean from absorbing energy from solar radiation. In contrast to snow covered ice, seawater has an albedo of only 0.07. As solar radiation gets stronger, the ocean absorbs most of the solar energy, which increases the water temperature. An increasing water temperature melts the ice from the bottom side, while the melting snow at the top decreases the albedo to 0.60 and lower [8]. As the ice gets weaker due to the melting from the bottom side, the ice breaks up and floes form, which then drift freely in the ocean.

This study will focus on the strength development of sea ice in the full-year pancake cycle, which has an effect on the formation, durability and break-up process of sea ice. This study presents the whole year cycle of freezing and thawing in 2019, enabling a direct comparison between the steps in the full-year pancake cycle.

Up to now, only a few tests have been conducted on sea ice in the Antarctic region. The maximum uniaxial compressive strength reported was 4.5 MPa in the melt season and the mean compressive strength was 2.35 MPa. In this case, the compressive strength was tested

immediately after sampling and showed the strongest layer in the center part of the ice floe [9]. Ice collected by Urabe and Inoue showed the same behavior, even though the samples were tested after long time storage in a cold room. The strongest layer was again in the center part of the floe with a maximum uniaxial compressive strength of about 2.5 MPa [10]. Only a few tests were conducted by Vaudrey with a maximum reported value of 9 MPa [11].

This is the first time, that the uniaxial compressive strength is determined in July, October and November of the same year with the same in situ testing equipment.

2 METHODS

The data provided in this study were collected during the SCALE Winter Cruise and SCALE Spring Cruise in 2019. Locations and ice concentrations for the different stations are displayed in Figure 9. The uniaxial compressive strength was determined with a hand stroke uniaxial compression test (GCTS PLT-2W Point Load Testing Device, GCTS Testing Systems, USA). Cores with a diameter of 9 cm and varying lengths were collected. These cores got cut into several cylindrical samples with a height of 13.5 cm. Even though the perfect relation of diameter to height is 1:2.5 it was decided to not reduce the diameter of the samples to avoid changes in the ice structure and proceed with the test as fast as possible after collection. This study kept the same strain rate for all samples in the ductile-to-brittle transition zone (10^{-3} 1/s) to focus on the sea ice strength development over depth.

3 RESULTS

The results are separated into the five stages of the full-year annual pancake cycle. Frazil ice is tested regarding its rheological properties, pancake ice, consolidated ice, and ice floes are tested using the uniaxial compression test device.

3.1 Frazil ice

Grease ice, which consists of lose frazil crystals and small flocs, is the first ice that forms in the freezing process in the MIZ of Antarctica. Frazil ice crystals grow under turbulent and supercooled conditions. The viscosity of grease ice was determined with a rheometer and showed a shear thinning behavior. A higher frazil ice concentration leads to a higher viscosity, which also indicates, that the ocean gets damped by a thicker frazil ice cover. Due to the completely different experimental approach the results for this set of experiments will be published elsewhere.

3.2 Pancake ice

Pancake ice develops from frazil ice. Pancake ice tested in July 2019 had a medium thickness of 0.36 m. The compressive strength increased from top to bottom, this is displayed in Figure 2. The minimum compressive strength for the pancake ice was 1.5 MPa and the maximum compressive strength was 3.1 MPa.

3.3 Pack ice (freezing period)

Three pack ice cores were cut into eight samples and tested in July 2019. The compressive strength was higher for the consolidated ice than for the pancake ice and showed a different profile over the depth (Figure 2). A relatively high compressive strength could be spotted close to the top, followed by a region with a lower compressive strength. The results for pancake ice and pack ice in the freezing period will be published elsewhere.

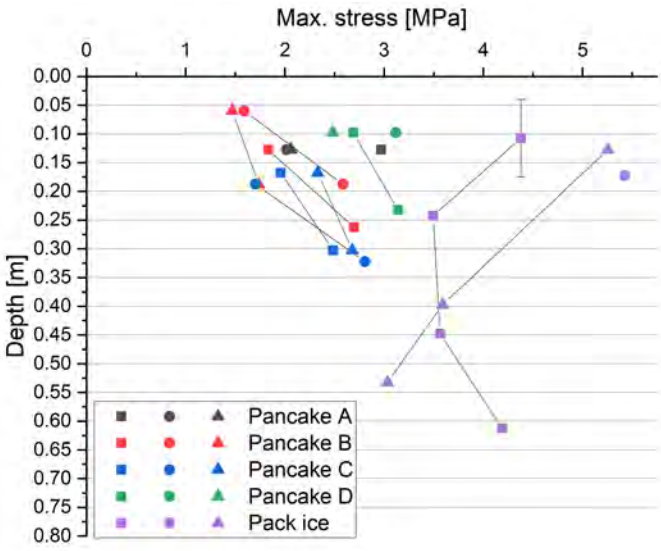


Figure 2: Sea ice strength in July 2019. Symbols connected with a line represent one core. The error bar indicates the length of a sample and is valid for all samples. Further results will be published elsewhere.

3.4 Pack ice (spring)

In total 44 pack ice samples were collected and tested at four different days for the compressive strength during the SCALE spring cruise 2019. The peak load for every sample from the pack ice stations MIZ2, MIZ3, MIZ6 and MIZ7 are displayed in Figure 3. MIZ2, MIZ6 and MIZ7 show an increase of the compressive strength over depth. MIZ3 shows a high compressive strength in a middle layer with a drop in strength beneath.

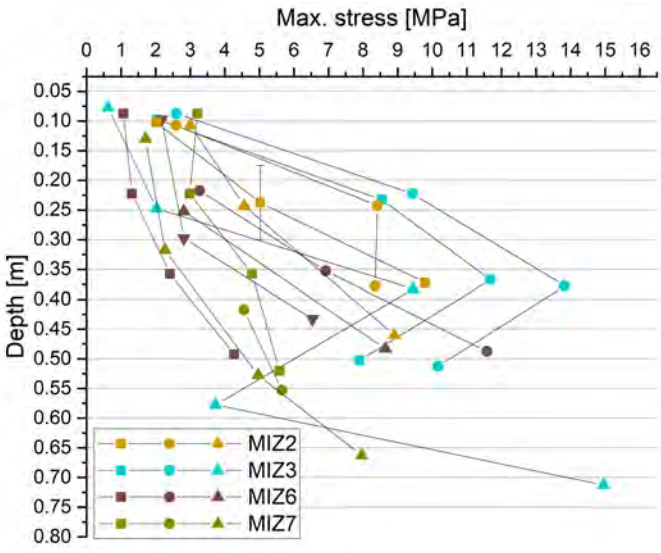


Figure 3: Maximum compressive load in October 2019 for MIZ2 (24.10.2019), MIZ3 (25.10.2019), MIZ6 (29.10.2019) and MIZ7 (30.10.2019). Symbols connected with a line represent one core. The failure occurs somewhere within the sample length of 13.5 cm without knowing the exact location. Therefor the error bar indicates the length of a sample and is valid for all samples shown in the figure.

The temperature gradient over the ice depth is shown in Figure 5. MIZ2 and MIZ3 have a slightly lower temperature at the top than MIZ6 and MIZ7. The bottom temperature is the

same for all stations. Salinity is displayed in Figure 5, showing that the highest salinity is at the top and decreases to a depth of about 30 cm. Below 30 cm all salinity profiles show scattering in the data points with no clear trend. The images of the samples after testing displayed in Figure 4 are typical for the sea ice collected at the stations MIZ2, MIZ3, and MIZ6.

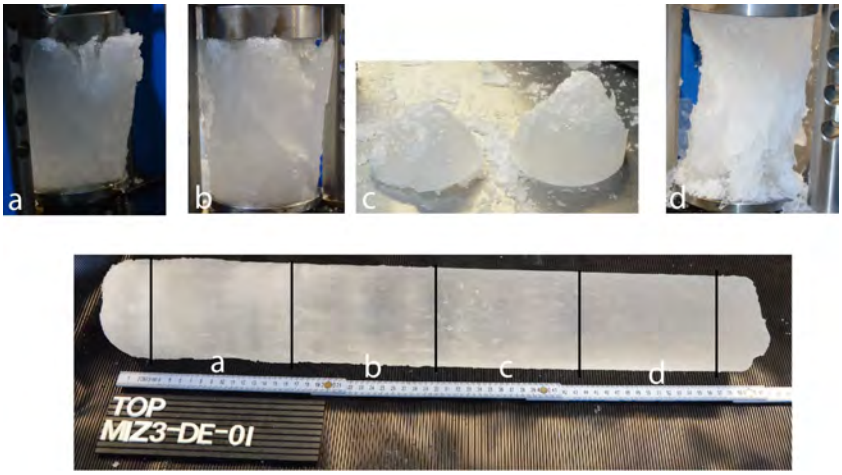


Figure 4: Bottom image of a core from station MIZ3 before preparing the samples. Top images: samples after the compression tests.

Different types of failure were obtained at stations MIZ2, MIZ3 and MIZ6. Sample MIZ3-DE-01-c (Figure 4) as well as sample MIZ3-DE-01-d (Figure 4) show a shear faulting failure. Whereas samples MIZ3-DE-01-a and MIZ3-DE-01-b (Figure 4) did not show terminal failure.

Shear faulting behavior results from confinement across the column [12]. Confinement is induced into the sample by a) the l/d ratio of 1.5, which is lower than the optimal l/d ratio, and b) the rough compression plates, preventing the sample to release stress laterally. The failure mechanism for shear faulting behavior is described in literature as follows: First parent cracks with an angle of 45° to the loading direction are induced into the sample, developing into wing cracks under increasing load. The growth of these cracks is trans-granular and followed by comb cracks. Comb cracks are unique for confined compression tests. They have one fixed end and a free end. If comb cracks are loaded by frictional drag across their free ends they fail. By this failure the load is shed further and starts a chain direction [13, 14]. Wing and comb cracks lead to a measured crack angle of 45° to 65° for all samples which showed a shear faulting behavior in the experiments presented in this study. The samples, which showed a shear faulting behavior during the compressive test, also exhibit the highest compressive strength. Cracks at an angle of 45° to the axis of maximum loading are also visible in the samples MIZ3-DE-01-a and MIZ3-DE-01-b (Figure 4), even though they did not show brittle final destruction of the sample. The samples did not fall apart even after the test, which suggests that the samples show a partwise brittle and ductile behavior. The appearance of 45° cracks to the direction of maximum loading, shows that the ice did not fail in a pure ductile way. Whereas the missing comb cracks, which would lead to an ultimate failure, suggests that the structure within the grain boundaries is not strong enough to allow secondary comb cracks to grow. This might be due to strain relaxation through creep, so that the stress cannot exceed the yield stress [11].

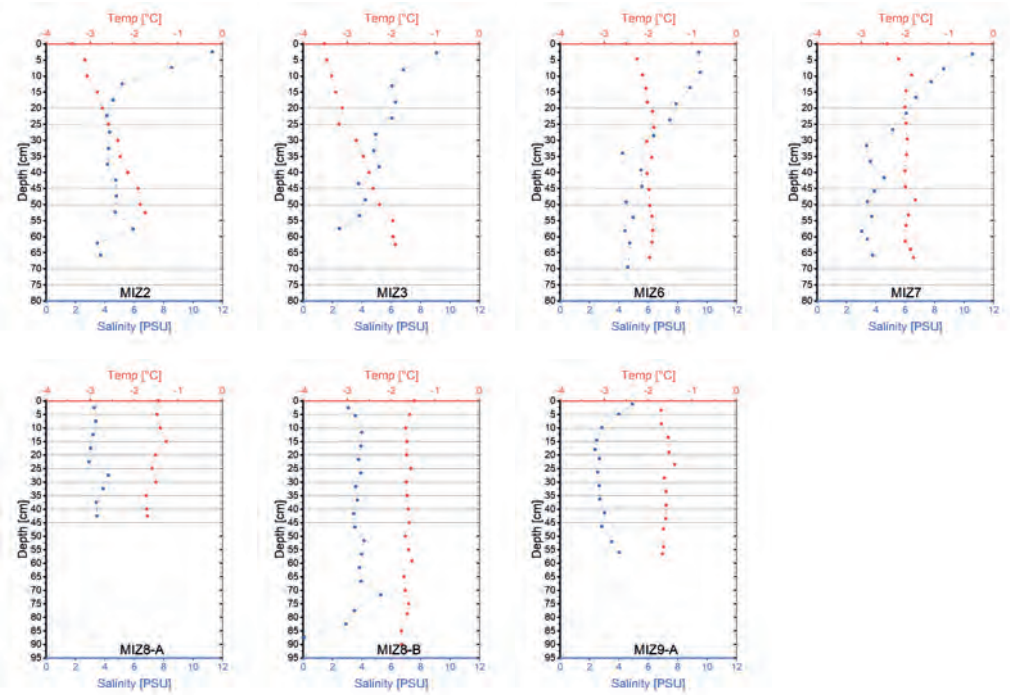


Figure 5: Temperature and salinity profiles for all stations. The first row shows the profiles for the pack ice stations, the second row the profiles of the floe stations. Salinity and temperature measurements were provided by Riesna Audh.

3.5 Ice floes

To reveal more information about the broken ice, eight cores from three different ice floes were tested. The maximum compressive strength is displayed in Figure 6 and shows that the compressive strength is only slightly increasing over depth. The temperature is constant over the ice core length and the salinity varies in a narrow range between 3 and 5 PSU.

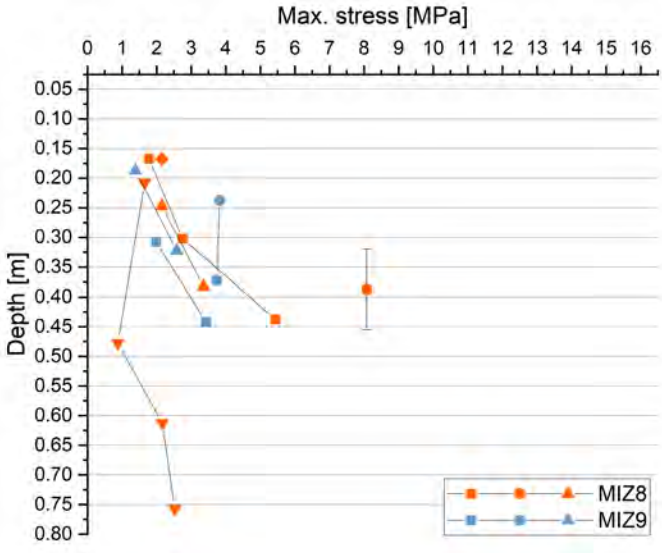


Figure 6: Maximum compressive load for ice floes in November 2019 at MIZ8 (01.11.2019) and MIZ9 (03.11.2019). Symbols connected with a line represent one core. The failure occurs somewhere within the sample length of 13.5 cm without knowing the exact location. Therefor the error bar indicates the length of a sample and is valid for all samples shown in the figure.

Figure 7 shows, at the bottom, an image of an uncut core from an ice floe and, at the top, images after the compression test had been conducted. Figure 7 was chosen, because it is representative for all floe stations from the spring cruise. It can be seen that the entire core is perforated with holes. The holes are distributed over the whole core length, they are visible at the top and bottom of the core. From the holes it can be concluded, that melting takes place from the bottom due to the warmer ocean as well as from the top, due to solar radiation. The top left-hand side sample (Figure 7 a)) deforms comparable to the top samples from Figure 4, whereas the bottom sample shows a different behavior than samples from previous stations. It looks like the top right hand side sample splits into several pieces of ice, along the brine channels. Suggesting, that the weakest bonding during the melting period is between the brine channels.

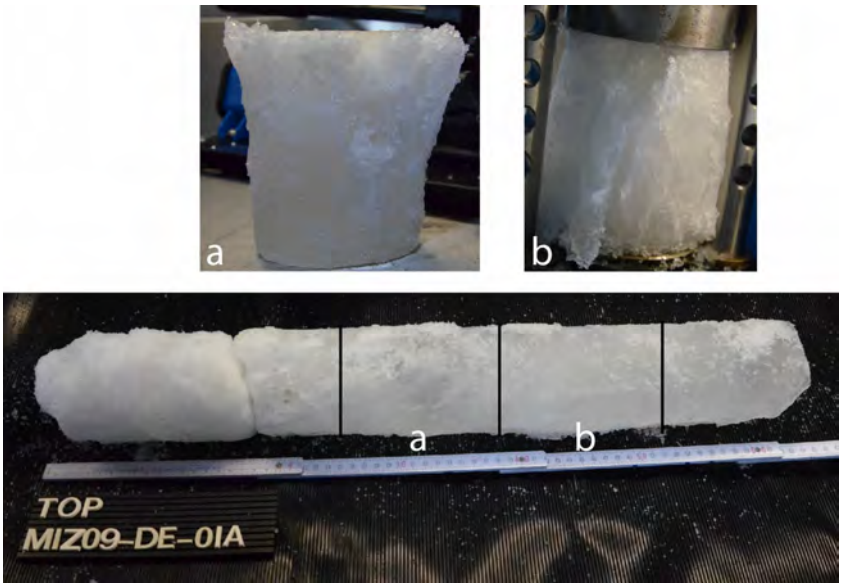


Figure 7: Bottom image of a core from station MIZ9 before preparing the samples. Top image: Samples after the compression test.

4 DISCUSSION

The compression strength of pancake ice is characterized by a low compressive strength at the top followed by an increase of strength over depth. The transition from pancake to pack ice is marked by a higher overall compressive strength. Furthermore, the compressive strength for the pack ice during freezing does not increase monotonously over depth but has a peak for the sample taken from the (Figure 2). This change in sea ice strength for the top sample from 3 MPa and lower for pancake ice to over 4 MPa for the pack ice can be explained by a temperature difference of about 6 °C. A temperature difference in this order of magnitude can lead to an increase of compressive strength by about 1.8 MPa [15].

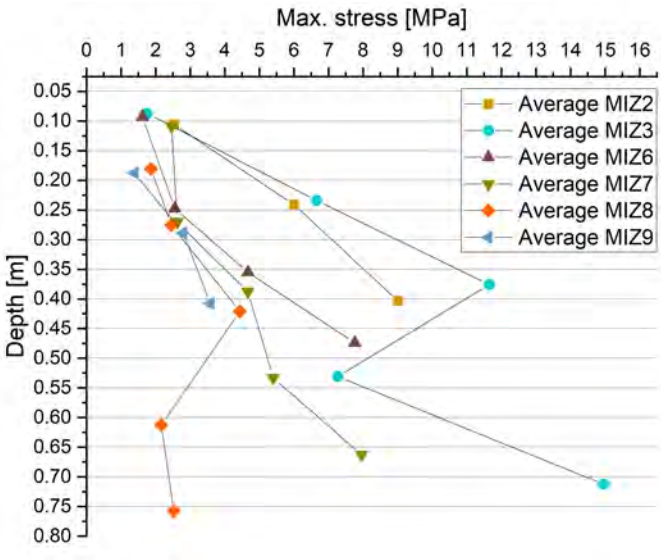


Figure 8: Average compressive strength at each station.

The next step in the full-year pancake cycle is the older pack ice during or shortly before melting. The fact that the ice at MIZ3 has the highest uniaxial compressive strength and the lowest temperature suggests, that it is the least melted station. Station MIZ2 shows a lower compressive strength and a slightly higher temperature compared to MIZ3. At the top samples it is not clear which station has a higher compressive strength in the layers underneath. But the samples in a depth between 20 to 25 cm show a smaller increase in compressive strength for MIZ2 than for MIZ3. The average compressive strength increases for samples from a depth between 10 and 15 cm to samples from a depth between 20 to 25 cm is 3.5 MPa for MIZ2 and 4.9 MPa for MIZ3. Cores collected at MIZ6 and MIZ7 do not show a strong increase from the first to the second sample, but still show an increase for the third sample. Even though the temperature does not differ in a depth of 40 cm and deeper, the compressive strength is further decreasing over depth (Figure 8). The last two ice floe stations MIZ8 and MIZ9 show a similar compressive strength as the ice from MIZ7. MIZ8 differs from MIZ7 by a weaker compressive strength below a depth of 45 cm. The ice floe at MIZ9 was already too short to test below 45 cm but shows the same compressive strength in the top part as MIZ6, MIZ7 and MIZ8. This leads to the assumption, that MIZ6 and MIZ7 are on the verge of breaking apart, a precise time or breaking mechanism cannot be pointed out in this study. It is suggested, that storms or waves will lead to the final ice break up [16, 17].

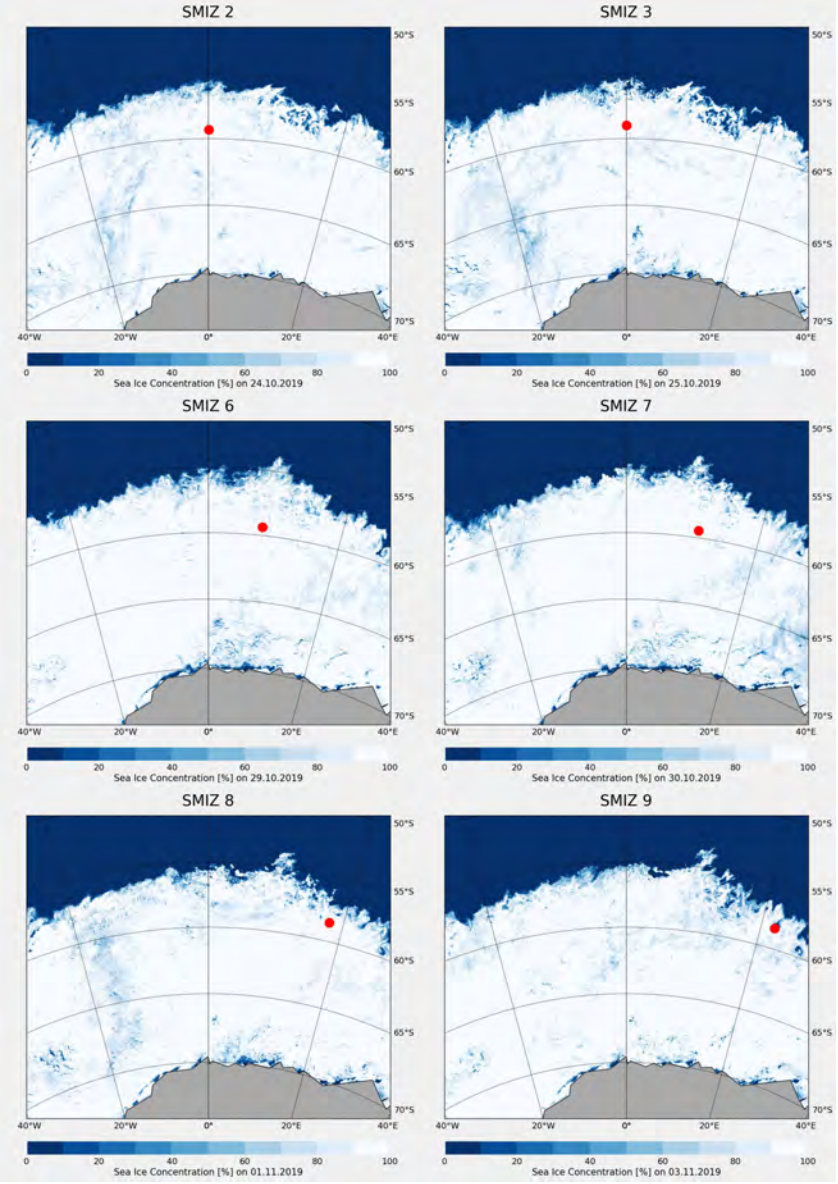


Figure 9: Sea ice concentration at the different stations. The ships position is marked with a red dot.

5 CONCLUSIONS

Data for the compressive strength of sea ice during the freezing and melting cycle of the Marginal Ice Zone of the Antarctic were presented. Samples from pancake ice, pack ice during freezing, pack ice during melting and ice floes were collected.

- The compressive strength of pancake ice increases over the sea ice depth.
- In contrast to the compressive strength in pancake ice, pack ice in the freezing period has a strong compressive strength at the top, followed by a weaker compressive strength underneath.
- Pack ice in the melting period has a comparable compressive strength for the top 15 cm, independent of the melting progress in the sea ice underneath. A maximum in compressive strength was measured in a depth between 35 to 40 cm.

- During the melting process, the compressive strength decreases until it reaches a constant strength over the sea ice depth. If the compressive strength is sufficiently decreased, the pack ice breaks apart and forms ice floes.
- The compressive strength in pack ice, which is estimated to break apart soon, and that of ice floes can be similar.

6 ACKNOWLEDGEMENTS

The SCALE cruises are funded by the South African National Research Foundation (NRF) through the South African National Antarctic Programme (SANAP), with contributions from the Department of Science and Innovation and the Department of Environmental Affairs. We are very grateful to the teams that have contributed to the success of the SCALE cruise in particular under the guidance of Marcello Vichi and Jörg Schröder.

REFERENCES

[1] K. Mezgec, B. Stenni, X. Crosta, V. Masson-Delmotte, C. Baroni, M. Braida, V. Ciardini, E. Colizza, R. Melis, M. C. Salvatore, M. Severi, C. Scarchilli, R. Traversi, R. Udisti, and M. Frezzotti, "Holocene sea ice variability driven by wind and polynya efficiency in the Ross Sea," *Nature Communications*, vol. 8, no. 1, p. 1334, Dec. 2017.

[2] G. W. Timco and W. F. Weeks, "A review of the engineering properties of sea ice," *Cold Regions Science and Technology*, vol. 60, no. 2, pp. 107–129, Feb. 2010.

[3] M. A. Lange, S. F. Ackley, P. Wadhams, G. S. Dieckmann, and H. Eicken, "Development of Sea Ice in the Weddell Sea," *Annals of Glaciology*, vol. 12, pp. 92–96, 1989.

[4] S. Nihashi and D. J. Cavalieri, "Observational evidence of a hemispheric-wide ice–ocean albedo feedback effect on Antarctic sea-ice decay," *Journal of Geophysical Research*, vol. 111, no. C12, p. C12001, Dec. 2006.

[5] M. J. Doble, "Pancake ice formation in the Weddell Sea," *Journal of Geophysical Research*, vol. 108, no. C7, p. 3209, 2003.

[6] L. A. Roach, M. M. Smith, and S. M. Dean, "Quantifying Growth of Pancake Sea Ice Floes Using Images From Drifting Buoys," *Journal of Geophysical Research: Oceans*, vol. 123, no. 4, pp. 2851–2866, Apr. 2018.

[7] B. R. Sutherland and N. J. Balmforth, "Damping of surface waves by floating particles," *Physical Review Fluids*, vol. 4, no. 1, p. 014804, Jan. 2019.

[8] D. K. Perovich and C. Polashenski, "Albedo evolution of seasonal Arctic sea ice: ALEDO EVOLUTION OF SEASONAL SEA ICE," *Geophysical Research Letters*, vol. 39, no. 8, pp. n/a–n/a, Apr. 2012.

[9] S. Kivimaa and P. Kosloff, "Compressive strength and structure of sea ice in the Weddell Sea, Antarctica," *Transactions on the Built Environment*, vol. 5, pp. 331–342, 1994.

[10] N. Urabe and M. Inoue, "Mechanical Properties of Antarctic Sea Ice," *Journal of Offshore Mechanics and Arctic Engineering*, vol. 110, no. 4, pp. 403–408, Nov. 1988.

[11] K. D. Vaudrey, "Ice engineering - Study of related properties of floating sea-ice sheets and summary of elastic and viscoelastic analyses," CIVIL ENGINEERING LABORATORY, Port Huencme, California, Tech. Rep. ADA051184, 1977.

[12] L. M. Wachter, C. E. Renshaw, and E. M. Schulson, "Transition in brittle failure mode in ice under low confinement," *Acta Materialia*, vol. 57, no. 2, pp. 345–355, Jan. 2009.

[13] E. M. Schulson, "Brittle failure of ice," *Engineering Fracture Mechanics*, vol. 68, no. 17-18, pp. 1839–1887, Dec. 2001.

[14] C. E. Renshaw and E. M. Schulson, "Universal behaviour in compressive failure of brittle materials," *Nature*, vol. 412, no. 6850, pp. 897–900, Aug. 2001.

[15] E. M. Schulson and P. Duval, "Structure of ice," in *Creep and Fracture of Ice*. Cambridge University Press, 2009, pp. 5–29.

[16] A. L. Kohout, M. J. M. Williams, S. M. Dean, and M. H. Meylan, "Storm-induced sea-ice breakup and the implications for ice extent," *Nature*, vol. 509, no. 7502, pp. 604–607, May 2014.

[17] J. J. Voermans, J. Rabault, K. Filchuk, I. Ryzhov, P. Heil, A. Marchenko, C. O. Collins III, M. Dabboor, G. Sutherland, and A. V. Babanin, "Experimental evidence for a universal threshold characterizing wave-induced sea ice break-up," *The Cryosphere*, vol. 14, no. 11, pp. 4265–4278, Nov. 2020.

The role of dynamic sea ice in a simplified general circulation model used for paleoclimate studies

Moritz Adam[*1], **Heather J. Andres**[2] **and Kira Rehfeld**[1,3]

[1] Institute of Environmental Physics, Heidelberg University
Heidelberg, Germany

[2] Memorial University of Newfoundland
St. John's, NL, Canada

[3] Geo- und Umweltforschungszentrum, Eberhard Karls Universität Tübingen
Tübingen, Germany

[*] Correspondence: Moritz Adam (madam@iup.uni-heidelberg.de)

Key words: Climate Modelling, Sea Ice Dynamics, Paleoclimatology

Abstract: *Observational records provide a strong basis for constraining sea ice models within a narrow range of climate conditions. Given current trends away from these conditions, models need to be tested over a wider range of climate states. The past provides many such examples based on paleoclimate data, including abrupt, large-amplitude climate events. However, the millennial-duration of typical paleoclimate simulations necessitates balancing the inclusion and sophistication of model processes against computational cost. This is why many simplified models used for multi-millennial simulation only feature representations of thermodynamic sea ice processes, while representing sea ice dynamics is essential for more complex general circulation models. We investigate the impact on climate mean states and variability of introducing sea ice dynamics into the simplified general circulation model PlaSim-LSG.*

We extend the default thermodynamic sea ice component in PlaSim-LSG with one that includes also dynamic sea ice processes. We adapt the structure and parallelization scheme of this new submodel originating from the MITgcm, a more complex state–of–the–art general circulation model. Then, we evaluate the impact of sea ice dynamics on the simulated climate. Comparing climatologies and the variability of the extended model to control simulations of the pre-existing setup, we find that the standard model overestimates sea ice extent, concentration and thickness. The extended model, however, is biased towards low sea ice amounts and extent. Modifying individual parameters in initial tests of the newly added component is not sufficient to compensate for this bias. Still, the general ability of the model to represent positive and negative biases of the sea ice cover provides a promising starting point for the tuning of PlaSim-LSG with sea ice dynamics. Eventually, the extended model can be used to investigate the role of sea ice for past climate oscillations.

1 INTRODUCTION

Paleoclimate simulations provide a test-bed to constrain climate models of different complexity over a much wider range than what is available from instrumental records [1, 2]. In addition, they provide an opportunity to verify concepts on mechanisms and tipping elements which led to abrupt climate oscillations in the past. Sea ice is closely linked with past abrupt climate transitions as found in model studies [3–6] and inferred from paleoclimate archives [7, 8].

A major limitation of transient paleoclimate simulations over multiple millennia with state–of–the–art general circulation models (GCMs), which represent the earth system to a great level of detail, are the high computational costs. Simplified GCMs still offer a reasonable representation of the atmosphere and ocean with a dynamical atmospheric core and a mixed-layer

https://doi.org/10.4995/YIC2021.2021.12383

or dynamic ocean component [9]. Depending on the questions and time scales of interest, they additionally feature representations of other components of the earth system, like sea ice thermodynamics or simple vegetation [10]. Yet, simplified GCMs are typically highly parametrized, have a relatively coarse spatial resolution to allow for moderate computational cost in simulations of multiple millennia, and are often specifically adapted to answer specific research questions with design decisions carefully weighing model complexity against computational costs [9–12]. As a result, simplified GCMs or models of intermediate complexity do not take all earth system processes into account in great detail and, other than in more complex state–of–the–art general circulation models, it is not always common to, for example, model the dynamics of sea ice. Simplified GCMs can, however, help to build a better understanding about which processes are actually needed to effectively resolve particular climate phenomena [11].

The Planet Simulator (PlaSim) [13–15] coupled to the Large Scale Geostrophic Ocean (LSG) [16] is a well-studied simplified GCM. It solves the primitive equations in the atmosphere, approximates the dynamical equations of the ocean under the assumptions of large spatial and temporal scales, and employs simplified parametrisations for processes like sea ice thermodynamics, greenhouse gas forcing, and land cover and vegetation [14]. However, it does not contain a component to model the dynamics of sea ice up to this point. PlaSim and its atmospheric core PUMA have been used in a wide range of applications from synchronization experiments [17] to entropy and hysteresis studies [18]. More recently, the model was used to study the dynamical landscape of climate [19] and in combination with LSG in a study on atmospheric contributions to abrupt climate changes in the past [20]. LSG has been extensively studied and was a part of CMIP1 [21] and of Paleoclimate Model Intercomparison Projects [e.g. 22–24].

Yet, it has been shown that under climate conditions of the Last Glacial Maximum, the time period of greatest land-based ice volume during the Last Glacial period, occuring around 21 kyr ago [25], PlaSim-LSG has pronounced biases with respect to CMIP5 simulations towards too low high-latitude winter temperatures over oceanic and snow-covered land regions [20]. Similar biases occur under present-day conditions during winter in high latitudes. The model overestimates climate sensitivity as is visible from transient simulations [26, 27] and simulates unrealistically large and thick amounts of sea ice. Dynamics of sea ice are crucial to realistically represent the sea ice thickness distribution, while sea ice thermodynamics are most relevant to enable feedbacks with earth system compartments [28, 29]. Thus, sea ice dynamics could potentially help to address the model biases by reducing the amount of too-thick multi-year sea ice present in the model. Also, representing sea ice in PlaSim-LSG in more detail for multi-millennial simulations of past climate could help to reveal the role of sea ice as an important moderating and tipping component in the onset and development of centennial- to millennial scale climate oscillations.

Here, we present our work integrating sea ice dynamics into PlaSim-LSG. While it is essential for more complex GCMs to represent the dynamics of sea ice, this is less common for simplified GCMs or earth system models of intermediate complexity used for coupled simulations of multiple millennia. We describe the pre-existing model configuration, and its newly extended capabilities for dynamic sea ice modelling in Section 2. While one of the primary motivations for these extensions to PlaSim-LSG comes from prospective multi-millennial paleoclimate simulations, we present and discuss initial simulations with the current state of the extended model under present-day climate conditions (Section 3.1). We choose present-day climate for initially constraining the extended model, because direct sea ice observations are available in this period. We test the impact of several key parameters of the sea ice dynamics component (Section 3.2), and evaluate the performance of the new model configuration (Section 3.3). We conclude with future perspectives in Section 4.

2 MODEL DESCRIPTION

2.1 Planet Simulator

PlaSim in version 17 is a simplified GCM which solves the wet primitive equations of the atmosphere in its dynamical core PUMA. We employ a T42 spectral resolution in this study (about 2.8° × 2.8°) and a vertical discretization of 10 layers. PlaSim performs calculations which are for example associated with the mixed-layer ocean, sea ice thermodynamics, and the surface energy balance on a 64 latitude × 128 longitude Gaussian grid [13–15, 20].

The thermodynamic sea ice model in PlaSim is based on the zero-layer Semtner [30] model which is used as well in several other GCMs like the MITgcm [31]. Other than in the standard version of PlaSim-LSG, snow on sea ice is represented following the description of snow on land in the appendix of Andres and Tarasov [20]. In addition to the configuration described there, we implement a simple representation of snow-covered ice, bare ice (meaning ice that is directly exposed to the atmosphere and not covered by snow), and melt pond fraction following the version 2 scheme for sea ice albedo of Køltzow [32]. The purpose of these extensions is to represent the sub-grid scale effects of melt ponds and snow cover on sea ice albedo and the surface energy balance [e.g. 28], which was not the case in PlaSim-LSG previously. We further introduce a globally conservative treatment of excess thermodynamic sea ice growth beyond the physical limits of the sea ice thickness parametrisation in the zero-layer model to improve model stability.

LSG is a 3d general circulation ocean model running at 2.5°× 5° horizontal resolution with 22 vertical layers [16, 22]. LSG implicitly solves the oceanic primitive equations assuming large spatial and temporal scales. This makes a longer integration step than for all other components in PlaSim possible. However, this benefit comes at the expense of not representing gravity waves and barotropic Rossby waves in the ocean [22]. While we run PlaSim at a time step of 20 minutes, one integration step of LSG is performed every 10 days. A mixed-layer ocean with a thickness of 50 m is coupled between LSG and the rest of PlaSim, allowing the ocean to respond to phenomena on shorter time scales than the LSG step. The mixed-layer ocean relaxes to the LSG solution under stationary atmospheric conditions, and mixes the solution from LSG and the thermal response to surface forcing when atmospheric conditions vary [described e.g. in 20].

PlaSim has been coupled to another ocean model, yielding the PlaSim-GENIE model [33]. This implementation replaced the thermodynamic sea ice component of PlaSim and the LSG ocean model by the GOLDSTEINSEAICE and GOLDSTEINOCEAN components. To represent sea ice dynamics, this model employs an advection scheme and uses Laplacian diffusion [34]. Conversely, we choose to retain the LSG model and extend PlaSim-LSG with a component to model sea ice dynamics which we adapt from the MITgcm [31, 35] (see Section 2.2). Unlike PlaSim-GENIE, this approach allows us to also resolve nonlinear viscous-plastic rheologies of sea ice. Another reason is that PlaSim-LSG is user-friendly, well-documented, and extensively studied. Finally, the LSG model has previously been shown to exhibit abrupt climate oscillations [36]. This makes it an ideal test-bed to study such oscillations during the Last Glacial Period.

2.2 Sea ice dynamics component

To model the dynamics of sea ice, we adapt those parts of the MITgcm's sea ice component [35] which solve the sea ice momentum equations of a variant of the nonlinear viscous-plastic (VP) sea ice model introduced by Hibler [37]. The momentum equations in the MITgcm's component are solved with the line-successive-over-relaxation (LSOR) method of Zhang and Hibler

[38] on an Arakawa C grid. Furthermore, we integrate the second- and third-order flux-limited volume- and area-conserving advection schemes from the MITgcm which are used to advect sea ice thickness, concentration, and snow cover [31, 35]. Ice–ocean and ice–atmosphere stresses are directly applied from PlaSim-LSG. While viscous-plastic rheologies with an elliptical yield curve and normal flow rule have been employed for many years in GCMs including MITgcm [e.g. 39, 40], it should be noted that they produce unphysical fracture angles. This is why current development efforts aim at using rheologies which result in better agreements of small-scale sea ice features with observations [e.g. 41, 42]. However, our motivation of incorporating sea ice dynamics into PlaSim-LSG is not to most accurately represent sea ice across a wide range of spatial scales. We rather aim to represent sea ice dynamics in this simplified GCM in sufficient complexity to reduce model biases and study its role in abrupt climate oscillations, which can be observed in multi-millennial simulations with the model. For this purpose, a well-tested and widely-applied sea ice component with elliptical yield curve is sufficient and affordable in terms of added computational costs.

To handle the coupling and interpolation between PlaSim's Gaussian grid and the Arakawa C grid of the dynamic sea ice component, we add and test an intermediate module. We use the pre-existing coupler of PlaSim and LSG to first interpolate any additionally needed oceanic fields to the PlaSim grid. In this process we extend the parallelization architecture of PlaSim to allow for fast handling of neighbouring grid areas used in the discretisations of the sea ice model ("halo exchange"). Additionally, we modify the MITgcm routine interfaces to match with the coding conventions of PlaSim where needed.

In the coupled model, the dynamic sea ice component is called sequentially in every step of PlaSim, with the oceanic stress forcing from LSG being updated at every LSG time step. Thickness categories used in the dynamic model are still represented as zero-layer in the thermodynamic component. This is the default option for the MITgcm sea ice model as well. Given the potential for substantial biases in zero-layer thermodynamic sea ice models, the MITgcm provides an option for the 3-layer model of Winton [43]. This is not available in our configuration due to the current implementation of sea ice thermodynamics in PlaSim.

3 PRELIMINARY RESULTS AND DISCUSSION

3.1 Impact on climatological model biases and simulated climate variability

Starting with the default PlaSim-LSG model parameter set and default parameter settings of the sea ice dynamics component, we run equilibrium simulations under present-day boundary conditions and radiative forcing (CO_2 concentration-equivalent of 360 ppm). Following an initial spin-up phase into a quasi-equilibrium state, we study climatologies over 70 years (150 for the control simulation with only thermodynamic sea ice). Compared to the model setup with only thermodynamic sea ice, we find a strongly decreased mean sea ice extent throughout the year for the model configuration which includes the new component for sea ice dynamics in both hemispheres. Fig. 1 shows this bias for Antarctica. Sea ice extent is below the 1981-2010 median observations for all months. As a result, the 2m temperature has a positive bias compared to reanalysis data (Fig. 2), strongly overcompensating the negative polar 2m temperature bias of the control simulation in the Northern High Latitudes but doing so only slightly in Antarctica. This may hint at the need for differing parametrisations of sea ice albedo for the two hemispheres.

Over all seasons, the mean sea ice thickness from the simulation with the extended model is greatly reduced compared to the configuration with only thermodynamic sea ice in most of the Antarctic Ocean, with thicker accumulations only in the Weddell Sea and Ross Sea (Fig. 1).

We observe a similar behaviour with greatly reduced sea ice thicknesses in the Arctic region (not shown). The configuration with only thermodynamic sea ice exhibits unrealistically thick sea ice under present-day conditions in parts of Antarctica and the Arctic. Thus, tuning of the coupled model should allow us to reach a realistic sea ice state in between the extremes of the configuration of PlaSim-LSG with only thermodynamic sea ice and the extended model.

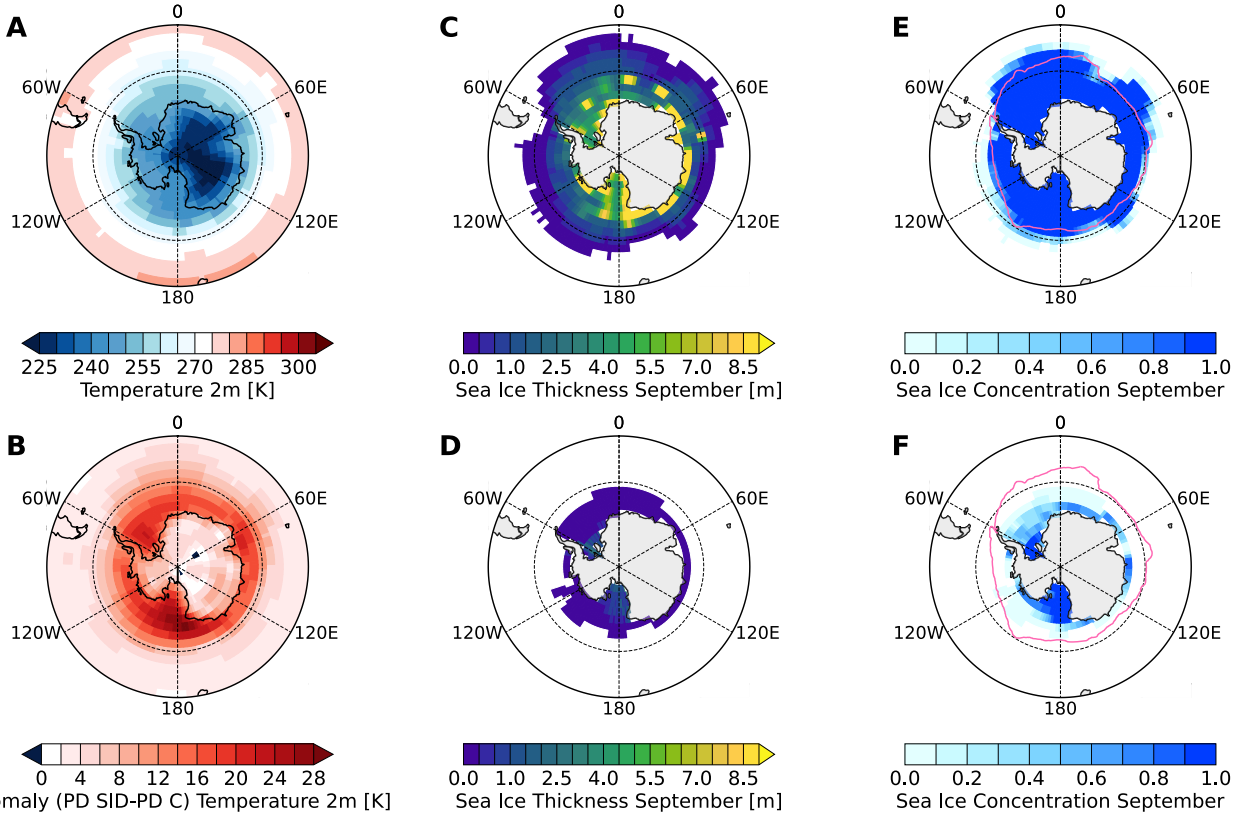


Figure 1: Climatogies of PlaSim-LSG for the present day control simulation (PD C) with only thermodynamic sea ice (top row) and for the simulation with sea ice dynamics under present-day conditions and using the default parametrizations (PD SID, bottom row). Panels show annual mean 2m-temperature of the standard configuration (**A**), annual mean 2m-temperature anomaly between the two configurations (simulation with sea ice dynamics minus control simulation, **B**), absolute sea ice thickness in September (**C, D**) and absolute sea ice concentration in September (**E, F**). The 1981-2010 median sea ice extent of September is indicated in magenta using the data of Cavalieri et al. [44].

Compared to reanalysis data, the model configuration with only thermodynamic sea ice overestimates mid to high latitude annual 2m temperature variability, measured in units of absolute 2m temperature standard deviation of the zonal 2m temperature average (Fig. 2). Conversely, the extended model with sea ice dynamics underestimates 2m temperature variability in Northern mid- to high latitudes. This negative bias in the temperature variability is a lot smaller in Southern mid-latitudes and remains positive for Southern high latitudes. Reduced temperature variability in a state with low amounts of sea ice is in line with previous findings which indicate that temperature anomalies could be amplified less under global warming scenarios which go along with major reductions in sea ice cover [45]. As for the biases of the mean temperature and sea ice states, the results for temperature variability are promising for achieving a realistic representation in between the extremes of the configurations with only thermodynamic sea ice and the one with sea ice dynamics through tuning.

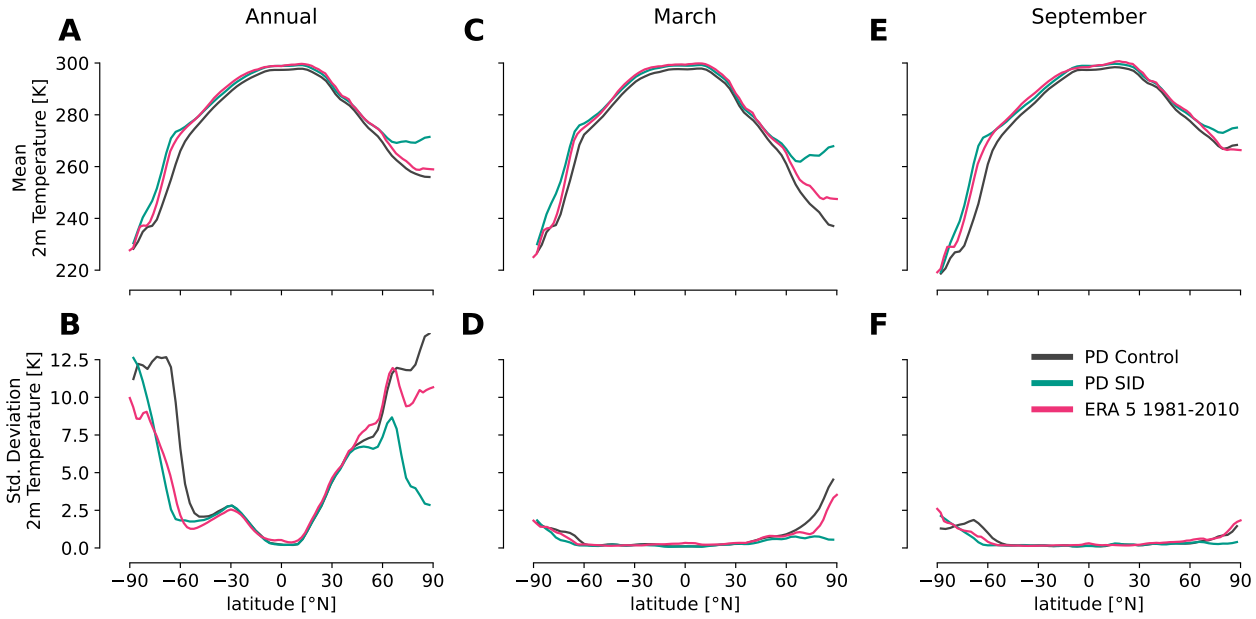


Figure 2: Zonal mean 2m-temperature (top row) and standard deviation (bottom row) for the entire annual cycle (panels A and B), and for monthly climatologies of March (panels C and D) and September (panels E and F) of the present-day control simulation (PD Control), the simulation with sea ice dynamics using default parameters (PD SID), and ERA 5 Reanalysis [46].

3.2 Effects of tuning

We conduct initial tests towards a more realistic representation of sea ice cover in the extended model. Therefore, we vary individual parameters of the dynamic sea ice component aiming at increases in mean sea ice extent, concentration and thickness distribution in the coupled model setup. We only consider key parameters of the sea ice dynamics component and do not conduct a comprehensive tuning procedure involving the entire coupled model. Table 1 lists these parameters which are related to the internal sea ice dynamics or the coupling to oceanic and atmospheric stresses. Parameters are varied equally in both hemispheres to test the general response of the coupled model. The purpose of these initial parameter tests is to gain an understanding for how the coupled model responds to parameter changes in the first place and to provide a basis for more systematic parameter tuning.

Overall we find little to no impact for individual parameter changes in 30 year-averaged simulation data. Sea ice concentration and thickness are slightly increased when changing the sea ice strength parameter P^* to $P^* = 3.1625 \times 10^4\,\mathrm{N/m^3}$ compared to the default parameters. P^* (see Table 1) is the main free parameter in the sea ice strength parametrization, which is why it is particularly suited for modifications [37]. The most notable impact can be found in and around the Beaufort Sea and East Siberian Sea in the Arctic (Fig. 3). For the Antarctic, sea ice extent and concentration are still generally too low. Thus, more realistic values cannot be achieved by variations in individual parameters of the sea ice component alone. Evaluating combinations of varied parameters at once and involving the thermodynamic sea ice component, the albedo parametrisation, and other components of the coupled model in a more rigorous process of parameter optimisation are possible next steps to improve the tuning of PlaSim, following e.g. Mehling et al. [47]. In addition, different parameterisations for the two hemispheres might need to be considered in the tuning. This is generally supported by previous findings indicating that the most accurate representations of sea ice by models which

Table 1: Main parameters related to the sea ice-internal dynamics, and to the coupling to ocean and atmosphere which can be subject to model tuning.

Parameter [units]	Variable name	Description	Default value	Tested values
C_w [10^{-3}]	`SEAICE_waterDrag`	Water drag for freely drifting sea ice	5.5	1.1, 2.4, 3.5, 6.2 (See Table 1 in [49] and the findings of [50])
k_2 [N/m^3]	`SEAICEbasalDragK2`	Parameter and implicit flag for basal stress parametrisation of landfast sea ice	0.0	15.0 (see [51] for additional parameters u_0, k_1, k_2 which have not been changed from defaults here)
C^* [1]	`SEAICE_cStar`	Empirical (exponential) scaling constant to couple sea ice thickness and strength following [37]	20.0	Not changed, same impact achievable with parameter P^*
P^* [10^4 N/m^3]	`SEAICE_strength`	Primary free parameter to couple sea ice thickness and strength following [37]	2.75	2.61, 2.89, 3.025, 3.1625 (correspond to $\sim \pm 5\%$ steps from MITgcm default)
κ [1]	`SEAICEstressFactor`	Overall coupling factor of sea ice and wind stress to ocean surface layer	1.0	0.9, 0.95, 1.05

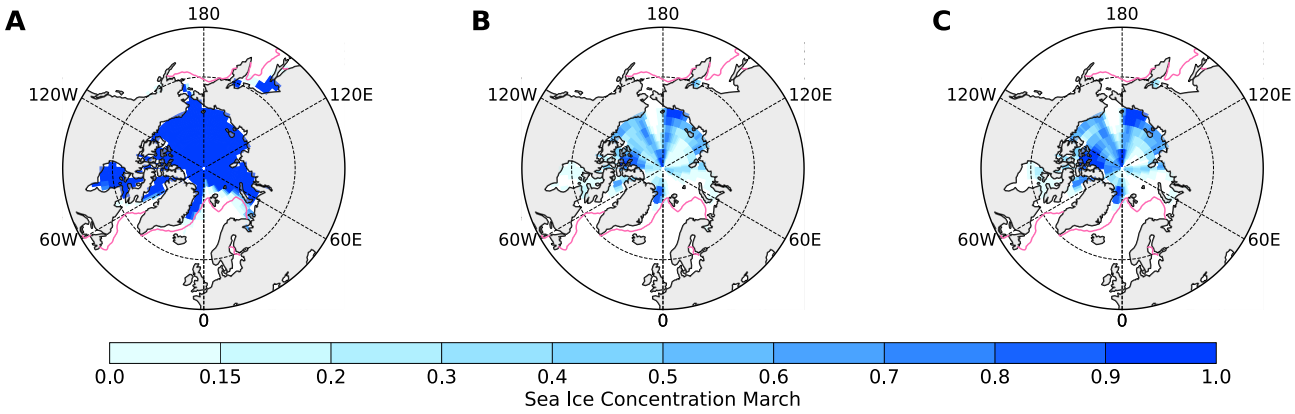


Figure 3: Sea ice concentration and extent in March for the PlaSim-LSG model with only thermodynamic sea ice (panel **A**), the default parametrisation of the extended model with sea ice dynamics (parameter $P^* = 2.75 \times 10^4\,\mathrm{N/m^3}$, **B**), and with an increased parameter $P^* = 3.1625 \times 10^4\,\mathrm{N/m^3}$ (**C**). The 1981-2010 median sea ice extent of March is indicated in magenta using the data of Cavalieri et al. [44].

comprise a VP rheology with an elliptical yield curve cannot be achieved with a single global value of the sea ice strength parameter [48]. A limiting factor for high values of P^* are the potentially inaccurate representations of small-scale sea ice features [42], which are, however, not the focus of our effort of representing sea ice dynamics in this simplified GCM used for multi-millennial simulations of past climate.

3.3 Model Performance

We compare the performance of the new model configuration to the different possible configurations of PlaSim-LSG and the core model PlaSim, for the hardware capabilities of a standard server (Intel Core i5-8600K, $6 \times 3.60 - 4.30\,\mathrm{GHz}$, simulations in this study on 4 cores) and a general purpose high performance computing cluster. Table 2 shows the total run time and the resulting simulation times achievable per day of simulation. The added component increases the simulation time per year by about 10 % compared to the standard PlaSim-LSG setup. For a T42 resolution we achieve a decent benefit when increasing the number of computing cores from eight to sixteen (35 % speed-up). Further increase to 32 cores does only decrease run time marginally. This hints at the primary limitation of performance of PlaSim-LSG in general. In the current implementation the LSG ocean component is not parallelized, although

it is sequentially coupled.

Table 2: Typical computational cost of different model setups. If not stated otherwise, the given numbers refer to a coupled PlaSim-LSG setup with only thermodynamic sea ice at T42 resolution (see Section 2).

Machine	Core Number and Clock Rate [GHz]	Time per simulation year [mn]	Simulation years per day
Standard Server	4×4.3	$\simeq 13.0$	111
Standard Server (with sea ice dynamics)	4×4.3	$\simeq 14.2$	101
Standard Server (only PlaSim core model)	2×4.3	$\simeq 6.4$	225
Standard Server (only PlaSim core model, T21)	2×4.3	$\simeq 2.7$	533
General purpose cluster	8×2.1 (0.25 node)	$\simeq 12.0$	122
General purpose cluster	16×2.1 (0.5 node)	$\simeq 7.8$	184
General purpose cluster	32×2.1 (1 node)	$\simeq 7.2$	199

4 CONCLUSION

We extended the simplified general circulation model PlaSim-LSG with a component for sea ice dynamics adapted from the MITgcm. While it is essential for state–of–the–art general circulation models to represent the coupled dynamics of sea ice, it is less common for simplified general circulation models, which are used for multi-millennial simulations of the past climate, to feature sea ice dynamics. The component now added to PlaSim-LSG solves the Hibler [37] sea ice momentum equations with a non-linear viscous-plastic rheology, and advects sea ice as a response to stress forcing, thereby adopting the most common representation of sea ice dynamics in more complex general circulation models. We studied climatological biases of 2m-temperature, and sea ice extent and thickness in the PlaSim-LSG configuration with only thermodynamic sea ice and with the new model extension under present-day climate conditions. The extended model presently under-represents sea ice extent and thickness compared to present-day observations, and exhibits ice-free summer months. Through this, the negative temperature bias of the standard model configuration in mid- to high latitudes is overcompensated. As expected, the reduced amount of sea ice leads to decreased temperature variability in mid- to high latitudes compared to the PlaSim-LSG version without thermodynamic sea ice. Thus, sea ice dynamics is of great importance for the mean state and variability of the high-latitude climate simulated by PlaSim-LSG.

Variations of individual parameters of the sea ice dynamics component have small to negligible impact on the sea ice bias of the extended model. More thorough tuning of the coupled model components simultaneously is required. However, while the extended model underestimates sea ice thickness and extent, the configuration with only thermodynamic sea ice overestimates them. Therefore we expect that a realistic present-day state in between these extremes can be reached through appropriate and comprehensive tuning of the coupled model. Introducing different parametrisations for the two hemispheres into the dynamic and thermodynamic sea ice components could provide an additional possibility to improve the simulated climate of the model.

Modelling sea ice dynamics adds about 10 % of runtime to the PlaSim-LSG model. This is reasonable given the comparably high degree of explicit formulations introduced into the model architecture to represent sea ice dynamics in more detail. The main bottleneck of the PlaSim-LSG combination remains the unparallelized LSG ocean component, which limits effective parallelization to sixteen cores. While additional tests for physical consistency, model stability under varying boundary conditions, and tuning remain, the extended model adds to the repertoire of different PlaSim-LSG configurations and allows us to study the impact of sea

ice dynamics on known biases of the model. Given the reasonable computational effort needed to run the extended model, it can potentially contribute to the understanding of mechanisms which led to past climate oscillations in multi-millennial simulations of the Last Glacial Cycle.

CODE AVAILABILITY

The most recent state of the extended model can be accessed on GitHub: https://www.github.com/paleovar/plasim17sid.

ACKNOWLEDGEMENTS

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project no. 395588486, and contributes to the PalMod project (https://www.palmod.de). The German Academic Scholarship Foundation (Studienstiftung) funded the stay of MA at Memorial University. We thank one anonymous reviewer for their valuable comments on the manuscript. MA thanks Lev Tarasov for enabling the research stay at Memorial University. We thank Frank Lunkeit for helpful discussions around the PlaSim-LSG model, Lev Tarasov for discussions on integrating the new sea ice component, and Elisa Ziegler for valuable comments on the manuscript. We acknowledge support by the state of Baden-Württemberg through bwHPC for computations on the bwUniCluster (https://www.bwhpc.de). We thank the MITgcm developers for making their extensively documented model code available. The NSIDC data on sea ice extent was downloaded from https://nsidc.org/data/seaice_index/archives (last access: 15.05.2021). The ERA5 surface temperature data was downloaded from the Copernicus Climate Data Store (last access: 19.04.2020).

REFERENCES

[1] P. Braconnot et al. *Nature Climate Change* 2 (2012), 417–424. DOI: 10.1038/nclimate1456.
[2] G. A. Schmidt et al. *Climate of the Past* 10.1 (2014), 221–250. DOI: 10.5194/cp-10-221-2014.
[3] C. Li et al. *Journal of Climate* 23.20 (2010), 5457 –5475. DOI: 10.1175/2010JCLI3409.1.
[4] A. Voigt and D. S. Abbot. *Climate of the Past* 8.6 (2012), 2079–2092. DOI: 10.5194/cp-8-2079-2012.
[5] G. Vettoretti and W. R. Peltier. *Geophysical Research Letters* 43.10 (2016), 5336–5344. DOI: 10.1002/2016GL068891.
[6] C. Li and A. Born. *Quaternary Science Reviews* 203 (2019), 1 –20. DOI: 10.1016/j.quascirev.2018.10.031.
[7] T. M. Dokken et al. *Paleoceanography* 28.3 (2013), 491–502. DOI: 10.1002/palo.20042.
[8] U. Hoff et al. *Nature Communications* 7.1 (2016), 12247. DOI: 10.1038/ncomms12247.
[9] G. Flato et al. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T. Stocker et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013.
[10] T. Hajima et al. *Progress in Earth and Planetary Science* 1.1 (2014), 1–25. DOI: 10.1186/s40645-014-0029-y.
[11] M. Claussen et al. *Climate Dynamics* 18.7 (2002), 579–586. DOI: 10.1007/s00382-001-0200-1.
[12] H. Goosse et al. *Geoscientific Model Development* 3.2 (2010), 603–633. DOI: 10.5194/gmd-3-603-2010.
[13] K. Fraedrich et al. *Meteorol. Z.* 14.3 (2005), 299–304. DOI: 10.1127/0941-2948/2005/0043.
[14] F. Lunkeit et al. Tech. rep. 2012. URL: https://www.mi.uni-hamburg.de/en/arbeitsgruppen/theoretische-meteorologie/modelle/sources/psusersguide.pdf.
[15] F. Lunkeit et al. Tech. rep. 2012. URL: https://www.mi.uni-hamburg.de/en/arbeitsgruppen/theoretische-meteorologie/modelle/sources/psreferencemanual-1.pdf.
[16] E. Maier-Reimer and U. Mikolajewicz. Tech. rep. Max-Planck-Institut fuer Meteorologie Hamburg, 1992. DOI: 10.2312/WDCC/DKRZ_Report_No02.
[17] F. Lunkeit. *Chaos* 11.1 (2001), 47–51. DOI: 10.1063/1.1338127.
[18] K. Fraedrich. *European Physical Journal Plus* 127.5 (2012). DOI: 10.1140/epjp/i2012-12053-7.
[19] G. Margazoglou et al. *Proceedings of the Royal Society A* 477.2250 (2021), 20210019. DOI: 10.1098/rspa.2021.0019. eprint: 2010.10374.
[20] H. J. Andres and L. Tarasov. *Climate of the Past* 15.4 (2019), 1621–1646. DOI: 10.5194/cp-15-1621-2019.

[21] C. Covey et al. *Global and Planetary Change* 37.1-2 (2003), 103–133. DOI: 10.1016/S0921-8181(02)00193-5.
[22] E. Maier-Reimer et al. *Journal of Physical Oceanography* 23.4 (1993), 731–757. DOI: 10.1175/1520-0485(1993)023⟨0731:MCOTHL⟩2.0.CO;2.
[23] F. Feser et al. *Theoretical and Applied Climatology* 65.1-2 (2000), 1–15. DOI: 10.1007/s007040050001.
[24] P. Braconnot et al. November (2004), 515–533. DOI: 10.1007/978-1-4020-2121-3_24.
[25] A. M. McCabe et al. *Science* 325.5941 (2009), 710–714. DOI: 10.1126/science.1172873.
[26] E. Ziegler et al. EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-11006, 2021. DOI: 10.5194/egusphere-egu21-11006.
[27] H. Andres and L. Tarasov. EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-13778, 2021. DOI: 10.5194/egusphere-egu21-13778.
[28] C. Haas. *Sea Ice: An Introduction to its Physics, Chemistry, Biology and Geology*. Ed. by D. N. Thomas and G. S. Dieckmann. Blackwell Science Ltd, 2003.
[29] J. Zhang and D. Rothrock. *Journal of Geophysical Research: Oceans* 105.C2 (2000), 3325–3338. DOI: 10.1029/1999jc900320.
[30] A. J. Semtner. *Journal of Physical Oceanography* 6.3 (1976), 379 –389. DOI: 10.1175/1520-0485(1976)006⟨0379:AMFTTG⟩2.0.CO;2.
[31] MITgcm Group. Last accessed: 14.10.2021. MIT/EAPS, Cambridge, MA 02139, USA, 1997-2021. URL: https://mitgcm.readthedocs.io/en/latest/.
[32] M. Køltzow. *Journal of Geophysical Research: Atmospheres* 112.D7 (2007). DOI: 10.1029/2006JD007693.
[33] P. B. Holden et al. *Geoscientific Model Development* 9.9 (2016), 3347–3361. DOI: 10.5194/gmd-9-3347-2016.
[34] N. R. Edwards and R. Marsh. *Climate Dynamics* 24.4 (2005), 415–433. DOI: 10.1007/s00382-004-0508-8.
[35] M. Losch et al. *Ocean Model.* 33.1-2 (2010), 129–144. DOI: 10.1016/j.ocemod.2009.12.008.
[36] G. Knorr and G. Lohmann. *Geochemistry, Geophysics, Geosystems* 8.12 (2007). DOI: 10.1029/2007GC001604.
[37] W. D. Hibler. *Journal of Physical Oceanography* 9.4 (1979), 815–846. DOI: 10.1175/1520-0485(1979)009⟨0815:adtsim⟩2.0.co;2.
[38] J. Zhang and W. D. Hibler. *Journal of Geophysical Research* 102.4 (1997), 412–415. DOI: 10.1029/96JC03744.
[39] D. Ferreira et al. *Journal of Climate* 24.4 (2011), 992–1012. DOI: 10.1175/2010JCLI3580.1.
[40] F. Zheng et al. *Advances in Atmospheric Sciences* 38.1 (2021), 29–48. DOI: 10.1007/s00376-020-9223-6.
[41] D. Ringeisen et al. *Cryosphere* 15.6 (2021), 2873–2888.
[42] A. Bouchat and B. Tremblay. *Journal of Geophysical Research: Oceans* 122.7 (2017), 5802–5825. DOI: 10.1002/2017JC013020.
[43] M. Winton. *Journal of Atmospheric and Oceanic Technology* 17.4 (2000), 525 –531. DOI: 10.1175/1520-0426(2000)017⟨0525:ARTLSI⟩2.0.CO;2.
[44] D. J. Cavalieri et al. Last accessed: 19.05.2021. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center, 1996, updated yearly. DOI: 10.5067/8GQ8LZQVL0VL.
[45] C. Huntingford et al. *Nature* 500.7462 (2013), 327–330. DOI: 10.1038/nature12310.
[46] H. Hersbach et al. *Quarterly Journal of the Royal Meteorological Society* 146.730 (2020), 1999–2049. DOI: 10.1002/qj.3803.
[47] O. Mehling et al. EGU General Assembly 2021, online, 19–30 Apr 2021, EGU21-1328, 2021. DOI: 10.5194/egusphere-egu21-1328.
[48] S. Juricke et al. *Journal of Climate* 26.11 (2013), 3785–3802. DOI: 10.1175/JCLI-D-12-00388.1.
[49] P. Lu et al. *Journal of Geophysical Research: Oceans* 116.C7 (2011). DOI: 10.1029/2010JC006878.
[50] H. D. B. S. Heorton et al. *Journal of Geophysical Research: Oceans* 124.8 (2019), 6388–6413. DOI: 10.1029/2018JC014881.
[51] J.-F. Lemieux et al. *Journal of Geophysical Research: Oceans* 120.4 (2015), 3157–3173. DOI: 10.1002/2014JC010678.

MODEL REDUCTION AND ARTIFICIAL INTELLIGENCE TECHNIQUES FOR SURROGATE AND DATA-ASSISTED MODELS IN COMPUTATIONAL ENGINEERING

Reducing computational time for FEM post-processing through the use of feedforward neural networks

Martin Zlatić[*], **Marko Čanađija**[†]

[*] Faculty of Engineering
University of Rijeka
Rijeka, Croatia
e-mail: mzlatic@riteh.hr

[†] Faculty of Engineering
University of Rijeka
Rijeka, Croatia
e-mail: marko.canadija@riteh.hr

Key words: Machine Learning, FEM, Postprocessing

Abstract: *With the recent surge in neural network usage, machine learning libraries have become more convenient to use and implement. In this paper the possibility of using neural networks in order to faster process displacements obtained from finite element calculation and replace existing post-processing procedures is investigated. The method is implemented on 2D membrane finite elements for their relative simplicity. A speed up is observed in comparison to traditional methods of post-processing. Possible further applications of this method are also presented in this paper.*

1 INTRODUCTION

As the performance of central processing units (CPUs) and graphics processing units (GPUs) increased in the past decades, neural networks have gained momentum since they can be implemented easier than ever before. This has given rise to code such as TensorFlow [1] and wrappers around the code to make it easier to use such as Keras [2]. Finite element calculations are commonly used by engineers in a plethora of fields [3] and the two are recently being combined and commonly used in describing constitutive models, multiscale simulations and other fields [4–12].

This paper has been inspired mainly by the work of Jung et al. [11] where neural networks are used to generate the finite element strain-displacement matrix in order to construct the element stiffness matrix. In this paper the possibility of calculating stress directly from nodal displacements using neural networks is presented as well as the speed increase over the FEM software Abaqus' post-processing. All the necessary data will be generated from Abaqus and Python with Keras will be used to train and evaluate neural networks.

2 PROBLEM STATEMENT

The goal of this paper is to correctly model linear elasticity trained on 2D membrane elements and directly obtain stress results from nodal displacements. A stiffness matrix of any finite element is given in eq. 1, where B is the strain displacement matrix and C is the material matrix.

$$K = \int_V B^T C B \, dV \tag{1}$$

In the paper by Jung [11] the strain displacement matrix is generated with neural networks, while in Huang et. al [12] only the material behaviour is captured. In Eq. 2 u are the nodal displacements of an element. Thus the need for having separate networks for generating a

https://doi.org/10.4995/YIC2021.2021.12473

strain displacement or material matrix can be eliminated and help speed up the calculation. Obtaining stresses from finite element calculations the following expression is used:

$$\sigma = CBu \tag{2}$$

3 DATA GENERATION AND PREPARATION

The data for training the neural network was obtained through the software Abaqus. First, a case was prepared through a script where a simple 4 point plate meshed with 2D membrane elements (type M3D4) was constrained at one edge and loaded on a different edge, Fig.1. The coordinates x_i, y_i on the plate were chosen randomly between 0.5 and 1.5 metres in their respective quadrants, and the nodal loads on the edge are all equal and are randomly chosen from $F_x^n \in \{-3000, 3000\}$ N , and $F_y^n \in \{-3000, 3000\}$ N with the individual nodal force then being $F^n = F_x^n \cdot i + F_y^n \cdot j$. A structured mesh of element size 50 mm was used. The force varies from case to case, as well as the edge to which the load or the constraint is applied. The minimum force applied to an edge can be 84 kN, while the maximum force can be 254 kN. The load direction also varies from case to case. In total 800 plates were auto-generated for obtaining training data, in total around 800 000 training samples.

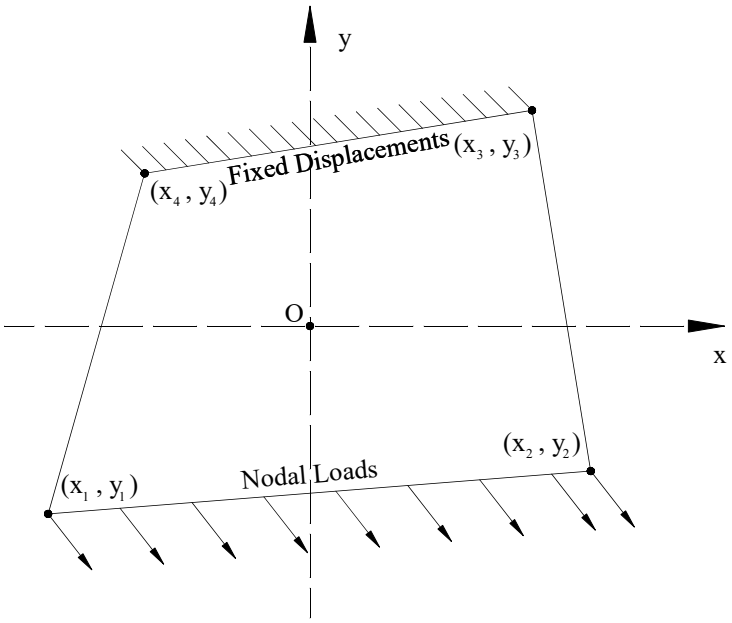


Figure 1: Plate geometry and boundary conditions for generating training data.

Once all the simulations are finished the data is processed in the following manner:

1. Nodal positions are obtained from the Abaqus input file
2. The intersection of the diagonals is found.
3. Distances between the intersection and nodes are found and stored in an auxiliary vector.
4. Displacements of the nodes are found and stored in an auxiliary vector.
5. Nodal positions and displacements and Poisson's ratio are added to the input vector for training.
6. Stress results from integration points are added into an output vector for training.

Data preparation is shown on Fig. 2. In total the input vector contains 17 values, nodal positions, displacements, and Poisson's ratio, see Eq. 3.

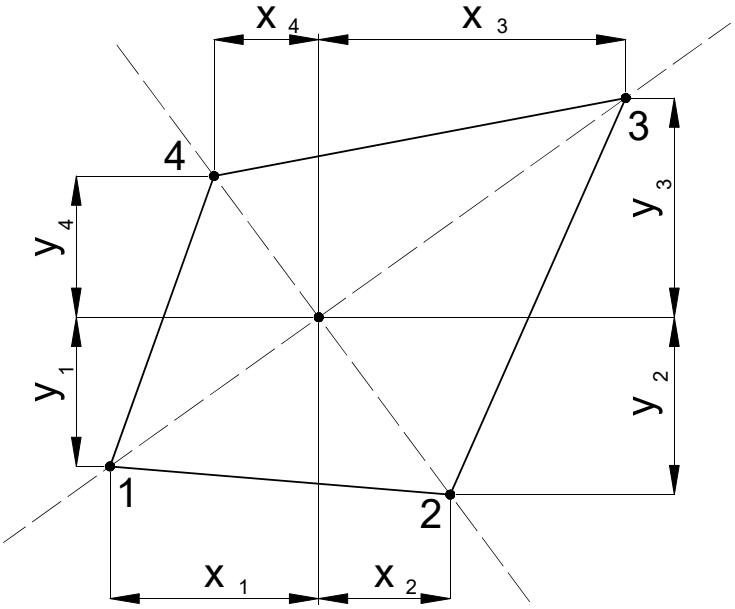


Figure 2: Illustration of diagonals intersection and nodal distances.

$$\boldsymbol{u} = (x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4, u_x^1, u_y^1, u_x^2, u_y^2, u_x^3, u_y^3, u_x^4, u_y^4, \nu) \tag{3}$$

The output vector contains 12 values, 2 normal stresses and 1 shear stress per integration point, see Eq. 4. The lower indices refer to the stress component, and the upper indices refer to the integration point.

$$\boldsymbol{\sigma} = (\sigma_x^1, \sigma_y^1, \tau_{xy}^1, \sigma_x^2, \sigma_y^2, \tau_{xy}^2, \sigma_x^3, \sigma_y^3, \tau_{xy}^3, \sigma_x^4, \sigma_y^4, \tau_{xy}^4,) \tag{4}$$

4 TRAINING AND EVALUATING THE NETWORK

The hyperparameters of the network (number of layers, neurons per layer, activation function, kernel initialization) were determined through trial and error. An illustration of a general feed-forward neural network is given in Fig. 3 The best performing hyperparameters were:

- Number of hidden layers: 2
- Neurons per layer: 100
- Activation function: Parametric Rectified Linear Unit (PReLU)
- Kernel initialization: Glorot normal [13]
- Kernel regularizer: L2 regularization
- Bias: None
- Optimizer: Adam
- Loss measure: Mean squared error

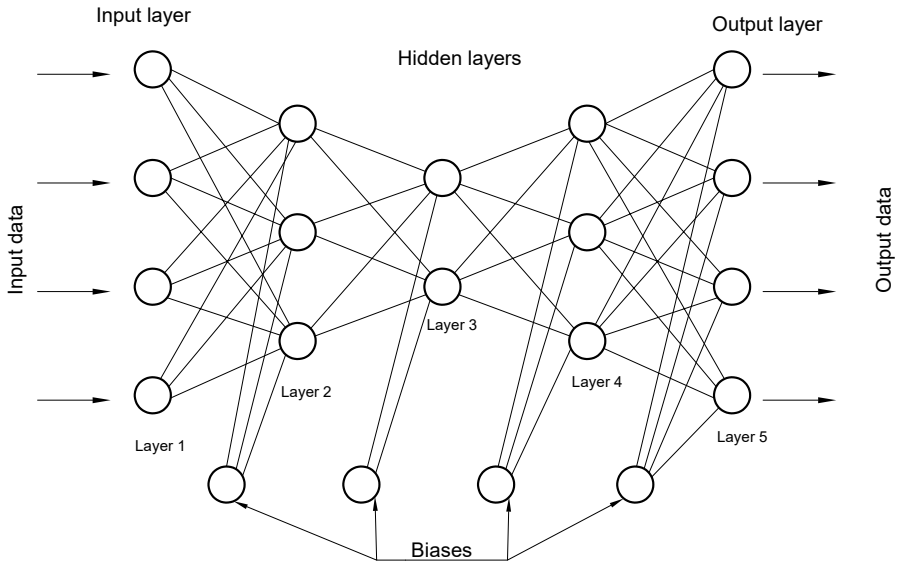


Figure 3: Illustration of a feed-forward neural network.

For each of the stress components a separate network was trained, due to this the output was subdivided into three outputs each consisting of 4 values.

$$\boldsymbol{\sigma_x} = (\sigma_x^1, \sigma_x^2, \sigma_x^3, \sigma_x^4) \tag{5}$$

$$\boldsymbol{\sigma_x} = (\sigma_y^1, \sigma_y^2, \sigma_y^3, \sigma_y^4) \tag{6}$$

$$\boldsymbol{\tau_{xy}} = (\tau_{xy}^1, \tau_{xy}^2, \tau_{xy}^3, \tau_{xy}^4) \tag{7}$$

The networks were then trained with early stopping enabled in case the validation loss does not improve for 10 epochs.

A few additional cases were generated that have not been in the training or validation set, these cases are declared to be the holdout set. As a general measure of accuracy the R^2 values are given in Table 1. The values were obtained on the holdout set.

Table 1: R^2 values for each stress component.

σ_x^1	σ_y^1	τ_{xy}^1	σ_x^2	σ_y^2	τ_{xy}^2	σ_x^3	σ_y^3	τ_{xy}^3	σ_x^4	σ_y^4	τ_{xy}^4
0.9906	0.9901	0.988	0.9915	0.975	0.978	0.9909	0.987	0.969	0.9905	0.975	0.969

A visual representation is given on Fig. 4. For brevity other plots like the one in Fig. 4 are not shown as they are very similar, as can be seen in Table 1.

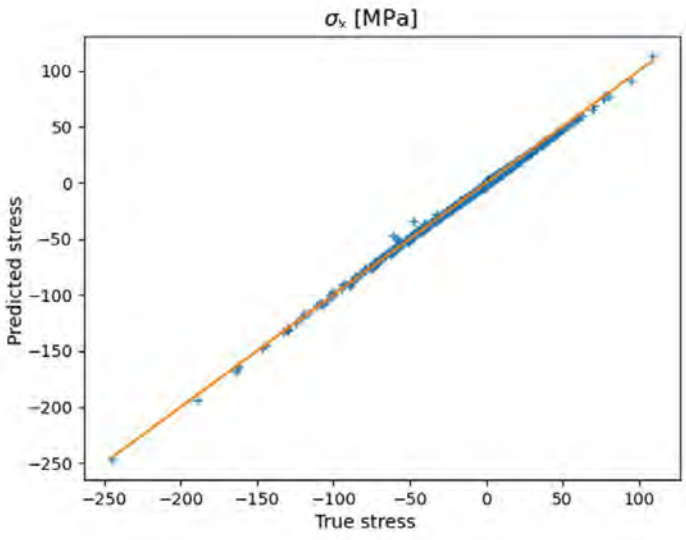


Figure 4: Plot of σ_x^1, predicted vs Abaqus stress.

5 TIME REDUCTION

Time was measured for Abaqus needed to complete the analysis of a plate with a set number of elements. Afterwards the simulation was rerun, but with the option to output stress results unchecked. The time difference between these two simulations is taken as the time necessary for Abaqus to post-process stress results. Then a dataset of the same size was processed by the previously obtained neural network and the execution time was measured. Given that 3 separate networks are used (one for each stress component) the execution time listed for the neural network is the total time for all 3 networks.

The time required for Abaqus to post-process stress results is 11 seconds while the execution time for the neural networks is 1.89 seconds. This translates into a time reduction of 82.8% or an acceleration of 5.82 times. Time required for saving the results from the networks to a file is also included in the neural network execution time (0.01 seconds per file save in NumPy).

6 CONCLUSION

Neural networks are a viable option for post-processing displacements of finite element calculations especially given the observed time reduction. In this paper they have been used in conjunction with 2D membrane finite elements and a linearly elastic material model. Applying neural networks to more complex material models such as those presented in Huang et al. [12] or du Bos et al. [9] and implementing them in non-linear solvers has the potential to reduce the computational time by a large margin.

Acknowledgments

This work has been fully supported by Croatian Science Foundation under the project IP-2019-04-4703. This support is gratefully acknowledged.

REFERENCES

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever,

K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[2] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[3] K. Bathe, *Finite element procedures*. Place of publication not identified: publisher not identified, 2006.

[4] J. Ghaboussi, D. A. Pecknold, M. Zhang, and R. M. Haj-Ali, "Autoprogressive training of neural network constitutive models," *International Journal for Numerical Methods in Engineering*, vol. 42, no. 1, pp. 105–126, may 1998.

[5] L. Liang, M. Liu, C. Martin, and W. Sun, "A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis," *Journal of The Royal Society Interface*, vol. 15, no. 138, p. 20170844, jan 2018.

[6] J. He, L. Li, J. Xu, and C. Zheng, "ReLU deep neural networks and linear finite elements," *Journal of Computational Mathematics*, vol. 38, no. 3, pp. 502–527, 2020.

[7] G. Capuano and J. J. Rimoli, "Smart finite elements: A novel machine learning application," *Computer Methods in Applied Mechanics and Engineering*, vol. 345, pp. 363–381, mar 2019.

[8] E. Haghighat, M. Raissi, A. Moure, H. Gomez, and R. Juanes, "A deep learning framework for solution and discovery in solid mechanics," 2020.

[9] M. L. du Bos, F. Balabdaoui, and J. N. Heidenreich, "Modeling stress-strain curves with neural networks: a scalable alternative to the return mapping algorithm," *Computational Materials Science*, vol. 178, p. 109629, jun 2020.

[10] P. Carrara, L. D. Lorenzis, L. Stainier, and M. Ortiz, "Data-driven fracture mechanics," *Computer Methods in Applied Mechanics and Engineering*, vol. 372, p. 113390, dec 2020.

[11] J. Jung, K. Yoon, and P.-S. Lee, "Deep learned finite elements," *Computer Methods in Applied Mechanics and Engineering*, vol. 372, p. 113401, dec 2020.

[12] D. Huang, J. N. Fuhg, C. Weißenfels, and P. Wriggers, "A machine learning based plasticity model using proper orthogonal decomposition," *Computer Methods in Applied Mechanics and Engineering*, vol. 365, p. 113008, jun 2020.

[13] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research - Proceedings Track*, vol. 9, pp. 249–256, 01 2010.

Comparison of numerical and experimental strains distributions in composite panel for aerospace applications

Waldemar Mucha*, Waclaw Kuś*, Júlio C. Viana† and João Pedro Nunes†

* Department of Computational Mechanics and Engineering
Silesian University of Technology
Gliwice, Poland
e-mail: waldemar.mucha@polsl.pl; waclaw.kus@polsl.pl

† Department of Polymer Engineering, IPC - Institute for polymers and composites
University of Minho
Guimarães, Portugal
e-mail: jcv@dep.uminho.pt; jpn@dep.uminho.pt

Key words: aerostructures, lightweight structures, strain measurement, finite element method, operational load monitoring, artificial intelligence

Abstract: *In structural applications of aerospace industry, weight efficiency, understood as minimal weight and maximal stiffness, is of great importance. This criterion can be achieved by composite lightweight structures. Typical structures for aforementioned applications are sandwich panels (e.g., with honeycomb core) and stiffened panels (e.g., with blade ribs, T-bar ribs, or hat ribs). In this paper, a hat-stiffened panel, made of carbon/epoxy woven composite, is considered. Results of experiments, consisting of loading the panel and measuring exciting forces and strains (using strain gages), are presented. The results are compared to strains distribution obtained from finite element model of the panel.*

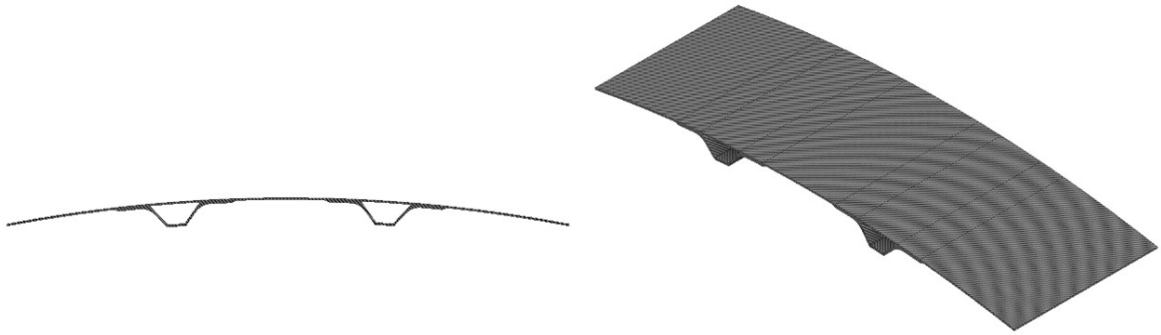
1 INTRODUCTION

Composite lightweight structures are popularly used for aerospace applications (aircraft skin, wings etc.). They are characterized by very good weight efficiency, which means low weight and high stiffness. Typical aerostructures are sandwich panels (e.g. with honeycomb core) and stiffened panels (e.g. with blade ribs, T-bar ribs, or hat ribs) [1–5].

In aerospace applications, such structures are often monitored in real-time in order to detect potential changes to material or geometric properties which could mean potential damage (Structural Health Monitoring) [6–8], or in order to estimate the remaining in-service life of the structures (Operational Loads Monitoring) [9–12]. The monitoring is often performed by means of embedded or surface mounted strain sensors (e.g., intrinsic optical fibers or strain gages) [13, 14].

In the following paper, a hat-stiffened panel, of geometry presented in Fig. 1, is considered. Dimensions of the panel are 597 x 204 x 29 mm. The material of the panel is a 10-layer laminate – woven carbon fiber / epoxy composite. Mechanical properties of a single layer, of thickness 230 μm, are presented in Table 1.

https://doi.org/10.4995/YIC2021.2021.12572



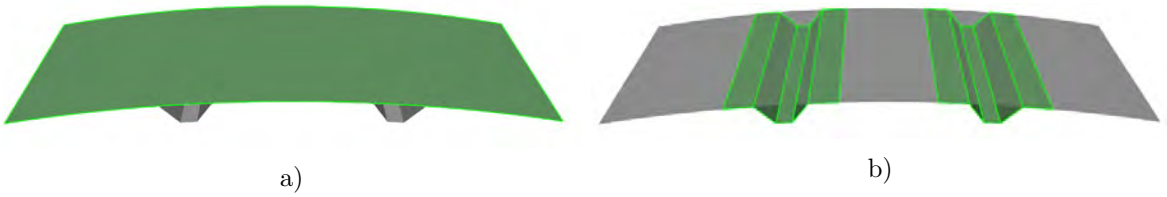
a) b)

Figure 1: Geometry of the hat-stiffened panel: a) front view, b) isometric view (597 x 204 x 29 mm).

Table 1: Mechanical properties of a single layer

Young Module	**[GPa]**	**Shear Module**	**[GPa]**	**Poisson's ratio**	**[]**
E1	64.70	G12	4.00	ν12	0.04
E2	64.70	G23	2.66	ν23	0.34
E3	7.17	G13	2.66	ν13	0.34

The geometry of the panel can be divided into the main curved part and the two ribs (Fig. 2). Each part contains 10 layers of carbon woven, therefore the panel is 2.3 mm thick (except the common part that is 4.6 mm thick). In the main part, layers 1, 3, 5, 6, 8, 10 have the carbon fibers of the woven parallel to the external edges of the panel, and layers 2, 4, 7, 9 rotated by the angle of 45°. In the ribs, all ten layers of the woven are of the same orientation – parallel to the external edges of the panel.



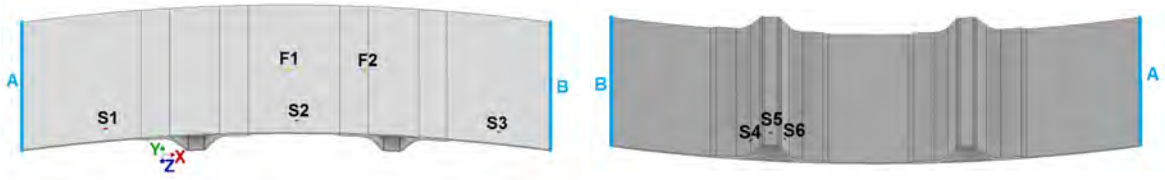
a) b)

Figure 2: Geometric regions of layer groups: a) main curved part, b) ribs.

In Section 2, the finite element model of the panel is presented, as well as results of numerical simulations for example load cases. Section 3 describes the experimental results of loading the panel and strain measurements. Strain distributions in numerical simulations and experiments are compared. Conclusions and idea of cyber-physical system for real-time monitoring of aerostructures using artificial intelligence techniques is presented in Section 4.

2 FINITE ELEMENT MODEL AND NUMERICAL RESULTS

The boundary conditions of the numerical model and characteristic points are presented in Fig. 3. On edge A displacements along axes X and Y were fixed. On edge B displacements along axis Y were fixed. Force is applied to two alternative points – F1 and F2. Displacements along axis Z are fixed in the point where load is applied. Six sensors (strain gages) are mounted at points S1-S6, to measure strains in longitudinal directions.



a) b)

Figure 3: Boudary conditions, strain sensors positions and load points: a) view from top, b) view from bottom.

Surface finite element model was created using ANSYS Workbench software, with ACP module. The model has got 18757 finite elements (of quadratic order) and 56584 nodes. The mesh is presented in Fig. 4. In points where loads are applied and strain sensors are mounted, the mesh is refined.

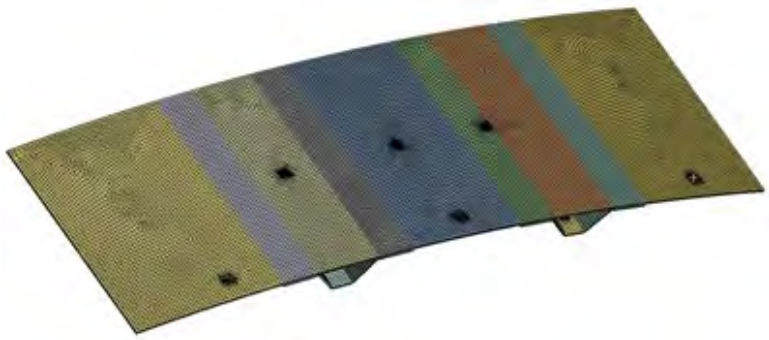


Figure 4: Finite element model.

Force of value 20 N was applied to points F1 and F2, sequentially. Displacements and strains distributions are presented in Fig. 5 and 6 for both load cases, respectively.

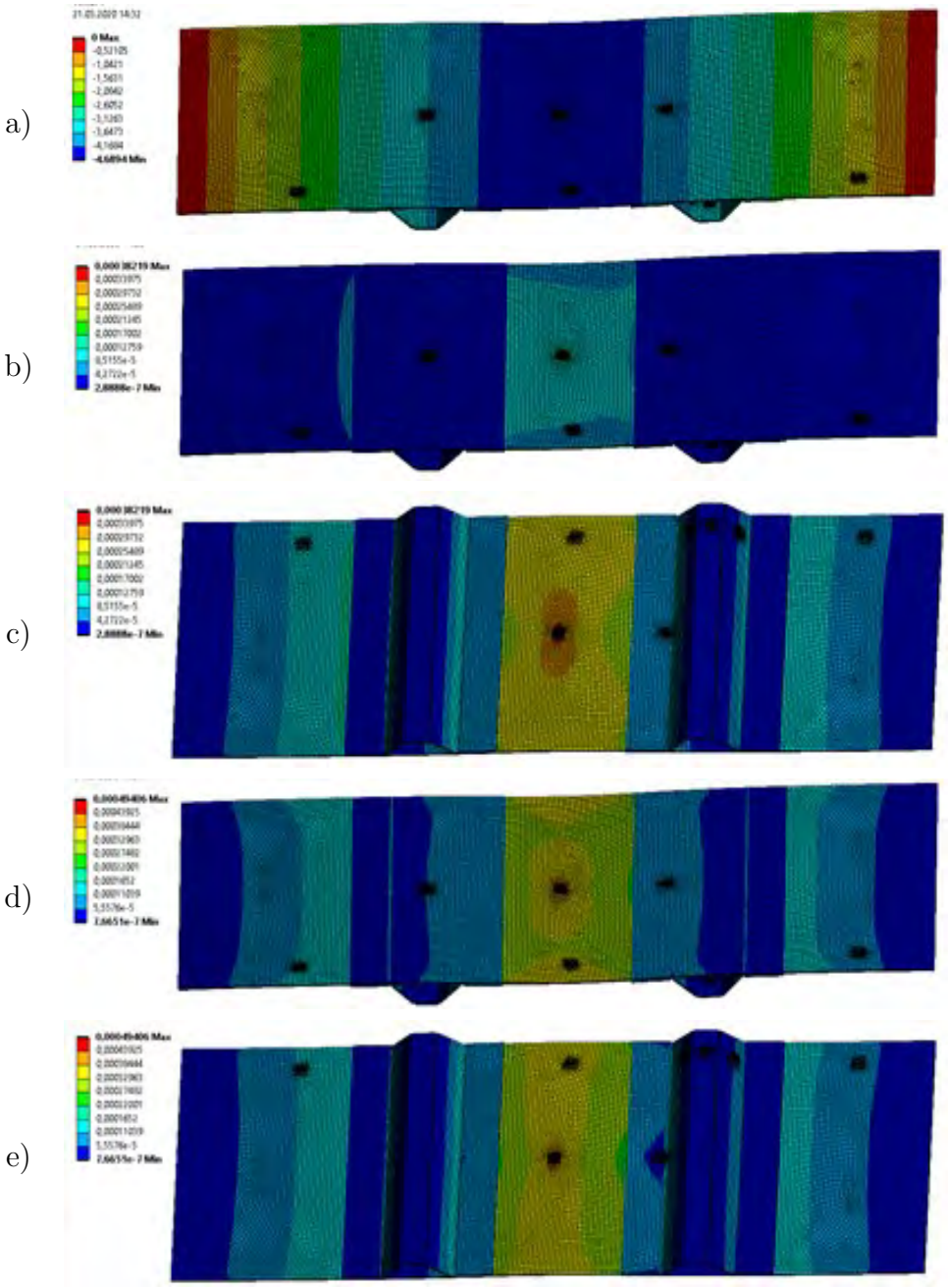


Figure 5: Results of finite element simulation, force applied at point F1 (central midpoint of the curved panel): a) vertical deformation [mm], b) maximum principal strain (top view), c) maximum principal strain (bottom view), d) von Mises equivalent strain (top view), e) von Mises equivalent strain (bottom view)

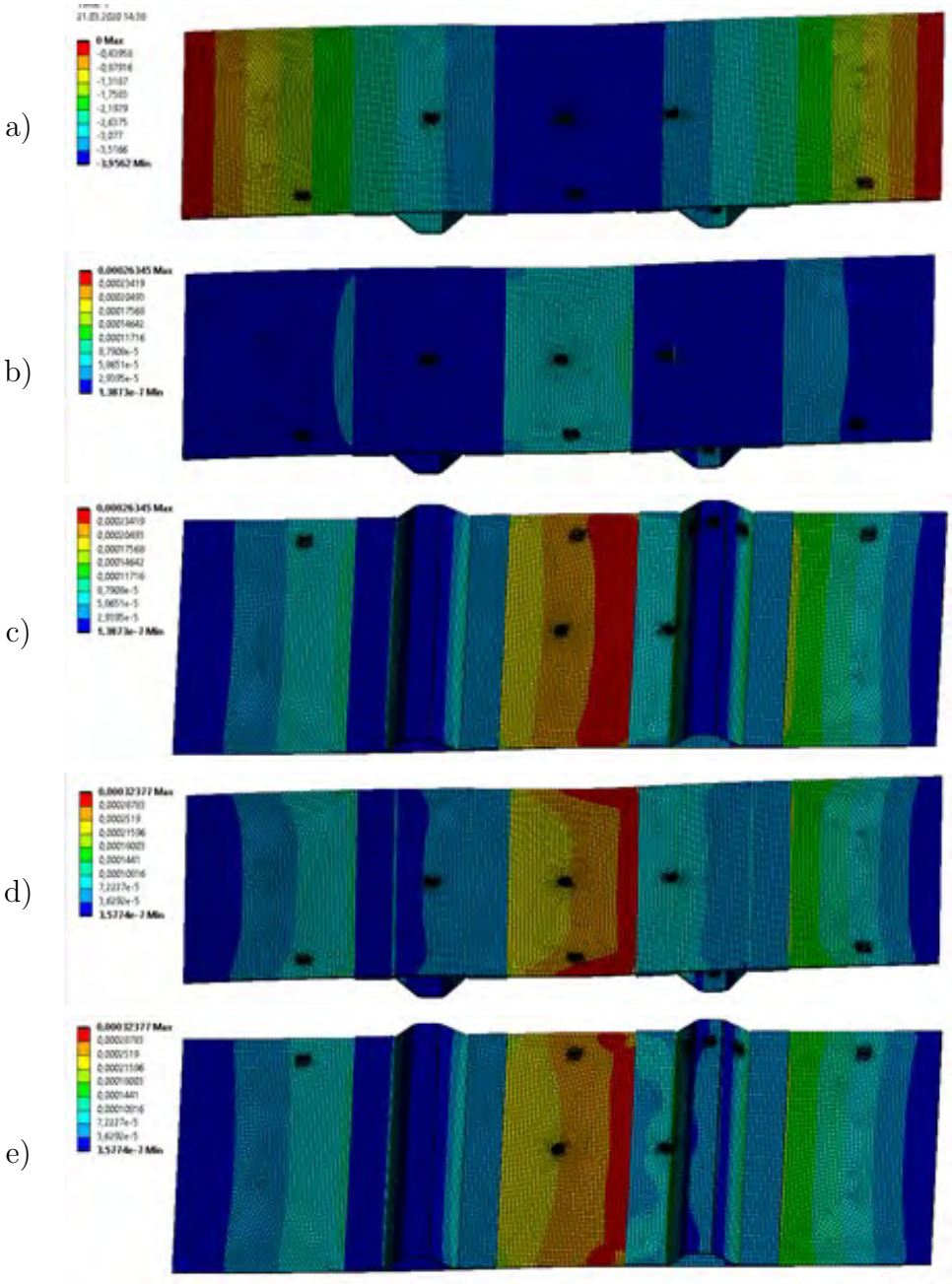


Figure 6: Results of finite element simulation, force applied at point F2 (near of the rib flap): a) vertical deformation [mm], b) maximum principal strain (top view), c) maximum principal strain (bottom view), d) von Mises equivalent strain (top view), e) von Mises equivalent strain (bottom view).

3 EXPERIMENTAL RESULTS

Experimental measurements of strains distributions were performed using a universal testing machine MTS Insight 10 with 500 N load cell and data acquisition system Hottinger Baldwin Messtechnik (HBM) MGCplus. Six strain gages and analog output of the applied load from the testing machine were connected to the acquisition system. Force was applied at points F1 and F2 in the range 0-20 N, with velocity of 0.5 mm/min. Photograph of the test stand during experiment is presented in Fig. 7.

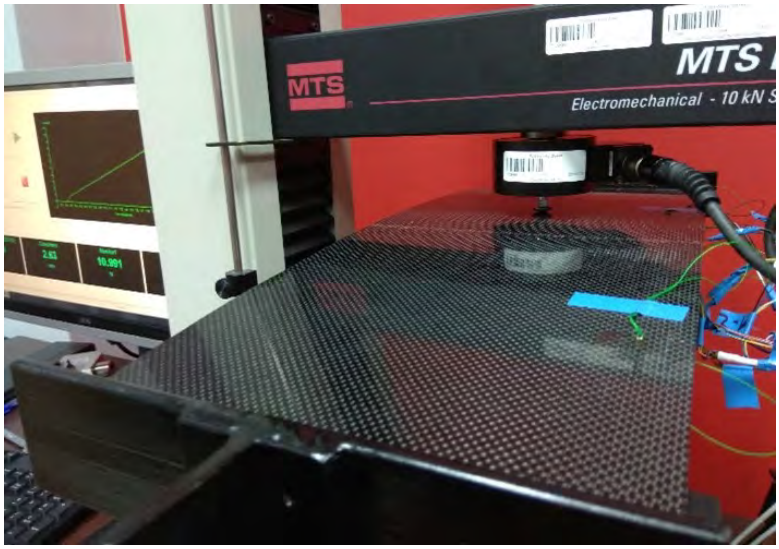


Figure 7: Photograph of experimental testing.

Strain values obtained from sensors S1-S6 during the experiments (for both load cases) were compared to numerical values. In the finite element model, the actives areas of the strain gages were modelled as separated surfaces, where average strains in local coordinate systems were computed. Comparison of the numerical and experimental data is visualized on force-strain plots, in Fig. 8. Table 2 summarizes the mean-square-errors between numerically computed and experimentally measured strains, for all strain sensors, for both load cases.

Table 2: Mean square error between numerical analysis and experimental data $[\mu m^2/m^2]$.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**
F1	6.45	13.62	2.40	4.08	0.60	1.44
F2	12.18	48.30	5.95	1.56	1.99	12.60

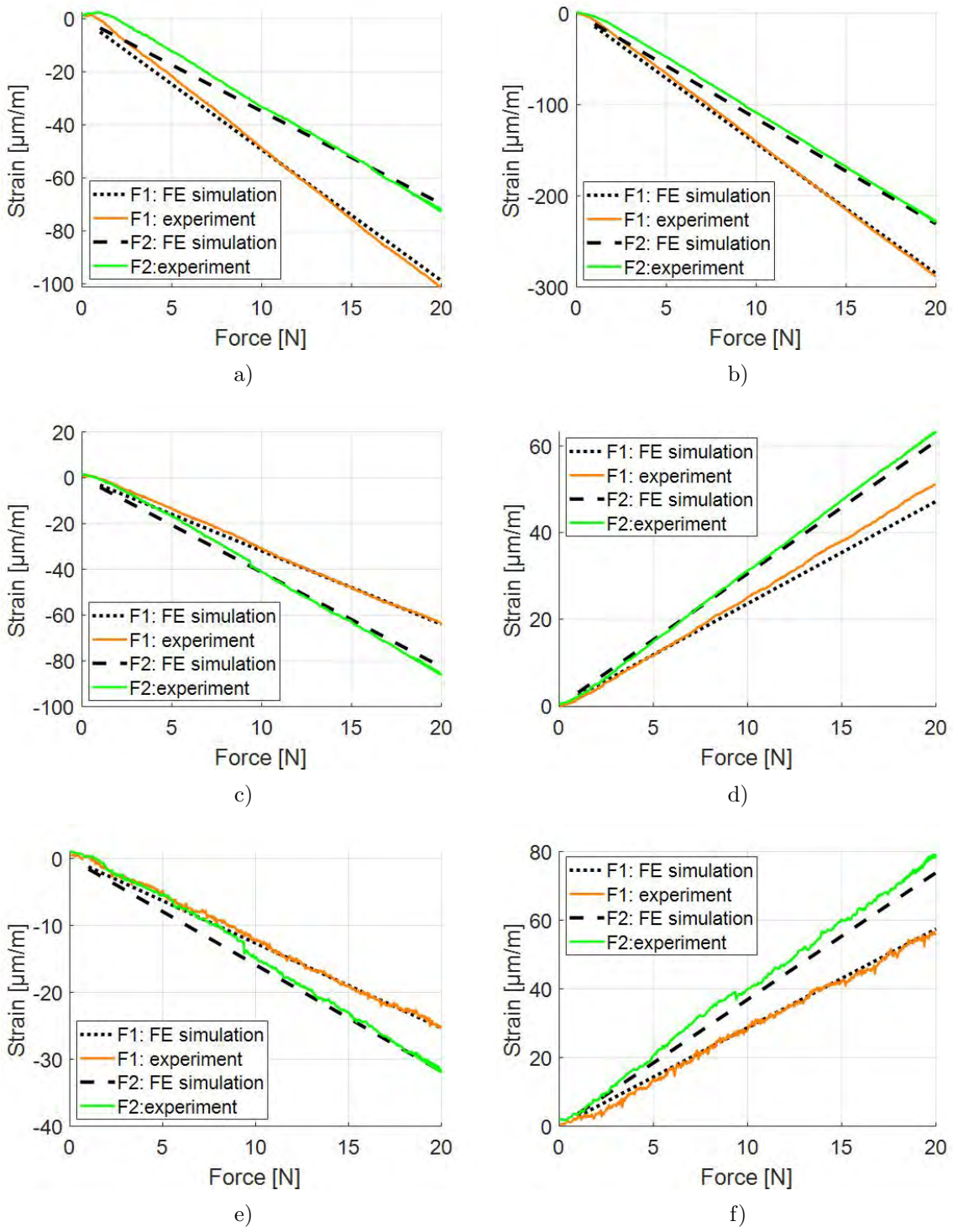


Figure 8: Comparison of strains obtained from experiments and numerical simulations: a) S1, b) S2, c) S3, d) S4, e) S5, f) S6

4 CONCLUSIONS

As one can see in Figures 8 and 9 and in Table 2, quite good agreements between numerical and experimental results were obtained. The finite element model is of satisfying accuracy. Minor differences between computed and measured strains can result from measurement accuracy of strain gages, material imperfections of the prototype panel (related to manufacturing process), minor geometrical differences between model and prototype.

The obtained results are for the authors a starting point for future research on real-time monitoring system. The idea is a cyber-physical system (CPS, system where computing unit controls physical components) [15], whose goal is real-time monitoring of aerostructures in order to predict possible damage, current loading state, or life span of the structure, based on strain measurements in selected areas. The number of critical points, where measurements should be taken, may be sometimes very high. In order to decrease the number of sensors, artificial intelligence techniques will be introduced. Artificial neural networks (ANNs) or deep learning networks (DLNs) will be trained, based on FE model of the structure, to give information on the whole structure, based on measurement data of only few points. The necessary condition is a high fidelity numerical model, that provides very similar data as the real object, like the model presented in the paper. Real-time computations using ANNs could be even performed in the microcontroller [16] on which the CPS is built.

REFERENCES

[1] G.-H. Kim, J.-H. Choi, and J.-H. Kweon, "Manufacture and performance evaluation of the composite hat-stiffened panel," *Composite Structures*, vol. 92, no. 9, pp. 2276–2284, 2010, fifteenth International Conference on Composite Structures.

[2] Y. M. Tang, A. F. Zhou, and K. C. Hui, "Comparison of fem and bem for interactive object simulation," *Computer-Aided Design*, vol. 38, no. 8, pp. 874–886, 2006.

[3] B. Zalewski and B. Bednarcyk, "Act payload shroud structural concept analysis and optimization," National Aeronautics and Space Administration, Cleveland, Ohio, USA, Tech. Rep., 2010, nASA/TM—2010-216942.

[4] K. Pravallika and M. Yugender, "Structural evaluation of aircraft stiffened panel," *International Journal of Science and Research*, vol. 5, no. 10, pp. 753–759, 2016, paper ID: ART20162160.

[5] M. P. Arunkumar, J. Pitchaimani, K. V. Gangadharan, and M. C. Lenin Babu, "Influence of nature of core on vibro acoustic behavior of sandwich aerospace structures," *Aerospace Science and Technology*, vol. 56, pp. 155–167, 2016.

[6] E. P. Carden and P. Fanning, "Vibration based condition monitoring: A review," *Structural Health Monitoring*, vol. 3, no. 4, pp. 355–377, 2004.

[7] M. Mitra and S. Gopalakrishnan, "Guided wave based structural health monitoring: A review," *Smart Materials and Structures*, vol. 25, no. 5, p. 053001, 2016.

[8] W. Fan and P. Qiao, "Vibration-based damage identification methods: A review and comparative study," *Structural Health Monitoring*, vol. 10, no. 1, pp. 83–111, 2011.

[9] N. Aldridge, P. Foote, and I. Read, "Operational load monitoring for aircraft & maritime applications," *Strain*, vol. 36, no. 3, pp. 123–126, 2008.

[10] A. Kurnyta, W. Zielinski, P. Reymer, and M. Dziendzikowski, "Operational load monitoring system implementation for su-22um3k aging aircraft," in *Structural Health Monitoring 2017*, Stanford, California, USA, 2017.

[11] V. Giurgiutiu, "Shm of fatigue degradation and other in-service damage of aerospace composites," in *Structural Health Monitoring of Aerospace Composites*, V. Giurgiutiu, Ed. Oxford: Academic Press, 2016, ch. 10, pp. 395–434.

[12] S. Willis, "Olm: A hands-on approach," in *ICAF 2009, Bridging the Gap between Theory and Operational Practice*, M. J. Bos, Ed. Dordrecht: Springer Netherlands, 2009, pp. 1199–1214.

[13] D. C. Betz, W. J. Staszewski, G. Thursby, and B. Culshaw, "Multi-functional fibre bragg grating sensors for fatigue crack detection in metallic structures," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 220, no. 5, pp. 453–461, 2006.

[14] S. Lecler and P. Meyrueis, "Intrinsic optical fiber sensor," in *Fiber Optic Sensors*, M. Yasin, S. W. Harun, and H. Arof, Eds. Rijeka: IntechOpen, 2012, ch. 3, pp. 53–76.

[15] W. Kuś and W. Mucha, "Memetic inverse problem solution in cyber-physical systems," in *Advances in Technical Diagnostics*, A. Timofiejczuk, B. E. Łazarz, F. Chaari, and R. Burdzik, Eds. Cham: Springer International Publishing, 2018, vol. 10, pp. 335–341.

[16] W. Mucha, "Real-time finite element simulations on arm microcontroller," *Journal of Applied Mathematics and Computational Mechanics*, vol. 16, no. 1, pp. 109–116, 2017.

ACKNOWLEDGEMENT

The research was partially funded from financial resources from the statutory subsidy of the Faculty of Mechanical Engineering, Silesian University of Technology, in 2021.

W.M. acknowledges the National Agency for Academic Exchange of Poland (under the Academic International Partnerships program, grant agreement PPI/APM/2018/1/00004) for supporting training in the University of Minho, which enabled execution of the study.

Model-order reduction for nonlinear dynamics including nonlinearities induced by damage

Alexandre Daby-Seesaram*†, Amélie Fau*, Pierre-Étienne Charbonnel† and David Néron*

* Université Paris-Saclay, ENS Paris-Saclay, CNRS, LMT
Laboratoire de Mécanique et Technologie, 91190, Gif-sur-Yvette, France
{alexandre.daby-seesaram, amelie.fau, david.neron}@ens-paris-saclay.fr

† DES - Service d'Études Mécaniques et Thermiques (SEMT), CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France
pierreetienne.charbonnel@cea.fr

Key words: LATIN-PGD, model-order reduction, damage, fragility curves, seismic risk assessment

Abstract: *Assessing the probability of failure of a structure under seismic loading requires the simulation of a great number of similar nonlinear computations. A model-order reduction strategy is proposed for decreasing the computational cost associated to each nonlinear simulation. In this contribution, the method is illustrated to evaluate the damage evolution in a primary circuit piping component of a pressurized water reactor, subjected to accidental seismic input. Piping components are described with a damageable elasto-plastic material exhibiting a preliminary damage pattern.*

1 INTRODUCTION

Fragility curves are one of the main tools for characterizing the resistance of civil engineering structures, such as nuclear facilities, to seismic hazard. These curves describe the probability that the response of a structure exceeds a given criterion, called "failure criterion", as a function of the expected seismic loading level. Their computational cost is expensive as a large number of loading scenarii must be considered to model seismic input variability, but also due to the inherent uncertainties (material parameters, geometry, modelling errors, etc.) that must be taken into account for reliability assessment. Their construction therefore falls into the scope of the *many-queries* problems where the need to reduce the numerical cost of each simulation is imperative. Another point is that the final aim of the study is to add a preliminary structural damage as a parameter of those charts. Decreasing the computational costs of solving large dimensional problems has long been studied and decreasing the dimension of the solution space has shown to be of great interest. Among the several existing methods, one finds *Model-Order Reduction* (MOR) techniques. Some of them (referred to as *a posteriori* methods) require beforehand the computation of a given reduced basis, while others (referred to as *a priori* methods) consist in building the reduced basis simultaneously with the computation. The first kind, including among others the use of Ritz vectors [1] or the Proper Orthogonal Decomposition (POD) [2], has greatly been studied in [3] where a wide range of reduced basis choices have been examined. In this last reference, computation time saving and robustness of the basis considered are looked over. It highlights that the choice of the basis proves to be decisive and when dealing with numerous computations, finding an ideal reduced basis on which to project the solutions to these various problems may not be an obvious task. To overcome this difficulty it is relevant to build the reduced basis on-the-fly, as the solver progresses, to optimize the choice of new modes. Such *a priori* model-order reduction methods include the *Proper Generalized Decomposition* (PGD) [4] which is used in the present work.

https://doi.org/10.4995/YIC2021.2021.13255

Herein, the focus is on the implementation of a strategy based on *a priori* model-order reduction for the calculation of the nonlinear dynamics problem at stake. Among the different possible approaches, the PGD coupled with the LATIN method [5] is particularly well suited for solving parameterized problems in nonlinear mechanics in order to build numerical charts [6]. The LATIN-PGD method is an iterative approach that seeks the solution of a given problem by building, in a greedy way, a dedicated reduced-order basis. This basis can be reused and enriched, allowing a good numerical efficiency. It has been applied to solve a wide range of problems in mechanics (and more recently for earthquake-engineering applications [7]). Here we develop this method to solve the low-frequency dynamics problem that arises when applying seismic loading to metallic piping structures with a nonlinear behaviour while taking into account their possibly pre-damaged state.

2 DYNAMIC EQUATIONS

The spatial domain on which the problem is written is denoted Ω. On that body of density ρ, body forces $\boldsymbol{f}_d$ and surface forces $\boldsymbol{F}_d$ are applied on Ω and on $\partial\Omega_2$ respectively while imposed displacements u_d are applied on $\partial\Omega_1$ as described in Fig. 1.

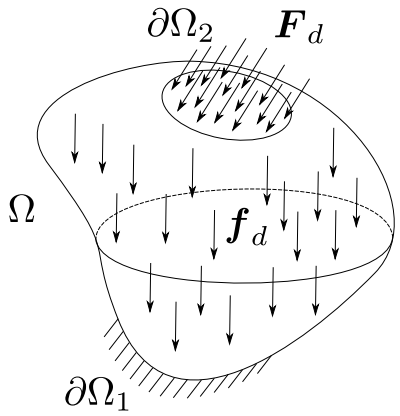


Figure 1: Reference problem on the domain Ω

Let us then define the three sets defining *admissibility*:

- $\mathcal{U} = \Big\{ \boldsymbol{u} \quad | \quad \boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left(\boldsymbol{\nabla}\ \boldsymbol{u} +^T \boldsymbol{\nabla}\ \boldsymbol{u}\right),\ \dot{\boldsymbol{u}}|_{t=0} = \boldsymbol{0},\ \boldsymbol{u}|_{t=0} = \boldsymbol{0} \text{ in } \Omega,$
$\boldsymbol{u} = \boldsymbol{u}_d \text{ on } \partial\Omega_1 \text{ and } \dot{\boldsymbol{u}} = \dot{\boldsymbol{u}}_d \text{ on } \partial\Omega_1 \big\},$
- $\mathcal{U}^0 = \Big\{ \boldsymbol{u} \quad | \quad \boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left(\boldsymbol{\nabla}\ \boldsymbol{u} +^T \boldsymbol{\nabla}\ \boldsymbol{u}\right),\ \dot{\boldsymbol{u}}|_{t=0} = \boldsymbol{0},\ \boldsymbol{u}|_{t=0} = \boldsymbol{0} \text{ in } \Omega,$
$\boldsymbol{u} = \boldsymbol{0} \text{ on } \partial\Omega_1 \text{ and } \dot{\boldsymbol{u}} = \boldsymbol{0} \text{ on } \partial\Omega_1 \big\},$
- $\mathcal{S} = \{\boldsymbol{\sigma} \quad | \quad \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + \boldsymbol{f}_d = \rho\ddot{\boldsymbol{u}} \text{ in } \Omega \text{ and } \boldsymbol{\sigma}\ \boldsymbol{n} = \boldsymbol{F}_d \text{ on } \partial\Omega_2\},$

$\mathcal{U}$ (respectively $\mathcal{U}^0$) is the kinematically admissible (respectively to zero) displacements set and $\mathcal{S}$ is the dynamically admissible stress set. $\boldsymbol{\sigma}$ is Cauchy's stress tensor while $\boldsymbol{\varepsilon}$ is the strain tensor. One then needs to find admissible displacement and stress fields $s = (\boldsymbol{u}, \boldsymbol{\sigma}) \in \mathcal{U} \times \mathcal{S}$ that also satisfy the constitutive relations.

The weak formulation of the dynamic equilibrium then reads

$$-\int_{\Omega\times I} \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{u}^*)\mathrm{d}\Omega\mathrm{d}t + \int_{\Omega\times I} \boldsymbol{f_d}\cdot\boldsymbol{u}^*\mathrm{d}\Omega\mathrm{d}t + \int_{\partial\Omega\times I} \boldsymbol{F_d}\cdot\boldsymbol{u}^*\mathrm{d}S\mathrm{d}t = \int_{\Omega\times I} \rho\ \ddot{\boldsymbol{u}}\cdot\boldsymbol{u}^*\mathrm{d}\Omega\mathrm{d}t \quad \forall \boldsymbol{u}^* \in \mathcal{U}^0. \quad (1)$$

In addition to this dynamic equation, the material behaviour of the structure is described through nonlinear equations which motivates the methodology introduced in this work.

3 DUCTILE DAMAGE MODEL INCLUDING CRACK CLOSURE EFFECT

The damage evolution in the structure is governed by a plastic model [8] with linear kinematic and isotropic hardening along with isotropic damage contribution [9]. In order to account for crack-closure effect, an effective stress tensor $\tilde{\boldsymbol{\sigma}}$ [10] is introduced to read

$$\tilde{\boldsymbol{\sigma}} = \frac{\boldsymbol{\sigma}_d}{1-D} + \left[\frac{\langle \sigma_H \rangle}{1-D} - \langle -\sigma_H \rangle\right] \mathbf{1}, \tag{2}$$

with $\boldsymbol{\sigma}_d$ the deviatoric part of Cauchy's stress $\boldsymbol{\sigma}$ and σ_H its hydrostatic part, D the damage variable, $\mathbf{1}$ the identity tensor and $\langle \square \rangle = \max(\square, 0)$ defining the positive part. Doing so leads to Hooke's relation between stress and elastic strain reading

$$\tilde{\boldsymbol{\sigma}} = \mathbb{K} : \boldsymbol{\varepsilon}^e \tag{3}$$

with $\mathbb{K}$ the Hooke's tensor. Thus, the damage variable is no more explicitly apparent in the elastic constitutive relation.

When the solicitation is high, some non reversibilities appear and plastic laws as well as new variables are required. The yield function $f_p(\boldsymbol{\sigma})$ describes the elastic domain. When the stress is small enough for the function to be negative then the material follows an elastic behaviour but when the yield function increases to the point that it reaches zero, non reversibilities appear and plasticity laws become necessary. The plasticity yield function f_p verifies

$$f_p \leq 0, \tag{4}$$

and is written using von Mises equivalent stress $J_2(\square)$ as follows

$$f_p = J_2\left(\frac{\boldsymbol{\sigma}}{1-D} - \boldsymbol{X}\right) - \sigma_y - R \tag{5}$$

with σ_y the yield stress, R the isotropic hardening variable and $\boldsymbol{X}$ the kinematic hardening tensor.

The model chosen to describe those irreversibilities, following the lines of [8], involves both kinematic and isotropic hardening. The elastic strain tensor reads $\boldsymbol{\varepsilon}^e = \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^p$ with $\boldsymbol{\varepsilon}^p$ being the plastic strain. The plasticity level is described by an internal variable called the cumulative plastic strain p which is a strictly increasing quantity. State equations relative to linear hardening read

$$\begin{cases} R = hp, \\ \boldsymbol{X} = \frac{2}{3} C \boldsymbol{\alpha} \end{cases} \tag{6}$$

with h the rate at which the isotropic hardening increases, C a material coefficient and $\boldsymbol{\alpha}$ the kinematics internal variable.

Lemaitre's damage evolution law reads as a function of the *elastic energy density* Y defined as

$$Y = \frac{1}{2}\boldsymbol{\varepsilon}^e : \mathbb{K} : \boldsymbol{\varepsilon}^e = R_\nu \frac{J_2(\tilde{\boldsymbol{\sigma}})^2}{2E}, \tag{7}$$

where the triaxiality function $R_\nu = \frac{2}{3}(1+\nu) + 3(1-2\nu)\langle \frac{\tilde{\sigma}_H}{J_2(\tilde{\boldsymbol{\sigma}})} \rangle^2$ is introduced.

In order to predict the temporal evolution of these quantities, evolution laws are needed. Plasticity evolution laws are derived using normality rule, leading to

$$\begin{cases} \dot{\boldsymbol{\varepsilon}}^p = \dot{p} \frac{3}{2} \frac{(\tilde{\boldsymbol{\sigma}} - \boldsymbol{X})_d}{J_2(\tilde{\boldsymbol{\sigma}} - \boldsymbol{X})}, \\ \dot{\boldsymbol{\alpha}} = \dot{p}(1-D) \left[\frac{3}{2} \frac{(\tilde{\boldsymbol{\sigma}} - \boldsymbol{X})_d}{J_2(\tilde{\boldsymbol{\sigma}} - \boldsymbol{X})}\right]. \end{cases} \tag{8}$$

As for the damage variable, the evolution is written as

$$\dot{D} = \dot{p}\left(\frac{Y}{S}\right)^{s}, \text{ if } W_s > W_D, \tag{9}$$

with W_s the so-called *corrected stored energy* [9], W_D a given energy threshold, s and S material parameters. This set of equations allows to describe finely the damage state evolution of the structure but the resulting problem is nonlinear and involves a large number of degrees of freedom.

4 THE LATIN-PGD

Using the finite element method, the mechanical problem gives a detailed description of the evolution of the quantities of interest but its computation requires solving nonlinear equations at each Gauss point at each given time step. That leads to a high computational cost that could be driven down by using model-order reduction techniques.

4.1 The PGD method

For solving linear problems, the idea of the PGD technique is to look for the solution as the sum of products of single-variable functions. Thus, a displacement field u is approximated by $u_N(\boldsymbol{x}, t)$ reading

$$u(\boldsymbol{x}, t) \approx u_N(\boldsymbol{x}, t) = \sum_{i=1}^{N} \overline{u}_i(\boldsymbol{x})\lambda_i(t). \tag{10}$$

The reduced basis $\{\overline{u}_i\}$ is not a priori known and is built during the computation using a greedy algorithm. New modes are added on the fly. In order to perform a PGD, the used greedy algorithm requires that one solves a linear space-time problem. Hence having a method turning the nonlinear problem into solving linear equations on such a domain is mandatory.

4.2 The LATIN solver

The LATIN method, first introduced in [11] has been singled out as it is an iterative non incremental solver that allows seeking a solution on the entire space-time domain while some of the computations involve linear equations. Each LATIN iteration is decomposed in so-called *local* and *global* stages. At the local stage, the nonlinear part of the constitutive relations is solved at each Gauss point and at the global stage, admissibility is imposed on the whole time-space domain leading to a linear problem. The PGD can be used for an efficient computation of the solution at the linear stage. Those solutions belong respectively to the manifold Γ gathering solutions of the nonlinear equations and the manifold $\mathcal{A}_d$ gathering solutions of the linear equations. The final solution s_{exact}, which is naturally found at the intersection of these two manifolds, is thought alternatively in both spaces $\mathcal{A}_d$ and Γ involving two search directions $\mathbb{H}^+$ and $\mathbb{H}^-$ linking the manifolds through Eq. (11),

$$\begin{cases} \left(\boldsymbol{\sigma}_{n+1} - \hat{\boldsymbol{\sigma}}_{n+1/2}\right) - \mathbb{H}^- \left(\boldsymbol{\varepsilon}_{n+1} - \hat{\boldsymbol{\varepsilon}}_{n+1/2}\right) = \mathbf{0}, \\ \left(\hat{\boldsymbol{\sigma}}_{n+1/2} - \boldsymbol{\sigma}_{n}\right) + \mathbb{H}^+ \left(\hat{\boldsymbol{\varepsilon}}_{n+1/2} - \boldsymbol{\varepsilon}_{n}\right) = \mathbf{0}. \end{cases} \tag{11}$$

This iterative scheme can be sketched by Fig. 2 where $\hat{s}_{n+1/2}$ is the solution belonging to Γ and s_{n+1} is a solution of $\mathcal{A}_d$, both computed at the $(n+1)^{\text{th}}$ stage of the method.

The solution can be initialized to a kinematically and dynamically admissible elastic solution. Then one loops over finding alternatively a solution in Γ and in $\mathcal{A}_d$ until a stopping criterion is reached. Such a criterion is satisfied when an error indicator

$$\eta^2 = \frac{\|\hat{s}_{n+1/2} - s_{n+1}\|^2}{1/2\|s_{n+1}\|^2 + 1/2\|\hat{s}_{n+1/2}\|^2}, \tag{12}$$

based on the distance between two consecutive solutions, is lower than a chosen threshold. The norm $\|s\|$ is defined as

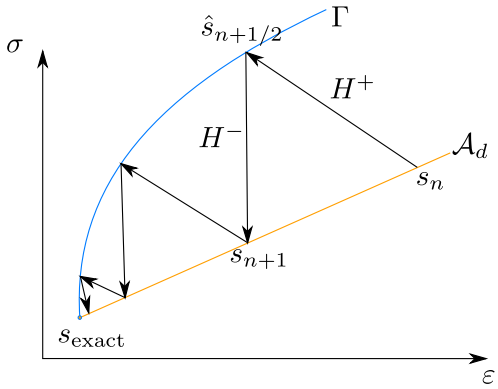


Figure 2: Working principle of the LATIN method, modified from [11]

$$\|s\|^2 = \int_{\Omega\times I} \boldsymbol{\varepsilon} : \mathbb{K} : \boldsymbol{\varepsilon}\ \mathrm{d}\Omega\mathrm{d}t + \int_{\Omega\times I} \boldsymbol{\sigma} : \mathbb{K}^{-1} : \boldsymbol{\sigma}\ \mathrm{d}\Omega\mathrm{d}t. \tag{13}$$

Such a methodology has recently been suggested for a dynamic resolution [7] where the material is considered to be described by a visco-plastic behaviour without considering damage evolution.

4.2.1 The global stage

The global stage consists in finding a solution in $\mathcal{A}_d$ which means solving the dynamic equilibrium defined by Eq. (1). Subtracting that equation written in two successive steps of the LATIN method gives the admissibiblity equation written in corrective terms reading

$$-\int_{\Omega\times I} \Delta\boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{u}^*)\mathrm{d}\Omega\mathrm{d}t = \int_{\Omega\times I} \rho\boldsymbol{\Delta\ddot{u}} \cdot \boldsymbol{u}^*\mathrm{d}\Omega\mathrm{d}t \quad \forall \boldsymbol{u}^* \in \mathcal{U}^0, \tag{14}$$

with $\Delta\square = \square^{n+1} - \square^n$.

To solve that equation the descending search direction given by Eq. (11) is injected in the latter, leading to,

$$\begin{aligned}\int_{\Omega\times I} \mathbb{H}^- : \boldsymbol{\varepsilon}(\boldsymbol{\Delta u}) : \boldsymbol{\varepsilon}(\boldsymbol{u}^*)\mathrm{d}\Omega\mathrm{d}t + \int_{\Omega\times I} \rho\Delta\ddot{\boldsymbol{u}} \cdot \boldsymbol{u}^*\mathrm{d}\Omega\mathrm{d}t \\ = \int_{\Omega\times I} \underbrace{\left[\left(\boldsymbol{\sigma}^n - \hat{\boldsymbol{\sigma}}^{n+1/2}\right) - \mathbb{H}^- : \left(\boldsymbol{\varepsilon}^n - \hat{\boldsymbol{\varepsilon}^{n+1/2}}\right)\right]}_{-\hat{\boldsymbol{f}}} : \boldsymbol{\varepsilon}(\boldsymbol{u}^*)\mathrm{d}\Omega\mathrm{d}t\ \forall \boldsymbol{u}^* \in \mathcal{U}^0.\end{aligned} \tag{15}$$

One may notice that terms in the second hand $\hat{\boldsymbol{f}}$ of that equation are already known quantities. The displacement field is the only unknown.

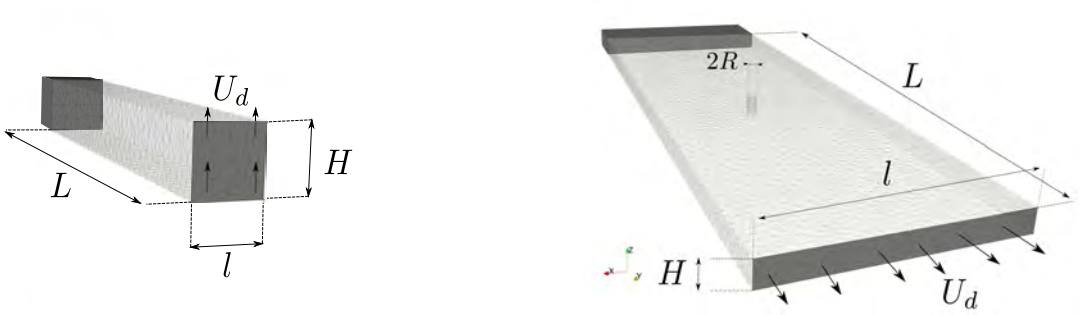
Because the global stage consists in solving linear equations over the whole time-space domain, a greedy algorithm can advantageously be set up in order to find the solution under a PGD form. To do so, the PGD decomposition $\square(\boldsymbol{x}, t) = \sum_{i=1}^{N} \overline{\square}^i(\boldsymbol{x})\lambda^i(t)$ of the displacement field is injected into the previous equation to compute the corrections, N being the number of PGD modes used to describe the solution.

4.2.2 The local stage

The local stage consists in solving the local and possibly nonlinear equations of the problem. That means finding $\hat{s}_{n+1/2} \in \Gamma$ knowing $s_n \in \mathcal{A}_d$. The ascendant search direction is chosen vertical, *i.e.* $\hat{\boldsymbol{\varepsilon}}_{n+1/2} = \boldsymbol{\varepsilon}_n$. Technically, the local stage consists in a radial feedback algorithm to compute the plastic and damage evolution of the structure while taking into account normality law and the von Mises criterion defined by Eq. (4).

5 NUMERICAL RESULTS

In order to illustrate the method, two cases are investigated. First the dynamic aspects of the problem are exposed with a 3D beam under flexion loading. Then a pre-damaged 3D plate with a hole is investigated with various initial damage states. Both geometries are described in Fig. 3 and their dimensions are summarized in Table 1.



(a) Geometry of the beam (b) Geometry of the plate

Figure 3: Two test cases geometries

Table 1: Dimensions of the geometries

Geometry	L	l	H	$U_d^{\max}$	R
Plate	60 mm	20 mm	2 mm	2 mm	1 mm
Beam	40 mm	8 mm	8 mm	5 mm	-

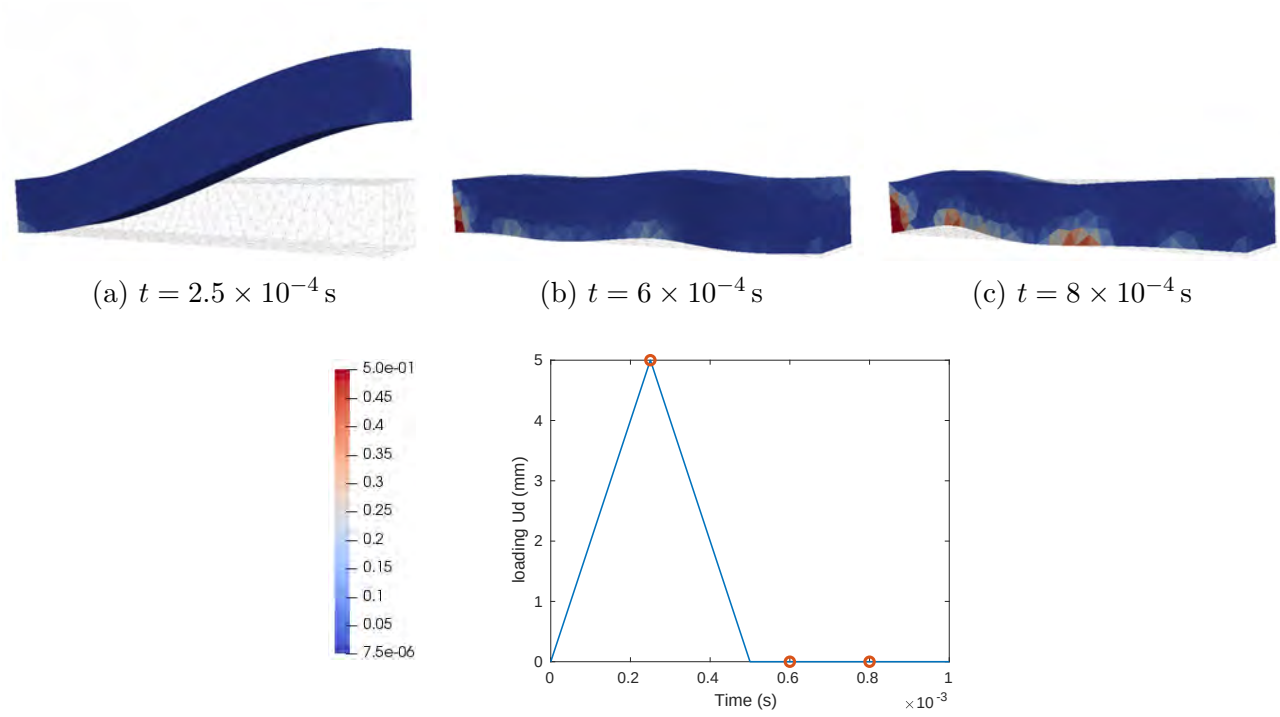
The material parameters are given in Table 2.

Table 2: Material parameters

Name	Parameters
Young's modulus	$E = 200\,\text{GPa}$
Poisson's ratio	$\nu = 0.3$
Kinematic hardening modulus	$C = 2.21 \times 10^4\,\text{MPa}$
Yield stress	$\sigma_y = 200\,\text{MPa}$
Isotropic hardening ratio	$h = 0\,\text{MPa}$
Damage law exponent	$s = 2$
Parameter for damage evolution	$S = 0.6\,\text{MPa}$
Density	$\rho = 7900\,\text{kg/m}^3$
Damage threshold energy	$W_D = 0\,\text{Jm}^3/\text{kg}$

5.1 Dynamic behaviour

A cantilever beam loaded by an imposed vertical displacement U_d at its end, as shown in Fig. 3a, is studied. The beam is submitted to a triangular load for the first half of the simulation then the displacement at the end of the beam is kept equal to zero for the second half. From an initial undamaged state, the damage increases along the beam. The damage maps at $t = 2.5 \times 10^{-4}\,\text{s}$, $t = 6 \times 10^{-4}\,\text{s}$ and $t = 8 \times 10^{-4}\,\text{s}$ are presented in Fig. 4a, 4b and 4c respectively. It can be noted that the first instant corresponds to the maximum amplitude of the external perturbation while the two other instants of interest are posterior to the external load, as shown in Fig. 4d.



(a) $t = 2.5 \times 10^{-4}\,\mathrm{s}$ (b) $t = 6 \times 10^{-4}\,\mathrm{s}$ (c) $t = 8 \times 10^{-4}\,\mathrm{s}$

(d) Loading

Figure 4: Evolution of the damage map in the beam

One may notice that, even though the last two showcased results (Fig. 4b and 4b) are taken when there is no more external loading, the damage map keeps on progressing. This evolution is therefore only due to inertial forces because waves propagate through the structure as observed in Fig. 4b and Fig. 4c.

The convergence of the LATIN-PGD implementation leading to this result is plotted in Fig. 5 which shows the evolution of the error indicator η with respect to the number of PGD modes. It may be noted that the number of modes is rather large at convergence. Indeed, currently the global stage of the method only consists in adding a PGD mode but a more efficient strategy would be to first update the temporal modes associated with the existing spatial modes and only add a supplementary PGD mode if that update did not prove to be effective enough.

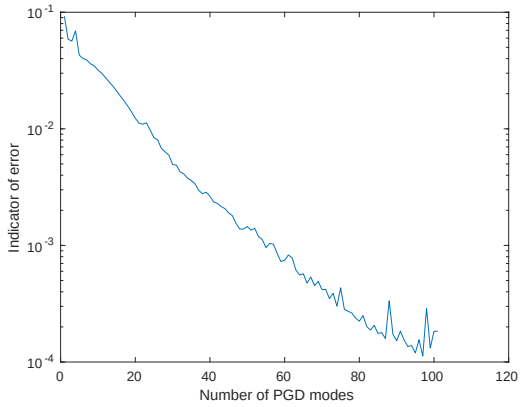
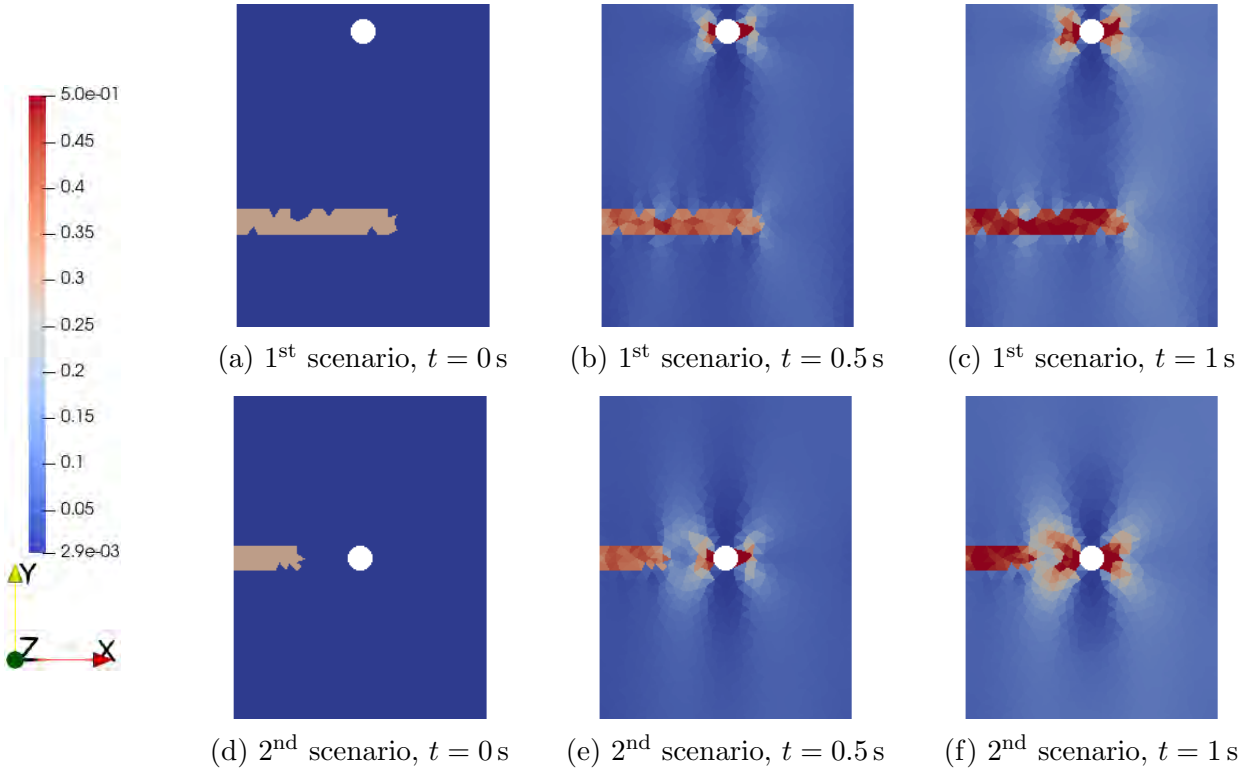


Figure 5: Evolution of the indicator of error for the beam scenario

5.2 Influence of the initial damage state

As previously stated our long term ambition is the construction of virtual charts in which the pre-damage is a parameter. To illustrate the role of a pre-damaged zone on the final solution two simulations with distinct initial damage states have been carried out. The common structure is a plate shown in Fig. 3a and the loading is a 1 s ramp loading directed along the y-axis. The first case scenario (illustrated in Fig. 6a, 6b and 6c) shows the structure with a pre-damaged zone below the hole while the second case scenario (illustrated in Fig. 6d, 6e and 6f) shows the structure with a pre-damaged zone facing the hole. Only a part of the whole structure is shown as to focus on the damaged zones, which are of interest.

When observing the damaged maps projected on the deformed structure in Fig. 6, one can see that damage tends to grow in the surrounding area of the initial damaged zone and near the hole. In the second case scenario similarly damage increases near the hole and the pre-damaged zone first. But a coalescence arises between those two zones. A significant influence of pre-damage zones is therefore observed.



(a) 1st scenario, $t = 0\,\mathrm{s}$ (b) 1st scenario, $t = 0.5\,\mathrm{s}$ (c) 1st scenario, $t = 1\,\mathrm{s}$

(d) 2nd scenario, $t = 0\,\mathrm{s}$ (e) 2nd scenario, $t = 0.5\,\mathrm{s}$ (f) 2nd scenario, $t = 1\,\mathrm{s}$

Figure 6: Evolution of the damage map in the plate

As it has been done for the plate, the convergence of the method is plotted in Fig. 7 which shows the evolution of the error indicator η while the number of PGD modes increases. The previous remark about the number of modes at convergence remains valid as a great number of modes is needed here too.

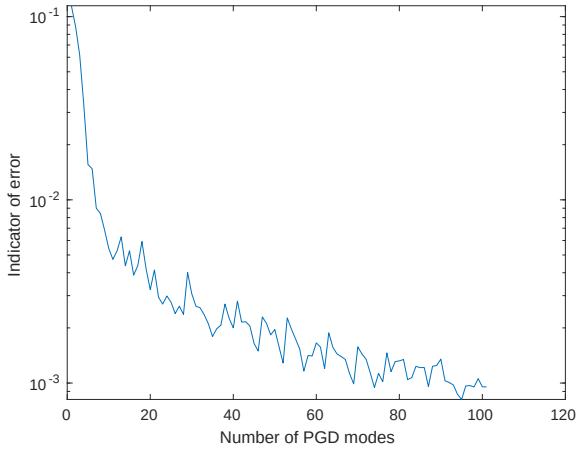


Figure 7: Evolution of the indicator of error for the plate scenario

A few other cases have been implemented showing similar convergence curves, proving the methodology robust enough to investigate a vast variety of scenarii.

6 CONCLUSIONS

The LATIN-PGD framework has been presented for damageable materials in dynamics. Predicting the damage evolution of a plastic structure under a dynamic loading is possible and the computation of the solution gives access to a PGD basis. The next step will be to implement the update strategy in the global stage in order to be able to take advantage of previously computed spatial modes. The LATIN-PGD methodology will then provide a favorable framework for the computation of fragility curves where the seismic performances of quasi-identical structures have to be computed for a family of similar inputs defining the seismic risk. Both initialization of the solution and re-use of reduced order PGD basis enable one to take full advantage of the possible redundancy contained in those virtual charts.

7 ACKNOWLEDGMENT

The SEISM Institute is deeply acknowledged for funding this research activity. This work is hosted by the NARSIS Project that is also thanked for giving the opportunity to study such thematic.

REFERENCES

[1] T.D Burton and W. Rhee. On the reduction of nonlinear structural dynamics models. *Journal of vibration and control*, 6(4):531–556, 2000.

[2] M. Kirby, J.-P. Boris, and L. Sirovich. A proper orthogonal decomposition of a simulated supersonic shear layer. *International journal for numerical methods in fluids*, 10(4):411–428, 1990.

[3] F. A. Lülf, D.-M. Tran, and R. Ohayon. Reduced bases for nonlinear structural dynamic systems: A comparative study. *Journal of sound and vibration*, 332(15):3897–3921, 2013.

[4] F. Chinesta, P. Ladevèze, and E. Cueto. A short review on model order reduction based on proper generalized decomposition. *Archives of Computational Methods in Engineering*, 18(4):395–404, 2011.

[5] P. Ladevèze. *Nonlinear computational structural mechanics: new approaches and non-incremental methods of calculation*. Mechanical engineering series. Springer, New York, 1999.

[6] D. Néron, P.-A. Boucard, and N. Relun. Time-space PGD for the rapid solution of 3D nonlinear parametrized problems in the many-query context. *International Journal for Numerical Methods in Engineering*, 103(4):275–292, 2015.

[7] S. Rodriguez, D. Néron, P.-E. Charbonnel, P. Ladevèze, and G. Nahas. Non incremental LATIN-PGD solver for nonlinear vibratoric dynamics problems. In *14ème Colloque National en Calcul des Structures, CSMA 2019*, Presqu'Île de Giens, France, May 2019.

[8] J. Lemaitre and J.-L. Chaboche. Mechanics of solid materials. *Cambridge university press*, 1994.

[9] J. Lemaitre and R. Desmorat. *Engineering Damage Mechanics: Ductile, Creep, Fatigue and Brittle Failures*. Springer Berlin / Heidelberg, 2005.

[10] M. Bhattacharyya, A. Fau, R. Desmorat, S. Alameddin, D. Néron, P. Ladevèze, and U. Nackenhorst. A kinetic two-scale damage model for high-cycle fatigue simulation using multi-temporal latin framework. *European Journal of Mechanics / A Solids*, 77, 2019.

[11] P. Ladevèze. Sur une famille d'algorithmes en mécanique des structures. *Comptes-rendus des séances de l'Académie des sciences. Série 2, Mécanique, physique, chimie, sciences de l'univers, sciences de la terre*, 300(2), 1985.

PHD OLYMPIADS

Block strategies to compute the lambda modes associated with the neutron diffusion equation

A. Carreño*, A. Vidal-Ferràndiz†, D. Ginestar† and G. Verdú*

* Instituto Universitario de Seguridad Industrial, Radifísica y Medioambiental (ISIRYM)
Universitat Politècnica de València
Valencia, Spain
e-mail: amcarsan@iqn.upv.es, gverdu@iqn.upv.es

† Instituto Universitario de Matemática Multidisciplinar (IMM)
Universitat Politècnica de València
Valencia, Spain
e-mail: anvifer2@imm.upv.es, dginesta@mat.upv.es

Key words: Block eigenvalue solvers, Neutron reactor system, Finite element method

Abstract: *Given a configuration of a nuclear reactor core, the spatial distribution of the power can be approximated by solving the λ-modes problem associated with the multigroup neutron diffusion equation. It is a partial generalized eigenvalue problem whose dominant eigenvalue characterises the criticality of the reactor and its associated eigenvector represents the distribution of the neutron flux in steady-state. The spatial discretization of the equation is made by using a continuous Galerkin high order finite element method. Usually, the matrices obtained from the discretization are huge and sparse. Moreover, they have a block structure given by the different number of energy groups. In this work, block strategies are developed to optimize the computation of the associated eigenvalue problems. First, different block eigenvalue solvers are studied. On the other hand, the convergence of these iterative methods mainly depends on the initial guess and the preconditioner used. In this sense, different multilevel techniques to accelerate the rate of convergence of this problem are proposed. A large three-dimensional benchmark shows the efficiency of the methodology proposed.*

1 INTRODUCTION

The computation of the dominant λ-modes associated with the neutron diffusion equation has an interest in nuclear engineering to study the criticality of a reactor and also to develop modal methods to integrate the time dependent equation. This equation is an approximation of the neutron transport equation that assumes that the neutron current is proportional to the gradient of the scalar neutron flux with a diffusion coefficient.

The λ-modes problem is discretized to yield a large algebraic generalized eigenvalue problem that has to be solved by using iterative methods to compute its dominant eigenvalues and their corresponding eigenvectors. In this work, a high order finite element method has been used for the spatial discretization of the λ-modes problem.

Krylov subspace methods, such as Arnoldi or Krylov-Schur method, can be applied to solve this non-symmetric eigenvalue problem [10, 12]. However, these iterative methods require reducing the generalized eigenvalue problem to an ordinary problem, and it implies solving many linear systems. Other methods to solve eigenvalue problems associated with nonsymmetric matrices are the gradient type methods, such as the Generalized Davidson method, that do not require solving linear systems involving the full operator. However, if there are clustered or degenerate eigenvalues, these methods may have problems to find all the eigenvalues.

In this work, a hybrid method is proposed that combines two types of solvers, the block inverse-free Arnoldi method (BIFPAM) and the modified generalized block Newton method (MGBNM). The BIFPAM was proposed in [8] for symmetric problems, but the authors nu-

https://doi.org/10.4995/YIC2021.2021.13470

merically showed that it also converges for this type of neutron problems where the dominant eigenvalues are positive. It does not need to solve linear systems. It improves the traditional steepest descent method by expanding the search direction to a Krylov subspace with the advantage of better approximation properties offered by Krylov subspaces. The MGBNM is a generalization of the modified block Newton method ([7]). It has a quadratic convergence, but it is very sensitive to the initial guess.

The structure of the rest of the paper is as follows. In section 2 the definition of λ-modes problem and the spatial discretization by using a high order finite element method is given. Section 3 briefly describes the eigenvalue solvers. Section 4 presents the multilevel strategy to improve the computational efficiency of the eigenvalue solvers. Section 5 presents the numerical results for the analysis of the methodology in a benchmark problem. Finally, Section 6 collects the main conclusions of the work.

2 THE λ-MODES PROBLEM

Given a configuration of a nuclear reactor core, it is possible to force its criticality dividing the neutron production rate by a positive number, λ, obtaining the known λ-modes problem [9],

$$\mathcal{L}\phi = \frac{1}{\lambda}\mathcal{M}\phi, \tag{1}$$

where $\mathcal{L}$ is the neutron loss operator, $\mathcal{M}$ is the neutron production operator and ϕ the neutron flux.

To solve the problem (1), a spatial discretization of the equations has to be selected. In this work, a high order Galerkin finite element method is used (see [12]) leading to an algebraic eigenvalue problem associated with the discretization of (1) with the following structure,

$$Ax = \lambda Bx\,, \tag{2}$$

where A and B are the matrices that appear from the discretization of $\mathcal{M}$ and $\mathcal{L}$, respectively. The vector x is the algebraic vector of weights associated with the neutron flux. For simplicity, the shape functions used are part of Lagrange finite elements. More details on the spatial discretization used and general boundary conditions can be found in [12]. The finite element method has been implemented using the open-source finite elements library Deal.II [2].

3 BLOCK SOLVERS

In this Section, several well-known eigenvalue solvers to solve the partial eigenvalue problem (2) are described. This list is not intended to be exhaustive, and other eigenvalue solvers appear in the neutron transport computations or the mathematical literature.

In nuclear computations, different strategies have been used to solve the generalized problem obtained from the discretization. First, we can transform the problem as an ordinary eigenvalue problem by using the inverse of B. The inverse of the matrix B is not computed and its product by a vector is applied by solving linear systems. Second, for the special case of the λ-modes and two energy groups, many works define an ordinary eigenvalue problem, but with half of the size of the original problem. Finally, in this work, we aim to apply direct methods for the generalized eigenvalue problem.

For this problem, we are interested in using block methods, that converge the set of eigenvectors in a block, in order to initialize the iterative methods if an initial set of approximated eigenvectors is provided.

Generalized Davidson method Davidson methods take one eigenvector and apply a correction as

$$x^{(i+1)} = x^{(i)} + t^{(i)}, \quad \text{where} \quad (A - \lambda^{(i)}B)t^{(i)} = -(A - \lambda^{(i)}B)x^{(i)}\,. \tag{3}$$

In particular, the generalized Davidson method estimates the correction by solving the problem

$$P_i t^{(i)} = -r^{(i)}, \text{ where } P_i \approx (A - \lambda^{(i)} B). \tag{4}$$

This method, although its convergence is slow, does not need to solve any linear system involving the full operator. This makes that the iterations are very cheap. In this case, the block implementation provided by the library SLEPC is used [6]. As preconditioner for this method, the ILU(0) factorization from the library PETSc is used [1].

Block Inverse-free preconditioned Arnoldi method (BIFPAM) This block method is proposed for symmetric and positive definite matrices [8]. However, we have shown that the convergence is also obtained for neutronic problems where the eigenvalues are positive numbers. Given a set of eigenvectors X_i, the following approximated eigenvectors X_{i+1} are obtained from the first Ritz q-eigenvectors of the small problem

$$Z^T A Z U = Z^T B Z U \Lambda, \text{ as } X_{i+1} = ZU, \tag{5}$$

where Z is a basis of [8]

$$\mathcal{K}_{d_k} := \bigcup_{m=1}^{q} K^i_{d_k,m}(P_{m,i}(A - \lambda_{m,i} B), X_{m,i}).$$

These bases are constructed with the Arnoldi method. It does not solve any linear system and it has a block implementation. However, this method must be preconditioned to improve the convergence. In this work, we test two preconditioners of the linear systems: the ILU(0) preconditioner and the GMG preconditioner [4].

Modified block generalized Newton method (MGBNM) The original method was proposed for ordinary eigenvalue problems and we have proposed two generalizations [4]. In this work, it assumes that a set of eigenvectors X can be decomposed as

$$X = ZQ, \text{ such that } Z^T Z = I_q\,. \tag{6}$$

Now, the Newton's method is applied to solve the non-linear problem

$$F(Z, K) := \begin{pmatrix} AZ - BZK \\ W^T Z - I_q \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{where } K \equiv Q^{-1} \Lambda Q \tag{7}$$

as

$$Z^{(k+1)} = Z^{(k)} - \Delta Z^{(k)}, \qquad K^{(k+1)} = K^{(k)} - \Delta K^{(k)}, \tag{8}$$

where the corrections are given by solving the q systems

$$\begin{pmatrix} A - \lambda_m^{(i)} B & B\bar{Z}^{(i)} \\ \bar{Z}^{(i)^T} & 0 \end{pmatrix} \begin{pmatrix} \Delta \bar{z}_m^{(i)} \\ -\Delta \lambda_m^{(i)} \end{pmatrix} = \begin{pmatrix} A\bar{z}_m^{(i)} - B\bar{z}_m^{(i)} \lambda_m^{(i)} \\ 0 \end{pmatrix}. \tag{9}$$

This block method has a quadratic convergence, then few linear systems must be solved in the computation. However, a 'good' initial guess must be provided to obtain convergence results. The linear systems are solved by using the GMRES method from the PETSc library [1] preconditioned with the block preconditioner developed for this method in [3].

Hybrid From the convergence histories of the BIFPAM and the MGBNM we have developed a hybrid method based on these methods [4]. We start from a set of sinitial eigenvectors, we then apply the BIFPAM method until a tolerance of 10^{-3} and then, with this solution we apply the MGBNM that has a quadratic convergence to reach the desired tolerance.

4 MULTILEVEL INITIALIZATION

Usually, the computation of the λ-modes for a realistic nuclear reactor requires much time to be solved. In this work, we propose a multilevel initialization to accelerate the convergences of the eigenvalue solvers.

It is well known that the convergence of iterative methods improves if better initial guesses are used. In this sense, it is proposed to use a multi-level method with two meshes: the initial mesh chosen from the spatial discretization, called the fine mesh and a coarse mesh obtained from the initial one considering a lower number of nodes. The solution obtained in the coarse mesh is used to generate an improved initial guess for the solution in the fine mesh.

In the coarse mesh, the materials and their corresponding cross-section must be redefined in each cell with a homogenization method. To solve the coarse eigenvalue problem we use Krylov-Schur method implemented in the library SLEPc [6]. The multi-level method can be summarized in Figure 1.

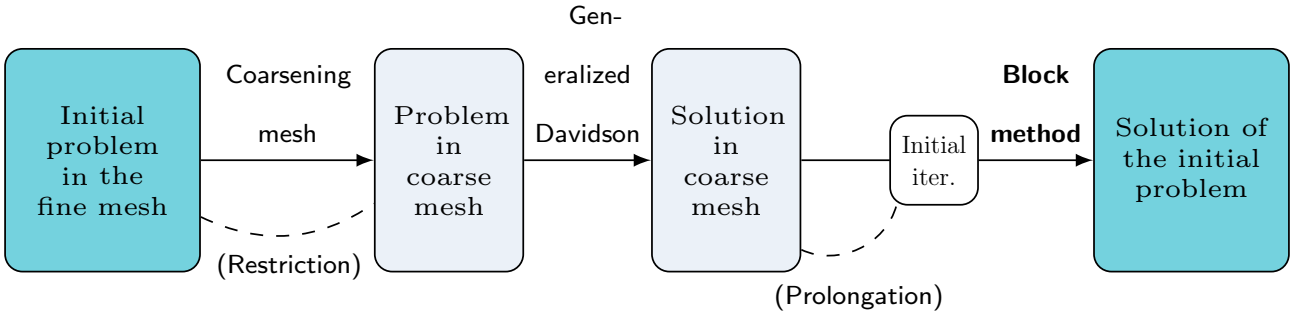


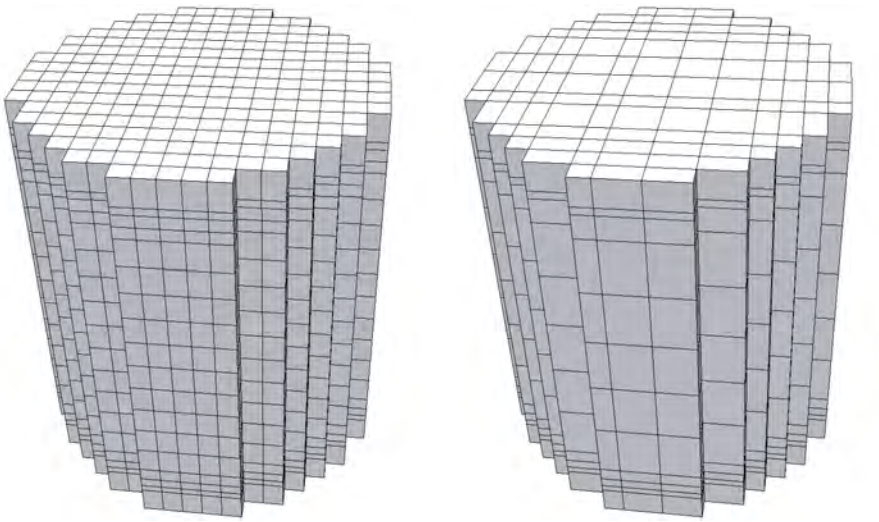
Figure 1: Scheme for the multilevel initialization.

5 NUMERICAL RESULTS

The three-dimensional NEACRP reactor is used to test the methodology described in this work [5]. The block strategies are tested to compute the dominant 4 λ-modes associated with the neutron diffusion equation. The initial mesh to discretize the reactor geometry has 3978 cells. Polynomials of degree 3 are used in the FEM, to have a problem of size 230 120 degrees of freedom. Tolerance for the eigenvalue solvers is set to obtain a residual error lower than 10^{-6}.

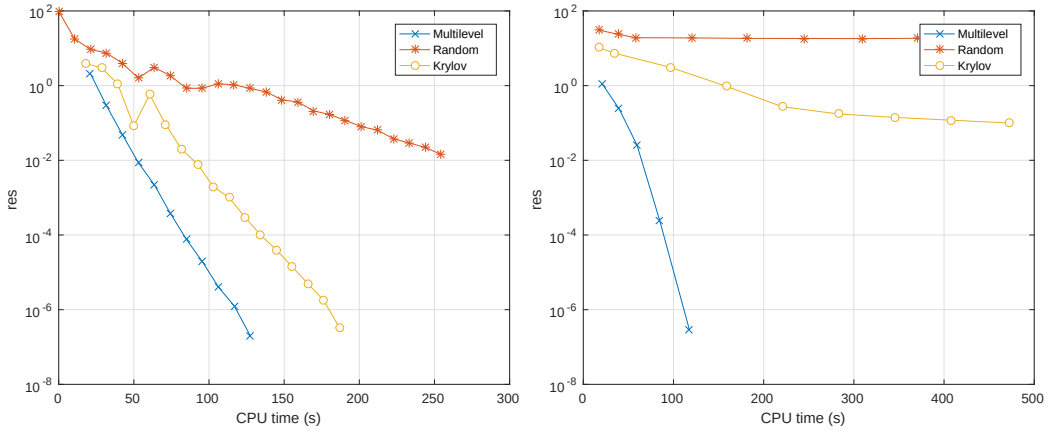
The methodology has been implemented in C++ based on data structures provided by the libraries Deal.II [2] and PETSc [1]. It has been incorporated to the open-source neutronic code `FEMFFUSION` [11]. For the computations, a computer with an Intel® Core™ i7-4790 @3.60GHz×8 processor with 32Gb of RAM running on Ubuntu GNU/Linux 18.04 LTS has been used.

First, multilevel initialization is analysed. It is compared with a Krylov initialization and a Random initialization. In the Krylov initialization, the vectors are obtained from a subspace of Krylov of dimension 10 associated to the matrix $A - \lambda_0^{(0)} B$. The Random initialization generates the q vectors using random numbers in the interval $[-1, 1]$. In both cases, the Gram-Schmidt orthogonalization and the generalized Rayleigh-Ritz process are then applied [4]. In the multilevel initialization, the simplified problem is defined by using a mesh of 1308 cells. Figure 2(a) shows the fine mesh used for the spatial discretization to solve the problem and Figure 2(a) represents the coarse mesh used to apply the multilevel initialization. The tolerance to solve for the simplified problem has been 10^{-3}. Figure 3 shows the convergence histories for the BIFPAM and the MBGNM with the different initializations. Both graphics reflect that the multilevel initialization, although it takes more time to obtain the initial guess, is a better strategy to initialize the block methods.



(a) Fine mesh (n. cells = 3978) (b) Coarse mesh (n. cells =1308)

Figure 2: Meshes for NEACRP reactor.



(a) BIFPAM (b) MGMNM

Figure 3: Convergence histories for the BIFPAM and the MGBNM using different initializations for the computation of the λ-modes of the NEACRP problem.

Now, the performance of the hybrid method is tested. Figure 4 compares the convergence histories of the MGBNM and the BIFPAM with the ILU and GMG preconditioner. It is deduced that the desired tolerance is reached quicker with the MGBNM. However, we would like to highlight that the convergence behaviour of BIFPAM-ILU is very similar to the one of BIFPAM-GMG and when the residual becomes smaller the convergence of the Newton method becomes faster.

Thus, it is proposed the hybrid method that initializes the algorithm with the BIFPAM method until $\text{res}_g = 10^{-2}$ and then, the MGBNM is applied. The BIFPAM has been set with the ILU preconditioner. Figure 5 compares the hybrid scheme with the MGBNM and the BIFPAM with ILU preconditioner. It is showed that the hybrid algorithm is an efficient scheme to compute 4 eigenvalues of the NEACRP problem.

Table 1 shows a comparison of the different eigenvalue solvers for the computation of several sets of eigenvalues of size q. In this computation, a semi-matrix free technique is used to avoid the full assemble of the matrices and then, the ILU preconditioner is substituted by the block Gauss-Seidel preconditioner [13]. All solvers are initialized with the multilevel technique. This

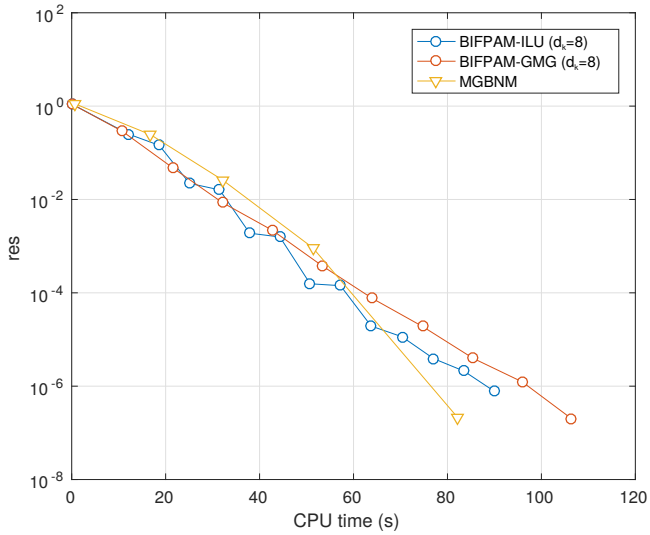


Figure 4: Convergence history for the fourth dominant eigenvalues of the NEACRP problem using the MGBNM and the BIFPAM.

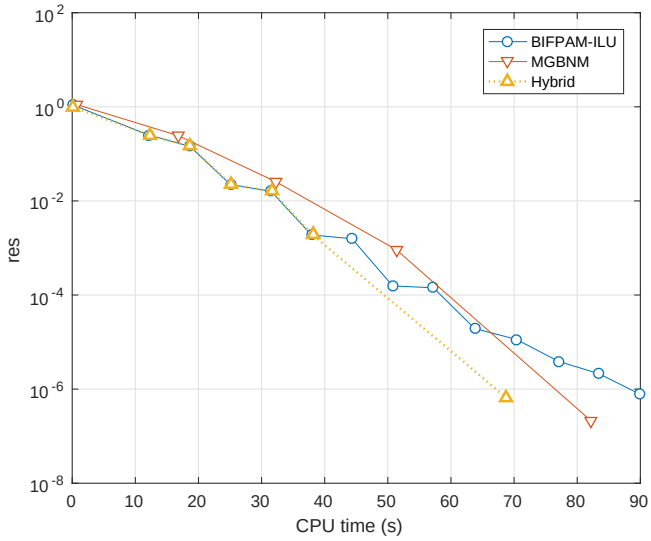


Figure 5: Convergence history of the BIFPAM with ILU preconditioner, the MGBNM and the hybrid method.

Table shows that the fastest results are obtained by applying the hybrid method, although, for a small number of eigenvalues, the BIFPAM is also very efficient.

Table 1: Computational times (s) obtained for the NEACRP reactor using the KSM method, the GDM, the BIFPAM, the MGBNM and the Hybrid method for different set of eigenvalues q.

q	GDM	BIFPAM	MGBNM	Hybrid
1	26	20	43	20
4	92	57	80	53
6	135	131	82	78

6 CONCLUSIONS

This work presents and compares several block eigenvalue solvers to compute a set of λ-modes associated with the neutron diffusion equation. Moreover, different strategies to improve the efficiency of these methods are described. First, numerical results show that the multilevel

initialization improves the efficiency of the methodologies. Regarding the eigenvalue solvers, one can deduce that the hybrid method (that combines the BIFPAM and the MGBNM) reduces the computational time to compute a set of modes in comparison with the BIFPAM, the MGBNM and the block Generalized Davidson.

7 ACKNOWLEDGEMENTS

This work has been partially supported by Spanish Ministerio de Economía y Competitividad under projects ENE2017-89029-P and MTM2017-85669-P. Furthermore, this work has been financed by the Generalitat Valenciana under the project PROMETEO/2018/035.

REFERENCES

[1] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, A. Dener, et al. PETSc users manual. 2019.

[2] W. Bangerth, T. Heister, and Kanschat G. `deal.II` *Differential Equations Analysis Library*. `http://www.dealii.org`.

[3] A. Carreño, L. Bergamaschi, A. Martinez, A. Vidal-Ferrándiz, D. Ginestar, and G. Verdú. Block preconditioning matrices for the newton method to compute the dominant λ-modes associated with the neutron diffusion equation. *Mathematical and Computational Applications*, 24(1):9, 2019.

[4] A. Carreño, A. Vidal-Ferràndiz, D. Ginestar, and G. Verdú. Block hybrid multilevel method to compute the dominant λ-modes of the neutron diffusion equation. *Annals of Nuclear Energy*, 121:513–524, 2018.

[5] H. Finnemann. A consistent nodal method for the analysis of space-time effects in large LWR's. Technical report, Technische Univ. Muenchen, Garching (F.R. Germany). Lab. fuer Reaktorregelung und Anlagensicherung, 1975.

[6] V. Hernandez, J.E. Roman, and V. Vidal. Slepc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):351–362, 2005.

[7] H. Lösche, R.and Schwetlick and G. Timmermann. A modified block Newton iteration for approximating an invariant subspace of a symmetric matrix. *Linear Algebra and its Applications*, 275:381 – 400, 1998.

[8] P. Quillen and Q. Ye. A block inverse-free preconditioned Krylov subspace method for symmetric generalized eigenvalue problems. *Journal of Computational and Applied Mathematics*, 233(5):1298–1313, 2010.

[9] W.M. Stacey. *Nuclear reactor physics*, volume 2. Wiley Online Library, 2007.

[10] G. Verdú, R. Miró, D. Ginestar, and V. Vidal. The implicit restarted Arnoldi method, an efficient alternative to solve the neutron diffusion equation. *Annals of nuclear energy*, 26(7):579–593, 1999.

[11] A. Vidal-Ferràndiz, A. Carreño, D. Ginestar, and G. Verdú. `FEMFFUSION`: A finite element method code for the neutron diffusion equation. `https://www.femffusion.imm.upv.es`, 2020.

[12] A. Vidal-Ferràndiz, R. Fayez, D. Ginestar, and G. Verdú. Solution of the lambda modes problem of a nuclear power reactor using an h–p finite element method. *Annals of Nuclear Energy*, 72:338–349, 2014.

[13] Antoni Vidal-Ferràndiz, Amanda Carreño, Damián Ginestar, and G Verdú. A block arnoldi method for the spn equations. *International Journal of Computer Mathematics*, 97(1-2):341–357, 2020.

Augmented fluid-structure interaction systems for viscoelastic pipelines and blood vessels

Giulia Bertaglia *

* Department of Mathematics and Computer Science
University of Ferrara
Ferrara, Italy
e-mail: giulia.bertaglia@unife.it

Key words: fluid–structure interaction, compliant ducts, viscoelastic effects, finite volume methods, IMEX Runge–Kutta schemes

Abstract: *In this work, innovative 1D hyperbolic models able to predict the behavior of the fluid-structure interaction mechanism that underlies the dynamics of flows in different compliant ducts are presented. Starting from the study of plastic water pipelines, the proposed tool is then applied to the biomathematical field to reproduce the mechanics of blood flow in both arteries and veins. With this aim, various different viscoelastic models have been applied and extended to obtain augmented fluid-structure interaction systems in which the constitutive equation of the material is directly embedded into the system as partial differential equation. These systems are solved recurring to Finite Volume Methods that take into account the recent evolution in the computational literature of hyperbolic balance laws systems. To avoid the loss of accuracy in the stiff regimes of the proposed systems, asymptotic-preserving Implicit-Explicit Runge-Kutta schemes are considered for the time discretization, which are able to maintain the consistency and the accuracy in the diffusive limit, without restrictions due to the scaling parameters.*

1 INTRODUCTION

Mathematical models and numerical methods are a powerful resource for better understanding phenomena and processes throughout the fluid dynamics field, allowing significant reduction in the costs, which would otherwise be required to perform laboratory experiments, and even allowing to obtain useful data that could not be gathered through measurements.

The correct characterization of the interactions that occur between the fluid and the wall that surrounds it is a fundamental aspect in all contexts involving deformable ducts, which requires the utmost attention at every stage of both the development of the computational method and the interpretation of the results and their application to cases of practical interest. Concerning flexible plastic pipes, which are playing an increasingly important role in hydraulic systems due to their cost-effectiveness and ease of installation, it has been demonstrated that the choice to characterize the fluid-structure interaction (FSI) behavior through a simple elastic law leads to consistent errors in the predictions of the pressure trends when studying hydraulic transients phenomena [7]. In fact, almost without exceptions, polymers manifest a viscoelastic behavior, responding to external forces in an intermediate way between the behavior of an elastic solid and a viscous liquid [19], and the adoption of a proper viscoelastic constitutive law for the definition of the FSI mechanism results fundamental [11].

Viscoelasticity is characterized by 3 primary features [14]:

1. *Creep*, which describes a material in continuous deformation over time when it is maintained under constant stress;

2. *Stress relaxation*, which refers to the decrease of stress over time when it is maintained under constant strain;

https://doi.org/10.4995/YIC2021.2021.13450

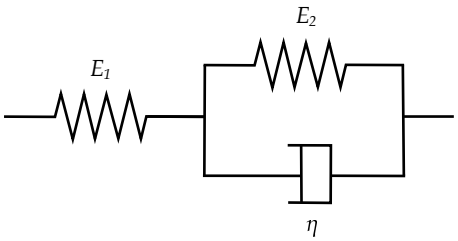


Figure 1: Scheme of the Standard Linear Solid Model with Kelvin-Voigt unit.

3. *Hysteresis*, which describes the dissipation of energy when a material undergoes cyclic loading and unloading.

Similarly, also biological tissues manifest viscoelastic properties. Thus, arteries and veins can be seen, with the due corrections specifically provided by hemodynamics, as highly flexible, viscoelastic tubes, tending almost to collapse under certain physiological conditions in the case of veins, hence leading to deal with highly non-linear systems [16]. Even though frequently, in hemodynamics models, the viscosity of vessels is neglected for simplicity, there is an increasing number of contributions showing the benefits of modeling the mechanical behavior of the vessel wall using a viscoelastic rheological characterization [1].

2 MATHEMATICAL MODELS

2.1 General one-dimensional models

The system of balance laws governing the motion of a compressible fluid through a flexible tube is obtained averaging the 3D compressible Navier-Stokes equations over the cross-section under the assumption of axial symmetry of the geometry of the conduct and of the flow. The resulting 1D non-linear hyperbolic system of partial differential equations (PDEs), composed by the continuity equation and by the momentum equation, reads [17]:

$$\partial_t(A\rho) + \partial_x(A\rho u) = 0 \tag{1a}$$

$$\partial_t(A\rho u) + \partial_x(A\rho u^2 + Ap) - p\,\partial_x A = F_R, \tag{1b}$$

where x is the space, t is the time, A is the cross-sectional area of the tube, ρ is the cross-sectional averaged density of the fluid, u is the averaged fluid velocity, p is the averaged fluid pressure and F_R is a model of the friction between fluid and tube wall, which can either account only for quasi-steady friction effects or both quasi-steady and unsteady ones (for further details the reader can refer to [2]).

Notice that when an incompressible fluid is considered (as for the case of blood flow studies) the system can be written as follows [16]:

$$\partial_t A + \partial_x(Au) = 0 \tag{2a}$$

$$\partial_t(Au) + \partial_x(Au^2) + \frac{A}{\rho}\,\partial_x p = \frac{F_R}{\rho}. \tag{2b}$$

To close system (1), an equation of state (EOS) and a constitutive law (also called *tube law*) must be introduced. In most of the technical applications it is usually sufficient to assume a barotropic behavior of the fluid, therefore $\rho = \rho(p)$. Nevertheless, taking into account cavitation phenomena may be necessary. An EOS for barotropic flows which accounts also for cavitation effects is presented in [10]. On the other hand, to solve system (2), only a proper tube law is needed.

2.2 The augmented fluid-structure interaction systems

The tube law describes the relationship between the tube cross-section and the internal pressure, containing all the information about the mechanical behavior of the pipe material. To correctly model the compliance and the flexibility of plastic ducts, in this work the Standard Linear Solid (SLS) model is considered, being the simplest viscoelastic rheological model able to describe the three main features of viscoelastic materials [14]. Hence, we assume that the mechanical behavior of the wall is defined by the interaction of a linear spring in series with a Kelvin-Voigt unit, composed of a linear spring in parallel with a linear dash-pot, as presented in Figure 1.

Evaluating the constitutive equation of the SLS model, expressed in terms of stress σ and strain ϵ,

$$\mathrm{d}_t\sigma = E_0\,\mathrm{d}_t\epsilon - \frac{1}{\tau_r}(\sigma - E_\infty\epsilon), \tag{3}$$

the three parameters of the model, namely the instantaneous Young modulus E_0, the asymptotic Young modulus E_∞, and the relaxation time τ_r, are so defined (referring to Figure 1):

$$E_0 = E_1, \qquad E_\infty = \frac{E_1E_2}{E_1+E_2}, \qquad \tau_r = \frac{\eta}{E_1+E_2}. \tag{4}$$

From equation (3), concerning a compressible fluid and a mildly non-linear system (1), applying Barlow's formula, introducing the linearized kinematic relation between the strain and the non-dimensional cross-sectional area rescaled with respect to its reference value $\alpha = \frac{A}{A_0} = (1+\epsilon^2) \approx 1+2\epsilon$, and recurring to the continuity equation (1a), the following PDE form of the SLS rheological law is obtained [2, 13]:

$$\partial_t A + d_1\,\partial_x(A\rho u) = S_1, \tag{5}$$

where

$$d_1 = \frac{2c_s^2}{2\rho c_s^2 + K\alpha}, \qquad S_1 = \frac{1}{\tau_r}\left[\frac{2A}{2\rho c_s^2 + K\alpha}(p-p_0) - \frac{E_\infty}{E_0}\frac{AK}{2\rho c_s^2 + K\alpha}(\alpha-1)\right].$$

Here, K represents the stiffness of the material, which accounts for the instantaneous Young modulus E_0, the wall thickness and the radius of the tube, $c_s = \sqrt{\frac{\partial p}{\partial \rho}}$ is the celerity contribute related to the compressibility of the fluid, which results equal to the sound speed when cavitation does not occur [2], and p_0 is the equilibrium pressure.

It can be observed that the relaxation time τ_r, and therefore the viscosity coefficient η, affects only the source term S_1. In fact, the viscous information about the FSI mechanism are all embedded in the term S_1, which defines viscoelastic damping effects. Interestingly, if we let $\tau_r \to 0$, entering in the diffusive and stiff regime of the system, from equation (3) we recover exactly the Laplace law, which is the standard elastic law used in literature [13]. Therefore, the hyperbolic augmented fluid-structure interaction (a-FSI) system for compressible fluids and mildly non-linear systems, capable of describing from simple elastic to viscoelastic FSI mechanisms, results

$$\partial_t(A\rho) + \partial_x(A\rho u) = 0 \tag{6a}$$
$$\partial_t(A\rho u) + \partial_x(A\rho u^2 + Ap) - p\,\partial_x A = F_R \tag{6b}$$
$$\partial_t A + d_1\,\partial_x(A\rho u) = S_1. \tag{6c}$$

Notice that to allow a formally correct treatment of possible discontinuous longitudinal changes of the reference cross-section or of the mechanical parameters of the wall, it is possible

to account for trivial equations which simply states that the interested variables are constant in time [2, 3].

A similar procedure can be followed also when considering blood flow models, hence an incompressible fluid and a highly non-linear setting, as in system (2), which leads to analogous results. Indeed, defining $\epsilon = \alpha^m - \alpha^n$, where parameters m and n are associated to the specific behavior of the vessel wall, whether arterial or venous [15], and using this definition in equation (3) together with Barlow's formula, recurring also to the continuity equation (2a), the following PDE of the SLS model is obtained [3, 4]:

$$\partial_t p + d_2\, \partial_x(Au) = S_2, \tag{7}$$

with

$$d_2 = \frac{K}{A}\left(m\alpha^m - n\alpha^n\right), \qquad S_2 = \frac{1}{\tau_r}\left[\frac{E_\infty}{E_0} K\left(\alpha^m - \alpha^n\right) - (p - p_0)\right].$$

The reader is invited to observe the similarities between d_1, S_1 and d_2, S_2. In particular, also in this configuration, the source term S_2 accounts for all the viscoelastic information of the FSI mechanism, and if we consider the diffusive limit letting $\tau_r \to 0$, we recover again the corresponding elastic tube law [3, 5]. Hence, the final hyperbolic a-FSI system for blood flow results:

$$\partial_t A + \partial_x(Au) = 0 \tag{8a}$$

$$\partial_t(Au) + \partial_x(Au^2) + \frac{A}{\rho}\partial_x p = \frac{F_R}{\rho} \tag{8b}$$

$$\partial_t p + d_2\, \partial_x(Au) = S_2 \tag{8c}$$

It is worth to underline that the choice of inserting the tube law in PDE form straight inside the system of equations results advantageous if compared to approaches generally followed in literature [1, 15]. Indeed, if the classical formulation is adopted choosing to characterize the FSI with the Kelvin-Voigt viscoelastic model (which, anyhow, lacks in the description of the relaxation process of the stress [14]), a second order derivative in space of the flow rate Au arises, which leads to deal with a non-hyperbolic system and consequent numerical issues.

Finally, to obtain more flexible models, it is possible to extend the number of Kelvin-Voigt units in the SLS configuration, obtaining the so-called Kelvin-Voigt chain [14]. Theoretically, the more elements we have, the more accurate our model will be in describing the real response of the material. Conversely, the more complex the model is, the more parameters that must be calibrated there are. The extension for the case of water pipelines is presented in details in [2].

3 NUMERICAL METHODS

Initially, to solve system (6), three different numerical schemes have been chosen and compared: the widely used Method of Characteristics (MOC) [7], an explicit path-conservative finite volume (FV) method associated with the Dumbser-Osher-Toro (DOT) Riemann solver [9], and a semi-implicit (SI) FV method specifically developed for axially symmetric compressible flows in compliant tubes [10].

On the other hand, to solve system (8), which can result *stiff* under physiological conditions, an Implicit-Explicit (IMEX) Runge-Kutta scheme, proposed for applications to hyperbolic systems with stiff relaxation terms, is considered [18]; while, for the space discretization, the same FV method with DOT solver previously mentioned is used. In particular, the second-order IMEX-SSP2(3,3,2) scheme is adopted [18]. The chosen numerical scheme is asymptotic preserving (AP) and asymptotic accurate in the zero-relaxation limit (i.e. when $\tau_r \to 0$), which

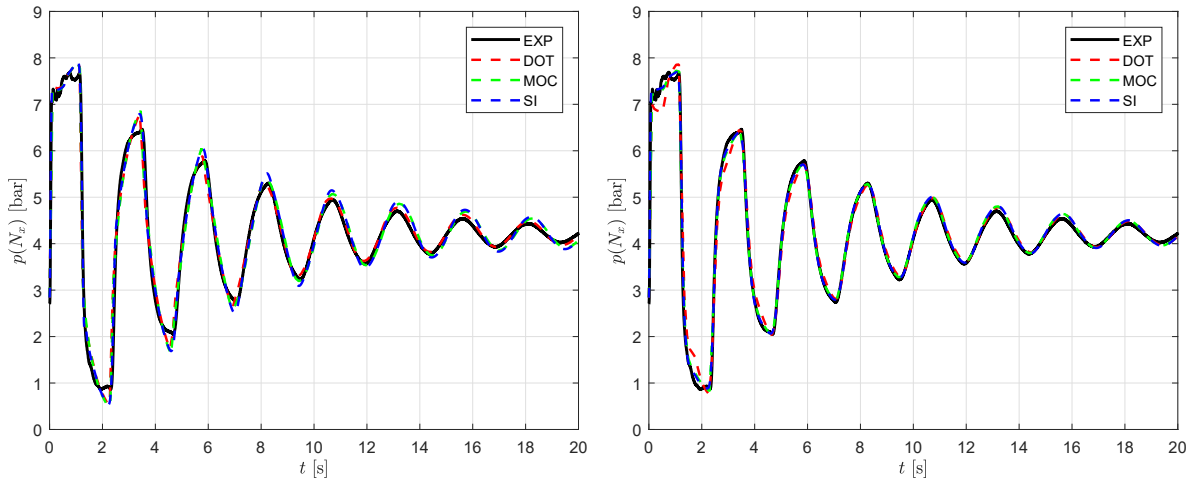


Figure 2: Comparison of the numerical results obtained with MOC, DOT and SI schemes against the experimental solution (EXP) of the water hammer test when using the SLS model (left) or the Kelvin-Voigt chain (right). Pressure $p(N_x)$ at the downstream end.

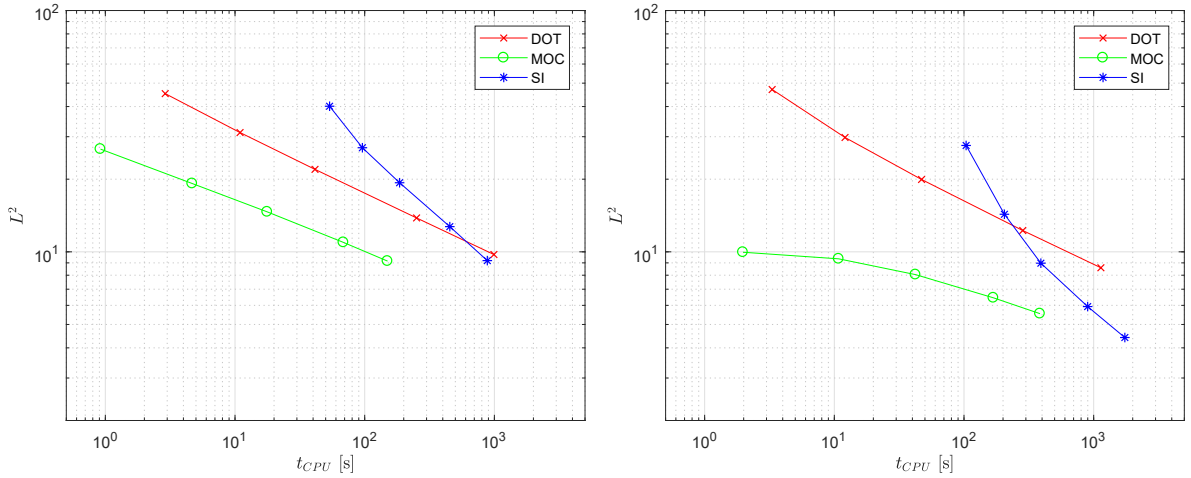


Figure 3: Results of the efficiency analysis for the water hammer test with the SLS model (left) and Kelvin-Voigt chain (right), in terms of L^2 norm with respect to the CPU time t_{CPU}.

allows to preserve the consistency of the scheme in the equilibrium, elastic limit as well as the order of accuracy, without restrictions due to the scaling parameters [5]. Another advantage of the chosen scheme lays in the possibility to analytically linearize each Runge-Kutta step to obtain a totally explicit algorithm, avoiding the adoption of iterative procedures like Newton-Raphson method, with a consequent consistent reduction of the computational cost [3].

4 NUMERICAL RESULTS AND DISCUSSION

To validate the proposed methodologies, different numerical tests have been designed. To compare the numerical methods used to solve system (6), a water hammer test case is here presented, with reference experimental data taken from [12], for which both the SLS model and the Kelvin-Voigt chain with 5 units are used.

Concerning the a-FSI blood flow model (8), targeted comparisons between numerical results and literature benchmarks have been performed with respect to close to reality test cases in single portions of vessels. In addition, patient-specific tests are considered, for which it has been possible to compare numerical results with available pressure data recorded in-vivo, from

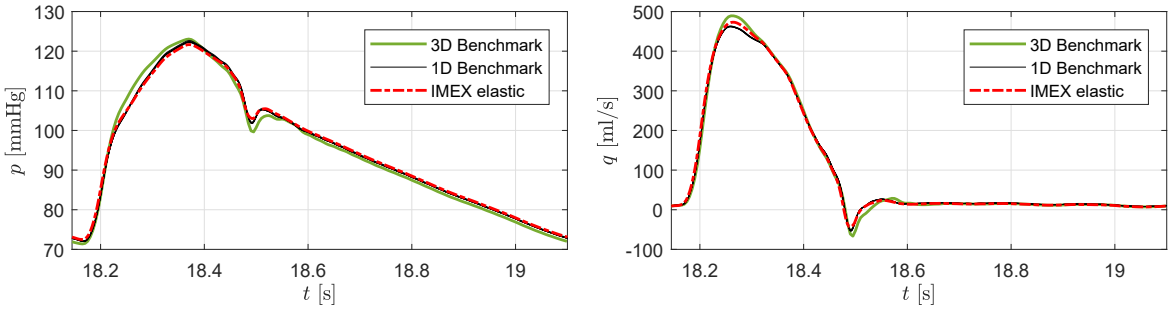


Figure 4: Baseline upper thoracic aorta case. Results obtained solving the 1D a-FSI system with the IMEX FV scheme with elastic tube law compared to six 1D and one 3D benchmark solutions. Results presented in terms of pressure at the midpoint (left) and flow rate at the midpoint (right).

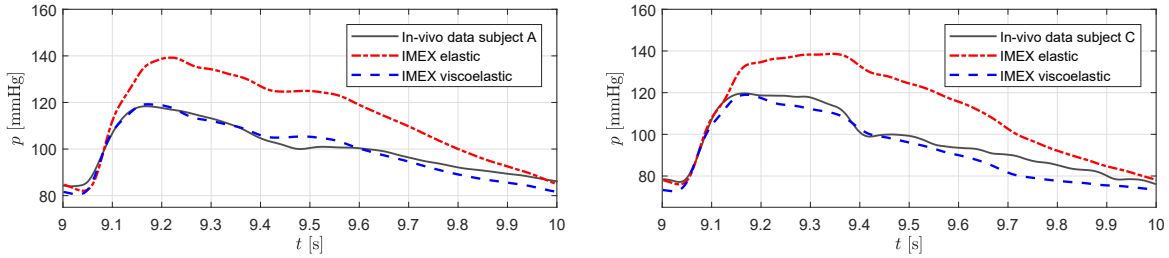


Figure 5: Patient-specific common carotid artery cases. Results obtained solving the 1D a-FSI system with the IMEX FV scheme, with elastic and viscoelastic tube law, for a 29 years old subject (left) and a 44 years old subject (right), in terms of pressure, compared with measured data.

different volunteers' common carotid arteries [4].

4.1 Water hammer tests

Following [12], a DN50 (22.0 mm radius) high-density polyethylene (HDPE) pipe is considered, with length 203.3 m and a flow rate of 2.0 l/s. In order to experimentally generate the transient wave, a closure maneuver was performed to a valve positioned downstream of the pipeline, with a closure time of 0.1 s. For this test, viscoelastic parameters have been calibrated using the SCE-UA (Shuffled Complex Evolution - University of Arizona) algorithm [8]. From Figure 2, it can be verified that the three numerical methods reproduce similar results, both using the SLS model or the extended Kelvin-Voigt chain. At the same time, it is observed that the increment of viscoelastic parameters does not return a consistent increase in the quality of the final result, weighing, on the other hand, in terms of computational cost and difficulty of calibration of the parameters. In fact, for the same water hammer test, an efficiency analysis has been computed to evaluate the performance of each numerical scheme adopted. Observing results shown in Figure 3, it is visible that the increment of viscoelastic parameters to characterize the material mechanics leads to an inevitable increment of computational costs not balanced by a comparable error reduction.

4.2 Blood flow tests

A baseline upper thoracic aorta test case is simulated, following [6], using a purely elastic wall model to allow comparisons with benchmark data available in literature. Figure 4 shows a comparison of the numerical results obtained solving the a-FSI system with the IMEX FV scheme with respect to six 1D and one 3D benchmarks [6]. It can be noticed that, for both pressure and flow rate, IMEX results are in perfect agreement with benchmarks.

Because no reference solutions of blood flow simulations on single vessels assuming a vis-

coelastic FSI are available in literature, flow velocity and pressure data measured in-vivo from four common carotids and two femoral arteries of volunteer subjects have been used to set up patient-specific test cases, to validate the proposed model in its viscoelastic configuration. The velocity wave extrapolated from each of the six subjects, obtained by Doppler measurements, is prescribed as inlet condition, while the post-processed pressure wave, measured recurring to the Arterial Tonometry technique, is used for comparisons with numerical results [4]. These simulations have been performed using both the elastic model and the viscoelastic law, to evaluate the effects of viscoelasticity in arterial hemodynamics. Viscoelastic parameters have been calibrated following [4]. Results of the test cases obtained for two patient-specific common carotid arteries are here reported in Figure 5, from which the excellent agreement between in-vivo measured and numerical pressure wave, obtained with the proposed methodology, can be observed. Indeed, these results confirm the capability of the proposed model to reproduce realistic pressure signals and the importance of taking into account the viscosity of the vessel wall in order not to overestimate systolic pressure values.

REFERENCES

[1] Alastruey, J., Khir, A. W., Matthys, K. S., Segers, P., Sherwin, S. J., Verdonck, P. R., Parker, K. H., Peiró, J. Pulse wave propagation in a model human arterial network: Assessment of 1-D visco-elastic simulations against in vitro measurements. *J. Biomech.*, 44(12):2250–2258 (2011).

[2] Bertaglia, G., Ioriatti, M., Valiani, A., Dumbser, M., Caleffi, V. Numerical methods for hydraulic transients in visco-elastic pipes. *J. Fluids Struct.*, 81:230–254 (2018).

[3] Bertaglia, G., Caleffi, V., Valiani, A. Modeling blood flow in viscoelastic vessels: the 1D augmented fluid–structure interaction system. *Comput. Methods Appl. Mech. Eng.*, 360(C):112772 (2020).

[4] Bertaglia, G., Navas-Montilla, A., Valiani, A., Monge García, M. I., Murillo, J., Caleffi, V. Computational hemodynamics in arteries with the one-dimensional augmented fluid-structure interaction system: viscoelastic parameters estimation and comparison with in-vivo data. *J. Biomech.*, 100(C):109595 (2020).

[5] Bertaglia, G., Caleffi, V., Pareschi, L., Valiani, A. Uncertainty quantification of viscoelastic parameters in arterial hemodynamics with the a-FSI blood flow model. *J. Comput. Phys.*, 430:110102 (2021).

[6] Boileau, E., Nithiarasu, P., Blanco, P. J., Müller, L. O., Fossan, F. E., Hellevik, L. R., Donders, W. P., Huberts, W., Willemet, M., Alastruey, J. A benchmark study of numerical schemes for one-dimensional arterial blood flow modelling. *Int. J. Numer. Method. Biomed. Eng.*, e02732:1–33 (2015).

[7] Covas, D. I. C., Stoianov, I., Mano, J. F., Ramos, H., Graham, N., Maksimovic, C. The dynamic effect of pipe-wall viscoelasticity in hydraulic transients. Part II - model development, calibration and verification. *J. Hydraul. Res.*, 43(1):56–70 (2005).

[8] Duan, Q. Y., Gupta, V. K., Sorooshian, S. Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory Appl.*, 76(3):501–521 (1993).

[9] Dumbser, M., Toro, E. F. A simple extension of the Osher Riemann solver to non-conservative hyperbolic systems. *J. Sci. Comput.*, 48:70–88 (2011).

[10] Dumbser, M., Iben, U., Ioriatti, M. An efficient semi-implicit finite volume method for axially symmetric compressible flows in compliant tubes. *Appl. Numer. Math.*, 89:24–44 (2015).

[11] Ferras, D., Manso, P., Schleiss, A., Covas, D. One-Dimensional Fluid–Structure Interaction Models in Pressurized Fluid-Filled Pipes: A Review. *Appl. Sci.*, 8(10):1844 (2018).

[12] Evangelista, S., Leopardi, A., Pignatelli, R., de Marinis, G. Hydraulic Transients in Viscoelastic Branched Pipelines. *J. Hydraul. Eng.*, 141(8):04015016 (2015).

[13] Leibinger, J., Dumbser, M., Iben, U., Wayand, I. A path-conservative Osher-type scheme for axially symmetric compressible flows in flexible visco-elastic tubes. *Appl. Numer. Math.*, 105:47–63 (2016).

[14] Lakes, R. Viscoelastic Materials. *Cambridge University Press* (2009).

[15] Montecinos, G. I., Müller, L. O., Toro, E. F. Hyperbolic reformulation of a 1D viscoelastic blood flow model and ADER finite volume schemes. *J. Comput. Phys.*, 266:101–123 (2014).

[16] Formaggia, L., Quarteroni, A., Veneziani, A. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system.* Springer (2009).

[17] Ghidaoui, M. S., Zhao, M., McInnis, D. A., Axworthy, D. H. A Review of Water Hammer Theory and Practice. *Appl. Mech. Rev.*, 58(1):49–76 (2005).

[18] Pareschi, L., Russo, G. Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. *J. Sci. Comput.*, 25:129–155 (2005).

[19] Shaw, M. T., MacKnight, W. J. *Introduction to Polymer Viscoelasticity: Third Edition.* John Wiley & Sons, Inc. (2005).