Advances in Mathematical Fluid Mechanics

Fernando Carapau
Ashwin Vaidya
Editors

Recent Advances in Mechanics and Fluid-Structure Interaction with Applications

The Bong Jae Chung Memorial Volume

Birkhäuser

Birkhäuser

Advances in Mathematical Fluid Mechanics

The *Advances in Mathematical Fluid Mechanics* series is a forum for the publication of high-quality, peer-reviewed research monographs and edited collections on the mathematical theory of fluid mechanics, with special regards to the Navier-Stokes equations and other significant viscous and inviscid fluid models. Titles in this series consider theoretical, numerical, and computational methods, as well as applications to science and engineering. Works in related areas of mathematics that have a direct bearing on fluid mechanics are also welcome. All manuscripts are peer-reviewed to meet the highest standards of scientific literature.

Fernando Carapau • Ashwin Vaidya
Editors

Recent Advances in Mechanics and Fluid-Structure Interaction with Applications

The Bong Jae Chung Memorial Volume

Birkhäuser

Editors

Fernando Carapau
Department of Mathematics and CIMA
University of Évora
Évora, Portugal

Ashwin Vaidya
Department of Mathematics
Montclair State University
Montclair, NJ, USA

ISSN 2297-0320 ISSN 2297-0339 (electronic)
Advances in Mathematical Fluid Mechanics
ISBN 978-3-031-14323-6 ISBN 978-3-031-14324-3 (eBook)
https://doi.org/10.1007/978-3-031-14324-3

Mathematics Subject Classification: 76Dxx, 74Fxx, 74F10, 74F20, 76Axx, 37Nxx, 37N25, 37N30, 76Mxx, 76M10, 76M12, 76Zxx, 74Bxx

This book is published under the imprint Birkhäuser, www.birkhauser-science.com by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

In loving memory of our dear friend and colleague Bong Jae Chung.
April 1, 1967–February 12, 2021

Preface

This book brings together current scholarship in the area of mechanics and its applications to various branches of mathematics, science, and engineering, specifically around themes of computation and modeling in fluid mechanics, in honor of our dear friend and colleague, Dr. Bong Jae Chung, a computational scientist who passed away on February 12, 2021. Bong Jae or Chung, as he was often referred to by his friends, was born in Daegu, South Korea, where he completed his undergraduate degree in physics at Kyung Hee University. He came to the United States in 1994 as a graduate student. After a brief stint in Georgia, he moved to Pittsburgh for his master's degree and eventually graduated from the University of Pittsburgh in 2004 with a PhD in mechanical engineering under the guidance of Professor Anne Robertson. He worked for several years as a postdoc and research professor at various universities including Johns Hopkins University (with Prof. Aleksander Popel), the University of North Carolina Chapel Hill (with Profs. Richard McLaughlin, Roberto Camassa, and Alberto Scotti), and George Mason University (with Prof. Juan Cebral) before securing a tenure track position in the Department of Mathematical Sciences and Department of Applied Mathematics & Statistics at Montclair State University, New Jersey, where he was employed for a little more than three years before his untimely passing.

Bong Jae was a prolific researcher with diverse interests ranging from problems of classical fluid mechanics, flows pertaining cerebral aneurysms, protein aggregation modeling, vortex-induced vibrations, and pattern formation in fluids to non-equilibrium thermodynamics. More recently, he had started working on modeling problems related to drug delivery, a topic on which his student Nicholas Jeffoupolous wrote a master's thesis[1]. Bong Jae also had a keen interest in experimental work having spent several months as a postdoc with Prof. George Klinzing in the Department of Chemical Engineering at the University of Pittsburgh, where he participated in the experiments related to pneumatic conveying, and also at the fluid dynamics laboratory at UNC Chapel-Hill, where he was part of the

[1] You can find his article with Bong Jae on this topic in Chap. 14

team studying vortex-induced vibrations. Bong Jae was an essential member of the Complex Fluids Laboratory at Montclair State University and was looking forward to getting involved in a variety of experiments and using our particle image velocimetry system to better understand wake vortex dynamics, which he was also modeling numerically.

This volume contains work by scholars from several countries who are experts in the different areas of theoretical and computational fluid mechanics and other areas of science in which Bong Jae shared keen interest. Many of the contributions here are by his mentors, friends, and collaborators and also scholars he wanted to work with in the future. To the extent possible, we have taken care to prepare the articles so that they are accessible and relevant not only to other researchers but also to graduate students, postdocs, and those wanting to pursue new lines of research in these areas of mechanics. For this reason, the papers have been prepared in a semi-tutorial style, where possible. While scholarship in the area of Fluid Structure Interaction (FSI) has been gaining ground, especially with developments in computational techniques and technology, most books in this area are restricted to very specific topics. The particular novelty and interesting aspect of this book lies in its interdisciplinarity, with contributions from mathematicians, physicists, mechanical and biomechanical engineers, and even psychologists, all bringing new perspectives to the study of mechanics.

This book is truly an eclectic mix of articles on various themes. We have therefore decided to organize the book into four thematic parts: (1) Theory, (2) Computations, (3) Experiments, and (4) Applications. In some cases where the papers fall in multiple categories, we have tried to assign them to a part we feel it best represents.

Part I on theoretical fluid mechanics consists of four papers which range from mathematical (existence of solutions) issues for fluids (Chap. 1: Berselli and Růžička) and fluid solid systems (Chap. 3: Galdi) to modeling the physics of fluids (Chap. 2: Carapau, Correia and Areias; Chap. 4: Camassa, Ding, McLaughlin, Overman, Parker, and Vaidya).[2] Part II on CFD and numerical methods features six papers. The first of these (Chap. 5: Bodnár, Keslerová, and Lancmanová) on the numerical methods for flow in branching channels was a repeating theme in much of Bong Jae's computational biomechanics work. Other papers in this part are focused on Galerkin methods in problems of plasticity (Chap. 6: Areias, Carapau, Lopes, and Rabczuk) and novel uses of modern computational techniques in mechanics such as use of machine learning techniques to understand emergence of patterns in kinetic models related to protein aggregation (Chap. 7: Pateras, Vaidya, and Ghosh), reduced order modeling (Chap. 8: Snyder, Mou, Liu, San, De Vita, and Iliescu), numerical issues in the modeling of viscoelastic fluid flows (Chap. 9: Pires and Bodnár), and cellular automata modeling of complex fluids (Chap. 10: Ramos, Carapau, and Correia). Part III on experiments includes two papers. The first of these

[2] It is worth mentioning that Bong Jae was very interested in the problem discussed in (Chap. 4) which was initiated by the authors during his postdoctoral days at UNC-CH in 2008–2009; he even helped with debugging the initial codes written for this work.

is on the thermomechanics of self-organization in dissipative systems (Chap. 11: De Bari and Dixon) on which Bong Jae had previously published and wished to contribute more to in the coming years. The second paper in this part is devoted to the use of vortex-induced vibrations towards hydrokinetic energy generation (Chap. 12: Wulandana and Haque), which was one of his primary research interests. In fact, a significant part of his computational effort after his arrival in New Jersey was devoted to development of numerical methods to study fluid-solid interactions, specifically vortex-induced oscillations. Part IV on applications of fluid mechanics covers areas of deep interest to Bong Jae including drug delivery (Chap. 13: Azhdari, Emami, and Ferreira; Chap. 14: Jefopolous and Chung), carbon sequestration (Chap. 15: Phouc and Massoudi), and Ocular flow (Chap. 16: Chung, Martinez, and Vaidya). Two of these chapters feature articles by Chung and his past students Brandon and Nicholas (Chaps. 14 and 16).

It is certainly worth mentioning that a special issue of this kind is rare. Such honor is reserved for the ”generals” of science not ”foot soldiers”. A commonly held view among scientists, whose essential sentiment is even expressed by the likes of David Bohm, is captured in the following statement:[3]

In the whole of human history, perhaps only a few people have achieved it [creativity]. Most of the rest of human action has been relatively mediocre, though it is interlaced with flashes of penetrating insight that help raise it above the level of mere humdrum.

We respectfully reject this viewpoint and the overarching hierarchical value system that it imposes on scientific contributions. It is being slowly recognized that creativity happens at all levels, and while we all admire and rely on the paradigm shifting, “wall-breaking” efforts to eliminate barriers to knowledge, there are those who do the same, one brick at a time. Their efforts are no less valuable, and collectively taken, such efforts are essential for the next great scientific transformation. When sincere and consistent, such work also deserves acknowledgment. I am therefore deeply appreciative of all colleagues who have volunteered to contribute to and supported this volume, in honor of a soldier of science; they remind us that knowledge seeking is a collective effort and every contribution has merit, much of it yet unforeseen.

We convey special thanks to Professor Giovanni Paolo Galdi for his help and encouragement in getting the book published in this series and to Professor Anne Robertson, Bong Jae’s PhD advisor and collaborator, for her encouragement and commitment to this project. Bong Jae expressed deep admiration for all his teachers and mentors and was deeply influenced by them, especially Dr. Robertson. On his behalf, we would therefore like to thank all his mentors, including Dr. Aleksander Popel (Biomedical Engineering, Johns Hopkins), Dr. Richard McLaughlin (Mathematics, UNC-Chapel Hill), Dr. Alberto Scotti (Marine Sciences, UNC-Chapel Hill), Dr. Roberto Camassa (Mathematics, UNC-Chapel Hill), and Dr. Juan Cebral (Biomedical Engineering, George Mason) for their mentorship and for the

[3] Bohm, D. (2004) On Creativity, editor Lee Nichol. London: Routledge.

intellectual stimulation they provided. We also thank Bong Jae's students Nicholas and Brandon, both of whom lost a mentor midway through their thesis project and yet persisted in completing the work and are featured in this book. We acknowledge the help and support of Mr. Chris Eder and Ms. Saveetha Balasubramaniam at Birskhauser-Springer for helping us see this volume through and for making this such a smooth process for us.

Montclair, NJ, USA
Évora, Portugal
June 2022

Ashwin Vaidya
Fernando Carapau

Personal Memories and Tributes

I feel fortunate to have served as Dr. Bong Jae Chung's doctoral advisor, and therefore had the opportunity to get to know him personally, watch him grow intellectually, share the joy of immersing in shared research on fluid and solid mechanics, experience his genuine kindness and see the happiness he drew from his wife and dear friends. Bong Jae was my second doctoral student. I first met him in the early stages of his graduate studies at the University of Pittsburgh as a member of his Masters' thesis committee. His research focused on the numerical study of freely moving bubbles in a stirred column. I was impressed by Bong Jae's determination to deeply understand this difficult topic and, with his advisor, Dr. Hwang's agreement, recruited him to my research group for his doctoral studies. One of the things that stood out to me even at that time, was his great love of learning and discovery. He was clever, determined, and ready to take on new and difficult topics.

Dr. Chung's doctoral research covered challenging topics involving theoretical and computational studies of cerebral aneurysms. His initial computational work evaluated flow in arterial bifurcations, where cerebral aneurysms are typically found. He built on these results to design, for the first time, an in vitro flow chamber to expose endothelial cells to the same wall shear stress field found at the apices of cerebral bifurcations. While most of his research and coursework was fluid mechanics, during the last year of his doctoral work, we began discussing possible ways of improving existing arterial wall models. Despite the fact that his background in solid mechanics was limited, he independently learned the material in the advanced graduate text, Theoretical Elasticity, by Green and Zerna. He then moved on to use this knowledge to apply the theory of small on large elastic deformations to the arterial system, for the first time. As a postdoctoral researcher at George Mason University, Dr. Chung had an extremely productive collaboration with Dr. Juan Cebral, one of the top computational biofluid dynamicists. This work led to important publications in the field of cerebral aneurysms, including Bong Jae's first author review article in the *Annals of Biomedical Engineering, CFD for evaluation and treatment planning of aneurysms: review of proposed clinical uses and their challenges*. It was a pleasure to continue to work with Bong Jae through joint research with Juan and Bong Jae on cerebral aneurysms.

Bong Jae maintained his focus on learning and developing new knowledge through all the different chapters of his life, while persistently working toward his goal of being a professor. His wife, Kelly Yoo, fondly described their shared love of camping and that on these trips, "he always had his back pack full with heavy books, research notes and computer." Like the rest of us, she appreciated Bong Jae's thirst for learning and sharing knowledge with others. Bong Jae will live on in all of our memories with love and deep respect.

Pittsburgh, PA, USA Anne M. Robertson

Chung and I were dear friends and close collaborators for nearly 24 years so this is a deep personal loss. We studied together and graduated a few days apart. We were most fortuitous to even share part of our postdoctoral experience at the University of North Carolina—Chapel Hill together and thought it miraculous that we would end up as faculty members in the same university. I have fond memories of our working deep into the night, engaging in exciting scientific and philosophical discussions and making elaborate future plans for exciting projects.

Chung was dedicated to his work and very passionate about it. He was extremely prolific, the rate and diversity of his contributions, especially in the last decade of his life are impressive (see the following pages for a full list of his publications). However, to him, his research and even teaching were not about achievements or reputation—it stemmed from a sincere joy of learning and sharing his knowledge with others. In his friendships also, he was about filling moments together with laughter and love; it did not matter what he was doing with his friends it was about making the interaction memorable. In all his encounters, he was about listening, not talking; about compassion, forgiveness and seeking the best in others. I greatly admired and appreciated his wisdom of kindness and simplicity. There are a great many reasons to mourn the loss of a friend and colleague, but we see this volume as a celebration of a humble, thoughtful, and passionate scientific life.

Montclair, NJ, USA Ashwin Vaidya

In September 2000 I started my PhD work at the Department of Mechanical Engineering and Materials Science, Pittsburgh, PA, USA under the supervision of Professors Anne M. Robertson and Adélia Sequeira (DMAT/IST, Portugal). As part of the work in this group, under the guidance of Professor Anne M. Robertson, I met my colleague Chung (as many of us referred to him), with whom I developed a solid friendship over the years, not only in scientific terms, but also in our personal lives. We were brothers. Chung was a warm, simple person and friend to everyone. His scientific observations and precious collaboration within the working group were appreciated by all. The scientific community and science itself have prematurely lost a good thinker. Chung, wherever you are, a big hug from your brother Fernando!

Évora, Portugal Fernando Carapau

Publications by Bong Jae Chung

1. B. Chung, A.M. Robertson and D.G. Peters, The numerical design of a parallel plate flow chamber for investigation of endothelial cell response to shear stress, *Computers and Structures*, 81, 535–546, 2003
2. B. Chung, P.C. Johnson and A.S. Popel, Application of Chimera grid to modeling cell motion and aggregation in a narrow tube, *International Journal for Numerical Methods in Fluids*, 53(1), 105–128, 2006
3. B.J. Chung, A. Vaidya and R. Wulandana, Stability of steady flow in a channel with linear temperature dependent viscosity, *International Journal of Applied Mathematics and Mechanics*, 2(1), 24–33, 2006
4. B. Chung and A. Vaidya, Optimal principle in fluid structure interaction, *Physica D*, 237, 2945–2951, 2008
5. R. Camassa, B. Chung, P. Howard, R.M. McLaughlin and A. Vaidya, Vortex induced oscillations of cylinders at low and intermediate Reynolds numbers, Advances in Mathematical Fluid Mechanics: A Tribute to Giovanni Paolo Galdi, A. Sequeira and R. Rannacher (Ed.), Springer Verlag, 2008
6. B. Chung, S. Kim, P. C. Johnson and A.S. Popel, Computational fluid dynamics of aggregating red blood cells in postcapillary venules, *Computer Methods in Biomechanics and Biomedical Engineering*, 12(4), 385–397, 2009
7. Z. Zeng, B.J. Chung, M. Durka and A.M. Robertson, An in vitro device for evaluation of cellular response to flows found at the apex of arterial bifurcations, *Advances in Mathematical Fluid Mechanics*, 631–657, 2010, https://doi.org/10.1007/978-3-642-04068-9-35
8. B.J. Chung and A. Vaidya, On the slow motion of a sphere in fluids with non-constant viscosities, *International Journal of Engineering Science*, 48(1), 78–100, 2010
9. B.J. Chung, G. Gipson, A. Shenoy and A. Vaidya, Image analysis of wake structure past finite cylinders, *International Journal of Imaging*, 4(A10), 18–32, 2010
10. S. Achuthan, B.J. Chung, P. Ghosh, V. Rangachari and A. Vaidya, A modified Stokes-Einstein‘s equation for A-beta aggregation, *BMC Bioinformatics*, 12(10), S13, 1–13, 2011

11. B.J. Chung and A. Vaidya, A Non-equilibrium pattern selection in particle sedimentation, *Applied Mathematics and Computation*, 218(7), 3451–3465, 2011
12. B.J. Chung and A. Vaidya, On the Affordances of the MaxEP Principle, *European Physical Journal B*, 87(20), 2014
13. B. Chung and J.R. Cebral, CFD for evaluation and treatment planning of aneurysms: review of proposed clinical uses and their challenges, *Annals of Biomedical Engineering*, 43(1), 122–38, 2014, https://doi.org/10.1007/s10439-014-1093-6
14. B. Chung, F. Mut, R. Karidvel, R. Lingineni, D.F. Kallmes and J.R. Cebral, Hemodynamic analysis of fast and slow aneurysm occlusions by flow diversion in rabbits, *Journal of NeuroInterventional Surgery*, 7, 931–935, 2015, https://doi.org/10.1136/neurintsurg-2014-011412
15. J.R. Cebral, X. Duan, B.J. Chung, C. Putman, K. Aziz and A.M. Robertson, Wall mechanical properties and hemodynamics of unruptured intracranial aneurysms, *American Journal of Neuroradiology*, 36, 1695–1703, 2015, https://doi.org/10.3174/ajnr.A4358
16. B. Chung, M. Cohrs, W. Ernst, G. Galdi and A. Vaidya, Wake-cylinder interactions of a hinged cylinder at low and intermediate Reynolds numbers, *Archives of Applied Mechanics*, 86(4), 627–641, 2015, https://doi.org/10.1007/s00419-015-1051-2
17. P. Berg, C. Roloff, O. Beuing, S.I. Sugiyama, N. Aristokleus, A. Anayiotos, N. Ashton, N. Bressloff, A. Brown, B.J. Chung, J.R. Cebral, G. Copelli, W. Fu, A. Qiao, A. Geers, S. Hodis, D. Dragomir-Daescu, E. Imdieke, M. Khan, Valen Sendstad, K. Kono, H. Meng, J. Xiang, P. Menon, P. Albal, O. Mierka, R. Munster, H. Morales, J. Osman, L. Goubergrits, J. Pallares, S. Cito, A. Passalacqua, S. Piskin, K. Pekkan, S. Ramalho, N. Marques, S. Sanchi, K. Schumacher, J. Sturgeon, H. Svihlova, J. Hron, G. Usera, M. Mendina, D. Steinman and G. Janiga, The Computational fluid dynamics rupture challenge 2013-phase II: Variability of hemodynamic simulations in two intracranial aneurysms, *Journal of Biomechanical Engineering*, 137/121008-1, 2015, https://doi.org/10.1115/1.4031794
18. W. Brinjikji, B. Chung, C. Jimenez, C. Putman, D.F. Kallmes and J.R. Cebral, Hemodynamic differences between unstable and stable unruptured aneurysms independent of size and location: pilot study, *Journal of NeuroInterventional Surgery*, 9, 376–380, 2017, https://doi.org/10.1136/neurintsurg-2016-012327
19. J.R. Cebral, X. Duan, P.S. Gade, B.J. Chung, F. Mut, K. Aziz and A.M. Robertson, Regional mapping of flow and wall characteristics of intracranial aneurysms, *Annals of Biomedical Engineering*, 44(12), 3553–3567, 2016, https://doi.org/10.1007/s10439-016-1682-7
20. B. Chung, D. Platt and A. Vaidya, The Mechanics of Clearance in a non-Newtonian Lubrication Layer, *International Journal of Non-Linear Mechanics*, 86, 133–145, 2016, https://doi.org/10.1016/j.ijnonlinmec.2016.08.010
21. J.R. Cebral, E. Ollikainen, B. Chung, F. Mut, V. Sippola, B.R. Jahromi, R. Tulamo, J. Hernesniemi, M. Niemela, A. Robertson and J. Frosen, Flow

conditions in the intracranial aneurysm lumen associate with inflammation and degenerative changes of the aneurysm wall, *American Journal of Neuroradiology*, 38(1), 119–126, 2017

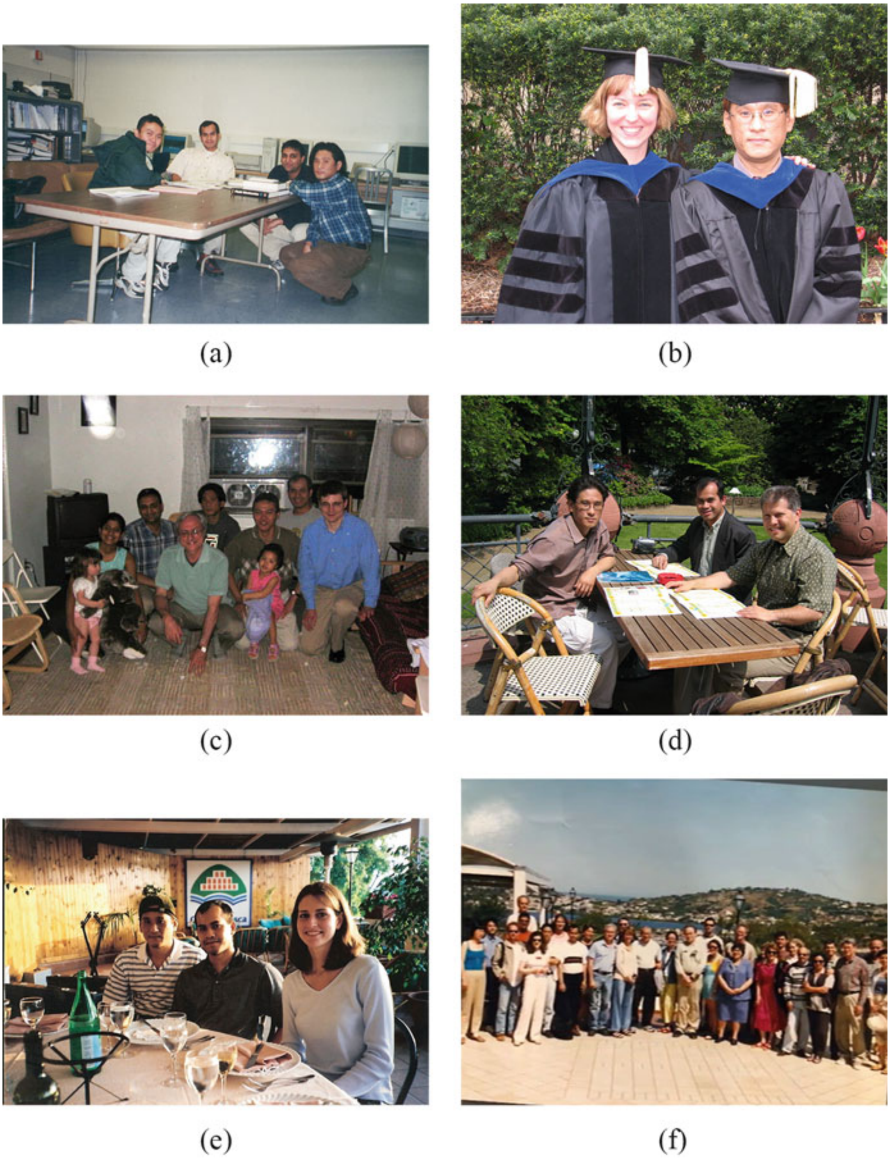
22. R. Doddasomayajula, B. Chung, F. Hamzei-Sichani, C. M. Putman and J.R. Cebral, Differences in hemodynamics and rupture rate of aneurysms at the bifurcation of the basilar and internal carotid arteries, *American Journal of Neuroradiology*, 38(3), 570–576, 2017, https://doi.org/10.3174/ajnr.A5088
23. D. Castillo, B. Chung, K. Schnitzer, K. Sorriano, H. Su and A. Vaidya, Metastable states in terminal orientation of hinged symmetric bodies in a flow, *International Journal of Engineering Science*, 111, 19–27, 2017, https://doi.org/10.1016/j.ijengsci.2016.11.004
24. J.R. Cebral, B.J. Chung, D. Ruijters, F. Nijnatten, F. Mut, P. Moret and S. Laurent, Understanding angiography-based aneurysm flow fields through comparison to computational fluid dynamics, *American Journal of Neuroradiology*, 38(6), 1180–1186, 2017, https://doi.org/10.3174/ajnr.A5158
25. B. Chung, F. Mut, W. Brinjikji, F. Hamzei-Sichani, C. Jimenez, C. Putman, D.F. Kallmes, P.M. Christopher, M. Pritz, F. Detmer and J.R. Cebral, Angio-architectures and hemodynamics characteristics of posterior communicating artery aneurysms and their association with rupture status, *American Journal of Neuroradiology*, 38(11), 2111-2118, 2017, https://doi.org/10.3174/ajnr.A5358
26. R. Doddasomayajula, B. Chung, F. Mut, C. M. Jimenez, F. Hamzei-Sichani, C. M. Putman and J. R. Cebral, Hemodynamic characteristics of ruptured and unruptured multiple aneurysms at mirror and ipsilateral locations, *American Journal of Neuroradiology*, 38(12), 2301–2307, 2017, https://doi.org/10.3174/ajnr.A5397
27. B.J. Chung, B. Ortega and A. Vaidya, Entropy production in a fluid-solid system far from thermodynamic equilibrium, *European Physical Journal E*, 40, 105, 2017, https://doi.org/10.1140/epje/i2017-11595-3
28. B. Waleed, B.J. Chung, J.R. Cebral, Y. Ding, J. Wald, F. Mut, Kadir, D. Kallmes, A. Rouchaud and G. Lanzino, Hemodynamic characteristics of stable and unstable vertebrobasilar dolichoectatic and fusiform aneurysms, *Journal of NeuroInterventional Surgery*, 10, 1102–1107, 2018
29. F.J. Detmer, B.J. Chung, F. Mut, M. Slawski, F. Hamzei-Sichani, C. Putman, C. Jimenez and J.R. Cebral, Development and internal validation of an aneurysm rupture probability model based on patient characteristics and aneurysm location, morphology, and hemodynamics, *International Journal of Computer Assisted Radiology and Surgery*, 13, 1767–1779, 2018, https://doi.org/10.1007/s11548-018-1837-0
30. F.J. Detmer, B.J. Chung, F. Mut, M. Pritz, M. Slawski, F. Hamzei-Sichani, D. Kallmes, C. Putman, C. Jimenez and J.R. Cebral, Development of a statistical model for discrimination of rupture status in posterior communicating artery aneurysms, *Acta Neurochir*, 160(8), 1643–1652, 2018, https://doi.org/10.1007/s00701-018-3595-8
31. B.J. Chung, M. Fernando, C.M. Putman, F. Hamzei-Sichani, W. Brinjiki, D. Kalmes, C.M. Jimenez and J.R. Cebral, Identification of hostile hemodynamics

and geometries of cerebral aneurysms: a case-control study, *American Journal of Neuroradiology*, 39(10), 1860–1866, 2018, https://doi.org/10.3174/ajnr.A5764

32. P. Berg and et al., Multiple Aneurysms AnaTomy CHallenge 2018 (MATCH)-Phase I: Segmentation, *Cardiovascular Engineering and Technology*, 9(4), 565–581, 2018, https://doi.org/10.1007/s13239/-018-00376-0
33. P. Berg and et al., Multiple Aneurysms AnaTomy Challenge 2018 (MATCH) - Phase II: Rupture Risk Assessment, *International Journal of Computer Assisted Radiology and Surgery*, 14, 1795–1804, 2019, https://doi.org/10.1007/s11548-019-01986-2
34. F. Detmer, S. Hadad, B.J. Chung, F. Mut, M. Slawski, N. Juchler, S. Hirsch, P. Bijlenga, Y. Uchiyama, S. Fujimura, M. Yamamoto, Y. Murayama, T. Hiroyuki, T. Koivisto, J. Frosen and J.R. Cebral, Extension of statistical learning for aneurysm rupture assessment to Finnish and Japanese populations using morphology, hemodynamics, and patient characteristics, *JNS Neurosurgical Focus*, 47(1), E16, 2019, https://doi.org/10.3171/2019.4.FOCUS19145
35. F. Mut, B.J. Chung, J. Chudyk, P. Lylyk, R. Kadirvel, D. Kallmes and J.R. Cebral, Image-based modeling of blood flow in cerebral aneurysms treated with intrasaccular flow diverting devices, *International Journal Numerical Methods in Biomedical Engineering*, 35, e3202, 2019, https://doi.org/10.1002/cnm.3202
36. J.R. Cebral, F. Detmer, B.J. Chung, J. Choque-Velasquez, B. Rezai, H. Lehto, R. Tulamo, J. Hernesniemi, M. Niemela, A. Yu, R. Williamson, K. Aziz, S. Sakur, S. Amin-Hanjani, F. Charbel, Y. Tobe, A. Robertson and J. Frosen, Local hemodynamic conditions associate with focal changes in the intracranial aneurysm wall, *AJNR Am. Journal of Neuroradiology*, 40(3), 510–516, 2019, https://doi.org/10.3174/ajnr.A5970
37. F.J. Detmer, B.J. Chung, C. Jimenez, F. Hamzei-Sichani, D. Kallmes, C. Putman and J.R. Cebral, Association of hemodynamics, morphology, and patient characteristics with aneurysm rupture stratified by aneurysm location, *Neuroradiology*, 61, 275–284, 2019, https://doi.org/10.1007/s00234-018-2135-9
38. J.R. Cebral, B.J. Chung, F. Mut, J. Chudyk, C. Bleise, E. Scrivano, P. Lylyk, R. Kadirvel and D. Kallmes, Analysis of flow dynamics and outcomes of cerebral aneurysms treated with intrasaccular flow diverting devices, *AJNR*, 40(9), 1511–1516, 2019, https://doi.org/10.3174/ajnr.A6169
39. P.S. Gade, R. Tulamo, K. Lee, F. Mut, E. Ollikainen, C. Chuang, B. Chung, M. Niemela, B.R. Jahromi, K. Aziz, A. Yu, F.T. Charbel, S. Amin-Hanjani, J. Frosen, J. Cebral and A.M. Robertson, Calcification in human intracranial aneurysms Is highly prevalent and displays both atherosclerotic and non-atherosclerotic types, *Arteriosclerosis, Thrombosis, and Vascular Biology*, 39(10), 2157–2167, 2019, https://doi.org/10.1161/ATVBAHA.119.312922
40. B.J. Chung and A. Vaidya, Self-organization in physical and biological systems: comment on "Morphogenesis as bayesian inference: A variational approach to pattern formation and control in complex biological systems" by F. Kuchling et

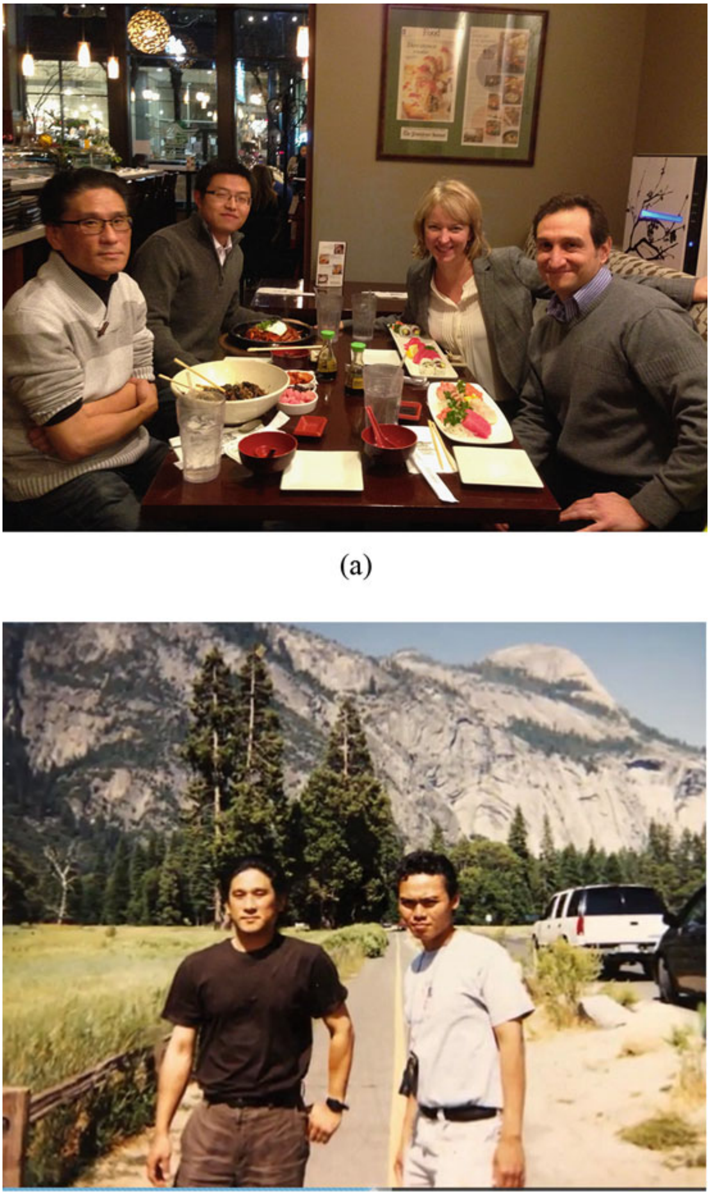
al., *Physics of Life Reviews*, 13, 115–118, 2019, https://doi.org/10.1016/j.plrev.2019.08.007

41. J. Araneo, B.J. Chung, M. Cristaldi, J. Pateras, A. Vaidya and R. Wulandana, Experimental control from wake induced autorotation with applications to energy harvesting, *International Journal of Green Energy*, 16(15), 1400–1413, 2019, https://doi.org/10.1080/15435075.2019.1671413
42. R. Wulandana, D. Foote, B.J. Chung and A. Vaidya, Vortex-induced autorotation potentials of bladeless turbine models, *International Journal of Green Energy*, 19(2), 190–200, 2022, https://doi.org/10.1080/15435075.2021.1941044
43. B.J. Chung, B. De Bari, J. Dixon, Dilip Kondepudi, Joseph Pateras, and Ashwin Vaidya, On the Thermodynamics of Self-Organization in Dissipative Systems: Reflections on the Unification of Physics and Biology, *Fluids*, 7(4), 141, 2022, https://doi.org/10.3390/fluids7040141



(a) (b) (c) (d) (e) (f)

Bong Jae with friends and colleagues and mentors in Pittsburgh between 2000–2004 (**a**, **b**, **c**) and Capo Miseno, Italy in 2001 (**d**, **e**, **f**)



(a)

(b)

Bong Jae with (**a**) colleagues in Fairfax, VA around 2008 and (**b**) fellow graduate student in California in 2000

Contents

Part I Theory

Part II Computation

Part III Experiments

Part IV Applications

Part I
Theory

Natural Second-Order Regularity for Systems in the Case $1 < p \le 2$ Using the A-Approximation

Luigi C. Berselli and Michael Růžička

1 Introduction

In this paper, we consider the boundary value problem associated with nonlinear elliptic systems:

$$\begin{cases} -\operatorname{div}\mathbf{S}(\mathbf{Du}) = \mathbf{f} & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega, \end{cases} \tag{1}$$

where the operator $\mathbf{S}$ depends on the symmetric gradient $\mathbf{Du}$ and has (p, δ)-structure (cf. Definition 3). Here, $\Omega \subset \mathbb{R}^3$ is a sufficiently smooth and bounded domain. The paradigmatic example for the operator in (1) is given by

$$\mathbf{S}(\mathbf{Du}) := (\delta + |\mathbf{Du}|)^{p-2}\mathbf{Du}, \quad \text{with} \quad \delta \ge 0,\ 1 < p < \infty\,. \tag{2}$$

Thus, problem (1) is a generalization to systems of the classical p-Laplace problem for scalars $\Delta_p u := \operatorname{div}(|\nabla u|^{p-2}\nabla u)$, which corresponds to the case $\delta = 0$. While the existence of weak solutions is a rather standard result—based on the theory of monotone operators—the regularity of solutions is more complicated and has been addressed for the case $1 < p \le 2$ by Seregin and Shilkin [22] (in the case of a flat boundary) and by the authors of the present paper in [8] (in a general smooth

L. C. Berselli (✉)
Dipartimento di Matematica, Università di Pisa, Pisa, Italy
e-mail: luigi.carlo.berselli@unipi.it

M. Růžička
Institute of Applied Mathematics, Albert-Ludwigs-University Freiburg, Freiburg, Germany
e-mail: rose@mathematik.uni-freiburg.de

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_1

domain). The proof is obtained by a classical strategy: the use of difference quotients to estimate partial derivatives in the tangential directions and ellipticity to recover normal derivatives. The main difficulties are those of justifying the calculations to make the argument rigorous. This has been done by means of (a) smoothing with the addition of an extra Laplace term $-\varepsilon\Delta\mathbf{u}^\varepsilon$ and (b) proving for the solution $\mathbf{u}^\varepsilon$ (of the approximate problem) estimates independent of $\varepsilon > 0$ to justify the limit as $\varepsilon \to 0$.

This approach cannot be used in the case $p > 2$, since the calculations—despite being formally very similar—are not justified. In fact, the added Laplacian term immediately implies estimates in $L^2(\Omega)$ for second-order partial derivatives of $\mathbf{u}^\varepsilon$, but this is still not enough to give proper meaning to all of the integrals appearing in the derivation of the various estimates.

To overcome this technical problem—very recently—we developed in [9] a theory based on the *(multiple) approximation* of the operator $\mathbf{S}$, which allows to treat the case $p > 2$, for all arbitrarily large p. The theory of the multiple approximation can also be applied in the case $1 < p \leq 2$ (in fact, a single approximation is enough in this case), providing an alternative proof for the results from [8, 22].

In this paper, we consider the case $1 < p \leq 2$, and we explain the modifications and simplifications of the theory with a "single" A-approximation. Even if the results we prove are not completely original, we believe it is important to explain them with great detail. This will be particularly interesting for students or younger researchers, since the developed method, which is highly flexible, can be adapted to several other problems. Even if we skip some details (which would make the presentation too long), we try to keep the presentation as much as possible self-contained. We refer with detailed citations to [7–10] for all missing technical details. We present a detailed presentation only in the elliptic case. Nevertheless, the method can be also applied to parabolic problems with minor modifications to recover in a different way results similar to those proved in [10] (see Sect. 6).

The main goal of this paper is to show how to prove a result of "natural" second-order regularity for weak solutions. This corresponds to proving—under possibly minimal assumptions on the data—that weak solutions (and not solutions with additional unproved properties) satisfy the following inequality:

$$\int\limits_\Omega (\delta + |\mathbf{Du}|)^{p-2}|\nabla\mathbf{Du}|^2\, d\mathbf{x} \leq C\,, \tag{3}$$

which can be equivalently rewritten as $\nabla\mathbf{F}(\mathbf{Du}) \in L^2(\Omega)$, where

$$\mathbf{F}(\mathbf{Du}) := (\delta + |\mathbf{Du}|)^{\frac{p-2}{2}}\,|\mathbf{Du}|\,. \tag{4}$$

The regularity coming from inequality (3) is called natural since if one restricts to the periodic case (and integration by parts can be done freely without boundary terms) this is formally obtained by multiplying the system in (1) by $-\Delta\mathbf{u}$ and performing straightforward integration by parts.

Remark 1 In the literature, the name natural is used to distinguish such regularity results from the so-called "optimal" second-order regularity (for which there exists also an intense research activity; see [1, 3, 12, 13]), which proves $\nabla \mathbf{S} \in L^2(\Omega)$, i.e.,

$$\int_\Omega \left|\nabla\big((\delta + |\mathbf{Du}|)^{p-2}\mathbf{Du}\big)\right|^2 d\mathbf{x} \le C\,.$$

The two notions of regularity are rather different in the spirit: the optimal regularity is linked with nonlinear versions of the singular integral theory, while the natural regularity is based on energy methods. The latter involves quasi-norms (cf. Barrett and Liu [2]), which are, among others, of crucial relevance for the numerical analysis of the problem, in particular, to prove optimal convergence rates for finite element discretizations.

In Sect. 2, we will give definitions of the missing notions and formulate general assumptions on the operator $\mathbf{S}$, covering the example (2) in the case $p \in (1, 2]$ and $\delta \in [0, \infty)$. Based on that, we consider the following notion of solution:

Definition 1 (Regular Solution) Let the operator $\mathbf{S}$ in (1) have (p, δ)-structure for some $p \in (1, \infty)$ and $\delta \in [0, \infty)$. We say that $\mathbf{u}$ is a regular solution to (1) if $\mathbf{u} \in W_0^{1,p}(\Omega)$ satisfies for all $\mathbf{w} \in W_0^{1,p}(\Omega)$

$$\int_\Omega \mathbf{S}(\mathbf{Du}) \cdot \mathbf{Dw}\, d\mathbf{x} = \int_\Omega \mathbf{f} \cdot \mathbf{w}\, d\mathbf{x}\,,$$

and fulfils

$$\mathbf{F}(\mathbf{Du}) \in W^{1,2}(\Omega)\,.$$

The main result we will prove with full details is the following:

Theorem 1 *Let the operator* $\mathbf{S}$ *in* (1)*, derived from a potential* U*, have* (p, δ)*-structure for some* $p \in (1, 2]$ *and* $\delta \in [0, \infty)$*. Let* $\Omega \subset \mathbb{R}^3$ *be a bounded domain with* $C^{2,1}$ *boundary. Assume that* $\mathbf{f} \in L^{p'}(\Omega)$*. Then, the system* (1) *has a unique regular solution with norms estimated only in terms of the characteristics of* $\mathbf{S}$*,* δ*,* Ω*, and* $\|\mathbf{f}\|_{p'}$*.*

The counterpart in the parabolic case (cf. Theorem 2) will be presented, without a detailed proof, in the final section. Moreover, for all results, we will study only the nondegenerate case $\delta > 0$. The degenerate case can be handled by a limiting argument, provided that the estimates do not degenerate as $\delta \to 0$, exactly as in [8, Sec. 3.2]. The proof of such estimates requires some changes with respect to the ones obtained in [9] (for $p > 2$) and in [10] (for $p < 2$, but with a different approximation) related to the initial condition. Such estimates are available in our setting (cf. Propositions 6, 9, 15). In fact, the limiting process $\delta \to 0$ depends only on the regularity available and is independent of the method used to prove it; hence, there is nothing to change with respect to the already available proof in [8].

Plan of the Paper In Sect. 2, we recall the main facts about N-functions and the A-approximation. Sections 3 and 4 are devoted to the proof of the existence and regularity for the solutions of the approximated problem. Especially Sect. 4 is crucial for the estimates independent of A. Section 5 explains the limiting process to come back from the A-approximate system to the original one. Finally, in Sect. 6, the corresponding results in the parabolic setting are presented.

2 On the A-Approximation of an Operator and Its Properties

In this section, we introduce the notation and the crucial properties of N-functions, which will be used to prove the relevant properties of A-approximated operators. We summarize and recall the main results already proved with full details in [9, 17], i.e., proofs of all statements in this section can be found in these references.

2.1 Notation

We use c, C to denote generic constants, which may change from line to line, but are not depending on the crucial quantities. Moreover, we write $f \sim g$ if and only if there exists constants $c, C > 0$ such that $c\, f \le g \le C\, f$.

We use the customary Lebesgue spaces $(L^p(\Omega), \|\,.\,\|_p)$, $p \in [1, \infty]$, and Sobolev spaces $(W^{k,p}(\Omega), \|\,.\,\|_{k,p})$, $p \in [1, \infty]$, $k \in \mathbb{N}$. We do not distinguish between scalar, vector-valued, or tensor-valued function spaces; however, we denote scalar functions by roman letters, vector-valued functions by small boldfaced letters, and tensor-valued functions by capital boldfaced letters. We denote by $|M|$ the *three*-dimensional Lebesgue measure of a measurable set M. As usual the gradient of a vector field $\mathbf{v} : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ is denoted as $\nabla\mathbf{v} = (\partial_i v_j)_{i,j=1,2,3} = (\partial_i \mathbf{v})_{i=1,2,3}$, while its symmetric part is denoted as $\mathbf{D}\mathbf{v} := \frac{1}{2}\big(\nabla\mathbf{v} + \nabla\mathbf{v}^\top\big)$. The derivative of functions defined on tensors, i.e., $U : \mathbb{R}^{3\times 3} \to \mathbb{R}$, is denoted as $\partial U = (\partial_{ij} U)_{i,j=1,2,3}$ where ∂_{ij} are the partial derivatives with respect to the canonical basis of $\mathbb{R}^{3\times 3}$.

2.2 N-Functions

A function $\varphi : \mathbb{R}^{\ge 0} \to \mathbb{R}^{\ge 0}$ is called an *N-function* if φ is continuous, convex, and strictly positive for $t > 0$ and satisfies[1]

[1] In the following, we use the convention that $\frac{\varphi'(0)}{0} := 0$.

$$\lim_{t\to 0^+}\frac{\varphi(t)}{t}=0\,,\qquad\qquad \lim_{t\to\infty}\frac{\varphi(t)}{t}=\infty\,.$$

If φ additionally belongs to $C^1(\mathbb{R}^{\geq 0})\cap C^2(\mathbb{R}^{>0})$ and satisfies $\varphi''(t)>0$ for all $t>0$, we call φ a *regular N-function*. In the rest of the paper, we restrict ourselves to this case and note that for a regular N-function we have $\varphi(0)=\varphi'(0)=0$. Moreover, φ' is increasing and $\lim_{t\to\infty}\varphi'(t)=\infty$. For details, we refer to [16, 18, 20, 21]. For a regular N-function φ, we define the *complementary function* φ^* via

$$\varphi^*(t):=\int_0^t(\varphi')^{-1}(s)\,ds\,.$$

One easily sees that φ^* is a regular N-function, too.

The Δ_2-*condition* plays an important role in Orlicz spaces. A nondecreasing function $\varphi:\mathbb{R}^{\geq 0}\to\mathbb{R}^{\geq 0}$ is said to satisfy the Δ_2-condition (in short $\varphi\in\Delta_2$), if for some constant $K\geq 2$ it holds

$$\varphi(2t)\leq K\varphi(t),\qquad \forall\, t\geq 0\,.$$

The Δ_2-constant (the smallest of such $K\geq 2$) of φ is denoted by $\Delta_2(\varphi)$.

It has been recently recognized that a fundamental role in regularity theory of problem similar to (1) is played by the notion of *balanced N-function* (cf. [9, 12, 21]). A regular N-function φ is called *balanced* if there exist constants $\gamma_1\in(0,1]$ and $\gamma_2\geq 1$ such that there holds

$$\gamma_1\,\varphi'(t)\leq t\,\varphi''(t)\leq\gamma_2\,\varphi'(t),\qquad \forall\, t>0\,.$$

The pair (γ_1,γ_2) is called *characteristics* of the balanced N-function φ. The property of being balanced transmits to φ^*, whose characteristics are $(\gamma_2^{-1},\gamma_1^{-1})$. Note that for a balanced N-function φ, we have the equivalences

$$\varphi(t)\sim\varphi'(t)\,t\sim\varphi''(t)\,t^2,\qquad \forall\, t>0$$

with constants of equivalence depending only on the characteristics of φ. In view of this, it is convenient to introduce the particular notation $a_\varphi:\mathbb{R}^{\geq 0}\to\mathbb{R}^{\geq 0}$, defined for regular N-functions φ via

$$a_\varphi(t):=\frac{\varphi'(t)}{t}\,.$$

Another important tool is *shifted N-functions* $\{\phi_a\}_{a\geq 0}$, defined for $t\geq 0$, by

$$\varphi_a(t):=\int_0^t\varphi_a'(s)\,ds\qquad\text{with}\qquad \phi_a'(t):=\phi'(a+t)\frac{t}{a+t}\,.$$

For an N–function $\phi \in \Delta_2$, there holds for all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ and all $t \geq 0$ that $\phi_{|\mathbf{P}|}\big(|\mathbf{P}-\mathbf{Q}|\big) \sim \phi_{|\mathbf{Q}|}\big(|\mathbf{P}-\mathbf{Q}|\big)$ with constants of equivalence depending only on $\Delta_2(\phi')$. The most relevant property for us is a change of shift.

Lemma 1 (Change of Shift) *Let ϕ be an N-function such that ϕ and ϕ^* satisfy the Δ_2-condition. Then, for all $\delta \in (0,1)$, there exists $c_\varepsilon = c_\varepsilon(\Delta_2(\phi'))$ such that for all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ and all $t \geq 0$ there holds*

$$\begin{aligned}\phi_{|\mathbf{P}|}(t) &\leq c_\varepsilon\, \phi_{|\mathbf{Q}|}(t) + \varepsilon\, \phi_{|\mathbf{P}|}\big(|\mathbf{P}-\mathbf{Q}|\big)\,,\\ \big(\phi_{|\mathbf{P}|}\big)^*(t) &\leq c_\varepsilon\, \big(\phi_{|\mathbf{Q}|}\big)^*(t) + \varepsilon\, \phi_{|\mathbf{P}|}\big(|\mathbf{P}-\mathbf{Q}|\big)\,.\end{aligned}$$

Proof These inequalities are proved in [21, Lemma 5.15, Lemma 5.18]. □

Finally, we introduce for $p \in (1,\infty)$ and $\delta \in [0,\infty)$ the function $\omega_{p,\delta} : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ via

$$\omega_{p,\delta}(t) := \int_0^t (\delta+s)^{p-2} s\, ds, \qquad \forall\, t \geq 0\,.$$

Remark 2 The function $\omega_{p,\delta}(t)$ is precisely the N-function associated with the canonical example for the operator $\mathbf{S}$ in (2). If p and δ are fixed (and to avoid confusion with shifted functions), we simply write $\omega(t) := \omega_{p,\delta}(t)$.

Clearly, ω is a regular N-function for all $p \in (1,\infty)$ and $\delta \in [0,\infty)$. More precisely, for $p \leq 2$, we have:

Lemma 2 *For any $p \in (1,2]$ and for any $\delta \in [0,\infty)$, there holds*

$$\begin{aligned}\omega(t) \leq (\omega)'(t)\, t\ &\leq 2^{p+1}\omega(t), \qquad && \forall\, t \geq 0\,,\\ (p-1)\,(\omega)'(t) \leq (\omega)''(t)\, t\ &\leq\ (\omega)'(t), && \forall\, t > 0\,.\end{aligned} \tag{5}$$

In particular, the function ω is a balanced N-function with characteristics $(p-1,1)$ and Δ_2-constant depending only on p. Moreover, also ω^* is a balanced N-function with characteristics $(1,(p-1)^{-1})$ and Δ_2-constant depending only on p.

For the shifts of ω and its complementary function ω^*, there hold for all $a \geq 0$ the equivalences $\omega_a(t) \sim (\delta+a+t)^{p-2}t^2$ and $(\omega_a)^*(t) \sim \big((\delta+a)^{p-1}+t\big)^{p'-2}t^2$.

2.3 Nonlinear Operators with (p,δ)-Structure

In this section, we collect the main properties of nonlinear operators derived from a potential and of operators having (p,δ)-structure.

Definition 2 (Operator Derived from a Potential) We say that an operator $\mathbf{S}: \mathbb{R}^{3\times 3} \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ is *derived from a potential* $U: \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$, and write $\mathbf{S} = \partial U$ if $\mathbf{S}(\mathbf{0}) = \mathbf{0}$ and for all $\mathbf{P} \in \mathbb{R}^{3\times 3} \setminus \{\mathbf{0}\}$ there holds

$$\mathbf{S}(\mathbf{P}) = \partial U(|\mathbf{P}^{\mathrm{sym}}|) = \frac{U'(|\mathbf{P}^{\mathrm{sym}}|)}{|\mathbf{P}^{\mathrm{sym}}|}\,\mathbf{P}^{\mathrm{sym}} = a_U(|\mathbf{P}^{\mathrm{sym}}|)\,\mathbf{P}^{\mathrm{sym}},$$

for some $U \in C^1(\mathbb{R}^{\geq 0}) \cap C^2(\mathbb{R}^{>0})$ satisfying $U(0) = U'(0) = 0$.

Definition 3 (Operator with a φ-Structure) Let the operator $\mathbf{S}\colon \mathbb{R}^{3\times 3} \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$, belonging to $C^0(\mathbb{R}^{3\times 3}; \mathbb{R}^{3\times 3}_{\mathrm{sym}}) \cap C^1(\mathbb{R}^{3\times 3} \setminus \{\mathbf{0}\}; \mathbb{R}^{3\times 3}_{\mathrm{sym}})$, satisfy $\mathbf{S}(\mathbf{P}) = \mathbf{S}\big(\mathbf{P}^{\mathrm{sym}}\big)$ and $\mathbf{S}(\mathbf{0}) = \mathbf{0}$. We say that $\mathbf{S}$ has *φ-structure* if there exist a regular N-function φ and constants $\gamma_3 \in (0, 1]$, $\gamma_4 > 1$ such that the inequalities

$$\sum_{i,j,k,l=1}^{3} \partial_{kl} S_{ij}(\mathbf{P})\, Q_{ij}\, Q_{kl} \geq \gamma_3\, a_\varphi(|\mathbf{P}^{\mathrm{sym}}|)\, |\mathbf{P}^{\mathrm{sym}}|^2\,,$$

$$\big|\partial_{kl} S_{ij}(\mathbf{P})\big| \leq \gamma_4\, a_\varphi(|\mathbf{P}^{\mathrm{sym}}|)\,,$$

are satisfied for all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{3\times 3}$ with $\mathbf{P}^{\mathrm{sym}} \neq \mathbf{0}$ and all $i, j, k, l = 1, 2, 3$. The constants γ_3, γ_4, and $\Delta_2(\varphi)$ are called the *characteristics* of $\mathbf{S}$ and will be denoted by $(\gamma_3, \gamma_4, \Delta_2(\varphi))$.

If $\varphi = \omega_{p,\delta}$ with $p \in (1, \infty)$ and $\delta \in [0, \infty)$ we say that $\mathbf{S}$ has *(p, δ)-structure* and call (γ_3, γ_4, p) its characteristics.

Closely related to an operator with φ-structure is the function $\mathbf{F}_\varphi\colon \mathbb{R}^{3\times 3} \to \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ defined via

$$\mathbf{F}_\varphi(\mathbf{P}) := \frac{\sqrt{\varphi'(|\mathbf{P}^{\mathrm{sym}}|)|\mathbf{P}^{\mathrm{sym}}|}}{|\mathbf{P}^{\mathrm{sym}}|}\,\mathbf{P}^{\mathrm{sym}} = \sqrt{a_\varphi(|\mathbf{P}^{\mathrm{sym}}|)}\,\mathbf{P}^{\mathrm{sym}}\,, \tag{6}$$

where the first representation holds only for $\mathbf{P}^{\mathrm{sym}} \neq \mathbf{0}$. In the special case of an operator $\mathbf{S}$ with (p, δ)-structure, we have (recall that $\omega = \omega_{p,\delta}$)

$$\mathbf{F}(\mathbf{P}) := \mathbf{F}_\omega(\mathbf{P}) = \sqrt{a_{\omega(|\mathbf{P}^{\mathrm{sym}}|)}}\,\mathbf{P}^{\mathrm{sym}} = \big(\delta + |\mathbf{P}^{\mathrm{sym}}|\big)^{\frac{p-2}{2}}\,\mathbf{P}^{\mathrm{sym}}\,,$$

which is consistent with the notation used in the previous literature, cf. (4).

If φ is a balanced N-function with characteristics (γ_1, γ_2), then $\mathbf{S} = \partial\varphi$ is an operator with φ-structure and with characteristics depending only on γ_1 and γ_2. The following result will be crucial for our investigations (cf. [21, Section 6]).

Proposition 1 *Let φ be a balanced N-function with characteristics (γ_1, γ_2). Let $\mathbf{S}$ have φ-structure with characteristics $(\gamma_3, \gamma_4, \Delta_2(\varphi))$ and let $\mathbf{F}_\varphi$ be defined in* (6)*. Then, we have for all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{3\times 3}$ that*

$$
\begin{aligned}
\big(\mathbf{S}(\mathbf{P})-\mathbf{S}(\mathbf{Q})\big)\cdot(\mathbf{P}-\mathbf{Q}) &\sim a_\varphi(|\mathbf{P}^{\mathrm{sym}}|+|\mathbf{P}^{\mathrm{sym}}-\mathbf{Q}^{\mathrm{sym}}|)\,|\mathbf{P}^{\mathrm{sym}}-\mathbf{Q}^{\mathrm{sym}}|^2\\
&\sim |\mathbf{F}_\varphi(\mathbf{P})-\mathbf{F}_\varphi(\mathbf{Q})|^2\,,\\
|\mathbf{S}(\mathbf{P})-\mathbf{S}(\mathbf{Q})| &\sim a_\varphi(|\mathbf{P}^{\mathrm{sym}}|+|\mathbf{P}^{\mathrm{sym}}-\mathbf{Q}^{\mathrm{sym}}|)\,|\mathbf{P}^{\mathrm{sym}}-\mathbf{Q}^{\mathrm{sym}}|\,,
\end{aligned}
$$

where the constants of equivalence depend only on $\gamma_1, \gamma_2, \gamma_3$*, and* γ_4.

In addition, the following result will be used to handle operators derived from a potential.

Proposition 2 *Let the operator* $\mathbf{S} = \partial U$*, derived from the potential* U*, have* φ*-structure, with characteristics* $(\gamma_3, \gamma_4, \Delta_2(\varphi))$*. If* φ *is a balanced N-function with characteristics* (γ_1, γ_2)*, then* U *is a balanced N-function satisfying for all* $t > 0$

$$
\frac{\gamma_3}{\gamma_2}\varphi''(t) \le U''(t) \le \frac{\gamma_4}{\gamma_1}\varphi''(t)\,.
$$

The characteristics of U *is equal to* $\big(\frac{\gamma_3}{\gamma_4}\frac{\gamma_1^2}{\gamma_2}, \frac{\gamma_4}{\gamma_3}\frac{\gamma_2^2}{\gamma_1}\big)$.

The significance of this proposition is that a general operator $\mathbf{S}$ derived from a potential U with (p, δ)-structure can be simply handled as the explicit example (2).

2.4 Approximation of a Nonlinear Operator

We now define the A-approximation of a function and of an operator and prove the relevant properties needed in the sequel. This approximation was introduced in [17] for $p > 2$ and generalized in the recent paper [9] to a so-called (A, q)-approximation, for some $q \ge 2$, which allows for a unified approach for all $p \in (1, \infty)$. The purpose of the A-approximation of a function is to have quadratic behavior near infinity (cf. [17, Lemma 2.22]), and consequently, one can take advantage of the standard Hilbertian theory.

Definition 4 (A**-Approximation of a Scalar Real Function)** Given a function $U \in C^1(\mathbb{R}^{\ge 0}) \cap C^2(\mathbb{R}^{>0})$ satisfying $U(0) = U'(0) = 0$, we define for $A \ge 1$ the *A-approximation* $U^A \in C^1(\mathbb{R}^{\ge 0}) \cap C^2(\mathbb{R}^{>0})$ via

$$
U^A(t) := \begin{cases} U(t) & t \le A\,,\\ \alpha_2\, t^2 + \alpha_1\, t + \alpha_0 & t > A\,. \end{cases}
$$

To ensure continuity up to second-order derivatives, the constants $\alpha_i = \alpha_i(U)$, $i = 0, 1, 2$, are given via

$$
\begin{aligned}
\alpha_2 &= \frac{1}{2}\, U''(A)\,,\\
\alpha_1 &= U'(A) - \, U''(A)\, A\,,\\
\alpha_0 &= U(A) - U'(A)\, A + \frac{1}{2} U''(A)\, A^2\,.
\end{aligned}
$$

Remark 3 If φ is a regular N-function, the definition of φ^A, together with the properties of φ, implies that there exists a constant $c(A, \varphi)$ such that for all $t \geq 0$ there holds

$$
a_{\varphi^A}(t) = \frac{(\varphi^A)'(t)}{t} \leq c(A, \varphi)\,.
$$

More precise (explicit) upper and lower bounds are given in (8) if $\varphi = \omega$.

We have the following relevant result (cf. [9, Lemma 2.42]) linking balanced functions with their A-approximations.

Lemma 3 *Let φ be a balanced N-function with characteristics (γ_1, γ_2). Then, for all $A \geq 1$, it holds that φ^A is also balanced with characteristics $\big(\gamma_1, \gamma_2\big)$.*

Concerning the homogeneity of the function ω for $1 < p \leq 2$ (similar results could be deduced also in the case $p > 2$), we have the following result:

Lemma 4 *Let $1 < p \leq 2$ and $\delta \in [0, \infty)$. The functions $\omega(t)$ and $\omega^A(t)$, for any $A \geq 1$, are balanced functions with characteristics $(p - 1, 1)$. Moreover, it holds for all $\lambda,\ t \geq 0$ that*

$$
\omega(\lambda\, t) \leq \max\{\lambda, \lambda^2\}\, \omega(t) \qquad \textit{and} \qquad \omega^A(\lambda\, t) \leq \max\{\lambda, \lambda^2\}\, \omega^A(t)\,. \tag{7}
$$

Proof The assertion on the characteristics for both functions follows directly from Lemmas 2 and 3 (which shows that they are unchanged by the A- approximation). The estimates in (7) are proved by observing that if ϕ is a regular N-function with characteristics (γ_1, γ_2), then it follows for all $t > 0$ that

$$
\frac{d}{dt} \log(\varphi'(t)) = \frac{\varphi''(t)}{\varphi'(t)} \leq \gamma_2 \frac{1}{t}\,,
$$

which implies, by integration with respect to t over $(s, \lambda s)$, with $\lambda > 1$ and $s > 0$, and using the exponential function, that

$$
\frac{\varphi'(\lambda\, s)}{\varphi'(s)} \leq \lambda^{\gamma_2}\,.
$$

A further integration with respect to s over $(0, t)$, $t > 0$, proves

$$\varphi(\lambda\, t) \le \lambda^{\gamma_2+1}\varphi(t)\,, \qquad \forall\, t > 0\,.$$

The case $t = 0$ is trivial, and in the case $0 \le \lambda \le 1$ and $t \ge 0$, the proof ends by observing that $\phi\big((1-\lambda)\,0 + \lambda\, t\big) \le \lambda\phi(t)$, by the convexity of ϕ. □

Next, we define the A-approximation of an operator derived from a potential.

Definition 5 (A-Approximation of an Operator Derived from a Potential) Let the operator $\mathbf{S} = \partial U$ be derived from the potential U. Then, we define for given $A \ge 1$ the *A-approximation* $\mathbf{S}^A := \partial U^A$ as the operator derived from the potential U^A, i.e., $\mathbf{S}^A$ satisfies $\mathbf{S}^A(\mathbf{0}) = \mathbf{0}$, and for all $\mathbf{P} \in \mathbb{R}^{3\times 3} \setminus \{\mathbf{0}\}$, there holds

$$\mathbf{S}^A(\mathbf{P}) := \partial U^A(|\mathbf{P}^{\rm sym}|) = \frac{(U^A)'(|\mathbf{P}^{\rm sym}|)}{|\mathbf{P}^{\rm sym}|}\,\mathbf{P}^{\rm sym} = a_{U^A}(|\mathbf{P}^{\rm sym}|)\,\mathbf{P}^{\rm sym}\,.$$

The properties of the operator $\mathbf{S}$ in Proposition 1 are inherited by the operator $\mathbf{S}^A$. More precisely, we have (cf. [9, Prop. 2.47]):

Proposition 3 *Let φ be a balanced N-function with characteristics (γ_1, γ_2). Let the operator $\mathbf{S} = \partial U$, derived from the potential U, have φ-structure with characteristics $(\gamma_3, \gamma_4, \Delta_2(\varphi))$. For $A \ge 1$, let φ^A and $\mathbf{S}^A$ be the A-approximation of φ and $\mathbf{S}$, respectively. Then, we have for all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{3\times 3}$ that*

$$\begin{aligned}(\mathbf{S}^A(\mathbf{P}) - \mathbf{S}^A(\mathbf{Q}))\cdot(\mathbf{P}-\mathbf{Q}) &\sim a_{\phi^A}(|\mathbf{P}^{\rm sym}| + |\mathbf{P}^{\rm sym} - \mathbf{Q}^{\rm sym}|)\,|\mathbf{P}^{\rm sym} - \mathbf{Q}^{\rm sym}|^2\,,\\ &\sim |\mathbf{F}_{\phi^A}(\mathbf{P}) - \mathbf{F}_{\phi^A}(\mathbf{Q})|^2,\\ |\mathbf{S}^A(\mathbf{P}) - \mathbf{S}^A(\mathbf{Q})| &\sim a_{\phi^A}(|\mathbf{P}^{\rm sym}| + |\mathbf{P}^{\rm sym} - \mathbf{Q}^{\rm sym}|)\,|\mathbf{P}^{\rm sym} - \mathbf{Q}^{\rm sym}|\,,\end{aligned}$$

with constants of equivalence depending only on $\gamma_1, \gamma_2, \gamma_3$, and γ_4.

Remark 4 For the limiting process, it is of fundamental relevance that in Proposition 3 the constants do not depend on $A \ge 1$. The details of the proof can be found in [9, Sec. 2].

Equivalent expressions for $\nabla\mathbf{F}(\mathbf{Du})$ (based on Proposition 3) play a crucial role in the proof of regularity of weak solutions. To this end, we define, for a sufficiently smooth operator $\mathbf{S} : \mathbb{R}^{3\times 3} \to \mathbb{R}^{3\times 3}_{\rm sym}$, the functions $\mathbb{P}_i^A : \mathbb{R}^{3\times 3} \to \mathbb{R}$, $i = 1, 2, 3$, via

$$\mathbb{P}_i^A(\mathbf{P}) := \partial_i\mathbf{S}^A(\mathbf{P})\cdot\partial_i\mathbf{P} = \sum_{j,k,l,m=1}^{3} \partial_{jk}S^A_{lm}(\mathbf{P})\,\partial_i P_{jk}\,\partial_i P_{lm}\,,$$

and emphasize that there is no summation over the index i.

The abovementioned properties of balanced functions allow us to deal with the problem associated with $\mathbf{S} = \partial U$ with (p, δ)-structure using the quantities related with ω and not with U itself, greatly simplifying both the presentation and the estimates, cf. Proposition 2. For this reason, we also introduce the following notation, consistent with (4):

$$\mathbf{F}^A(\mathbf{P}) := \mathbf{F}_{\omega^A}(\mathbf{P}) \qquad \text{and} \qquad a^A(t) := a_{\omega^A}(t)\,.$$

Using this notation, we have the following result (cf. [9, Prop. 2.49], [8, Prop. 2.4]):

Proposition 4 *Let the operator $\mathbf{S} = \partial U$, derived from the potential U, have (p, δ)-structure for some $p \in (1, \infty)$ and $\delta \in [0, \infty)$, with characteristics (γ_3, γ_4, p). If for a vector field $\mathbf{v}\colon \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ there holds $\mathbf{F}^A(\mathbf{Dv}) \in W^{1,2}(\Omega)$, then we have for $i = 1, 2, 3$ and a.e. in Ω the following equivalences:*

$$\begin{aligned} |\partial_i \mathbf{F}^A(\mathbf{Dv})|^2 &\sim a^A(|\mathbf{Dv}|)\,|\partial_i \mathbf{Dv}|^2 \\ &\sim \mathbb{P}_i^A(\mathbf{Dv})\,, \\ |\partial_i \mathbf{S}^A(\mathbf{Dv})|^2 &\sim a^A(|\mathbf{Dv}|)\,\mathbb{P}_i^A(\mathbf{Dv})\,, \end{aligned}$$

where the constants of equivalence depend only on γ_3, γ_4, and p.

In view of this proposition, it is important to have upper and lower bounds for a^A in order to control various quantities related to $\mathbf{F}^A$, in terms of $\mathbf{F}$. Crucial in this respect is the following result (cf. [9, Lem. 2.69]):

Lemma 5 *For $p \in (1, 2]$, $\delta > 0$, and $A \geq 1$, the function $a^A(t)$ is nonincreasing, and for all $t \geq 0$, there holds*

$$\begin{aligned} (p-1)\,a(t) \leq a^A(t) \leq \delta^{p-2}\,, \\ (p-1)\,(\delta + A)^{p-2} \leq a^A(t)\,. \end{aligned} \tag{8}$$

Proof The statement is clear for $t \leq A$ using $a^A(t) = a(t) = (\delta + t)^{p-2}$, $0 \leq \delta$, $t \leq A$, and $p \leq 2$. For $t \geq A$, we have $a^A(t) = \omega''(A) + \frac{\omega'(A) - \omega''(A)\,A}{t}$. Thus, we get that $a^A(A) = (\delta + A)^{p-2}$, $\lim_{t\to\infty} a^A(t) = (\delta + A)^{p-3}\big(\delta + (p-1)A\big)$, and $(a^A)'(t) = -\frac{\omega'(A) - \omega''(A)\,A}{t^2} \leq 0$ in view of (5), and $p \leq 2$. This yields

$$(\delta + A)^{p-2} \geq a^A(t) \geq (\delta + A)^{p-3}((p-1)A + \delta) \geq (p-1)\,(\delta + A)^{p-2}\,,$$

which implies the assertions using $\delta^{p-2} \geq (\delta + A)^{p-2}$ and $(\delta + A)^{p-2} \geq (\delta + t)^{p-2}$ in view of $t \geq A$, and $p \leq 2$. □

In the sequel, we will use frequently the following consequences (cf. [9, Cor. 2.71]):

Corollary 1 *Let the operator* $\mathbf{S}$*, derived from the potential* U*, have* (p, δ)*-structure for some* $p \in (1, 2]$ *and* $\delta > 0$*, with characteristics* (γ_3, γ_4, p)*. Then, there holds for all* $t \geq 0$ *that*

$$\frac{(p-1)}{2}(\delta + A)^{p-2}t^2 \leq \omega^A(t)\,,$$

$$(p-1)\,\omega(t) \leq \omega^A(t) \leq \frac{\delta^{p-2}}{2}\,t^2\,,$$

$$(\omega^A)^*(t) \leq (p-1)\,(\Delta_2(\omega^*))^M\,\omega^*(t)\,,$$

where $M \in \mathbb{N}_0$ *is chosen such that* $(p-1)^{-1} \leq 2^M$*. Moreover, for all* $\mathbf{P} \in \mathbb{R}^{3\times 3}$*, there holds*

$$|\mathbf{F}^A(\mathbf{P})|^2 \sim \omega^A(|\mathbf{P}^{\text{sym}}|)\,,$$

$$c\,|\mathbf{F}(\mathbf{P})|^2 \leq |\mathbf{F}^A(\mathbf{P})|^2\,,$$

$$|\mathbf{S}^A(\mathbf{P})| \leq c\,\delta^{p-2}|\mathbf{P}^{\text{sym}}|\,,$$

with constants c depending only on γ_3, γ_4*, and* p*.*

Corollary 2 *Under the assumptions of Proposition 4, there exists* $c(p, \gamma_i) > 0$ *such that*

$$c(p, \gamma_i)|\nabla\mathbf{F}(\mathbf{Dv})|^2 \leq |\nabla\mathbf{F}^A(\mathbf{Dv})|^2.$$

Proof This follows immediately from Proposition 4, Lemma 5, and [8, Prop. 2.4]. □

3 On the Existence and Uniqueness of Regular Solutions for the Approximate Problem

In this section, we introduce the approximate problem and prove existence, uniqueness, and regularity of its solutions. In fact, to prove Theorem 1, we use an approximate problem, obtained by replacing the operator $\mathbf{S} = \partial U$ with (p, δ)-structure by $\mathbf{S}^A = \partial U^A$ which has $(2, \delta)$ structure, i.e., we study

$$\begin{cases} -\operatorname{div}\mathbf{S}^A(\mathbf{D}\mathbf{u}^A) = \mathbf{f} & \text{in } \Omega\,, \\ \mathbf{u}^A = \mathbf{0} & \text{on } \partial\Omega\,. \end{cases} \tag{9}$$

This system can be treated by standard techniques typical for linear equations. This procedure yields various estimates independent of $A \geq 1$ for the solution $\mathbf{u}^A$, which will enable us to pass to the limit $A \to \infty$ and to show that the limit $\mathbf{u} = \lim_{A\to\infty} \mathbf{u}^A$ will be a regular solution of the original problem (1).

The first standard result concerns the existence and uniqueness of weak solutions for (9).

Proposition 5 *Let the operator $\mathbf{S} = \partial U$, derived from the potential U, have (p, δ)-structure for some $p \in (1, 2]$ and $\delta \in (0, \infty)$. Assume that $\mathbf{f} \in L^{p'}(\Omega)$. Let $\mathbf{S}^A$ be as in Definition 5. Then, the approximate problem (9) possesses a unique weak solution, i.e., $\mathbf{u}^A \in W_0^{1,2}(\Omega)$ with $\mathbf{F}^A(\mathbf{Du}^A) \in L^2(\Omega)$ satisfies for all $\mathbf{w} \in W_0^{1,2}(\Omega)$*

$$\int\limits_\Omega \mathbf{S}^A(\mathbf{Du}^A)\cdot \mathbf{Dw}\, d\mathbf{x} = \int\limits_\Omega \mathbf{f}\cdot\mathbf{w}\, d\mathbf{x}\,. \tag{10}$$

This solution satisfies the estimate

$$\begin{aligned} &\|\mathbf{F}^A(\mathbf{Du}^A)\|_2^2 + (p-1)(\delta + A)^{p-2}\|\mathbf{Du}^A\|_2^2 \\ &\quad + (p-1)\big(\|\mathbf{F}(\mathbf{Du}^A)\|_2^2 + \|\mathbf{Du}^A\|_p^p\big) \leq C \int\limits_\Omega \omega^*(|\mathbf{f}|)\, d\mathbf{x}\,, \end{aligned} \tag{11}$$

with C depending only on the characteristics of $\mathbf{S}$ and Ω.

Remark 5 The energy-type estimate (11), which is obtained by testing with $\mathbf{u}^A$, implies that: (i) $\mathbf{u}^A \in W_0^{1,2}(\Omega)$ with norms depending on A; (ii) $\mathbf{u}^A \in W_0^{1,p}(\Omega)$ with norms bounded uniformly with respect to $A \geq 1$.

Proof of Proposition 5 The proof is based on a classical Faedo-Galerkin approximation of (9). The existence of Galerkin solutions $\mathbf{u}_k^A$, for $k \in \mathbb{N}$, follows by a standard argument based on Brouwer fixed point theorem. Passing to the limit as $k \to \infty$ (for A fixed) can be done within the standard theory of monotone operators (Minty-Browder theory). Since this is a fully standard argument, we just derive the a priori estimates necessary for this procedure.

By using $\mathbf{u}_k^A$ as test function in the Galerkin approximation for $\mathbf{u}_k^A$, we get

$$\begin{aligned} c\,\|\mathbf{F}^A(\mathbf{Du}_k^A)\|_2^2 &\leq c_\varepsilon \int\limits_\Omega (\omega^A)^*(|\mathbf{f}|)\, d\mathbf{x} + \varepsilon \int\limits_\Omega \omega^A(|\mathbf{u}_k^A|)\, d\mathbf{x} \\ &\leq c_\varepsilon \int\limits_\Omega (\omega^A)^*(|\mathbf{f}|)\, d\mathbf{x} + \varepsilon\, C \int\limits_\Omega \omega^A(|\mathbf{Du}_k^A|)\, d\mathbf{x}\,, \end{aligned}$$

where we used in the first line Proposition 3 with $\mathbf{Q} = \mathbf{0}$ together with Young inequality and in the second line

$$\int_\Omega \omega^A(|\mathbf{u}_k^A|)\,d\mathbf{x} \le C_P \int_\Omega \omega^A(|\nabla \mathbf{u}_k^A|)\,d\mathbf{x} \le C_P C_K \int_\Omega \omega^A(|\mathbf{D}\mathbf{u}_k^A|)\,d\mathbf{x}\,,$$

which follows from modular versions of Poincaré and Korn inequalities in Orlicz spaces (see [5, 9, 23]). Moreover, we absorb the last term on the right-hand side of the previous estimate using $\int_\Omega \omega^A(|\mathbf{D}\mathbf{u}_k^A|)\,d\mathbf{x} \sim \|\mathbf{F}^A(\mathbf{D}\mathbf{u}_k^A)\|_2^2$ in view of Corollary 1. Note that all constants are independent of $A \ge 1$ and depend only on the characteristics of $\mathbf{S}$ and on Ω. Moreover, from Corollary 1, it also follows that

$$\int_\Omega (\omega^A)^*(|\mathbf{f}|)\,d\mathbf{x} \le c(p) \int_\Omega \omega^*(|\mathbf{f}|)\,d\mathbf{x} \le C(p)\Big(\delta^p + \int_\Omega |\mathbf{f}|^{p'}\,d\mathbf{x}\Big)\,, \tag{12}$$

where the last estimate shows that the right-hand side in (11) is finite. Hence, after the limiting procedure $k \to \infty$, we arrive at

$$\|\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)\|_2^2 \le C \int_\Omega \omega^*(|\mathbf{f}|)\,d\mathbf{x}\,, \tag{13}$$

for some C independent of A and δ. Uniqueness follows from the strict monotonicity of $\mathbf{S}^A$ (cf. Proposition 3). By using the estimates in Corollary 1 and the definition of $\mathbf{F}^A$, we derive from (13) the various terms in the estimate (11), which ends the proof. □

3.1 Description and Properties of the Boundary

We assume that the boundary $\partial\Omega$ is of class $C^{2,1}$, that Is, for each point $P \in \partial\Omega$, there are local coordinates such that in these coordinates we have $P = 0$ and $\partial\Omega$ is locally described by a $C^{2,1}$-function, i.e., there exist $R_P,\ R'_P \in (0,\infty),\ r_P \in (0,1)$, and a $C^{2,1}$-function $g_P : B^2_{R_P}(0) \to B^1_{R'_P}(0)$ such that

(b1) $\mathbf{x} \in \partial\Omega \cap (B^2_{R_P}(0) \times B^1_{R'_P}(0)) \iff x_3 = g_P(x_1,x_2)\,,$

(b2) $\Omega_P := \{(x',x_3)\,\big|\,x' = (x_1,x_2) \in B^2_{R_P}(0),\ g_P(x') < x_3 < g_P(x') + R'_P\} \subset \Omega$,

(b3) $\nabla g_P(0) = \mathbf{0}$, and $\forall\, x' = (x_1,x_2)^\top \in B^2_{R_P}(0) \quad |\nabla g_P(x')| < r_P\,,$

where $B^k_r(0)$ denotes the k-dimensional open ball with center 0 and radius $r > 0$. We also define, for $0 < \lambda < 1$, the open sets $\lambda\,\Omega_P \subset \Omega_P$ as

$$\lambda\,\Omega_P := \{(x',x_3)\,\big|\,x' = (x_1,x_2)^\top \in B^2_{\lambda R_P}(0),\ g_P(x') < x_3 < g_P(x') + \lambda R'_P\}\,.$$

To localize near $\partial\Omega \cap \partial\Omega_P$, for $P \in \partial\Omega$, we fix smooth functions $\xi_P : \mathbb{R}^3 \to \mathbb{R}$ such that

$$(\ell 1) \qquad \chi_{\frac{1}{2}\Omega_P}(\mathbf{x}) \le \xi_P(\mathbf{x}) \le \chi_{\frac{3}{4}\Omega_P}(\mathbf{x})\,,$$

where $\chi_A(\mathbf{x})$ is the indicator function of the measurable set A. For the remaining interior estimate, we localize by a smooth function $0 \le \xi_0 \le 1$ with $\operatorname{spt}\xi_0 \subset \Omega_0$, where $\Omega_0 \subset \Omega$ is an appropriate open set such that $\operatorname{dist}(\partial\Omega_0, \partial\Omega) > 0$. Since the boundary $\partial\Omega$ is compact, we can use an appropriate finite sub-covering which, together with the interior estimate, yields the global estimate.

Let us introduce the tangential derivatives near the boundary. To simplify the notation, we fix $P \in \partial\Omega$, $h \in (0, \frac{R_P}{16})$ and simply write $\xi := \xi_P$, $g := g_P$. We use the standard notation $\mathbf{x} = (x', x_3)^\top$ and denote by $\mathbf{e}^i$, $i = 1, 2, 3$ the canonical orthonormal basis in $\mathbb{R}^3$. In the following lowercase Greek letters, take values 1, and 2. For a function f with $\operatorname{spt} f \subset \operatorname{spt}\xi$, we define for $\alpha = 1, 2$ tangential translations:

$$f_\tau(x', x_3) = f_{\tau_\alpha}(x', x_3) := f\big(x' + h\,\mathbf{e}^\alpha, x_3 + g(x' + h\,\mathbf{e}^\alpha) - g(x')\big)\,,$$

tangential differences $\Delta^+ f := f_\tau - f$ and tangential difference quotients $d^+ f := h^{-1}\Delta^+ f$. For simplicity, we denote $\nabla g := (\partial_1 g, \partial_2 g, 0)^\top$ and use the operations $(\cdot)_\tau$, $(\cdot)_{-\tau}$, $\Delta^+(\cdot)$, $\Delta^+(\cdot)$, $d^+(\cdot)$ and $d^-(\cdot)$ also for vector-valued and tensor-valued functions, intended as acting component-wise.

We will use the following properties of the difference quotients, all proved in [4]. Let $\mathbf{v} \in W^{1,1}(\Omega)$ be such that $\operatorname{spt}\mathbf{v} \subset \operatorname{spt}\xi$. Then

$$\begin{aligned} \nabla d^\pm \mathbf{v} &= d^\pm \nabla \mathbf{v} + (\partial_3 \mathbf{v})_\tau \otimes d^\pm \nabla g\,, \\ \mathbf{D} d^\pm \mathbf{v} &= d^\pm \mathbf{D}\mathbf{v} + (\partial_3 \mathbf{v})_\tau \overset{s}{\otimes} d^\pm \nabla g\,, \\ \operatorname{div} d^\pm \mathbf{v} &= d^\pm \operatorname{div}\mathbf{v} + (\partial_3 \mathbf{v})_{\pm\tau} d^\pm \nabla g\,, \\ \nabla \mathbf{v}_{\pm\tau} &= (\nabla \mathbf{v})_{\pm\tau} + (\partial_3 \mathbf{v})_{\pm\tau} d^\pm \nabla g\,, \end{aligned} \tag{14}$$

where $(\mathbf{v} \otimes \mathbf{w})_{ij} := v_i w_j$, $i, j = 1, 2, 3$, and $\mathbf{v} \overset{s}{\otimes} \mathbf{w} := \frac{1}{2}\big(\mathbf{v} \otimes \mathbf{w} + (\mathbf{v} \otimes \mathbf{w})^\top\big)$. Moreover, we have also the following properties: If $\operatorname{spt} g \subset \operatorname{spt}\xi$, then there holds

$$(d^- g)_\tau = -d^+ g\,, \quad (d^+ g)_{-\tau} = -d^- g\,, \quad d^- g_\tau = -d^+ g\,,$$

and if $\operatorname{spt} g \cup \operatorname{spt} f \subset \operatorname{spt}\xi$, then we have

$$d^\pm(fg) = f_{\pm\tau}\, d^\pm g + (d^\pm f)\, g\,.$$

As for the classical difference quotients, L^q-uniform bounds (with respect to $h > 0$) for $d^+ f$ imply that $\partial_\tau f$ belongs to $L^q(\operatorname{spt}\xi)$.

Lemma 6 *If* $f \in W^{1,1}(\Omega)$*, then we have for* $\alpha = 1, 2$

$$d^+ f \to \partial_\tau f = \partial_{\tau_\alpha} f := \partial_\alpha f + \partial_\alpha g\, \partial_3 f \qquad \text{as } h \to 0\,, \tag{15}$$

almost everywhere in $\operatorname{spt}\xi$*, (cf. [17]). If we define, for* $0 < h < R_P$

$$\Omega_{P,h} = \Big\{ \mathbf{x} \in \Omega_P \,\Big|\, x' \in B^2_{R_P-h}(0) \Big\}\,,$$

and, if $f \in W^{1,q}_{\text{loc}}(\mathbb{R}^3)$*, for* $1 \le q < \infty$*, then*

$$\int\limits_{\Omega_{P,h}} |d^+ f|^q \, d\mathbf{x} \le c \int\limits_{\Omega_P} |\partial_\tau f|^q \, d\mathbf{x}\,.$$

Moreover, if $d^+ f \in L^q(\Omega_{P,h_0})$*,* $1 < q < \infty$*, and if*

$$\exists\, c_1 > 0 : \quad \int\limits_{\Omega_{P,h_0}} |d^+ f|^q \, d\mathbf{x} \le c_1, \qquad \forall\, h_0 \in (0, R_P) \text{ and } \forall\, h \in (0, h_0)\,,$$

then $\partial_\tau f \in L^q(\Omega_P)$ *and*

$$\int\limits_{\Omega_P} |\partial_\tau f|^q \, d\mathbf{x} \le c_1\,.$$

The following variants of formula of integration by parts will often be used.

Lemma 7 *Let* $\operatorname{spt} g \cup \operatorname{spt} f \subset \operatorname{spt}\xi = \operatorname{spt}\xi_P$ *and* $0 < h < \frac{R_P}{16}$*. Then*

$$\int\limits_{\Omega} f g_{-\tau} \, d\mathbf{x} = \int\limits_{\Omega} f_\tau g \, d\mathbf{x}\,.$$

Consequently, $\int\limits_{\Omega} f d^+ g \, d\mathbf{x} = \int\limits_{\Omega} (d^- f) g \, d\mathbf{x}$*. Moreover, if in addition* f *and* g *are smooth enough and at least one vanishes on* $\partial\Omega$*, then*

$$\int\limits_{\Omega} f \partial_\tau g \, d\mathbf{x} = -\int\limits_{\Omega} (\partial_\tau f) g \, d\mathbf{x}\,.$$

3.2 Regularity Results with Possible Dependencies on A

We start proving spatial regularity for the approximate problem. The estimates, which will be proved for first-order derivatives of $\nabla\mathbf{u}^A$ and $\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)$ in this first step, are uniform with respect to $A \geq 1$:

(i) In the interior of Ω
(ii) For tangential derivatives near the boundary

On the contrary, the estimates depend on A in the normal direction near the boundary $\partial\Omega$. Nevertheless, this allows later on to use the equations point-wise and to prove (in a different way) estimates independent of $A \geq 1$ even near the boundary, allowing then to pass to the limit with $A \to \infty$.

By using the translation method, we obtain the following results, which will be proved below:

Proposition 6 *Let the operator $\mathbf{S} = \partial U$, derived from the potential U, have (p, δ)-structure for some $p \in (1, 2]$, and $\delta \in (0, \infty)$, with characteristics (γ_3, γ_4, p). Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with $C^{2,1}$ boundary, and assume that $\mathbf{f} \in L^{p'}(\Omega)$. Then, the unique weak solution $\mathbf{u}^A \in W_0^{1,2}(\Omega)$ of the approximate problem (9) satisfies*

$$\int_\Omega \xi_0^2|\nabla\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 + \omega^A\big(\xi_0^2|\nabla^2\mathbf{u}^A|\big) + (\delta + A)^{p-2}\xi_0^2|\nabla^2\mathbf{u}^A|^2\,d\mathbf{x} \leq c_0\,,$$

$$\int_\Omega \xi_P^2|\partial_\tau\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 + \omega^A\big(\xi_P^2|\partial_\tau\nabla\mathbf{u}^A|\big) + (\delta + A)^{p-2}\xi_P^2|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} \leq c_P\,, \tag{16}$$

where $c_0 = c_0(\delta, \|\mathbf{f}\|_{p'}, \|\xi_0\|_{1,\infty}, \gamma_3, \gamma_4, p)$, while the constant related to the neighborhood of P is such that $c_P = c_P(\delta, \|\mathbf{f}\|_{p'}, \|\xi_P\|_{1,\infty}, \|g_P\|_{C^{2,1}}, \gamma_3, \gamma_4, p)$. Here, $\xi_0(\mathbf{x})$ is a cutoff function with support in the interior of Ω, and for arbitrary $P \in \partial\Omega$, the tangential derivative is defined locally in Ω_P by (15).

By using Proposition 6 and the ellipticity of $\mathbf{S}^A$, we can write, for a.e. $\mathbf{x} \in \Omega$, the missing partial derivatives in the normal direction (which is locally $\mathbf{e}_3$ after a rotation of coordinates) in terms of the tangential ones. By employing the previous results, we obtain estimates also for the partial derivatives in the $\mathbf{e}_3$-direction, but with a critical dependence on the approximation parameter A.

Proposition 7 *Under the assumptions of Proposition 6, there exists a constant $C_1 > 0$ such that, provided in the local description of the boundary there holds $r_P < C_1$ in (b3), where $\xi_P(\mathbf{x})$ is a cutoff function with support in Ω_P, there holds*

$$\int_\Omega \xi_P^2 |\partial_3 \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 + \omega^A\big(\xi_P^2 |\partial_3 \mathbf{D}\mathbf{u}^A|\big)\, d\mathbf{x} \le C_A\,,$$

where $C_A = C_A(\delta, \|\mathbf{f}\|_{p'}, \|\xi_P\|_{1,\infty}, \|g_P\|_{C^{2,1}}, \gamma_3, \gamma_4, p, A)$.

Before starting the proof of these two propositions, we generalize [7, Lemma 3.11], originally proved for $\phi = \omega$ (with $1 < p \le 2$) to $\phi = \omega^A$. The main properties used are convexity of ω^A, that $(\omega^A)''$ is nonincreasing, and the equivalence properties from Lemma 3.

Lemma 8 *Let* $p \in (1, 2]$ *and* $\delta \ge 0$. *Then, for* ξ *and* g *as above and for any* $\mathbf{v} \in W_0^{1,2}(\Omega)$, *we have*

$$\int_\Omega \omega^A\big(\xi|\nabla d^+\mathbf{v}|\big) + \omega^A\big(\xi|d^+\nabla\mathbf{v}|\big)\, d\mathbf{x} \le c \int_\Omega \xi^2 \big|d^+\mathbf{F}^A(\mathbf{D}\mathbf{v})\big|^2\, d\mathbf{x}$$
$$+\, c(\|\xi\|_{1,\infty}, \|g\|_{C^{1,1}}) \int_{\Omega\cap\operatorname{spt}\xi} \omega^A\big(|\nabla\mathbf{v}|\big)\, d\mathbf{x}\,,$$

with constants not depending on δ *and* A.

Proof The proof is carried out by adapting that of [7, Lem. 3.11]. First, we use the following identity:

$$\xi\, \nabla d^+\mathbf{v} = \nabla(\xi\, d^+\mathbf{v}) - \nabla\xi \otimes d^+\mathbf{v}\,,$$

and consequently we get, by using (7), that

$$\int_\Omega \omega^A(\xi|\nabla d^+\mathbf{v}|)\, d\mathbf{x} \le c \int_\Omega \omega^A(|\mathbf{D}(\xi\, d^+\mathbf{v})|)\, d\mathbf{x} + c(\|g\|_{C^{0,1}}, \|\xi\|_{1,\infty}) \int_{\Omega\cap\operatorname{spt}\xi} \omega^A(|\nabla\mathbf{v}|)\, d\mathbf{x}\,,$$

where we also used Korn's inequality for N-functions (cf. [14, Thm. 6.10]), with a constant independent of $A \ge 1$, and the following inequality (cf. [9, Sec. 3.2]):

$$\int_{\Omega\cap\operatorname{spt}\xi} \omega^A(|d^\pm\mathbf{v}|)\, d\mathbf{x} \le c \int_{\Omega\cap\operatorname{spt}\xi} \omega^A(|\nabla\mathbf{v}|)\, d\mathbf{x}\,. \tag{17}$$

Using the identities

$$\mathbf{D}(\xi\, d^+\mathbf{v}) = \xi\, \mathbf{D}(d^+\mathbf{v}) + \nabla\xi \overset{s}{\otimes} d^+\mathbf{v} = \xi\, d^+\mathbf{D}\mathbf{v} + (\partial_3\mathbf{v})_\tau \overset{s}{\otimes} d^+\nabla g + \nabla\xi \overset{s}{\otimes} d^+\mathbf{v}\,,$$

the properties of ω^A, ξ, g, and (17), we obtain

$$\begin{aligned}\int\limits_{\Omega}\omega^A(\xi|\nabla d^+\mathbf{v}|)\,d\mathbf{x} &\le c\int\limits_{\Omega}\omega^A(\xi|d^+\mathbf{D}\mathbf{v}|)\,d\mathbf{x}\\ &\qquad + c(\|\xi\|_{1,\infty},\|g\|_{C^{1,1}})\int\limits_{\Omega\cap\operatorname{spt}\xi}\omega^A(|\nabla\mathbf{v}|)\,d\mathbf{x}\,.\end{aligned} \tag{18}$$

We focus on the first term on the right-hand side of (18). Using a change of shift as in Lemma 1 yields

$$\omega^A(\xi|d^+\mathbf{D}\mathbf{v}|)\le c\big(\omega^A_{|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|}(\xi|d^+\mathbf{D}\mathbf{v}|)+\omega^A(|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|)\big)\,. \tag{19}$$

By the fact that ω^A is balanced with characteristics depending only on p (cf. Lemmas 2 and 3), we get

$$\begin{aligned}\omega^A_{|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|}(\xi|d^+\mathbf{D}\mathbf{v}|) &\sim (\omega^A_{|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|})'(\xi|d^+\mathbf{D}\mathbf{v}|)\,\xi|d^+\mathbf{D}\mathbf{v}|\\ &= \frac{(\omega^A)'(|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|+\xi|d^+\mathbf{D}\mathbf{v}|)}{|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|+\xi|d^+\mathbf{D}\mathbf{v}|}\,\xi^2|d^+\mathbf{D}\mathbf{v}|^2\\ &= a^A(|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|+\xi|d^+\mathbf{D}\mathbf{v}|)\,\xi^2|d^+\mathbf{D}\mathbf{v}|^2.\end{aligned}$$

Next, since a^A is nonincreasing for $p\in(1,2]$ (see Lemma 5), we get

$$\begin{aligned}\omega^A_{|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|}(\xi|d^+\mathbf{D}\mathbf{v}|) &\le a^A(|\mathbf{D}\mathbf{v}|+|\Delta^+\mathbf{D}\mathbf{v}|)\,\xi^2|d^+\mathbf{D}\mathbf{v}|^2\\ &\sim \xi^2|d^+\mathbf{F}^A(\mathbf{D}\mathbf{v})|^2\,.\end{aligned}$$

Inserting this into (19), we obtain from (18) that

$$\begin{aligned}&\int\limits_{\Omega}\omega^A(\xi|\nabla d^+\mathbf{v}|)\,d\mathbf{x}\le\\ &\le c\int\limits_{\Omega}\xi^2\big|d^+\mathbf{F}^A(\mathbf{D}\mathbf{v})\big|^2\,d\mathbf{x}+c(\|\xi\|_{1,\infty},\|g\|_{C^{1,1}})\int\limits_{\Omega\cap\operatorname{spt}\xi}\omega^A(|\nabla\mathbf{v}|)\,d\mathbf{x}\,.\end{aligned} \tag{20}$$

Next, we observe that (14) and (7) yield

$$\begin{aligned}\int\limits_{\Omega}\omega^A(\xi|d^+\nabla\mathbf{v}|)\,d\mathbf{x} &\le \int\limits_{\Omega}\omega^A(\xi|\nabla d^+\mathbf{v}|)\,d\mathbf{x}+\int\limits_{\Omega}\omega^A(\xi|\partial_3\mathbf{v}||d^+\nabla g|)\,d\mathbf{x}\\ &\le \int\limits_{\Omega}\omega^A(\xi|\nabla d^+\mathbf{v}|)\,d\mathbf{x}+c(\|\xi\|_{\infty},\|g\|_{C^{1,1}})\int\limits_{\Omega\cap\operatorname{spt}\xi}\omega^A(|\nabla\mathbf{v}|)\,d\mathbf{x}\,.\end{aligned}$$

This shows that also the term $\int_\Omega \omega^A(\xi\,|d^+\nabla\mathbf{v}|)\,d\mathbf{x}$ can be estimated by the right-hand side of (20), ending the proof. □

We can now proceed with the proof of regularity in the tangential directions and in the interior.

Proof of Proposition 6 We obtain estimates for tangential derivatives by considering limits of increments in the tangential directions cf. [7, 9]. Fix $P \in \partial\Omega$ and use in Ω_P

$$\mathbf{w} = d^-(\xi^2 d^+(\mathbf{u}^A|_{\frac{1}{2}\Omega_P}))\,,$$

where $\xi := \xi_P$, $g := g_P$, and $h \in (0, \frac{R_P}{16})$, as a test function in the weak formulation (10) of Problem (9). This yields

$$\begin{aligned}
&\int\limits_\Omega \xi^2 d^+\mathbf{S}^A(\mathbf{Du}^A)\cdot d^+\mathbf{Du}^A\,d\mathbf{x} = \\
&= -\int\limits_\Omega \mathbf{S}^A(\mathbf{Du}^A)\cdot\big(\xi^2 d^+\partial_3\mathbf{u}^A - (\xi_{-\tau}d^-\xi + \xi d^-\xi)\partial_3\mathbf{u}^A\big)\overset{s}{\otimes} d^-\nabla g\,d\mathbf{x} \\
&\quad -\int\limits_\Omega \mathbf{S}^A(\mathbf{Du}^A)\cdot\xi^2(\partial_3\mathbf{u}^A)_\tau \overset{s}{\otimes} d^-d^+\nabla g - \mathbf{S}^A(\mathbf{Du}^A)\cdot d^-\big(2\xi\nabla\xi \overset{s}{\otimes} d^+\mathbf{u}^A\big)\,d\mathbf{x} \\
&\quad +\int\limits_\Omega \mathbf{S}^A((\mathbf{Du}^A)_\tau)\cdot\big(2\xi\,\partial_3\xi\, d^+\mathbf{u}^A + \xi^2 d^+\partial_3\mathbf{u}^A\big)\overset{s}{\otimes} d^+\nabla g\,d\mathbf{x} \\
&\quad +\int\limits_\Omega \mathbf{f}\cdot d^-(\xi^2 d^+\mathbf{u}^A)\,d\mathbf{x} =: \sum_{j=1}^{8} I_j\,.
\end{aligned} \tag{21}$$

The properties of $\mathbf{S}^A$, Proposition 3, and Lemma 8 imply the following estimate:

$$\begin{aligned}
&\int\limits_\Omega \xi^2\big|d^+\mathbf{F}^A(\mathbf{Du}^A)\big|^2 + \omega^A\big(\xi\,|d^+\nabla\mathbf{u}^A|\big)\,d\mathbf{x} \le \\
&\le c\int\limits_\Omega \xi^2 d^+\mathbf{S}^A(\mathbf{Du}^A)\cdot d^+\mathbf{Du}^A\,d\mathbf{x} + c(\|\xi\|_{1,\infty}, \|g\|_{C^{1,1}})\int\limits_{\Omega\cap\operatorname{spt}\xi}\omega^A\big(|\nabla\mathbf{u}^A|\big)\,d\mathbf{x}\,.
\end{aligned}$$

The terms I_1–I_7 in (21) are estimated exactly as in [7, (3.17)–(3.22)], while I_8 is estimated as the term I_{15} in [7, (4.20)]. Thus, we get, by using also Corollary 1,

$$\begin{aligned}
&\int\limits_\Omega (\delta + A)^{p-2}\xi^2\big|d^+\nabla\mathbf{u}^A\big|^2 + \xi^2\big|d^+\mathbf{F}^A(\mathbf{Du}^A)\big|^2 + \omega^A(\xi\,|d^+\nabla\mathbf{u}^A|)\,d\mathbf{x} \le \\
&\le c(\|\mathbf{f}\|_{p'}, \|\xi\|_{2,\infty}, \|g\|_{C^{2,1}}, \delta)\,.
\end{aligned}$$

This proves the second estimate in (16) by Lemma 6, since the constant on the right-hand side does not depend on $h > 0$.

The first estimate in (16) is proved in the same way with many simplifications, since in the interior one can consider directly standard translations in all the coordinate directions. □

For the Proof of Proposition 7, the following observation will be crucial.

Remark 6 The obtained estimate $(16)_1$, Proposition 4, and Lemma 5 imply that $\mathbf{u}^A \in W^{2,2}_{\mathrm{loc}}(\Omega)$ (with estimates depending on A) and that the system (9) is well-defined point-wise a.e. in Ω.

Proof of Proposition 7 To estimate the derivatives in the $\mathbf{e}_3$-direction, we use equation (9) point-wise a.e. in Ω, which is justified by Remark 6. Denoting, for $\alpha, \gamma = 1, 2$, $A_{\alpha\gamma} := \partial_{\gamma 3} S^A_{\alpha 3}(\mathbf{Du}^A)$, $\mathfrak{b}_\gamma := \partial_3 D_{\gamma 3}\mathbf{u}^A$, and[2] $\mathfrak{f}_\alpha := f_\alpha + \partial_{33} S^A_{\alpha 3}(\mathbf{Du}^A)\partial_3 D_{33}\mathbf{u}^A + \partial_{\gamma\sigma} S^A_{\alpha 3}(\mathbf{Du}^A)\partial_3 D_{\gamma\sigma}\mathbf{u}^A + \sum_{k,l=1}^3 \partial_{kl} S^A_{\alpha\beta}(\mathbf{Du}^A)\partial_\beta D_{kl}\mathbf{u}^A$, we can rewrite the first two equations in (9) as follows:

$$-2A_{\alpha\gamma}\mathfrak{b}_\gamma = \mathfrak{f}_\alpha \qquad \text{a.e. in } \Omega\,.$$

We employ this equality separately on each Ω_P in order to use the notion of tangential derivative. By straightforward manipulations (cf. [7, Sections 3.2 and 4.2]) we get a.e. in Ω_P

$$a^A(|\mathbf{Du}^A|)\,|\mathfrak{b}| \leq c\,\Big(|\mathbf{f}| + |\mathbf{f}|\|\nabla g\|_\infty + a^A(|\mathbf{Du}^A|)\,\big(|\partial_\tau \nabla \mathbf{u}^A| + \|\nabla g\|_\infty |\nabla^2 \mathbf{u}^A|\big)\Big)\,.$$

Note that we can deduce from this inequality information about $\tilde{\mathfrak{b}}_\gamma := \partial^2_{33}\mathbf{u}^A_\gamma$, because $|\mathfrak{b}| \geq 2|\tilde{\mathfrak{b}}| - |\partial_\tau \nabla \mathbf{u}^A| - \|\nabla g\|_\infty |\nabla^2 \mathbf{u}^A|$. Adding on both sides, for $\alpha = 1, 2$ and $i, k = 1, 2, 3$, the term

$$a^A(|\mathbf{Du}^A|)\,\big(|\partial_\alpha \partial_i u^A_k| + |\partial^2_{33} u^A_3|\big)\,,$$

we finally arrive, a.e. in Ω_P at the inequality

$$\begin{aligned} &a^A(|\mathbf{Du}^A|)|\nabla^2 \mathbf{u}^A| \leq \\ &\leq c\,\Big(|\mathbf{f}| + |\mathbf{f}|\|\nabla g\|_\infty + a^A(|\mathbf{Du}^A|)\,\big(|\partial_\tau \nabla \mathbf{u}^A| + \|\nabla g\|_\infty |\nabla^2 \mathbf{u}^A|\big)\Big)\,, \end{aligned}$$

where, due to the results proved in Sect. 2, the constant c only depends on the characteristics of $\mathbf{S}$. Next, we can choose the open sets Ω_P in such a way that $\|\nabla g_P(x')\|_{\infty,\Omega_P}$ is small enough, so that we can absorb the last term from the right-hand side, which yields

[2] Recall that we use the summation convention over repeated Greek lowercase letters from 1 to 2.

$$a^A(|\mathbf{Du}^A|)\,|\nabla^2\mathbf{u}^A| \le c\,\Big(|\mathbf{f}| + a^A(|\mathbf{Du}^A|)\,|\partial_\tau\nabla\mathbf{u}^A|\Big) \qquad \text{a.e. in } \Omega_P\,,$$

where again the constant c only depends on the characteristics of $\mathbf{S}$. Dividing both sides by the quantity $\sqrt{a^A(|\mathbf{Du}^A|)} \neq 0$ (which is nonzero by the fact that $\delta > 0$ and the properties of ω^A) and raising the result to the power 2, we get a.e. in Ω_P

$$a^A(|\mathbf{Du}^A|)|\nabla^2\mathbf{u}^A|^2 \le c\,\frac{|\mathbf{f}|^2}{a^A(|\mathbf{Du}^A|)} + c\,a^A(|\mathbf{Du}^A|)|\partial_\tau\nabla\mathbf{u}^A|^2\,. \tag{22}$$

Note that both sides are finite a.e. and, for the moment, we know that the left-hand side belongs at least to $L^1_{loc}(\Omega_P)$.

Concerning the first term on the right-hand side, we note that Lemma 5 and the definition of a^A imply

$$\frac{1}{a^A(t)} \le \frac{1}{(p-1)}\frac{1}{a(t)} = \frac{1}{(p-1)}\frac{1}{(\delta+t)^{p-2}}\,.$$

Using this estimate and Hölder inequality, we get, with a constant c independent on A,

$$\begin{aligned}\int\limits_\Omega \frac{|\mathbf{f}|^2}{a^A(|\mathbf{Du}^A|)}\,d\mathbf{x} &\le c(p)\|\mathbf{f}\|^2_{p'}\|\delta + |\mathbf{Du}^A|\|^{2-p}_p \\ &\le c\big(\|\mathbf{f}\|^{p'}_{p'} + \delta^p + \|\mathbf{F}(\mathbf{Du}^A)\|_2^2\big)\,.\end{aligned} \tag{23}$$

For the second term on the right-hand side of (22), we use that, in view of Lemma 5, there holds:

$$\begin{aligned}\int\limits_\Omega \xi_P^2 a^A(|\mathbf{Du}^A|)|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} &\le \delta^{p-2}\int\limits_\Omega \xi_P^2|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} \\ &= \frac{\delta^{p-2}}{(\delta+A)^{p-2}}\,(\delta+A)^{p-2}\int\limits_\Omega \xi_P^2|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} \\ &\le \left(1+\frac{A}{\delta}\right)^{2-p} c_P\,,\end{aligned}$$

where the final estimate follows from the already proved results on tangential derivatives in Proposition 6.

Hence, multiplying (22) by ξ_P^2 and integrating over the proper sub-domain

$$\Omega_{P,\varepsilon} := \big\{\mathbf{x} \in \Omega_P \,\big|\, g_P + \varepsilon < x_3 < g_P + R'_P,\ \text{for } 0 < \varepsilon < R'_P\big\}\,,$$

we get, also using (11) and (12),

$$
\begin{aligned}
&\int\limits_{\Omega_{P,\varepsilon}} \xi_P^2 a^A(|\mathbf{Du}^A|)|\nabla^2\mathbf{u}^A|^2\,d\mathbf{x} \leq \\
&\leq c\int\limits_{\Omega_{P,\varepsilon}} \frac{|\mathbf{f}|^2}{a^A(|\mathbf{Du}^A|)}\,d\mathbf{x} + c\int\limits_{\Omega_{P,\varepsilon}} \xi_P^2 a^A(|\mathbf{Du}^A|)|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} \\
&\leq c\int\limits_{\Omega} \frac{|\mathbf{f}|^2}{a^A(|\mathbf{Du}^A|)}\,d\mathbf{x} + c\int\limits_{\Omega} \xi_P^2 a^A(|\mathbf{Du}^A|)|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x} \\
&\leq C\left(\delta^p + \|\mathbf{f}\|_{p'}^{p'} + (1+A\delta^{-1})^{2-p}\right).
\end{aligned}
$$

Since this estimate is independent of $\varepsilon > 0$, the above inequality shows, by monotone convergence, that also $\int_\Omega \xi_P^2 a^A(|\mathbf{Du}^A|)|\nabla^2\mathbf{u}^A|^2\,d\mathbf{x} \leq C(\|\mathbf{f}\|_{p'}, \delta, \delta^{-1}, A)$, ending the proof. □

Remark 7 The reader should notice that the dependence on A is mainly due to the fact that we have a stress tensor depending on the symmetric gradient. To use Korn inequality in Lemma 8, we have to pay the price of estimates depending on A. In the case of a stress tensor depending on the full gradient, this step can be skipped (see the results in [6] where the considered problem has an additional term with 2-structure and the A-approximation is not needed).

By collecting the results of the Propositions 6 and 7, we get the following result:

Proposition 8 *Let the operator* $\mathbf{S} = \partial U$*, derived from the potential* U*, have* (p,δ)*-structure for some* $p \in (1,2]$ *and* $\delta \in (0,\infty)$*. Let* $\Omega \subset \mathbb{R}^3$ *be a bounded domain with* $C^{2,1}$ *boundary, and let* $\mathbf{f} \in L^{p'}(\Omega)$*. Then, the unique weak solution* $\mathbf{u}^A \in W_0^{1,2}(\Omega)$ *of problem* (9) *satisfies*

$$
\int\limits_{\Omega} |\nabla \mathbf{F}^A(\mathbf{Du}^A)|^2\,d\mathbf{x} \leq c(A,\delta^{-1}),
$$

where c *depends also on the characteristics of* $\mathbf{S}$*,* δ*,* $\|\mathbf{f}\|_{p'}$*,* $|\Omega|$*, and the* $C^{2,1}$*-norms of the local description of* $\partial\Omega$*. In particular, the above estimate implies that* $\mathbf{u}^A \in W^{2,2}(\Omega)$*.*

Proof The proof is simply obtained by observing that $\overline{\Omega}$ is a compact set. After having fixed all Ω_P small enough (depending on P) to perform the calculations leading to Proposition 7, we can extract a finite covering of sets Ω_P and consequently prove the uniform nature of estimates in terms of $P \in \partial\Omega$. To show that $\mathbf{u}^A \in W^{2,2}(\Omega)$, we use Proposition 4 and Lemma 5. □

4 Estimates Uniform with Respect to A for the Solutions of the Approximate Problem

We now sketch the proof of the estimate of $\nabla \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)$, which is independent of A. Moreover, we also improve the δ-dependence of the estimates. We adapt to the A-approximation the same procedure already used in [8], obtaining the steady counterpart to the case $1 < p \le 2$ of the results proved in [9, Sec. 3].

Proposition 9 *Let the same hypotheses as in Theorem 1 be satisfied with $\delta > 0$, and let the local description g_P of the boundary and the localization function ξ_P satisfy $(b1)$–$(b3)$ and $(\ell 1)$ (cf. Sect. 3.1). Then, there exists a constant $C_2 > 0$ such that the regular solution $\mathbf{u}^A \in W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$ of the approximate problem (9) satisfies for every $P \in \partial\Omega$*

$$\int\limits_{\Omega} \xi_P^2 |\partial_3 \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 \, d\mathbf{x} \le C\,,$$

provided $r_P < C_2$ in $(b3)$, with C depending on the characteristics of $\mathbf{S}$, δ, $\|\mathbf{f}\|_{p'}$, $\|\xi_P\|_{1,\infty}$, $\|g_P\|_{C^{2,1}}$, and C_2.

Proof We will not give the full proof of this result, since it is very similar to that of [8, Prop. 3.2]. For the reader's convenience, we just explain the main steps.

Fix an arbitrary point $P \in \partial\Omega$ and a local description $g = g_P$ of the boundary and the localization function $\xi = \xi_P$ as before. Proposition 4 yields that there exists a constant C_0, depending only on the characteristics of $\mathbf{S}$ such that

$$\frac{1}{C_0} |\partial_3 \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 \le \mathbb{P}_3^A(\mathbf{D}\mathbf{u}^A) \qquad \text{a.e. in } \Omega\,.$$

We now work directly with $\mathbb{P}_3^A(\mathbf{D}\mathbf{u}^A)$ to deduce estimates for $|\partial_3 \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2$. Note that, since $\mathbf{u}^A$ is a regular solution of (9), all calculations are justified. Thus, using the definition of $\mathbb{P}_3^A(\mathbf{D}\mathbf{u}^A)$ and the symmetries of $\mathbf{S}^A$ and $\mathbf{D}\mathbf{u}^A$, we obtain

$$\begin{aligned} &\frac{1}{C_0} \int\limits_{\Omega} \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2 \, d\mathbf{x} \le \qquad (24) \\ &\le \int\limits_{\Omega} \xi^2 \partial_3 \mathbf{S}^A_{\alpha\beta}(\mathbf{D}\mathbf{u}^A)\, \partial_3 D_{\alpha\beta} \mathbf{u}^A \, d\mathbf{x} + \int\limits_{\Omega} \xi^2 \partial_3 \mathbf{S}^A_{3\alpha}(\mathbf{D}\mathbf{u}^A)\, \partial_\alpha D_{33} \mathbf{u}^A \, d\mathbf{x} \\ &\quad + \int\limits_{\Omega} \sum_{j=1}^{3} \xi^2 \partial_3 \mathbf{S}^A_{j3}(\mathbf{D}\mathbf{u}^A)\, \partial_3^2 u_j^N \, d\mathbf{x} \\ &=: \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3\,. \end{aligned}$$

The most critical term is $\mathcal{J}_1$ which is estimated, for any $\lambda > 0$, as follows

$$|\mathcal{J}_1| \le \lambda \int\limits_\Omega \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c_{\lambda^{-1}} \left(1 + \|\nabla g\|_\infty^2\right) \sum_{\beta=1}^{2} \int\limits_\Omega \xi^2 |\partial_\beta \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x}$$
$$+ \int\limits_\Omega \xi^2 |\partial_3 \mathbf{S}^A(\mathbf{Du}^A)| \, |\nabla^2 g| \, |\mathbf{Du}^A| \, d\mathbf{x} + \left| \int\limits_\Omega \xi^2 \partial_3 \mathbf{S}^A_{\alpha\beta}(\mathbf{Du}^A) \, \partial_\alpha \partial_{\tau_\beta} u_3^A \, d\mathbf{x} \right| .$$

In the last but one term we multiply and divide by $\sqrt{a^A(|\mathbf{Du}^A|)}$, use Proposition 4, Young inequality, and $a^A(|\mathbf{Du}^A|)|\mathbf{Du}^A|^2 \sim |\mathbf{F}^A(|\mathbf{Du}^A|)|^2$ (cf. Proposition 3), yielding that it is estimated by

$$\lambda \int\limits_\Omega \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c_{\lambda^{-1}} \, \|\nabla^2 g\|_\infty^2 \int\limits_\Omega |\mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} .$$

To handle the last term in the above estimate of $\mathcal{J}_1$ we perform a crucial partial integration. This avoids to have terms with the quantity $\partial_3 \mathbf{S}^A(\mathbf{Du}^A)$ which cannot be estimated in terms of tangential derivatives. Let us explain the main idea beyond this step. Observe that, by neglecting the localization ξ, integration by parts gives

$$\int\limits_\Omega \partial_3 \mathbf{S}^A_{\alpha\beta}(\mathbf{Du}^A) \, \partial_\alpha \partial_{\tau_\beta} u_3^A \, d\mathbf{x} = \int\limits_\Omega \partial_\alpha \mathbf{S}^A_{\alpha\beta}(\mathbf{Du}^A) \, \partial_3 \partial_{\tau_\beta} u_3^A \, d\mathbf{x}$$
$$= \int\limits_\Omega \partial_\alpha \mathbf{S}^A_{\alpha\beta}(\mathbf{Du}^A) \, \partial_{\tau_\beta} D_{33} \mathbf{u}^A \, d\mathbf{x} .$$

We next multiply and divide the integrand on the right-hand side by $\sqrt{a^A(|\mathbf{Du}^A|)}$, using Proposition 4, Young inequality, and the definition of the tangential derivatives, yielding that

$$\left| \int\limits_\Omega \partial_\alpha \mathbf{S}^A_{\alpha\beta}(\mathbf{Du}^A) \, \partial_{\tau_\beta} D_{33} \mathbf{u}^A \, d\mathbf{x} \right| \le$$
$$\le c \sum_{\alpha=1}^{2} \int\limits_\Omega |\partial_\alpha \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c \sum_{\beta=1}^{2} \int\limits_\Omega |\partial_{\tau_\beta} \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x}$$
$$\le c \sum_{\alpha=1}^{2} \int\limits_\Omega |\partial_{\tau_\alpha} \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c \, \|\nabla g\|_\infty^2 \int\limits_\Omega |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} .$$

The presence of the localization leads to several additional lower-order terms, which all can be easily handled as in [8]. To treat $\mathcal{J}_2$, we multiply and divide by

$\sqrt{a^N(|\mathbf{Du}^A|)}$, using Proposition 4 and Young inequality, to show that, for any given $\lambda > 0$, it holds

$$|\mathcal{J}_2| \leq \lambda \int_0^t \int_\Omega \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c_{\lambda^{-1}} \sum_{\beta=1}^{2} \int_0^t \int_\Omega \xi^2 |\partial_\beta \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \,,$$

for some constant $c_{\lambda^{-1}}$ depending on λ^{-1}. To handle the term $\mathcal{J}_3$, we use Eq. (9). All terms are handled exactly as in [8, Prop. 3.2], and thus, we skip the details here. All together we arrive at the following: estimate

$$\begin{aligned} |\mathcal{J}_1| + |\mathcal{J}_2| + |\mathcal{J}_3| \leq \; & \big(\lambda + c_{\lambda^{-1}} \, \|\nabla g\|_\infty^2\big) \int_\Omega \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \\ & + c_{\lambda^{-1}} \sum_{\beta=1}^{2} \int_\Omega \xi^2 |\partial_{\tau_\beta} \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \\ & + c_{\lambda^{-1}} \big(1 + \|\nabla \xi\|_\infty^2\big) \int_\Omega |\mathbf{F}^A(|\mathbf{Du}^A|)|^2 \, d\mathbf{x} + c_{\lambda^{-1}} \int_\Omega \frac{|\mathbf{f}|^2}{a^A(|\mathbf{Du}^A|)} \, d\mathbf{x} \,. \end{aligned}$$

Now, we first choose $\lambda > 0$ smaller than $(4C_0)^{-1}$, and then we choose the covering of the boundary $\partial\Omega$ such that $c_{\lambda^{-1}} \, \|\nabla g\|_\infty^2 \leq (4C_0)^{-1}$, in order to absorb in the left-hand side of (24) the term involving $\partial_3 \mathbf{F}^A(\mathbf{Du}^A)$. By using the estimate (23) already proved for the term with the external force, we get

$$\begin{aligned} \int_\Omega \xi^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \leq c \sum_{\beta=1}^{2} & \int_\Omega \xi^2 |\partial_{\tau_\beta} \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \\ & + c \int_\Omega |\mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} + c\big(\delta^p + \|\mathbf{f}\|_{p'}^{p'}\big) \,, \end{aligned}$$

with constants depending only on the characteristics of $\mathbf{S}$, $\|g\|_{C^{2,1}}$, and $\|\xi\|_{1,\infty}$. The uniform estimates (11) and (16) for the right-hand side allow us to end the Proof of Proposition 9. □

Proposition 10 *Let the operator* $\mathbf{S} = \partial U$*, derived from the potential* U*, have* (p, δ)*-structure for some* $p \in (1, 2]$ *and* $\delta \in (0, \infty)$*. Let* $\Omega \subset \mathbb{R}^3$ *be a bounded domain with* $C^{2,1}$ *boundary, and let* $\mathbf{f} \in L^{p'}(\Omega)$*. Then, the unique weak solution* $\mathbf{u}^A \in W_0^{1,2}(\Omega)$ *of the problem* (9) *satisfies*

$$\int_\Omega |\nabla \mathbf{F}(\mathbf{Du}^A)|^2 + |\nabla \mathbf{F}^A(\mathbf{Du}^A)|^2 \, d\mathbf{x} \leq C \,,$$

where C depends on the characteristics of $\mathbf{S}$, δ, $\|\mathbf{f}\|_{p'}$, $|\Omega|$, *and the* $C^{2,1}$*-norms of the local description of* $\partial\Omega$. *In particular, the above estimate implies that* $\mathbf{u}^A$ *is uniformly bounded with respect to* $A \geq 1$ *in* $W^{2,\frac{3p}{p+1}}(\Omega)$.

Proof The assertion follows in the same way as in the Proof of Proposition 8. The estimate for $\nabla\mathbf{F}(\mathbf{D}\mathbf{u}^A)$ in $L^2(\Omega)$ follows from Corollary 2. It implies in turn that $\mathbf{u}^A$ is bounded uniformly with respect to $A \geq 1$ in $W^{2,\frac{3p}{p+1}}(\Omega)$ by using [11, Lem. 4.5]. □

5 Passing to the Limit

The final step concerns passing to the limit $A \to \infty$. The unique solution of (1) will be obtained as

$$\mathbf{u} := \lim_{A\to\infty} \mathbf{u}^A \,,$$

with the limit taken in appropriate function spaces.

Remark 8 It will be needed to extract several subsequences, but we still write simply $A \to \infty$ to avoid using too heavy notation.

By uniform—with respect to A—estimates in $W^{1,2}(\Omega)$ in Proposition 10, it directly follows that $\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)$ has a weak limit which we denote as $\widehat{\mathbf{F}} \in W^{1,2}(\Omega)$. Moreover, Proposition 10 also yields that $\|\nabla^2\mathbf{u}^A\|_{3p/(p+1)} \leq C$, with a constant independent of A. Hence, the compact Sobolev embedding $W^{2,\frac{3p}{p+1}}(\Omega) \hookrightarrow\hookrightarrow W^{1,1}(\Omega)$ implies the strong convergence of gradients in $L^1(\Omega)$. This also implies that $\mathbf{D}\mathbf{u}^A(\mathbf{x}) \to \mathbf{D}\mathbf{u}(\mathbf{x})$ for almost every $\mathbf{x} \in \Omega$.

Combining these two facts with $\lim_{A\to\infty}\mathbf{F}^A(\mathbf{P}) = \mathbf{F}(\mathbf{P})$, which is valid uniformly with respect to any compact set in $\mathbb{R}^{3\times 3}$, and the lower semicontinuity of the norm, it follows that

$$\lim_{A\to\infty} \mathbf{F}^A(\mathbf{D}\mathbf{u}^A) = \mathbf{F}(\mathbf{D}\mathbf{u}) \qquad \text{weakly in } W^{1,2}(\Omega) \text{ and a.e. in } \Omega\,,$$

and also

$$\int\limits_{\Omega} |\nabla\mathbf{F}(\mathbf{D}\mathbf{u})|^2\, d\mathbf{x} \leq C\,.$$

Observe that $\widehat{\mathbf{F}} = \mathbf{F}(\mathbf{D}\mathbf{u})$ since weak limit in Lebesgue spaces and the a.e. limit coincide.

It remains to be proved that $\mathbf{u}$ is the unique solution of (1). From the construction of $\mathbf{S}^A$, it follows $\mathbf{S}^A(\mathbf{P}) \to \mathbf{S}(\mathbf{P})$, uniformly with respect to any compact set in $\mathbb{R}^{3\times 3}$.

This fact, coupled with the almost everywhere convergence of $\mathbf{D}\mathbf{u}^A$, implies that

$$\lim_{A\to\infty} \mathbf{S}^A(\mathbf{D}\mathbf{u}^A(\mathbf{x})) = \mathbf{S}(\mathbf{D}\mathbf{u}(\mathbf{x})) \qquad a.e.\, \mathbf{x} \in \Omega\,,$$

which is nevertheless *not* enough to infer directly that

$$\lim_{A\to\infty} \int_\Omega \mathbf{S}^A(\mathbf{D}\mathbf{u}^A)\cdot \mathbf{D}\mathbf{w}\, d\mathbf{x} = \int_\Omega \mathbf{S}(\mathbf{D}\mathbf{u})\cdot \mathbf{D}\mathbf{w}\, d\mathbf{x}, \qquad \forall\, \mathbf{w} \in C_0^\infty(\Omega)\,,$$

and to pass to the limit in the weak formulation. To this end, we need, for instance, additionally a uniform bound on $\mathbf{S}^A(\mathbf{D}\mathbf{u}^A)$ in $L^q(\Omega)$, for some $q > 1$. This would imply that $\mathbf{S}^A(\mathbf{D}\mathbf{u}^A) \rightharpoonup \widehat{\mathbf{S}}$ in $L^q(\Omega)$ and that the limit will satisfy $\widehat{\mathbf{S}} = \mathbf{S}(\mathbf{D}\mathbf{u})$, again by the identification of weak and almost everywhere limits in the Lebesgue spaces.

Observe that from the definition of $\mathbf{S}^A$, we have (cf. Proposition 3 and Lemma 5) for $p \in (1, 2]$ that

$$|\mathbf{S}^A(\mathbf{D}\mathbf{u}^A)| \le c\,\delta^{p-2}|\mathbf{D}\mathbf{u}^A|\,.$$

On the other hand, the estimate $\mathbf{F}(\mathbf{D}\mathbf{u}^A) \in W^{1,2}(\Omega)$, which is uniform with respect to A, implies by the Sobolev embedding $W^{1,2}(\Omega) \hookrightarrow L^6(\Omega)$ that $\|\mathbf{F}(\mathbf{D}\mathbf{u}^A)\|_6 \le C$. Using the properties of $\mathbf{F}$, it follows for $p \in (1, 2]$ that (cf. Proposition 3 and Lemma 5)

$$\|\mathbf{D}\mathbf{u}^A\|_{3p} \le C\,.$$

Hence, we get that $\mathbf{S}^A(\mathbf{D}\mathbf{u}^A)$ is bounded uniformly in $L^{3p}(\Omega)$ for $1 < p \le 2$. This finally allows us to pass to the limit in the weak formulation, showing that $\mathbf{u}$ solves (1). Thus, we proved Theorem 1.

6 On the Time-Dependent Problem

In this section, we state the natural regularity results in the time-dependent case. These results can be proved by adapting the method used in the steady situation. Thus, we just give the statements of the needed results and explain necessary changes.

Remark 9 Results from this section are partially contained in [10], where they are proved with a different approximation method and under more restrictive assumptions on the data, which however yield also regularity in time, i.e., it is proved there that in addition $\frac{\partial}{\partial t}\mathbf{F}(\mathbf{D}\mathbf{u}) \in L^2(\Omega)$ holds. Here, we are keeping the minimal assumptions to prove the natural regularity with respect to the spatial variables. However, additional assumptions on the data would allow to fully recover all the regularity results from [10]. The result presented in this section is the ($p \le 2$)-counterpart of [9, Thm. 3.4].

We now show how the initial boundary value problem

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \operatorname{div} \mathbf{S}(\mathbf{D}\mathbf{u}) = \mathbf{f} & \text{in } I \times \Omega\,, \\ \mathbf{u} = \mathbf{0} & \text{on } I \times \partial\Omega\,, \\ \mathbf{u}(0) = \mathbf{u}_0 & \text{in } \Omega\,, \end{cases} \tag{25}$$

where $I = (0, T)$, for some $T > 0$, can be handled by adapting the tools used in the steady case. First, we introduce the notion of a regular solution.

Definition 6 (Regular Solution) Let the operator $\mathbf{S} = \partial U$ in (25), derived from the potential U, have (p, δ)-structure for some $p \in (1, \infty)$ and $\delta \in [0, \infty)$. Let $\Omega \subset \mathbb{R}^3$ be a bounded domain with $C^{2,1}$ boundary, and let $I = (0, T)$, $T \in (0, \infty)$, be a finite time interval. Then, we say that $\mathbf{u}$ is a regular solution of (25) if $\mathbf{u} \in L^p(I; W_0^{1,p}(\Omega))$ satisfies for all $\psi \in C_0^\infty(0, T)$ and all $\mathbf{w} \in W_0^{1,p}(\Omega)$

$$\int_0^T \Big(\frac{\partial \mathbf{u}(t)}{\partial t}, \mathbf{w}\Big)\, \psi(t) + (\mathbf{S}(\mathbf{D}\mathbf{u}(t)), \mathbf{D}\mathbf{w})\, \psi(t)\, dt = \int_0^T (\mathbf{f}(t), \mathbf{w})\, \psi(t)\, dt\,,$$

and fulfils

$$\begin{aligned} \mathbf{u} &\in L^\infty(I; W_0^{1,p}(\Omega)) \cap W^{1,2}(I; L^2(\Omega))\,, \\ \mathbf{F}(\mathbf{D}\mathbf{u}) &\in L^\infty(I; L^2(\Omega)) \cap L^2(I; W^{1,2}(\Omega))\,. \end{aligned}$$

To formulate clearly the dependence on the data in the various Estimates, we introduce the quantity

$$|||\mathbf{u}_0, \mathbf{f}|||^2 := \int_\Omega |\mathbf{u}_0|^2 + |\mathbf{D}\mathbf{u}_0(\mathbf{x})|^p\, d\mathbf{x} + \int_0^T \int_\Omega |\mathbf{f}(t, \mathbf{x})|^{p'} + |\mathbf{f}(t, \mathbf{x})|^2\, d\mathbf{x}\, dt\,.$$

As in the steady case, we replace the operator $\mathbf{S}$ in (25) by $\mathbf{S}^A$ as in Definition 5. The next step is the construction of a "strong solution." This is done by means of a Galerkin approximation of the A-approximation of problem (25), using a priori estimates obtained by formally testing with $\mathbf{u}^A$ and $\frac{\partial \mathbf{u}^A}{\partial t}$. In fact, we proceed as in the Proof of Proposition 5 and [9, Prop. 3.7] and use also Corollary 1. However, in contrast to the case $p > 2$ in [9, Prop. 3.7], we also need to approximate the initial condition $\mathbf{u}_0$ to obtain a priori estimates independent of δ^{-1}. In fact, if we would not do so, we have to handle the term $\omega^A(|\mathbf{D}\mathbf{u}_0|)$ which results from testing with

$\frac{\partial \mathbf{u}^A}{\partial t}$. This could be done by using Corollary 1 yielding $\omega^A(|\mathbf{D}\mathbf{u}_0|) \le c\,\delta^{p-2}|\mathbf{D}\mathbf{u}_0|^2$, which produces an undesired δ^{-1} dependence. We avoid this by approximating $\mathbf{u}_0$ in $W_0^{1,p}(\Omega) \cap L^2(\Omega)$ by an $\mathbf{u}_0^A \in W_0^{1,\infty}(\Omega)$ satisfying

$$\|\mathbf{D}\mathbf{u}_0^A\|_\infty \le A\,. \tag{26}$$

This could be done by using the "convolution-translation" method, which finds its introduction probably in the work of Puel and Roptin [19] and was rediscovered many times for different applications to partial differential equations or simply by appealing to standard properties of Sobolev functions and mollification. In fact, from the proof of [15, Thm. 5.5.2] and standard transformation and covering arguments, it follows that for $\mathbf{u}_0 \in W_0^{1,p}(\Omega) \cap L^2(\Omega)$ there exists a sequence $(\mathbf{w}_n) \subset W_0^{1,p}(\Omega) \cap L^2(\Omega)$ and $n_0 \in \mathbb{N}$ such that $\operatorname{supp}\mathbf{w}_n \subset \Omega_{\frac{1}{n}} := \{\mathbf{x} \in \Omega \,\big|\, \operatorname{dist}(\mathbf{x}, \partial\Omega) > \frac{1}{n}\}$, $n \ge n_0$, $\|\mathbf{D}\mathbf{w}_n\|_p \le 2\,\|\mathbf{D}\mathbf{u}_0\|_p$, $\|\mathbf{w}_n\|_2 \le 2\,\|\mathbf{u}_0\|_2$, $n \ge n_0$, and $\mathbf{w}_n \to \mathbf{u}_0$ in $W_0^{1,p}(\Omega) \cap L^2(\Omega)$. Thus, we can mollify with a standard mollification kernel ρ, which yields (for $0 < \varepsilon_n < 1/2n$) a sequence $\mathbf{v}_n := \rho_{\varepsilon_n} * \mathbf{w}_n$ belonging to $C_0^\infty(\Omega)$ and converging to $\mathbf{u}_0$ in $W_0^{1,p}(\Omega) \cap L^2(\Omega)$. We can choose ε_n such that it is a decreasing null sequence. Moreover, Hölder inequality yields

$$\begin{aligned}\|\mathbf{D}\mathbf{v}_n\|_\infty &= \|\rho_{\varepsilon_n} * \mathbf{D}\mathbf{w}_n\|_\infty \\ &\le \frac{1}{\varepsilon_n^{3/p}}\|\rho\|_{p'}\|\mathbf{D}\mathbf{w}_n\|_p \le \frac{2}{\varepsilon_n^{3/p}}\|\rho\|_{p'}\|\mathbf{D}\mathbf{u}_0\|_p =: A_n \nearrow \infty\,.\end{aligned}$$

For $A \in [A_n, A_{n+1})$, $n \ge n_0$, we set $\mathbf{u}_0^A := \mathbf{v}_n$, which satisfies (26) and converges to $\mathbf{u}_0$ in $W_0^{1,p}(\Omega) \cap L^2(\Omega)$.

Now, we can formulate the result showing the existence of a "strong solution."

Proposition 11 *Let the operator $\mathbf{S} = \partial U$, derived from the potential U, have (p,δ)-structure for some $p \in (1,2]$ and $\delta \in [0,\infty)$. Assume that $\mathbf{u}_0 \in W_0^{1,p}(\Omega) \cap L^2(\Omega)$ and $\mathbf{f} \in L^{p'}(I \times \Omega)$. Let $\mathbf{S}^A$ be as in Definition 5, and let $\mathbf{u}_0^A$ be as constructed above, satisfying (26). Then, for all $A \ge A_{n_0}$, the approximate problem*

$$\begin{cases} \dfrac{\partial \mathbf{u}^A}{\partial t} - \operatorname{div}\mathbf{S}^A(\mathbf{D}\mathbf{u}^A) = \mathbf{f} & \text{in } I \times \Omega\,, \\ \mathbf{u}^A = \mathbf{0} & \text{on } I \times \partial\Omega\,, \\ \mathbf{u}^A(0) = \mathbf{u}_0^A & \text{in } \Omega\,, \end{cases} \tag{27}$$

possesses a unique strong solution $\mathbf{u}^A$, i.e., $\mathbf{u}^A \in W^{1,2}(I; L^2(\Omega))$ with $\mathbf{F}^A(\mathbf{D}\mathbf{u}^A) \in L^\infty(I; L^2(\Omega))$, which satisfies for all $\psi \in C_0^\infty(0,T)$ and all $\mathbf{w} \in W_0^{1,2}(\Omega)$

$$\int_0^T \Big(\frac{\partial \mathbf{u}^A(t)}{\partial t}, \mathbf{w}\Big)\, \psi(t) + (\mathbf{S}^A(\mathbf{D}\mathbf{u}^A(t)), \mathbf{D}\mathbf{w})\, \psi(t)\, dt = \int_0^T (\mathbf{f}(t), \mathbf{w})\, \psi(t)\, dt\,.$$

In addition, the solution $\mathbf{u}^A$ *satisfies the estimate*

$$\operatorname*{esssup}_{t\in I} \Big(\|\mathbf{u}^A(t)\|_2^2 + \|\mathbf{F}^A(\mathbf{D}\mathbf{u}^A(t))\|_2^2 + (\delta + A)^{p-2}\|\nabla \mathbf{u}^A(t)\|_2^2 + \|\mathbf{F}(\mathbf{D}\mathbf{u}^A(t))\|_2^2\Big) \\ + \int_0^T \Big\|\frac{\partial \mathbf{u}^A(s)}{\partial t}\Big\|_2^2\, ds \le C\,\big(\delta^p + |||\mathbf{u}_0, \mathbf{f}|||^2\big)\,,$$

with C depending only on the characteristics of $\mathbf{S}$ *and* Ω*.*

Proof We do not give the full proof, which is a combination of Proposition 5 and [9, Prop. 3.7]. It differs from [9, Prop. 3.7] mainly in the approximation of the initial condition. Thus, we just derive the a priori estimates.

Formally taking $\mathbf{u}^A$ as test function in (25), we directly get

$$\frac{1}{2}\|\mathbf{u}^A(t)\|_2^2 + \int_0^t \int_\Omega \omega^A(|\mathbf{D}\mathbf{u}^A(s)|)\, d\mathbf{x}ds \le \frac{1}{2}\|\mathbf{u}_0^A\|_2^2 + C \int_0^t \int_\Omega \omega^*(|\mathbf{f}(s)|)\, d\mathbf{x}ds\,,$$

where the external force is treated (for a.e. $t \in I$) as in the proof of Proposition 5 (note that for this estimate, the special choice of $\mathbf{u}_0^A$ is not essential).

The second estimate is obtained by testing (25) by the time derivative of $\mathbf{u}^A$. In this way, we get

$$\int_0^t \Big\|\frac{\partial \mathbf{u}^A(s)}{\partial t}\Big\|_2^2\, ds + \int_\Omega \omega^A(|\mathbf{D}\mathbf{u}^A(t)|)\, d\mathbf{x} \le c \int_\Omega \omega^A(|\mathbf{D}\mathbf{u}_0^A|)\, d\mathbf{x} + c \int_0^t \|\mathbf{f}(s)\|_2^2\, ds\,.$$

The problem is that ω^A has a quadratic growth, while $\mathbf{D}\mathbf{u}_0$ belongs to $L^p(\Omega)$. To resolve this, we take advantage of the special approximation $\mathbf{u}_0^A$. In view of (26) and the definition of ω^A, we get

$$\omega^A(|\mathbf{D}\mathbf{u}_0^A(\mathbf{x})|) = \omega(|\mathbf{D}\mathbf{u}_0^A(\mathbf{x})|) \qquad \text{for a.e. } \mathbf{x} \in \Omega\,.$$

Consequently, testing (27) with $\mathbf{u}^A$ and $\frac{\partial \mathbf{u}^A}{\partial t}$ results in

$$\frac{1}{2}\|\mathbf{u}^A(t)\|_2^2+\int\limits_\Omega \omega^A(|\mathbf{D}\mathbf{u}^A(t)|)\,d\mathbf{x}+\int\limits_0^t\int\limits_\Omega \omega^A(|\mathbf{D}\mathbf{u}^A(s)|)\,d\mathbf{x}+\int\limits_0^t\Big\|\frac{\partial\mathbf{u}^A(s)}{\partial t}\Big\|_2^2\,ds$$

$$\leq\frac{1}{2}\|\mathbf{u}_0^A\|_2^2+\int\limits_\Omega \omega(|\mathbf{D}\mathbf{u}_0^A|)\,d\mathbf{x}+C\int\limits_0^t\int\limits_\Omega \omega^*(|\mathbf{f}(s)|)\,d\mathbf{x}ds+C\int\limits_0^t\|\mathbf{f}(s)\|_2^2\,ds\,.$$

The assertion follows, using the estimates from Corollary 1, the properties of the approximation $\mathbf{u}_0^A$, estimate (12), and the definition of $|||\mathbf{u}_0,\mathbf{f}|||$. □

By using the same tools employed in Sect. 3, one can prove the regularity in the interior and for tangential derivatives (with estimates independent of A). Also, regularity in normal direction follows analogously, but the estimates depend on A. More precisely, by adapting the translation method used in the Proof of Proposition 6, the result below can be proved:

Proposition 12 *Let the operator $\mathbf{S}=\partial U$, derived from the potential U, have (p,δ)-structure for some $p\in(1,2]$ and $\delta\in(0,\infty)$, with characteristics (γ_3,γ_4,p). Let $\Omega\subset\mathbb{R}^3$ be a bounded domain with $C^{2,1}$ boundary, and let $\mathbf{u}_0\in W_0^{1,p}(\Omega)\cap L^2(\Omega)$ and $\mathbf{f}\in L^{p'}(I\times\Omega)$. Then, the unique strong solution $\mathbf{u}^A$ of the approximate problem (27) satisfies for a.e. $t\in I$:*

$$\int\limits_0^t\int\limits_\Omega \xi_0^2|\nabla\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2+\omega^A\big(\xi_0^2|\nabla^2\mathbf{u}^A|\big)+(\delta+A)^{p-2}\xi_0^2|\nabla^2\mathbf{u}^A|^2\,d\mathbf{x}\,ds\leq c_0\,,$$

$$\int\limits_0^t\int\limits_\Omega \xi_P^2|\partial_\tau\mathbf{F}^A(\mathbf{D}\mathbf{u}^A)|^2+\omega^A\big(\xi_P^2|\partial_\tau\nabla\mathbf{u}^A|\big)+(\delta+A)^{p-2}\xi_P^2|\partial_\tau\nabla\mathbf{u}^A|^2\,d\mathbf{x}\,ds\leq c_P\,,$$

where $c_0=c_0(\delta,|||\mathbf{u}_0,\mathbf{f}|||,\|\xi_0\|_{1,\infty},\gamma_3,\gamma_4,p)$, while the constant related to the neighborhood of P is such that $c_P=c_P(\delta,|||\mathbf{u}_0,\mathbf{f}|||,\|\xi_P\|_{1,\infty},\|g_P\|_{C^{2,1}},\gamma_3,\gamma_4,p)$.

By using Proposition 12 and ellipticity of $\mathbf{S}^A$, we can write, for a.e. $(t,\mathbf{x})\in I\times\Omega$, the missing partial derivatives in the normal direction (which is locally $\mathbf{e}_3$ after a rotation of coordinates) in terms of the tangential ones, obtaining the following result:

Proposition 13 *Under the assumptions of Proposition 12, there exists a constant $C_1>0$ such that, provided in the local description of the boundary, there holds $r_P<C_1$ in (b3), where $\xi_P(\mathbf{x})$ is a cutoff function with support in Ω_P, and then*

$$\int_0^t \int_\Omega \xi_P^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2 + \omega^A\big(\xi_P^2 |\partial_3 \mathbf{Du}^A|\big)\, d\mathbf{x}\, ds \le C_A\,,$$

where $C_A = C_A(\delta, |||\mathbf{u}_0, \mathbf{f}|||, \|\xi_P\|_{1,\infty}, \|g_P\|_{C^{2,1}}, \gamma_3, \gamma_4, p, \omega, A)$.

Next, we improve the estimate in the normal direction in the sense that we will show that they are bounded uniformly with respect to the parameter $A \ge A_{n_0}$. At this stage, the time derivative is treated as an L^2-term on the right-hand side, while an $L^{p'}$-estimate would be needed to estimate it properly. This can be overcome by appropriate integration by parts. This step involves multiplying the equations by $\xi_P^2 \partial_{33}^2 \mathbf{u}^A$ and integrating by parts over the whole domain. To this end, the following technical result is used to justify the treatment of the time derivative:

Lemma 9 *Let* $\partial\Omega \in C^{2,1}$ *and let* $\mathbf{v} \in L^2(I; W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)) \cap W^{1,2}(I; L^2(\Omega))$. *Then, for all* $t \in [0, T]$, *it holds*

$$-\int_0^t \int_\Omega \frac{\partial \mathbf{v}}{\partial t} \partial_{33}^2 \mathbf{v}\, d\mathbf{x}\, dt = \frac{1}{2}\|\partial_3 \mathbf{v}(t)\|_2^2 - \frac{1}{2}\|\partial_3 \mathbf{v}(0)\|_2^2\,.$$

Note that this result requires that $\mathbf{u}_0 \in W_0^{1,2}(\Omega)$, and starting from this point, we need further regularity of the initial condition. With Lemma 9, one can prove the following result:

Proposition 14 *Let the same hypotheses as in Proposition 12 be satisfied and assume also* $\mathbf{u}_0 \in W_0^{1,2}(\Omega)$. *Let the local description* g_P *of the boundary and the localization function* ξ_P *satisfy* $(b1) - (b3)$ *and* $(\ell 1)$ *(cf. Sect. 3.1). Then, there exists a constant* $C_2 > 0$ *such that the unique strong solution* $\mathbf{u}^A$ *of the approximate problem* (27) *satisfies for every* $P \in \partial\Omega$ *and a.e.* $t \in I$

$$\int_0^t \int_\Omega \xi_P^2 |\partial_3 \mathbf{F}^A(\mathbf{Du}^A)|^2\, d\mathbf{x}\, ds \le C\,,$$

provided $r_P < C_2$ *in* $(b3)$, *with* C *depending on the characteristics of* $\mathbf{S}$, δ, $|||\mathbf{u}_0, \mathbf{f}|||$, $\|\mathbf{Du}_0\|_2$, $\|\xi_P\|_{1,\infty}$, $\|g_P\|_{C^{2,1}}$, *and* C_2.

From Propositions 11 and 14, we deduce in the same way as in the Proof of Proposition 10:

Proposition 15 *Under the assumption of Proposition 14, the unique strong solution* $\mathbf{u}^A$ *of the approximate problem* (27) *satisfies*

$$\operatorname*{esssup}_{t\in I}\Big(\|\mathbf{u}^A(t)\|_2^2+\|\mathbf{F}(\mathbf{Du}^A(t))\|_2^2\Big)+\int_0^T\int_\Omega |\nabla\mathbf{F}(\mathbf{Du}^A)(s)|^2+\Big|\frac{\partial\mathbf{u}^A(s)}{\partial t}\Big|_2^2\,d\mathbf{x}\,ds\le C\,,$$

with C depending on the characteristics of $\mathbf{S}$, δ, $|||\mathbf{u}_0,\mathbf{f}|||$, $\|\mathbf{Du}_0\|_2$, and the $C^{2,1}$-norms of the local description of $\partial\Omega$. In particular, $\mathbf{u}^A$ is uniformly bounded with respect to $A\ge 1$ in $L^p(I;W^{2,\frac{3p}{p+1}}(\Omega))$.

Finally, passing to the limit as $A\to\infty$ can be performed in a way similar to that used in the steady case: Observe that the bound on the time derivative allows us to use the Aubin-Lions lemma to infer the (space-time) convergence:

$$\mathbf{Du}^A\to\mathbf{Du}\qquad \text{a.e. in } I\times\Omega,\quad\text{and strongly in } L^2(I\times\Omega).$$

The rest of the argument requires minor changes to prove finally the following result:

Theorem 2 *Let the operator $\mathbf{S}$ in (25), derived from a potential U, have (p,δ)-structure for some $p\in(1,2]$ and $\delta\in(0,\infty)$. Let $\Omega\subset\mathbb{R}^3$ be a bounded domain with $C^{2,1}$ boundary. Assume that $\mathbf{u}_0\in W_0^{1,2}(\Omega)$ and $\mathbf{f}\in L^{p'}(I\times\Omega)$. Then, the system (25) has a unique regular solution with norms estimated only in terms of the characteristics of $\mathbf{S}$, δ, Ω, $\|\mathbf{u}_0\|_{1,2}$, and $\|\mathbf{f}\|_{p'}$.*

Acknowledgement Luigi C. Berselli was partially supported by a grant of the group GNAMPA of INdAM.

References

1. A.Kh. Balci, A. Cianchi, L. Diening, V. Maz'ya, A pointwise differential inequality and second-order regularity for nonlinear elliptic systems. Math. Ann. **383**(3-4), 1–50 (2022)
2. J.W. Barrett, W.B. Liu, Finite element approximation of the parabolic p-Laplacian. SIAM J. Numer. Anal. **31**, 413–428 (1994)
3. L. Beck, G. Mingione, Lipschitz bounds and nonuniform ellipticity. Comm. Pure Appl. Math. **73**(5), 944–1034 (2020)
4. H. Beirão da Veiga, P. Kaplický, M. Růžička, Boundary regularity of shear–thickening flows. J. Math. Fluid Mech. **13**, 387–404 (2011)
5. L. Belenki, L.C. Berselli, L. Diening, M. Růžička, On the finite element approximation of p-Stokes systems. SIAM J. Numer. Anal. **50**(2), 373–397 (2012)
6. L.C. Berselli, C.R. Grisanti, On the regularity up to the boundary for certain nonlinear elliptic systems. Discrete Contin. Dyn. Syst. Ser. S **9**(1), 53–71 (2016)
7. L.C. Berselli, M. Růžička, Global regularity properties of steady shear thinning flows. J. Math. Anal. Appl. **450**(2), 839–871 (2017)
8. L.C. Berselli, M. Růžička, Global regularity for systems with p-structure depending on the symmetric gradient. Adv. Nonlinear Anal. **9**(1), 176–192 (2020)

9. L.C. Berselli, M. Růžička, Natural second-order regularity for parabolic systems with stress tensor with (p, δ)-structure and depending only on the symmetric gradient. Calc. Var. Partial Differential Equations **61**(4) Paper No. 137, 49 pp. (2022)
10. L.C. Berselli, M. Růžička, Space-time discretization for nonlinear parabolic systems with p-structure. IMA J. Numer. Analy. **42**(1), 260–299 (2022)
11. L.C. Berselli, L. Diening, M. Růžička, Existence of strong solutions for incompressible fluids with shear dependent viscosities. J. Math. Fluid Mech. **12**(1), 101–132 (2010)
12. A. Cianchi, V.G. Maz'ya, Optimal second-order regularity for the p-Laplace system. J. Math. Pures Appl. **132**(9), 41–78 (2019)
13. A. Cianchi, V.G. Maz'ya, Second-order regularity for parabolic p-Laplace problems. J. Geom. Anal. **30**(2), 1565–1583 (2020)
14. L. Diening, M. Růžička, K. Schumacher, A decomposition technique for John domains. Ann. Acad. Sci. Fenn. Math. **35**(1), 87–114 (2010)
15. L.C. Evans, *Partial Differential Equations*. Graduate Studies in Mathematics, vol. 19 (American Mathematical Society, Providence, 1998)
16. M.A. Krasnoselskiĭ, J.B. Rutickiĭ, *Convex Functions and Orlicz Spaces*. Translated from the first Russian edition by L.F. Boron (P. Noordhoff, Groningen, 1961)
17. J. Málek, J. Nečas, M. Růžička, On weak solutions to a class of non–Newtonian incompressible fluids in bounded three-dimensional domains. the case $p \geq 2$. Adv. Diff. Equ. **6**, 257–302 (2001)
18. J. Musielak, *Orlicz Spaces and Modular Spaces* (Springer, Berlin, 1983)
19. J.P. Puel, M.C. Roptin, Théorème de densité résultant du lemme de friedrichs. Technical Report, Université de Rennes, 1967. Rapport de stage dirigé par C. Goulaouic, DEA
20. M.M. Rao, Z.D. Ren, *Theory of Orlicz Spaces*. Monographs and Textbooks in Pure and Applied Mathematics, vol. 146 (Dekker, New York, 1991)
21. M. Růžička, L. Diening, Non–Newtonian fluids and function spaces, in *Nonlinear Analysis, Function Spaces and Applications, Proceedings of NAFSA 2006 Prague*, vol. 8 (2007), pp. 95–144
22. G.A. Seregin, T.N. Shilkin, Regularity of minimizers of some variational problems in plasticity theory. Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **243**(Kraev. Zadachi Mat. Fiz. i Smezh. Vopr. Teor. Funktsii. 28):270–298, 342–343 (1997)
23. G. Talenti, Nonlinear elliptic equations, rearrangements of functions and Orlicz spaces. Ann. Mat. Pura Appl. **120**(4), 160–184 (1979)

Three-Dimensional Velocity Field Using the Cross-Model Viscosity Function

Fernando Carapau, Paulo Correia, and Pedro Areias

1 Introduction

Let us consider the constitutive equation for an incompressible and homogeneous linearly viscous fluid where the Cauchy stress tensor is given by

$$\boldsymbol{T} = -p\boldsymbol{I} + 2\mu\boldsymbol{D}, \tag{1}$$

where p is the hydrostatic pressure, μ the constant viscosity, and $\boldsymbol{D}$ the symmetric part of the velocity gradient, also called the rate of deformation tensor

$$\boldsymbol{D} := \frac{1}{2}\Big(\nabla\boldsymbol{\vartheta} + \big(\nabla\boldsymbol{\vartheta}\big)^T\Big), \tag{2}$$

where[1] $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\boldsymbol{x}, t)$ is the three-dimensional velocity field, $\nabla\boldsymbol{\vartheta}$ is the spatial velocity gradient, and $\big(\nabla\boldsymbol{\vartheta}\big)^T$ denotes the transpose of $\nabla\boldsymbol{\vartheta}$. The fluids that comply with Eq. (1) are known in the scientific literature as Newtonian fluids. On the other hand, there are fluids for which the viscosity is not constant, and it may depend on

[1] Let $\boldsymbol{x} = (x_1, x_2, x_3)$ be the rectangular space Cartesian coordinates (for convenience, we set $x_3 = z$) and t is the time variable.

F. Carapau (✉) · P. Correia
Departamento de Matemática and CIMA, Universidade de Évora, Évora, Portugal
e-mail: flc@uevora.pt; pcorreia@uevora.pt

P. Areias
Departamento de Engenharia Mecânica, Universidade Técnica de Lisboa, IST, Lisboa, Portugal
e-mail: pedro.areias@tecnico.ulisboa.pt

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_2

certain parameters, as pressure and/or shear rate. These fluids for which the viscosity is not constant are known as non-Newtonian fluids.

For many real fluids, the viscosity of the flow changes with the intensity of the rate of deformation tensor (see, for example, [1]). This change of the viscosity can be very large in some fluids, and it cannot be ignored. Throughout this work, we will consider that the viscosity only depends on the intensity of the shear rate. The simplest way to model such behavior is to introduce in (1) the viscosity as a function of shear rate:

$$\mu(|\dot{\gamma}|) : \mathbb{R}^+ \to \mathbb{R}^+,$$

where $\dot{\gamma}$ is a scalar measure of the rate of shear defined by

$$|\dot{\gamma}| = \sqrt{2\boldsymbol{D} : \boldsymbol{D}}.$$

Therefore, the Cauchy stress tensor in (1) takes the form

$$\boldsymbol{T} = -p\boldsymbol{I} + \mu(|\dot{\gamma}|)\Big(\nabla\boldsymbol{\vartheta} + \big(\nabla\boldsymbol{\vartheta}\big)^T\Big). \tag{3}$$

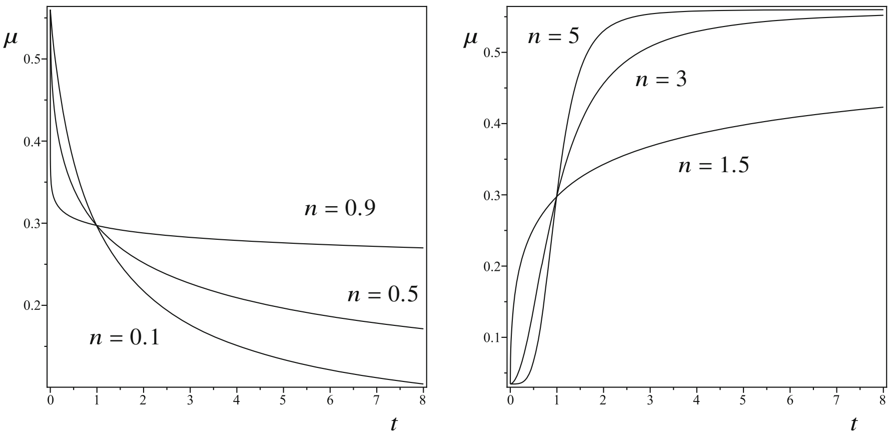
The class of non-Newtonian fluids satisfying condition (3) is called generalized Newtonian fluids (or quasi-Newtonian). In general, we can divide the generalized Newtonian fluid into two subclass: the shear-thinning (or pseudoplastic) fluids where the viscosity decreases with the increasing shear rate and the shear-thickening (or dilatant) fluids for which the viscosity increases with the increasing shear rate. The shear-thinning behavior is commonly observed in real fluids, for example, suspensions, emulsions, polymeric fluids (see, for example, [2–4]). The shear-thickening behavior is less common, although it can be observed at highly loaded suspensions, for example, starch, plaster, and a few unusual polymeric fluids (see, for example, [2–4]).

Next, we will present the specific viscosity function under study in this work, that is, the cross model, where the viscosity function in (3) is given by

$$\mu(|\dot{\gamma}|) = \mu_\infty + \frac{\mu_0 - \mu_\infty}{1 + (k|\dot{\gamma}|)^{1-n}}. \tag{4}$$

Here, parameters k and n are called the consistency index and the flow index (positive constants), respectively. In this model, we consider fluids with bounded low μ_0 and high limiting viscosities μ_∞. Considering, $n = 1$ in Eq. (4), the Cauchy stress tensor (3) corresponds to the Newtonian fluid behavior with $\mu = (\mu_\infty + \mu_0)/2$. Moreover, if $n < 1$, we obtain

$$\lim_{|\dot{\gamma}|\to\infty} \mu(|\dot{\gamma}|) = \mu_\infty, \quad \lim_{|\dot{\gamma}|\to 0} \mu(|\dot{\gamma}|) = \mu_0,$$



(a) (b)

Fig. 1 Cross model: (**a**) shear-thinning viscosity and (**b**) shear-thickening viscosity. Both cases with values $k = 1.007$ s, $\mu_0 = 0.56$ poise, and $\mu_\infty = 0.0345$ poise (see, for example, [5, 6]), for different values of flow index

and the fluid shows us a shear-thinning behavior (see Fig. 1). If $n > 1$, then

$$\lim_{|\dot{\gamma}|\to\infty} \mu(|\dot{\gamma}|) = \mu_0, \quad \lim_{|\dot{\gamma}|\to 0} \mu(|\dot{\gamma}|) = \mu_\infty,$$

and we have a shear-thickening fluid behavior (see Fig. 1).

Numerical simulations relating to a three-dimensional model for a homogeneous incompressible fluid based on the Cauchy stress tensor (3) with viscosity function (4), for a given geometry, require a high computational effort. In this sense, theories that allow us to reduce the complexity of the problems under study by reducing variables are important. A possible simplification is to consider the evolution of average flow quantities using simpler one-dimensional models. Usually, classical one-dimensional models are obtained by imposing additional assumptions related to the nonlinear convective acceleration and the viscous dissipation terms. These closure approximations are typically based on assuming a purely axial flow with a field dependence on axial variables (see, for example, [7–9]). In this work, we present an alternative theory to reduce the three-dimensional model under study to a one-dimensional system of ordinary differential equations, which depend only on time and on a single spatial variable, by using the Cosserat theory associated with fluid dynamics (see Caulk and Naghdi [10]). The basis of this theory (see Duhem [11]) and Eugène and François Cosserat [12]) is to consider an additional structure of deformable vectors (called directors) assigned to each point on a spatial curve (the Cosserat curve). The use of directors in continuum mechanics goes back to Duhem [11], who regarded a body as a collection of points, together with associated directions. This theory has also been used by several authors in studies of rods, plates, and shells (see, for example, [13–17]). An analogous hierarchical theory related to fluid dynamics has been developed by Caulk and Naghdi [10] and Green et al. [18–20]. Recently, this hierarchical theory has been applied to models associated with hemodynamics (see Robertson and Sequeira [21] and Carapau and Sequeira [22]). Regarding the swirling motion, this hierarchical theory was used to study several models (see Caulk and Naghdi [10] and Carapau et al. [23–25]). Also, this hierarchical theory has been applied for specific models related to non-Newtonian fluids under different geometries and perspectives (see Carapau [26, 27], Carapau and Correia [28], and Carapau et al. [29, 30]). This alternative approach theory has been validated by the works of Caulk and Naghdi [10], Robertson and Sequeira [21], Carapau and Sequeira [22, 29], and Carapau [27].

The advantage of using the Cosserat theory related to fluid dynamics is not so much getting an approximation of the three-dimensional system but rather in using it as an independent framework to predict some properties of the three-dimensional problem under study. The main features of the director theory are as follows: it incorporates all components of the linear momentum equation; it is a hierarchical theory, making it possible to increase the accuracy of the model; the system of equations is closed at each order and therefore unnecessary to make assumptions about the form of the nonlinear and viscous terms; invariance under superposed rigid body motions is satisfied at each order; the wall shear stress enters directly as a dependent variable in the formulation; and the director theory has been shown to be useful for modeling flow in curved tubes, considering many more directors than in the case of a straight tube. A detailed discussion about Cosserat theory, related to fluid dynamics, can be found in [10, 18–20]. The three-dimensional numerical study of the flow associated with an incompressible fluid that follows the constitutive equation (3) with viscosity function (4) in a circular cross-section tube with constant radius is in fact a challenging and complex study in terms of computational effort and infeasible in many relevant issues. Our one-dimensional approach is obtained by integrating the linear momentum equation over the cross section of the tube, taking the three-dimensional velocity field approximation provided by the Cosserat theory. This procedure yields a one-dimensional system, depending only on time and a single spatial variable, which is the axis of the symmetrical flow. This velocity field approximation satisfies exactly both the incompressibility condition and the kinematic boundary condition. Based on the work of Caulk and Naghdi (see [10]), we consider the three-dimensional velocity field $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}(\boldsymbol{x}, t)$ approximated by:[2]

$$\boldsymbol{\vartheta} = \boldsymbol{v} + \sum_{N=1}^{k} x_{\alpha_1} \dots x_{\alpha_N} \boldsymbol{W}_{\alpha_1 \dots \alpha_N}, \tag{5}$$

with

$$\boldsymbol{v} = v_i(z,t)\,\boldsymbol{e}_i, \quad \boldsymbol{W}_{\alpha_1 \dots \alpha_N} = W^i_{\alpha_1 \dots \alpha_N}(z,t)\,\boldsymbol{e}_i. \tag{6}$$

[2] In the sequel, Latin indices take the values 1, 2, and 3 and Greek indices 1 and 2, and we use the convention of summing over repeated indices.

In condition (5), $\boldsymbol{v}$ denotes the velocity along the axis of symmetry z at time t, $x_{\alpha_1}\dots x_{\alpha_N}$ are the polynomial weighting functions with order k, the vectors $\boldsymbol{W}_{\alpha_1\dots\alpha_N}$ are the director velocities which are symmetric with respect to their indices, and $\boldsymbol{e}_i$ are the associated unit basis vectors. We remark that the number k identifies the order in the hierarchical theory and is related to the number of directors. In applications, these director velocities are associated with physical characteristics of the fluid. Considering the velocity field approximation (5) with nine directors (see [10]), i.e., $k = 3$ in (5) and the constitutive condition (3) with viscosity function (4) in our one-dimensional model, we obtain the unsteady equation for mean pressure gradient depending on the volume flow rate, Womersley number, and viscosity parameters over a finite section of a straight, rigid, and impermeable tube with constant circular cross section. Attention is focused on some numerical simulations for constant and nonconstant mean pressure gradient using a Runge-Kutta method. In particular, given a specific data, we get information about the volume flow rate, and consequently we can illustrate the three-dimensional velocity field behavior on the circular cross section of the tube.

2 Governing Equations

Taking into account the constitutive condition (3) with viscosity function (4), we consider the motion of a homogeneous incompressible generalized Newtonian fluid without body forces inside straight rigid and impermeable rectilinear tube with circular cross section of constant radius (see Fig. 2). The boundary of the fluid is defined by the surface scalar constant function ϕ, which is related to the circular cross-section straight tube by the following relationship:

$$\phi^2 = x_1^2 + x_2^2. \tag{7}$$

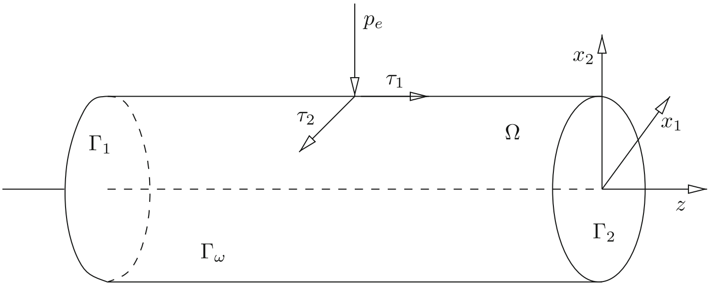


Fig. 2 Fluid domain Ω with normal and tangential components of the surface traction vector p_e and τ_1, τ_2 with constant circular cross section ϕ along the axis of symmetry z. The boundary $\partial\Omega$ is composed by the proximal cross section Γ_1, by the distal cross section Γ_2, and by the lateral wall of the tube Γ_w

Therefore, the equations of motion, considering conservation of linear momentum and mass, are given in $\Omega \times (0, T)$ by

$$\begin{cases} \rho\Big(\dfrac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla \boldsymbol{\vartheta}\Big) = \nabla \cdot \boldsymbol{T}, \\ \nabla \cdot \boldsymbol{\vartheta} = 0, \\ \boldsymbol{T} = -p\boldsymbol{I} + \big(\mu_\infty + \dfrac{\mu_0 - \mu_\infty}{1 + (k|\dot{\gamma}|)^{1-n}}\big)\Big(\nabla\boldsymbol{\vartheta} + \big(\nabla\boldsymbol{\vartheta}\big)^T\Big), \quad \boldsymbol{t}_w = \boldsymbol{T} \cdot \boldsymbol{n}, \end{cases} \tag{8}$$

with the initial condition

$$\boldsymbol{\vartheta}(\boldsymbol{x}, 0) = \boldsymbol{\vartheta}_0(\boldsymbol{x}) \quad \text{in} \quad \Omega, \tag{9}$$

and the homogeneous Dirichlet boundary condition

$$\boldsymbol{\vartheta}(\boldsymbol{x}, t) = 0 \quad \text{on} \quad \Gamma_w \times (0, T), \tag{10}$$

where ρ is the constant density of fluid. Equation $(8)_1$ represents the balance of linear momentum, and $(8)_2$ is the incompressibility condition. The constitutive equation appears in $(8)_3$ and $\boldsymbol{t}_w$ denotes the stress vector on the surface whose outward unit normal vector is $\boldsymbol{n}(\boldsymbol{x}, t) = n_i(\boldsymbol{x}, t)\boldsymbol{e}_i$. The components of the outward unit normal vector to the surface ϕ are given by

$$n_1 = \frac{x_1}{\phi}, \quad n_2 = \frac{x_2}{\phi}, \quad n_3 = 0. \tag{11}$$

The theoretical study of the model (8)–(10), namely, existence, uniqueness, and regularity of classical and weak solutions, still poses some difficulties. In this work, we are interested in computational simulations of the model (8)–(10), using the director approach related to fluid dynamics. Since Eq. (7) defines a material surface, the three-dimensional velocity field $\boldsymbol{\vartheta}$ must satisfy the kinematic condition[3]

$$\frac{d}{dt}\big(\phi^2 - x_1^2 - x_2^2\big) = 0,$$

i.e.,

$$-x_1\vartheta_1 - x_2\vartheta_2 = 0, \tag{12}$$

on the boundary defined by (7). Averaged quantities such as volume flow rate and pressure are needed to study one-dimensional models. Consider $S = S(z, t)$ a generic axial section of the domain Ω at time t defined by the spatial variable z, bounded by the circle defined by (7), and let $A(z, t)$ be the area of this section $S(z, t)$. Then, the volume flow rate Q is defined by

[3] The material time derivative is given by $\frac{d}{dt}(\,\cdot\,) = \frac{\partial}{\partial t}(\,\cdot\,) + \boldsymbol{\vartheta} \cdot \nabla(\,\cdot\,)$.

$$Q(z,t) = \int_{S(z,t)} \vartheta_3(\boldsymbol{x},t)da, \tag{13}$$

and the average pressure $\bar{p}$ by

$$\bar{p}(z,t) = \frac{1}{A(z,t)} \int_{S(z,t)} p(\boldsymbol{x},t)da. \tag{14}$$

Next, considering (5), it follows (see [10]) that the approximation of the three-dimensional velocity field $\boldsymbol{\vartheta} = \vartheta_i(\boldsymbol{x},t)\boldsymbol{e}_i$ using nine directors is given by

$$\begin{aligned}\boldsymbol{\vartheta} = &\Big[x_1(\xi + \sigma(x_1^2 + x_2^2)) - x_2(\omega + \eta(x_1^2 + x_2^2))\Big]\boldsymbol{e}_1\\ &+ \Big[x_1(\omega + \eta(x_1^2 + x_2^2)) + x_2(\xi + \sigma(x_1^2 + x_2^2))\Big]\boldsymbol{e}_2\\ &+ \Big[v_3 + \gamma(x_1^2 + x_2^2)\Big]\boldsymbol{e}_3,\end{aligned} \tag{15}$$

where $\xi, \omega, \gamma, \sigma, and \eta$ are scalar functions of the spatial variable z and time t. The physical significance of these scalar functions in (15) is the following: γ is related to transverse shearing motion, ω and η are related to rotational motion (also called swirling motion) about $\boldsymbol{e}_3$, while ξ and σ are related to transverse elongation. We use nine directors because it is the minimum number for which the incompressibility condition and the kinematic boundary conditions on the lateral surface of the tube are satisfied pointwise. Using the velocity approach (15), the kinematic conditions (12) on the lateral boundary reduce to

$$-\phi^2(\xi + \phi^2\sigma) = 0, \tag{16}$$

and the incompressibility condition given by Eq. $(8)_2$ becomes

$$(v_3)_z + 2\xi + (x_1^2 + x_2^2)(\gamma_z + 4\sigma) = 0, \tag{17}$$

where the subscripted variable denotes partial differentiation. For Eq. (17) to hold at every point in the fluid, the velocity coefficients must satisfy the separate conditions:

$$(v_3)_z + 2\xi = 0, \quad \gamma_z + 4\sigma = 0. \tag{18}$$

Hence, the boundary condition (12) and the incompressibility condition given by Eq. $(8)_2$ are satisfied exactly by the velocity field (15) if we impose the conditions (16) and (18). On the wall boundary of the rigid tube, we impose the no-slip boundary condition requiring that the velocity field (15) vanishes identically on the surface (7), i.e., condition (10) is satisfied. Thus, it follows that

$$\xi + \phi^2\sigma = 0, \quad \omega + \phi^2\eta = 0, \quad v_3 + \phi^2\gamma = 0. \tag{19}$$

Therefore, Eq. (16) is satisfied identically, and the two incompressibility conditions (18) reduce to

$$(v_3)_z + 2\xi = 0, \quad (\phi^2 v_3)_z = 0. \tag{20}$$

Considering the flow in a rigid tube with constant circular cross section given by surface (7) without swirling motion (i.e., $\omega = \eta = 0$), conditions (13), (15), (19), and (20), then the volume flow rate Q is just a function of time t, given by

$$Q(t) = \frac{\pi}{2}\, \phi^2 v_3(z,t), \tag{21}$$

and, consequently, the velocity field (15) can be rewritten as

$$\boldsymbol{\vartheta}(\boldsymbol{x},t) = \frac{2Q(t)}{\pi\phi^2}\Big(1 - \frac{x_1^2 + x_2^2}{\phi^2}\Big)\boldsymbol{e}_3, \tag{22}$$

and the initial condition (9) is satisfied when we consider in computational simulations $Q(0) = \text{const.}$

To simplify the computational effort, it is convenient to introduce the stress vector $\boldsymbol{t}_w$ on the lateral surface in terms of its outward unit normal $\boldsymbol{n}$ and in terms of the components of the surface traction vector τ_1, τ_2 and p_e in the form (see [10])

$$\boldsymbol{t}_w = \tau_1 \boldsymbol{\lambda} - p_e \boldsymbol{n} + \tau_2 \boldsymbol{e}_\theta, \tag{23}$$

where τ_1 is the wall shear stress, while $\boldsymbol{\lambda}$ and $\boldsymbol{e}_\theta$ are the unit tangent vectors defined by

$$\boldsymbol{\lambda} = \boldsymbol{n} \times \boldsymbol{e}_\theta, \quad \boldsymbol{e}_\theta = (x_\alpha/\phi)\boldsymbol{e}_{\alpha\beta}\boldsymbol{e}_\beta, \tag{24}$$

with $\boldsymbol{e}_{11} = \boldsymbol{e}_{22} = 0$ and $\boldsymbol{e}_{12} = -\boldsymbol{e}_{21} = 1$. Using conditions (11) and (24), the expression for the stress vector (23) can be rewritten in terms of its rectangular Cartesian components as

$$\boldsymbol{t}_w = \frac{1}{\phi}(-p_e x_1 - \tau_2 x_2)\boldsymbol{e}_1 + \frac{1}{\phi}(-p_e x_2 + \tau_2 x_1)\boldsymbol{e}_2 + \tau_1 \boldsymbol{e}_3. \tag{25}$$

Next, instead of the momentum equation $(8)_1$ be verified pointwise in the fluid, we impose the following integral conditions (see [10]):

$$\int_S \Big[\nabla \cdot \boldsymbol{T} - \rho\Big(\frac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla \boldsymbol{\vartheta}\Big)\Big] da = 0, \tag{26}$$

$$\int_S \Big[\nabla \cdot \boldsymbol{T} - \rho\Big(\frac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla \boldsymbol{\vartheta}\Big)\Big] x_{\alpha_1} \ldots x_{\alpha_N} da = 0, \tag{27}$$

where $N = 1, 2, 3$. Using the divergence theorem and a form of Leibniz rule, Eqs. (26) and (27) for nine directors can be reduced to the following vector equations:

$$\frac{\partial \boldsymbol{h}}{\partial z} + \boldsymbol{f} = \boldsymbol{a}, \tag{28}$$

and

$$\frac{\partial \boldsymbol{m}^{\alpha_1 \ldots \alpha_N}}{\partial z} + \boldsymbol{l}^{\alpha_1 \ldots \alpha_N} = \boldsymbol{k}^{\alpha_1 \ldots \alpha_N} + \boldsymbol{b}^{\alpha_1 \ldots \alpha_N}, \tag{29}$$

where $\boldsymbol{h}$, $\boldsymbol{k}^{\alpha_1 \ldots \alpha_N}$, $\boldsymbol{m}^{\alpha_1 \ldots \alpha_N}$ are resultant forces defined by

$$\boldsymbol{h} = \int_S \boldsymbol{T}_3 da, \quad \boldsymbol{k}^{\alpha} = \int_S \boldsymbol{T}_{\alpha} da, \quad \boldsymbol{k}^{\alpha\beta} = \int_S \Big(\boldsymbol{T}_{\alpha} x_{\beta} + \boldsymbol{T}_{\beta} x_{\alpha}\Big) da, \tag{30}$$

$$\boldsymbol{k}^{\alpha\beta\gamma} = \int_S \Big(\boldsymbol{T}_{\alpha} x_{\beta} x_{\gamma} + \boldsymbol{T}_{\beta} x_{\alpha} x_{\gamma} + \boldsymbol{T}_{\gamma} x_{\alpha} x_{\beta}\Big) da, \tag{31}$$

and

$$\boldsymbol{m}^{\alpha_1 \ldots \alpha_N} = \int_S \boldsymbol{T}_3 x_{\alpha_1} \ldots x_{\alpha_N} da. \tag{32}$$

The quantities $\boldsymbol{a}$ and $\boldsymbol{b}^{\alpha_1 \ldots \alpha_N}$ are inertia terms defined by

$$\boldsymbol{a} = \int_S \rho \Big(\frac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla \boldsymbol{\vartheta}\Big) da, \tag{33}$$

$$\boldsymbol{b}^{\alpha_1 \ldots \alpha_N} = \int_S \rho \Big(\frac{\partial \boldsymbol{\vartheta}}{\partial t} + \boldsymbol{\vartheta} \cdot \nabla \boldsymbol{\vartheta}\Big) x_{\alpha_1} \ldots x_{\alpha_N} da, \tag{34}$$

and $\boldsymbol{f}$, $\boldsymbol{l}^{\alpha_1 \ldots \alpha_N}$, which arise due to surface traction on the lateral boundary, are defined by

$$\boldsymbol{f} = \int_{\partial S} \boldsymbol{t}_w \, ds, \tag{35}$$

$$\boldsymbol{l}^{\alpha_1 \ldots \alpha_N} = \int_{\partial S} \boldsymbol{t}_w \, x_{\alpha_1} \ldots x_{\alpha_N} ds. \tag{36}$$

Next, we will derive the equation for the mean pressure gradient using the computed values for the quantities (30)–(36) in Eqs. (28)–(29) according to [10].

3 Main Results and Simulations

The computational effort to calculate the quantities (30)–(36) related to the constitutive equation (8)$_3$ for any index flow n (i.e., shear-thinning viscosity and shear-thickening viscosity) is difficult to handle. This difficulty is related to computational problems arising from the calculation of integrals with singularities. However, for some positive integer values of n, the difficulty can be overcome. Therefore, considering the choice $n = 3$ on Eq. (8)$_3$, the equation for the mean pressure gradient will be obtained using the resulting quantities from (30) to (36) on Eqs. (28)–(29).

In sequence, using the velocity field (22), the surface (7), the volume flow rate (21), and the stress vector (25) in Eqs. (30)–(36), we can explicitly calculate the forces $\boldsymbol{h}$, $\boldsymbol{k}^{\alpha}$, $\boldsymbol{k}^{\alpha\beta}$, $\boldsymbol{k}^{\alpha\beta\gamma}$, $\boldsymbol{m}^{\alpha_1 \ldots \alpha_N}$, the inertia terms $\boldsymbol{a}$, $\boldsymbol{b}^{\alpha_1 \ldots \alpha_N}$, and the surface tractions $\boldsymbol{f}$, $\boldsymbol{l}^{\alpha_1 \ldots \alpha_N}$. Hence, plugging these solutions into Eqs. (28)–(29) and using Eq. (14), by solving a linear system, we get the unsteady equation for the average pressure gradient, given by

$$\begin{aligned}\bar{p}_z(z,t) = &-\frac{4\rho}{3\pi\phi^2}Q_t(t) - \frac{8\mu_0}{\pi\phi^4}Q(t)\\ &+ \big(\mu_0 - \mu_\infty\big)\Big[\frac{\pi^3\phi^8}{64\,k^4\,Q^3(t)} ln\Big(\frac{32k^2Q^2(t) + \pi^2\phi^6}{\pi^2\phi^6}\Big)\\ &- \frac{\pi\phi^2}{2\,k^2\,Q(t)}\Big],\end{aligned} \tag{37}$$

Integrating condition (37) over a finite section of the tube between z_1 and z_2 with $z_1 < z_2$, we obtain the mean pressure gradient over the interval $[z_1, z_2]$ at time t, given by

$$\begin{aligned}G(t) = &\frac{4\rho}{3\pi\phi^2}Q_t(t) + \frac{8\mu_0}{\pi\phi^4}Q(t) + \big(\mu_0 - \mu_\infty\big)\Big[\frac{\pi\phi^2}{2\,k^2\,Q(t)}\\ &- \frac{\pi^3\phi^8}{64\,k^4\,Q^3(t)} ln\Big(\frac{32k^2Q^2(t) + \pi^2\phi^6}{\pi^2\phi^6}\Big)\Big],\end{aligned} \tag{38}$$

where

$$G(t) = \frac{\bar{p}(z_1,t) - \bar{p}(z_2,t)}{z_2 - z_1}.$$

Next, let us consider the following dimensionless variables:

$$\hat{t} = \omega_0 t, \quad \hat{Q}(\hat{t}) = \frac{2\rho}{\pi\phi k}Q(t), \quad \hat{G}(\hat{t}) = \frac{\rho^3\phi^7}{k^4}G(t), \tag{39}$$

where ω_0 is the characteristic frequency for unsteady flows. In the cases where a steady volume flow rate is specified, the nondimensional volume flow rate $\hat{Q}$ is identical to the classical Reynolds number used for flow in tubes (see Robertson and Sequeira [21]). Substituting the new variables (39) in Eq. (38), we obtain the nondimensional mean pressure gradient:

$$\hat{G}(\hat{t}) = \frac{2}{3}\mathcal{W}_o^2 \hat{Q}_{\hat{t}}(\hat{t}) + 4\mathcal{A}_\mu \hat{Q}(\hat{t}) + \mathcal{B}_\mu \Big[\frac{1}{\hat{Q}(\hat{t})} - \frac{1}{8}\frac{C_\mu}{\hat{Q}^3(\hat{t})} ln\Big(8\frac{\hat{Q}^2(\hat{t})}{C_\mu} + 1 \Big) \Big], \tag{40}$$

where $\mathcal{W}_o = \phi^3\sqrt{\rho^3\omega_0/k^3}$ is the Womersley number, which is the most commonly used parameter to reflect the pulsatility of the flow and $\mathcal{A}_\mu$, $\mathcal{B}_\mu$, and C_μ are viscosity parameters, given by

$$\mathcal{A}_\mu = \frac{\mu_0\rho^2\phi^4}{k^3}, \ \mathcal{B}_\mu = \frac{(\mu_0 - \mu_\infty)\rho^4\phi^4}{k^7}, \ C_\mu = \frac{\rho^2\phi^4}{k^4}. \tag{41}$$

Moreover, using $(39)_2$ and the dimensionless variables

$$\hat{x}_1 = \frac{x_1}{\phi}, \quad \hat{x}_2 = \frac{x_2}{\phi}, \quad \hat{z} = \frac{z}{\phi}, \quad \hat{\boldsymbol{\vartheta}}(\hat{\boldsymbol{x}}, \hat{t}) = \frac{\phi\rho}{k}\boldsymbol{\vartheta}(\boldsymbol{x}, t), \tag{42}$$

at the velocity equation (22), we get the nondimensional three-dimensional velocity field:

$$\hat{\boldsymbol{\vartheta}}(\hat{\boldsymbol{x}}, \hat{t}) = \hat{Q}(\hat{t})\Big(1 - (\hat{x}_1^2 + \hat{x}_2^2)\Big)\boldsymbol{e}_3. \tag{43}$$

In the next section, we present numerical simulations associated with the Eqs. (40) and (43) for specific flow regimes, considering

$$\mathcal{A}_\mu \to 1, \quad \mathcal{B}_\mu \to 0, \quad C_\mu \neq 0, \tag{44}$$

in order to reduce the computational effort.

3.1 Constant Mean Pressure Gradient

In Fig. 3, we can observe the behavior of the unsteady volume flow rate solution given by (40) obtained using a Runge-Kutta method with constant mean pressure gradient $\hat{G}(\hat{t}) = 1$ when we increase the Womersley number. Therefore, we note that the amplitude of the solution in the initial transient phase increases and becomes less pronounced as the Womersley number increases. In this particular case of a constant mean pressure gradient, the volume flow rate given by (40)

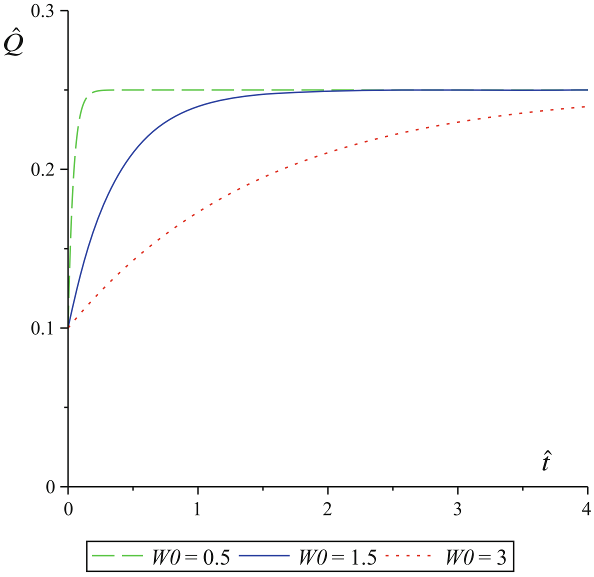


Fig. 3 Unsteady volume flow rate given by Eq. (40) with constant mean pressure gradient $\hat{G}(\hat{t}) = 1$ where $\hat{Q}(0) = 0.1$ and $\mathcal{W}_o = (0.5; 1.5; 3)$ for shear-thickening fluids with $n = 3$

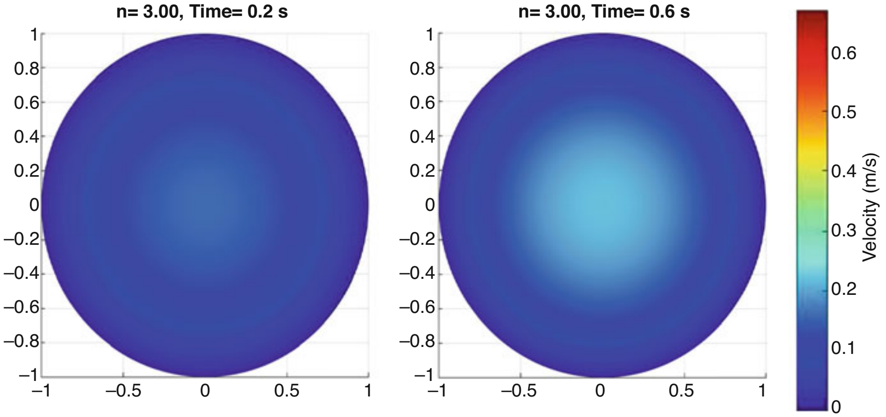


Fig. 4 Three-dimensional velocity field (43) where the volume flow rate is obtained by (40) with $\hat{G}(\hat{t}) = 1$, $\hat{Q}(0) = 0.1$, $\mathcal{W}_o = 1.5$, and $n = 3$ (shear-thickening fluid). Time parameters: $\hat{t} = 0.2, \hat{t} = 0.6$

converges toward to the steady-state solution, converging faster for small values of the Womersley number, i.e., when $\mathcal{W}_o \rightarrow 0$.

Moreover, with the information of the volume flow rate given by (40), obtained for certain flow regimes, we can return to the three-dimensional problem to obtain the behavior of the three-dimensional velocity field (43) in time on the circular cross section of the tube. Figures 4 and 5 illustrate the three-dimensional velocity field (43) behavior in the circular cross section of the tube when we increase the time parameters.

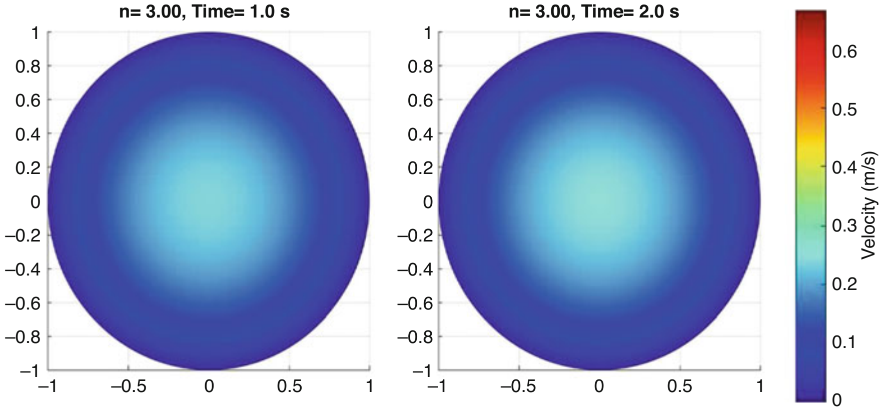


Fig. 5 Three-dimensional velocity field (43) where the volume flow rate is obtained by (40) with $\hat{G}(\hat{t}) = 1$, $\hat{Q}(0) = 0.1$, $\mathcal{W}_o = 1.5$, and $n = 3$ (shear-thickening fluid). Time parameters: $\hat{t} = 1, \hat{t} = 2$

3.2 Nonconstant Mean Pressure Gradient

Let us consider the nonconstant mean pressure gradient function, given by

$$\hat{G}(\hat{t}) = 1 + \frac{\sin^2(\hat{t})}{e^{\hat{t}}}, \tag{45}$$

which shows an interesting behavior (see Fig. 6). More specifically, it shows a strong variation in the initial stage and after the initial transient phase has small fluctuations, which tend to decrease with time. In Fig. 7, we can observe the behavior of the unsteady volume flow rate solution given by (40) obtained using a Runge-Kutta method with nonconstant mean pressure gradient (45), when we increase the Womersley number $\mathcal{W}_o = (0.5; 1.5; 3)$. In the initial phase of transition, we can verify the variation of the volume flow rate with the increase of the Womersley number, but with time the volume flow rate tends to stabilize regardless of the period of variation of the nondimensional parameter.

Finally, with the information of the volume flow rate given by (40), obtained for certain flow regimes with nonconstant pressure gradient (45), we can return to the three-dimensional problem to obtain the behavior of the three-dimensional velocity field (43) in time on the circular cross section of the tube (see Figs. 8 and 9).

Fig. 6 Nonconstant mean pressure gradient given by Eq. (45)

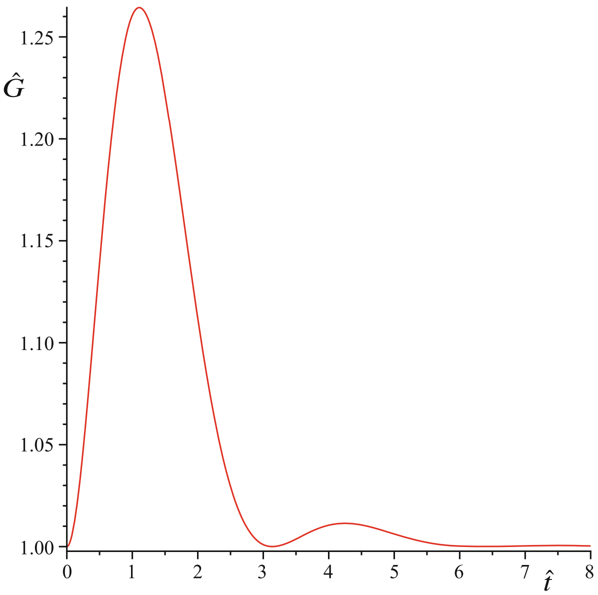
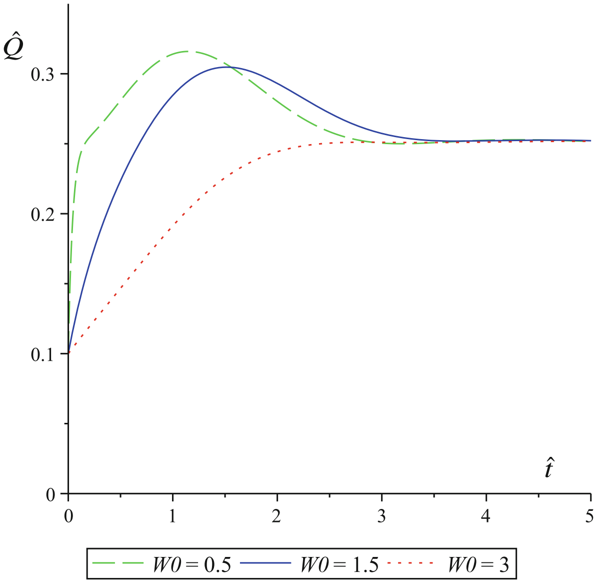


Fig. 7 Unsteady volume flow rate given by Eq. (40) with nonconstant mean pressure gradient (45) where $\hat{Q}(0) = 0.1$ and $\mathcal{W}_o = (0.5; 1.5; 3)$ for shear-thickening fluids with $n = 3$



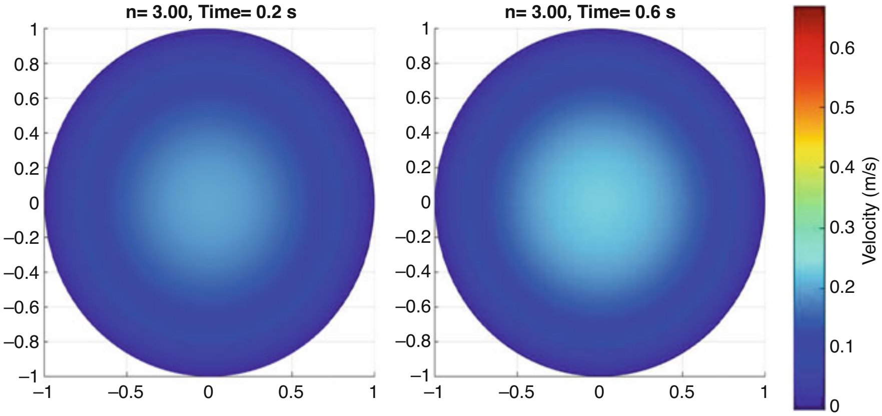


Fig. 8 Three-dimensional velocity field (43) where the volume flow rate is obtained by (40) with nonconstant mean pressure gradient (45), $\hat{Q}(0) = 0.1$, $\mathcal{W}_o = 0.5$, and $n = 3$ (shear-thickening fluid). Time parameters: $\hat{t} = 0.2$, $\hat{t} = 0.6$

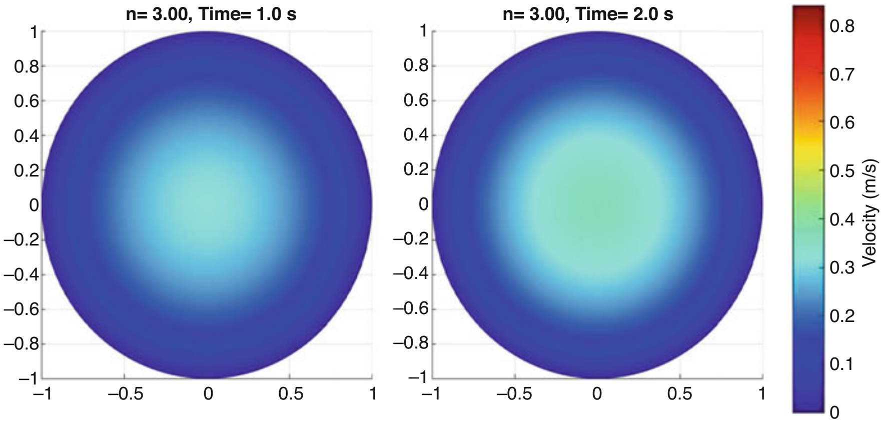


Fig. 9 Three-dimensional velocity field (43) where the volume flow rate is obtained by (40) with nonconstant mean pressure gradient (45), $\hat{Q}(0) = 0.1$, $\mathcal{W}_o = 0.5$, and $n = 3$ (shear-thickening fluid). Time parameters: $\hat{t} = 1$, $\hat{t} = 2$

4 Conclusions

Based on the works [10, 21, 22, 27, 29], we are facing a one-dimensional theory relevant to the study of physical problems involving the flow of Newtonian and non-Newtonian fluids under different geometries and perspectives, being a valid alternative to the classics one-dimensional models. The nature of Eq. (40) shows us in general the difficulty and the challenge of studying the flow of an incompressible fluid where the viscosity varies with the shear rate. In this work, based on a one-dimensional model obtained by using the Cosserat theory, we studied the behavior of the unsteady volume flow rate and the unsteady three-dimensional velocity field of an incompressible fluid where the viscosity function was given by Eq. (4), i.e., by cross-model viscosity function. Our one-dimensional approach is difficult to implement for any power index n, the difficulty being associated with computational problems due to the singularities presented in the integral calculus caused by constitutive equation $(8)_3$. In this sense, it was not possible to obtain a general equation for the mean pressure gradient involving the volume flow rate, Womersley number, power index n, and viscosity parameters. Based on the computational work and considering $n = 3$, we obtain specific ordinary differential equation to the mean pressure gradient involving the volume flow rate, Womersley number, and viscosity parameters. Using a Runge-Kutta method to solve the ordinary differential equation, we present the behavior of the unsteady volume flow rate by fixing the mean pressure gradient for specific flow regimes. Furthermore, we illustrate the three-dimensional velocity field behavior related to the model (8)–(10). Future work related to the Cosserat theory, which we are currently under study, include fluid-structure interaction, curved tubes, and the case of tubes with branches or bifurcations.

Acknowledgments This work has been partially supported by Centro de Investigação em Matemática e Aplicações (CIMA), through the grant UIDB/04674/2020 of FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

1. S.L. Rosen, *Fundamental Principles of Polymeric Materials*, 2nd edn. (Wiley, Hoboken, 1993)
2. B.R. Bird, R.C. Armstrong, O. Hassager, *Dynamics of Polymeric Liquids*, vol. 1, 2nd edn. (Wiley, Hoboken, 1987)
3. N.P. Cheremisinoff, *Rheology and Non-Newtonian Flows*. Encyclopedia of Fluid Mechanics, ed. by N.P. Cheremisinoff , vol. 7 (Springer, Berlin, 1986)
4. E. Marušić-Paloka, Steady flow of a non-newtonian fluid in unbounded channels and pipes. Math. Models Methods Appl. Sci. **10**(9), 1425–1445 (2000)
5. Y.I. Cho, K.R. Kensey, Effects of non-newtonian viscosity of blood on flows in a diseased arterial vessel, part 1: steady flows. Biorheology **28**, 41–262 (1991)
6. K.K. Yeleswarapu, Evaluation of continuum models for characterizing the constitutive behavior of blood. Ph.D. Thesis, University of Pittsburgh, 1996
7. T. Huges, J. Lubliner, On the one-dimensional theory of blood flow in the larger vessels, Math. Biosci. **18**, 161–170 (1973)
8. S.J. Sherwin, V. Franke, J. Peiró, K. Parker, One-dimensional modelling of a vascular network in space-time variables. J. Eng. Math. **47**, 217–250 (2003)
9. L. Formaggia, D. Lampont, A. Quarteroni, One-dimensional models for blood flow in arteries. J. Eng. Math. **47**, 251–276 (2003)
10. D. Caulk, P.M. Naghdi, Axisymmetric motion of viscous fluid flow inside a slender surface of revolution. J. Appl. Mech. **54**, 190–196 (1987)
11. P. Duhem, Le potentiel thermodynamique et la pression hydrostatique. Ann. École Norm **10**, 187–230 (1893)
12. E. Cosserat, F. Cosserat, Sur la théorie des corps minces. Compt. Rend. **146**, 169–172 (1908)

13. J.L. Ericksen, C. Truesdell, Exact theory of stress and strain in rods and shells. Arch. Rat. Mech. Anal. **1**(1), 295–323 (1958)
14. C. Truesdell, R. Toupin, *The Classical Field Theories of Mechanics*, ed. by Handbuch der Physik III (Springer, Berlin, 1960), pp. 226–793
15. A.E. Green, N. Laws, P.M. Naghdi, Rods, plates and shells. Proc. Camb. Phil. Soc. **64**(1), 895–913 (1968)
16. A.E. Green, P.M. Naghdi, M.L. Wenner, On the theory of rods II. Developments by direct approach. Proc. R. Soc. Lond. A **337**(1), 485–507 (1974)
17. P.M. Naghdi, The theory of shells and plates, in *Flügg's Handbuch der Physik*, vol. VIa/2, edn. (Springer, Berlin, 1972), pp. 425–640
18. A.E. Green, P.M. Naghdi, A direct theory of viscous fluid flow in channels. Arch. Ration. Mech. Analy. **86**, 39–63 (1984)
19. A.E. Green, P.M. Naghdi, A direct theory of viscous fluid flow in pipes I: basic general developments. Phil. Trans. R. Soc. Lond. A **342**(1), 525–542 (1993)
20. A.E. Green, P.M. Naghdi, A direct theory of viscous fluid flow in pipes: II flow of incompressible viscous fluid in curved pipes. Phil. Trans. R. Soc. Lond. A **342**(1), 543–572 (1993)
21. A.M. Robertson, A. Sequeira, A director theory approach for modeling blood flow in the arterial system: an alternative to classical 1D models. Math. Models Methods Appl. Sci. **15**(6), 871–906 (2005)
22. F. Carapau, A. Sequeira, 1D models for blood flow in small vessels using the cosserat theory. WSEAS Trans. Math. **5**(1), 54–62 (2006)
23. F. Carapau, Axisymmetric swirling motion of viscoelastic fluid flow inside a slender surface of revolution. IAENG Eng. Lett. **17**(4), 238–245 (2009)
24. F. Carapau, J. Janela, A one-dimensional model for unsteady axisymmetric swirling motion of a viscous fluid in a variable radius straight circular tube. Int. J. Eng. Sci. **72**, 107–116 (2013)
25. F. Carapau, J. Janela, P. Correia, S. Vila, Numerical solvability of a cosserat model for the swirling motion of a third-grade fluid in a constant radius straight circular tube. Int. J. Appl. Math. Stat. **57**(2), 1–15 (2018)
26. F. Carapau, One-dimensional viscoelastic fluid model where viscosity and normal stress coefficients depend on the shear rate. Nonlinear Analy. Real World Appl. **11**, 4342–4354 (2010)
27. F. Carapau, 1D viscoelastic flow in a circular straight tube with variable radius. Int. J. Appl. Math. Stat., No. D10 **19**, 20–39 (2010)
28. F. Carapau, P. Correia, Numerical simulations of a third-grade fluid flow on a tube through a contraction. Eur. J. Mech. B/Fluids **65**, 45–53 (2017)
29. F. Carapau, A. Sequeira, Axisymmetric motion of a second order viscous fluid in a circular straight tube under pressure gradients varying exponentially with time. WIT Trans. Eng. Sci. **52**, 409–419 (2006)
30. F. Carapau, P. Correia, T. Rabczuk, P. Areias, One-dimensional model for the unsteady flow of a generalized third-grade viscoelastic fluid. Neural Comput. Appl. **32**(16), 12881–12894 (2020)

Small Forced Oscillation of a Rigid Body in a Viscous Liquid

Giovanni P. Galdi

Ὅν οί ϑεοὶ φιλοῦσιν, ὰποϑνήσϖει νέος.
He whom the gods love dies young
MENANDER 342-291 BC

1 Introduction

Viscous flow around oscillating bodies is a problem of significant relevance in fluid mechanics, e.g., [1, 10] and the literature there cited. As a matter of fact, its range of application is rather wide, since it may be of interest at different scales: from microfluidics [8] to design of marine and land vehicles [1] and from bubble dynamics [9] to stability of structures [3]. Concerning the latter, of particular importance is the phenomenon of forced oscillation of suspension bridges, induced by the vortex shedding of the fluid (air), which reflects into an oscillatory regime of the wake. When the frequency of the wake approaches the natural structural frequency of the body (the bridge), a resonant phenomenon may occur that could lead to structural failure. An infamous example of this phenomenon is the well-known collapse of the Tacoma Narrows Bridge.

As a first step toward furnishing a rigorous mathematical analysis of this phenomenon, in the joint work [2], we have started to analyze the simple model problem where a two-dimensional rectangular structure is subject to a unidirectional restoring elastic force and immersed in the two-dimensional channel flow of

G. P. Galdi (✉)
Department of Mechanical Engineering and Materials Sciences, University of Pittsburgh,
Pittsburgh, PA, USA
e-mail: galdi@pitt.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_3

a Navier-Stokes liquid driven by a time-independent Poiseuille flow. The main objective in [2] is to investigate the existence of possible equilibrium configurations of the structure, at least for "small" data. In [2], an analogous question was also addressed when the rectangle is allowed to rotate around its center and subject to a restoring elastic torque.

The natural, second step to be undertaken is then to suppose that, more generally, the driving mechanism is time periodic of prescribed period T ("T-periodic") and study the corresponding forced oscillation of the structure with a view to possible occurrence of resonance. The main goal of this paper is to provide an introductory contribution in that direction.

More precisely, let $\mathscr{S}$ be a rigid body subject to an elastic restoring force **R** that we assume to be applied to its center of mass, G. $\mathscr{S}$ is in the flow of Navier-Stokes liquid driven by a T-periodic, uniform velocity $\boldsymbol{U} = \boldsymbol{U}(t)$ at "large" distance from $\mathscr{S}$. To avoid "boundary effects" that could be irrelevant to the study we have in mind, we suppose that the liquid fills the whole space, Ω, outside $\mathscr{S}$. Moreover, we assume that on $\mathscr{S}$ a torque is applied that prevents it from rotating. Under these conditions, the system of equations governing the motion of the coupled system body liquid, in a frame $\mathcal{F} \equiv \{G, \boldsymbol{e}_i\}$ attached to $\mathscr{S}$, is given by [4]

$$\begin{gathered}\left.\begin{aligned}&\partial_t\boldsymbol{\vartheta} + (\boldsymbol{\vartheta} - \boldsymbol{\chi})\cdot\nabla\boldsymbol{\vartheta} = \nu\Delta\boldsymbol{\vartheta} - \nabla p\\ &\operatorname{div}\boldsymbol{\vartheta} = 0\end{aligned}\right\}\ \text{in}\ \Omega\times\mathbb{R}\,,\\ \boldsymbol{\vartheta}(x,t) = \boldsymbol{\chi}(t)\,,\ (x,t)\in\partial\Omega\times\mathbb{R}\,;\ \ \lim_{|x|\to\infty}\boldsymbol{\vartheta}(x,t) = \boldsymbol{U}(t)\,,\ t\in\mathbb{R}\,,\\ M\dot{\boldsymbol{\chi}} + \rho\int_{\partial\Omega}\mathbb{T}(\boldsymbol{\vartheta},p)\cdot\boldsymbol{n} = \boldsymbol{R}\ \text{ in}\ \mathbb{R}\,.\end{gathered}\tag{1}$$

Here, $\boldsymbol{\vartheta}$ and $\rho\, p$ are velocity and pressure fields of the liquid and ρ and ν its density and kinematic viscosity, while M and $\boldsymbol{\chi} = \boldsymbol{\chi}(t)$ are mass of $\mathscr{S}$ and velocity of G, respectively. Furthermore,

$$\mathbb{T}(z,\psi) := 2\nu\,\mathbb{D}(z) - \psi\,\mathbb{I}\,,\quad \mathbb{D}(z) := \frac{1}{2}\left(\nabla z + (\nabla z)^{\top}\right)\,,$$

with $\mathbb{I}$ identity matrix, is the Cauchy stress tensor, and, finally, $\boldsymbol{n}$ is the unit outer normal at $\partial\Omega$.

The relevant question we want to investigate is whether (1) admits the existence of T-periodic solutions and, more importantly, whether or not such existence holds for *arbitrary* value of T (> 0). The response to the latter will provide the information on whether resonance does or does not occur.

In this paper, we will consider only "small oscillations" of $\mathscr{S}$, which translates into the assumption of creeping flow for the liquid (Stokes approximation). Further, we take $\boldsymbol{R}$ to be a linear function of the displacement $\boldsymbol{\xi} := \int \boldsymbol{\chi}(s)\mathrm{d}s$ with respect to a fixed point, namely,

$$\boldsymbol{R} = -\ell\,\boldsymbol{\xi}\,,\ \ \ell\in\mathbb{R}_+\,.$$

As a consequence, (1) becomes

$$\begin{aligned} &\left.\begin{aligned}\partial_t \boldsymbol{\vartheta} &= \nu\Delta\boldsymbol{\vartheta} - \nabla p \\ \operatorname{div}\boldsymbol{\vartheta} &= 0\end{aligned}\right\} \text{ in } \Omega\times\mathbb{R}\,, \\ &\boldsymbol{\vartheta}(x,t) = \dot{\boldsymbol{\xi}}(t)\,,\ (x,t)\in\partial\Omega\times\mathbb{R}\,;\quad \lim_{|x|\to\infty}\boldsymbol{\vartheta}(x,t) = \boldsymbol{U}(t)\,,\ t\in\mathbb{R}\,, \\ &\ddot{\boldsymbol{\xi}} + \omega_0^2\,\boldsymbol{\xi} + \varpi\int_{\partial\Omega}\mathbb{T}(\boldsymbol{\vartheta},p)\cdot\boldsymbol{n} = \mathbf{0} \text{ in } \mathbb{R}\,, \end{aligned} \tag{2}$$

where

$$\omega_0^2 := \frac{\ell}{M}\,,\quad \varpi := \frac{\rho}{M}\,.$$

Our primary objective is then to investigate whether, under suitable regularity assumption on $\boldsymbol{U}$, problem (2) has one (and only one) T-periodic solution in a suitable function class. The main result, formulated in Theorem 1 in the following section, states that, provided $\boldsymbol{U}$ is T-periodic and smooth enough, (2) is uniquely solvable for T-periodic $(\boldsymbol{\vartheta}, p, \boldsymbol{\xi})$ in a suitable function class, *whatever* $T > 0$. Thus, in particular, resonance is ruled out, at least in the creeping flow approximation. Even though well known, it is worth emphasizing that in the absence of liquid, such a result is not true. The method we use relies on a combination of the ideas developed in [7] and [6]. The crucial point in the proof of the theorem is to establish a uniform bound of the generic Fourier mode of the displacement (the "amplitude" of the oscillation) in terms of the data, namely, $\boldsymbol{U}$ (see (34)). As expected, such a bound is lost in the limit $\nu \to 0$. As it becomes clear from the proof, a result entirely analogous to that given in Theorem 1 can be obtained if, in addition, a T-periodic (smooth enough) force is acting on $\mathscr{S}$.

We believe that our approach could be extended to study the full nonlinear problem (1), as well as applied to investigate the fundamental question of vortex-induced oscillation of the body (Hopf bifurcation), which, in fact, is the original motivation of our work. These investigations will be the object of future research.

The plan of the paper is as follows: After recalling some known results in Sect. 2, in the following Sect. 3, we reformulate the problem in terms of its averaged (over a period) and oscillatory components (see (7)–(8)) and then give the proof of our main finding in Theorem 1.

2 Preliminary Results

We begin to recall some basic notation. By Ω, we indicate a domain of $\mathbb{R}^3$, complement of the closure of a bounded domain Ω_0 of class C^2. As customary, $L^q = L^q(\Omega)$ is the Lebesgue space with norm $\|\cdot\|_q$, and $W^{m,2} = W^{m,2}(\Omega)$ denotes Sobolev space, $m \in \mathbb{N}$, with norm $\|\cdot\|_{m,2}$. Furthermore, $D^{m,2} = D^{m,2}(\Omega)$ are homogeneous Sobolev spaces with seminorm $|u|_{m,2} := \sum_{|l|=m} \|D^l u\|_2$, whereas $D_0^{1,2} = D_0^{1,2}(\Omega)$ is the completion of $C_0^\infty(\Omega)$ in the norm $|\cdot|_{1,2}$.

A function $u : \Omega \times \mathbb{R} \mapsto \mathbb{R}^3$ is T*-periodic*, $T > 0$, if $u(\cdot, t+T) = u(\cdot t)$, for a.a. $t \in \mathbb{R}$, and we set $\overline{u} := \frac{1}{T}\int_0^T u(t)\mathrm{d}t$. Let B be a function space endowed with seminorm $\|\cdot\|_B$ and $T > 0$. Then, $L^2(0,T;B)$ is the class of functions $u : (0,T) \to B$ such that

$$\|u\|_{L^2(B)} := \Big(\int_0^T \|u(t)\|_B^2\Big)^{\frac{1}{2}} < \infty.$$

Likewise, we put

$$W^{1,2}(0,T;B) = \Big\{u \in L^2(0,T;B) : \partial_t u \in L^2(0,T;B)\Big\}.$$

For simplicity, we write $L^2(B)$ for $L^2(0,T;B)$, etc. Moreover, we define the Banach spaces

$$\begin{aligned}
&L^2_\sharp := \{\boldsymbol{\xi} \in L^2(0,T),\ \boldsymbol{\xi} \text{ is } T\text{-periodic}\},\\
&W^k_\sharp := \{\boldsymbol{\xi} \in L^2_\sharp(0,T),\ d^l\boldsymbol{\xi}/dt^l \in L^2(0,T),\ \ l=1,\dots,k\},\\
&\mathcal{L}^2_\sharp := \{\boldsymbol{u} \in L^2(L^2);\ \boldsymbol{u} \text{ is } T\text{-periodic, with } \overline{\boldsymbol{u}} = \boldsymbol{0}\},\\
&\mathcal{W}^2_\sharp := \{\boldsymbol{u} \in W^{1,2}(L^2) \cap L^2(W^{2,2});\ \boldsymbol{u} \text{ is } T\text{-periodic, with } \overline{\boldsymbol{u}} = \boldsymbol{0}\},
\end{aligned}$$

along with norms

$$\begin{aligned}
&\|\boldsymbol{\xi}\|_{L^2_\sharp} := \|\boldsymbol{\xi}\|_{L^2(0,T)}, \quad \|\boldsymbol{\xi}\|_{W^k_\sharp} := \|\boldsymbol{\xi}\|_{W^{k,2}(0,T)},\\
&\|\boldsymbol{u}\|_{\mathcal{L}^2_\sharp} := \|\boldsymbol{u}\|_{L^2(L^2)}, \quad \|\boldsymbol{u}\|_{\mathcal{W}^2_\sharp} := \|\boldsymbol{u}\|_{W^{1,2}(L^2)} + \|\boldsymbol{u}\|_{L^2(W^{2,2})}.
\end{aligned}$$

Before addressing the resolution of our problem, we need to recall some known results. The following one is proved in [4, Lemma 4.9]:

Lemma 1 *Suppose* $\boldsymbol{u} \in L^6(\Omega) \cap D^{1,2}(\Omega)$, *with* $\operatorname{div}\boldsymbol{u} = 0$ *in* Ω *and* $\boldsymbol{u}|_{\partial\Omega} = \boldsymbol{u}_* \in \mathbb{R}^3$. *Then, there exists a numerical constant* c_0 *such that*

$$|\boldsymbol{u}_*| \le c_0|\Omega_0|^{-\frac{1}{6}}\|\mathbb{D}(\boldsymbol{u})\|_2\,.$$

The proof of the next lemma is given in [6, Lemma 5.1].

Lemma 2 *Consider the boundary-value problems, with* $i = 1,2,3$, $k \in \mathbb{Z}\backslash\{0\}$, *and* $\omega := 2\pi/T$:

$$\left.\begin{aligned}
&\mathrm{i}\,k\,\omega\,\boldsymbol{h}_k^{(i)} = \nu\Delta\boldsymbol{h}_k^{(i)} - \nabla\gamma_k^{(i)}\\
&\operatorname{div}\boldsymbol{h}_k^{(i)} = 0
\end{aligned}\right\}\ \text{in } \Omega\,,$$
$$\boldsymbol{h}_k^{(i)}|_{\partial\Omega} = \boldsymbol{e}_i\,, \tag{3}$$

The following properties hold:

(i) *There is one and only one solutions* $(\boldsymbol{h}_k^{(i)}, \gamma_k^{(i)}) \in W^{2,2}(\Omega) \times W^{1,2}(\Omega)$. *This solution satisfies the estimates*

$$\|\boldsymbol{h}_k^{(i)}\|_2 \le c\,; \quad \|\nabla \boldsymbol{h}_k^{(i)}\|_2 \le c\,|k|^{\frac{1}{2}}\,; \quad |\boldsymbol{h}_k^{(i)}|_{2,2} \le c\,|k|\,, \tag{4}$$

where $c = c(\omega, \nu) > 0$.

(ii) *The matrix* $\mathbb{B}$ *defined by components*[1]

$$(\mathbb{B})_{\ell i} = \int_{\partial\Omega} \mathbb{T}_{\ell j}(\boldsymbol{h}_k^{(i)}, \gamma_k^{(i)}) n_j$$

satisfies the condition (with $^* \equiv$ *c.c.)*

$$\boldsymbol{\zeta}^* \cdot \mathbb{B} \cdot \boldsymbol{\zeta} = \mathrm{i}\, k\, \omega\, \|\zeta_i \boldsymbol{h}^{(i)}\|_2^2 + 2\nu \|\mathbb{D}(\zeta_i \boldsymbol{h}^{(i)})\|_2^2\,, \tag{5}$$

for all $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3) \in \mathbb{C}^3$.

3 Main Result

This section is entirely devoted to the proof of the main contribution of this paper, stated later on in Theorem 1. To reach this goal, we begin to rewrite (2) in terms of its averaged and oscillatory components. Thus, assuming T-periodicity and setting

$$\boldsymbol{\vartheta}(x,t) = \overline{\boldsymbol{\vartheta}}(x) + \mathbf{v}(x,t),\ p(x,t) = \overline{p}(x) + \mathsf{p}(x,t),\ \boldsymbol{\xi} = \overline{\boldsymbol{\xi}} + \boldsymbol{\sigma}(t),\ \boldsymbol{U}(t) = \overline{\boldsymbol{U}} + \mathbf{U}(t) \tag{6}$$

problem (2) can be formally split into the following two problems:

$$\begin{gathered}
\left.\begin{aligned} \nu\Delta\overline{\boldsymbol{\vartheta}} &= \nabla\overline{p} \\ \operatorname{div}\overline{\boldsymbol{\vartheta}} &= 0 \end{aligned}\right\} \text{ in } \Omega\,, \\
\overline{\boldsymbol{\vartheta}}(x) = \mathbf{0}\,,\ x \in \partial\Omega\,, \quad \lim_{|x|\to\infty} \overline{\boldsymbol{\vartheta}}(x) = \overline{\boldsymbol{U}} \\
\omega_0^2\, \overline{\boldsymbol{\xi}} + \varpi \int_{\partial\Omega} \mathbb{T}(\overline{\boldsymbol{\vartheta}}, \overline{p}) \cdot \boldsymbol{n} = \mathbf{0}\,,
\end{gathered} \tag{7}$$

and

$$\begin{gathered}
\left.\begin{aligned} \partial_t \mathbf{v} &= \nu\Delta\mathbf{v} - \nabla\mathsf{p} \\ \operatorname{div}\mathbf{v} &= 0 \end{aligned}\right\} \text{ in } \Omega\times\mathbb{R}\,, \\
\mathbf{v}(x,t) = \dot{\boldsymbol{\sigma}}(t)\,,\ (x,t) \in \partial\Omega\times\mathbb{R}\,, \quad \lim_{|x|\to\infty} \mathbf{v}(x,t) = \mathbf{U}(t)\,,\ t \in \mathbb{R}\,, \\
\ddot{\boldsymbol{\sigma}} + \omega_0^2\, \boldsymbol{\sigma} + \varpi \int_{\partial\Omega} \mathbb{T}(\mathbf{v}, \mathsf{p}) \cdot \boldsymbol{n} = \mathbf{0} \text{ in } \mathbb{R}\,.
\end{gathered} \tag{8}$$

[1] Unless otherwise stated, we assume summation over repeated indices.

We next perform the lift of the vector U as follows: Set

$$\boldsymbol{W}(t) := x_3\mathsf{U}_2(t)\boldsymbol{e}_1 + x_1\mathsf{U}_3(t)\boldsymbol{e}_2 + x_2\mathsf{U}_1(t)\boldsymbol{e}_3\,. \tag{9}$$

Clearly,

$$\operatorname{curl} \boldsymbol{W} = \mathsf{U}(t)\,. \tag{10}$$

Let $\phi(x)$ be a smooth cutoff function that is 1 for $|x| \geq 2R$ and 0 for $|x| \leq R$, R sufficiently large, and define

$$\boldsymbol{w}(x,t) := \operatorname{curl}\big(\phi(x)\boldsymbol{W}(t)\big)\,.$$

In view of (10), we deduce

$$\boldsymbol{w}(x,t) = \phi(x)\mathsf{U}(t) - \boldsymbol{W} \times \nabla\phi(x) \tag{11}$$

so that $\boldsymbol{w}$ is a T-periodic solenoidal vector function that is equal to $\mathsf{U}(t)$ for $|x| \geq 2R$ and equal to 0 for $|x| \leq R$. Therefore, introducing the field

$$\boldsymbol{u}(x,t) := \mathsf{v}(x,t) - \boldsymbol{w}(x,t)\,, \tag{12}$$

we deduce that (8) is equivalent to

$$\begin{gathered}
\left.\begin{aligned}
&\partial_t \boldsymbol{u} = \nu\Delta\boldsymbol{u} - \nabla\mathsf{p} + \boldsymbol{f}\\
&\operatorname{div}\boldsymbol{u} = 0
\end{aligned}\right\}\ \text{in}\ \Omega\times\mathbb{R}\,,\\
\boldsymbol{u}(x,t) = \dot{\boldsymbol{\sigma}}(t)\,,\ (x,t)\in\partial\Omega\times\mathbb{R}\,;\quad \lim_{|x|\to\infty}\boldsymbol{u}(x,t) = \boldsymbol{0}\,,\ t\in\mathbb{R}\,,\\
\ddot{\boldsymbol{\sigma}} + \omega_0^2\boldsymbol{\sigma} + \varpi\int_{\partial\Omega}\mathbb{T}(\boldsymbol{u},\mathsf{p})\cdot\boldsymbol{n} = \varpi\,\boldsymbol{F}\ \text{ in}\ \mathbb{R}\,,
\end{gathered} \tag{13}$$

where

$$\boldsymbol{f} = \boldsymbol{f}(x,t) := -\partial_t\boldsymbol{w} + \nu\Delta\boldsymbol{w}\,,\quad \boldsymbol{F} = \boldsymbol{F}(t) := -\int_{\partial\Omega}\mathbb{T}(\boldsymbol{w},0)\cdot\boldsymbol{n}\,. \tag{14}$$

Notice that by (9) and (11), we obtain, on the one hand,

$$\overline{\boldsymbol{f}} \equiv \overline{\boldsymbol{F}} \equiv \boldsymbol{0}\,. \tag{15}$$

and, on the other hand,

$$\|\boldsymbol{f}\|_{\mathcal{L}^2_\sharp} + \|\boldsymbol{F}\|_{L^2_\sharp} \leq C\,\|\mathsf{U}\|_{W^1_\sharp}\,, \tag{16}$$

with $C = C(\Omega,\nu) > 0$. We are now in a position to prove the following theorem:

Theorem 1 *Let $T > 0$ arbitrary, and let $\boldsymbol{U} := (\overline{\boldsymbol{U}} + \boldsymbol{\mathsf{U}}) \in W^1_\sharp$. Then, problem (2) has one and only one solution $(\boldsymbol{\vartheta}, p, \boldsymbol{\xi})$ such that*

$$\begin{aligned}&(\overline{\boldsymbol{\vartheta}} - \overline{\boldsymbol{U}}, \overline{p}) \in [L^6(\Omega) \cap D^{1,2}_0(\Omega) \cap D^{2,2}(\Omega)] \times W^{1,2}(\Omega)\,,\\ &\big(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}} - \boldsymbol{\mathsf{U}}, p - \overline{p}\big) \in \mathcal{W}^2_\sharp \times L^2(D^{1,2})\,,\ \ \boldsymbol{\xi} \in W^2_\sharp\,.\end{aligned} \tag{17}$$

Furthermore, this solution satisfies the estimates

$$\begin{aligned}&\|\overline{\boldsymbol{\vartheta}} - \overline{\boldsymbol{U}}\|_6 + |\nabla\overline{\boldsymbol{\vartheta}}\|_{1,2} + |\overline{\boldsymbol{\xi}}| + \|\overline{p}\|_{1,2} \le C_1\,|\overline{\boldsymbol{U}}|\\ &\|\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}} - \boldsymbol{\mathsf{U}}\|_{\mathcal{W}^2_\sharp} + \|\boldsymbol{\xi}\|_{W^2_\sharp} + \|p - \overline{p}\|_{L^2(D^{1,2})} \le C_2\|\boldsymbol{\mathsf{U}}\|_{W^2_\sharp}\,,\end{aligned} \tag{18}$$

where $C_i = C_i(\Omega, \nu, T, \omega_0, \varpi) > 0, i = 1, 2.$

Proof Throughout the proof, by c_i, $i = 1, 2, \ldots, C$, we denote positive constants that, at most, may have a similar parameter dependence as the constants C_i defined above. From classical results on the Stokes problem [5, TheoremV.5.3 and IV.5.1], we infer that, for any $\overline{\boldsymbol{U}} \in \mathbb{R}^3$, there exists a unique solution $(\overline{\boldsymbol{\vartheta}}, \overline{p})$ to $(7)_{1,2,3}$ in the class specified by $(17)_1$ that, in addition, obeys the estimate

$$\|\overline{\boldsymbol{\vartheta}} - \overline{\boldsymbol{U}}\|_6 + \|\nabla\overline{\boldsymbol{\vartheta}}\|_{1,2} + \|\overline{p}\|_{1,2} \le c_1\,|\overline{\boldsymbol{U}}|\,. \tag{19}$$

Moreover, by [5, Theorem II.9.1], we infer that $\overline{\boldsymbol{\vartheta}} - \overline{\boldsymbol{U}}$ obeys $(7)_4$ as well. Finally, by well-known trace theorems, we show that

$$\left|\int_{\partial\Omega} \mathbb{T}(\overline{\boldsymbol{\vartheta}}, \overline{p}) \cdot \boldsymbol{n}\right| \le c_2\,\big(\|\nabla\overline{\boldsymbol{\vartheta}}\|_{1,2} + \|\overline{p}\|_{1,2}\big)$$

so from the latter and (19), we may choose $\overline{\boldsymbol{\xi}}$ as in $(7)_4$ and deduce $(18)_1$. We now pass to the resolution of problem (13). To this end, we set

$$\boldsymbol{u} := \boldsymbol{\mathsf{z}} + \boldsymbol{\mathsf{w}}\,,\ \ \mathsf{p} := \tau + \mathsf{q} \tag{20}$$

where $\boldsymbol{\mathsf{z}}$ and $\boldsymbol{\mathsf{w}}$ satisfy the following set of equations:

$$\left.\begin{aligned}&\partial_t\boldsymbol{\mathsf{z}} - \nu\Delta\boldsymbol{\mathsf{z}} = -\nabla\tau + \boldsymbol{f}\\ &\operatorname{div}\boldsymbol{\mathsf{z}} = 0\end{aligned}\right\}\ \text{in } \Omega \times \mathbb{R}$$
$$\boldsymbol{\mathsf{z}}|_{\partial\Omega} = \mathbf{0} \tag{21}$$

and

$$\left.\begin{aligned}&\partial_t\boldsymbol{\mathsf{w}} - \nu\Delta\boldsymbol{\mathsf{w}} = -\nabla\mathsf{q}\\ &\operatorname{div}\boldsymbol{\mathsf{w}} = 0\end{aligned}\right\}\ \text{in } \Omega \times \mathbb{R}$$
$$\boldsymbol{\mathsf{w}}|_{\partial\Omega} = \dot{\boldsymbol{\sigma}}\,; \tag{22}$$
$$\ddot{\boldsymbol{\sigma}} + \omega_0^2\,\boldsymbol{\sigma} + \varpi\int_{\partial\Omega} \mathbb{T}(\boldsymbol{\mathsf{w}}, \mathsf{q}) \cdot \boldsymbol{n} = \varpi\,\boldsymbol{F} - \varpi\int_{\partial\Omega} \mathbb{T}(\boldsymbol{\mathsf{z}}, \tau) \cdot \boldsymbol{n} := \varpi\,\mathcal{F}\,.$$

Since $\boldsymbol{f}$ satisfies (15) and (16), by [7], it follows that there exists a unique solution $(\mathsf{z}, \tau) \in \mathcal{W}^2_\sharp \times L^2(D^{1,2})$ that, in addition, obeys the inequality

$$\|\mathsf{z}\|_{\mathcal{W}^2_\sharp} + \|\tau\|_{L^2(D^{1,2})} \leq c_3\, \|\boldsymbol{f}\|_{\mathcal{L}^2_\sharp} \leq C\, \|\mathsf{U}\|_{W^1_\sharp}\,. \tag{23}$$

Since, by trace theorem,

$$\|\int_{\partial\Omega} \mathbb{T}(\mathsf{z}, \tau) \cdot \boldsymbol{n}\|_{L^2_\sharp} \leq c\, \left(\|\mathsf{z}\|_{\mathcal{W}^2_\sharp} + \|\tau\|_{L^2(D^{1,2})}\right)\,, \tag{24}$$

by assumption, the properties of (z, τ), (15), (16), and (23) we may infer that the function $\mathcal{F}$ in (22) is in $L^2_\sharp$, with $\overline{\mathcal{F}} = \mathbf{0}$ and that, in addition,

$$\|\mathcal{F}\|_{L^2_\sharp} \leq C\; \|\mathsf{U}\|_{W^1_\sharp}\,. \tag{25}$$

Thus, in order to find solutions to (22), we formally expand w, q, and $\boldsymbol{\sigma}$, in Fourier series as follows:

$$\mathsf{w}(x,t) = \sum_{k\in\mathbb{Z}} \mathsf{w}_k(x)\, \mathrm{e}^{\mathrm{i}k\,\omega\,t}, \quad \mathsf{q}(x,t) = \sum_{k\in\mathbb{Z}} \mathsf{q}_k(x)\, \mathrm{e}^{\mathrm{i}k\,\omega\,t}, \tag{26}$$

$$\boldsymbol{\sigma}(t) = \sum_{k\in\mathbb{Z}} \boldsymbol{\sigma}_k\, \mathrm{e}^{\mathrm{i}k\,\omega\,t}, \quad \mathsf{w}_0 \equiv \nabla\mathsf{q}_0 \equiv \boldsymbol{\sigma}_0 \equiv \mathbf{0},$$

where $(\mathsf{w}_k, \mathsf{q}_k, \boldsymbol{\sigma}_k)$ solve the problem $(k \neq 0)$

$$\left.\begin{aligned} \mathrm{i}\,k\,\omega\,\mathsf{w}_k &= \nu\Delta\mathsf{w}_k - \nabla\mathsf{q}_k \\ \operatorname{div}\mathsf{w}_k &= 0 \end{aligned}\right\} \text{ in } \Omega \qquad \mathsf{w}_k|_{\partial\Omega} = \mathrm{i}k\boldsymbol{\sigma}_k\,, \tag{27}$$

with the further condition

$$\left(-k^2\,\omega^2 + \omega_0^2\right)\boldsymbol{\sigma}_k + \varpi \int_{\partial\Omega} \mathbb{T}(\mathsf{w}_k, \mathsf{q}_k) \cdot \boldsymbol{n} = \varpi\mathcal{F}_k\,, \tag{28}$$

where $\{\mathcal{F}_k\}$ are Fourier coefficients of $\mathcal{F}$ with $\mathcal{F}_0 \equiv \mathbf{0}$. For each fixed $k \in \mathbb{Z}\backslash\{0\}$, a solution to (27)–(28) is given by[2]

$$\mathsf{w}_k = \sum_{i=1}^{3} \mathrm{i}\,k\,\boldsymbol{\sigma}_{ki}\boldsymbol{h}_k^{(i)}\,, \quad \mathsf{q}_k = \sum_{i=1}^{3} \mathrm{i}\,k\,\boldsymbol{\sigma}_{ki}\gamma_k^{(i)}\,, \tag{29}$$

[2] No summation over k.

with $(\boldsymbol{h}_k^{(i)}, \gamma_k^{(i)})$ given in Lemma 2, and where $\boldsymbol{\sigma}_k$ solve the equations

$$\left(-k^2\,\omega^2+\omega_0^2\right)\boldsymbol{\sigma}_k+\sum_{i=1}^{3}\mathrm{i}\,k\,\varpi\,\sigma_{ki}\int_{\partial\Omega}\mathbb{T}(\boldsymbol{h}_k^{(i)},\gamma_k^{(i)})\cdot\boldsymbol{n}=\varpi\,\mathcal{F}_k\,, \tag{30}$$

which, with the notation of Lemma 2(ii), can be equivalently rewritten as

$$\mathbb{M}\cdot\boldsymbol{\sigma}_k=\varpi\,\mathcal{F}_k\,,\quad \mathbb{M}:=(-k^2\,\omega^2+\omega_0^2)\mathbb{I}+\mathrm{i}\,k\,\varpi\,\mathbb{B}\,. \tag{31}$$

The matrix $\mathbb{M}$ is invertible for all $k\neq 0$. In fact, using (5), for all $\boldsymbol{\zeta}\in\mathbb{C}^3$, we show

$$\boldsymbol{\zeta}^*\cdot\mathbb{M}\cdot\boldsymbol{\zeta}=(-k^2\,\omega^2+\omega_0^2)\,|\boldsymbol{\zeta}|^2-k^2\omega\,\varpi\,\|\zeta_i\boldsymbol{h}_k^{(i)}\|_2^2+\mathrm{i}\,k\,\varpi\,\nu\|\mathbb{D}(\zeta_i\boldsymbol{h}_k^{(i)})\|_2^2\,.$$

Thus, assuming $\mathbb{M}\cdot\boldsymbol{\zeta}=\boldsymbol{0}$, it follows

$$\mathbb{D}(\zeta_i\boldsymbol{h}_k^{(i)})\equiv 0\,. \tag{32}$$

However, by the properties of $\boldsymbol{h}_k^{(i)}$, we obtain that $\zeta_i\boldsymbol{h}_k^{(i)}|_{\partial\Omega}=\boldsymbol{\zeta}$, which by Lemma 1, the embedding $W^{1,2}(\Omega)\subset L^6(\Omega)$, and (32) implies $\boldsymbol{\zeta}=\boldsymbol{0}$, namely, 0 is not an eigenvalue of $\mathbb{M}$. As a result, for the given $\mathcal{F}_k$, (31) has one and only one solution $\boldsymbol{\sigma}_k$. If we now dot-multiply both sides of (31) by $\boldsymbol{\sigma}_k^*$ and use again (5), we deduce

$$(-k^2\,\omega^2+\omega_0^2)\,|\boldsymbol{\sigma}_k|^2-k^2\omega\,\varpi\,\|\sigma_{ki}\boldsymbol{h}_k^{(i)}\|_2^2+\mathrm{i}\,k\,\varpi\,\nu\|\mathbb{D}(\sigma_{ki}\boldsymbol{h}_k^{(i)})\|_2^2=\varpi\,(\mathcal{F}_k,\boldsymbol{\sigma}_k^*)\,,$$

which, in turn, furnishes

$$\begin{aligned} k\,\nu\|\mathbb{D}(\sigma_{ki}\boldsymbol{h}_k^{(i)})\|_2^2&=\Im[(\mathcal{F}_k,\boldsymbol{\sigma}_k^*)]\,,\\ (k^2\,\omega^2-\omega_0^2)\,|\boldsymbol{\sigma}_k|^2+k^2\omega\varpi\,\|\sigma_{ki}\boldsymbol{h}_k^{(i)}\|_2^2&=\varpi\,\Re[(\mathcal{F}_k,\boldsymbol{\sigma}_k^*)]\,. \end{aligned} \tag{33}$$

Recalling that $\sigma_{ki}\boldsymbol{h}_k^{(i)}|_{\partial\Omega}=\boldsymbol{\sigma}_k$, by $(33)_1$, Schwarz inequality, and Lemma 1, we show the crucial estimate

$$|\boldsymbol{\sigma}_k|\leq\frac{c_0}{\nu\,|\Omega_0|^{\frac{1}{6}}}|\mathcal{F}_k|\,,\quad |k|\geq 1\,. \tag{34}$$

Again by Schwarz inequality, from $(33)_2$, we get, in particular,

$$k^2\omega^2\,|\boldsymbol{\sigma}_k|\leq\varpi\,|\mathcal{F}_k|+\omega_0^2\,|\boldsymbol{\sigma}_k|\,,$$

and so combining the latter with (34), we conclude

$$k^2|\boldsymbol{\sigma}_k| \le \left(\varpi + \frac{c_0\,\omega_0^2}{\nu\,|\Omega_0|^{\frac{1}{6}}}\right)\frac{|\mathcal{F}_k|}{\omega^2} := C_0\,|\mathcal{F}_k|\,,\quad |k|\ge 1\,. \tag{35}$$

From (35), it immediately follows that

$$\|\boldsymbol{\sigma}\|^2_{W^2_\sharp} = \sum_{|k|\ge 1}(|k|^4+|k|^2+1)|\boldsymbol{\sigma}_k|^2 \le 3C_0^2\sum_{|k|\ge 1}|\mathcal{F}_k|^2 = 3C_0^2\|\mathcal{F}\|^2_{L^2_\sharp}\,. \tag{36}$$

Moreover, from (29), (36), and (4), we infer

$$\begin{aligned}\|\mathbf{w}\|^2_{\mathcal{W}^2_\sharp} &= \sum_{|k|\ge 1}\Big[(|k|^2+1)\|\mathbf{w}_k\|_2^2 + \|\nabla\mathbf{w}_k\|_2^2 + \|D^2\mathbf{w}_k\|_2^2\Big] \\ &\le c_4\sum_{|k|\ge 1}(|k|^4+|k|^2+1)|\boldsymbol{\sigma}_k|^2 \le c_5\,\|\mathcal{F}\|^2_{L^2_\sharp},\end{aligned} \tag{37}$$

so that, combining (36), (37), $(22)_1$, and (25), we obtain

$$\|\mathbf{w}\|_{\mathcal{W}^2_\sharp} + \|\boldsymbol{\sigma}\|_{W^2_\sharp} + \|\nabla\mathbf{q}\|_{L^2(D^{1,2})} \le c_6\,\|\mathbf{U}\|_{W^1_\sharp}\,. \tag{38}$$

We now observe that from (11), (12), and (20), we have

$$\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}} - \mathbf{U} = \mathbf{z} + \mathbf{w} + (\phi - 1)\mathbf{U} - \boldsymbol{W}\times\nabla\phi\,.$$

Thus, in view of (9) and the properties of ϕ, we deduce

$$\|\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}} - \mathbf{U}\|_{\mathcal{W}^2_\sharp} \le \|\mathbf{z}+\mathbf{w}\|_{\mathcal{W}^2_\sharp} + c_7\|\mathbf{U}\|_{W^1_\sharp}\,. \tag{39}$$

As a result, combining (39) with (23) and (38) allows us to conclude that $(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}} - \mathbf{U},\, p - \overline{p})$ is in the class $(17)_2$ and that it satisfies $(18)_2$. Finally, by [5, Theorem II.9.2], we infer that $\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}}$ satisfies also the asymptotic condition in $(8)_4$, which thus completes the existence part of the theorem. To show uniqueness, it is enough to show that, in the stated function class, the problem

$$\begin{aligned} &\left.\begin{aligned}\partial_t\boldsymbol{\vartheta} - \nu\Delta\boldsymbol{\vartheta} &= -\nabla p\\ \operatorname{div}\boldsymbol{\vartheta} &= 0\end{aligned}\right\}\ \text{in } \Omega\times\mathbb{R}\,,\\ &\boldsymbol{\vartheta}|_{\partial\Omega} = \dot{\boldsymbol{\xi}}\,,\quad \lim_{|x|\to\infty}\boldsymbol{\vartheta}(x,t) = \mathbf{0}\,,\ t\in\mathbb{R}\,,\\ &\ddot{\boldsymbol{\xi}} + \omega_0^2\,\boldsymbol{\xi} + \varpi\int_{\partial\Omega}\boldsymbol{T}(\boldsymbol{\vartheta},p)\cdot\boldsymbol{n} = \mathbf{0}\,,\end{aligned} \tag{40}$$

has only the zero solution. This is easily established. In fact, if we dot-multiply both sides of $(40)_1$ by $\boldsymbol{\vartheta}$, integrate by parts over Ω, and use $(40)_{3,5}$, we show

$$\frac{1}{2}\frac{d}{dt}(\varpi\,\|\boldsymbol{\vartheta}(t)\|_2^2 + |\dot{\boldsymbol{\xi}}(t)|^2 + \omega_0^2\,|\boldsymbol{\xi}(t)|^2) + 2\nu\,\varpi\,\|\mathbb{D}(\boldsymbol{\vartheta}(t))\|_2^2 = 0\,.$$

Integrating both sides of this equation from 0 to T and employing the T-periodicity imply

$$\|\mathbb{D}(\boldsymbol{\vartheta}(t))\|_2 \equiv 0. \tag{41}$$

From (41), we derive, in particular, $\|\mathbb{D}(\overline{\boldsymbol{\vartheta}})\|_2 \equiv 0$, which, since $\overline{\boldsymbol{\vartheta}}|_{\partial\Omega} = \mathbf{0}$, furnishes

$$\overline{\boldsymbol{\vartheta}} \equiv \mathbf{0}. \tag{42}$$

By (41)–(42), we then infer

$$\|\mathbb{D}(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})\|_2 \equiv \|\mathbb{D}(\boldsymbol{\vartheta}(t))\|_2 \equiv 0\,. \tag{43}$$

However, by assumption, we have that for all $t \in [0, T]$, $\boldsymbol{\vartheta}(t) \equiv (\boldsymbol{\vartheta}(t) - \overline{\boldsymbol{\vartheta}}) \in W^{1,2}(\Omega) \subset L^6(\Omega)$. Therefore, owing to Lemma 1, (43), and (42), we conclude $\boldsymbol{\vartheta} \equiv \nabla p \equiv \mathbf{0}$.

Acknowledgement I wish to thank warmly Dr. Ariana Ferrari for helpful conversations.

References

1. R.D. Blevins, *Flow Induced Vibrations* (Van Nostrand Reinhold, New York, 1977)
2. D. Bonheure, G.P. Galdi, F. Gazzola, Equilibrium configuration of a rectangular obstacle immersed in a channel flow. C. R. Math. Acad. Sci. Paris **358**, 887–896 (2020); Updated version in arXiv:2004.10062v2 (2021)
3. C. Dyrbye, S.O. Hansen, *Wind Loads on Structures* (Wiley, New York, 1997)
4. G.P. Galdi, On the motion of a rigid body in a viscous liquid: A mathematical analysis with applications, in *Handbook of Mathematical Fluid Dynamics*, vol. I (North-Holland, Amsterdam, 2002), pp. 653–791
5. G.P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Steady-State Problems*. Springer Monographs in Mathematics, 2nd edn. (Springer, New York, 2011)
6. G.P. Galdi, On the self-propulsion of a rigid body in a viscous liquid by time-periodic boundary data. J. Math. Fluid Mech. **22** Paper No. 61, 34 p. (2020)
7. G.P. Galdi, M. Kyed, Time–periodic flow of a viscous liquid past a body, in *Partial Differential Equations in Fluid Mechanics*. London Mathematical Society Lecture Note Series, vol. 452 (Cambridge University Press, Cambridge, 2018), pp. 20–49

8. A. Ostrovsky, Y.A. Stepanyants, Dynamics of particles and bubblers under the action of acoustic radiation force, in *Chaotic, Fractional, and Complex Dynamics: New Insights and Perspectives*, ed. by M. Edelman, M. Elbert, M.A.F. Sanjuan (Springer, Berlin, 2017), pp. 205–230
9. N.A. Pelekasis, J.A. Tsamopoulos, Bjerknes forces between two bubbles. Part 1. Response to a step change in pressure. J. Fluid Mech. **254**, 467–499; Part 2. Response to an oscillatory pressure field. J. Fluid Mech. **254**, 501–527 (1993)
10. C.H.K. Williamson, S. Govardhan, Vortex-induced vibrations. Ann. Rev. Fluid Mech. **36**, 413–55 (2004)

Critical Density Triplets for the Arrestment of a Sphere Falling in a Sharply Stratified Fluid

Roberto Camassa, Lingyun Ding, Richard M. McLaughlin, Robert Overman, Richard Parker, and Ashwin Vaidya

1 Introduction

Stratified fluids are those in which the background, equilibrium density field varies with height. Such systems occur naturally in many environments including lakes, oceans, and the Earth's atmosphere, as well as on other planets. Sedimentation of particles in stratified fluids ubiquitously occurs in natural environments [22] and plays a vital role in marine snow [21, 24], oil spill properties [2, 8], and distributions of dense microplastics [18] in the oceans and most recently in marine particulate aggregation [9].

Here, we focus on an interesting phenomenon that occurs when particle cross density interfaces between two fluids of different densities. In a work by Abaid et al. [1], the experimental sedimentation of a sphere in stratified saltwater was studied, and an intriguing bounce phenomenon was first documented in which a dense sphere falling in the fluid momentarily stopped and began to rise before ultimately falling. The momentary levitation of the sphere yields a prolonged settling time, which can contribute to the accumulation of particulate matter in the vicinity of strong density transition layers in the environment, e.g., haloclines or thermoclines [10, 11, 13, 21, 32].

We remark that there are three important factors to this bounce phenomenon. The first parameter is the Reynolds number $\frac{Ua}{\nu}$, where U is a characteristic velocity, a is the radius of the sphere, and ν is the kinematic viscosity. Several articles in the

R. Camassa · L. Ding · R. M. McLaughlin (✉) · R. Overman · R. Parker
Department of Mathematics, University of North Carolina, Chapel Hill, NC, USA
e-mail: camassa@amath.unc.edu; dingly@live.unc.edu; rmm@email.unc.edu

A. Vaidya
Department of Mathematics, Montclair State University, Montclair, NJ, USA
e-mail: vaidyaa@mail.montclair.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_4

literature [4, 5, 7] investigated the gravitational settling particles at low Reynolds numbers regime ($Re = 0.001$) where a complete first principle-based theory is possible. In the low Reynolds number, no bounce is observed, while the original work by Abaid et al. [1] involved Reynolds numbers in the hundreds. The second parameter is the relative thickness of the fluid density transition layer h/a which is characterized by the ratio of the layer thickness h to the particle radius a. Several studies [27, 30] have explored gravitational particle settling in sharply stratified fluids but reported no bounce phenomenon. In [27] and [30], the parameter $h/a \gg 1$ which takes values 60 and 20, respectively, whereas in the work of Abaid et al. [1], this parameter was much smaller, taking values around 3. In this paper, we focus on this parameter regime and explore the dependence of the bounce phenomenon on layer thickness experimentally. The third factor is the relation between the top fluid density ρ_1, bottom fluid density ρ_2, and sphere density ρ_b. The sphere rises into the upper fluid when its density is lower than that of the bottom fluid. When the sphere density is considerably higher than the fluid densities, the sphere penetrates the interface without bouncing back. As a result, predicting the range of sphere densities for which motion reversal is conceivable with known top and bottom fluid densities is intriguing.

Toward that goal, we are interested in using experiments and theory to determine a critical density triplet $(\rho_1, \rho_2, \rho_b^*)$ with the constraint $\rho_b^* \geq \rho_2 \geq \rho_1$. For any sphere with the density $\rho_b^* \geq \rho_b \geq \rho_2$, the falling sphere will bounce; if the sphere density equals the critical density, $\rho_b = \rho_b^*$, the falling sphere will just stop momentarily but not rise before ultimately descending to the tank bottom. Additionally, increasing the sphere density such that $\rho_b > \rho_b^*$ with fixed ρ_1 and ρ_2 or, equivalently, decreasing ρ_2 with fixed ρ_1 and $\rho_b = \rho_b^*$ prevents the sphere from stopping. As previously stated, we concentrate on cases with relatively high Reynolds numbers (between 20 and 450, based on the terminal velocities in the bottom and top layers, respectively) and a sharply stratified fluid ($h/a < 4$, $h \sim 0.9\,\text{cm}$, and $a = 0.25\,\text{cm}$). There are very few studies attempting to estimate these critical densities in the literature. Perhaps the first attempt was by Camassa et al. [3]. In that work, they proposed a coarse criterion for the critical density triplet which is based upon estimating the enhanced buoyancy through an asymptotic calculation of the drift volume induced by a sphere traveling a finite but large distance. Here, we aim to improve the critical density estimation based on the potential flow assumption and the system's potential energy for the levitation phenomenon of a sedimenting sphere in such a parameter regime. By analyzing the monotonicity of the potential energy curve, we establish an estimation that depends on the sphere and fluid density, the initial position of the sphere, and the thickness of the fluid density transition layer. Last, we anticipate this study could have applications in separating particles with different densities.

The paper is organized as follows: In Sect. 2, we present the setup of the model and formulate the energy equation of the system. In Sect. 3, we document the details of the experimental procedure and the critical density obtained by the experimental method. The linear regression of the experimental data shows critical densities satisfy the relation $\rho_b^* = 1.03\rho_2 - 0.0295\rho_1$. Additionally, we demonstrate that thicker layer transitions are less capable of arresting the sphere. In Sect. 4, we provide a criterion to estimate the critical sphere density with given top and the bottom fluid density, the layer thickness, and the initial position of the sphere. We document the details of our numerical method in Appendix.

2 Setup and Governing Equations

2.1 Setup Description

We consider a sphere with the density ρ_b and radius a sedimenting in a two-layer unbounded homogeneous fluid imposed upon the regions above and below an artificial interface as sketched in Fig. 1. The top and bottom fluid densities are ρ_1 and ρ_2, respectively ($\rho_2 > \rho_1$). At time $t = 0$, the interface of two fluid layers is centered at $z = 0$. The sphere starts at $(0, z_b(0))$, $z_b(0) = z_0 > a$ with an initial velocity $(0, v_b(0))$.

Due to the complex nature of the fluid flow around the body, the energy expressions of interest must be determined numerically by evaluating the evolution of the fluid interface. We make the following simplifying assumptions: First, the sphere penetrates the interface very fast, sufficiently so as not to generate any waves. The interface of the two-fluid layers stays sharp at the end of the experiment. Second, we assume the fluids are inviscid and irrotational. Based on these assumptions, the velocity field induced by the falling sphere can be modeled by the three-dimensional potential flow. To take advantage of the axial symmetry, we adopt the cylindrical coordinate system $r = \sqrt{x^2 + y^2}$. The position $(r(t), z(t))$ of a passive tracer in the fluid satisfies the equation:

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{v_b(t)a^3(r^2 - 2\tilde{z}^2)}{2(\tilde{z}^2 + r^2)^{5/2}}, \quad \frac{\mathrm{d}r}{\mathrm{d}t} = \frac{-3v_b(t)a^3\tilde{z}r}{2(\tilde{z}^2 + r^2)^{5/2}}, \tag{1}$$

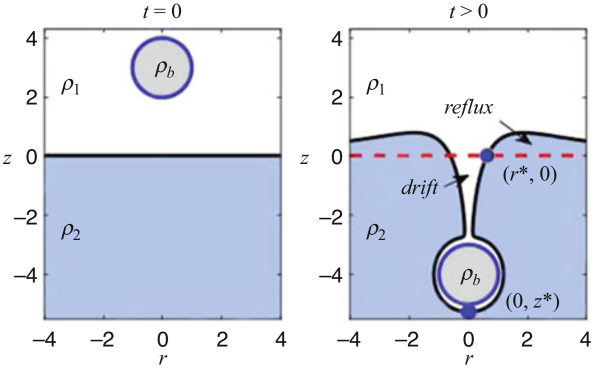


Fig. 1 The setup at initial time $t = 0$ and a finite time $t > 0$. The white, light blue, and gray regions are occupied by the top fluid, bottom fluid, and sphere, respectively. The black solid line is the interface between the top and the bottom fluid

where $\tilde{z} = z(t) - \left(z_0 + \int\limits_0^t v_b(s)\mathrm{d}s\right)$ and $(0, v_b(t))$ is the velocity of the sphere. Here, we observe an important property of this model which will greatly simplify the numerical calculation: the resulting interface shape from passive advection is independent of the time history of the sphere trajectory. To see this, rescaling time via $\int\limits_0^t v_b(s)\mathrm{d}s = k$ in Eq. (1) results in:

$$\frac{\mathrm{d}z}{\mathrm{d}k} = \frac{-a^3(r^2 - 2\tilde{z}^2)}{2(\tilde{z}^2 + r^2)^{5/2}}, \quad \frac{\mathrm{d}r}{\mathrm{d}k} = \frac{-3a^3\tilde{z}r}{2(\tilde{z}^2 + r^2)^{5/2}}, \quad \tilde{z} = z - (z_0 + k), \tag{2}$$

which is the equation of the tracer in the case that the sphere moves with the unit speed with respect to the pseudo-time k. Thus, the interface resulting from any sphere motion ending at the same position is identical. There are two options to explore here. First, one could explore the consequences of employing energy conservation to self-consistently evolve the sphere and fluid under the assumptions of potential flow. Second, one could study the potential energy stored in the fluid as a function of the three densities through a sphere moving at a constant speed. We will study the latter option here in this paper because of the independence of path history and its direct theoretical implications and discuss the limitations of the former in the conclusion section. Without loss of generality, we assume the sphere has a constant speed in the numerical simulation.

Last, we assume the fluid density linearly depends on the concentration of the solute, for example, the sodium chloride solution [17]. Since the solute is passively advected by the fluid flow and the diffusion is negligible in the experimental timescale, the fluid density field satisfies the advection equation:

$$\partial_t \rho + \mathbf{u} \cdot \nabla \rho = 0, \tag{3}$$

where $\mathbf{u}$ is the velocity field provided in Eq. (1). In this study, we consider the following initial density profile:

$$\rho(r, z, 0) = \rho_I(z) = \begin{cases} \rho_1 & \frac{L_\rho}{2} \le z, \\ \rho_1 + \frac{2(\rho_2 - \rho_1)}{L_\rho^2}\left(\frac{L_\rho}{2} - x\right)^2\left(\frac{x}{L_\rho} + 1\right) & -\frac{L_\rho}{2} < z < \frac{L_\rho}{2}, \\ \rho_2 & z \le -\frac{L_\rho}{2}, \end{cases} \tag{4}$$

which has a continuous first-order derivative. As $L_p \to 0$, Eq. (7) converges to the step function:

$$\rho(r, z, 0) = \begin{cases} \rho_1 & 0 \le z, \\ \rho_2 & z < 0. \end{cases} \tag{5}$$

2.2 Nondimensionalization

We nondimensionalize the equations and formulae via the following change of variables:

$$ar' = r,\ az' = z,\ \frac{a}{U}t' = t,\ a^4 g\rho_1 P' = P,\ \rho_1 \rho' = \rho,\ aL'_\rho = L_\rho,\ \mathrm{Re} = \frac{Ua}{\nu}. \tag{6}$$

We can distinguish the dimensional and dimensionless variables by the units after their values. The variables in Sect. 3 are in the dimensional form. Without specification, the variables in the next subsection and Sect. 4 are dimensionless. Hence, we can drop the prime without confusion. The nondimensionalized initial density profile is:

$$\rho(r,z,0) = \rho_I(z) = \begin{cases} 1 & \frac{L_\rho}{2} \leq z, \\ 1 + \frac{2(\rho_2 - 1)}{L_\rho^2}\left(\frac{L_\rho}{2} - x\right)^2\left(\frac{x}{L_\rho} + 1\right) & -\frac{L_\rho}{2} < z < \frac{L_\rho}{2}, \\ \rho_2 & z \leq -\frac{L_\rho}{2}. \end{cases} \tag{7}$$

2.3 The Potential Energy

Our goal is to use the energy of the sphere-fluid system to capture the arrestment of the spherical body as it moves through the two fluids. A full-scale dynamic explanation of the phenomena is very complex. We believe that our explanation provides an alternative and simpler explanation for the levitation phenomenon.

The total mechanical energy of the system at any instant of time can be given by:

$$E(t) = P_1(t) + P_2(t) + P_b(t) + K_b(t) + K_f(t), \tag{8}$$

where $P_1(t)$ and $P_2(t)$ are the potential energies of the top and bottom fluids, respectively. $P_b(t)$ is the potential energy of the body. $K_b(t)$ and $K_f(t)$ are the kinetic energies of the body and fluid, respectively. It is possible to express the kinetic energy of the fluid in terms of the added mass [15, 28].

First, we consider the extremely sharp stratification, namely, $L_\rho = 0$. A consequence of the law of conservation of mass is that the sphere penetrating into the bottom fluid causes a displacement of the top layer into the bottom and vice versa, the bottom fluid into the top layer. Therefore, in estimating the change in potential energies, the exact shape of the interface at a given time instant must be known. It is possible to write $\Delta P_1 = P_1(t) - P_1(0)$ as a result of gained volume by the top fluid in the drift region and a lost volume in the reflux region. Similarly, $\Delta P_2 = P_2(t) - P_2(0)$ can be thought of as potential energy of the bottom fluid due to a volume lost in the drift region and gained volume in the reflux region. We can therefore think of the net change in potential energy of the entire fluid as coming only from the drift and reflux volume regions in addition to the gravitational potential energy contribution of the fluid displaced by the body. The change of potential energy can then be written in the form:

$$\begin{aligned}\Delta P &= \Delta P_1 + \Delta P_2 + \Delta P_b \\ &= \pi(\rho_2 - 1)\left(\int_{r^*(z_b)}^{\infty} r z(r)^2 \mathrm{d}r - \int_{z^*(z_b)}^{0} r(z)^2 z \mathrm{d}z\right) + \frac{4\pi}{3}(\rho_b - 1)(z_b(t) - z_b(0)),\end{aligned} \tag{9}$$

where $(r^*(z_b), 0)$ are the coordinates of the point of zero Lagrangian displacement and $(0, z^*(z_b))$ is the lowest point on the interface (see Fig. 1). The asymptotic expansion of ΔP in the limit of some parameters is available via the asymptotic expansions provided in [3, 20, 33, 34]. In other cases, we have to compute ΔP numerically.

Second, for the case with a nonzero density transition layer thickness, we prefer to use the results in the zero-layer thickness case rather than solve the full advection Equation (3). We consider N artificial interfaces at $z = z_i$ which satisfy $-\frac{L_\rho}{2} = z_N < z_{N-1} < \ldots < z_2 < z_1 = \frac{L_\rho}{2}$. The fluid density between nth and $(n+1)$th layer is approximated by the density at the middle point $\rho_{I,n} = \rho_I((z_{n+1} + z_n)/2)$. Since the density field is passively advected by the flow, we can divide the $(N + 1)$ layer system into N independent two-layer systems while conserving the total potential energy (see the schematic in Fig. 2). The first system consists of two fluids separated by a sharp interface located at $z = z_1$. The top fluid density is 1 and the bottom fluid density is $\rho_{I,1}$. The interface in nth system $(n > 1)$ is located at $z = z_n$. Top and bottom fluid densities are 0 and $\rho_{I,n} - \rho_{I,n-1}$, respectively. The $(N + 1)$th system only contains a sphere centered at z_b with the density $\rho_b - 1$. Clearly, the summation of the potential energy of these $N + 1$ systems equals the potential energy of the original system. Since each system has a two-layer stratified fluid, we can apply the previous conclusion (9) and obtain the following expression of the change in potential energy:

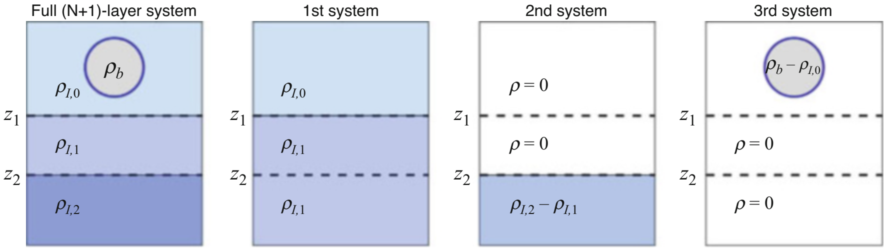


Fig. 2 A schematic of decomposing the full $(N+1)$ layer system into N two-layer systems and a system that only contains the sphere. The potential energy in the full system equals the sum of the potential energy in all subsystems. Here, $N = 2$, $t = 0$, $\rho_{I,0} = 1$, $\rho_{I,N} = \rho_2$. z_n is the height of the n-th interface

$$\begin{aligned}\Delta P &= \frac{4\pi}{3}(\rho_b - 1)(z_b(t) - z_b(0)) \\ &+ \lim_{N\to\infty}\sum_{n=1}^{N}\pi\left(\rho_{I,n} - \rho_{I,n-1}\right)\left(\int_{r_n^*(z_b)}^{\infty} r z_n(r)^2 \mathrm{d}r - \int_{z_n^*(z_b)}^{0} r_n(z)^2 z \mathrm{d}z\right),\end{aligned} \tag{10}$$

where $r_n(z)$, $z_n(r)$, $r_n^*(z_b)$, and $z_n^*(z_b)$ are associated with the interface starts at $z = z_n$ and $\rho_{I,0} = 1$, $\rho_{I,N} = \rho_2$. In the numerical simulation, we distribute z_n uniformly, and $N = 30 \sim 40$ is enough to obtain desirable results.

3 Experimental Methods and Results

Our experimental study involved dropping several spherical beads into a tank containing density-stratified liquids by varying the fluid and sphere densities and the layer thickness. In the following sections, we detail the exact procedure followed for various aspects and stages of the experimental study.

3.1 Tank, Bath, and Camera Setup

As presented in Fig. 3, the setup consists of two experimental tanks placed within a thermal bath. The outer tank (thermal bath tank) is regulated by a Thermo Scientific NESLAB RTE-7 Digital Plus Refrigerated Bath. The thermal bath is maintained at 19 degrees Celsius while taking data. Each of the inner experimental tanks consists of two Plexiglas sides for ease of viewing, a Plexiglas bottom, and two sides made of copper plates to assist in thermalization. The copper plates are coated in a protective sealant to prevent corrosion. The bottom layer of fluid is prepared in one of the inner tanks and the top layer in the other. The tank with the bottom layer of fluid is filled only halfway as the experiment will be run in this tank after pouring the top layer. The outer tank is a glass fish tank for ease of viewing. The NESLAB machine is connected to the fish tank via flexible PVC tubing. Two solid 1-foot sections of PVC pipe were glued into opposing corners of the fish tank. The flexible tubing is run over the top of the tank and down through these PVC pipes so that the input and output flow can be placed parallel to the experimental tank sides in an effort to minimize any vibrations upon the experimental tanks.

As stated earlier, there are two inner experimental tanks where saltwater solutions are prepared and one outer tank for the thermal bath. The two inner tanks have the outer dimensions of 7” by 7” by 12.5”. The two copper plates are each 0.25” thick, and the two Plexiglas sides and bottom are each 0.5” thickness. This results in inner dimensions of 6.5” by 6” by 12”. The thermal bath tank was simply a standard glass aquarium measuring 24.5” by 12.5” by 16.5”. On the rear of the outer tank behind the experimental tank in which the bead is dropped in, a background is

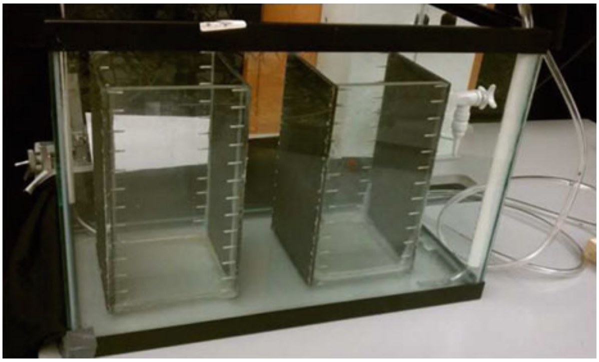


Fig. 3 Setup with two inner experimental tanks and outer tank as a thermal bath which is connected to a recirculating thermally controlled reservoir

placed. A variety of backgrounds were used. For the human viewer, a black and white checkerboard pattern of 0.635 cm by 0.635 cm squares produced the best visualization of the layer transition. The compression of the squares is easy to discern, and this gives a very clear view of the layer. For the purposes of tracking the bead in the DataTank script, a checkerboard pattern with 1 mm by 1 mm squares produced cleaner data. A solid background was never used, but this would most likely produce data with even less noise in the script, but such a background makes it much more difficult to discern a clean transition with the human eye.

A Sony HD camcorder is used for the duration of filming on the project. The camera is set up a meter in front of the inner tank the experiment will be run in. The camera is leveled to be on alignment with the water-air interface in the tank which contains the bottom layer. The alignment is performed using the lines on the front and back of tank. At the interface level, the lines on the front pane of the tank should be directly in front of the lines on the rear pane of the tank. Also, the lines on the rear of the tank should stick out from the end of the lines on the front of the tank by equal amounts on both left and right sides of the tank at the interface level (see Figs. 4 and 5). After completing the alignment phase, a meterstick is put in the center of the tank where the bead will fall, and the camera is focused on the smallest demarcations on the stick.

3.2 *Stratification Setup*

The density setup begins by filling the tanks with deionized water. Diamond Crystal Extra Coarse Solar Salt is then added to the water to bring the solution up to the desired density. An aquarium fishnet is used to hold the salt in the tank while mixing. The salt dissolves faster using the fishnet since there is a greater surface area of salt exposed to the water as opposed to being piled on the bottom. Also having

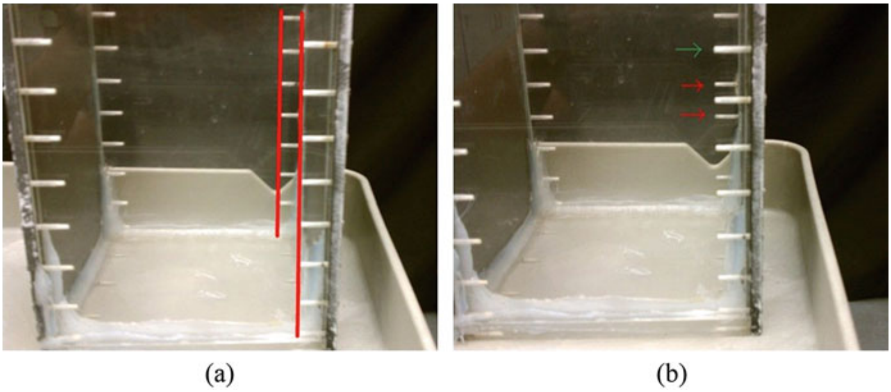


Fig. 4 **(a)** Vertical alignment: the tip of the screw holes on the back wall should be inside the top of the screw holes on the front wall. Also, the distance between the line of tips on both the back wall and the line of tips on the front wall (lines shown in red) should be the same for both the left and the right side of the tank. **(b)** The green arrow shows good alignment horizontally. The back screw hole is directly behind the front screw hole. Red lines show these screw holes are not aligned with the screw hole on the front of the tank

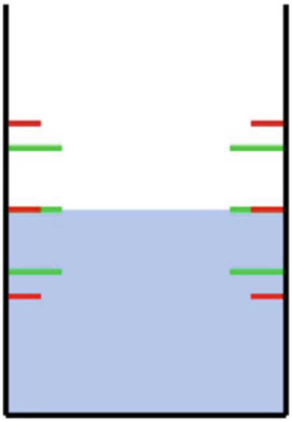


Fig. 5 Optimal setup: green lines are screw holes on the back wall, and red are those on the front. The distance between the hole ends on the front and the back of the tank (red and green liens) is equal on both the left and the right side of the tank. Also, at the level of the water-air interface (before top layer is poured), the rear screw holes are directly behind those of the front. Both above and below the water interface line the screw holes on the back wall appear before those of the front wall. They are “inside” the next set of screw holes on the front wall, and the difference in height between this set of red and green lines is equal for all four such sets

the salt in the fishnet allows quick removal of the salt to avoid overshooting the target density. While fixing the density, the temperature must be maintained at 19 degrees Celsius. Both solutions (top and bottom layer) are brought up to the desired densities in separate tanks. To ensure the solution is fully mixed, once attaining the target density, another sample is tested to ensure the density is true.

Pouring the layer is the most delicate part of the experiment as the two saltwater solutions will mix very easily. Care must be taken to pour the layer carefully and to not bump the tanks while the layer is being poured. The layer is poured through a diffuser (Fig. 6). The diffuser is simply a combination of two types of foam. The porous, spongelike center allows the layer being poured to slowly settle on top of the bottom layer. The outer foam (blue Styrofoam insulation board) keeps the diffuser buoyant enough to float up as the water level rises. Before using the diffuser, the diffuser must be primed with the top layer solution. For this reason, the tank with the top layer solution should be filled to the top even though the tank with the bottom layer is only filled half way. If the diffuser is primed with deionized water and the top layer has a density different from deionized water, then as the top layer is poured through the diffuser, there will be a strong tendency for the density to drift away from the original density as it flows through. Therefore, the diffuser needs to be primed with a few cups of the same solution which will eventually be poured through the diffuser. Along these lines, the diffuser must also be cleaned with deionized water after use to prevent salt accumulation. After attaining both desired densities and priming the diffuser, the diffuser is placed on the top of the experimental tank, floating on top of the bottom layer. A syringe is then used to gently pour the top layer through the diffuser. The general idea is to pour very slowly at the start (a rapid drip from a syringe) and then speed up as the distance between the diffuser and interface layer increases. For a more quantifiable rate, using 60-

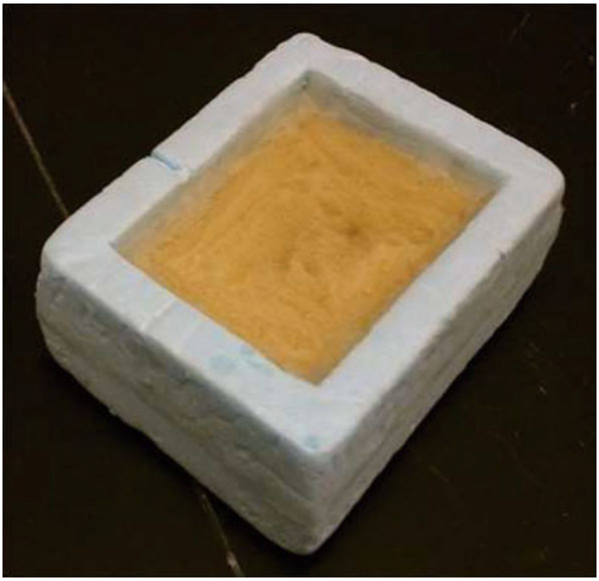


Fig. 6 The diffuser for pouring the stratified fluid

ml syringes with the exit hole widened to 8 mm the first pour through the diffuser should take approximately 45 s. The final pour through the diffuser should take 1.5 s. At the start, going too quickly will mix the interface, resulting in a poor transition layer. Once the diffuser sits an inch or so above of the interface, going too slowly just allows the interface more time to diffuse. The goal when pouring the layer for this experiment is to keep the layer thickness (see Layer Profiling) under 1.00 cm. When the tank is filled to just shy of the top, carefully lift the diffuser straight up, keeping it level while doing so to prevent water from pouring suddenly from it. Any big drips or sudden movements while removing the diffuser will disturb the layer. The purpose of filling the experimental tank all the way up is to ensure that the bead has enough time to reach its terminal velocity in the solution before encountering the transition layer.

The beads dropped are made of glass and have diameters of 4–5 mm. The beads have a very slight peak on one end as a result of manufacturing. Although very slight, this little extra glass makes this point the heaviest part of the bead. Therein, when dropping the bead, care is taken to orientate the bead so this point is on the bottom. Otherwise, the bead will spin in an effort to orientate itself in this manner as it falls. Before releasing the bead but while holding the bead under the surface of the water, care is taken to remove any air bubbles adhering to the bead by rolling the bead between two fingers. The beads were manufactured by the American Density Materials. To accurately measure their precise density, we used bisection search with Archimedes method using different tanks of saltwater in insulated containers. Fluid densities were accurately measured using an Anton Paar DMA 4500 Densitometer.

Layer profiling is performed using a conductivity probe attached to a Velmex, Inc. high-precision UniSlide. The layer is profiled after the bead has been dropped because the probe disturbs the layer upon passing through it. Once the slider is clamped to the tank to ensure it doesn't move while the probe is being lowered, measurements of conductivity and temperature are taken at 0.1-cm increments beginning at approximately a centimeter above the interface and continuing to approximately a centimeter below (until the readings level off). The layer thickness is calculated as follows: After all readings have been taken, the conductivity and temperature readings are converted to densities using previous experimental tables for saltwater solutions. Layer thickness is quantified by focusing on the change from the lowest density solution (top fluid) to the highest density fluid (bottom fluid). The layer thickness is measured as the distance between the points of 10 and 90% changes in density. More precisely, we define L_{10} to be the height at which the density profile takes the value of the top density plus 10% of the total density variation and similarly define L_{90} to be the height at which the density is top density plus 90% of the total density variation. Then, the layer thickness is given by $L = |L_{10} - L_{90}|$.

For the entirety of the experimental data (except for the section on the effects of diffusion time on layer thickness), the layer thickness was kept under 1.1 cm. The majority of the runs had a layer thickness of 0.85–0.9 cm. The final step in the experiment is to film a meterstick in the tank, right where the bead fell. This is used for attaining a scale in the script which calculates the minimum velocities.

3.3 Experimental Results

We repeated the experiment with hundreds of combinations of the various sphere and fluid densities to experimentally search for the critical density triplet $(\rho_1, \rho_2, \rho_b^*)$ as defined in Sect. 1. This is an extremely labor-intensive task as each measurement for one bottom density requires preparing the salt solution with the desired density, pouring an entirely fresh layer, and measuring the density profiles, which takes hours. With given top fluid and sphere densities, finding one critical bottom density takes at least ten independent fresh tanks.

Table 1 shows the critical density triplet $(\rho_1, \rho_2, \rho_b^*)$ and related experimental parameters. We calculate the sphere speed by a DataTank script. Because of the spatial and temporal resolution, as well as camera noise, the speed will never be exactly zero. Hence, we report the minimum speed in each experiment and adopt a consistent criterion to determine the arrestment. We have two observations from

Table 1 Critical density triplet $(\rho_1, \rho_2, \rho_b^*)$ and related experimental parameters. The fourth column is the minimum speed during the whole falling process. The radius of the spheres is 0.25 cm

Bead density(g/cc)	Top density(g/cc)	Bottom density(g/cc)	Min velocity (cm/s)	Layer thickness(cm)
1.0901	0.997	1.08680	0.069	0.814641521
±	0.997	1.08678	0.055	0.830600263
0.0001	0.997	1.08683	0.065	0.874706739
1.07495	0.997	1.07166	0.056	0.843531654
±	0.997	1.07182	0.001	0.903259318
0.0001	0.997	1.07170	0.039	0.904773401
1.05018	0.997	1.04805	0.029	0.819058413
±	0.997	1.04805	0.052	0.817808592
0.0002	0.997	1.04803	0.098	0.840865709
1.03997	0.997	1.03761	0.018	0.896928129
±	0.997	1.03759	0.06	0.795575335
0.0002	0.997	1.03761	0.051	0.905041625
1.03506	0.997	1.03335	0.006	0.885285341
±	0.997	1.03335	0.013	0.87707262
0.00015	0.997	1.03333	0.052	0.850024643
1.02017	0.997	1.01886	0.048	0.874341231
±	0.997	1.01882	0.035	0.882575525
0.0001	0.997	1.01883	0.067	0.890235316
1.0901	1.02998	1.08725	0.02	0.959324843
±	1.03006	1.08726	0.01	0.950414969
0.0001	1.03000	1.08728	0.05	0.98831011
1.0901	1.03999	1.08755	0.04	0.955474843
±	1.04001	1.08755	0.05	0.940466969
0.0001	1.04000	1.08755	0.06	0.99331213

Table 1. First, from the first six rows in the table, we see that ρ_2 increases as ρ_b increases when ρ_1 is fixed. Second, from the first row and the last two rows in the table, we see ρ_2 slightly decreases as ρ_1 increases with a fixed sphere density. More interestingly, the linear regression yields the following formula between $\frac{\rho_b^*}{\rho_1}$ and $\frac{\rho_2}{\rho_1}$:

$$\frac{\rho_b^*}{\rho_1} = a_1 \frac{\rho_2}{\rho_1} + a_2, \tag{11}$$

where $a_1 = 1.03$ and $a_2 = -0.0295$ with 95% confidence bounds, respectively, $(1.019, 1.042)$ and $(-0.04183, -0.01718)$. The summed square of residuals is 4.5369e–07, and the R-square value is 0.99987, which indicates a strong statistical linear relation. One can also see the linear relation in Fig. 11 which will be elaborated in the next section.

A brief inquiry as to the effects of diffusion time on the resulting layer thickness and critical density was carried out. The purpose of this was mainly to convince those carrying out the experiment that a difference in layer thickness between say 0.85 and 0.95 cm would not drastically distort the data. This is important because of the difficulty of ascertaining a precise and repeatable layer thickness every time. The process was simply pouring a tank and waiting for a certain amount of time before dropping the bead and measuring the layer thickness. All runs were done with the same bead of the density of 1.03997 g/cc. The critical density found with deionized water on top and no wait time between finishing pouring of the layer and dropping the bead was 1.03761g/cc with a layer thickness of 1.0072 cm. With a 2-hour wait time, the critical density was found to be 1.03770 g/cc with a layer thickness of 1.4998 cm. With a 3-hour wait time, the critical density was 1.03772 g/cc with a layer thickness of 1.6747 cm. This shows that after 3 hours of diffusion, the critical density is shifted by about 0.0001 g/cc. Hence, this provides some comfort for the few minutes of difference in time pouring the layer for each run, although more work on this could be done.

We could obtain a rough estimation of layer growth rate by assuming the density profile is:

$$\rho(z) = \rho_1 + \frac{\rho_2 - \rho_1}{2} \left(\operatorname{erf}\left(\frac{-z}{2\sqrt{\kappa(t + t_0)}} \right) + 1 \right),$$

where

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} \mathrm{d}t.$$

The 10th and 90th percentiles of erf(z) are around -0.90619 and 0.90619, respectively. Based on our definition of the density transition layer thickness, we have $L = 3.62478\sqrt{\kappa(t + t_0)}$. Applying the linear regression on L^2 with respect to t, namely, $L^2 = a_1 t + a_2$, we have $a_1 = 0.0001666$ and $a_2 = 1.023$ with

95% confidence bounds (0.0001122, 0.0002209) and (0.6159, 1.431), respectively. The summed square of residuals is 0.0011, and the R-square value is 0.9993. Comparing these two expressions of the layer thickness L, we obtain $t_0 = 6445.87$ s and $\kappa = 1.197809 \times 10^{-5}$ cm^2/s. The molecular diffusivity κ computed here is close to the diffusivity of NaCl reported in the literature [31] which is around $1.3 \times 10^{-5} \sim 1.6 \times 10^{-5}$ cm^2/s.

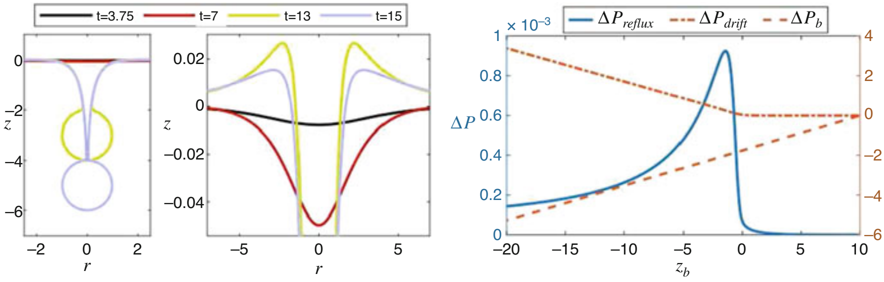
It is worth noting that because of the intrusive manner of layer profiling with a probe, each of these data points was from separate runs. Since each of these runs was from separated pours, the starting layer thickness differed between the three.

4 Critical Density and Energy Criterion

Camassa et al. [3] proposed that for a reversal of motion to occur, the averaged density of the sphere and drift fluid must necessarily be less than the density of the bottom layer fluid, in order to have negative buoyancy in the system, which leads to a coarse criterion for the critical density triplet $\bar{\rho}_b = (1 + c)\rho_2 - c\rho_1$, where c is the ratio of drift volume to the sphere volume. In the case that sphere travels from positive infinity to negative infinity, $c = \frac{1}{2}$. This criterion also shows a linear dependence between the critical densities which agrees with the experimental observations. If $\rho_1 = 0.997$ g/cc and $\rho_2 = 1.0376$ g/cc, Table 1 shows $\rho_b^* = 1.04$ g/cc, while this criterion predicts $\bar{\rho}_b = 1.0579$ g/cc, which has the relative difference $\frac{\bar{\rho}_b - \rho_1}{\rho_b^* - \rho_1} - 1 \approx 0.4163$. Considering that the sphere travels a finite distance, c could be smaller at roughly 0.44. Therefore, this reduces the relative difference to 0.35, which is still a relatively large difference. Of course, such large errors are to be expected as applying the drift volume directly to the sphere's buoyancy is at best a coarse approximation.

In this section, instead of considering the drift volume, we aim to improve the critical density estimation based on the system's potential energy. With numerical simulations, we next show that the potential energy as a function of the sphere position can change from a non-monotonic to a monotonic function of position as the sphere density increases. This observation will provide a criterion to estimate the critical sphere density such that any sphere with a density higher than the critical value cannot arrest. Since we have nondimensionalized the problem, we set the radius and sphere speed to be unity in the numerical simulations below.

Figure 7a shows a typical interface evolution as the sphere falls. The scale of the reflux region is small compared to the drift region. Figure 7b shows the variation of the change in potential energy contributions due to the sphere, the reflux, and drift regions as indicated in Eq. (9). The contribution due to the drift region can be seen to far exceed that due to the reflux region as would be expected in free space. After all, the drift volume is carried to infinity along with the moving sphere. While the drift contribution increases monotonically in the range of values computed, the reflux contribution peaks as the sphere just crosses the initial interface and then decays.



(a) (b)

Fig. 7 (**a**) The figure shows the time evolution of the interface based on the numerical simulation, as the sphere falls from the upper layer to the lower one. The drift region is very apparent in the left panel. The right panel shows a zoomed-in view of the reflux region at different times. (**b**) The dual y-axis chart shows, in the same simulation as panel (**a**), the change of potential energy contributed by the reflux region ΔP_{reflux} (left axis), drift region ΔP_{drift} (right axis), and body ΔP_b (right axis) when the center of the body z_b at different positions. Notice the scale difference in y-direction. The parameters are $\rho_2 = 1.04$, $\rho_b = 1.042$, and $z_0 = 10$

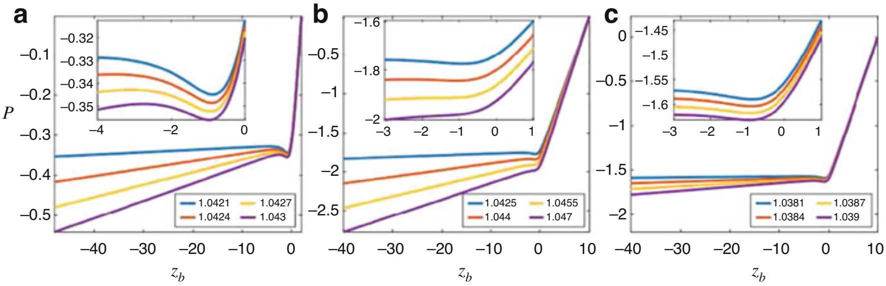


Fig. 8 The figure denotes the net potential energy variation in the entire system for different parameters. The parameters are $\rho_2 = 1.042$ in panel, (**a**, **b**), $\rho_2 = 1.038$ in panel (**c**), $z_b(0) = z_0 = 2$ in panel (**a**), and $z_0 = 10$ in panel (**b**, **c**). The legend shows the corresponding value of ρ_b for each curves. The insets are zoomed-in versions of each picture

In Fig. 8, we plot the total potential energy versus z_b with $\rho_2 = 1.042$ in panel (a, b) and $\rho_2 = 1.038$ in panel (c) while varying the sphere density ρ_b. The insets are zoomed-in view of the original picture near the critical points. The different columns correspond to different starting points for the sphere, namely, $z_b(0) = z_0 = 2$ in panel (a) and $z_0 = 10$ in panel (b, c). Figure 8 is very telling; as ρ_b is varied, we see significant variations in the types of curves produced: (1) The early part of each of these curves begins with the potential energy of the system decreasing with z_b caused by the falling sphere whose potential energy decreases. The potential energy contributions of the fluid are nonexistent at this stage; (2) the next phase of this curve occurs when the sphere reaches the interface. The sphere's potential energy continues to decrease; however, the drift and reflux regions now take place, resulting in a positive contribution to the potential energy which could counter the negative values of the sphere. The density of the sphere now becomes important henceforth. For the case that ρ_b is slightly larger than ρ_2 as is seen clearly in Fig. 8a and c, the positive energy of the fluid wins resulting in a minimum in the energy curve. Then the energy curve rises briefly as the sphere penetrates the interface and soon after begins to fall again as the fluid's potential energy fails to overcome that of the sphere. For the case when ρ_b is sufficiently larger than ρ_2, we see that the energy curve can be completely dominated by the potential energy of the sphere which shows monotonically decreasing behavior. Therefore, we denote the sphere density where the transient between these two cases happens as $\bar{\rho}_b$, which provides an estimation for ρ_b^* in the critical density triplet.

Mathematically, when the sphere density reaches the critical value, $\bar{\rho}_b$, there exists a degenerate critical point on the curve, $z_b = z_b^*$, such that:

$$\partial_{z_b} P\Big|_{z_b=z_b^*} = \partial_{z_b}^2 P\Big|_{z_b=z_b^*} = 0. \tag{12}$$

Equivalently, $\partial_{z_b} P$ is nonpositive for $z_b < z_0$ and only equals zero at one point $z_b = z_b^*$. Next, we first consider the density profiles with the zero density transition layer thickness and then study the profile with nonzero-layer thickness.

4.1 Zero Density Transition Layer Thickness

We start with the case $L_\rho = 0$. We explore the dependence of the critical density $\bar{\rho}_b$ for several parameters such as the bottom fluid density ρ_2 and the initial position z_0. First, as the initial position is closer to the interface, the sphere entrains the light fluid and therefore is harder to levitation. Figure (a) clearly shows the critical density asymptotically converges as the initial position moving further away from the interface. Similarly, Fig. 9b shows the critical point $z_b^*(z_0)$ also converges as $z_0 \to \infty$.

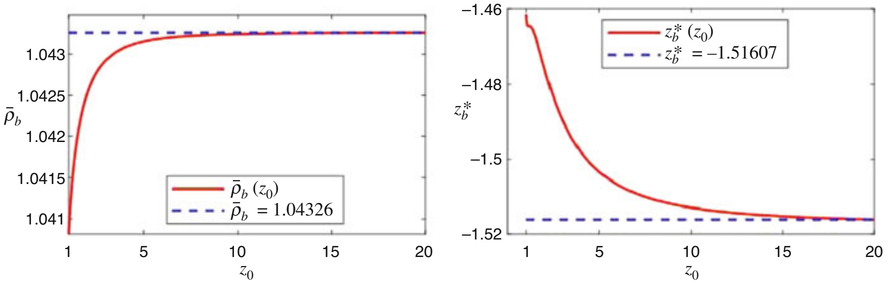
To further investigate the criterion (12), we take the derivative of the potential energy equation (9) with respect to z_b and setting it to zero, which yields:

$$\pi(\rho_2 - 1)\partial_{z_b}\left(\int_{r^*(z_b)}^{\infty} r z(r)^2 \mathrm{d}r - \int_{z^*(z_b)}^{0} r(z)^2 z \mathrm{d}z\right) + \frac{4\pi}{3}(\rho_b - 1) = 0. \tag{13}$$

The second equality in Eq. (9) becomes:

$$\pi(\rho_2 - 1)\partial_{z_b}^2\left(\int_{r^*(z_b)}^{\infty} r z(r)^2 \mathrm{d}r - \int_{z^*(z_b)}^{0} r(z)^2 z \mathrm{d}z\right) = 0. \tag{14}$$

Numerically solving the above equation yields the critical position $z_b = z_b^*$. Substituting it back to Eq. (13) gives the critical density:



(a) (b)

Fig. 9 **(a)** The red solid curve shows the variations of critical density $\bar{\rho}_b$ for a unit sphere as a function of initial position z_0 with $\rho_2 = 1.04$. The blue dashed line is the asymptote of the $\bar{\rho}_b(z_0)$ as $z_0 \to \infty$. **(b)** The red solid curve shows the critical point z_b^* as a function of initial position z_0. The blue dashed line shows an asymptote of $z_b^*(z_0)$

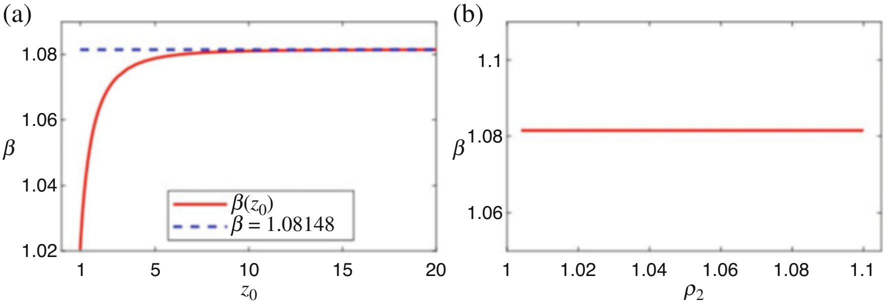


Fig. 10 **(a)** The red solid line shows the variation of the nondimensionless parameter β as a function of the initial position z_0. The blue dashed line is the asymptote of the $\beta(z_0)$ as $z_0 \to \infty$. **(b)** The figure shows that nondimensionless parameter β is independent of the bottom fluid density ρ_2

$$\bar{\rho}_b = (\rho_2 - 1)\beta + 1, \quad \beta = \frac{3}{4}\,\partial_{z_b}\left(\int_{r^*(z_b)}^{\infty} rz(r)^2 \mathrm{d}r - \int_{z^*(z_b)}^{0} r(z)^2 z \mathrm{d}z\right)\Bigg|_{z_b = z_b^*}, \tag{15}$$

where β can be numerically computed. We call attention to three important properties of β. First, β is dimensionless. Second, as demonstrated in Fig. 10a, β increases as the sphere initial position z_0 increases. Third, β is independent of ρ_2 and ρ_b, which is verified in Fig. 10b. Also note that in this model the depth z_b^* is independent of the sphere density as is evident from Eq. (14).

Equation (15) shows that $\bar{\rho}_b$ linearly depends on ρ_2 , which closely resembles Equation (11) from the experiments with slightly different coefficients. Additionally, Eq. (15) indicates that the difference between the critical sphere density and the bottom fluid density increases proportionally with the density differences in the

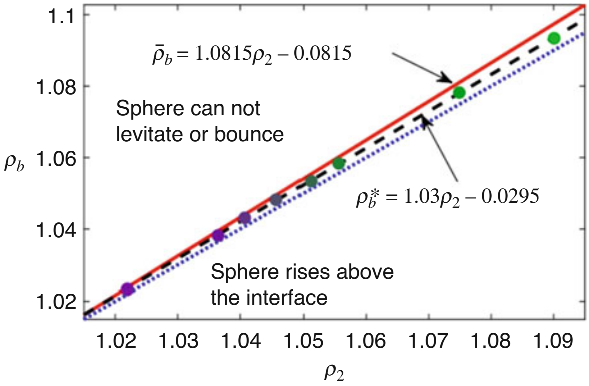


Fig. 11 Comparison of experimental results and theoretical prediction. The blue dotted line represents a line of slope 1, and the red solid lines are the critical sphere densities obtained by solving equation (12) numerically with $z_0 = 20$. The coordinates of the color dots are the nondimensionalized critical density triplet $(1, \rho_2, \rho_b)$ from Table 1. The black dashed line is the linear regression (11) of the experimental data. All dots are bounded by the red and blue lines

fluid and also with the value of $\beta - 1$ where $\beta > 1$. Therefore, we have equality, namely, $\bar{\rho}_s = \rho_2$ iff $\rho_2 = \rho_1$.

Now, we are ready to compare our numerically obtained values with those obtained from our experiments, which are presented in Fig. 11. When $z_0 = 20$, $\bar{\rho}_b$ can be expressed as:

$$\bar{\rho}_b = 1.0815\rho_2 - 0.0815. \tag{16}$$

We have two comments about this formula. First, $\bar{\rho}_b$ provided in the above equation shows a relative difference $\frac{\bar{\rho}_b(\rho_2)-1}{\rho_b^*(\rho_2)-1} - 1 \leq 0.043$ for all $\rho_2 \in [1, 1.1]$, which is a great improvement compared with the criterion proposed in article [3]. Second, the experimental values of critical density from Table 1 and the linear regression (11) consistently fall inside theoretical critical window $[\rho_2, \bar{\rho}_b]$ within the experimental parameter regime $1 < \rho_2 \leq 1.1$. Figure 11 demonstrates the sphere density $\bar{\rho}_b$ obtained from the energy curve constitutes an upper bound for the density ρ_b^* in the critical density triplet. This behavior is very reminiscent of a van der Waals-type pressure-volume curve [19].

4.2 *Nonzero Density Transition Layer Thickness*

Now, let us switch the attention to the case with nonzero density transition layer thickness. With a similar procedure, we have:

$$\bar{\rho}_b = (\rho_2 - 1)\beta + 1,$$

$$\beta = \frac{3}{4} \lim_{N\to\infty} \sum_{n=1}^{N} \frac{\rho_{I,n} - \rho_{I,n-1}}{\rho_2 - 1} \partial_{z_b} \left(\int_{r_n^*(z_b)}^{\infty} r z_n(r)^2 \mathrm{d}r - \int_{z_n^*(z_b)}^{0} r_n(z)^2 z \mathrm{d}z \right) \Bigg|_{z_b = z_b^*}, \tag{17}$$

where z_b^* is the location for the summation reaches the minimum value, which also solves the equation:

$$\lim_{N\to\infty} \sum_{n=1}^{N} \frac{\rho_{I,n} - \rho_{I,n-1}}{\rho_2 - 1} \partial_{z_b}^2 \left(\int_{r_n^*(z_b)}^{\infty} r z_n(r)^2 \mathrm{d}r - \int_{z_n^*(z_b)}^{0} r_n(z)^2 z \mathrm{d}z \right) \Bigg|_{z_b = z_b^*} = 0. \tag{18}$$

According to Eq. (7), we have $\rho(z_i) - \rho(z_j) = (\rho_2 - 1) f(z_i, z_j)$ for $z_i, z_j \in [-\frac{L_\rho}{2}, \frac{L_\rho}{2}]$ where $f(z_i, z_j)$ is independent of ρ_2. Hence, the dimensionless parameter β defined in Eq. (17) is independent of ρ_2 and only depends on L_p and the initial position of the sphere.

The layer thickness L in Sect. 3 is measured as the distance between the points of 10 and 90% changes in density. For the density profile provided in Eq. (7), we have $L \approx 0.6084 L_p$. The layer thickness presented in Table 1 is around $L = \frac{0.85\,\text{cm}}{0.25\,\text{cm}} \approx 3.4$. Hence, we numerically evaluate β in Eq. (17) with $L_p = 3.4/0.6084 \approx 5.6$ and obtain:

$$\bar{\rho}_b = 1.0289\rho_2 - 0.0289. \tag{19}$$

This estimation shows a relative difference $\frac{\bar{\rho}_b(\rho_2)-1}{\rho_b^*(\rho_2)-1} - 1 \leq 0.0078$ for all $\rho_2 \in [1, 1.1]$.

Figure 12 shows β and $\bar{\rho}_b$ decrease as the layer thickness increases which qualitatively captures the trend observed in experiments. We have two remarks: First, one can observe this trend with three fluid layers in the simulation. Second, the decreasing rate of $\bar{\rho}_b$ with respect to the layer thickness is relatively larger than the experimental observation: In the experiment, the nondimensionalized ρ_b^* changes around 0.0002 as the nondimensionalized layer thickness L increases from 4 to 6, while Fig. 12 shows $\bar{\rho}_b$ changes 0.0004 as L_ρ increases from 4/0.6084 to 6/0.6084. The difference in the change rate is because our potential energy-based criterion doesn't consider the complex nature of the fluid flow, for example, the viscous fluid layer around the body. However, even for the large L_ρ, $\bar{\rho}_b$ is an accurate estimation with less than 1% relative difference which is better than the estimation provided in Eq. (16) under the zero-layer thickness assumption.

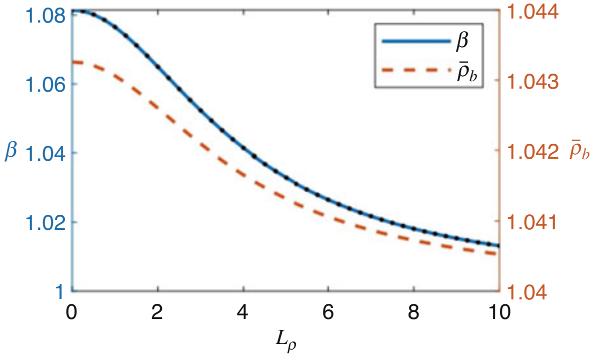


Fig. 12 The blue solid line and red dashed line show the variation of the nondimensionless parameter β (left axis) and estimated critical sphere density $\bar{\rho}_b$ (right axis) provided in Eq. (17) as functions of L_ρ, respectively. The parameters are $\rho_2 = 1.04$ and $z_0 = 20 + L_\rho/2$. For convergence, we need the number of artificial layers N to increase as L_ρ increases. We use $N = 30$ when $L_\rho = 1$ and $N = 80$ when $L_\rho = 10$. To show β is independent of ρ_2, we repeat the simulation for $\rho_2 = 1.05$ and then plot the resulted β with black dots which is fully overlapped with the solid blue curve for the case $\rho_2 = 1.04$

5 Conclusion and Discussion

We have studied the constraints of the fluid and sphere densities for producing a bouncing or levitation when a rigid sphere falls in a two-layer stratified fluid. Experiments focus on cases with relatively high Reynolds numbers (between 20 and 450) and sharply stratified fluid ($h/a < 4$). We explore the critical density triplet $(\rho_1, \rho_2, \rho_b^*)$ as defined in Sect. 1 with experimental and theoretical method. The main results are summarized as follows:

First, experiments show that the increasing of fluid density transition layer decreases the difference between ρ_2 and ρ_b^*. Second, when the relative fluid density transition layer thickness h/a is around 3∼4, the linear regression of the experimental shows that the dimensionless ratio ρ_b^*/ρ_1 increases linearly as ρ_2/ρ_1 increases. Third, based on the monotonicity of the potential energy curve, we identified a critical sphere density $\bar{\rho}_b$ which could be the estimation of ρ_b^*. With the zero-layer thickness assumption, the estimation $\bar{\rho}_b$ is an upper bound of the experimental measured ρ_b^* with less than 0.043 relative difference $\frac{\bar{\rho}_b(\rho_2)-\rho_1}{\rho_b^*(\rho_2)-\rho_1} - 1$ within the experimental parameter regime $\rho_2/\rho_1 \in [1, 1.1]$. Next, we demonstrated that $\bar{\rho}_b$ decreases as the layer thickness increases. When the layer thickness matches the experimental value $\frac{h}{a} \approx 3.4$, we obtain a more accurate estimation $\bar{\rho}_b$ which has a 0.0078 relative difference within the same experimental parameter regime, which is a great improvement compared with the estimation proposed in article [3].

Future research include a number of directions: First, the fluid layer surrounding the sphere could play a role in the settling dynamics [1, 6]. In particular, the work using a vertically towed fishing line in stratification [6] provides a starting point toward estimating the size of this boundary layer. We expect that including the viscous drag from the fluid layer into the model could yield a more accurate prediction of the critical densities. Second, many articles numerically studied the particles settling in an unbounded stratified fluid [12, 14, 29], while fewer studies have addressed the case with sharply stratified fluid and relative high Reynolds numbers. We plan to investigate the complicated dynamics of the momentary levitation discussed in this paper using a direct numerical simulation of the Navier-Stokes equation and understand the effect from the rigid boundaries. Third, we are interested in generalizing the theory to a dual problem, namely, the rise of droplets in a sharply stratified fluid [16, 23, 25], which is important in the study of the oil spill [2, 8].

Acknowledgments We acknowledge funding received from the following National Science Foundation Grant No. DMS-1910824 and Office of Naval Research Grant No ONR N00014-18-1-2490. Partial support for Lingyun Ding is gratefully acknowledged from the National Science Foundation Grant No. NSF-DMS-1929298 from the Statistical and Applied Mathematical Sciences Institute.

Appendix

Numerical Method

In this section, we document the details of the numerical calculation of the drift and reflux contributions to the potential energy and associated issues.

As the sphere penetrates the interface and deforms it, there is considerable stretching of the mesh in the region around the sphere, due to the potential nature of the flow. The uniform mesh on the interface cannot resolve the dynamics efficiently. Hence, for simplicity, we adopt a nonuniform mesh, which takes the parameterization:

$$x(s) = \begin{cases} 0 & s = 0, \\ e^{\frac{s}{r_1}} & 0 < s \leq r_1, \, y(s) = 0, \\ k_1(s - r_1) + 1 & r_1 < s, \end{cases} \tag{20}$$

where r_1 and k_1 are constants selected to resolve the interface evolution profile, which varies for different initial position of the sphere and the duration of the evolution. The mesh points cluster exponentially at the neighborhood of zero and distributed uniformly when they are far away from zero. The exponential profile of the initial mesh in the immediate vicinity of the particle provides a high density of meshes where the stretching is maximum.

A fourth-order explicit Runge-Kutta method with typical step size $\Delta t = 10^{-3}$ was used to compute the time evolution of the interface region as the sphere moved through the layers by solving the initial value problem with the velocity field provided in Eq. (1).

We approximate the interface by the cubic spline with the boundary condition "not-a-knot" to ensure fourth order accuracy in the interface tracking stage. The point of zero Lagrangian displacement $(r^*(z_b), 0)$ is calculated by solving the root of the spline function.

The integrals in Eq. (9) are evaluated by the trapezoidal rule. To achieve higher accuracy, one can adopt the spline-based quadrature rules described in [26, 35]. The potential energy as a function of the sphere position is approximated by a fifth-order spline function. Since differentiation could introduce unexpected oscillations when the data is not smooth enough, instead of solving $\partial^2_{z_b} P(z_b^*) = 0$ for the critical point z_b^*, we calculate z_b^* by finding the minimum value of $\partial_{z_b} P$.

We verified that all numerical results were not sensitive to an increase of either spatial or temporal resolution, therefore establishing the convergence of the numerical scheme.

References

1. N. Abaid, D. Adalsteinsson, A. Agyapong, R.M. McLaughlin, An internal splash: Levitation of falling spheres in stratified fluids. Phys. Fluids **16**, 1567–1580 (2004)
2. D. Adalsteinsson, R. Camassa, S. Harenberg, Z. Lin, R.M. McLaughlin, K. Mertens, J. Reis, W. Schlieper, B. White, Y. Liu, et al., Subsurface trapping of oil plumes in stratification: Laboratory investigations. Monitor. Model. Deepwater Horizon Oil Spill A Record-Breaking Enterprise **1**, 257–262 (2011)
3. R. Camassa, R.M. McLaughlin, M.N. Moore, A. Vaidya, Brachistochrones in potential flow and the connection to Darwin's theorem. Phys. Lett. A **372**, 6742–6749 (2008)
4. R. Camassa, C. Falcon, J. Lin, R.M. McLaughlin, R. Parker, Prolonged residence times for particles settling through stratified miscible fluids in the stokes regime. Phys. Fluids **21**, 031702 (2009)
5. R. Camassa, C. Falcon, J. Lin, R.M. McLaughlin, N. Mykins, A first-principle predictive theory for a sphere falling through sharply stratified fluid at low Reynolds number. J. Fluid Mech. **664**, 436–465 (2010)
6. R. Camassa, R.M. McLaughlin, M.N. Moore, K. Yu, Stratified flows with vertical layering of density: experimental and theoretical study of flow configurations and their stability. J. Fluid Mech. **690**, 571–606 (2012)
7. R. Camassa, S. Khatri, R.M. McLaughlin, J.C. Prairie, B. White, S. Yu, Retention and entrainment effects: experiments and theory for porous spheres settling in sharply stratified fluids. Phys. Fluids **25**, 081701 (2013)
8. R. Camassa, Z. Lin, R. McLaughlin, K. Mertens, C. Tzou, J. Walsh, B. White, Optimal mixing of buoyant jets and plumes in stratified fluids: theory and experiments. J. Fluid Mech. **790**, 71–103 (2016)
9. R. Camassa, D.M. Harris, R. Hunt, Z. Kilic, R.M. McLaughlin, A first principle mechanism for particulate aggregation and self-assembly in stratified fluids. Nat. Commun. **10**, 1–8 (2019)
10. S.A. Condie, M. Bormans, The influence of density stratification on particle settling, dispersion and population growth. J. Theoret. Biol. **187**, 65–75 (1997)
11. D. Deepwell, B.R. Sutherland, Cluster formation during particle settling in stratified fluid. Phys. Rev. Fluids **7**, 014302 (2022)

12. D. Deepwell, R. Ouillon, E. Meiburg, B.R. Sutherland, Settling of a particle pair through a sharp, miscible density interface. Phys. Rev. Fluids **6**, 044304 (2021)
13. K. Denman, A. Gargett, Biological-physical interactions in the upper ocean: the role of vertical and small scale transport processes. Ann. Rev. Fluid Mech. **27**, 225–256 (1995)
14. A. Doostmohammadi, S. Dabiri, A.M. Ardekani, A numerical study of the dynamics of a particle settling at moderate Reynolds numbers in a linearly stratified fluid. J. Fluid Mech. **750**, 5–32 (2014)
15. I. Eames, S. Belcher, J. Hunt, Drift, partial drift and Darwin's proposition. J. Fluid Mech. **275**, 201–223 (1994)
16. J. Farhadi, A. Sattari, P. Hanafizadeh, Passage of a rising bubble through a liquid-liquid interface: A flow map for different regimes. Canadian J. Chem. Eng. **100**, 375–390 (2022)
17. R.E. Hall, The densities and specific volumes of sodium chloride solutions at 25°. J. Washington Acad. Sci. **14**, 167–173 (1924)
18. A.J. Jamieson, L. Brooks, W.D. Reid, S. Piertney, B.E. Narayanaswamy, Linley, T., Microplastics and synthetic particles ingested by deep-sea amphipods in six of the deepest marine ecosystems on earth. Roy. Soc. Open Sci. **6**, 180667 (2019)
19. D.K. Kondepudi, et al., *Introduction to Modern Thermodynamics*, vol. 666. (Wiley Chichester, Chichester, 2008)
20. M. Lighthill, Drift. J. Fluid Mech. **1**, 31–53 (1956)
21. S. MacIntyre, A.L. Alldredge, C.C. Gotschalk, Accumulation of marines now at density discontinuities in the water column. Limnol. Oceanograp. **40**, 449–468 (1995)
22. J. Magnaudet, M.J. Mercier, Particles, drops, and bubbles moving across sharp interfaces and stratified layers. Ann. Rev. Fluid Mech. **52**, 61–91 (2020)
23. T.L. Mandel, L. Waldrop, M. Theillard, D. Kleckner, S. Khatri, et al., Retention of rising droplets in density stratification. Phys. Rev. Fluids **5**, 124803 (2020)
24. J.C. Prairie, K. Ziervogel, C. Arnosti, R. Camassa, C. Falcon, S. Khatri, R.M. McLaughlin, B.L. White, S. Yu, Delayed settling of marine snow at sharp density transitions driven by fluid entrainment and diffusion-limited retention. Marine Ecology Progr. Ser. **487**, 185–200 (2013)
25. V.A. Shaik, A.M. Ardekani, Drag, deformation, and drift volume associated with a drop rising in a density stratified fluid. Phys. Rev. Fluids **5**, 013604 (2020)
26. A. Sommariva, M. Vianello, Gauss–green cubature and moment computation over arbitrary geometries. J. Comput. Appl. Math. **231**(2), 886–896 (Elsevier, 2009)
27. A. Srdić-Mitrović, N. Mohamed, H. Fernando, Gravitational settling of particles through density interfaces. J. Fluid Mech. **381**, 175–198 (1999)
28. G.I. Taylor, The energy of a body moving in an infinite fluid, with an application to airships. Proc. R. Soc. Lond. Ser. A Containing Papers Math. Phys. Char. **120**, 13–21 (1928)
29. C.R. Torres, H. Hanazaki, J. Ochoa, J. Castillo, M. Van Woert, Flow past a sphere moving vertically in a stratified diffusive fluid. J. Fluid Mech. **417**, 211–236 (2000)
30. L. Verso, M. van Reeuwijk, A. Liberzon, Transient stratification force on particles crossing a density interface. Int. J. Multiphase Flow **121**, 103109 (2019)
31. V. Vitagliano, P.A. Lyons, Diffusion coefficients for aqueous solutions of sodium chloride and barium chloride. J. Amer. Chem. Soc. **78**, 1549–1552 (1956)
32. E. Widder, S. Johnsen, S. Bernstein, J. Case, D. Neilson, Thin layers of bioluminescent copepods found at density discontinuities in the water column. Marine Biol. **134**, 429–437 (1999)
33. C.S. Yih, New derivations of Darwin's theorem. J. Fluid Mech. **152**, 163–172 (1985)
34. C.S. Yih, Evolution of Darwinian drift. J. Fluid Mech. **347**, 1–11 (1997)
35. Q. Zhang, L. Ding, Lagrangian flux calculation through a fixed planar curve for scalar conservation laws. SIAM J. Sci. Comput. **41**(6), A3596–A3623 (SIAM, 2019). https://doi.org/10.1137/18M1210885

Part II
Computation

Numerical Investigation of Incompressible Fluid Flow in Planar Branching Channels

Tomáš Bodnár, Radka Keslerová, and Anna Lancmanová

1 Introduction

This work is motivated by the flow of blood and air in biomedical applications [25, 27, 15]. The blood flow in circulatory system and the air flow in the respiratory system share some physical similarities [14, 28, 20]. From the mathematical modeling point of view, both problems can be seen (with certain level of simplification) as flow of incompressible viscous fluid in a system of branching channels. The fluid is flowing through channels that are characterized by a multilevel (almost fractal-like) branching with secondary branches of different size and orientation with respect to the main channel. Such channel pattern is characteristic, however not exclusive, to the biomedical systems in living organisms. It can also be found in many industrial and environmental problems. The complicated configuration of the channels leads to numerous problems related to geometry description, its discretization (grid generation), and mathematical formulation of the associated problem including suitable boundary conditions.

The description and discretization of the channel geometry are usually done using the standard grid generation performed on the part of the space occupied by the fluid, i.e., the interior of the channel. The grid can be either structured (in simple cases) or rather unstructured for realistic description of larger representative parts of the circulatory or respiratory systems. This approach is quite common; however, it is associated with certain drawbacks. This includes rather nontrivial and expensive grid generation and necessity of the grid regeneration in case of even small geometry modifications. Also, the CFD solvers for general unstructured

T. Bodnár (✉) · R. Keslerová · A. Lancmanová
Department of Technical Mathematics, Faculty of Mechanical Engineering, Czech Technical University in Prague, Prague, Czech Republic
e-mail: Tomas.Bodnar@fs.cvut.cz; Radka.Keslerova@fs.cvut.cz; Anna.Lancmanova@fs.cvut.cz

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_5

grids are more complicated, making it difficult to implement any nonstandard mathematical models or boundary conditions.

Most of the problems associated with standard methods using wall-fitted grids can be avoided while using the immersed boundary method. In this case, a larger domain is discretized, typically in the form of cuboid in 3D or rectangle in 2D space. A grid (usually Cartesian, i.e., orthogonal equispaced) is constructed in the whole such domain where also the model equations are solved. The specific channel geometry is only represented at the level of the mathematical model being used, one model in the region occupied by the fluid and another one elsewhere. The switch between the models is simply realized using some kind of indicator (characteristic) function specifying the interior and exterior parts of the considered channel. In this case, due to very simple grid structure and domain shape, the CFD solver can be very simple. Implementation of alternative mathematical models or boundary conditions is quite easy and straightforward. Also, the changes in geometrical configuration of the channel are rather easy, only requiring re-definition of the characteristic function describing the fluid region. No grid changes or code adjustments are needed.

The main aim in this present work is to compare the results of a standard finite volume based method [16, 15] which used the wall-fitted grid, with a much simpler finite difference code working on regular Cartesian grid [18] while employing a generic implementation of immersed boundary method. As a test case, a channel with single branch inclined at different angles was chosen, similar to geometry used in [27] or [21]. The results of both codes are compared to see whether the simple finite difference and immersed boundary method based code can match the essential flow characteristics resolved by the older standard finite volume code used in some of our previous studies [3, 4, 15, 2].

2 Mathematical Model

The flow of incompressible, homogeneous Newtonian fluid can be described by the system of Navier-Stokes equations. It represents balance of mass and linear momentum.

2.1 Governing Equations

The incompressible Navier-Stokes equations can be written in a conservative (divergence) vector form:

$$\frac{\partial \boldsymbol{u}}{\partial t} + \nabla\cdot(\boldsymbol{u}\otimes\boldsymbol{u}) = -\frac{1}{\rho}\nabla p + \nu\Delta\boldsymbol{u}\ , \tag{1}$$

where ρ is the (constant) density, ν is the (constant) kinematic viscosity, and p, $\boldsymbol{u}$ is the pressure and velocity field, respectively. This form is directly derived from integral version of balance laws and is thus often considered as the most general differential form of the governing system. It can be directly used in finite volume discretization. Alternatively, this system is often rewritten in the nonconservative (convective) form as:

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = -\frac{1}{\rho}\nabla p + \nu \Delta \boldsymbol{u}. \tag{2}$$

This form is typically used in finite difference and finite element discretizations.

In both cases, the velocity field $\boldsymbol{u}$ obeys the incompressibility (divergence-free) constraint ($\nabla \cdot \boldsymbol{u} = 0$).

3 Numerical Methods

The numerical methods for both the finite difference and the finite volume in-house codes are briefly described here. We start from the description of the artificial compressibility method that is shared by both codes. For other ways of pressure-velocity coupling, see our previous works [17] and [18].

3.1 Artificial Compressibility Method

The artificial (or pseudo-)compressibility method is one of the simplest and most frequently used methods for velocity-pressure coupling in incompressible fluid flow simulations [8, 7]. It allows to calculate pressure from velocity field and to enforce the incompressibility, i.e., the divergence-free constraint. The method is based on a direct analogy with compressible flows, where time derivative of density is present in the continuity (mass balance) equation:

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{u}) = 0\,. \tag{3}$$

Such term can equivalently be expressed in terms of pressure (using the state equation), considering the dependence of density on pressure $\rho = \rho(p)$, which leads to reformulated continuity equation:

$$\frac{\partial \rho}{\partial p}\frac{\partial p}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla} \rho + \rho\, \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0\,. \tag{4}$$

The first term in (4) is associated with the speed of sound c by $\frac{\partial p}{\partial \rho} = c^2$, considering an adiabatic change of state. Adopting this physical argument and returning back to homogeneous fluids with constant density, a modified continuity equation can be written:

$$\frac{1}{\beta^2}\frac{\partial p}{\partial t} + \rho\, \nabla \cdot \boldsymbol{u} = 0\,, \tag{5}$$

with β being artificial (pseudo-)speed of sound. This adjustable parameter has finite value, which should be suitably chosen depending on the case solved, to ensure the pressure-velocity coupling and convergence of the numerical solution. For further details concerning the choice and effect of the artificial speed of sound β, see, for example, [11] or [29]. Because the added term containing the time derivative of pressure is purely artificial, the method is often just used solving steady problems. In such cases, the nonphysical term vanishes in the steady state, and the divergence-free velocity field is recovered.

3.2 Finite Difference Solver

The finite difference approximation of governing equations is a natural choice because of the use of immersed boundary method on Cartesian (structured, orthogonal, equispaced) grids. In such case, the discretization is extremely simple, allowing for easy implementation and modification of various numerical methods and algorithms.

In discretization of the governing system, both the nonconservative (convective) and the conservative (divergence) form of the equations can be used. The conservative system (1), including the continuity equation (divergence-free constraint), can be (for homogeneous case with constant density) written in vector form as:

$$\boldsymbol{D}\mathbf{W}_t + \mathbf{F}_x + \mathbf{G}_y + \mathbf{H}_z = \mathbf{R}_x + \mathbf{S}_y + \mathbf{T}_z, \tag{6}$$

where $\boldsymbol{D} = diag\,(0, 1, 1, 1)$ and $\mathbf{W} = col(p, u, v, w)$ are the vectors of unknowns.

In case of artificial compressibility method, the diagonal matrix $\boldsymbol{D}$ is replaced by $\boldsymbol{D}_\beta = diag\left(\frac{1}{\rho\beta^2}, 1, 1, 1\right)$ including the artificial compressibility parameter β:

$$\mathbf{F} = \begin{pmatrix} u \\ u^2 + p/\rho \\ u\,v \\ u\,w \end{pmatrix}, \qquad \mathbf{G} = \begin{pmatrix} v \\ v\,u \\ v^2 + p/\rho \\ v\,w \end{pmatrix}, \qquad \mathbf{H} = \begin{pmatrix} w \\ w\,u \\ w\,v \\ w^2 + p/\rho \end{pmatrix}, \tag{7}$$

$$\mathbf{R} = \begin{pmatrix} 0 \\ \nu u_x \\ \nu v_x \\ \nu w_x \end{pmatrix}, \qquad \mathbf{S} = \begin{pmatrix} 0 \\ \nu u_y \\ \nu v_y \\ \nu w_y \end{pmatrix}, \qquad \mathbf{T} = \begin{pmatrix} 0 \\ \nu u_z \\ \nu v_z \\ \nu w_z \end{pmatrix}. \tag{8}$$

In laminar case (i.e., with constant viscosity ν), the right-hand side of equation (6) can further be simplified to contain the viscous terms in the form of the Laplacian of velocity:

$$\boldsymbol{D}\mathbf{W}_t + \mathbf{F}_x + \mathbf{G}_y + \mathbf{H}_z = \nu \boldsymbol{D}\Delta\mathbf{W} \tag{9}$$

Again, in case of artificial compressibility method, the diagonal matrix $\boldsymbol{D}$ on the left-hand side is replaced by $\boldsymbol{D}_\beta$, adding the time derivative of pressure to the continuity equation. Similar modification on the right-hand side is only used exceptionally, to add extra stabilization (regularization) term proportional to Laplacian of pressure (see, for example, [26, 3]).

The nonconservative form of (9) or (2), respectively, is:

$$\boldsymbol{D}\mathbf{W}_t + u\mathbf{W}_x + v\mathbf{W}_y + w\mathbf{W}_z = -\frac{1}{\rho}\widehat{\nabla} p + \nu \boldsymbol{D}\Delta\mathbf{W}\,, \tag{10}$$

where $\widehat{\nabla} = col\left(0, \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)$ is the extended gradient operator, which if applied to pressure field leads to $\widehat{\nabla} p = col(0,\ p_x,\ p_y,\ p_z)$.

3.2.1 Immersed Boundary Method

The immersed boundary method is a simple way that allows to simulate flows in and around complex geometries, without the need to construct shape-specific grids for each individual problem configuration. More details and further references on the principle and various versions of the immersed boundary method can, for example, be found in [23, 24] or [22, 10] and [6].

The version used in this work is probably the simplest of all possible implementations of the immersed boundary method. The situation can be described using the schematic pictures of the grids used for finite volume (FVM) and finite difference (FDM) methods in this work. The typical (structured in this case) grid used FVM simulations has a multiblock structure, with individual grids constructed for each block. These grids have simple structure with the grid lines fitted to physical or artificial boundaries of the computational domain. This results in grids that are aligned (parallel) to boundaries. This allows to achieve optimal spatial accuracy of the applied numerical schemes. Also, the grid points coincide with the boundary, so the implementation of boundary conditions is quite straightforward. This situation is shown in Figs. 1(a) and 2(a). Such grid can easily be refined close to the wall, but

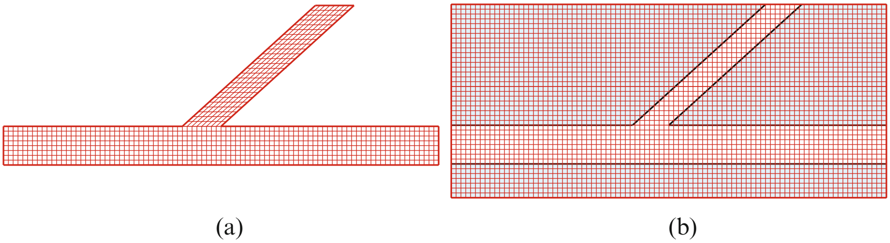


Fig. 1 Grid structure for finite volume and finite difference simulations. (**a**) Wall-fitted grid for FVM. (**b**) Immersed boundary grid for FDM

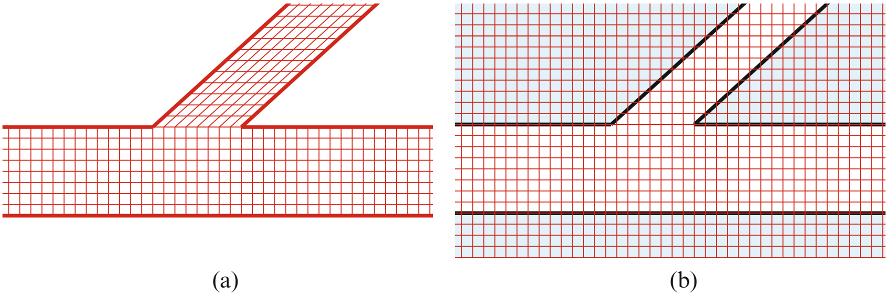


Fig. 2 Detail of the grid for finite volume and finite difference simulations. (**a**) Wall-fitted grid for FVM. (**b**) Immersed boundary grid for FDM

sudden changes in the directions and cell sizes at the interfaces of individual blocks might be problematic (and difficult to treat numerically).

A completely different grid (and discretization) concept is used for the presented immersed boundary finite difference code. Instead of constructing the grid just for the interior of the channel (occupied by a fluid), a larger rectangular domain is chosen, containing the physical domain of interest (the channel). This situation is shown in Fig. 1(b). This whole rectangle now represents a computational domain, and a Cartesian (i.e., regular, orthogonal equispaced) grid is constructed for the whole (rectangular) domain. In this configuration, the physical boundaries (channel walls in the considered case) no more coincide with boundaries of the grid as it is shown in Figs. 1(b) and 2(b).

In the immersed boundary FDM method, the governing equations are discretized in the whole rectangular domain, and suitable boundary conditions are only imposed on its boundary. The unknown fields of velocity and pressure are sought in all internal points of the domain, distinguishing the points inside of the fluid domain (marked by white color in Fig. 2(b)) and inside of the solid domain (marked by light blue color in Fig. 2(b)). The values of the velocity in the solid region are then reset to zero (which corresponds to solid in rest), and no special treatment is applied to pressure. In fact, the whole calculation of (fluid) velocity in the points inside the solid region can be skipped, directly set to zero, so the discrete governing equations are only solved in the points in the fluid region. This approach was already tested in [19], where some additional references can be found. The advantage of this approach is its simplicity, where very basic discretization formulas can lead to highest accuracy, due to Cartesian (undeformed, unstretched) grid. The simulations shown further prove that even this simplest version of the immersed boundary method gives quite good results comparable with those obtained by more complicated and refined finite volume codes.

3.2.2 Lax-Friedrichs Scheme

The Lax-Friedrichs scheme is one of the simplest classical schemes used for numerical discretization of conservation laws and in simulation of transport phenomena. It is an explicit method, employing central in space discretization, that is formally of first order in both space and time. Despite of its rather low theoretical accuracy, it is well known for its simplicity (no need for upwinding) and robustness. It is well known that in its basic form the scheme it contains quite strong internal numerical diffusion. This nonphysically high diffusivity can however be significantly reduced and individually adjusted for each solved case [17, 18].

Besides of its robustness, the Lax-Friedrichs scheme allows for very easy and straightforward implementation of pressure-velocity coupling methods, including the artificial compressibility, pressure correction, or operator-splitting methods. Some of these methods are more difficult to implement in the other predictor-corrector or multistage methods we have used in our study.

When the conservative formulation is used, the (modified) Lax-Friedrichs scheme can conveniently be written in a vector form. Here, it is written in 2D version as it was used in our simulations:

$$\begin{aligned}\mathbf{W}^{n+1}_{i,j} = (1-\zeta)\mathbf{W}^{n}_{i,j} + \zeta\left(\frac{\mathbf{W}^{n}_{i+1,j}+\mathbf{W}^{n}_{i-1,j}+\mathbf{W}^{n}_{i,j+1}+\mathbf{W}^{n}_{i,j-1}}{4}\right)+ \\ + \Delta t\ \boldsymbol{D}_{\beta}^{-1}\Bigg[-\frac{\mathbf{F}^{n}_{i+1,j}-\mathbf{F}^{n}_{i-1,j}}{2\Delta x}-\frac{\mathbf{G}^{n}_{i,j+1}-\mathbf{G}^{n}_{i,j-1}}{2\Delta y}+ \\ + \nu\,\boldsymbol{D}\left(\frac{\mathbf{W}^{n}_{i+1,j}-2\mathbf{W}^{n}_{i,j}+\mathbf{W}^{n}_{i-1,j}}{\Delta x^2}+\frac{\mathbf{W}^{n}_{i,j+1}-2\mathbf{W}^{n}_{i,j}+\mathbf{W}^{n}_{i,j-1}}{\Delta y^2}\right)\Bigg]\end{aligned} \tag{11}$$

This formal simplicity comes mainly from the fact that the pressure is hidden in the inviscid fluxes **F** and **G** (and **H**). For the nonconservative version of the Lax-Friedrichs scheme, the component vise form is rather used leading to (12), (13), and (14):

$$\begin{aligned} p_{i,j}^{n+1} &= (1-\zeta)p_{i,j}^{n} + \zeta\left(\frac{p_{i+1,j}^{n}+p_{i-1,j}^{n}+p_{i,j+1}^{n}+p_{i,j-1}^{n}}{4}\right) + \\ &+ \rho\beta^2\Delta t\left[-\frac{u_{i+1,j}^{n}-u_{i-1,j}^{n}}{2\Delta x} - \frac{v_{i,j+1}^{n}-v_{i,j-1}^{n}}{2\Delta y}\right], \end{aligned} \tag{12}$$

$$\begin{aligned} u_{i,j}^{n+1} &= (1-\zeta)u_{i,j}^{n} + \zeta\left(\frac{u_{i+1,j}^{n}+u_{i-1,j}^{n}+u_{i,j+1}^{n}+u_{i,j-1}^{n}}{4}\right) + \\ &+ \Delta t\left[-u_{i,j}^{n}\frac{u_{i+1,j}^{n}-u_{i-1,j}^{n}}{2\Delta x} - v_{i,j}^{n}\frac{u_{i,j+1}^{n}-u_{i,j-1}^{n}}{2\Delta y} - \frac{1}{\rho}\frac{p_{i+1,j}^{n}-p_{i-1,j}^{n}}{2\Delta x}\right. \\ &\left.+ \nu\left(\frac{u_{i+1,j}^{n}-2u_{i,j}^{n}+u_{i-1,j}^{n}}{\Delta x^2} + \frac{u_{i,j+1}^{n}-2u_{i,j}^{n}+u_{i,j-1}^{n}}{\Delta y^2}\right)\right], \end{aligned} \tag{13}$$

$$\begin{aligned} v_{i,j}^{n+1} &= (1-\zeta)v_{i,j}^{n} + \zeta\left(\frac{v_{i+1,j}^{n}+v_{i-1,j}^{n}+v_{i,j+1}^{n}+v_{i,j-1}^{n}}{4}\right) + \\ &+ \Delta t\left[-u_{i,j}^{n}\frac{v_{i+1,j}^{n}-v_{i-1,j}^{n}}{2\Delta x} - v_{i,j}^{n}\frac{v_{i,j+1}^{n}-v_{i,j-1}^{n}}{2\Delta y} - \frac{1}{\rho}\frac{p_{i,j+1}^{n}-p_{i,j-1}^{n}}{2\Delta y}\right. \\ &\left.+ \nu\left(\frac{v_{i+1,j}^{n}-2v_{i,j}^{n}+v_{i-1,j}^{n}}{\Delta x^2} + \frac{v_{i,j+1}^{n}-2v_{i,j}^{n}+v_{i,j-1}^{n}}{\Delta y^2}\right)\right]. \end{aligned} \tag{14}$$

3.2.3 MacCormack Scheme

The MacCormack scheme is a relatively simple step-up from the first-order Lax-Friedrichs method which offers theoretically second order of accuracy in both space and time. It is a specific variant of Lax-Wendroff class of methods that is written in the predictor-corrector form using asymmetric forward/backward discretization stencil to approximate spatial derivatives to provide finally a central (second order) approximation. The increased formal accuracy of the method comes at the price of doubled computational cost (for both CPU and memory requirements). This method is in general also a bit less robust than the previously described Lax-Friedrichs scheme, because it contains significantly less numerical diffusion, which may lead to nonphysical oscillations in computational field (and need for extra numerical stabilization). Despite of these shortcomings, the MacCormack scheme is a simple representative of second-order methods that can bring some additional comparative advantage into the simulations performed within the scope of this work.

To describe the MacCormack scheme, it's better to start from a rearranged equation (9), where all terms except the time derivative are placed on the right-hand side:

$$\mathbf{W}_t = \boldsymbol{D}_\beta^{-1}\left[-\left(\mathbf{F}_x+\mathbf{G}_y+\mathbf{H}_z\right)+\nu\boldsymbol{D}\Delta\mathbf{W}\right] \tag{15}$$

Now, in order to update in time the values of the vector $\mathbf{W}^n$ to $\mathbf{W}^{n+1}$, the aim is to construct an approximation of $\mathbf{W}_t$ from (15). This approximation is built differently in predictor (e.g., using backward differences) and in corrector (using forward differences). Then the final update is performed using linear combination of the two values obtained. This procedure is formalized in the following steps:

Predictor Step The spatial derivatives of inviscid fluxes in (15) are discretized using backward differences, while central differencing is used for viscous terms:

$$\begin{aligned}\left(\frac{\partial \mathbf{W}}{\partial t}\right)^n_{i,j} = \boldsymbol{D}_\beta^{-1}\Bigg[&- \frac{\mathbf{F}^n_{i,j} - \mathbf{F}^n_{i-1,j}}{\Delta x} - \frac{\mathbf{G}^n_{i,j} - \mathbf{G}^n_{i,j-1}}{\Delta y} + \\ &+ \nu\, \boldsymbol{D}\left(\frac{\mathbf{W}^n_{i+1,j} - 2\mathbf{W}^n_{i,j} + \mathbf{W}^n_{i-1,j}}{\Delta x^2} + \frac{\mathbf{W}^n_{i,j+1} - 2\mathbf{W}^n_{i,j} + \mathbf{W}^n_{i,j-1}}{\Delta y^2}\right)\Bigg],\end{aligned} \tag{16}$$

where values of all variables on the right side are known at the current time level n. Using the approximate value $\left(\frac{\partial \mathbf{W}}{\partial t}\right)^n_{i,j}$, the preliminary (marked by tilde) values of unknown $\mathbf{W}^{n+1}$ are predicted from the first two terms of the corresponding Taylor series, i.e., using explicit Euler time-stepping:

$$\widetilde{\mathbf{W}}^{n+1}_{i,j} = \mathbf{W}^n_{i,j} + \Delta t \left(\frac{\partial \mathbf{W}}{\partial t}\right)^n_{i,j}, \tag{17}$$

where the value of the first term on the right-hand side is known at current time level and the second term was previously evaluated from (16). The obtained auxiliary values of $\widetilde{\mathbf{W}}^{n+1}$ are used in the corrector step.

Corrector Step Here, the predicted values $\widetilde{\mathbf{W}}^{n+1}$ are used to calculate another approximation of $\mathbf{W}_t$ from (15) but this time applying the forward in space differentiation of inviscid fluxes, while central differencing is used again for viscous fluxes:

$$\begin{aligned}\left(\frac{\partial \widetilde{\mathbf{W}}}{\partial t}\right)^{n+1}_{i,j} = \boldsymbol{D}_\beta^{-1}\Bigg[&- \frac{\widetilde{\mathbf{F}}^{n+1}_{i+1,j} - \widetilde{\mathbf{F}}^{n+1}_{i,j}}{\Delta x} - \frac{\widetilde{\mathbf{G}}^{n+1}_{i,j+1} - \widetilde{\mathbf{G}}^{n+1}_{i,j}}{\Delta y} + \\ &+ \nu\, \boldsymbol{D}\left(\frac{\widetilde{\mathbf{W}}^{n+1}_{i+1,j} - 2\widetilde{\mathbf{W}}^{n+1}_{i,j} + \widetilde{\mathbf{W}}^{n+1}_{i-1,j}}{\Delta x^2} + \frac{\widetilde{\mathbf{W}}^{n+1}_{i,j+1} - 2\widetilde{\mathbf{W}}^{n+1}_{i,j} + \widetilde{\mathbf{W}}^{n+1}_{i,j-1}}{\Delta y^2}\right)\Bigg].\end{aligned} \tag{18}$$

Again, all the terms on the right-hand side can easily be evaluated from the known values $\widetilde{\mathbf{W}}^{n+1}_{i,j}$.

Variables Update Having now the two approximate values of $\mathbf{W}_t$ from predictor and corrector steps, the final approximation can be built as an average of the two values, representing an approximation of $\mathbf{W}_t$ at the time level $n + 1/2$:

$$\left(\frac{\partial \mathbf{W}}{\partial t}\right)_{i,j}^{n+1/2} = \frac{1}{2}\left[\left(\frac{\partial \mathbf{W}}{\partial t}\right)_{i,j}^{n} + \left(\frac{\partial \widetilde{\mathbf{W}}}{\partial t}\right)_{i,j}^{n+1}\right] \tag{19}$$

Finally, the values of the unknown variable vector $\mathbf{W}^{n+1}$ can be obtained by making a forward time step of the length Δt from $\mathbf{W}^n$ using the approximate time derivative $\mathbf{W}_t^{n+1/2}$:

$$\mathbf{W}_{i,j}^{n+1} = \mathbf{W}_{i,j}^{n} + \Delta t \left(\frac{\partial \mathbf{W}}{\partial t}\right)_{i,j}^{n+1/2} \tag{20}$$

In principle, the MacCormack method is thus similar to Heun's method for solution of ordinary differential equations. Also here, it leads to increased (second order) accuracy with respect to time. The spatial accuracy of MacCormack method is the same (second order) as for the Lax-Wendroff scheme due to the use of central approximation of all spatial derivatives [9].

The whole scheme can alternatively (and equivalently) be rewritten in more common form, allowing direct comparison with Lax-Friedrichs scheme. The expressions for predicted and corrected values are shown in (21) and (22):

$$\begin{aligned}\widetilde{\mathbf{W}}_{i,j} = \mathbf{W}_{i,j}^{n} + \Delta t\ \boldsymbol{D}_\beta^{-1}\Bigg[&- \frac{\mathbf{F}_{i,j}^{n} - \mathbf{F}_{i-1,j}^{n}}{\Delta x} - \frac{\mathbf{G}_{i,j}^{n} - \mathbf{G}_{i,j-1}^{n}}{\Delta y} + \\ &+ \nu\, \boldsymbol{D}\left(\frac{\mathbf{W}_{i+1,j}^{n} - 2\mathbf{W}_{i,j}^{n} + \mathbf{W}_{i-1,j}^{n}}{\Delta x^2} + \frac{\mathbf{W}_{i,j+1}^{n} - 2\mathbf{W}_{i,j}^{n} + \mathbf{W}_{i,j-1}^{n}}{\Delta y^2}\right)\Bigg]\end{aligned} \tag{21}$$

$$\begin{aligned}\mathbf{W}_{i,j}^{n+1} = \frac{1}{2}\left(\mathbf{W}_{i,j}^{n} + \widetilde{\mathbf{W}}_{i,j}\right) + \frac{\Delta t}{2}\ \boldsymbol{D}_\beta^{-1}\Bigg[&- \frac{\widetilde{\mathbf{F}}_{i+1,j}^{n} - \widetilde{\mathbf{F}}_{i,j}^{n}}{\Delta x} - \frac{\widetilde{\mathbf{G}}_{i,j+1}^{n} - \widetilde{\mathbf{G}}_{i,j}^{n}}{\Delta y} + \\ &+ \nu\, \boldsymbol{D}\left(\frac{\widetilde{\mathbf{W}}_{i+1,j}^{n} - 2\widetilde{\mathbf{W}}_{i,j}^{n} + \widetilde{\mathbf{W}}_{i-1,j}^{n}}{\Delta x^2} + \frac{\widetilde{\mathbf{W}}_{i,j+1}^{n} - 2\widetilde{\mathbf{W}}_{i,j}^{n} + \widetilde{\mathbf{W}}_{i,j-1}^{n}}{\Delta y^2}\right)\Bigg]\end{aligned} \tag{22}$$

It's good to note, that the choice of backward differencing in predictor and forward differencing in corrector steps was arbitrary, and it could have been chosen the other way round and shouldn't affect the results [11, 1].

The nonconservative version of the MacCormack scheme is again rather written in component vise form separately for unknown pressure and velocity, so instead of predictor in form (21), we obtain (23), (24), and (25):

$$\widetilde{p}_{i,j} = p^n_{i,j} - \beta^2 \rho\, \Delta t \left[\frac{u^n_{i,j} - u^n_{i-1,j}}{\Delta x} + \frac{v^n_{i,j} - v^n_{i,j-1}}{\Delta y} \right] \tag{23}$$

$$\begin{aligned} \widetilde{u}_{i,j} &= u^n_{i,j} + \\ &+ \Delta t \left[- u^n_{i,j} \frac{u^n_{i,j} - u^n_{i-1,j}}{\Delta x} - v^n_{i,j} \frac{u^n_{i,j} - u^n_{i,j-1}}{\Delta y} - \frac{1}{\rho} \frac{p^n_{i,j} - p^n_{i-1,j}}{\Delta x} + \right. \\ &\left. + \nu \left(\frac{u^n_{i+1,j} - 2u^n_{i,j} + u^n_{i-1,j}}{\Delta x^2} + \frac{u^n_{i,j+1} - 2u^n_{i,j} + u^n_{i,j-1}}{\Delta y^2} \right) \right] \end{aligned} \tag{24}$$

$$\begin{aligned} \widetilde{v}_{i,j} &= v^n_{i,j} + \\ &+ \Delta t \left[- u^n_{i,j} \frac{v^n_{i,j} - v^n_{i-1,j}}{\Delta x} - v^n_{i,j} \frac{v^n_{i,j} - v^n_{i,j-1}}{\Delta y} - \frac{1}{\rho} \frac{p^n_{i,j} - p^n_{i,j-1}}{\Delta y} + \right. \\ &\left. + \nu \left(\frac{v^n_{i+1,j} - 2v^n_{i,j} + v^n_{i-1,j}}{\Delta x^2} + \frac{v^n_{i,j+1} - 2v^n_{i,j} + v^n_{i,j-1}}{\Delta y^2} \right) \right] \end{aligned} \tag{25}$$

These predicted fields $\widetilde{p}$, $\widetilde{u}$, and $\widetilde{v}$ are updated in the corrector step, where instead of (22) we obtain (26), (27), and (28):

$$p^{n+1}_{i,j} = \frac{1}{2} \left(p^n_{i,j} + \widetilde{p}_{i,j} \right) - \beta^2 \rho \frac{\Delta t}{2} \left[\frac{\widetilde{u}_{i+1,j} - \widetilde{u}_{i,j}}{\Delta x} + \frac{\widetilde{v}_{i,j+1} - \widetilde{v}_{i,j}}{\Delta y} \right] \tag{26}$$

$$\begin{aligned} u^{n+1}_{i,j} &= \frac{1}{2} \left(u^n_{i,j} + \widetilde{u}_{i,j} \right) + \\ &+ \frac{\Delta t}{2} \left[- \widetilde{u}_{i,j} \frac{\widetilde{u}_{i+1,j} - \widetilde{u}_{i,j}}{\Delta x} - \widetilde{v}_{i,j} \frac{\widetilde{u}_{i,j+1} - \widetilde{u}_{i,j}}{\Delta y} - \frac{1}{\rho} \frac{\widetilde{p}_{i+1,j} - \widetilde{p}_{i,j}}{\Delta x} + \right. \\ &\left. + \nu \left(\frac{\widetilde{u}_{i+1,j} - 2\widetilde{u}_{i,j} + \widetilde{u}_{i-1,j}}{\Delta x^2} + \frac{\widetilde{u}_{i,j+1} - 2\widetilde{u}_{i,j} + \widetilde{u}_{i,j-1}}{\Delta y^2} \right) \right] \end{aligned} \tag{27}$$

$$\begin{aligned} v^{n+1}_{i,j} &= \frac{1}{2} \left(v^n_{i,j} + \widetilde{v}_{i,j} \right) + \\ &+ \frac{\Delta t}{2} \left[- \widetilde{u}_{i,j} \frac{\widetilde{v}_{i+1,j} - \widetilde{v}_{i,j}}{\Delta x} - \widetilde{v}_{i,j} \frac{\widetilde{v}_{i,j+1} - \widetilde{v}_{i,j}}{\Delta y} - \frac{1}{\rho} \frac{\widetilde{p}_{i,j+1} - \widetilde{p}_{i,j}}{\Delta y} + \right. \\ &\left. + \nu \left(\frac{\widetilde{v}_{i+1,j} - 2\widetilde{v}_{i,j} + \widetilde{v}_{i-1,j}}{\Delta x^2} + \frac{\widetilde{v}_{i,j+1} - 2\widetilde{v}_{i,j} + \widetilde{v}_{i,j-1}}{\Delta y^2} \right) \right] \end{aligned} \tag{28}$$

3.2.4 Numerical Stabilization

The MacCormack scheme belongs to the family of central, second-order schemes that are prone to numerical oscillations in the presence of sharp solution gradients. In the presented immersed boundary method implementation, such nonphysical numerical oscillations often appear at the fluid-solid interface, i.e., in the proximity of the channel walls. In order to keep these numerical oscillations under control, a fourth-order artificial viscosity is implemented in the code, applied just to pressure field. The added smoothing (regularization) term has the form $\varepsilon h^4 \Delta^2 p$, where h stands for the grid cell size and $\Delta^2 p$ is the bi-Laplacian of pressure containing the fourth-order spatial derivatives. This numerical stabilization process can be added as an extra smoothing step in the algorithm, after the corrector step. So the final smoothed values $\hat{p}_{i,j}^{n+1}$ at the time level $n+1$ are obtained as:

$$\hat{p}_{i,j}^{n+1} = p_{i,j}^{n+1} + D^4 p_{i,j}^{n} \quad . \tag{29}$$

The added numerical viscosity has the form $D^4 p_{i,j}^n = D_x^4 p_{i,j}^n + D_y^4 p_{i,j}^n, \quad$ with :

$$\begin{aligned} D_x^4 p_{i,j}^n &= \varepsilon\,(p_{i-2,j}^n - 4p_{i-1,j}^n + 6p_{i,j}^n - 4p_{i+1,j}^n + p_{i+2,j}^n) \quad , \\ D_y^4 p_{i,j}^n &= \varepsilon\,(p_{i,j-2}^n - 4p_{i,j-1}^n + 6p_{i,j}^n - 4p_{i,j+1}^n + p_{i,j+2}^n) \quad . \end{aligned} \tag{30}$$

The coefficient ε should be suitably chosen for the optimal smoothing properties. Typically, the fourth-order stabilization is aimed at suppression of high-frequency (point-to-point) spatial oscillations and is less strong than the more commonly used second-order numerical diffusion stabilization. For more details, see, for example, [12] or the discussion of numerical diffusion in [5].

3.3 *Finite Volume Solver*

The finite volume discretization is used in waste majority of commercial CFD solvers. In this work, it was used as a reference for comparison and validation of the newly developed finite difference solver. Within the presented study, the finite volume method was used in two codes. First is an in-house developed simple 2D code, while second is the more general open-source package OpenFOAM. Here, we only describe our own finite volume code, while we refer the reader to the online documentation of the OpenFOAM solver (www.openfoam.com).

In finite volume method, we usually start directly from the conservative form of governing equations, i.e., from (1) or (6), respectively. In most cases, the inviscid and viscous fluxes are being treated separately, each using specific discretization. For example, upwinding or some higher-order reconstruction is applied in the discretization of inviscid fluxes. In our case, however, the same central scheme is used for both inviscid and viscous terms discretization, and thus the governing system (6) can be rewritten in divergence-like form:

$$\boldsymbol{D}\mathbf{W}_t + (\mathbf{F} - \mathbf{R})_x + (\mathbf{G} - \mathbf{S})_y + (\mathbf{H} - \mathbf{T})_z = 0 \quad . \tag{31}$$

Considering the artificial compressibility method (i.e., replacing $\boldsymbol{D}$ by $\boldsymbol{D}_\beta$), it takes the form:

$$\mathbf{W}_t = -\boldsymbol{D}_\beta \left[(\mathbf{F} - \mathbf{R})_x + (\mathbf{G} - \mathbf{S})_y + (\mathbf{H} - \mathbf{T})_z\right] \quad . \tag{32}$$

Applying the finite volume discretization on the spatial derivatives on the right-hand side of (32), a semi-discrete system is obtained, consisting of ordinary differential equations for time evolution of approximate values of $\mathbf{W}$ at individual grid cells.

In two-dimensional case (using structured grid), it is obtained by integrating (32) over each grid cell $\Omega_{i,j}$, using Green's theorem on the right-hand side:

$$\int_{\Omega_{i,j}} \mathbf{W}_t d\Omega = -\boldsymbol{D}_\beta \oint_{\partial\Omega_{i,j}} \left[(\mathbf{F} - \mathbf{R})n_x + (\mathbf{G} - \mathbf{S})n_y\right] dS \quad , \tag{33}$$

and defining the cell (averaged) value $\mathbf{W}_{i,j} = \frac{1}{|\Omega_{i,j}|} \int_{\Omega_{i,j}} \mathbf{W}\, d\Omega$, which leads to:

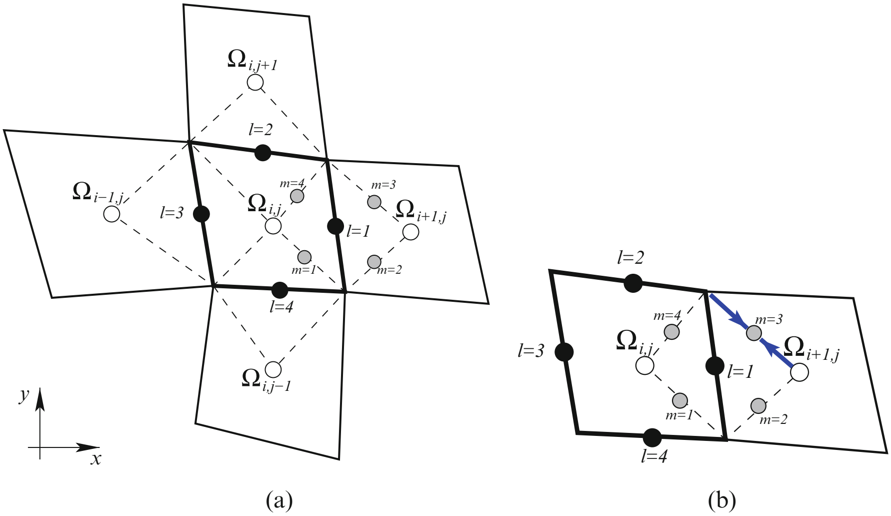
$$\frac{d\mathbf{W}_{i,j}}{dt} = -\underbrace{\frac{\boldsymbol{D}_\beta}{|\Omega_{i,j}|} \oint_{\partial\Omega_{i,j}} \left[(\mathbf{F} - \mathbf{R})n_x + (\mathbf{G} - \mathbf{S})n_y\right] dS}_{\mathcal{L}\mathbf{W}_{i,j}} \ . \tag{34}$$

The spatial discretization operator $\mathcal{L}\mathbf{W}_{i,j}$ approximates the integral over the cell boundary $\partial\Omega_{i,j}$. The values of inviscid fluxes are simply interpolated from the neighboring cell centers to the boundary. The approximate values of viscous fluxes at boundary points can be obtained by Green's theorem for integration over the boundary of dual (diamond-shaped) control volumes, using both centroids and vertices of the primary grid (see Fig. 3). The result in semi-discrete system of ODE's has the form:

$$\frac{d\mathbf{W}_{i,j}}{dt} = -\mathcal{L}\mathbf{W}_{i,j} \tag{35}$$

and can be solved, for example, by a Runge-Kutta multistage method:

$$\begin{aligned}
\mathbf{W}^{(0)}_{i,j} &= \mathbf{W}^{n}_{i,j} \ , \\
\mathbf{W}^{(r+1)}_{i,j} &= \mathbf{W}^{(0)}_{i,j} - \alpha_{(r)} \Delta t \mathcal{L}\mathbf{W}^{(r)}_{i,j} \ , \qquad r = 1, \ldots, s \\
\mathbf{W}^{n+1}_{i,j} &= \mathbf{W}^{(s)}_{i,j} \ .
\end{aligned} \tag{36}$$



(a) (b)

Fig. 3 Grid configuration for approximation of inviscid and viscous fluxes. (**a**) Primary and secondary cells. (**b**) Approximation of viscous fluxes

This specific method belongs to the family of low-storage methods, where always only the values from the previous one stage are required. This is a great advantage in case of large, high-resolution simulations, where memory efficiency can be limiting. The three-stage explicit RK scheme used to obtain results presented here had coefficients $\alpha_{(1)} = 1/2$, $\alpha_{(2)} = 1/2$, $\alpha_{(3)} = 1$. More details on this type of finite volume discretization and associated Runge-Kutta methods can be found, for example, in [13, 3, 15, 2].

4 Numerical Simulations

The numerical simulations performed in this study had two main objectives: first, to compare the outputs of different methods and codes to verify that the newly developed immersed boundary code is sufficiently accurate and robust and, second, to investigate the flow in the branching area of the channel depending on the angle of attachment of the secondary branch.

4.1 Test Case Description

4.1.1 Domain Geometry

For the immersed boundary implementation of finite difference method, the two-dimensional (2D) computational domain was chosen as a rectangle in $x - y$ plane with dimensions $30D \times 10D$. The numerical simulations were performed on a structured, orthogonal (Cartesian) grid with different number of equidistant nodes. The standard grid had 1200×200 cells.

The coordinate system has been chosen to have the origin at the edge of branching. The whole domain as well as the channel geometry is shown in Fig. 4. The width (diameter) of the main horizontal channel is denoted by D, and the width of the inclined (oblique) branch inclined at the angle α was chosen to be $D/2$. The same configuration was kept for all simulations, just changing the angle α by setting it to values 30°, 60°, 90°, 120°, and 150°.

For finite volume simulations, just the interior of the channel (marked by white color in Fig. 4) was used to construct the grid. See Sect. 3.2.1 for details.

4.1.2 Boundary Conditions

The boundary setup was chosen as simple as possible in order to allow for easy extension for non-Newtonian and turbulent flows, where the velocity profiles are not a priori known even for the simple channel with Poiseuille-like flow. Therefore, we have opted for the flow to be defined by pressure drop to be prescribed between inlet and outlet parts of the boundary. So only different values of pressure were prescribed at different inlet/outlet parts of the boundary. Otherwise, the homogeneous Neumann condition was prescribed for velocity components on those

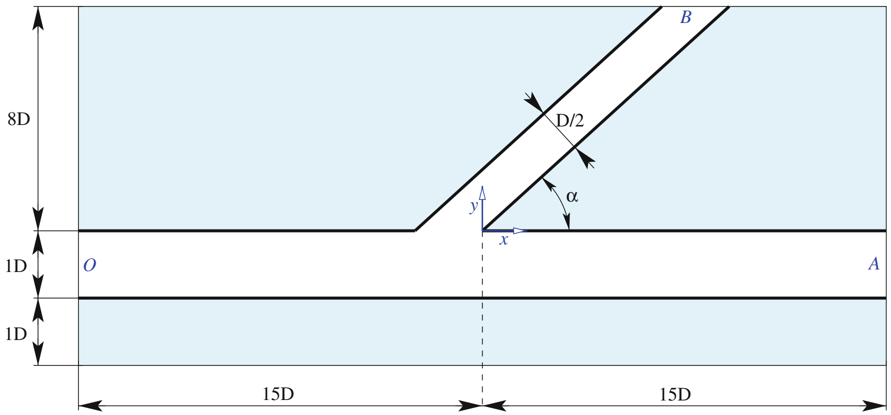


Fig. 4 Computational domain of a branched pipe

parts of boundary to mimic a fully developed flow. On the channel wall of course, the no-slip, i.e., homogeneous, Dirichlet condition $\boldsymbol{u} = (0, 0)$ was prescribed for velocity.

To establish a pressure drop between inlet and outlet of the channel, the pressure was set to zero at outlet, and a suitable (positive) value of pressure was set at the inlet. In order to achieve some realistic flow conditions, we have chosen all conditions similar to flow of blood in common carotid artery. From its diameter $D = 6\,\mathrm{mm}$, fluid viscosity, and channel length (without considering the branch), it was possible to find suitable inlet pressure from Poiseuille solution, to achieve similar velocity as it's found in the real blood vessel of the same size. This led us to the choice of inlet pressure $p_{in} = 60\,\mathrm{Pa}$, while at the outlets we prescribed pressures $p_A = p_B = 0$.

For the immersed boundary method, the velocity is set to zero inside the part of the domain occupied by the solid material, so no special treatment is needed on the channel walls for pressure or velocity. See Sect. 3.2.1 for details of the implementation.

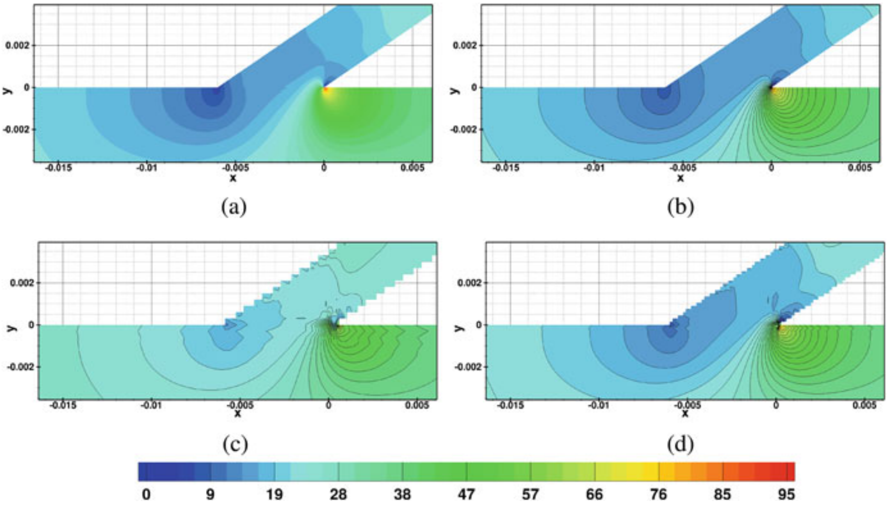
4.2 Numerical Results

The aim of presented numerical results is to demonstrate the applicability of the chosen methods and their settings for the considered class of problems. The newly developed finite difference method (FDM)-based immersed boundary code is compared with in-house finite volume method (FVM) and open-source finite volume code OpenFOAM. Both FVM methods share the same grid. For the FDM method with immersed boundary channel representation, two different grids were used. The standard coarser grid had resolution 1200×200 cells, while the finer grid had doubled of the cells in the vertical y direction, i.e., having 1200×400 cells.

Figures 5, 6, and 7 show the comparison of pressure and velocity fields obtained using all the considered codes for the case of oblique branching at angle $\alpha = 30°$. The pressure fields in Fig. 5 have very similar character, and except the FDM results on coarse grid, all results are almost identical. The comparison of horizontal velocity fields in Fig. 6 reveals that the in-house FVM code and FDM code on finer grid provide almost identical results. The OpenFOAM results predict a bit higher velocity in the main channel, while the FDM code on coarse grid predicts lower velocity. The same level of agreement between the results can also be seen in the comparison of vertical velocity fields for the same case shown in Fig. 7.

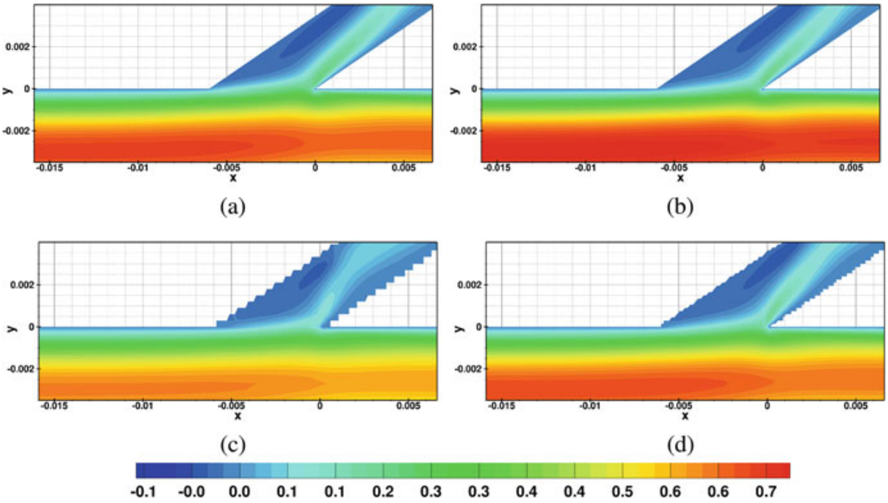
It is interesting to see that the level of agreement between the results changes for different angles α of the secondary branch. The comparison of pressure and velocity fields in the case of $\alpha = 60°$ is shown in Figs. 8, 9, and 10. Here, it seems the OpenFOAM results are closest to the FDM on finer grid (see the comparison of the horizontal velocity contours in Fig. 9).

The comparison of results in the case of $\alpha = 90°$ (shown in Figs. 11, 12, and 13) shows that even the results obtained by FDM on the coarse grid are almost



(a) (b) (c) (d)

Fig. 5 Pressure field in detail for the case $\alpha = 30°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

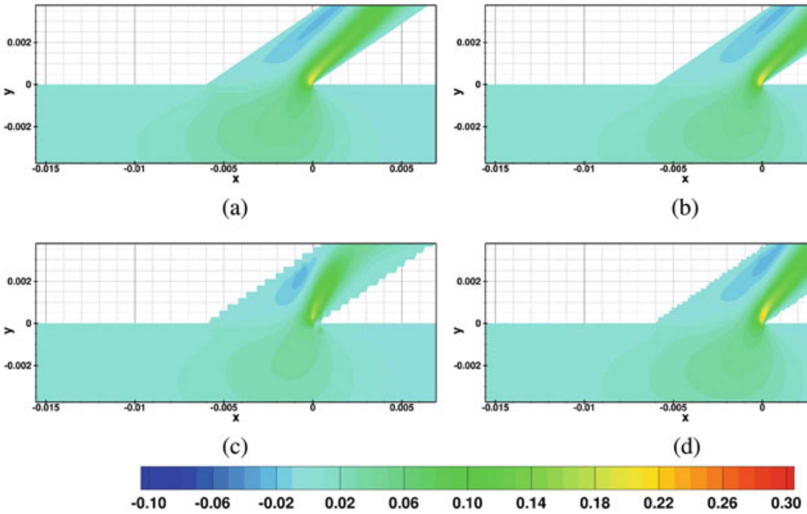


(a) (b) (c) (d)

Fig. 6 Horizontal velocity u in detail for the case $\alpha = 30°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

identical to other methods. The orthogonality of the grid allows for optimal use of all computational points and leads to highest accuracy of numerical approximation.

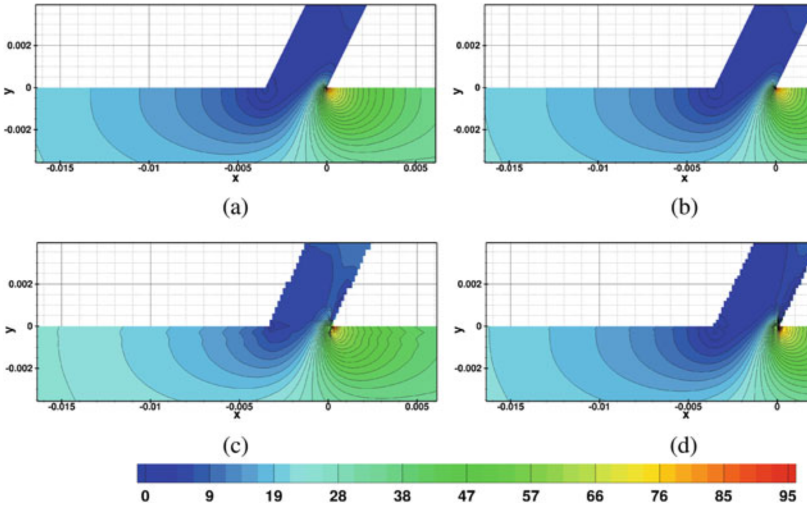
All the three mentioned cases, i.e., for $\alpha = 30°, 60°, 90°$ can be compared according to the velocity profiles presented in Fig. 14 showing the velocity magnitude at the inlet and outlet sections of the main channel and the outlet section of the secondary branch. In general, the results are in a very good agreement. The



(a) (b)

(c) (d)

Fig. 7 Vertical velocity v in detail for the case $\alpha = 30°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid



(a) (b)

(c) (d)

Fig. 8 Pressure field in detail for the case $\alpha = 60°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

FDM method works better on finer grid (as expected). The main differences between the results are at the outlet of the secondary branch where the profiles differ most. This is probably because the details of the implementation of boundary conditions differ for each code. For FDM methods, the variables are extrapolated along the grid lines, i.e., in the y direction. For the in-house FVM code, such extrapolation is

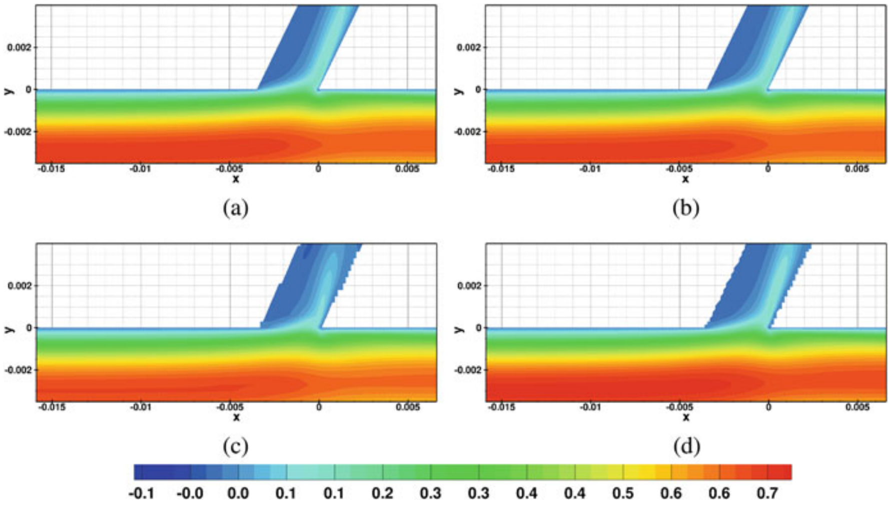


Fig. 9 Horizontal velocity u in detail for the case $\alpha = 60°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

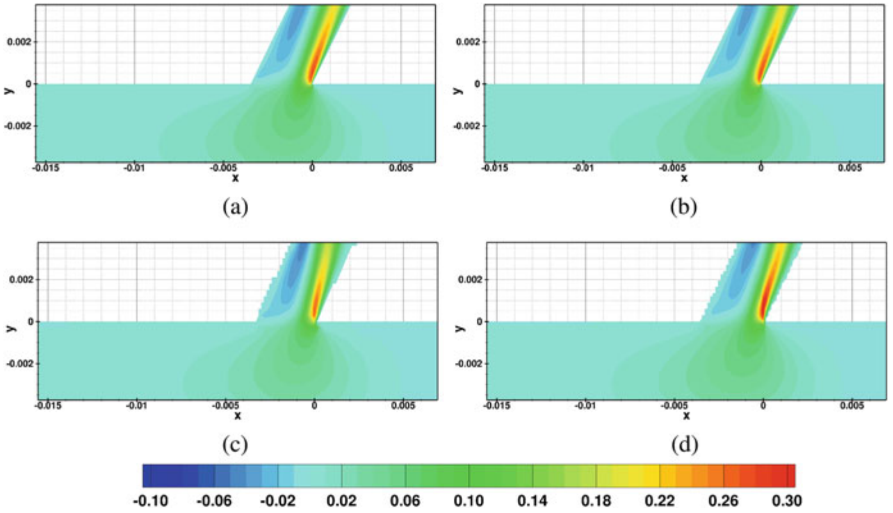


Fig. 10 Vertical velocity v in detail for the case $\alpha = 60°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

performed to ghost cells which are constructed as a prolongation of the actual grid close to the boundary. It results in extrapolation along the oblique grid lines which are parallel to the walls of the branch. On the other hand, the OpenFOAM technique imposes the normal derivative directly on the cell boundary face, without the need to construct any ghost cells. Thus, certain local small differences in the obtained

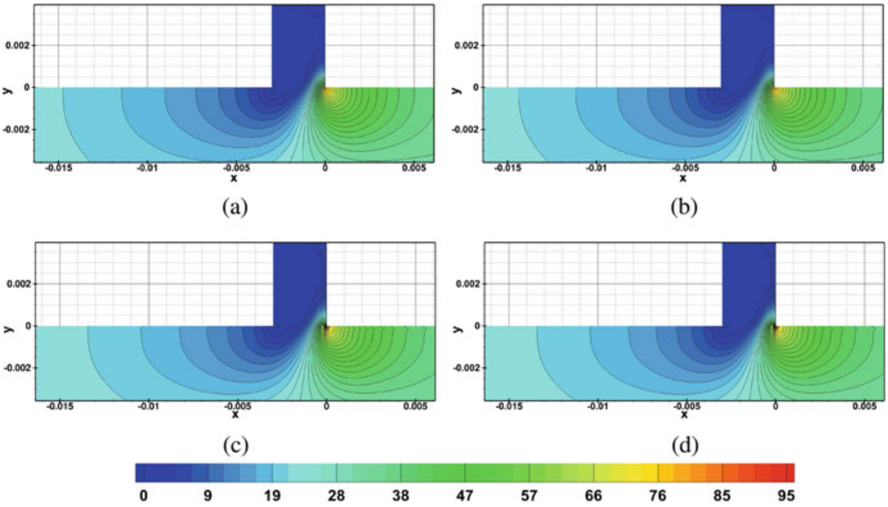


Fig. 11 Pressure field in detail for the case $\alpha = 90°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

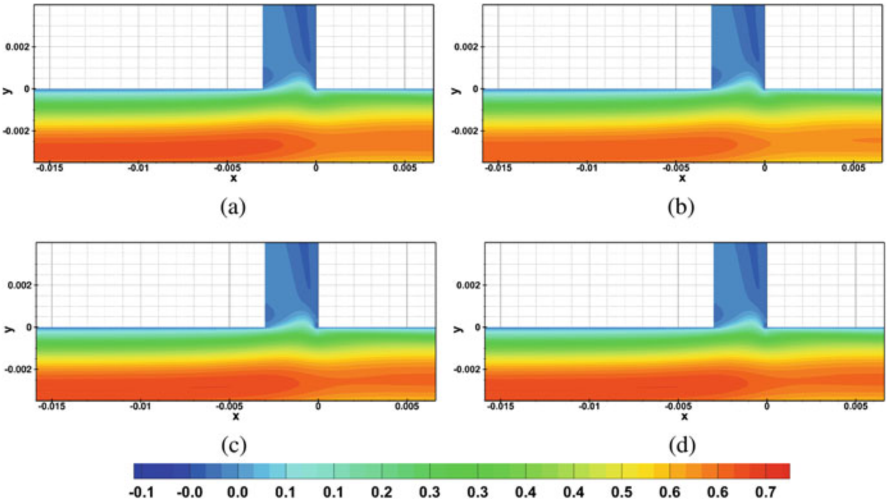


Fig. 12 Horizontal velocity u in detail for the case $\alpha = 90°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

solutions are expected. Largest differences are again observed for the case $\alpha = 30°$ where the grids differ most and thus also the differences in the implementation of boundary conditions become more important. This, however, doesn't seems to have any significant effect on the main flow features. For the comparison of profiles including the angles $\alpha = 120°, 150°$, see Fig. 14.

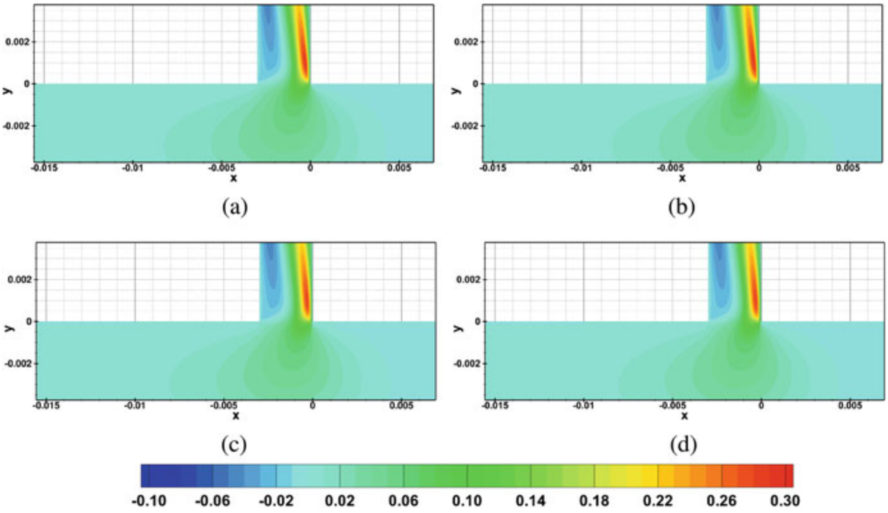
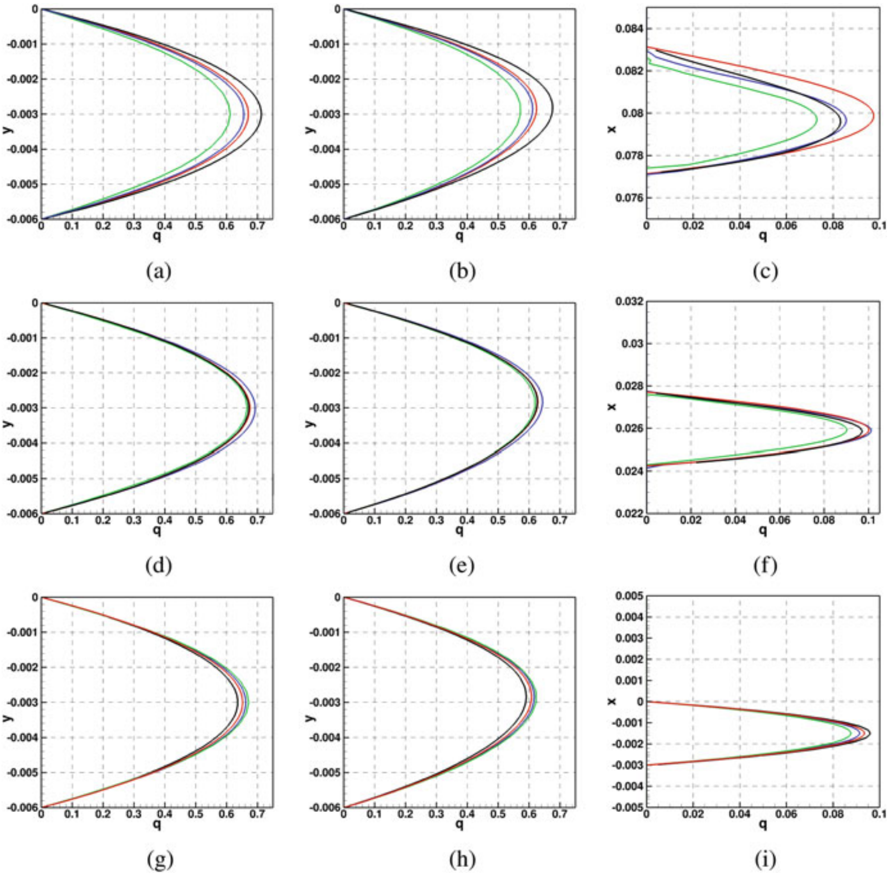


Fig. 13 Vertical velocity v in detail for the case $\alpha = 90°$, different solvers, and grids. (**a**) FVM in-house. (**b**) FVM OpenFOAM. (**c**) FDM coarse grid. (**d**) FDM finer grid

From now on, we will only focus on the comparison of our in-house FVM code and the immersed boundary FDM code on finer grid (unless specified otherwise). The pressure field for the case $\alpha = 30°$ is shown in Fig. 15, with FVM results in the left column and FDM results in the right column. The global view at the complete channel shows that in the main channel as well as in the secondary (oblique) branch, the pressure field (away from the branching region) behaves like in Poiseuille flow, with linear pressure distribution along the channel axis. The detailed look at the branching region shows that the pressure is very well captured by the FDM method, despite of quite rough representation of the oblique branch walls by the simple implementation of the immersed boundary method. Also, the comparison of velocity components (Fig. 16), velocity magnitude, and streamlines (Fig. 17) shows very good agreement between the results obtained using both, FVM and FDM methods.

Closer look at the performance of the FDM method on coarse and finer grid is provided in Fig. 18 for pressure fields and velocity magnitude in Fig. 19. Apparently, for $\alpha = 90°$, the results on the coarse and finer grid are almost identical, but the more the branch deviates from this ideal position, the differences between the results become more apparent.

Similar comparison for cases with different branching angle α is shown in Figs. 20 and 21 for finite volume method (in the left column) and finite difference method (in the right column). Also here, the mutual agreement between the two methods depends on the angle α, with best results (smallest solution differences) achieved for angles close to $\alpha = 90°$, while in the cases $\alpha = 30°$ and $\alpha = 150°$ the differences are more pronounced. The comparison of streamlines for FVM and FDM solutions (shown in Fig. 22) shows that both methods captured properly the



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 14 Profiles of velocity magnitude for various branching angles α. In-house FVM code red em dashed hypen, OpenFOAM FVM code black em dashed hypen, FDM code on coarse grid green em dashed hypen, FDM code on finer grid blue em dashed hypen. (**a**) $\alpha = 30°$—channel inlet. (**b**) $\alpha = 30°$—channel outlet. (**c**) $\alpha = 30°$—branch outlet. (**d**) $\alpha = 60°$—channel inlet. (**e**) $\alpha = 60°$—channel outlet. (**f**) $\alpha = 60°$—branch outlet. (**g**) $\alpha = 90°$—channel inlet. (**h**) $\alpha = 90°$—channel outlet. (**i**) $\alpha = 90°$—branch outlet

vortices in regions of separated flow. There are no visible differences, and only close inspection can reveal some small shifts in the position of reattachment points.

The comparison of velocity profiles shown in Fig. 23 confirms the overall very good mutual agreement between the results obtained by different codes. In some cases, the results are almost identical (so it seems some profile line is missing, but it's not). The main differences appear at the outlet from the secondary branch, at the angles far from the case $\alpha = 90°$, where the grids are optimal and orthogonal. The performance of the FDM method can be substantially be improved by refining the grid. But in any case, even in the present version of the grid, the simple

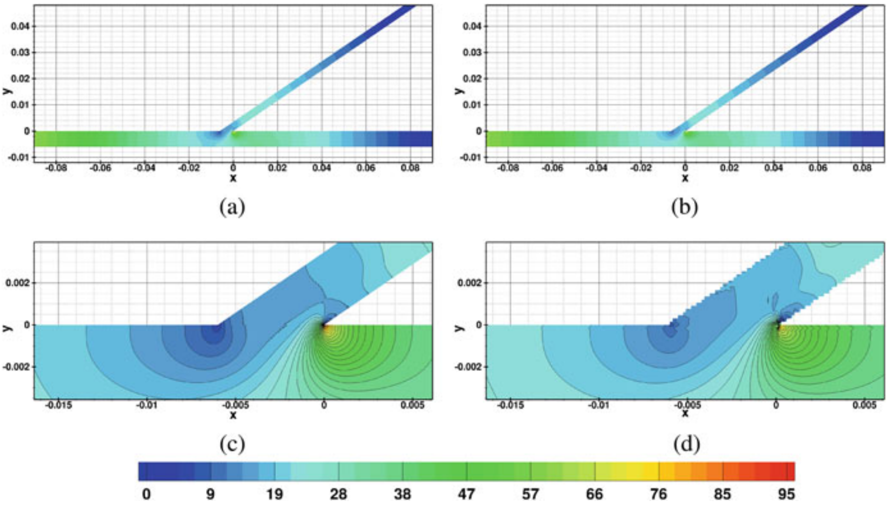


Fig. 15 Pressure field for the case $\alpha = 30°$—finer grid simulations. (**a**) Pressure—whole channel—FVM. (**b**) Pressure—whole channel—FDM. (**c**) Pressure—detail—FVM. (**d**) Pressure—detail—FDM

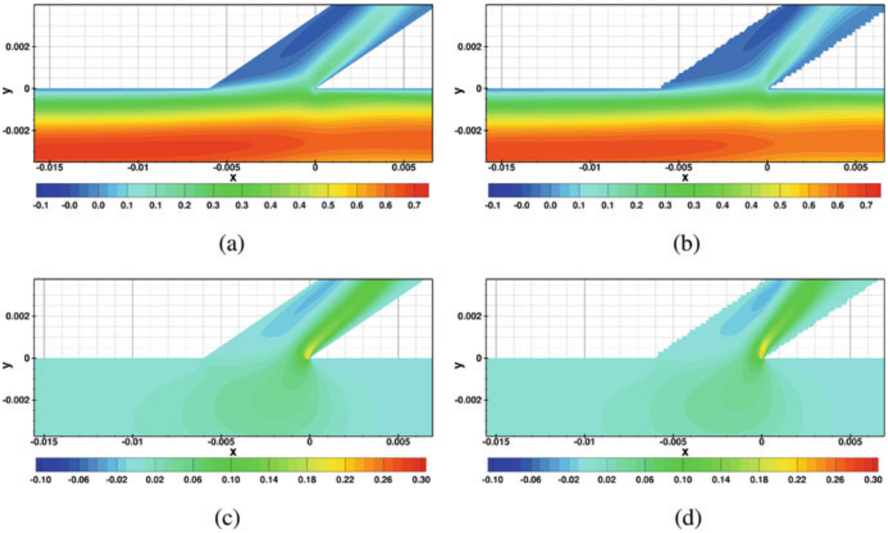
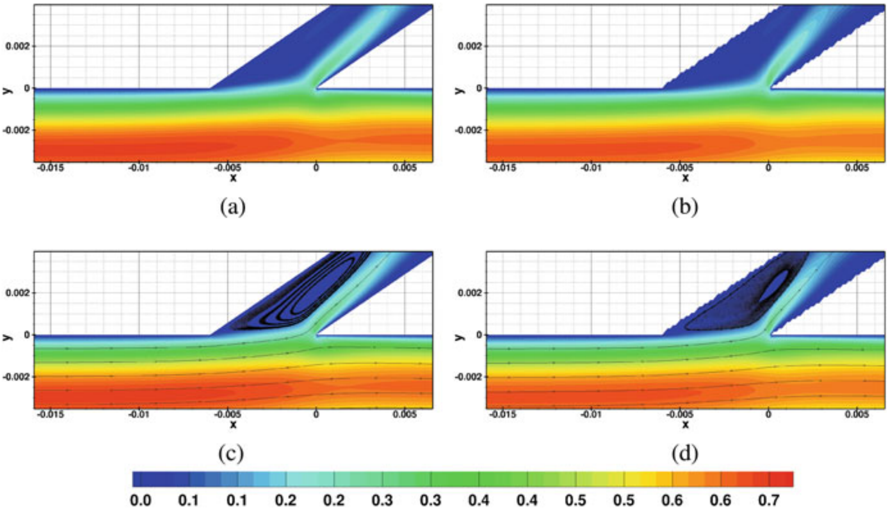


Fig. 16 Velocity field for the case $\alpha = 30°$—finer grid simulations. (**a**) Horizontal velocity u—detail—FVM. (**b**) Horizontal velocity u—detail—FDM. (**c**) Vertical velocity v—detail—FVM. (**d**) Vertical velocity v—detail—FDM



(a)

(b)

(c)

(d)

Fig. 17 Velocity magnitude and streamlines for the case $\alpha = 30°$—finer grid simulations. (**a**) Velocity magnitude—detail—FVM. (**b**) Velocity magnitude—detail—FDM. (**c**) Streamlines—detail—FVM. (**d**) Streamlines—detail—FDM

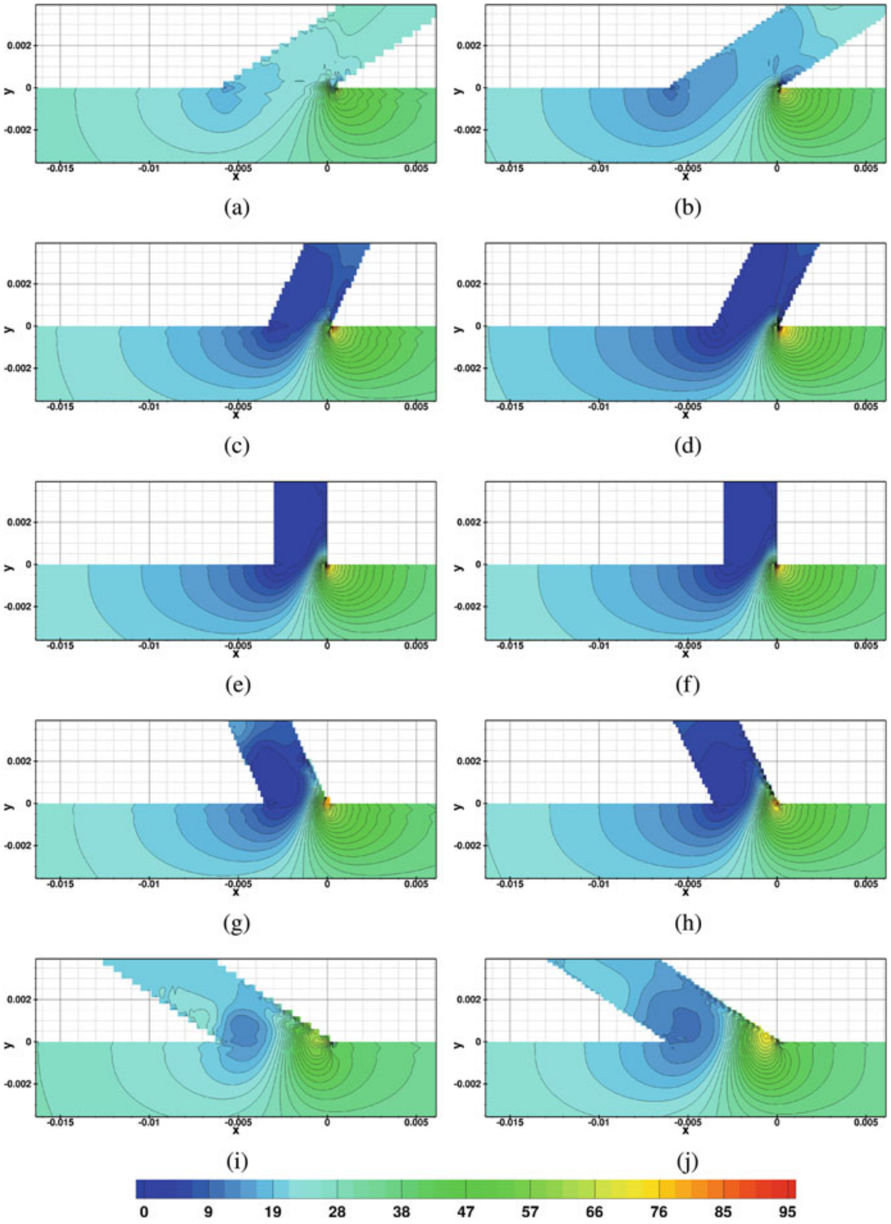
immersed boundary FDM method performs very well, which is sufficient for the future intended tests of reduced order model boundary conditions.

5 Conclusions and Remarks

The main aim of this work was to develop and validate simple finite difference code, employing immersed boundary method, to simulate the flow of viscous fluid flow in branching channels. The new code is intended for future testing of some nonstandard boundary conditions based on reduced order models. The presented series of numerical simulations clearly showed despite of the (intentional) simplicity of the chosen finite difference discretization and grid, the results provided by the code are on par with the outputs of more advanced finite volume in-house as well as open-source alternatives.

It was found that for the FDM code working on Cartesian grid with immersed boundary method, special attention should be paid to grid resolution to properly capture all essential physical features of the flow in the oblique branches. Although the results obtained on coarse and finer grid are qualitatively very similar (showing the same flow structure), some of the quantitative parameters (like maximum velocity or flow rate) may differ.

In the presented comparison, a simple pressure-based setup was chosen, where the flow is driven only by the prescribed pressure differences between inlet and outlet boundaries of the channel branches. Such setup is very sensitive to



(a)

(b)

(c)

(d)

(e)

(f)

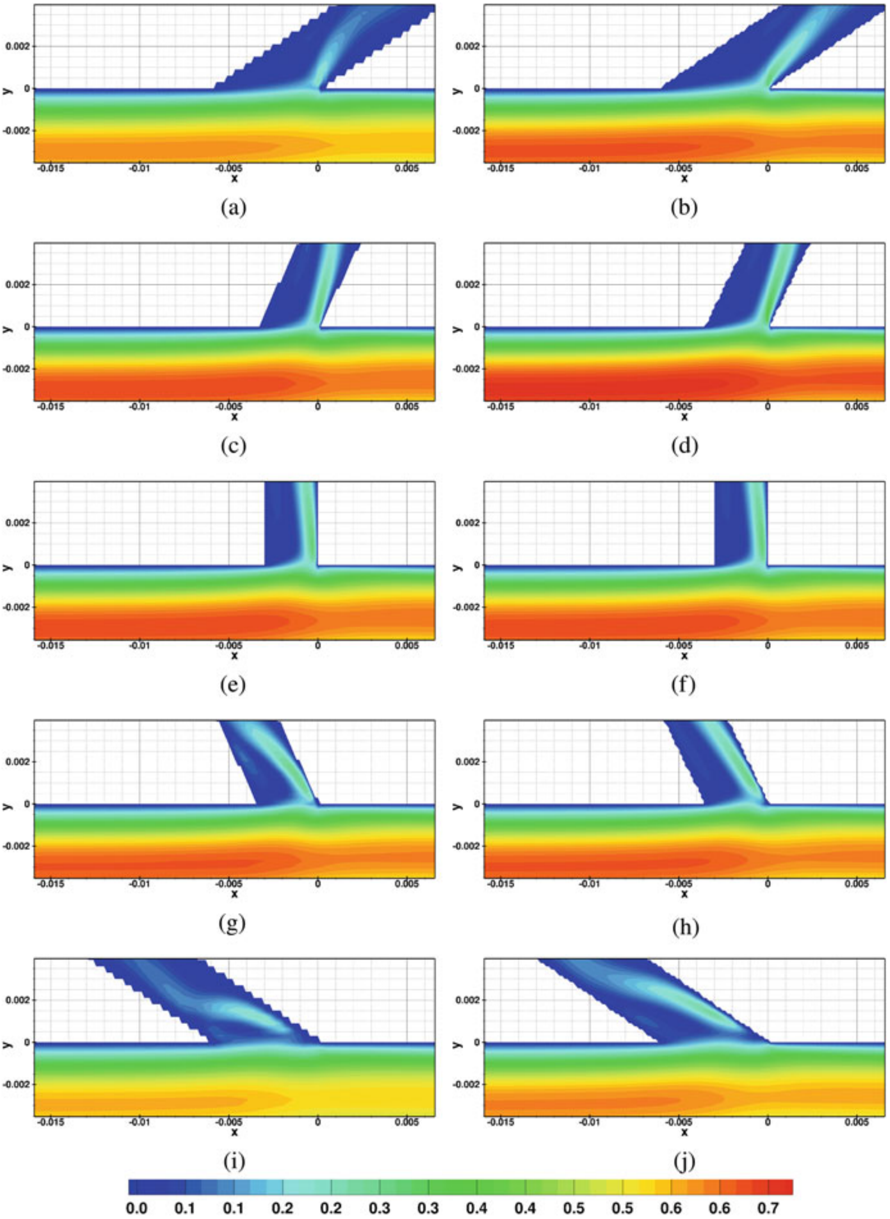
(g)

(h)

(i)

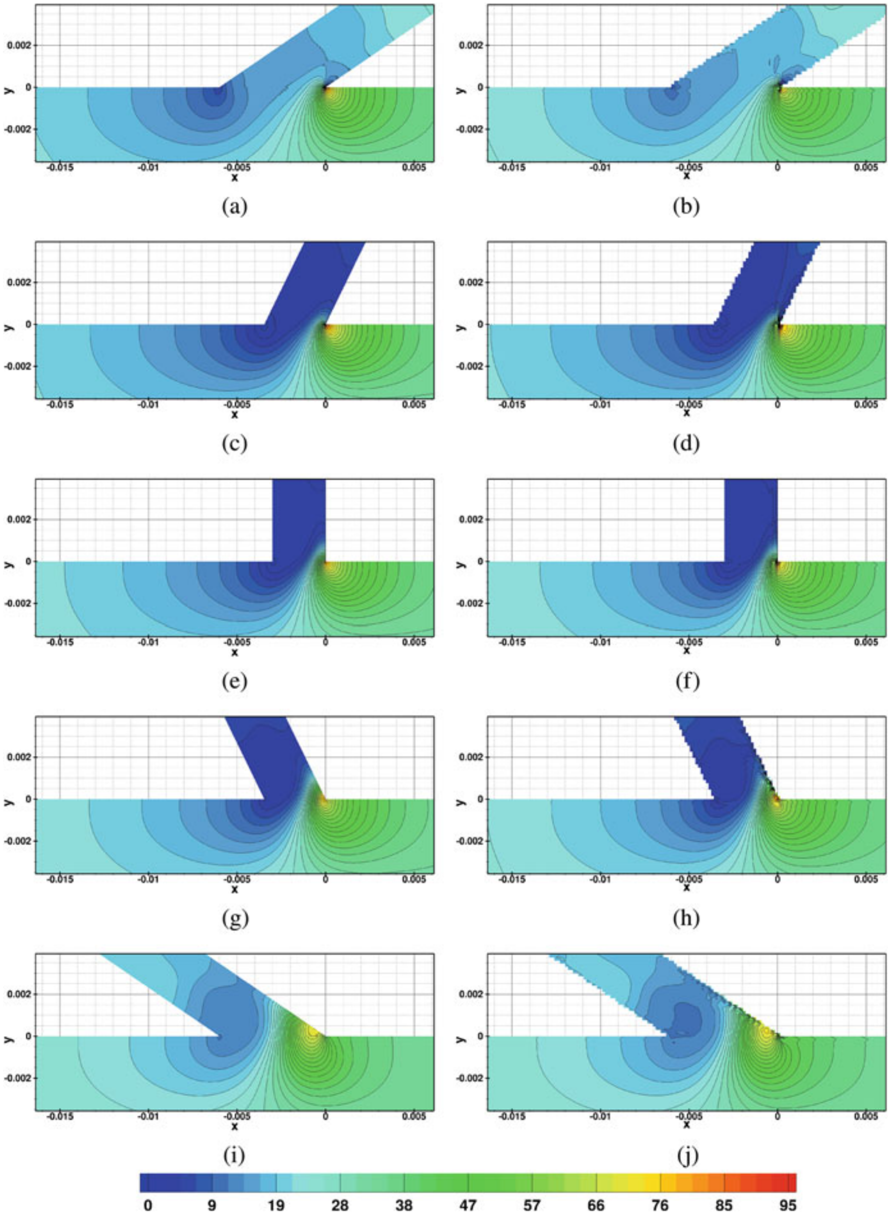
(j)

Fig. 18 Pressure field comparison for coarse and finer grid results obtained using FDM method. (**a**) $\alpha = 30°$—FDM—coarse grid. (**b**) $\alpha = 30°$—FDM—finer grid. (**c**) $\alpha = 60°$—FDM—coarse grid. (**d**) $\alpha = 60°$—FDM—finer grid. (**e**) $\alpha = 90°$—FDM—coarse grid. (**f**) $\alpha = 90°$—FDM—finer grid. (**g**) $\alpha = 120°$—FDM—coarse grid. (**h**) $\alpha = 120°$—FDM—finer grid. (**i**) $\alpha = 150°$—FDM—coarse grid. (**j**) $\alpha = 150°$—FDM—finer grid



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Fig. 19 Velocity magnitude comparison for coarse and finer grid results obtained using FDM method. (**a**) $\alpha = 30°$—FDM—coarse grid. (**b**) $\alpha = 30°$—FDM—finer grid. (**c**) $\alpha = 60°$—FDM—coarse grid. (**d**) $\alpha = 60°$—FDM—finer grid. (**e**) $\alpha = 90°$—FDM—coarse grid. (**f**) $\alpha = 90°$—FDM—finer grid. (**g**) $\alpha = 120°$—FDM—coarse grid. (**h**) $\alpha = 120°$—FDM—finer grid. (**i**) $\alpha = 150°$—FDM—coarse grid. (**j**) $\alpha = 150°$—FDM—finer grid



(a) (b)

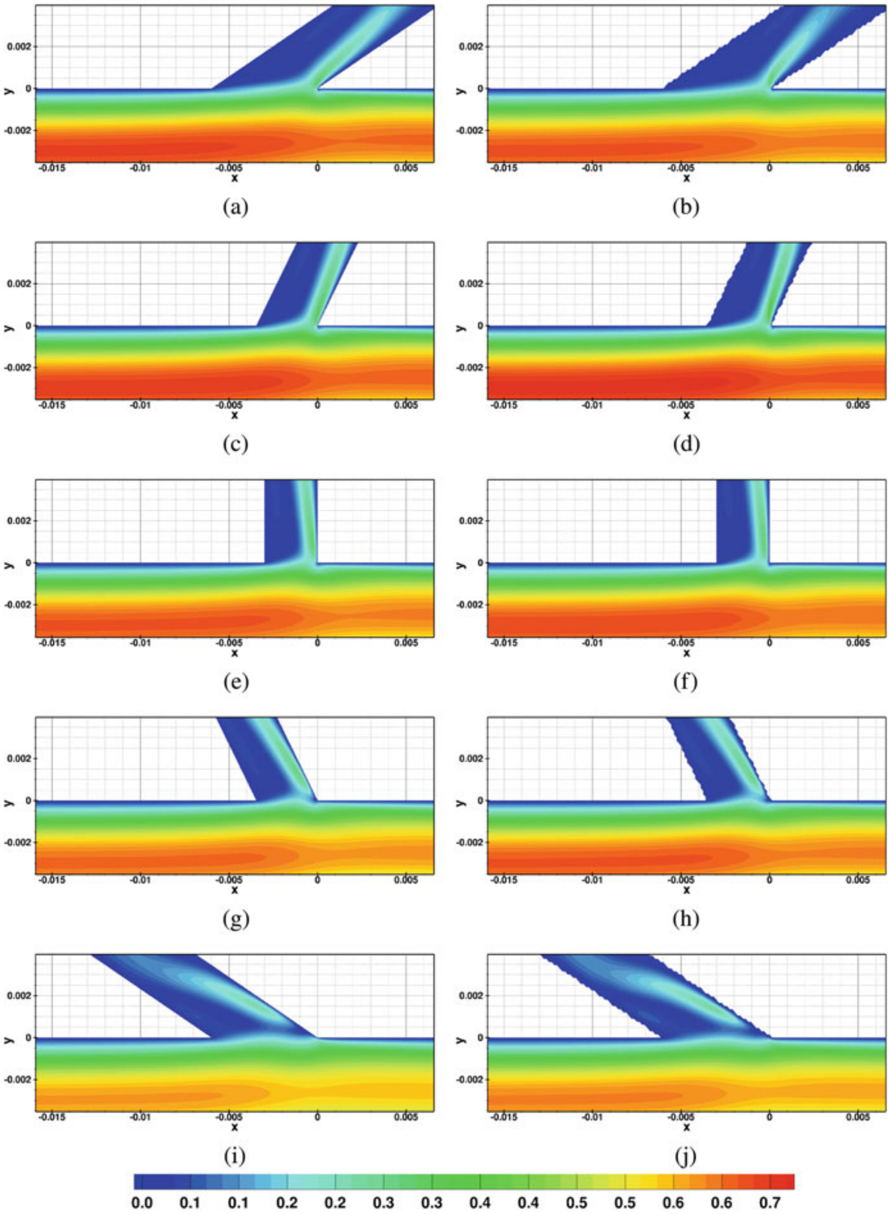
(c) (d)

(e) (f)

(g) (h)

(i) (j)

Fig. 20 Pressure field detail comparison for various branching angles α. (**a**) $\alpha = 30°$—FVM. (**b**) $\alpha = 30°$—FDM. (**c**) $\alpha = 60°$—FVM. (**d**) $\alpha = 60°$—FDM. (**e**) $\alpha = 90°$—FVM. (**f**) $\alpha = 90°$—FDM. (**g**) $\alpha = 120°$—FVM. (**h**) $\alpha = 120°$—FDM. (**i**) $\alpha = 150°$—FVM. (**j**) $\alpha = 150°$—FDM



(a)

(b)

(c)

(d)

(e)

(f)

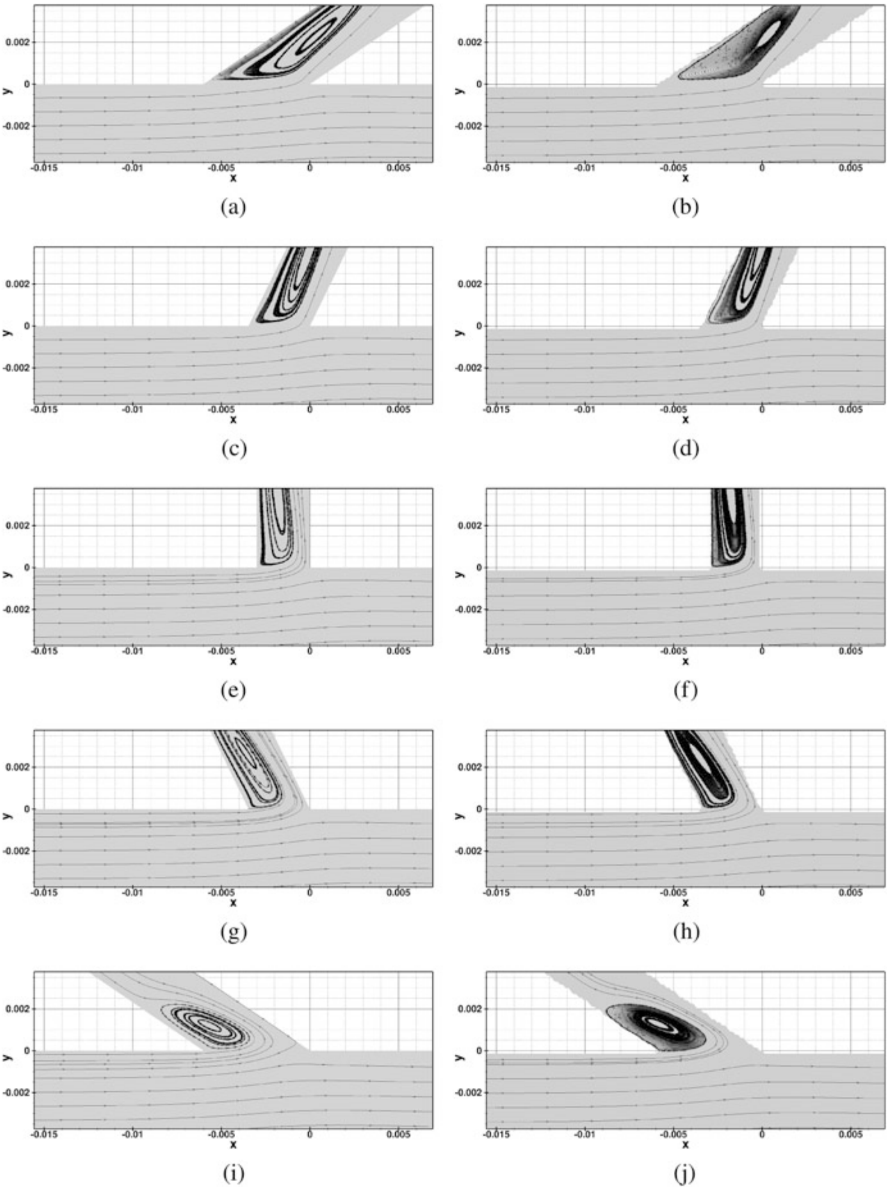
(g)

(h)

(i)

(j)

Fig. 21 Velocity magnitude detail comparison for various branching angles α. (**a**) $\alpha = 30°$—FVM. (**b**) $\alpha = 30°$—FDM. (**c**) $\alpha = 60°$—FVM. (**d**) $\alpha = 60°$—FDM. (**e**) $\alpha = 90°$—FVM. (**f**) $\alpha = 90°$—FDM. (**g**) $\alpha = 120°$—FVM. (**h**) $\alpha = 120°$—FDM. (**i**) $\alpha = 150°$—FVM. (**j**) $\alpha = 150°$—FDM



(a)

(b)

(c)

(d)

(e)

(f)

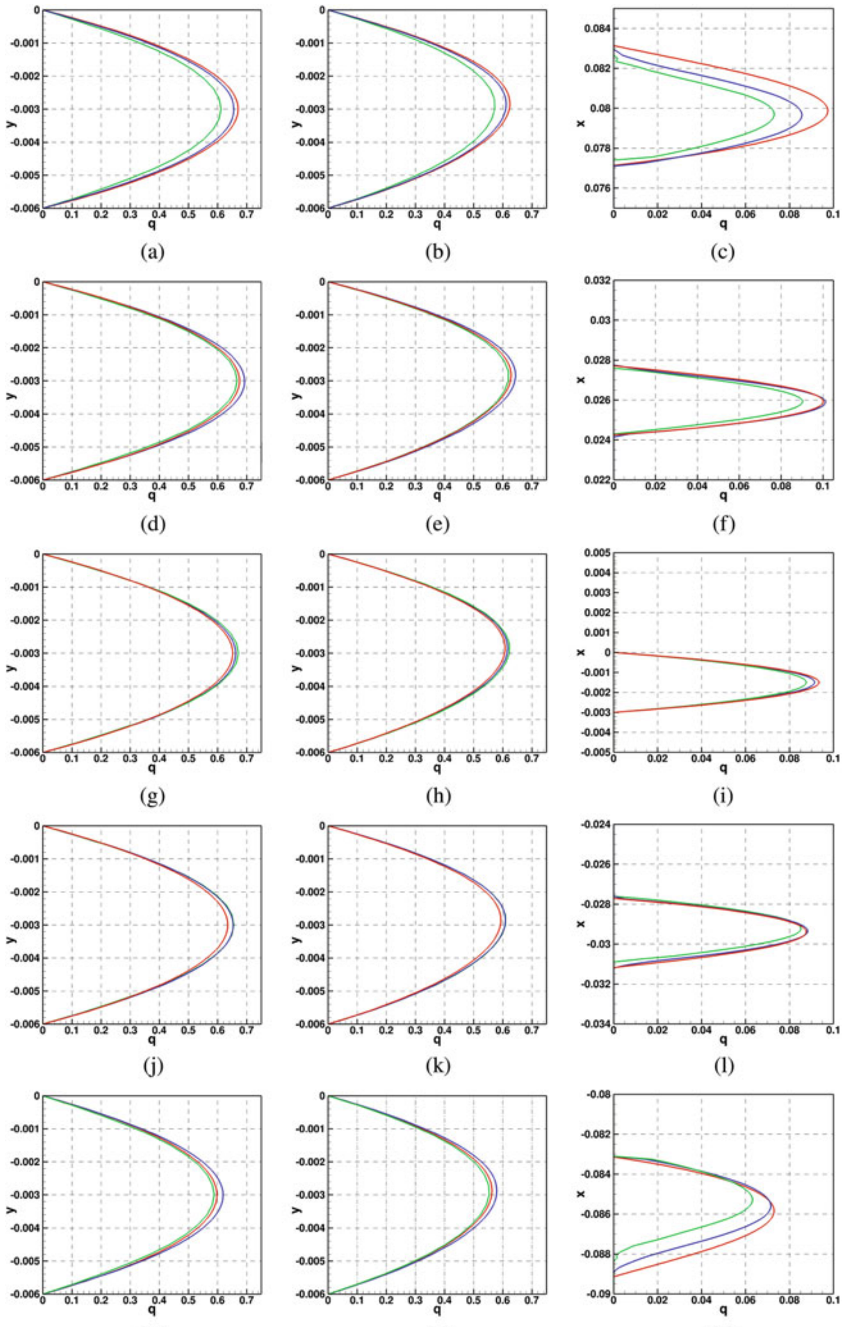
(g)

(h)

(i)

(j)

Fig. 22 Streamlines detail comparison for various branching angles α. (**a**) $\alpha = 30°$—FVM. (**b**) $\alpha = 30°$—FDM. (**c**) $\alpha = 60°$—FVM. (**d**) $\alpha = 60°$—FDM. (**e**) $\alpha = 90°$—FVM. (**f**) $\alpha = 90°$—FDM. (**g**) $\alpha = 120°$—FVM. (**h**) $\alpha = 120°$—FDM. (**i**) $\alpha = 150°$—FVM. (**j**) $\alpha = 150°$—FDM



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Fig. 23 Profiles of velocity magnitude for various branching angles α. In-house FVM code red em dashed hypen, FDM code on coarse grid green em dashed hypen, FDM code on finer grid blue em dashed hypen. (**a**) $\alpha = 30°$—channel inlet. (**b**) $\alpha = 30°$—channel outlet. (**c**) $\alpha = 30°$—branch outlet. (**d**) $\alpha = 60°$—channel inlet. (**e**) $\alpha = 60°$—channel outlet. (**f**) $\alpha = 60°$—branch outlet. (**g**) $\alpha = 90°$—channel inlet. (**h**) $\alpha = 90°$—channel outlet. (**i**) $\alpha = 90°$—branch outlet . (**j**) $\alpha = 120°$—channel inlet. (**k**) $\alpha = 120°$—channel outlet. (**l**) $\alpha = 120°$—branch outlet. (**m**) $\alpha = 150°$—channel inlet. (**n**) $\alpha = 150°$—channel outlet. (**o**) $\alpha = 150°$—branch outlet

the numerical method, grid structure, and the way the boundary conditions are imposed. This sensitivity is due to the fact that the flow rate in channel branches is a priori unknown and the flow field only develops due to pressure difference forcing. The discretization artifacts, such as the numerical diffusion and dispersion, can thus significantly affect the flow resistance, and thus the resulting flow can be significantly altered. In this context, the agreement between the numerical predictions of the three considered methods and codes can be judged as very good.

The simplicity of the immersed boundary approach and finite difference discretization allows for very simple testing of various numerical methods and computational setups. The accuracy of this FDM approach proved to be sufficient for this purpose. In addition, the in-house finite volume code for structured grids and the open-source (OpenFOAM) finite volume code for arbitrary grids facilitate the future implementation of the tested models into more advanced codes dealing with realistic three-dimensional geometries.

Our future work will focus on the extension of the presented comparison for unsteady flows and non-Newtonian fluids, which is crucial for the intended investigation of various biomedical applications.

Acknowledgments The financial support for the present work was partly provided by the *Praemium Academiae* of Š. Nečasová and RVO:67985840; partly by the *Czech Technical University, Faculty of Mechanical Engineering,* under the *Grant No. SGS22/148/OHK2/3T/12*; and partly by the European Regional Development Fund Project Center for Advanced Applied Science No.CZ.02.1.01/0.0/0.0/16 019/0000778.

References

1. J.D. Anderson, *Computational Fluid Dynamics - The Basics with Applications* (McGraw-Hill, New York, 1995)
2. L. Beneš, P. Louda, R. Keslerová, K. Kozel, J. Štigler, Numerical simulations of flow through channels with T-junction. Appl. Math. Comput. **219**(13), 7225–7235 (2013)
3. T. Bodnár, A. Sequeira, Numerical simulation of the coagulation dynamics of blood. Comput. Math. Methods Med. **9**(2), 83–104 (2008)
4. T. Bodnár, A. Sequeira, Numerical study of the significance of the non-Newtonian nature of blood in steady flow through a stenosed vessel, in ed. by R. Rannacher, A. Sequeira, *Advances in Mathematical Fluid Mechanics* (Springer, Berlin, 2010), pp. 83–104
5. T. Bodnár, Ph. Fraunié, K. Kozel, Modified equation for a class of explicit and implicit schemes solving one-dimensional advection problem. Acta Polytechnica **61**(SI), 49–58 (2021)
6. J.-I. Choi, R.C. Oberoi, J.R. Edwards, J.A. Rosati, An immersed boundary method for complex incompressible flows. J. Comput. Phys. **224**, 757–784 (2007)
7. A.J. Chorin, A numerical method for solving incompressible viscous flows problems. J. Comput. Phys. **2**(1), 12–26 (1967)
8. A.J. Chorin, The numerical solution of the Navier-Stokes equations for an incompressible fluid. Bull. Amer. Math. Soc. **73**, 928–931 (1967)
9. C.A.J. Fletcher, *Computational Techniques for Fluid Dynamics*. Springer Series in Computational Physics, vol. 1–2, 2nd edn. (Springer, Berlin, 1991)
10. R. Ghias, R. Mittal, H. Dong, A sharp interface immersed boundary method for compressible viscous flows. J. Comput. Phys. **225**, 528–553 (2007)

11. C. Hirsch, *Numerical Computation of Internal and External Flows*, vol. 1, 2 (Wiley, Hoboken, 1988)
12. A. Jameson, Time dependent calculations using multigrid, with applications to unsteady flows past airfoils and wings, in *AIAA 10th Computational Fluid Dynamics Conference, Honolulu*, June 1991. AIAA Paper 91-1596 (1991)
13. A. Jameson, W. Schmidt, E. Turkel, Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes, in *AIAA 14th Fluid and Plasma Dynamic Conference, Palo Alto*, June 1981. AIAA paper 81-1259 (1981)
14. M.Y. Kang, J. Hwang, J.W. Lee, Effect of geometric variations on pressure loss for a model bifurcation of the human lung airway. J. Biomechan. **44**(6), 1196–1199 (2011)
15. R. Keslerová, D. Trdlička, Numerical solution of viscous and viscoelastic fluids flow through the branching channel by finite volume scheme. J. Phys. Conf. Ser. **633**, 012128 (2015)
16. K. Kozel, R. Keslerová, Numerical study of viscous and viscoelastic fluids flow. J. Math-for-Industry **3**(3), 27–32 (2011)
17. A. Lancmanová, Comparison of numerical methods for unsteady simulations of incompressible fluids flows. Master's Thesis, Czech Technical University in Prague, 2020. (in Czech).
18. A. Lancmanová, T. Bodnár, Steady incompressible flow through a branched channel, in ed. by T. Bodnár, T. Neustupa, D. Šimurda, *Proceedings Topical Problems of Fluid Mechanics 2021*, Institute of Thermomechanics CAS (2021), pp. 87–94
19. A. Lancmanová, T. Bodnár, R. Keslerová, Numerical validation of a simple immersed boundary solver for branched channels simulations, in ed. by D. Šimurda, T. Bodnár, *Proceedings Topical Problems of Fluid Mechanics 2022*. Institute of Thermomechanics CAS (2022), pp. 127–134
20. Y. Liu, R.M.C. So, C.H. Zhang, Modeling the bifurcating flow in an asymmetric human lung airway. J. Biomech. **36**(7), 951–959 (2003)
21. B.J. Medhi, V. Agrawal, A. Singh, Experimental investigation of particle migration in suspension flow through bifurcating microchannels. AIChE J. **64**(6), 2293–2307 (2018)
22. R. Mittal, G. Iaccarino, Immersed boundary methods. Ann. Rev. Fluid Mech. **37**, 239–261 (2005)
23. C.S. Peskin, Numerical analysis of blood flow in the heart. J. Comput. Phys. **25**, 220–252 (1977)
24. C.S. Peskin, The immersed boundary method. Acta Numer. **11**, 1–39 (2002)
25. K. Pradhan, A. Guha, Fluid dynamics of a bifurcation. Int. J. Heat Fluid Flow **80**, 108483 (2019)
26. R. Rannacher, *Numerik 3: Probleme der Kontinuumsmechanik und ihre numerische Behandlung*, chapter FE-Methoden für inkompressible Strömungen (Heidelberg University Publishing, Heidelberg, 2017), pp. 203–278
27. M.M. Reddy, A. Singh, Flow of concentrated suspension through oblique bifurcating channels. AIChE J. **60**(7), 2692–2704 (2014)
28. Y. Shang, J. Dong, L. Tian, K. Inthavong, J. Tu, Detailed computational analysis of flow dynamics in an extended respiratory airway model. Clinical Biomech. **61**, 105–111 (2019)
29. E. Turkel, Preconditioned methods for solving the incompressible and low speed compressible equations. J. Comput. Phys. **72**(2), 277–298 (1987)

Consistent $\overline{C}$ Element-Free Galerkin Method for Finite Strain Analysis

P. Areias, F. Carapau, J. Carrilho Lopes, and T. Rabczuk

1 Introduction

Simulations of engineering material processing technology are supported by elastoplastic analyses. Two constitutive requirements are important in this context: (1) the quality of the stress values present in the yield functions depends on the smoothness of the displacements and crucially on mesh distortion [1] and (2) quasi-incompressibility conditions in metal plasticity and polymers are difficult to satisfy with reasonable support sizes in meshless methods [2]. Compared with displacements, errors in stresses are a magnitude higher, even without accounting for incompressibility. High-order (quadratic and cubic) finite elements are typically not adopted in finite strain elastoplastic analysis due to well-known shortcomings:

- High-order elements are adversely impacted by mesh distortion. Convergence rate is changed by distortion [1]. Adaptive remeshing is required more often with high-order elements.
- Problems requiring high-order derivatives impose dedicated techniques or isogeometric formulations [3].

P. Areias
Instituto Superior Técnico, Lisbon, Portugal
e-mail: pedro.areias@tecnico.ulisboa.pt

F. Carapau (✉) · J. C. Lopes
Universidade de Évora, Colégio Luis António Verney, Évora, Portugal
e-mail: flc@uevora.pt

T. Rabczuk
Bauhaus-University Weimar, Weimar, Germany
e-mail: timon.rabczuk@uni-weimar.de

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_6

- Although stress quality improves with the order of the complete polynomial, in finite element methods, stresses are still discontinuous at inter-element boundaries [4]. Plasticity results are dependent on the quality of the stresses, which is compromised even in high-order finite elements.
- The use of finite elements for quasi-incompressible problems requires specialized techniques (see, for example, [5–7]).

Note that Rabczuk, Belytschko, and Xiao [8] proved that a Lagrangian kernel is required for stability,[1] but classical finite strain plasticity algorithms (e.g., [9, 10]) combined with EFG are based on configuration updating (see [11]). A comprehensive presentation of developments in meshless methods (including EFG) was recently published by J.-S. Chen et al. [12]. A related development combining partition of unity and least squares is described in Cai et al. [13]. Several remedies are described, in particular for boundary conditions. Therefore, meshless methods, in particular with quadratic and cubic bases and satisfying the Kronecker delta condition, perfectly fit these applications:

- Since no isoparametric mapping is used, mesh distortion sensitivity is attenuated with respect to finite elements.
- Stresses are continuous, as long as all terms participating in the shape functions are differentiable.
- Contact algorithms are relatively simplified.
- Quasi-incompressibility can be directly addressed by changing the polynomial basis.
- Strain localization problems can be directly addressed via strain-gradient methods.

Several applications have been published with meshless discretization for finite strain plasticity [11], but not at the same scale of finite elements. The reputation for difficult-to-impose boundary conditions still affects EFG, although developments in interpolation have resurrected interest in the question of the Kronecker delta property (see [14]). In contrast with finite strain plasticity, hyperelastic implementations of EFG are common, and recent papers report realistic results with high degree of continuity (see [15]). In this paper are the following:

A newly developed fully anisotropic elastoplastic framework based on the iteration for $\boldsymbol{C}_e$ [16] does not require the explicit form of the deformation gradient. This motivates a revisiting of the moving least squares/EFG approach. Another effect that is often reported in the context of EFG is the volumetric locking in quasi-incompressible applications, [11, 17]. This is addressed here by the following techniques:

- Selective quadrature for the right Cauchy-Green tensor $\boldsymbol{C}$, with reduced quadrature in $\det \boldsymbol{C}$ and full quadrature in $\widehat{\boldsymbol{C}} = \det [\boldsymbol{C}]^{-1/3}\, \boldsymbol{C}$

[1] Strictly in particle methods, but stabilized particle methods share properties with EFG.

- Selective interpolation for these terms, with a higher-order polynomial being adopted for $\widehat{\boldsymbol{C}}$

In terms of discretization, this work adopts the following techniques:

- Ab initio definition of the shape functions and derivatives for the entire analysis.
- Parameterized quadrature and interpolation functions for the deviatoric and volumetric parts of the right Cauchy-Green tensor $\boldsymbol{C}$.
- Quasi-singular weight functions (see [18–20]).
- Quadrature points are defined in tetrahedra.
- Lagrangian diffuse derivatives are adopted.
- Constitutive integration making use of the Mandel stress tensor and iteration on $\boldsymbol{C}_e$ [16].

Volumetric locking has been diagnosed in element-free Galerkin methods by Dolbow and Belytschko [21] where a mixed displacement-pressure formulation was proposed in the small strain case. Within the RKPM family of W.K. Liu's group, a pressure projection method was proposed, where pressure is re-interpolated using fewer points and a specific patch [22]. Applications were made with incompressible hyperelasticity. More conventional F-bar formulations have been used in the context of particle methods with explicit integration by Wu et al. [23]. In the small strain case, Recio , Jorge and Dinis [24] have applied $\overline{B}$ and Enhanced strain techniques to an EFG formulation. For implicit integration, an incremental finite deformation version was adopted by Coombs et al. [25]. In neither of these papers the closed-form expressions for the equilibrium and Jacobian were presented in the finite strain case. In the incremental case (see [25]), expressions are significantly simplified, and results for moderate plastic deformations are shown in that paper. In Moutsanidis et al. [26], an F-bar implementation is presented for the conforming reproducing kernel method. Navas et al. [27], in order to avoid the locking involved in the fluid phase of the porous media, devised a B-bar algorithm.

This paper is organized as follows: Sect. 2 presents the interpolation, based on moving least squares and diffuse derivatives, as well as the algorithm to guarantee a sufficiently small support radius. Section 3 presents the discretization based on the total Lagrangian approach, including the partition of $\boldsymbol{C}$ with its first and second variations. This is followed by Sect. 4 where the constitutive integration, fitting the developments of Sect. 3, is described in detail. In Sect. 5, three benchmark tests are presented, and finally conclusions are drawn in Sect. 6.

2 Interpolation

2.1 General Approach for Moving Least Squares

Interpolation with a polynomial basis and least squares fitting was introduced by P. Lancaster and K. Salkauskas [18]. Herein, classical derivations are followed (see

[19, 28, 29]). We introduce m as the number of terms in the polynomial basis, n as the number of supporting nodes, and D as support radius. For a given node K,, the distance to a given point with coordinates $\boldsymbol{X}$ is identified as $s_K(\boldsymbol{X})$. Let us consider a $q-$tuple of nonnegative integers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q) \in \mathbb{N}_0^q$. We write the absolute value as the sum $|\boldsymbol{\alpha}| = \sum_{i=1}^{q} \alpha_i$. We consider the set of all polynomials of degree equal or less than p as:

$$\mathcal{P}_p = \left\{ p_\alpha(\boldsymbol{X}) = X_1^{\alpha_1} \cdots X_q^{\alpha_q} \mid |\boldsymbol{\alpha}| \leq p \right\}. \tag{1}$$

We now introduce a polynomial basis as an array of elements of $\mathcal{P}_p$:

$$\boldsymbol{p}(\boldsymbol{X}) = \{p_1(\boldsymbol{X}), p_2(\boldsymbol{X}), \cdots, p_m(\boldsymbol{X})\} \quad p_i \in \mathcal{P}_p \tag{2}$$

with $\#\boldsymbol{q}(\boldsymbol{X}) = {}^{(p+q)!}/_{p!q!} = m$. We therefore use m elements of $\mathcal{P}$ for the polynomial basis. The direct form (2) is known to produce conditioning difficulties. Therefore, we adopt a normalized and shifted form using a complete basis:

$$\begin{aligned} \boldsymbol{p}(\boldsymbol{X}) = \Bigg\{ 1, \frac{(X_1 - \overline{X}_1)}{D}, \frac{(X_2 - \overline{X}_2)}{D}, \frac{(X_3 - \overline{X}_3)}{D}, \\ \frac{(X_1 - \overline{X}_1)(X_2 - \overline{X}_2)}{D^2}, \frac{(X_1 - \overline{X}_1)(X_3 - \overline{X}_3)}{D^2}, \frac{(X_2 - \overline{X}_2)(X_3 - \overline{X}_3)}{D^2}, \\ \frac{(X_1 - \overline{X}_1)^2}{D^2}, \frac{(X_2 - \overline{X}_2)^2}{D^2}, \frac{(X_3 - \overline{X}_3)^2}{D^2}, \ldots \Bigg\}. \end{aligned} \tag{3}$$

We use $\overline{\boldsymbol{X}}$ as a centroid of the nodes within the $D-$radius of $\boldsymbol{X}$. Given a point with coordinates $\boldsymbol{X}$, the approximation weight of another point with coordinates $\boldsymbol{X}_I$ depends on the distance between the points $s_I(\boldsymbol{X}) = \|\boldsymbol{X} - \boldsymbol{X}_I\|$. The notation $w[s_I(\boldsymbol{X})]$ is introduced to represent this weight function of $\boldsymbol{X}$. From this basis, an $m \times n$ $\boldsymbol{P}$ Vandermonde matrix is defined by its elements as follows:

$$P_{iJ} = p_i(\boldsymbol{X}_J) \quad i = 1, \ldots, m, \qquad J = 1, \ldots, n \tag{4}$$

The components of weight matrix, which is a function of the supporting points and the coordinates $\boldsymbol{X}$, are given by:

$$W_{IJ}(\boldsymbol{X}) = \delta_{IJ} w[s_I(\boldsymbol{X})] \qquad I, J = 1, \ldots, n \tag{5}$$

Applying the traditional least squares arguments [28] leads to the following format for the n-dimensional shape function array $\boldsymbol{N}(\boldsymbol{X})$:

$$\boldsymbol{N}(\boldsymbol{X}) = \boldsymbol{p}(\boldsymbol{X}) \cdot \boldsymbol{A}^{-1}(\boldsymbol{X}) \cdot \boldsymbol{B}(\boldsymbol{X}) \tag{6}$$

where $\boldsymbol{A}(\boldsymbol{X})$ is the $m \times m$ moment matrix $\boldsymbol{A}(\boldsymbol{X}) = \boldsymbol{B}(\boldsymbol{X}) \cdot \boldsymbol{P}^T$ and $\boldsymbol{B}(\boldsymbol{X})$ is the $m \times n$ linear combination matrix $\boldsymbol{B}(\boldsymbol{X}) = \boldsymbol{P} \cdot \boldsymbol{W}(\boldsymbol{X})$. We make use of the $\boldsymbol{Q} \cdot \boldsymbol{R}$ decomposition of $\sqrt{\boldsymbol{W}(\boldsymbol{X})} \cdot \boldsymbol{P}^T$:

$$\sqrt{\boldsymbol{W}(\boldsymbol{X})} \cdot \boldsymbol{P}^T = \boldsymbol{Q}(\boldsymbol{X}) \cdot \boldsymbol{R}(\boldsymbol{X}) \tag{7}$$

where $\boldsymbol{Q}(\boldsymbol{X})$ is an orthogonal matrix and $\boldsymbol{R}(\boldsymbol{X})$ is an upper triangular matrix [30]. A classical Gram-Schmidt algorithm for the $\boldsymbol{Q} \cdot \boldsymbol{R}$ decomposition is used (see [31]). For our application, only $\boldsymbol{R}(\boldsymbol{X})$ is required. It is straightforward to obtain, from (6), the final form of the shape function array:

$$\boldsymbol{N}(\boldsymbol{X}) = \boldsymbol{p}(\boldsymbol{X}) \cdot \boldsymbol{R}^{-1}(\boldsymbol{X}) \cdot \boldsymbol{R}^{-T}(\boldsymbol{X}) \cdot \boldsymbol{B}(\boldsymbol{X}). \tag{8}$$

Therefore, this operation is relatively inexpensive since it consists of two triangular solves. Omitting the dependence on $\boldsymbol{X}$, we have:

$$\boldsymbol{R}^T \cdot \boldsymbol{U}_1 = \boldsymbol{B} \tag{9}$$

$$\boldsymbol{R} \cdot \boldsymbol{U}_2 = \boldsymbol{U}_1 \tag{10}$$

where $\boldsymbol{U}_2$ is a $m \times n$ matrix, which suffices to define the shape functions. Reintroducing the dependence on $\boldsymbol{X}$, the result is:

$$\boldsymbol{N}(\boldsymbol{X}) = \boldsymbol{p}(\boldsymbol{X}) \cdot \boldsymbol{U}_2(\boldsymbol{X}). \tag{11}$$

The interpolated value $\phi(\boldsymbol{X})$ is obtained by linear combination of nodal values $\boldsymbol{\phi} = \{\phi_1, \phi_2, \cdots, \phi_n\}$ $\phi(\boldsymbol{X}) = \boldsymbol{N}(\boldsymbol{X}) \cdot \boldsymbol{\phi}$. In terms of components, Eq. (6) is written as:

$$N_L(\boldsymbol{X}) = p_j(\boldsymbol{X}) U_{2jL}(\boldsymbol{X}) \qquad L = 1, \ldots, n; \quad j, k = 1, \ldots, m \tag{12}$$

First derivative of $N_L(\boldsymbol{X})$ with respect to coordinates X_m, $m = 1, 2, 3$ is here denoted as:

$$\begin{aligned} N'_L(\boldsymbol{X}) =& p'_j(\boldsymbol{X}) U_{2jL}(\boldsymbol{X}) \\ & - p_j(\boldsymbol{X}) A^{-1}_{jl}(\boldsymbol{X}) A'_{lp}(\boldsymbol{X}) U_{2pL}(\boldsymbol{X}) \\ & + p_j(\boldsymbol{X}) A^{-1}_{jk}(\boldsymbol{X}) B'_{kL}(\boldsymbol{X}) \end{aligned} \tag{13}$$

where:

$$B'_{kL}(\boldsymbol{X}) = P_{kJ} W'_{JL}(\boldsymbol{X}) \tag{14}$$

$$A'_{lp}(\boldsymbol{X}) = B'_{lL}(\boldsymbol{X}) P_{pL}. \tag{15}$$

In terms of $p'_j(\boldsymbol{X})$ and $W'_{JI}(\boldsymbol{X})$, Eq. (13) can be written as a sum of two terms:

$$N'_L(\boldsymbol{X}) = N^{\star}_L(\boldsymbol{X}) + N^{\bullet}_L(\boldsymbol{X}) \tag{16a}$$

where:

$$N^{\star}_L(\boldsymbol{X}) = p'_j(\boldsymbol{X})\, U_{2jL}(\boldsymbol{X}) \tag{16b}$$

and:

$$N^{\bullet}_L(\boldsymbol{X}) = p_j(\boldsymbol{X})\, A^{-1}_{jl}(\boldsymbol{X})\, P_{lM} W'_{MQ}(\boldsymbol{X}) \left[\delta_{QL} - P_{pQ} U_{2pL}(\boldsymbol{X})\right]. \tag{16c}$$

It is a tradition to identify (16b) as the diffuse derivative (see Nayroles, Touzot, and Villon [32]).

2.2 Quasi-Singular Weight Function

Singular weight functions are known to produce an interpolation satisfying the Kronecker delta property [18]. Quasi-singular functions have been adopted to approximate this property [19]. The following quasi-singular weight function is introduced (see, for example, [19, 20]):

$$w\left[s_I(\boldsymbol{X})\right] = \begin{cases} \left[{}^{s_I^2(\boldsymbol{X})}/_{D^2} + \mathrm{tol}^2\right]^{-1} - \left[1 + \mathrm{tol}^2\right]^{-1} & s_I \leq D \\ 0 & s_I > D \end{cases} \tag{17}$$

where tol $\in \mathbb{R}^+$ is a tolerance parameter. The maximum value of $w[s_I]$ is obtained as:

$$w[0] = 1/(\mathrm{tol}^2 + \mathrm{tol}^4). \tag{18}$$

Here, we adopt tol $= 1\times 10^{-3}$. The Kronecker delta property is approximately satisfied:

$$N_I(\boldsymbol{X}_J) \cong \delta_{IJ}. \tag{19}$$

Derivatives of $w[s_I]$ with respect to s_I are trivially given by

$$\frac{\mathrm{d}w[s_I]}{\mathrm{d}s_I} = -\frac{2D^2 s_I}{\left(D^2\mathrm{tol}^2 + s_I^2\right)^2}. \tag{20}$$

Strong versions of this weighting are available (see M. Dehghan, [33]) but involve an intricate implementation.

3 Discrete Equilibrium Equations

In finite element technology, two papers introduced a consistent formulation for the so-called mean dilatation technique [6, 34] which was invented by Nagtegaal et al. [35]. A straightforward total Lagrangian implementation is followed (see, for example, [36]). We make use of the definition of the right Cauchy-Green tensor:

$$\boldsymbol{C}\left(\boldsymbol{X}_h\right)=\boldsymbol{F}^T\left(\boldsymbol{X}\right)\cdot\boldsymbol{F}\left(\boldsymbol{X}\right). \tag{21}$$

A partition into volumetric and deviatoric parts is required for selective quadrature. Omitting the dependence on $\boldsymbol{X}_h$, the derivatoric Cauchy-Green tensor follows from the Flory [37] decomposition:

$$\widehat{\boldsymbol{C}}=\det\left[\boldsymbol{C}\right]^{-1/3}\boldsymbol{C}. \tag{22}$$

Introducing the variation symbol δ and taking advantage of the symmetry of $\boldsymbol{C}$, the variation of $\widehat{\boldsymbol{C}}$ is calculated as (see also Appendix Section "First and Second Variations of $\det\left[\boldsymbol{C}\right]$"):

$$\delta\widehat{\boldsymbol{C}}=\left(\det\left[\boldsymbol{C}\right]^{-1/3}\mathcal{I}-\frac{1}{3}\widehat{\boldsymbol{C}}\otimes\boldsymbol{C}^{-1}\right):\delta\boldsymbol{C} \tag{23}$$

where $\mathcal{I}$ is the symmetric fourth-order identity tensor, i.e., $[\mathcal{I}]_{ijkl}=\frac{1}{2}\left(\delta_{ij}\delta_{kl}+\delta_{ik}\delta_{jl}\right)$. This variation will be required later in the formation of the weak form of equilibrium. Newton-Raphson iteration requires the second variation of $\widehat{\boldsymbol{C}}$. For the second variation of $\widehat{\boldsymbol{C}}$, we adopt the time derivative notation, which results in:

$$\begin{aligned}\delta\dot{\widehat{\boldsymbol{C}}}=&\left(\det\left[\boldsymbol{C}\right]^{-1/3}\mathcal{I}-\frac{1}{3}\widehat{\boldsymbol{C}}\otimes\boldsymbol{C}^{-1}\right):\delta\dot{\boldsymbol{C}}\\&+\dot{\boldsymbol{C}}:\left(-\frac{1}{3}\det\left[\boldsymbol{C}\right]^{-4/3}\boldsymbol{C}^{-1}\otimes\mathcal{I}\right):\delta\boldsymbol{C}\\&+\dot{\boldsymbol{C}}:\left(\frac{1}{9}\widehat{\boldsymbol{C}}\otimes\boldsymbol{C}^{-1}\otimes\boldsymbol{C}^{-1}-\frac{1}{3}\det\left[\boldsymbol{C}\right]^{-1/3}\mathcal{I}\otimes\boldsymbol{C}^{-1}\right):\delta\boldsymbol{C}\\&+\dot{\boldsymbol{C}}:\mathtt{T}:\delta\boldsymbol{C}\end{aligned} \tag{24}$$

where T is a sixth-order tensor which is defined in terms of components as:

$$[\mathsf{T}]_{mnijkl} = \frac{1}{3} C^{-1}_{km} C^{-1}_{nl} \widehat{C}_{ij}. \tag{25}$$

Given the decomposition, we assume an independent $\det\left[\overline{\boldsymbol{C}}\right]$ which we denote as $\theta_C = \det\left[\overline{\boldsymbol{C}}\right]$. In this case, we define a combined right Cauchy-Green tensor:

$$\boldsymbol{C}_\star = \left(\frac{\theta_C}{\det[\boldsymbol{C}]}\right)^{1/3} \boldsymbol{C}. \tag{26}$$

Interpolation for $\overline{\boldsymbol{C}}$ makes use of a lower-order polynomial and/or fewer quadrature points. The specific form (26) was proposed by Simo et al. [6] with a clear significance: in the context of low-order finite elements, to replace an over-constrained imposition of $\det[\boldsymbol{C}] \cong 1$ by an independent field θ_C. Here, $\overline{\boldsymbol{C}}$ can follow a distinct quadrature rule or a distinct interpolation. The first variation of $\boldsymbol{C}_\star$ is calculated as:

$$\delta\boldsymbol{C}_\star = \left(\frac{\theta_C}{\det[\boldsymbol{C}]}\right)^{1/3} \delta\boldsymbol{C} + \frac{1}{3}\left[\left(\boldsymbol{C}_\star \otimes \overline{\boldsymbol{C}}^{-1}\right) : \delta\overline{\boldsymbol{C}} - \left(\boldsymbol{C}_\star \otimes \boldsymbol{C}^{-1}\right) : \delta\boldsymbol{C}\right]. \tag{27}$$

Using the time derivative notation, an analogous form is obtained:

$$\dot{\boldsymbol{C}}_\star = \left(\frac{\theta_C}{\det[\boldsymbol{C}]}\right)^{1/3} \dot{\boldsymbol{C}} + \frac{1}{3}\left[\left(\boldsymbol{C}_\star \otimes \overline{\boldsymbol{C}}^{-1}\right) : \dot{\overline{\boldsymbol{C}}} - \left(\boldsymbol{C}_\star \otimes \boldsymbol{C}^{-1}\right) : \dot{\boldsymbol{C}}\right]. \tag{28}$$

The time derivative of $\delta\boldsymbol{C}_\star$ is obtained from (27) as:

$$\begin{aligned}\delta\dot{\boldsymbol{C}}_\star = & -\dot{\theta}_C \frac{2}{9}\theta_C^{-5/3}\widehat{\boldsymbol{C}}\delta\theta_C - \frac{1}{3}\theta_C^{-2/3}\widehat{\boldsymbol{C}}\delta\dot{\theta}_C \\ & + \frac{1}{3}\theta_C^{-2/3}\left(\dot{\theta}_C\delta\widehat{\boldsymbol{C}} + \delta\theta_C\dot{\widehat{\boldsymbol{C}}}\right) + \theta_C^{1/3}\delta\dot{\widehat{\boldsymbol{C}}}.\end{aligned}$$

These expressions are error-prone to implement manually and therefore have been implemented in Mathematica [38] with the AceGen add-on, developed by Korelc [39]. The Mathematica sheets and corresponding Fortran 90 source codes are available in GitHub (see [40]). For a given point $\boldsymbol{X}_h$ with discrete support $\Omega_{\boldsymbol{X}_h}$, we have:

$$\boldsymbol{F}(\boldsymbol{X}_h) = \frac{\mathrm{d}\boldsymbol{x}_h}{\mathrm{d}\boldsymbol{X}_h} = \sum_{L\in\Omega_{\boldsymbol{X}_h}} \left(\frac{\mathrm{d}N_L(\boldsymbol{X}_h)\,\boldsymbol{x}_L}{\mathrm{d}\boldsymbol{X}_h}\right). \tag{29}$$

In terms of components and omitting the dependence on $\boldsymbol{X}_h$, we obtain the components of $\boldsymbol{F}$ as:

$$F_{ij} = \frac{\mathrm{d}N_L}{\mathrm{d}X_j} x_{iL}. \tag{30}$$

Using the variation symbol, δ, we introduce the variation of $\boldsymbol{F}$, in the equilibrium sense, as:

$$\delta F_{ij} = \frac{\mathrm{d}N_L}{\mathrm{d}X_j} \delta x_{iL}. \tag{31}$$

Introducing the notation $N_{jL} = {}^{\mathrm{d}N_L}/_{\mathrm{d}X_j}$ for the shape function derivatives, the following results for $\boldsymbol{C}$ and its first and second variations are obtained:

$$\begin{aligned}
C_{ij} &= N_{iK} N_{jL} x_{kK} x_{kL} \Rightarrow \\
\delta C_{ij} &= N_{iK} N_{jL} \left(x_{kL} \delta x_{kK} + x_{kK} \delta x_{kL} \right) \\
\dot{C}_{ij} &= N_{iK} N_{jL} \left(x_{kL} \dot{x}_{kK} + x_{kK} \dot{x}_{kL} \right) \\
\delta \dot{C}_{ij} &= N_{iK} N_{jL} \left(\dot{x}_{kL} \delta x_{kK} + \dot{x}_{kK} \delta x_{kL} \right).
\end{aligned}$$

Note that besides the node indices K and L, the index k is also muted. Equilibrium is established in a weak form by the use of the second Piola-Kirchhoff stress $\boldsymbol{S}_\star$ and the spatial configuration variation $\delta \boldsymbol{x}$:

$$\frac{1}{2} \int_{\Omega_0} \boldsymbol{S}_\star : \delta \boldsymbol{C}_\star \, \mathrm{d}\Omega_0 \;=\; \boldsymbol{f}_{\mathrm{ext}} \cdot \delta \boldsymbol{x} \tag{32}$$

where $\boldsymbol{S}_\star \equiv \boldsymbol{S}_\star \left(\boldsymbol{C}_\star \right)$ where $\boldsymbol{C}_\star$ was calculated as shown in (27). For the application of Newton-Raphson iteration, we require the first variation of (32). As discussed previously, to avoid confusion with the variation symbol δ, we use the time derivative to denote the variation of equilibrium. By taking this time derivative variation, the tangent modulus $\mathcal{C}$ is employed to read:

$$\frac{1}{2} \int_{\Omega_0} \boldsymbol{S}_\star : \delta \dot{\boldsymbol{C}}_\star \mathrm{d}\Omega_0 + \frac{1}{4} \int_{\Omega_0} \delta \boldsymbol{C}_\star : \mathcal{C} : \dot{\boldsymbol{C}}_\star \, \mathrm{d}\Omega_0 = \boldsymbol{f}_{\mathrm{ext}} \cdot \delta \boldsymbol{x} \;-\; \frac{1}{2} \int_{\Omega_0} \boldsymbol{S}_\star : \delta \boldsymbol{C}_\star \, \mathrm{d}\Omega_0 \tag{33}$$

where $\boldsymbol{f}_{\mathrm{ext}}$ is the external load vector and is the nodal velocity vector. Note that in the implementation, the second derivative of $\boldsymbol{C}_\star$ is required in $\delta \dot{\boldsymbol{C}}_\star$. In Voigt form (see [41]), we have the following internal force and tangent stiffness:

$$\boldsymbol{f}_L = \int_{\Omega_0} \boldsymbol{B}_L^T \cdot I_6 \cdot \hat{\boldsymbol{S}}_\star \, \mathrm{d}\Omega_0 \tag{34}$$

$$\boldsymbol{K}_{KL} = \int_{\Omega_0} \boldsymbol{B}_K^T \cdot I_6 \cdot \mathcal{C} \cdot I_6 \cdot \boldsymbol{B}_L \, \mathrm{d}\Omega_0 + \int_{\Omega_0} \check{\boldsymbol{S}}_\star \cdot I_6 \cdot \boldsymbol{B}_L^\star \, \mathrm{d}\Omega_0. \tag{35}$$

Matrices $\boldsymbol{B}$ and $\boldsymbol{B}^{\star}$ are implemented in [40], and I_6 is a diagonal matrix containing 1 for indices 11 and 22 and 33 and 2 for indices 44, 55, and 66. In contrast with advanced finite element formulations [42, 43], these are classical and direct derivations. In addition, shape functions and corresponding derivatives are calculated once, at the start of the solution process.

4 Hyperelasticity/Plasticity Using the Elastic Mandel Stress Tensor

4.1 Formulation

The Mandel stress tensor approach to finite strain plasticity is adopted [44, 45]. We make use of the Kröner-Lee decomposition [46–48]:

$$\boldsymbol{F} = \boldsymbol{F}_e \cdot \boldsymbol{F}_p. \tag{36}$$

Using (36), the velocity gradient is determined by its definition and then partitioned as follows:

$$\boldsymbol{L} = \dot{\boldsymbol{F}} \cdot \boldsymbol{F}^{-1} = \boldsymbol{L}_e + \boldsymbol{F}_e \cdot \boldsymbol{L}_p \cdot \boldsymbol{F}_e^{-1} \tag{37}$$

with $\boldsymbol{L}_e = \dot{\boldsymbol{F}}_e \cdot \boldsymbol{F}_e^{-1}$ the elastic velocity gradient and $\boldsymbol{L}_p = \dot{\boldsymbol{F}}_p \cdot \boldsymbol{F}_p^{-1}$ the plastic velocity gradient. The second Piola-Kirchhoff stress is a function of the elastic part of $\boldsymbol{F}$ by means of $\boldsymbol{C}_e = \boldsymbol{F}_e^T \cdot \boldsymbol{F}_e$ (cf. [49] page 166), the second Piola-Kirchhoff stress at the intermediate configuration is given by $\boldsymbol{S}_e\left(\boldsymbol{C}_e\right)$ (see [50]), from which energy consistency results in a specific form for the second Piola-Kirchhoff stress $\boldsymbol{S} = \boldsymbol{F}_p^{-1} \cdot \boldsymbol{S}_e\left(\boldsymbol{C}_e\right) \cdot \boldsymbol{F}_p^{-T}$. In the hyperelastic case, a strain energy density function $\psi\left(\boldsymbol{C}_e\right)$ exists such as:

$$\boldsymbol{S}_e\left(\boldsymbol{C}_e\right) = 2\frac{\mathrm{d}\psi\left(\boldsymbol{C}_e\right)}{\mathrm{d}\boldsymbol{C}_e}. \tag{38}$$

The Neo-Hookean model is used, with the following strain energy density function:

$$\psi(\mathbf{C}_e) = \frac{\mu}{2}\left[\mathrm{tr}\left(\mathbf{C}_e\right) - 3\right] - \mu \log\sqrt{\det\left(\mathbf{C}_e\right)} + \frac{\lambda}{2}\left[\log\sqrt{\det\left(\mathbf{C}_e\right)}\right]^2. \tag{39}$$

The flow law follows similar arguments [45], with the initial plastic deformation gradient corresponding to the identity, $\left[\boldsymbol{F}_p\right]_0 = \boldsymbol{I}$. Agreeing with standard derivations on plasticity, a yield function ϕ is introduced, as well as a plastic multiplier $\dot{\gamma}$. Introducing the notation $\boldsymbol{Q}_p = \boldsymbol{F}_p^{-1}$, we summarize the constitutive system as:

$$\boldsymbol{S} = \boldsymbol{Q}_p \cdot \boldsymbol{S}_e\left(\boldsymbol{C}_e\right) \cdot \boldsymbol{Q}_p^T \tag{40}$$

$$\dot{\boldsymbol{Q}}_p = -\dot{\gamma}\,\boldsymbol{Q}_p \cdot \boldsymbol{N}\left[\boldsymbol{T}_e\right] \tag{41}$$

$$\left[\boldsymbol{Q}_p\right]_0 = \boldsymbol{I} \tag{42}$$

$$\prec \phi\left(\boldsymbol{T}_e\right) + \dot{\gamma} \succ - \dot{\gamma} = 0 \tag{43}$$

with $\prec \bullet \succ = \frac{\bullet + |\bullet|}{2}$ being the unit ramp function. In (41), the Mandel stress [44] $\boldsymbol{T}_e$ is given by:

$$\boldsymbol{T}_e = \boldsymbol{C}_e \cdot \boldsymbol{S}_e\left(\boldsymbol{C}_e\right). \tag{44}$$

Assuming an associated flow law [48], we have the flow vector $\boldsymbol{N}\left(\boldsymbol{T}_e\right)$ determined from the derivative of $\phi\left(\boldsymbol{T}_e\right)$:

$$\boldsymbol{N}(\boldsymbol{T}_e) = {}^{\mathrm{d}\phi(\boldsymbol{T}_e)}/_{\mathrm{d}\boldsymbol{T}_e}. \tag{45}$$

When hardening is present, power equivalence provides the effective plastic strain rate $\dot{\varepsilon}_p$ as a function of the yield stress σ_y:

$$\dot{\varepsilon}_p = \dot{\gamma}\frac{\boldsymbol{T}_e : \boldsymbol{N}(\boldsymbol{T}_e)}{\sigma_y}. \tag{46}$$

4.2 Constitutive Integration

For the constitutive integration, we use superscripts n and $n+1$ to identify two consecutive time steps and Δt as the time step size. Applying the backward Euler method for $\dot{\boldsymbol{Q}}_p$ and $\dot{\gamma}$ results in:

$$\boldsymbol{Q}_p^{n+1} = \boldsymbol{Q}_p^n \cdot \underbrace{\left[\boldsymbol{I} + \Delta\gamma\widehat{\boldsymbol{N}}\left(\boldsymbol{C}_e^{n+1}\right)\right]^{-1}}_{\left[\Delta\widehat{\boldsymbol{Q}}(\boldsymbol{C}_e^{n+1},\Delta\gamma)\right]^{-1}} \tag{47}$$

$$\gamma^{n+1} = \gamma^n + \underbrace{\dot{\gamma}^{n+1}\Delta t}_{\Delta\gamma}. \tag{48}$$

We now define the elastic trial Cauchy-Green tensor as $\boldsymbol{C}_e^\star = \left[\boldsymbol{Q}_p^n\right]^T \cdot \boldsymbol{C}^{n+1} \cdot \boldsymbol{Q}_p^n$. Introducing the function $\widehat{\boldsymbol{C}}_e^\star\left(\boldsymbol{C}^{n+1}\right) = \left(\boldsymbol{Q}_p^n\right)^T \cdot \boldsymbol{C}^{n+1} \cdot \boldsymbol{Q}_p^n$, the constitutive system for $\Delta\gamma > 0$ consists of the following equations:

$$\underbrace{\left[\Delta\widehat{\boldsymbol{Q}}\left(\boldsymbol{C}_e^{n+1},\Delta\gamma\right)\right]^T\cdot\boldsymbol{C}_e^{n+1}\cdot\left[\Delta\widehat{\boldsymbol{Q}}\left(\boldsymbol{C}_e^{n+1},\Delta\gamma\right)\right]-\widehat{\boldsymbol{C}}_e^{\star}\left(\boldsymbol{C}^{n+1}\right)}_{\boldsymbol{r}_c\left(\boldsymbol{C}_e^{n+1},\Delta\gamma,\boldsymbol{C}^{n+1}\right)}=\mathbf{0} \tag{49}$$

$$\phi_\star\left[\boldsymbol{C}_e^{n+1}\cdot\hat{\boldsymbol{S}}_e\left(\boldsymbol{C}_e^{n+1}\right)\right]=0. \tag{50}$$

Since $\boldsymbol{C}_e^{n+1}$ is symmetric, Voigt notation can be used, $\mathbf{C}_e^{n+1}=\text{Voigt}\left[\boldsymbol{C}_e^{n+1}\right]$ and $\mathbf{r}_c\left(\mathbf{C}_e^{n+1},\Delta\gamma,\mathbf{C}^{n+1}\right)=\text{Voigt}\left[\boldsymbol{r}_c\left(\boldsymbol{C}_e^{n+1},\Delta\gamma,\boldsymbol{C}^{n+1}\right)\right]$. Omitting the function arguments for conciseness, the Newton-Raphson iteration for $\mathbf{C}_e^{n+1}$ (Voigt form) and $\Delta\gamma$ is written as:

$$\underbrace{\begin{bmatrix}\frac{\partial\mathbf{r}_c}{\partial\mathbf{C}_e^{n+1}} & \frac{\partial\mathbf{r}_c}{\partial\Delta\gamma}\\ \frac{\partial\phi}{\partial\mathbf{C}_e^{n+1}} & 0\end{bmatrix}}_{\boldsymbol{J}}\underbrace{\begin{Bmatrix}\Delta\mathbf{C}_e^{n+1}\\ \Delta\Delta\gamma\end{Bmatrix}}_{\Delta\mathbf{Y}}=-\underbrace{\begin{Bmatrix}\mathbf{r}_c\left(\mathbf{C}_e^{n+1},\Delta\gamma,\mathbf{C}^{n+1}\right)\\ \phi_\star\left[\mathbf{C}_e^{n+1}\cdot\hat{\mathbf{S}}_e\left(\mathbf{C}_e^{n+1}\right)\right]\end{Bmatrix}}_{\mathbf{r}} \tag{51}$$

with $\mathbf{Y}=\left\{\mathbf{C}_e^{n+1}\ \Delta\gamma\right\}^T$ being the constitutive unknowns for this problem. Following $\mathbf{C}_e^{n+1}$, $\boldsymbol{Q}_p^{n+1}$ is determined by (47), and the second Piola-Kirchhoff stress at step $n+1$ is given in tensor notation by:

$$\check{\boldsymbol{S}}^{n+1}\left(\underbrace{\mathbf{C}_e^{n+1},\Delta\gamma}_{\mathbf{Y}}\right)=\boldsymbol{Q}_p^n\cdot\left[\Delta\widehat{\boldsymbol{Q}}\left(\mathbf{C}_e^{n+1},\Delta\gamma\right)\right]^{-1}\cdot\hat{\boldsymbol{S}}_e\left(\mathbf{C}_e^{n+1}\right)$$
$$\cdot\left\{\left[\Delta\widehat{\boldsymbol{Q}}\left(\mathbf{C}_e^{n+1},\Delta\gamma\right)\right]^{-1}\right\}^T\cdot\left(\boldsymbol{Q}_p^n\right)^T. \tag{52}$$

Stress sensitivity, the determination of the consistent modulus, with $\mathbf{S}^{n+1}=\text{Voigt}\left[\boldsymbol{S}^{n+1}\right]$, is determined as follows:

$$\frac{\mathrm{d}\mathbf{S}^{n+1}}{\mathrm{d}\mathbf{C}^{n+1}}=\frac{\partial\widehat{\mathbf{S}}^{n+1}}{\partial\mathbf{C}_e^{n+1}}\cdot\frac{\mathrm{d}\mathbf{C}_e^{n+1}}{\mathrm{d}\mathbf{C}^{n+1}}+\frac{\partial\widehat{\mathbf{S}}^{n+1}}{\partial\Delta\gamma}\frac{\mathrm{d}\Delta\gamma}{\mathrm{d}\mathbf{C}^{n+1}}. \tag{53}$$

In (53), a single product dot $\cdot$ is adopted for double contraction of quantities in Voigt form. From (53), we can conclude that $\mathcal{C}$ is determined as a function of the solution of (51), since:

$$\mathrm{d}\mathbf{Y}/\mathrm{d}\mathbf{C}^{n+1}=-\boldsymbol{J}^{-1}\cdot\frac{\partial\mathbf{r}}{\partial\mathbf{C}^{n+1}} \tag{54}$$

therefore, stress sensitivity is simply given by:

$$\frac{\mathrm{d}\mathbf{S}^{n+1}}{\mathrm{d}\mathbf{C}^{n+1}} = -\left({}^{\mathrm{d}\widehat{\mathbf{S}}^{n+1}}\!/_{\mathrm{d}\mathbf{Y}}\right) \cdot \left({}^{\mathrm{d}\mathbf{Y}}\!/_{\mathrm{d}\mathbf{C}^{n+1}}\right). \tag{55}$$

The effective plastic strain rate follows the integration of (46):

$$\varepsilon_p^{n+1} = \varepsilon_p^n + \Delta\gamma \frac{\boldsymbol{T}_e : \boldsymbol{N}(\boldsymbol{T}_e)}{\sigma_y}. \tag{56}$$

4.3 Specific Yield Function

The nondimensional yield function is given by:

$$\phi_\star(\boldsymbol{T}_e) = \frac{\sigma_{\mathrm{eq}}(\boldsymbol{T}_e)}{\sigma_y} - 1 \tag{57}$$

where, as a prototype equivalent stress, a specific Hill48 criterion (1948 [51]) is adopted. The general form of the Hill48 equivalent stress σ_{eq} is written as:

$$\sigma_{\mathrm{eq}}(\boldsymbol{T}_e) = \Big[F(T_{22} - T_{33})^2 + G(T_{33} - T_{11})^2 + H(T_{11} - T_{22})^2 \tag{58}$$

$$+ 2S_1\left(T_4^s\right)^2 + 2S_2\left(T_5^s\right)^2 + 2S_3\left(T_6^s\right)^2\Big]^{1/2} \tag{59}$$

where the subscript e of $\boldsymbol{T}_e$ is omitted for conciseness. In (58), the superscript s is adopted to indicate a symmetrized quantity. For example, $T_6^s = {}^1\!/_2\,(T_{23} + T_{32})$. Introducing the yield ratios, $\boldsymbol{y} = \{y_1, \ldots, y_6\}$ as constitutive data, we have for F, G, H, $S_{1,\ldots,3}$:

$$F = \frac{1}{2}\left({}^1\!/_{y_2^2} + {}^1\!/_{y_3^2} - {}^1\!/_{y_1^2}\right) \tag{60}$$

$$G = \frac{1}{2}\left({}^1\!/_{y_1^2} + {}^1\!/_{y_3^2} - {}^1\!/_{y_2^2}\right)$$

$$H = \frac{1}{2}\left({}^1\!/_{y_1^2} + {}^1\!/_{y_2^2} - {}^1\!/_{y_3^2}\right)$$

$$S_k = {}^3\!/_2(y_{k+3}^2) \quad k = 3, \ldots, 6.$$

We note that many other yield criteria can be used, since any specific form of $\sigma_{\mathrm{eq}}(\boldsymbol{T}_e)$ can inserted.

5 Numerical Tests

Numerical tests were performed with the code from the leading author, SimPlas [52], and the specific source code for $\overline{C}-$EFG was created using Mathematica [38] with the AceGen add-on [36, 39]. Source code for the equations in this work is available via GitHub [40].

5.1 Straight Cantilever Beam with Closed-Form Solution

We start with the Timoshenko and Goodier [53] cantilever beam in small strain elasticity. Two values of the Poisson coefficient are adopted: $\nu = 0.3$ and $\nu = 0.49999$. The quasi-incompressible case is here specified with a plane strain assumption. A comparison with the MINI element by D. Arnold [5] is performed. The beam is represented in Fig. 1.

We use the first slope boundary condition by Timoshenko and Goodier [53] who obtained the solution for the displacement in the plane stress case:

$$\boldsymbol{u}(x, y) = \frac{P}{4c^3 E} \left\{ \begin{array}{c} \frac{E}{G}\left(y^3 - 3c^2 y\right) + 3y(l-x)(l+x) - \nu y^3 \\ (l-x)^2(2l+x) + 3\nu x y^2 \end{array} \right\}.$$

Introducing this solution into the strain components and making use of Hooke's law, we calculate the strain energy per unit thickness as:

$$U = \frac{1}{2}\int_0^l \left[\int_{-c}^{+c} \left(\varepsilon_{xx}\sigma_{xx} + \varepsilon_{yy}\sigma_{yy} + \gamma_{xy}\tau_{xy}\right) \mathrm{d}y\right] \mathrm{d}x. \quad (61)$$

The plane strain case is obtained replacing E by $E/1-\nu^2$ and ν by $\nu/1-\nu$. Strain energy per unit thickness is given by:

$$U_{\text{plane strain}} = \frac{P^2\left[6c^2 E l + 5Gl^3\left(1-\nu^2\right)\right]}{20c^3 GE}.$$

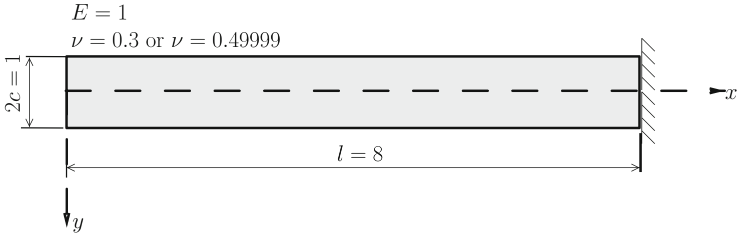


Fig. 1 Timoshenko and Goodier cantilever beam [53] with fixed support

Table 1 Closed-form and converged solutions for the cantilever beam

Closed-form solutions (specialized from [53])				
	Plane stress		Plane strain	
Poisson coefficient:	$v(0)$	U	$v(0)$	U
$\nu = 0.3$	2048	1036.48	1863.68	944.32
$\nu = 0.49999$	2048	1038.40	1536.02	782.41
Converged solutions ($h = 0.005$)				
	Plane stress		Plane strain	
Poisson coefficient:	$v(0)$	U	$v(0)$	U
$\nu = 0.3$	⋆	⋆	1880.90	940.33
$\nu = 0.49999$	⋆	⋆	1542.00	770.08

$$U_{\text{plane stress}} = \frac{P^2\left(6c^2El + 5Gl^3\right)}{20c^3GE}.$$

Results are given in Table 1 for both cases. This table also shows the converged results for $h = 0.005$ obtained with a mixed finite element formulation [52]. Only the plane strain case will be addressed, since it is more demanding in terms of convergence.

Displacement results as a function of the characteristic mesh size h are summarized in Table 2, with the following cases being considered:

1. $\nu = 0.3$ with full quadrature (3 Gauss points per triangle for both $\boldsymbol{C}$ and $\overline{\boldsymbol{C}}$).
2. $\nu = 0.49999$ with full quadrature.
3. $\nu = 0.49999$ with selective quadrature (3 Gauss points per triangle for $\boldsymbol{C}$ and 1 Gauss point for $\overline{\boldsymbol{C}}$).

The following notation is adopted for the polynomials:

1. $1 \leq p_0 \leq 3$ is the degree of polynomial adopted for $\overline{\boldsymbol{C}}$.
2. $1 \leq p_1 \leq 3$ with $p_1 \geq p_0$ is the degree of polynomial adopted for $\boldsymbol{C}$.

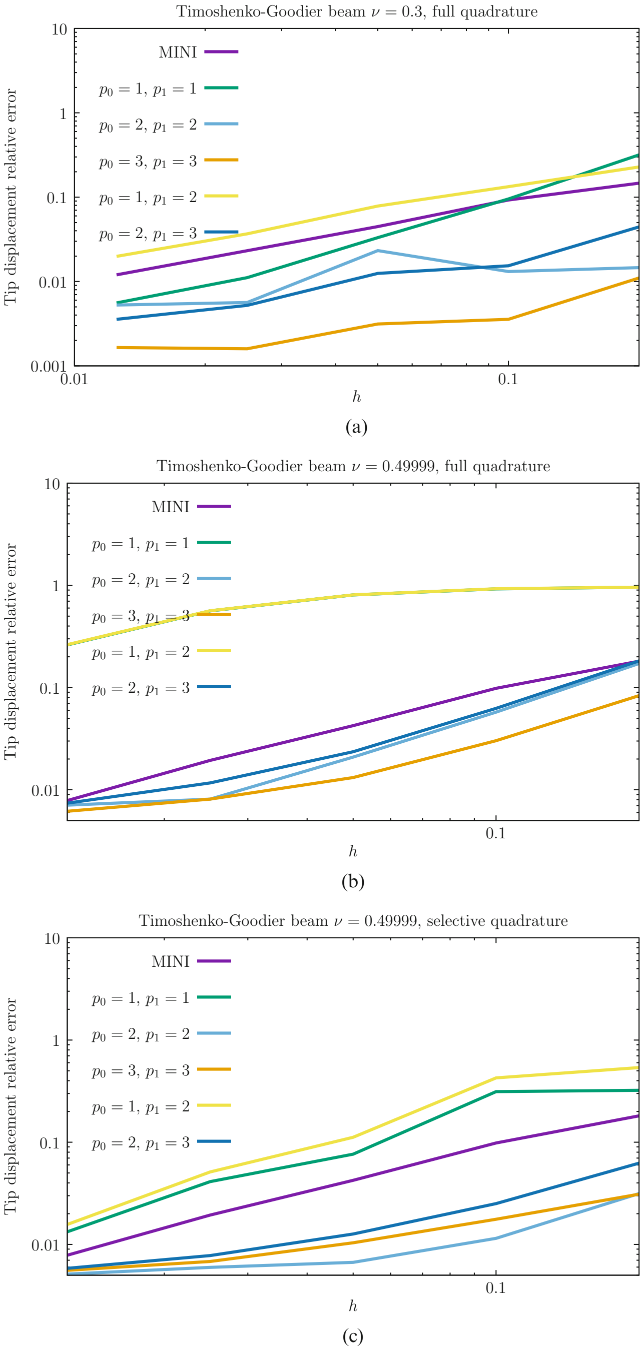
From the observation of Table 2, we conclude that:

1. For the compressible case, all formulations behave acceptably, with the exception of $p_0 = p_1 = 1$ which results in excessive displacements. In addition, with $p_0 = p_1 = 2$, we can conclude that results are non-monotonous.
2. Using full quadrature for the quasi-incompressible case, two combinations exhibit severe volumetric locking: $p_0 = p_1 = 1$ and $p_0 = 1$, $p_1 = 2$.
3. Using selective quadrature for the quasi-incompressible case only $p_0 = 1$, $p_1 = 2$ exhibits locking. Both $p_0 = p_1 = 3$ and $p_0 = 2$ and $p_1 = 3$ are acceptable formulations.

Tip displacement error convergence is shown in Fig. 2 for $\nu = 0.3$ and $\nu = 0.49999$. The latter is considered with full and selective quadrature. Energy error convergence is determined for both the compressible and quasi-incompressible cases in Table 3 for $p_0 = 2$ and $p_1 = 3$.

Table 2 Timoshenko and Goodier cantilever beam: numerical results for $v(0)$

$\nu = 0.3$							
h	T3	MINI	$p_0 = 1, p_1 = 1$	$p_0 = 2, p_1 = 2$	$p_0 = 3, p_1 = 3$	$p_0 = 1, p_1 = 2$	$p_0 = 2, p_1 = 3$
0.0125	1880.3	1903.5	1891.4	1890.8	1884.0	1918.3	1887.6
0.0250	1878.5	1924.5	1901.8	1891.5	1883.9	1950.0	1890.7
0.0500	1871.5	1965.2	1943.3	1922.4	1886.8	2028.7	1904.4
0.1000	1842.4	2055.1	2059.9	1905.6	1887.6	2131.3	1909.8
0.2000	1758.6	2156.8	2478.1	1908.4	1901.6	2310.1	1964.4
$v = 0.49999$ full quadrature							
h	T3	MINI	$p_0 = 1, p_1 = 1$	$p_0 = 2, p_1 = 2$	$p_0 = 3, p_1 = 3$	$p_0 = 1, p_1 = 2$	$p_0 = 2, p_1 = 3$
0.0125	606.8	1554.1	1138.6	1531.1	1532.5	1136.4	1530.6
0.0250	277.4	1571.9	674.1	1525.5	1529.5	671.5	1524.0
0.0500	208.8	1607.4	299.4	1509.8	1521.7	297.7	1505.7
0.1000	185.1	1693.7	116.9	1453.4	1495.4	114.2	1445.4
0.2000	166.9	1821.2	62.1	1277.0	1413.4	61.9	1260.8
$v = 0.49999$ selective quadrature							
h	T3	MINI	$p_0 = 1, p_1 = 1$	$p_0 = 2, p_1 = 2$	$p_0 = 3, p_1 = 3$	$p_0 = 1, p_1 = 2$	$p_0 = 2, p_1 = 3$
0.0125	606.8	1554.1	1521.5	1534.1	1533.4	1517.8	1533.0
0.0250	277.4	1571.9	1478.5	1532.8	1531.5	1463.1	1530.0
0.0500	208.8	1607.4	1424.1	1531.7	1526.0	1369.6	1522.5
0.1000	185.1	1693.7	1059.6	1524.3	1514.8	885.0	1503.3
0.2000	166.9	1821.2	1045.3	1493.5	1494.2	712.5	1445.8


Timoshenko-Goodier beam $\nu = 0.3$, full quadrature

(a)

Timoshenko-Goodier beam $\nu = 0.49999$, full quadrature

(b)

Timoshenko-Goodier beam $\nu = 0.49999$, selective quadrature

(c)

Fig. 2 Timoshenko-Goodier cantilever beam: tip displacement convergence for $\nu = 0.3$ and $\nu = 0.49999$. Results from the MINI element [5] are also included for comparison. (**a**) $\nu = 0.3$, full quadrature. (**b**) $\nu = 0.49999$, full quadrature. (**c**) $\nu = 0.49999$, selective quadrature

Table 3 Timoshenko and Goodier cantilever beam: numerical results for U in the plane strain case with selective quadrature and $p_0 = 2$ and $p_1 = 3$

h	$\nu = 0.3$	$\upsilon = 0.49999$
0.0125	943.7	766.5
0.0250	945.1	765.2
0.0500	949.5	762.1
0.1000	957.1	756.3
0.2000	981.8	743.5

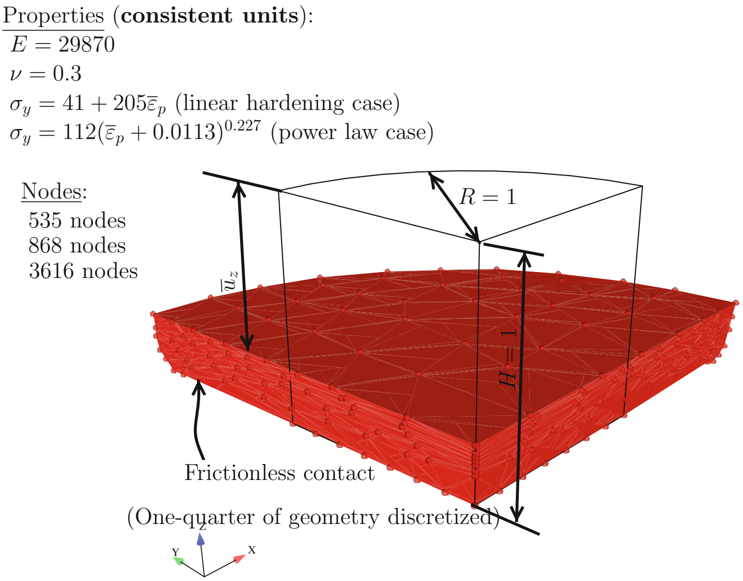

Properties (**consistent units**):
$E = 29870$
$\nu = 0.3$
$\sigma_y = 41 + 205\bar{\varepsilon}_p$ (linear hardening case)
$\sigma_y = 112(\bar{\varepsilon}_p + 0.0113)^{0.227}$ (power law case)
Nodes:
535 nodes
868 nodes
3616 nodes
$\bar{u}_z$
$R = 1$
$H = 1$
Frictionless contact
(One-quarter of geometry discretized)
Z
Y
X

Fig. 3 Billet upsetting test: relevant data and notation

5.2 *Billet Upsetting Test*

We make use of the upsetting test reported by M.A. Puso and J. Solberg [54] in its two elastoplastic versions (linear and power hardening). Geometry, boundary conditions, and constitutive properties are shown in Fig. 3. Three uniform meshes are adopted for comparison, containing 535, 868, and 3616 nodes. Nodes are forced to remain above a horizontal plane by a non-penetration condition. Of the two cases reported in [54], the elastoplastic case described is the most demanding, and it was found that only their nodal integrated and stabilized UT4s provided stable and accurate results. Using a cubic basis ($p_0 = p_1 = 3$), Fig. 4 shows the very smooth contour plots for ε_p and hydrostatic σ_H. All three factors contribute to a more flexible behavior: finer meshes are less stiff, larger supports produce softer behavior, and quadratic basis produces results beneath the reaction displacement curve reported in [54]. Using uniform quadrature and uniform interpolation, Fig. 5 shows the results compared with the reported in [54]. We test three basis dimensions:

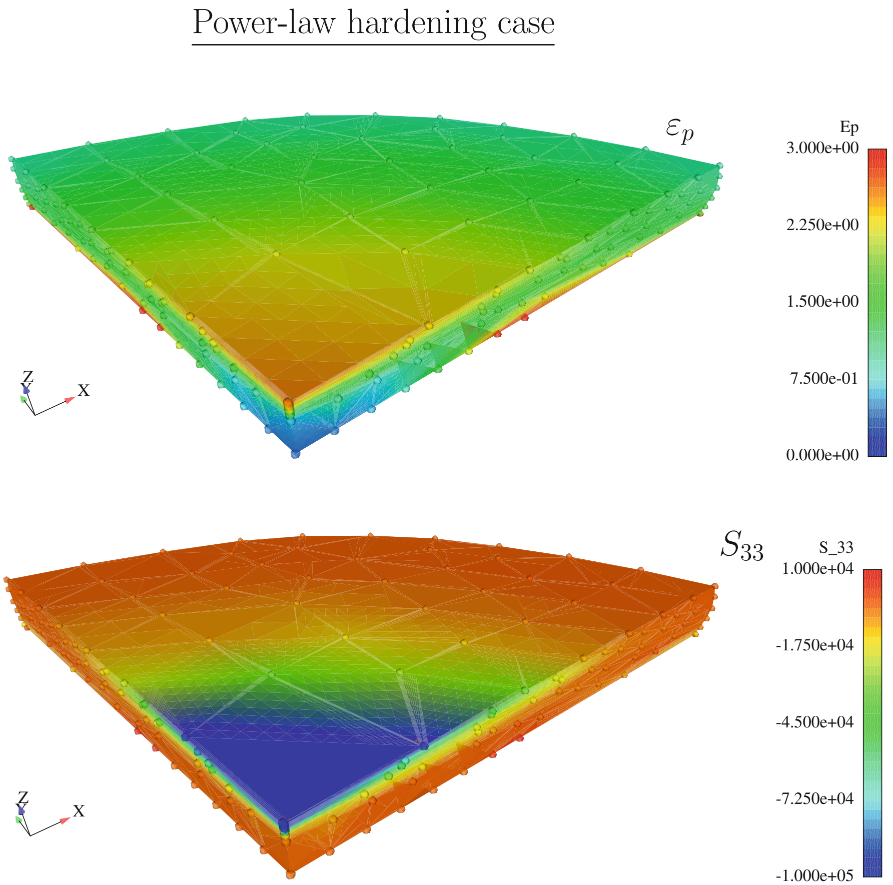


Fig. 4 Upsetting test: contour plots (ε_p, S_{33}) for the power-law hardening case and 535 nodes

linear, quadratic, and cubic with $n = 25$. In terms of quadrature, both 1 and 4 Gauss points are tested. Reduced quadrature produces exceedingly flexible results, as shown in Fig. 5. This conclusion leads us to favor either full quadrature (4 points in both terms) or selective quadrature (4 points for the deviatoric terms and 1 point for the volumetric term). Focusing on the polynomial bases, Fig. 6 shows the effect of p_0 and p_1 on the displacement-reaction behavior. The following conclusions are taken:

- In contrast with displacement-based finite elements, increasing the polynomial degree does not produce more flexible results.
- In contrast with finite elements, uniform reduced quadrature does not produce hourglassing/point instabilities. However, significant loss of stiffness is observed, which precludes its use in the quasi-incompressible case.

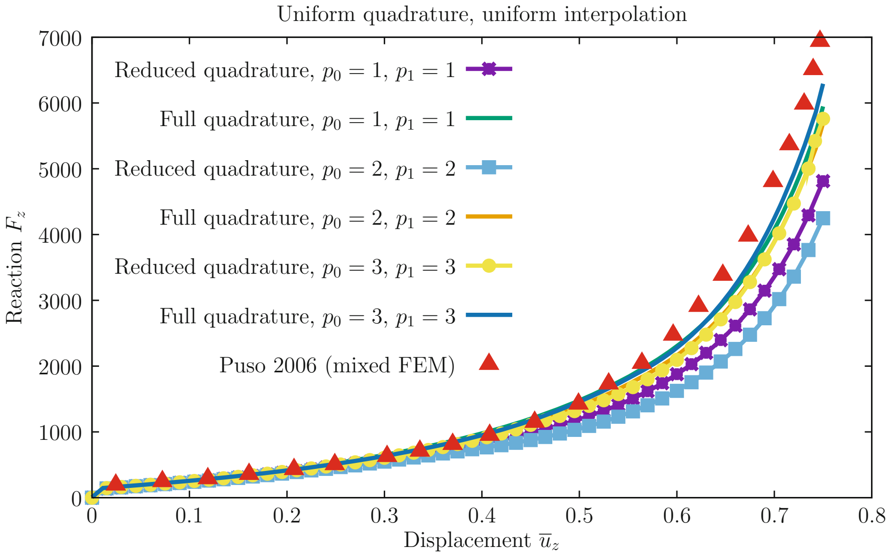


Fig. 5 Upsetting test, linear hardening, $n = 25$: effect of dimension of polynomial basis for uniform quadrature/uniform basis

- The polynomial degree of the deviatoric term, p_1, is important in terms of results.

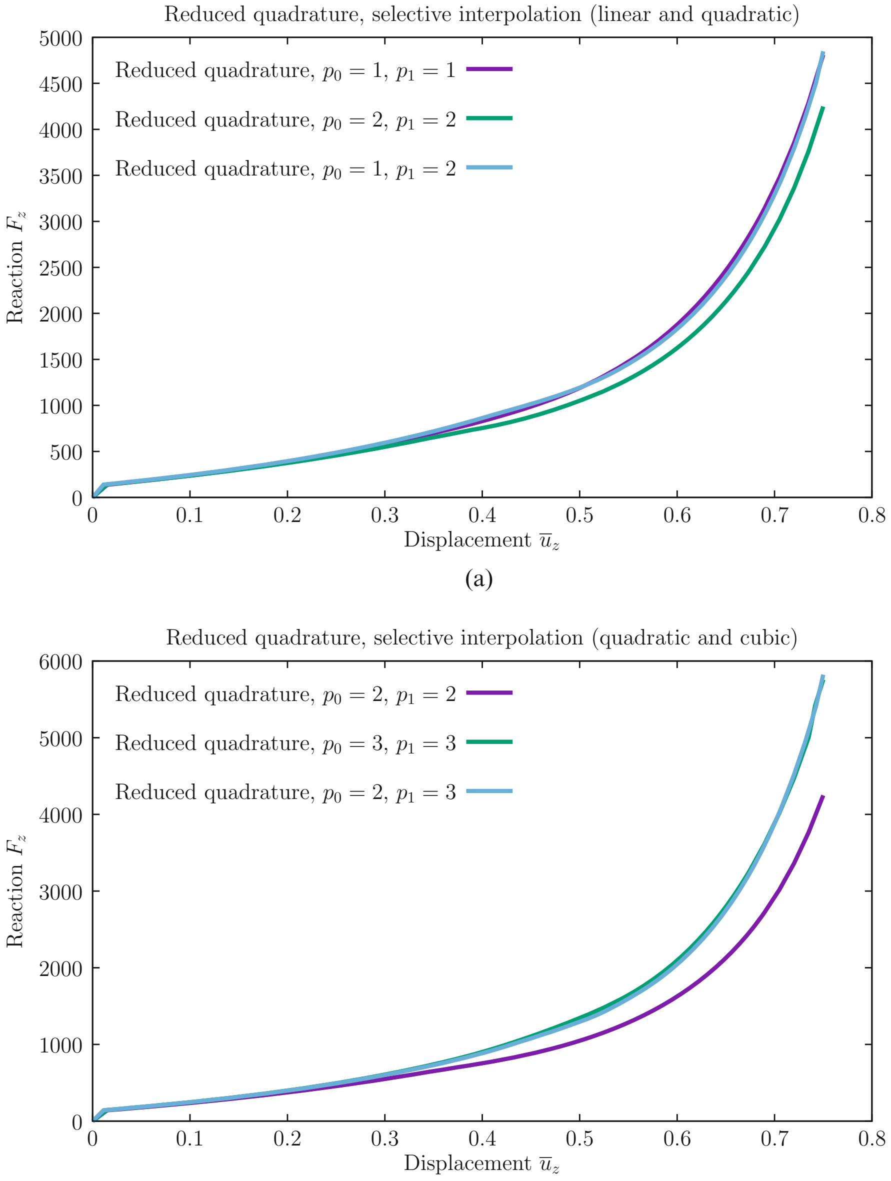
Selective interpolation is now contemplated, with Fig. 7 showing the effect of combining distinct bases. We conclude that the deviatoric term p_1 is crucial for the results.

Combining selective quadrature with full interpolation, results show that significant differences exist by changing the basis (see Fig. 8). In terms of mesh convergence, excellent results are obtained, as Fig. 9 shows. In our experience, this is one of the advantages of meshless methods.

Finally, to complete the test of Puso and Solberg [54], the power-law hardening is tested in Fig. 10

5.3 Tension Test

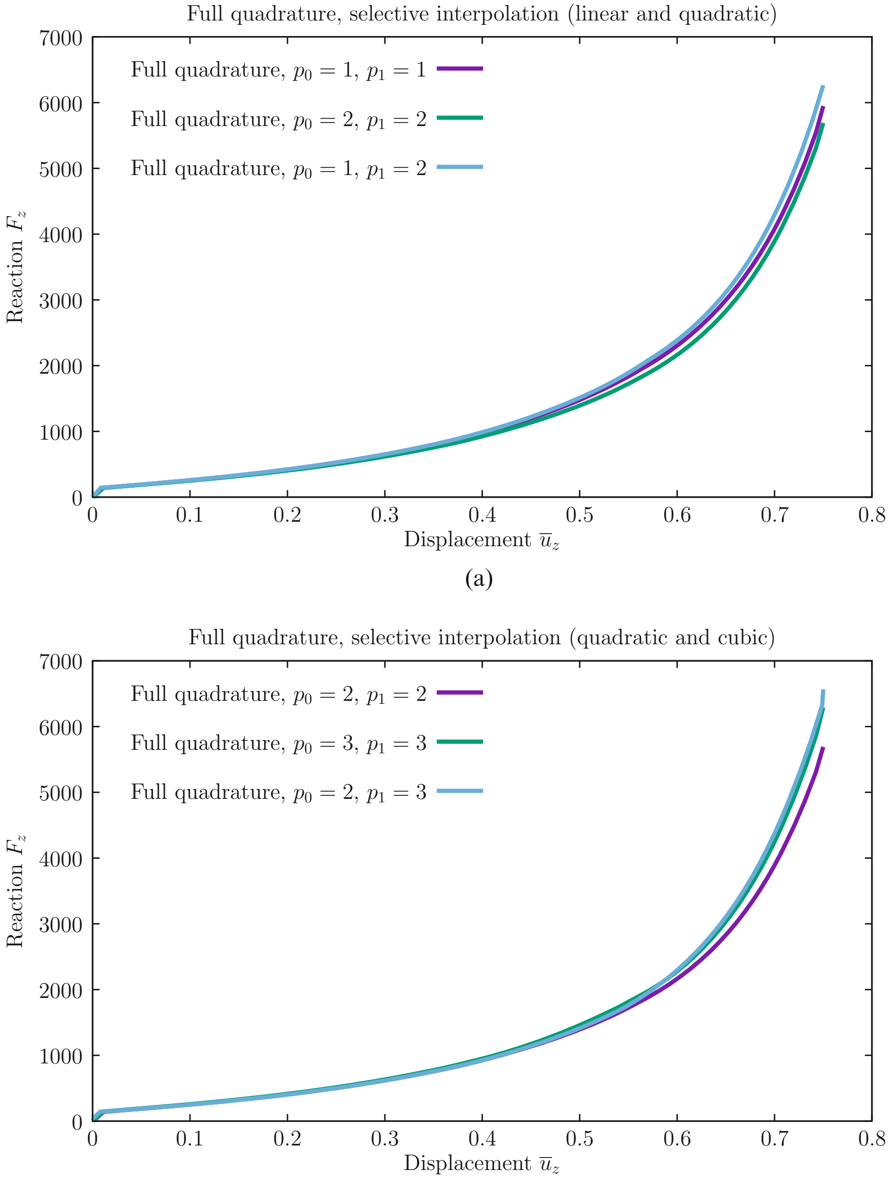
We apply the $\overline{C}$-EFG method to the tension test discussed by Simo and co-workers in the context of J_2 plasticity [9] (see also the 1993 reference [55] where the test is described in detail). Geometry, boundary conditions, and material properties are summarized in Fig. 11, along with the two cases of nodal distribution, structured and unstructured, as this was found to have an effect on the results. The contour plot



(a)

(b)

Fig. 6 Upsetting test, $n = 25$, effect of selective polynomial basis for the deviatoric (p_1) and volumetric (p_0) terms. Reduced quadrature. (**a**) Linear and quadratic bases. (**b**) Quadratic and cubic bases. (**c**)



(a)

(b)

Fig. 7 Upsetting test, $n = 25$, effect of selective polynomial basis for the deviatoric (p_1) and volumetric (p_0) terms. Full quadrature. (**a**) Linear and quadratic bases. (**b**) Quadratic and cubic bases

of the effective plastic strain, given by Eq. (56), is shown in Fig. 12 for two values of $\boldsymbol{y}$. The specific yield stress σ_y is given by the hardening law shown in Fig. 11.

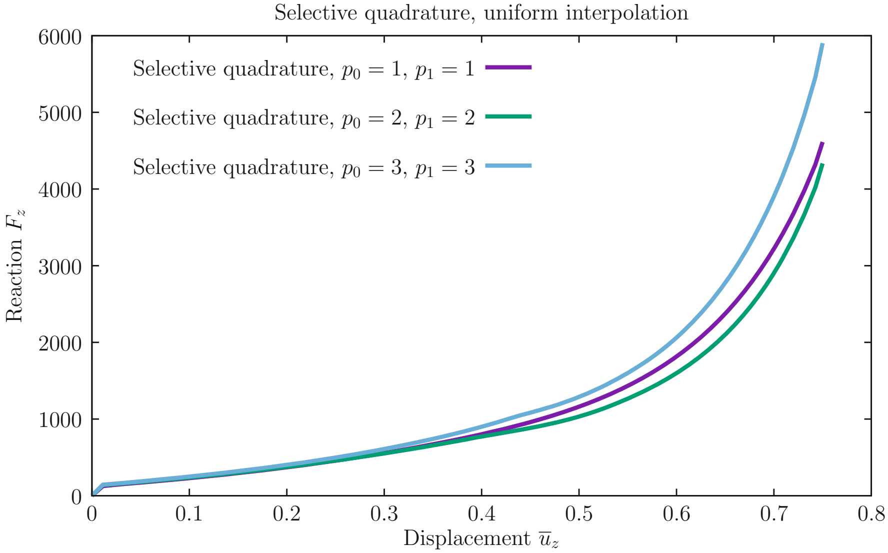


Fig. 8 Upsetting test, $n = 25$, effect of selective quadrature for uniform interpolation

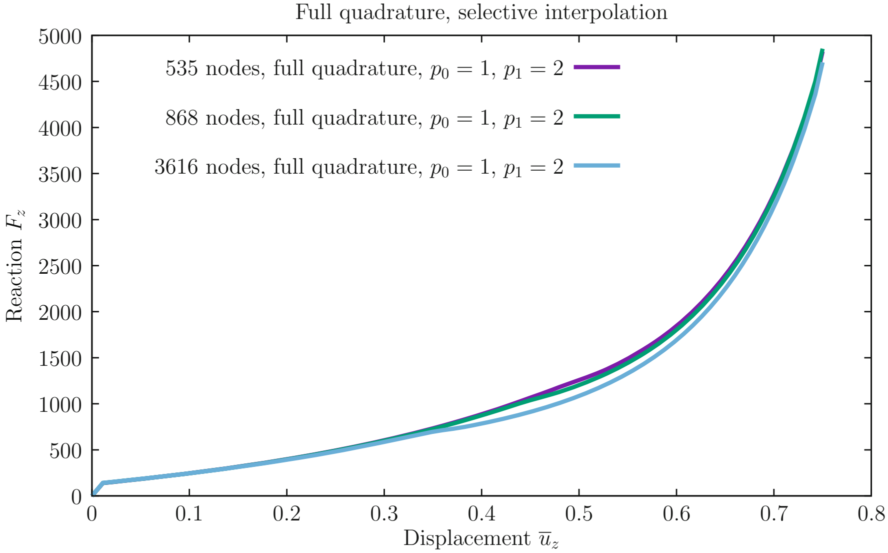


Fig. 9 Effect of a number of nodes using full quadrature and selective interpolation

Compared to mixed FE formulations, results are distinct. When compared with enhanced assumed strain hexahedra, specifically Simo and Armero [7, 55], both the initial plastic behavior and the post-localization behavior are different (see Fig. 13). We note that two significant differences exist: (1) Simo and Armero adopted a

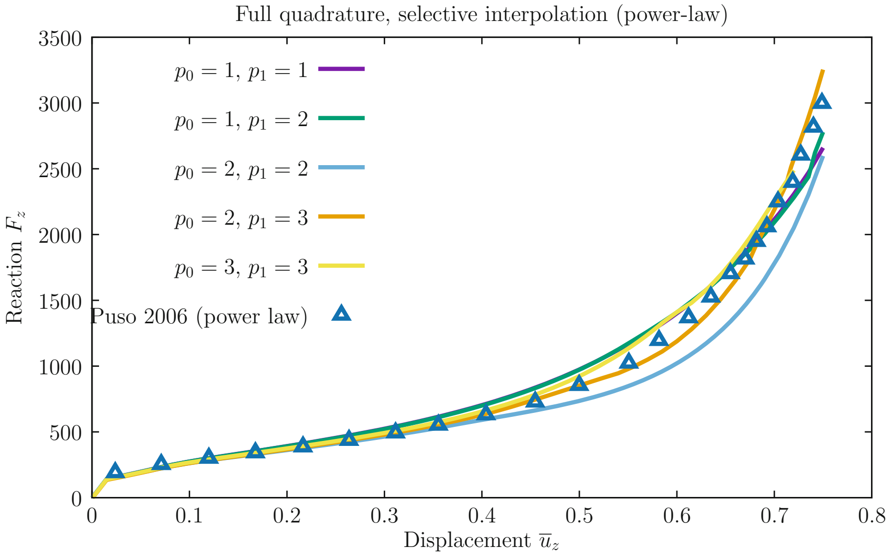


Fig. 10 Results for power-law hardening

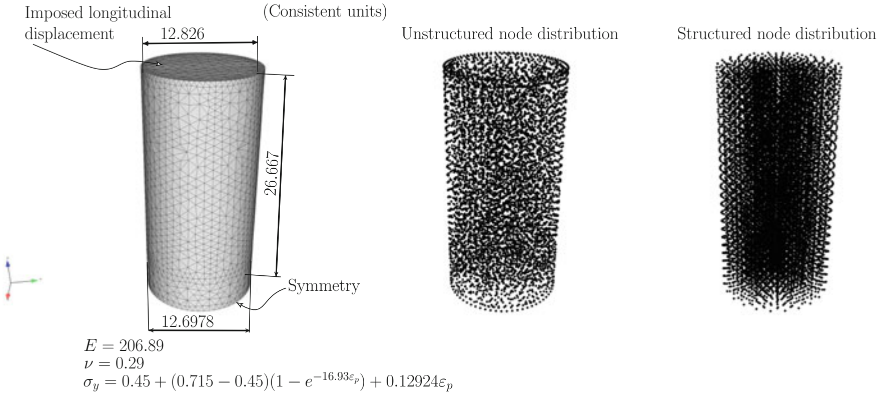

Imposed longitudinal
displacement
(Consistent units)
12.826
26.667
Symmetry
12.6978
$E = 206.89$
$\nu = 0.29$
$\sigma_y = 0.45 + (0.715 - 0.45)(1 - e^{-16.93\varepsilon_p}) + 0.12924\varepsilon_p$
Unstructured node distribution
Structured node distribution

Fig. 11 Relevant dimensions and mesh for the Neo-Hookean/Hill48 tension test

formulation based on the Kirchhoff stress tensor and radial-return mapping for J_2 plasticity and (2) hexahedra tend to reproduce the incompressibility condition with sharper stretching. MINI elements (see, [5]) are also used for comparison, as Fig. 13 shows. When compared with the MINI runs, much coarser meshes are used in EFG for similar results. In contrast with the previous examples, finer node distributions result in a sharper localization region, with lower reactions for higher displacements. For the structured mesh with 3760 nodes, Fig. 14a shows the advantages of using $p_0 = 2$ and $p_1 = 3$ in terms of post-localization. When adopting an unstructured

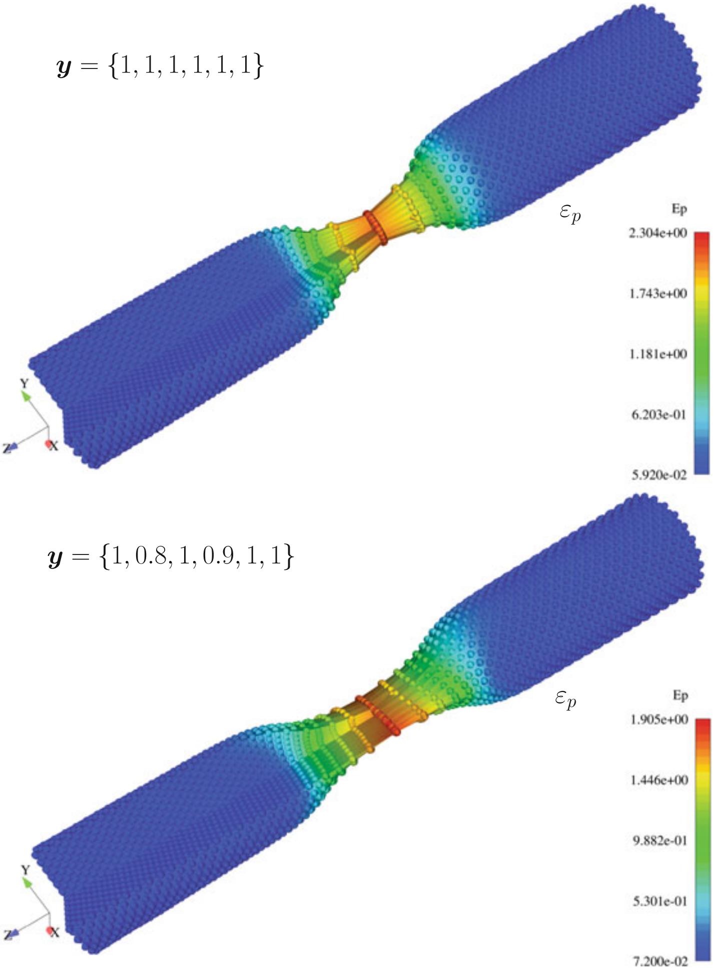


Fig. 12 Tension test: deformed configurations for both yield functions ({1, 1, 1, 1, 1, 1} and {1, 0.8, 1, 0.9, 1, 1} with the corresponding effective plastic strain colors

node distribution, a less pronounced post-localization behavior is exhibited (see Fig. 14b).

6 Conclusions

In the context of $\overline{C}$ decomposition and by parameterizing the quadrature and the degree of the polynomial basis, we developed a discretization scheme with the following distinctive features:

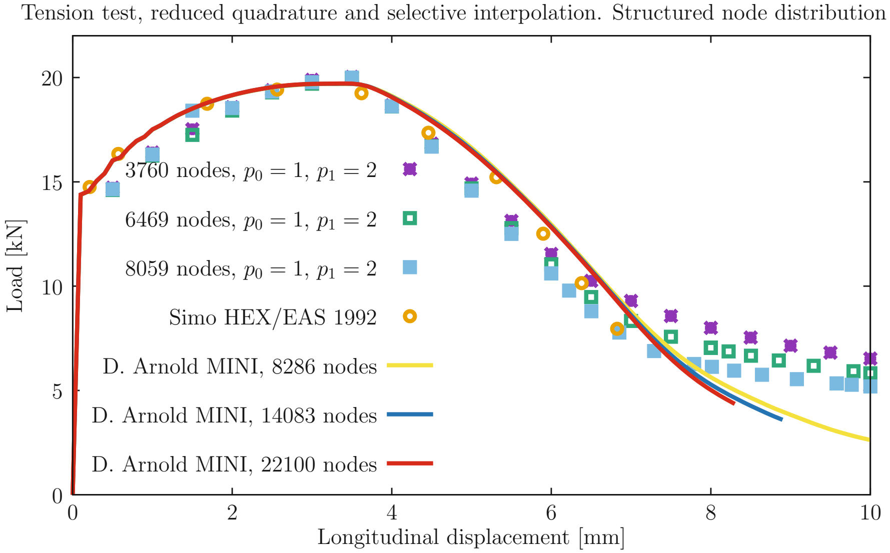
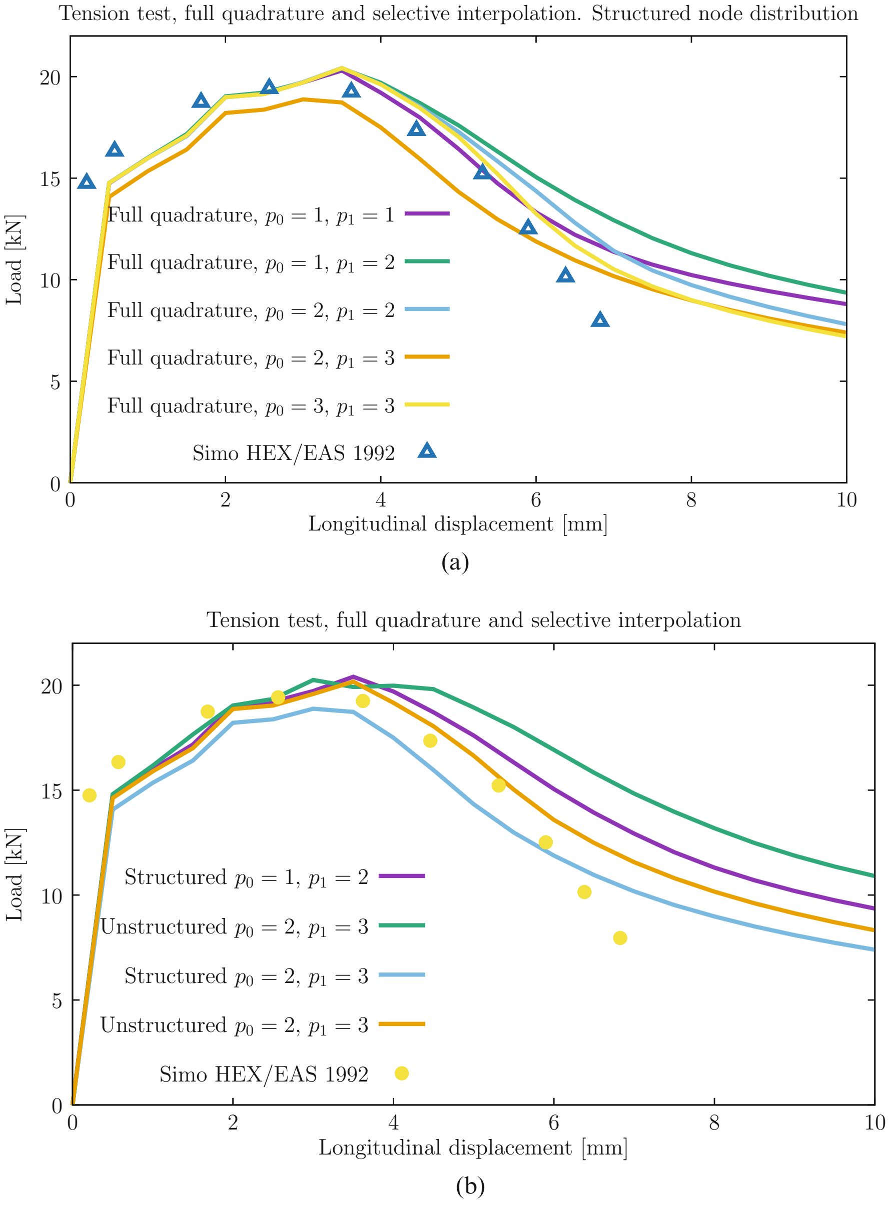


Fig. 13 Comparison with advanced finite element technology [7, 55] and effect of node density on the results

- An initial perturbation of internal FE nodal positions is performed for efficiency reasons (low n).
- From linear up to cubic shape, functions are adopted for the volumetric and deviatoric terms of the right Cauchy-Green tensor. Lagrangian diffuse derivatives are defined ab initio for the entire analysis.
- A pre-established nodal support is imposed, and a tetrahedra integration with 1 or 4 quadrature points for $\boldsymbol{C}$ and $\overline{\boldsymbol{C}}$ is adopted.
- Constitutive integration makes use of the Mandel stress tensor and iteration on $\boldsymbol{C}_e$ [16].

Implementation is straightforward and was performed in SimPlas [52] with AceGen [39] and Mathematica [38]. Three benchmark tests were performed, which allow the following conclusions:

- Even with small supports and coarse meshes, results are highly competitive with established finite elements if either selective interpolation or selective quadrature are adopted. This holds for the quasi-incompressible case where special finite elements are adopted.
- Numerical testing shows that the ideal combination is $p_0 = 2$ and $p_1 = 3$ with either selective or full quadrature.
- Finite strain plasticity solutions are very robust, with large strains being possible without loss of convergence or instabilities.
- The finite strain formulation is simpler than with mixed finite elements and on par with displacement-based FEM. Source code is available at GitHub, cf. [40].



(a)

(b)

Fig. 14 Effect of selective interpolation and structured/unstructured node distribution. (**a**) Effect of selective interpolation. (**b**) Effect of structured/unstructured node distribution

Acknowledgments The authors acknowledge the support of **FCT**, through **IDMEC**, under **LAETA**, project **UIDB/50022/2020.**

Appendix

First and Second Variations of det $[\boldsymbol{C}]$

For the determinant of $\boldsymbol{C}$, the following relations hold, which follow from established results and application of the chain rule:

$$\delta \det[\boldsymbol{C}] = \det[\boldsymbol{C}]\,\boldsymbol{C}^{-1} : \delta\boldsymbol{C} \tag{62}$$

$$\begin{aligned} \mathrm{d}\delta \det[\boldsymbol{C}] = {} & \mathrm{d}\boldsymbol{C} : \left(\det[\boldsymbol{C}]\,\boldsymbol{C}^{-1} \otimes \boldsymbol{C}^{-1}\right) : \delta\boldsymbol{C} \\ & + \det[\boldsymbol{C}] : \mathrm{d}\delta\boldsymbol{C} - \mathrm{d}\boldsymbol{C} : \mathsf{S} : \delta\boldsymbol{C} \end{aligned} \tag{63}$$

where S is a fourth-order tensor with components $\mathsf{S}_{klij} = C^{-1}_{ik} C^{-1}_{lj}$. Given the $n-$th power of det $[\boldsymbol{C}]$, det $[\boldsymbol{C}]^n$, it follows that:

$$\delta \det[\boldsymbol{C}]^n = n \det[\boldsymbol{C}]^{n-1}\,\delta \det[\boldsymbol{C}] \tag{64}$$

$$\mathrm{d}\delta \det[\boldsymbol{C}]^n = n(n-1)\det[\boldsymbol{C}]^{n-2}\,\delta \det[\boldsymbol{C}]\,\mathrm{d}\det[\boldsymbol{C}] + n\det[\boldsymbol{C}]^{n-1}\,\mathrm{d}\delta\det[\boldsymbol{C}] \tag{65}$$

with $\delta \det[\boldsymbol{C}]$ being given by (62) and $\mathrm{d}\delta \det[\boldsymbol{C}]$ by (63). Given $\theta_C = \det\left[\overline{\boldsymbol{C}}\right]$, similar expressions are obtained for $\delta\theta_C$ and $\mathrm{d}\delta\theta_C$.

References

1. N.S. Lee, K.J. Bathe, Effects of element distortions on the performance of isoparametric elements. Int. J. Numer. Meth. Eng. **36**, 3553–3576 (1993)
2. J. Dolbow, N. Moës, T. Belytschko, Modeling fracture in Mindlin-Reissner plates with the extended finite element method. Int. J. Solids Struct. **37**, 7161–7183 (2000)
3. X. Peng, E. Atroshchenko, P. Kerfriden, S.P.A. Bordas, Isogeometric boundary element methods for three dimensional static fracture and fatigue crack growth. Comp. Method. Appl. Mech. Eng. **316**, 151–185 (2017)
4. K.-J. Bathe, *Finite Element Procedures* (Prentice-Hall, Hoboken, 1996)
5. D.N. Arnold, F. Brezzi, M. Fortin, A stable finite element for the Stokes equations. Calcolo **XXI**(IV), 337–344 (1984)
6. J.C. Simo, R.L. Taylor, K.S. Pister, Variational and projection methods for the volume constraint in finite deformation elastoplasticity. Comp. Method. Appl. Mech. Eng. **51**, 177–208 (1985)
7. J.C. Simo, F. Armero, Geometrically non-linear enhanced strain mixed methods and the method of incompatible modes. Int. J. Numer. Meth. Eng. **33**, 1413–1449 (1992)
8. T. Rabczuk, T. Belytschko, S.P. Xiao, Stable particle methods based on Lagrangian kernels. Comp. Method Appl. Mech. Eng. **193**, 1035–1063 (2004)

9. J.C. Simo, Algorithms for static and dynamic multiplicative plasticity that preserve the classical return mapping schemes of the infinitesimal theory. Comp. Method Appl. Mech. Eng. **99**, 61–112 (1992)
10. J.C. Simo, T.J.R. Hughes, *Computational Inelasticity*, Corrected Second Printing Edition (Springer, Berlin, 2000)
11. R. Rossi, M.K. Alves, On the analysis of an EFG method under large deformations and volumetric locking. Comput. Mech. **39**, 381–399 (2007)
12. J.-S. Chen, M. Hillman, S.-W. Chi, Meshfree methods: Progress made after 20 years. J. Eng. Mech-ASCE **143**(4), 04017001 (2017)
13. Y. Cai, X. Zhuang, C. Augarde, A new partition of unity finite element free from the linear dependence problem and possessing the delta property. Comp. Method Appl. Mech. Eng. **199**, 1036–1043 (2010)
14. B. Boroomand, S. Parand, Towards a general interpolation scheme. Comp. Method Appl. Mech. Eng. **381**, 113830 (2021)
15. G. Bourantas, B.F. Zwick, G.R. Joldes, A. Wittek, K. Miller, Simple and robust element-free Galerkin method with almost interpolating shape functions for finite deformation elasticity. Appl. Math. Model **96**, 284–303 (2021)
16. P. Areias, T. Rabczuk, J. Ambrósio, Extrapolation and c_e-based implicit integration of anisotropic constitutive behavior. Int. J. Numer. Meth. Eng. **122**, 1218–1240 (2021)
17. A. Huerta, S.F. Méndez, Locking in the incompressible limit for the element-free galerin method. Int. J. Numer. Meth. Eng. **51**, 1361–1383 (2001)
18. P. Lancaster, K. Salkauskas, Surfaces generated by moving least squares methods. Math. Comput. **37**(155), 141–158 (1981)
19. T. Most, C. Bucher, A moving least squares weighting function for the element-free Galerkin method which almost fulfills essential boundary conditions. Struct. Eng. Mech. **21**(3), 315–332 (2005)
20. T. Most, C. Bucher, New concepts for moving least squares: an interpolation non-singular weighting function and weighted nodal least squares. Eng. Anal. Bound Elem. **32**, 461–470 (2008)
21. J. Dolbow, T. Belytschko, Volumetric locking in the element free galerkin method. Int. J. Numer. Meth Eng. **46**, 925–942 (1999)
22. J.-S. Chen, S. Yoon, H.-P. Wang, W.K. Liu, An improved reproducing kernel particle method for nearly incompressible finite elasticity. Comp. Method Appl. Mech. Eng. **181**, 117–145 (2000)
23. C.-T. Wu, S.-W. Chi, M. Koishi, Y. Wu, Strain gradient stabilization with dual stress points for the meshfree nodal integration method in inelastic analyses. Int. J. Numer. Meth Eng. **107**, 3–30 (2016)
24. D.P. Recio, R.M. Natal Jorge, L.M.S. Dinis, Locking and hourglass phenomena in an element-free Galerkin context: the B-bar method with stabilization and an enhanced strain method. Int. J. Numer. Meth Eng. **68**, 1329–1357 (2006)
25. W.M. Coombs, T.J. Charlton, M. Cortis, C.E. Augarde, Overcoming volumetric locking in material point methods. Comp. Method Appl. Mech. Eng. **333**, 1–21 (2018)
26. G. Moutsanidis, J.J. Koester, M.R. Tupek, J.-S. Chen, Y. Bazilevs, Treatment of near-incompressibility in meshfree and immersed-particle methods. Comput. Part Mech. **7**, 309–327 (2020)
27. P. Navas, S. López-Querol, R.C. Yu, B. Li, B-bar based algorithm applied to meshfree numerical schemes to solve unconfined seepage problems through porous media. Int. J. Numer. Meth Eng. **40**, 962–984 (2016)
28. T. Belytschko, Y.Y. Lu, L. Gu, Element-free galerkin methods. Int. J. Numer. Meth Eng. **37**, 229–256 (1994)
29. T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, P. Krysl, Meshless methods: an overview and recent developments. Comp. Method Appl. Mech. Eng. **139**, 3–47 (1996)
30. G.J. Golub, C.F. Van Loan, *Matrix Computations*, 3rd edn. (Johns Hopkins, Baltimore, 1996)
31. P. Areias, EFG MLS (2021). https://github.com/PedroAreiasIST/EFG

32. B. Nayroles, G. Touzot, P. Villon, Generalizing the finite element method: Diffuse. Comput. Mech. **10**, 307–318 (1992)
33. M. Dehghan, M. Abbaszadeh, Interpolating stabilized moving least squares (MLS) approximation for 2D elliptic interface problems. Comp. Method Appl. Mech. Eng. **328**, 775–803 (2018)
34. B. Moran, M. Ortiz, C.F. Shih, Formulation of implicit finite element methods for multiplicative finite deformation plasticity. Int. J. Numer. Meth Eng. **29**, 483–514 (1990)
35. J.C. Nagtegaal, D.M. Parks, J.R. Rice, On numerically accurate finite element solutions in the fully plastic range. Comp. Method Appl. Mech. Eng. **4**, 153–177 (1974)
36. P. Wriggers, *Nonlinear Finite Element Methods* (Springer, Berlin, 2008)
37. P.J. Flory, Elasticity of polymer networks cross-linked in states of strain. Trans. Faraday Soc. **56**, 722–743 (1960)
38. Wolfram Research Inc. Mathematica (2007)
39. J. Korelc, Multi-language and multi-environment generation of nonlinear finite element codes. Eng. Comput. **18**(4), 312–327 (2002)
40. P. Areias, F-bar in meshless (2020). https://github.com/PedroAreiasIST/fbar
41. T. Belytschko, W.K. Liu, B. Moran, *Nonlinear Finite Elements for Continua and Structures* (Wiley, Hoboken, 2000)
42. P. Areias, J.M.A. César de Sá, C.A. Conceição António, A.A. Fernandes, Analysis of 3D problems using a new enhanced strain hexahedral element. Int. J. Numer. Meth Eng. **58**, 1637–1682 (2003)
43. P. Areias, C. Tiago, J. Carrilho Lopes, F. Carapau, P. Correia, A finite strain Raviart-Thomas tetrahedron. Eur. J. Mech. A-Solid **80**, 103911 (2020)
44. J. Mandel, Equations constitutives et directeurs dans les milieux plastiques et viscoplastiques. Int. J. Solids Struct. **9**, 725–740 (1973)
45. B. Eidel, F. Gruttmann, Elastoplastic orthotropy at finite strains: multiplicative formulation and numerical implementation. Comput. Mater. Sci. **28**, 732–742 (2003)
46. E. Kröner, Allgemeine kontinuumstheorie der versetzungen und eigenspannungen. Arch. Ration Mech. Analy. **4**, 273–334 (1960)
47. E.H. Lee, Elasto-plastic deformation at finite strains. J. Appl. Mech-ASME **36**, 1–6 (1969)
48. J. Lubliner, *Plasticity Theory* (Macmillan, London, 1990)
49. M.E. Gurtin, *An Introduction to Continuum Mechanics*. Mathematics in Science and Engineering, vol. 158 (Academic, New York, 1981)
50. J. Mandel, *Foundations of Continuum Thermodynamics*, Chapter Thermodynamics and Plasticity (MacMillan, London, 1974), pp. 283–304
51. R. Hill, A theory of yielding and plastic flow of anisotropic metals. Proc. R. Soc. Lond. **193**, 281–297 (1948)
52. P. Areias, Simplas. http://www.simplassoftware.com. Portuguese Software Association (ASSOFT) registry number 2281/D/17.
53. S. Timoshenko, J.N. Goodier, *Theory of Elasticity*, 2nd edn. (McGraw-Hill Book Company, New-York, 1951)
54. M.A. Puso, J. Solberg, A stabilized nodally integrated tetrahedral. Int. J. Numer. Methods Eng. **67**, 841–867 (2006)
55. J.C. Simo, F. Armero, R.L. Taylor, Improved versions of assumed strain tri-linear elements for 3D finite deformation problems. Comp. Method Appl. Mech. Eng. **110**, 359–386 (1993)

Physics-Informed Bias Method for Multiphysics Machine Learning: Reduced Order Amyloid-β Fibril Aggregation

Joseph Pateras, Ashwin Vaidya, and Preetam Ghosh

1 Multiphysics Modeling

Deriving reduced order models for multiphysical systems is a hallmark of modern science which affords viability to the computational analysis of many physical, chemical, biological, geological, etc. systems. As system complexity grows, so too must the dimensionality of our models. Creating complex enough models for increasingly complex problems is a vexing concern for researchers. Take, for instance, the three-body problem [9]. In classical mechanics, it is easy enough to model the orbit of two bodies interacting with Newton's laws of motion and gravitation. However, add a third or perhaps n-many bodies, and the problem becomes substantially harder; difficulty increases so much so that this problem is an important consideration in space mission design [9]. By increasing our understanding of the three-body system's physics and tweaking our modeling approach, we can also create novel approaches to obtain better results and draw interesting conclusions [1, 12].

Generally, multiphysics problems are presented here as a large class of problems, which might be computationally difficult and are apt targets of the physics-informed ML (PIML) approach. There are innumerably many techniques to providing modeling solutions to multiphysics problems. Numerical solutions to PDEs are obtained using finite elements or differences, spectral, meshless, or any variety of methods. If many numerical methods must be reproduced for a large number of

J. Pateras · P. Ghosh (✉)
Virginia Commonwealth University, Richmond, VA, USA
e-mail: paterasj@vcu.edu; pghosh@vcu.edu

A. Vaidya
Montclair State University, Montclair, NJ, USA
e-mail: vaidyaa@montclair.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_7

iterations and/or on increasingly nonlinear, complex systems, the costs of traditional multiphysical modeling, discussed in Sect. 2, can become prohibitive to expanding research.

In this work, the focus is on a particular biophysical model with relevancy to the study of Alzheimer's disease. By applying the training set bias method outlined in Sect. 2, we are able to reliably reproduce previous numerical solutions [5, 6] in a considerably small fraction of the time and with improving accuracy as we introduce more information to the biasing method paradigm.

1.1 Amylod-β Fibril Aggregation

The runaway aggregation of toxic amyloid proteins is a condition underpinning many serious health conditions commonly referred to as amyloid diseases. One such protein, which aggregates in our brains near neurons, is called Amyloid-β (Aβ). Modeling the formation of these toxic plaques is important as they can only be observed directly in the human brain during an autopsy. This was exactly how in the year 1906 Dr. Alois Alzheimer noted the presence of these plaques in the brain of a patient exhibiting dementia-like symptoms now synonymous with Alzheimer's disease.

The model presented in Fig. 1 is a reduced order representation of Aβ aggregation. The entire amyloid system is extremely complex and a full-scale analysis is difficult to imagine. The model in Fig. 1 describes individual monomers of Aβ aggregating to form a nucleation size oligomer of size n and an ordered step of aggregation to form a full fibril containing m-many individual Aβ proteins. A_1 is the monomer species of Aβ. n-many A_1 proteins come together to form an experimentally observed intermediate nucleation size, A_n [4]. Each A also has a corresponding "prime" species, A', signifying an aggregation reaction has occurred in the presence of an environmental catalyst L, such as a fatty acid or a surfactant. Both healthy and toxic pathways culminate in a post-nucleation size oligomer of size m (A_m and A'_m, respectively)[2]. The reactions in Eqs. (1)–(6) describe the interactions between the differently sized aggregates of Aβ. Each double-ended arrow in Fig. 1 describes a reversible aggregation reaction or a switching between toxic and healthy pathways. It is observed that fully realized fibrils no longer mutate between healthy and toxic.

$$A_1 + L \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} A'_1, \tag{1}$$

$$n * A_1 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} A_n, \tag{2}$$

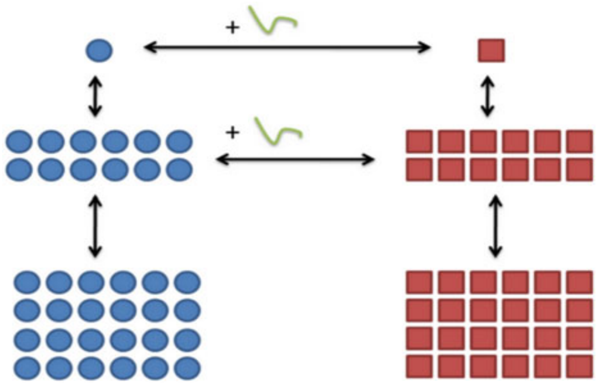


Fig. 1 A reduced order aggregation model for Aβ amyloid formation

$$n * A_1' \underset{k_3^-}{\overset{k_3^+}{\rightleftharpoons}} A_n', \tag{3}$$

$$A_n + L \underset{k_4^-}{\overset{k_4^+}{\rightleftharpoons}} A_n', \tag{4}$$

$$\frac{m}{n} * A_n \underset{k_5^-}{\overset{k_5^+}{\rightleftharpoons}} A_m, \tag{5}$$

$$\frac{m}{n} * A_n' \underset{k_6^-}{\overset{k_6^+}{\rightleftharpoons}} A_m'. \tag{6}$$

The reactions in Eqs. (1)–(6) are derived from the law of mass action, where the various k values describe the rates of each reaction. The law of mass action can be used to model many various chemical or physical systems like the previously presented Aβ model [6] or in a myriad of other diverse fields including but certainly not limited to particle flocking behavior in fluid-surface interactions [3], model-based epidemiology [13], or radioligand binding studies [10].

The general framework of a multiphysics problem apt for a PIML study is easily definable, is quite broad, and is certainly not limited to mass action models. The system of study can be:

1. Any system about which we have knowledge of some physical, mechanical, chemical, or properties otherwise
2. Any system which is traditionally modeled with complex multiphysical assumptions
3. Any system whose breadth of study is limited by availability of data and/or the computational expense of modeling

1.2 Complex Multiphysics Modeling Costs

Generally, solutions to highly nonlinear models introduce significant computational expense and fidelity concerns in numerical solutions. Researchers are widely creative in addressing both issues. New and improved models help quantify systems more reliably, and enhanced computation techniques decrease cost. However, the increasing complexities of our models will only increase proportionally as research probes deeper understanding of physical systems.

1.3 Amyloid-*β* Aggregation Model

By employing the law of mass action, we can convert the list of chemical equations into a system of differential equations describing the concentration of each $A\beta$ species. With nondimensionalization and characteristic choices of rate k_1^- and concentration A_1, we arrive at the nondimensional system of equations:

$$\frac{dB_1}{ds} = n\alpha_1 B_n - n\alpha_2 B_1^n + B_1^{'} - \alpha_3 B_1, \tag{7}$$

$$\frac{dB_1^{'}}{ds} = n\beta_1 B_n^{'} - n\beta_2 B_1^{'n} + \alpha_3 B_1 - B_1^{'}, \tag{8}$$

$$\frac{dB_n}{ds} = \alpha_2 B_1^n - \alpha_1 B_n + \frac{m}{n}\alpha_5 B_m + \beta_4 B_n^{\prime} - \alpha_4 B_n - \frac{m}{n}\beta_3 B_n^{\frac{m}{n}}, \tag{9}$$

$$\frac{dB_n^{'}}{ds} = \beta_2 B_1^{'n} - \beta_1 {B'}_n + \alpha_4 B_n + \frac{m}{n}\beta_5 B_m^{\prime} - \frac{m}{n}\beta_6 {B'}_n^{\frac{m}{n}} - \beta_4 {B'}_n, \tag{10}$$

$$\frac{dB_m}{ds} = \beta_3 B_n^{\frac{m}{n}} - \alpha_5 B_m, \tag{11}$$

$$\frac{dB_m^{\prime}}{ds} = \beta_6 {B'}_n^{\frac{m}{n}} - \beta_5 {B'}_m. \tag{12}$$

Here, $B_1, B_n,\ and\ B_m$ represent healthy oligomers of sizes $1, n, andm$, and $B_1^{\prime}, B_n^{\prime},\ and\ B_m^{\prime}$ are toxic oligomers. The various nondimensional rate constants are given by α_i or β_i. The equations are solved using MATLAB's built-in *ode45* function. Steady-state values are obtained for each Aβ species to draw conclusions about the aggregation process' dependence upon environmental conditions. Figure 2a depicts the aggregation of monomers over time modeled by the governing equations (7)–(12). The computations of the governing equations are performed repeatedly for various initial conditions—depicted by Fig. 2b—and for differing values of rate constants representing environmental factors.

By comparing the steady-state concentrations of each species, the dominance of toxic or healthy oligomers is deduced. Named and defined in Table 1 are

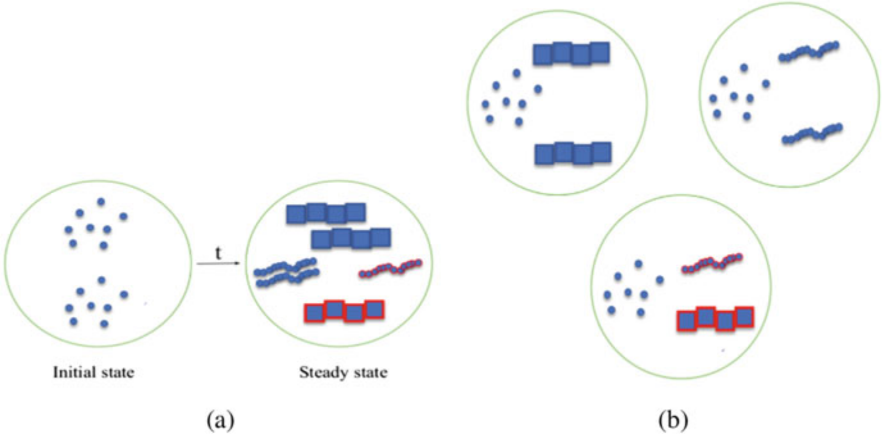


Fig. 2 (**a**) Depiction of the aggregation process and (**b**) an example of the variety of pathological initial conditions to be considered

Table 1 The eight possible steady states when evaluating the competition of healthy and toxic species

State conditions	State name
$B_1 > B_1'$ and $B_n > B_n'$ and $B_m > B_m'$	hhh
$B_1 > B_1'$ and $B_n > B_n'$ and $B_m < B_m'$	hht
$B_1 > B_1'$ and $B_n < B_n'$ and $B_m > B_m'$	hth
$B_1 < B_1'$ and $B_n > B_n'$ and $B_m > B_m'$	thh
$B_1 < B_1'$ and $B_n < B_n'$ and $B_m > B_m'$	tth
$B_1 > B_1'$ and $B_n < B_n'$ and $B_m < B_m'$	htt
$B_1 < B_1'$ and $B_n > B_n'$ and $B_m < B_m'$	tht
$B_1 < B_1'$ and $B_n < B_n'$ and $B_m < B_m'$	ttt

eight possible combinations when comparing healthy vs toxic oligomers at each aggregation level. The outcome state—denoted xxx—describes the toxicity of the chosen input conditions. For example, thh describes toxic monomers dominating in comparison to healthy monomers, while healthy ologimers dominate the nucleation and post-nucleation domains.

The excruciating computational expense comes from the desire to solve for the dominant state in *many* sets of conditions. Monte Carlo or other parameter sampling techniques are important to multiphysics modeling. In the Aβ example, for instance, to properly inform, clinical probes at controlling the runaway amyloid process models should encompass a wide possibility of initial conditions. Additionally, to fully capture the possible range of environmental factors like catalysts or patient predisposition, models need to incorporate various parameter regimes. To produce numerical results reconcilable with in vitro experiments and with breadth wide enough to draw conclusions about seeding conditions, Ghosh et al. [6] solve Eqs. (7) to (12) for 4000 rate parameter combinations to produce a phase space of just one set of initial conditions. With an average iteration duration of ∼three seconds[1] on a single processor, one initial condition takes about 8 days to resolve. The computation here is embarrassingly parallel; however, work is bound by hardware and still consistently burdened with increasing model complexity.

As discussed in Sect. 1, multiphysics models encompass such wide fields as planetary physics and protein aggregation. Large parametric sweeps like Monte Carlo methods are ubiquitous. Large sampling methods are used to guide flight trajectory risks in unmanned aerial systems [11]. Considering the costs with large parametric sweeps on complex nonlinear differential equations, one can imagine the computational complexity of controlling aircraft or of solving x-many n-body problems required to put humans into atmospheric orbit and further.

2 Physics-Informed Machine Learning

Machine learning approaches are popular for their ability to transcend many of the costs presented in Sect. 2. ML approaches to modeling multiphysics systems can explore high-dimensional feature spaces for correlations. Deep ML architectures offer creative ways to extract features from multi-fidelity data. With ML approaches like neural networks gaining popularity, they have been used in all types of research—including flight control [8].

However, some problems, like the Aβ aggregation problem, lack the massive empirical data to train an ML model or are prohibited by the computational expense of reliable simulation data. Physics-informed machine learning harnesses the computational advantages of ML and accelerates training and improves model generalization by integrating the model with systemic information. Karniadakis et al. [7] describe three principles of PIML: observational biases, inductive biases, and learning biases. Each of the methods integrates physically relevant information like symmetry, conservation laws, or system dynamics into the model's data augmentation procedures, architecture, or training procedures, respectively [7]. The Aβ example will employ a standard densely connected neural network. The training data will be iteratively augmented to capture segments of data where feature occurrence is proportional to feature significance. This observationally biased neural network is shown to increase in accuracy as essentially a physics-informed data augmentation is performed.

[1] Computation described in detail in Sect. 3 Table 2.

2.1 Amyloid-β Fibril Aggregation

The neural network architecture used to predict the dominant steady-state aggregation is defined using the *keras* package. The network contains an input and output layer of six and eight dimensions, respectively, with a densely connected intermediate layer of 60 nodes. The dense layer uses the rectified linear activation function. The *Nadam* optimizer is used. The input to the model is the initial concentrations of Aβ protein species. The output is the fraction of the rate parameter phase space dominated by each steady-state outcome.

The goal of this study is to show that a neural network trained on informed data augmentations can provide steady-state simulations with fidelity. This particular neural network is trained on various initial seedings of Aβ protein species, as depicted in Fig. 2b. Initial seedings are defined by the initial concentration of each Aβ species, given by B_1, B_1', B_n, B_n', $and B_m$, B_m'. For example, 1, 0, 0, 0, 0, 0 defines the case where only healthy monomers of Aβ are present at the onset of simulations. The data for the study is obtained by solving Eqs. (7)–(12) with MATLAB's *ode*45. 1000 randomly varied initial seedings are solved for steady-state species concentrations. 70% of the initial data is used for training, while 30% is withheld for testing. After training and testing the initial data, 700 augmented training data points are created, and metrics are compared to the previous iteration until the convergence of accuracy.

The source of the physics-informed intervention is the previous knowledge about the system's dependency on initial conditions. It is observed in Ghosh et al. [6] that the pathological outcome is dependent more heavily on certain seedings than others. In the trained neural network, the same trends are noticed when a sensitivity analysis is performed on the trained model, with respect to well-known seeding conditions.

Concisely, step one is to generate initial training and testing data. Next is to train the neural network on the initial training data. At this stage, we output the test accuracy. Then, by comparing the trained model's prediction to a well-known seeding, a sensitivity analysis is performed. The training set is augmented such that feature occurrence is proportional to feature sensitivity. The model is retrained on the new data, and updated test accuracy is reported.

The simplicity of the neural architecture is matched by the narrow scope of the Aβ seedings. The example training set is populated by randomly perturbed seedings whose steady-state conditions are already well known. Random small changes about the well-understood initial seedings make up our training set. The purpose of this example is to show a simple case of PIML aptly relieving the computational expense of repetitive differential equation solvers while avoiding the need for massive amounts of data with physics-informed automatic training set generation.

The first two iterations of the observational bias method are described above, and results are reported in Sect. 3. The general informed data augmentation process is depicted in Fig. 3.

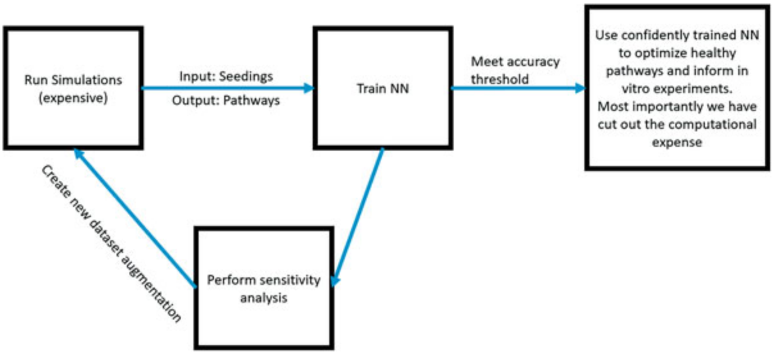

Run Simulations (expensive)
Input: Seedings
Output: Pathways
Train NN
Meet accuracy threshold
Use confidently trained NN to optimize healthy pathways and inform in vitro experiments. Most importantly we have cut out the computational expense
Create new dataset augmentation
Perform sensitivity analysis

Fig. 3 The automatic feedback framework of the observational bias method

3 Case Study: Training Set Bias Method for Modeling Amyloid-β Aggregation

The initial training is performed with a learning rate of 0.01 and a batch size of 60. Final loss converges to 0.2874 with 1000 epochs. The augmented data is trained under identical conditions for a final loss of 0.2462. On a batch size of 25, the test accuracy for iteration one is 97.6%. Parameter importance is determined as the mean absolute error of predicting the most well-studied seeding case, the base case beginning with only monomers over 2000 random small incremental shifts in each input. The occurrence of each feature in the following training set is then proportional to its importance—thus reinforcing the ML process with knowledge of the steady-state behavior of the system. The model is then retrained under identical conditions, with testing accuracy in iteration 2, on the same testing set, is 98.3%. The training processes can be repeated many times to see consistent increase in accuracy.

Table 2 shows just how much time can be saved by implementing the PIML approach, with 98.3% accuracy in just three iterations. It is key to understanding that the runtime of making predictions and training the simple neural architecture is negligible compared to the time it takes to generate training sets. The bottleneck in runtime is still the time spent numerically solving the governing equations. However, the PIML approach allows us to take a small subset of data and extrapolate predictions with a reasonable accuracy rate. Furthermore, the possibility of running more data augmentation iterations is only impeded by the need to recreate datasets by again computing the governing equations.

4 Conclusions

The example of modeling Amyloid-β aggregation has been proposed as a problem where mathematical modeling is key for anatomical reasons and whose breadth of research is prohibited by computational expense. The use of a neural network to

Table 2 Comparison of the first and second iteration of the PIML approach to standalone differential equation-based modeling. The neural model is trained and evaluated in a Google Colaboratory notebook on a single 2.2 GHz Intel CPU with 13 GB RAM, and MATLAB computations are performed on one local 3.6 GHz Intel CPU with 15 GB usable RAM

		Time to resolve	Time to resolve
Method	Testing accuracy	One initial seeding (approx.)	100 Initial seedings (approx.)
PIML 1	97.6%	2.1 days	2.1 days
PIML 2	98.3%	3.5 days	3.5 days
$ode45$	–	8.3 days	830 days

accurately predict dominant steady-state concentrations and the ability for physics-informed data augmentation to improve accuracy of the said model is displayed in Sect. 3. Most importantly, the PIML approach affords significant speedup and opens the pathological probing of the Aβ model to much larger parameter spaces.

Generally, this is one example of expensive multiphysical modeling where a physics-informed machine learning model could reliably produce modeling results with significant advantages in terms of computational expense.

Acknowledgments This work was partially supported by NSF CBET 1802588 (to P.G.) and NSF CBET 1802641 (to A.V.)

References

1. A. Chenciner, A remarkable periodic solution of the three-body problem in the case of equal masses. Ann. Math. 881–901 (2000)
2. D. Dean, Strain-specific Fibril Propagation by an Aβ Dodecamer. Sci. Rep. **7**, 40787 (2017)
3. B. De Bari, A thermodynamic analysis of end-directed particle flocking in chemical systems. Commun. Nonlinear Sci. Numer. Simul. **106**, 106107 (2021)
4. P. Ghosh, Determination of critical nucleation number for a single nucleation amyloid-β aggregation model. Math. Biosci. **273**, 70–79 (2016)
5. P. Ghosh, A game-theoretic approach to deciphering the dynamics of amyloid-β aggregation along competing pathways. R. Soc. Open Sci. **7**(4), 191814 (2020)
6. P. Ghosh, A network thermodynamic analysis of amyloid aggregation along competing pathways. Appl. Math. Comput. **393**, 125778 (2021)
7. G.E. Karniadakis, Physics-informed machine learning. Nat. Rev. **3**(6), 422–440 (2021)
8. B.S. Kim, Nonlinear flight control using neural networks. J. Guidance Control Dyn. **20**(1), 26–33 (1997)
9. W.S. Koon, Dynamical systems, the three-body problem and space mission design, in: 99: In 2 Volumes, ed. by W.S. Koon, M.W. Marsden, J.E., Ross, S.D. Equadiff (World Scientific, Singapore 2000), pp. 881–901
10. H.J. Motulsky, The kinetics of competitive radioligand binding predicted by the law of mass action. Mol. Pharmacol. **25**(1), 1–9 (1985)
11. E. Rudnick-Cohen, Modeling unmanned aerial system (UAS) risks via Monte Carlo simulation, in *ICUAS* (IEEE, Piscataway, 2019), pp 1296–1305
12. K. Sitnikov, The existence of oscillatory motions in the three-body problem. Dokl. Akad. Nauk SSSR **133**(2), 303–306 (1960)
13. E.B. Wilson, The law of mass action in epidemiology. Proc. Nat. Acad. Sci. U.S.A. **31**(1), 24 (1945)

Reduced Order Model Closures: A Brief Tutorial

William Snyder, Changhong Mou, Honghu Liu, Omer San, Raffaella De Vita, and Traian Iliescu

1 Introduction

Reduced order models (ROMs) are computational models whose dimensions are orders of magnitude lower than the dimensions of the full order models (FOMs) (i.e., models obtained from classical numerical methods, e.g., the finite element method). Because ROMs are relatively low-dimensional, their computational cost is orders of magnitude lower than the computational cost of FOMs. Thus, ROMs represent a promising alternative to FOMs in computationally intensive applications, e.g., digital twins of wind farms and real-time surgical procedures. ROMs are expected to play a key role in establishing mathematical modeling foundations for digital twins of many engineering, healthcare, and environmental systems. Indeed, if ROM results are nearly indistinguishable from the corresponding FOM results, then they can contribute as predictive tools in emerging digital twin infrastructures. However, despite being successfully used in simple, academic test problems, ROMs have not made a significant impact in complex, practical applications.

One of the main hurdles in the ROMs' development is their notorious inaccuracy when they are used in the *under-resolved* regime, i.e., when the ROM's dimension (i.e., its number of degrees of freedom (DOF)) is not large enough to capture

W. Snyder · H. Liu · R. De Vita · T. Iliescu (✉)
Virginia Tech, Blacksburg, VA, USA
e-mail: swilli9@vt.edu; hhliu@vt.edu; devita@vt.edu; iliescu@vt.edu

C. Mou
University of Wisconsin-Madison, Madison, WI, USA
e-mail: cmou3@wisc.edu

O. San
Oklahoma State University, Stillwater, OK, USA
e-mail: osan@okstate.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_8

the complex dynamics of the underlying system. To illustrate the under-resolved regime, think of the numerical simulation of the flow around a wind farm. This simulation with a FOM (e.g., the finite element method) generally requires millions (if not billions) of DOF. Thus, performing shape optimization or real-time control of the wind farm flow, which would require many individual FOM runs, is not feasible. Replacing the costly FOM with a ROM would be a natural choice. However, in order to represent the turbulent flow dynamics in the wind farm simulation would require thousands or tens of thousands of DOF in the ROM. Despite the ROM's cost being much lower than the FOM cost, it is still too high to allow the use of the ROM in real-time control applications, where thousands of ROM runs would be required. Thus, a practical choice would be to use much cheaper ROMs, i.e., ROMs with much fewer (e.g., hundreds or even tens) DOF. However, these low-dimensional ROMs, although computationally efficient (and, therefore, practical), generally yield inaccurate results. The reason is simple: these ROMs do not have enough DOF to represent the complex dynamics of a complex flow such as the turbulent wind farm flow.

The above discussion yields the following two important conclusions:

1. The under-resolved ROM regime is critical in realistic, complex applications.
2. Under-resolved ROMs produce inaccurate results.

These conclusions naturally lead to the following question:

Q0

How do we fix the under-resolved ROMs?

The answer to Q0 is simple:

A0

We develop good ROM closure models, i.e., correction terms that increase the standard ROM's accuracy.

To our knowledge, the first (and only) survey of ROM closure models was performed in [1], where the authors discuss dozens of ROM closures for fluids that have been developed over the last four decades. We are not aware, however, of a tutorial on ROM closures. This paper takes a first step at filling that gap.

This brief tutorial on ROM closures (also known as parameterizations [5, 10, 11, 16, 17, 34, 53] and hidden dynamics [40, 41]) is structured as a sequence of simple questions and answers that lead the reader from a simple PDE to projection ROMs, and then to ROM closures. Our paper is aimed at first year graduate students and advanced undergraduate students. Thus, we strive to keep the technical details to a level that is easily understood by students with a standard background in differential equations and numerical methods. We also emphasize that our goal in this tutorial is not to explain the "how," but the "why." That is, we carefully explain the principles used to develop ROM closures, without focusing on particular approaches (which are carefully discussed in [1]).

The rest of the paper is organized as follows: In Sect. 2, we illustrate the ROM closure modeling concept for a three-dimensional toy problem. In Sect. 3, we present the general algorithm used to develop the classical Galerkin ROM. In Sect. 4, we first present the ROM closure problem, and then we discuss its solution, i.e., the ROM closure model. In Sect. 5, we construct the data-driven variational multiscale ROM, in which available data is used to build the ROM closure model. In Sect. 6, we illustrate how closure modeling can significantly increase the ROM accuracy in the numerical simulation of fluid flows. In Sect. 7, we survey current mathematical results for ROM closure modeling. Finally, in Sect. 8, we present conclusions and future research avenues.

2 A Crash Course in ROM Closure: A Toy Problem

Before carefully presenting the ROM closure modeling in the next sections, we illustrate the underlying *concepts* and *principles* for a *toy problem*. These concepts and principles are broadly illustrated in the schematic in Fig. 1, which is adapted from Fig. 1 in [3].

To present our toy problem, we first assume that the FOM solution, $\boldsymbol{u}^{FOM}$, can be accurately approximated by only three ROM basis functions:

$$\boldsymbol{u}^{FOM}(x,t) \approx a_1(t)\boldsymbol{\varphi}_1(\boldsymbol{x}) + a_2(t)\boldsymbol{\varphi}_2(\boldsymbol{x}) + a_3(t)\boldsymbol{\varphi}_3(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \boldsymbol{\varphi}_3$ are the ROM basis functions, and a_1, a_2, a_3 are the sought time-dependent coefficients. Of course, for complex systems, one should use many more (e.g., hundreds and even thousands of) ROM basis functions to accurately approximate $\boldsymbol{u}^{FOM}$. However, to graphically illustrate the need for closure modeling in our toy problem, we assume that three ROM basis functions are enough.

Next, we use the three ROM basis functions in the Galerkin framework to construct the *Galerkin ROM (G-ROM)*. Details regarding the G-ROM construction are given in Sect. 3. For the purpose of the toy problem illustration in this section, we just note that the resulting G-ROM is a three-dimensional dynamical system that can be written as follows:

$$\begin{bmatrix} \dot{a}_1 \\ \dot{a}_2 \\ \dot{a}_3 \end{bmatrix} = \begin{bmatrix} F_1(a_1, a_2, a_3) \\ F_2(a_1, a_2, a_3) \\ F_3(a_1, a_2, a_3) \end{bmatrix}, \tag{2}$$

where F_1, F_2, and F_3 are the components of the ROM operators, e.g., vectors, matrices, and tensors, which are presented in Sect. 3. Since the three ROM basis

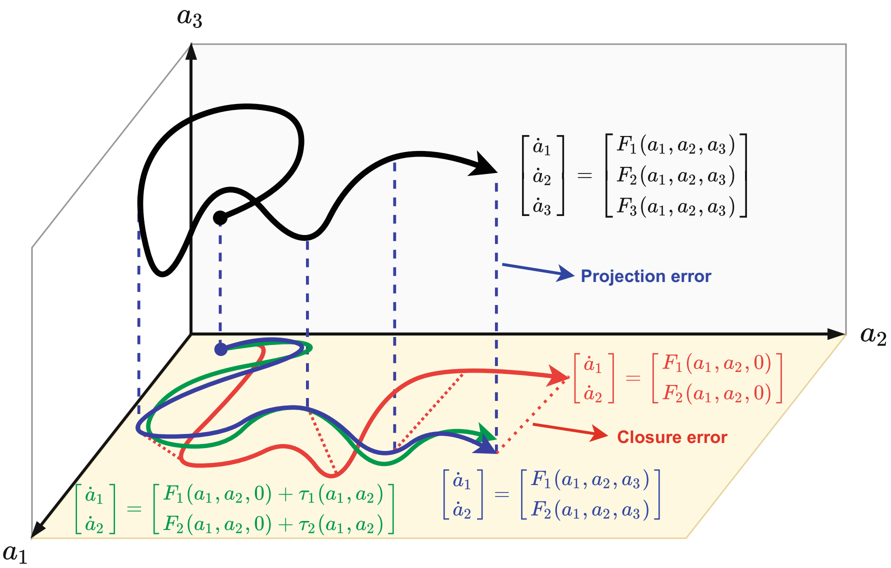

a_3
a_2
a_1
$\begin{bmatrix}\dot a_1\\ \dot a_2\\ \dot a_3\end{bmatrix} = \begin{bmatrix}F_1(a_1,a_2,a_3)\\ F_2(a_1,a_2,a_3)\\ F_3(a_1,a_2,a_3)\end{bmatrix}$
Projection error
$\begin{bmatrix}\dot a_1\\ \dot a_2\end{bmatrix} = \begin{bmatrix}F_1(a_1,a_2,0)\\ F_2(a_1,a_2,0)\end{bmatrix}$
Closure error
$\begin{bmatrix}\dot a_1\\ \dot a_2\end{bmatrix} = \begin{bmatrix}F_1(a_1,a_2,0)+\tau_1(a_1,a_2)\\ F_2(a_1,a_2,0)+\tau_2(a_1,a_2)\end{bmatrix}$
$\begin{bmatrix}\dot a_1\\ \dot a_2\end{bmatrix} = \begin{bmatrix}F_1(a_1,a_2,a_3)\\ F_2(a_1,a_2,a_3)\end{bmatrix}$

Fig. 1 A schematic representation of the ROM closure modeling for a three-dimensional toy problem. The goal is to reduce the three-dimensional G-ROM (2) (black curve and equations) to the most accurate two-dimensional ROM. The I-ROM (3) (blue curve and equations) is the most accurate ROM obtained in the Galerkin framework, but it is not closed (since it depends on a_3). The two-dimensional G-ROM (4) (red curve and equations) is closed, but it is not accurate (since we simply ignore the a_3 contribution). The two-dimensional G-ROM supplemented with a closure model (5) (green curve and equations) is closed and more accurate than the two-dimensional G-ROM (4) since the closure terms $\tau_1(a_1, a_2)$ and $\tau_2(a_1, a_2)$ aim at steering the green curve toward the blue curve

functions yield an accurate approximation of the FOM solution in (1), the three-dimensional G-ROM in (2) is expected to yield an accurate approximation to $\boldsymbol{u}^{FOM}$. That is, solving the three-dimensional G-ROM (2) for a_1, a_2, a_3, and then plugging these values back into (1) yields an accurate approximation to $\boldsymbol{u}^{FOM}$. In Fig. 1, the time evolution of the solution of the accurate three-dimensional G-ROM (2) is represented as the black curve.

At this point, we invoke the need to *reduce the computational cost* of the three-dimensional G-ROM (2). Specifically, we aim at constructing a *two-dimensional* ROM that is as accurate as possible (preferably, as accurate as the three-dimensional G-ROM (2)). For our toy problem (1), this amounts to constructing a dynamical system for a_1 and a_2 (assuming that the first two ROM basis functions dominate the third, as is often the case; see Sect. 3).

Of course, reducing the ROM dimension from three to two does not yield such a great reduction of computational time. We emphasize, however, that we consider this reduction only to illustrate the ROM closure modeling concept for our toy problem. In practical settings, ROMs reduce the FOM dimension by orders of magnitude.

The most natural way to construct an accurate two-dimensional ROM is to keep only the first two equations in (2) and discard the third equation:

$$\begin{bmatrix} \dot{a}_1 \\ \dot{a}_2 \end{bmatrix} = \begin{bmatrix} F_1(a_1, a_2, a_3) \\ F_2(a_1, a_2, a_3) \end{bmatrix}. \tag{3}$$

Mathematically, this amounts to first using a Galerkin expansion for all three ROM basis functions (i.e., using (1)), and then using a Galerkin projection onto only the first two basis functions (instead of projecting onto all three basis functions, as done in (2)).

In Fig. 1, the time evolution of the solution of the efficient, two-dimensional ROM (3) is represented as the blue curve. Of course, since we perform a Galerkin projection only onto the first two basis functions, we incur an error, which we denote as the (Galerkin) *projection error* (the blue dashed lines in Fig. 1). Nevertheless, it stands to reason that, in the Galerkin framework with the basis $\{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \boldsymbol{\varphi}_3\}$, the two-dimensional ROM (3) is the most accurate two-dimensional ROM we can hope to get. This is why we call the two-dimensional ROM (3) the *ideal ROM (I-ROM)*. However, the two-dimensional I-ROM (3) has a big problem: It is *not closed* since the equations for a_1 and a_2 depend on a_3. *This is the ROM closure problem.*

So how do we solve the ROM closure problem? *The easiest way to solve the ROM closure problem is to simply ignore it.* That is, we can simply ignore the a_3 contribution to the dynamics in (3):

$$\begin{bmatrix} \dot{a}_1 \\ \dot{a}_2 \end{bmatrix} = \begin{bmatrix} F_1(a_1, a_2, 0) \\ F_2(a_1, a_2, 0) \end{bmatrix}. \tag{4}$$

The ROM in (4) is two-dimensional and closed (since the equations depend only on a_1 and a_2). In Fig. 1, the time evolution of the solution of this two-dimensional ROM (4) is represented as the red curve. Of course, since in (4) we simply ignored the a_3 contribution to the correct dynamics of a_1, a_2 given by (3), we incur an error, which is generally called the *closure error* (the red dashed lines in Fig. 1).

Remark 1 (Galerkin Closure is a Relative Concept) We note that if we start with just two ROM basis functions $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$, the Galerkin ROM framework (which is presented in Sect. 3 and outlined in Algorithm 1) yields a two-dimensional G-ROM that satisfies exactly the equations in (4). Thus, *the ROM closure concept is relative to the ROM space used in the Galerkin framework*:

- If we start with two basis functions, the Galerkin method yields the two-dimensional G-ROM (4), which is closed.
- If, however, we start with the larger (three-dimensional) ROM space spanned by $\boldsymbol{\varphi}_1$, $\boldsymbol{\varphi}_2$, and $\boldsymbol{\varphi}_3$, the discussion in this section shows that the most accurate two-dimensional ROM obtained by a direct truncation of the three-dimensional G-ROM (2) (i.e., the I-ROM (3)) is not closed.

Remark 2 (Galerkin Closure is a General Concept) We emphasize that, although our discussion focuses exclusively on ROMs, *the Galerkin closure is a general concept that is associated with the classical Galerkin framework*. Thus, there is no surprise that, over half a century, closure has been addressed in different contexts: large eddy simulation (LES) [6], variational multiscale (VMS) methods [24], subgrid-scale (SGS) methods [21, 33], and nonlinear Galerkin (NG) methods [20].

At this point, it is probably a good idea to summarize our discussion. As illustrated in the schematic in Fig. 1, the reader interested in constructing the most accurate two-dimensional G-ROM has reached a *crossroads*:

- On the one hand, the I-ROM (3) is the most accurate two-dimensional ROM that we can get by using the Galerkin framework, but it is not closed.
- On the other hand, the G-ROM (4) is closed, but we are incurring the closure error.

This is as far as the classical Galerkin framework can take us. We're stuck. So what do we do next?

The answer, as many times in numerical methods, is to take a middle of the road approach. Specifically, we construct a *ROM closure model* and add it to the G-ROM (4):

$$\begin{bmatrix} \dot{a}_1 \\ \dot{a}_2 \end{bmatrix} = \begin{bmatrix} F_1(a_1, a_2, 0) + \tau_1(a_1, a_2) \\ F_2(a_1, a_2, 0) + \tau_2(a_1, a_2) \end{bmatrix}, \tag{5}$$

where $\tau_1(a_1, a_2)$, $\tau_2(a_1, a_2)$ are the components of the ROM closure model, i.e., correction terms that aim at steering the inaccurate G-ROM (4) as close as possible to the accurate (but not closed) I-ROM (3). In Fig. 1, the time evolution of the solution of the closed ROM (5) is represented as the green curve.

How do we construct the ROM closure model in (5)? We answer this question in Sect. 5. But first, in Sect. 3, we present the main steps in the G-ROM construction.

3 Galerkin ROM (G-ROM)

Over the past four decades, projection ROMs have been used in the numerical simulation of fluid flows [8, 22, 23, 39, 42, 49]. In this tutorial, we exclusively consider projection ROMs that use numerical or experimental data to find the "best" basis, which is then used together with the Galerkin method to construct the ROM. In this section, we present the main steps in the construction of the Galerkin ROM.

To illustrate the Galerkin ROM construction, we start with a generic PDE for the dynamics of a variable of interest, $\boldsymbol{u}$:

$$\boldsymbol{u}_t = \boldsymbol{f}(\boldsymbol{u}), \tag{6}$$

Algorithm 1 Galerkin ROM (G-ROM) algorithm

```
1: Use numerical or experimental data to construct modes $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_R\}$, which represent the
   recurrent spatial structures in the system (6).
2: Choose the dominant modes $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$, $r \leq R$, as ROM basis functions.
3: Use a Galerkin expansion $\boldsymbol{u}_r(\boldsymbol{x}, t) = \sum_{j=1}^{r} a_j(t)\, \boldsymbol{\varphi}_j(\boldsymbol{x})$.
4: Replace $\boldsymbol{u}$ with $\boldsymbol{u}_r$ in (6), and then on both sides of (6) take the inner product with each mode
   $\boldsymbol{\varphi}_i$, $i = 1, \ldots, r$. That is, perform a Galerkin projection of the PDE (6) onto the ROM space
   $\mathbf{X}^r := \text{span}\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$. The obtained *Galerkin ROM (G-ROM)* is of the form

       $$\overset{\bullet}{\boldsymbol{a}} = \boldsymbol{F}(\boldsymbol{a}), \qquad (7)$$

   where $\boldsymbol{a}(t) = (a_i(t))_{i=1,\ldots,r}$ is the vector of coefficients in the Galerkin expansion in step 3
   and $\boldsymbol{F}$ comprises the ROM operators.
5: In the offline stage, compute the ROM operators (e.g., vectors, matrices, and tensors), which
   are preassembled from the ROM basis.
6: In the online stage, repeatedly use the G-ROM (7) for longer time intervals.
```

equipped with appropriate boundary conditions and initial conditions. In Algorithm 1, we list the main steps in the Galerkin ROM construction.

Remark 3 (ROM=d2G) The main steps in the G-ROM (7) construction presented in Algorithm 1 are straightforward. In principle, they are the same steps as those used to construct classical Galerkin methods, e.g., the finite element method (FEM). The fundamental difference between the G-ROM and the FEM is that the former uses a *data-driven basis*, whereas the latter uses a universal basis (i.e., piecewise polynomials). Thus, one could think of the projection ROMs that we discuss in this tutorial as *data-driven Galerkin (d2G) methods*.

Next, we explain some of the steps in Algorithm 1.

ROM Basis (Step 1)

To construct the ROM basis, we first collect *snapshots* from the simulation of the FOM. If we are interested in time prediction (as in the numerical illustration in Sect. 6), the snapshots can be FEM approximations of (6) at the time instances $t_1, \ldots, t_M$, i.e., $\boldsymbol{u}_h^1, \ldots, \boldsymbol{u}_h^M$, respectively. (If (6) depends on parameters, we can also build a ROM basis for parameter prediction [22, 42].) Next, we use these snapshots to construct the modes $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_R\}$, which represent the recurrent spatial structures in the system described by (6). Different approaches can be used to construct the ROM basis functions, e.g., (i) the proper orthogonal decomposition (POD) [8, 23, 31, 49, 51]; (ii) the reduced basis method (RBM) [22, 42]; (iii) the proper generalized decomposition (PGD) [15]; and (iv) clustering [9]. In this tutorial, to fix ideas, we exclusively use the POD to generate the ROM basis.

For a careful presentation of the POD basis, the reader is referred to, e.g., [23] (for a physical presentation) and to [51] (for a mathematical presentation). In this paper, however, we only briefly discuss the *qualitative* properties of the POD basis functions, which we will later use in our numerical illustration in Sect. 6. The reason for our brief qualitative discussion of the POD basis is that *ROM closure modeling* *does not depend on the particular type of ROM basis functions used.* That is, our presentation of ROM closure modeling remains the same for any type of ROM basis used in a Galerkin framework, whether it is POD, RBM, or PGD.

The main principle used to construct the G-ROM basis can be stated as follows: *Use the available snapshots to find the ROM basis that "best" represents the system's dynamics.* Since this is the "best" basis, for certain problems, one can hope to use much fewer basis functions to construct the G-ROM than to construct, e.g., FEM models. For example, instead of using millions or even billions of basis functions as in FEM simulations, one can hope to use tens or hundreds basis functions in the G-ROM construction. This choice of "best" basis yields computational models (i.e., ROMs) whose dimension can be orders of magnitude lower than the dimension of FEM models. (This also explains the term "reduced" in the ROM terminology.)

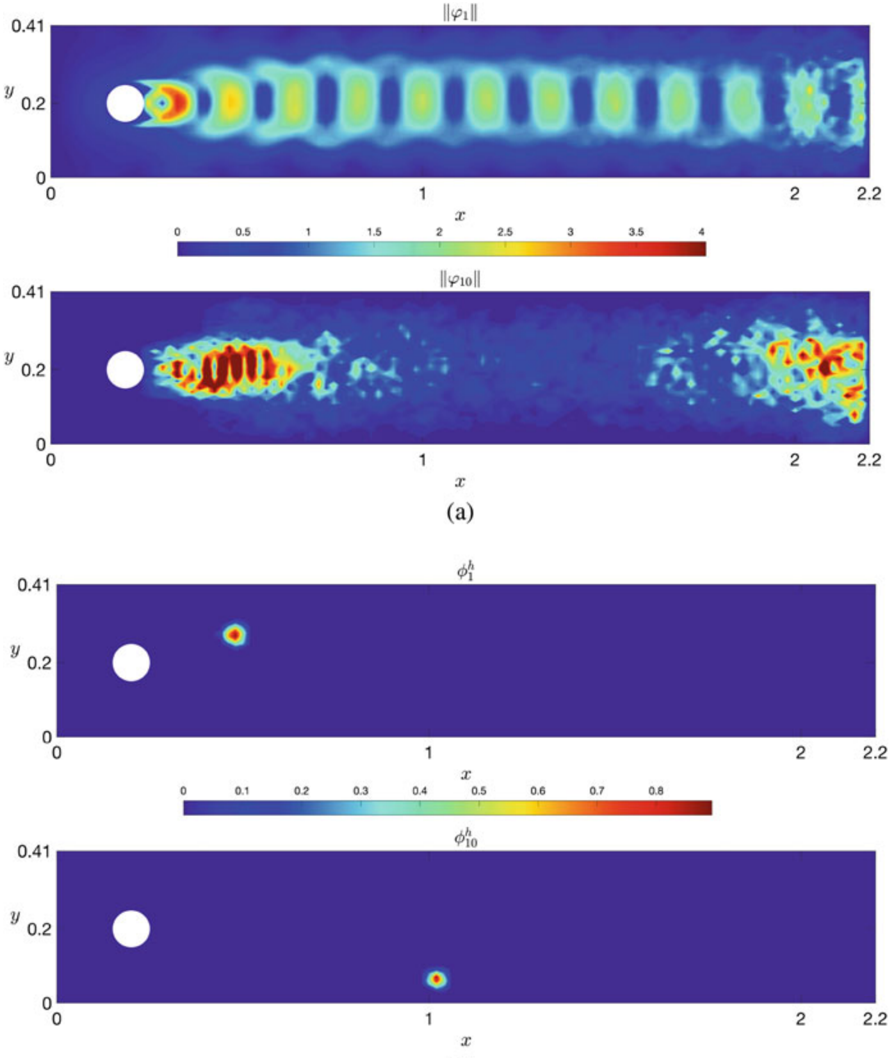
Of course, a natural question is what the "best" ROM basis means. In fact, there are many proposals for the "best" ROM basis, and each proposal yields a different class of ROMs (e.g., POD, RBM, or PGD, to name just a few). For example, given a set of snapshots, the POD basis is the orthonormal basis that yields the minimum projection error with respect to a chosen norm (e.g., the L^2 norm) [51].

However, independent of the approach used to construct them, the ROM basis functions generally share several qualitative features. To illustrate this, in Fig. 2 we plot the Euclidian norm of two POD basis functions, $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_{10}$, and two FEM basis functions, $\boldsymbol{\phi}_1^h$ and $\boldsymbol{\phi}_{10}^h$, for a 2D flow past a circular cylinder [37]. One can clearly see the significant differences between the POD basis functions (top two plots) and the FEM basis functions (bottom two plots). Indeed, the POD basis functions have global support (i.e., they can be nonzero over the entire computational domain), whereas the FEM basis functions have local support (i.e., they are one at one mesh point and zero everywhere else). To further illustrate the different characteristics of the POD basis, in Fig. 3 we plot the Euclidian norm of two POD basis functions, $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_{10}$, for soft tissue modeling [48]. Comparing these two POD basis functions with the POD basis functions in the top two plots of Fig. 2, we can clearly see that different physical systems (i.e., the soft tissue in Fig. 3 and the flow in Fig. 2) yield fundamentally different POD basis functions. We emphasize that this is in complete contrast with classical numerical methods, such as the FEM. Indeed, the FEM basis functions are *universal* basis functions, i.e., they have the same shape (piecewise polynomials and local support) for all the problems. In contrast, the POD basis functions (and ROM basis functions in general) change their shape when we change the problem. This can be clearly seen by comparing the top two plots of Fig. 2 with the plots of Fig. 3.

Galerkin ROM Construction (Steps 2–6)

To illustrate the G-ROM construction, we use the Navier-Stokes equations (NSE) as a mathematical model:

$$\frac{\partial \boldsymbol{u}}{\partial t} - Re^{-1} \Delta \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla p = \boldsymbol{0}\,, \tag{8}$$



(a)

(b)

Fig. 2 2D flow past a circular cylinder: (**a**) Euclidian norm of ROM basis functions $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_{10}$ at mesh points. (**b**) Euclidian norm of FEM basis functions ϕ_1^h and ϕ_{10}^h at mesh points. Note that the ROM basis functions are fundamentally different from the FEM basis functions: The former have global support, whereas the latter have local support

$$\nabla \cdot \boldsymbol{u} = 0\,, \tag{9}$$

where $\boldsymbol{u}$ is the velocity, p the pressure, and Re the Reynolds number. We consider the NSE posed on a bounded spatial domain in either $\mathbb{R}^2$ or $\mathbb{R}^3$, and supplemented with homogeneous Dirichlet boundary conditions and an appropriate

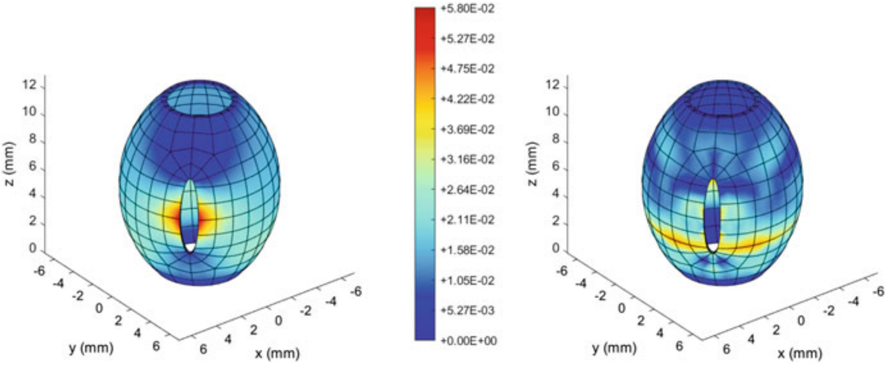


Fig. 3 Soft tissue modeling: Euclidian norm of ROM basis functions $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_{10}$ at mesh points

initial condition. The NSE (8)–(9) can be cast in the general form (6) by choosing $\boldsymbol{f}(\boldsymbol{u}) = Re^{-1}\Delta\boldsymbol{u} - \boldsymbol{u}\cdot\nabla\boldsymbol{u}$ (after applying the Leray projection, which maps the vector field into the divergence-free subspace of the underlying state space) [50].

To construct the G-ROM for the NSE, we follow Steps 2–6 in Algorithm 1. That is, we choose the first r basis functions from the modes constructed in Step 1, use a Galerkin truncation

$$\boldsymbol{u}_r(\boldsymbol{x},t) = \sum_{j=1}^{r} a_j(t)\,\boldsymbol{\varphi}_j(\boldsymbol{x}), \tag{10}$$

replace $\boldsymbol{u}$ with $\boldsymbol{u}_r$ in the NSE (8), and project the resulting PDE onto the ROM space, $\boldsymbol{X}^r$. Furthermore, we apply the divergence theorem to the diffusion term and the pressure term. This yields the G-ROM [37]:

$$\overset{\bullet}{\boldsymbol{a}} = A\,\boldsymbol{a} + \boldsymbol{a}^{\top}\,B\,\boldsymbol{a}, \tag{11}$$

where $\boldsymbol{a}(t)$ is the vector of unknown coefficients $a_j(t)$, $1 \le j \le r$ in the Galerkin expansion (10). The ROM operator A in (11) is an $r \times r$ matrix that corresponds to the diffusion term in the NSE (i.e., $-Re^{-1}\Delta\boldsymbol{u}$) and has entries

$$A_{im} = -Re^{-1}\left(\nabla\boldsymbol{\varphi}_m, \nabla\boldsymbol{\varphi}_i\right), \quad 1 \le i, m \le r, \tag{12}$$

where $(\cdot,\cdot)$ denotes the L^2 inner product. The ROM operator B in (11) is an $r\times r\times r$ tensor that corresponds to the nonlinear term in the NSE (i.e., $\boldsymbol{u}\cdot\nabla\boldsymbol{u}$) and has entries

$$B_{imn} = -\left(\boldsymbol{\varphi}_m \cdot \nabla\boldsymbol{\varphi}_n, \boldsymbol{\varphi}_i\right), \quad 1 \le i, m, n \le r. \tag{13}$$

We note that the pressure term in the G-ROM (11) vanishes since we assumed that the ROM modes are discretely divergence-free (which is the case if, e.g., the snapshots are discretely divergence-free). ROMs that provide a pressure approximation are discussed in, e.g., [18, 22].

Once the matrix A and tensor B are assembled in the offline stage, the G-ROM (11) is a relatively low-dimensional, efficient dynamical system that can be used in the online stage for longer time intervals (or more parameter values, e.g., Re [22, 42]).

4 The Closure Problem and Its Solution: The Closure Model

This section has two goals: In Sect. 4.1, we motivate the need for ROM closure modeling in the under-resolved regime, i.e., we describe the ROM closure problem. In Sect. 4.2, we show how to solve the ROM closure problem, i.e., we show how to construct a ROM closure model. To this end, we give the definition of the ROM closure model, show that using the exact closure model (i.e., using the ideal ROM) increases the ROM accuracy, and finally outline the main steps in the ROM closure model construction.

4.1 The Closure Problem

The G-ROM (11) constructed in Sect. 3 is appealing from the computational point of view: The G-ROM can significantly reduce the dimension (and, thus, the computational cost) of classical numerical discretization (e.g., FEM) models by orders of magnitude. So one can ask the following natural question:

? Q1

What is wrong with G-ROM?

The short answer to Q1 is: It depends on the resolution. Specifically:

> A1

It depends on whether we are in the *resolved* regime or the *under-resolved* regime.

- In the *resolved* regime (i.e., when there are enough ROM basis functions $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$ to accurately represent the underlying dynamics), the G-ROM produces accurate results.

- In the *under-resolved* regime (i.e., when there are not enough ROM basis functions $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$ to accurately represent the underlying dynamics), the G-ROM produces inaccurate results.

  But then one can ask the following questions:

? **Q2**

Why is the under-resolved regime important? Why do we need to worry about it?

> **A2**

Many important applications (e.g., atmospheric boundary layer flows, digital twins of wind farms, and anisotropic and heterogeneous biological tissues) are centered around *multiscale* systems that require a large number of ROM basis functions. However, to ensure a low computational cost in these applications, under-resolved G-ROMs are generally used.

4.2 The Closure Model

In Sect. 4.1, we defined the ROM closure problem, and we explained why it is important. In this section, we present the solution to the ROM closure problem. That is, we answer the following question:

? **Q3**

What is the solution to the closure problem?

> **A3**

The solution to the closure problem is the closure model. That is, replace the G-ROM (11) with

$$\overset{\bullet}{\boldsymbol{a}} = \boldsymbol{F}(\boldsymbol{a}) + \boldsymbol{\tau}(\boldsymbol{a}), \tag{14}$$

where $\boldsymbol{\tau}(\boldsymbol{a})$ is the closure model, which represents the effect of the discarded ROM modes $\{\boldsymbol{\varphi}_{r+1}, \ldots, \boldsymbol{\varphi}_R\}$ on the ROM dynamics.

Note that A3 is a vague definition, which begs the following questions: What exactly does “model the effect” mean? What exactly does $\boldsymbol{\tau}(\boldsymbol{a})$ in (14) actually model?

Answering these natural questions is not straightforward. To do so, we need to *extend the Galerkin framework*. This sounds like a daunting task, but it turns out to be relatively simple. The “trick” is to *rethink the space* we use in the Galerkin framework:

In the resolved regime, the ROM space $\boldsymbol{X}^r := \text{span}\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$ is the only space we will ever need, since everything happens in $\boldsymbol{X}^r$. Thus, in the resolved regime, G-ROM should (and generally does) work just fine.

However, in the under-resolved regime we need *two* spaces: (i) the *resolved space* $\boldsymbol{X}^r$, and (ii) the *unresolved space* $\boldsymbol{X}^{r'} := \text{span}\{\boldsymbol{\varphi}_{r+1}, \ldots, \boldsymbol{\varphi}_R\}$. To keep the ROM dimension (and, therefore, its computational cost) low, we want to work in the resolved space, $\boldsymbol{X}^r$. However, to increase the ROM accuracy, we should do our best to model the contribution to the ROM dynamics made by the dynamics in the unresolved space, $\boldsymbol{X}^{r'}$. But this sounds like a lot of work (both in terms of modeling and computation). So the following is a natural question:

? Q4

Does $\boldsymbol{X}^{r'}$ have a significant effect on the ROM dynamics?

> A4

Yes.

The answer A4 is simple. In Sect. 4.2.1, we introduce the ideal ROM, which adds the *exact* closure term to the classical G-ROM. The ideal ROM results clearly show why the effect of $\boldsymbol{X}^{r'}$ should be modeled. Specifically, we show that the ideal ROM results are dramatically more accurate than the G-ROM results. Thus, we conclude that modeling the exact ROM closure term is beneficial to ROM accuracy.

4.2.1 The Ideal ROM (I-ROM)

To present the ideal ROM, we first need to define the spaces of resolved ROM scales (i.e., $\boldsymbol{X}^r$) and unresolved ROM scales (i.e., $\boldsymbol{X}^{r'}$). To this end, we extend the variational multiscale (VMS) framework proposed by Hughes and his group two decades ago in the FEM context. We note, however, that there are other ways of defining the spaces of resolved and unresolved ROM scales, e.g., spatial filtering [37].

First, we leverage the orthonormality of the ROM basis functions and construct the two orthogonal spaces, $\boldsymbol{X}^r$ and $\boldsymbol{X}^{r'}$, as follows:

$$\boldsymbol{X}^r := \text{span}\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\} \qquad \text{and} \qquad \boldsymbol{X}^{r'} := \text{span}\{\boldsymbol{\varphi}_{r+1}, \ldots, \boldsymbol{\varphi}_R\}. \tag{15}$$

The space $\boldsymbol{X}^r$ represents the space of the *resolved ROM scales*, i.e., the spatial scales that are explicitly approximated by a given r-dimensional ROM. In contrast, the space $\boldsymbol{X}^{r'}$ represents the space of the *unresolved ROM scales*, i.e., the spatial scales that are not explicitly approximated by the chosen ROM. We note that since the ROM basis functions are generally ordered from the most important to the least important (with respect to a physical criterion, e.g., kinetic energy [23]), the decomposition in (15) is natural. We also note that since we are concerned with the under-resolved regime that often occurs in practical applications, we consider the case when $r \ll R$.

The next step in the construction of the ideal ROM is to extend the Galerkin framework to the space $\boldsymbol{X}^R := \boldsymbol{X}^r \oplus \boldsymbol{X}^{r'}$, which is the *maximal ROM space* (i.e., the space spanned by all the snapshots). Thus, we use the ROM approximation of both resolved and unresolved scales, i.e., we utilize $\boldsymbol{u}_R \in \boldsymbol{X}^R$ defined as

$$\boldsymbol{u}_R = \sum_{j=1}^{R} a_j \, \boldsymbol{\varphi}_j = \sum_{j=1}^{r} a_j \, \boldsymbol{\varphi}_j + \sum_{j=r+1}^{R} a_j \, \boldsymbol{\varphi}_j = \boldsymbol{u}_r + \boldsymbol{u}' , \tag{16}$$

where $\boldsymbol{u}_r \in \boldsymbol{X}^r$ represents the *resolved ROM component* of $\boldsymbol{u}$, and $\boldsymbol{u}' \in \boldsymbol{X}^{r'}$ represents the *unresolved ROM component* of $\boldsymbol{u}$. Next, we plug $\boldsymbol{u}_R$ in the generic equation (6), project the resulting equation onto $\boldsymbol{X}^r$, and use the ROM basis orthogonality to show that $\left(\boldsymbol{u}_{R,t} \, , \boldsymbol{\varphi}_i\right) = \left(\boldsymbol{u}_{r,t} \, , \boldsymbol{\varphi}_i\right)$, $\forall \, i = 1, \ldots, r$, where $\boldsymbol{u}_{R,t}$ and $\boldsymbol{u}_{r,t}$ are the time derivatives of $\boldsymbol{u}_R$ and $\boldsymbol{u}_r$, respectively. Following these steps, we obtain the *ideal ROM (I-ROM)*:

$$\left(\boldsymbol{u}_{r,t} \, , \boldsymbol{\varphi}_i\right) = \left(\boldsymbol{f}(\boldsymbol{u}_r) \, , \boldsymbol{\varphi}_i\right) + \underbrace{\left(\boldsymbol{f}(\boldsymbol{u}_R) \, , \boldsymbol{\varphi}_i\right) - \left(\boldsymbol{f}(\boldsymbol{u}_r) \, , \boldsymbol{\varphi}_i\right)}_{\boldsymbol{\tau}^{I-ROM} = \text{ ideal ROM closure term}} , \quad \forall \, i = 1, \ldots, r. \tag{17}$$

The last two terms in (17) yield the *ideal ROM closure term*, $\boldsymbol{\tau}^{I-ROM}$, which represents the effect of the discarded ROM modes $\{\boldsymbol{\varphi}_{r+1}, \ldots, \boldsymbol{\varphi}_R\}$ onto the dynamics of the resolved ROM scales, $\boldsymbol{u}_r$. Using the expansion (16), the I-ROM (17) can be written as the following dynamical system for the vector of ROM coefficients of the resolved scales:

$$\overset{\bullet}{\boldsymbol{a}} = \boldsymbol{F}(\boldsymbol{a}) + \boldsymbol{\tau}^{I-ROM}(a_1, \ldots, a_r, a_{r+1}, \ldots, a_R). \tag{18}$$

The above discussion clearly shows that, from a mathematical point of view, the correct equations satisfied by the coefficients of the resolved ROM scales are the I-ROM equations (18) instead of the G-ROM equations (11). However, we need to

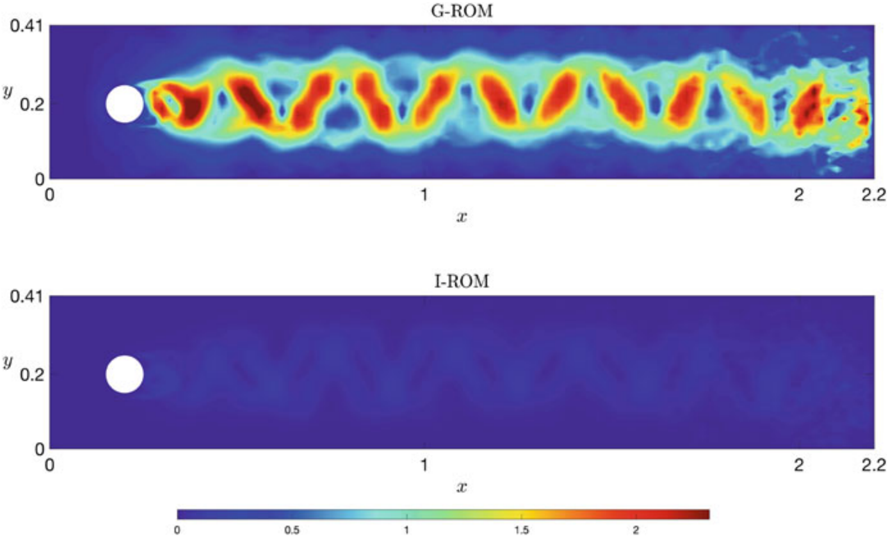


Fig. 4 2D flow past a circular cylinder. The Euclidian norm of the error, $\boldsymbol{u}^{FOM} - \boldsymbol{u}^{ROM}$, at mesh points for G-ROM (11) (top) and I-ROM (17) (bottom). The I-ROM error is significantly lower than the G-ROM error, which illustrates the potential benefit of ROM closure modeling

ask ourselves whether this mathematical framework has a *practical impact* (i.e., we need to ask question Q4). Specifically, we need to check whether the I-ROM results are better than the G-ROM results.

In Fig. 4, we present results for the I-ROM (18) and the G-ROM (11) in the numerical simulation of a two-dimensional flow past a circular cylinder. These plots clearly show that the I-ROM performs significantly better than the classical G-ROM. Thus, these results suggest that including a model for the I-ROM closure term, $\boldsymbol{\tau}^{I-ROM}$, could increase the ROM accuracy.

Remark 4 (The Closure Model Increases Accuracy) There is a lot of confusion in the ROM community (and not only) regarding the role of the closure model. In this section, we tried to emphasize that the main role of the ROM closure model is to increase the *accuracy* of the G-ROM. Indeed, in Eq. (14), adding the closure term, $\boldsymbol{\tau}(\boldsymbol{a})$, to the classical G-ROM yields a more accurate model (in the extended Galerkin framework).

That being said, in many important practical applications (e.g., convection-dominated flows), the G-ROM's inaccuracy often manifests itself in the form of *spurious numerical oscillations*. Thus, a popular misconception (at least in computational fluid dynamics) is that the only role of the ROM closure model is to eliminate/alleviate these numerical oscillations, i.e., to increase the numerical stability of the G-ROM.

However, we emphasize that, while numerical stability of the model is *necessary* (indeed, if the model is accurate, then it has to be stable), it is *not sufficient*. For example, we can add a very large stabilization term to the classical G-ROM. This, most likely, will stabilize the model, but will also degrade its accuracy.

To summarize, we emphasize that ROM closure modeling is not simply about adding numerical stabilization. Instead, ROM closure modeling is about adding the *"right"* amount of numerical stabilization (i.e., the amount of stabilization that makes the model accurate).

4.2.2 Closure Model Construction

The I-ROM results in Sect. 4.2.1 clearly show that the effect of $\boldsymbol{X}^{r'}$ should be modeled. We emphasize, however, that the I-ROM itself does *not* represent a practical solution since it depends on the coefficients of the discarded ROM modes, $a_{r+1}, \ldots, a_R$, which we do not model in our ROM (since we work in $\boldsymbol{X}^r$).

? **Q5**

How do we make the I-ROM (18) practical?

> **A5**

We construct a closure model, $\boldsymbol{\tau}$, which is an approximation in $\boldsymbol{X}^r$ of the I-ROM closure term, $\boldsymbol{\tau}^{I-ROM}$:

$$\boldsymbol{\tau}^{I-ROM}(a_1, \ldots, a_r, a_{r+1}, \ldots, a_R) \approx \boldsymbol{\tau}(a_1, \ldots, a_r). \tag{19}$$

Since $\boldsymbol{\tau}$ in (19) lives in $\boldsymbol{X}^r$, it can be computed with the available ROM data, and, thus, can be used in practical computations.

Remark 5 (Closure=Correction) Equation (14) shows that the closure model, $\boldsymbol{\tau}$, in (19) can be interpreted as a correction term that is added to the G-ROM (11) to correct its dynamics in $\boldsymbol{X}^R$. So do we really need I-ROM in order to construct the closure model? In Sect. 5, we will show that the I-ROM is needed when we construct data-driven ROM closures. Furthermore, we note that the I-ROM derivation explains the closure model terminology. Indeed, $\boldsymbol{\tau}^{I-ROM}(a_1, \ldots, a_r, a_{r+1}, \ldots, a_R)$ shows that the I-ROM (17) is closed in $\boldsymbol{X}^R$, but not in $\boldsymbol{X}^r$.

ROM closure models are of three types: (i) *Functional*, which use physical insight to construct the closure model; (ii) Structural, which use mathematical tools; and (iii) *Data-driven*, which use available data. The three types of ROM closure models are surveyed in [1]. In this tutorial, we take a different approach and, for clarity of presentation, focus on data-driven approaches, which have experienced a tremendous development over the last few years. Specifically, in the next section, we present the data-driven variational multiscale ROM closure model.

5 The Data-Driven Variational Multiscale ROM (D2-VMS-ROM)

In this section, we illustrate how data-driven modeling can be leveraged to construct the ROM closure model. Specifically, we outline the main steps in the construction of one data-driven ROM closure model, i.e., the data-driven variational multiscale ROM (D2-VMS-ROM) that was proposed in [37] (see also [52]). To this end, we follow the presentation in Section 2.3 in [37] to construct the two-scale D2-VMS-ROM. (We note that a three-scale D2-VMS-ROM was also proposed and tested in [37].)

To build the D2-VMS-ROM, we start with the I-ROM (18). As explained in answer A5, to construct the ROM closure model we need to find an approximation $\boldsymbol{\tau}(a_1, \ldots, a_r)$ for the I-ROM closure term in (18), $\boldsymbol{\tau}^{I-ROM}(a_1, \ldots, a_r, a_{r+1}, \ldots, a_R)$. The construction of the data-driven ROM closure model consists of two steps: (i) postulating a model form ansatz; and (ii) solving a least squares problem to determine the coefficients of the model form. Next, we outline these two steps.

5.1 Model Form Ansatz

The first step in the construction of the data-driven ROM closure model is to *postulate a model form (ansatz)*. Specifically, we approximate the I-ROM closure term $\boldsymbol{\tau}^{I-ROM}$ with $\boldsymbol{g}(\boldsymbol{u}_r)$, where $\boldsymbol{g}$ is a *generic* function whose coefficients/parameters still need to be determined:

$$\boxed{\tau_i^{I-ROM} \overset{(17)}{=} \left(\boldsymbol{f}(\boldsymbol{u}_R), \boldsymbol{\varphi}_i\right) - \left(\boldsymbol{f}(\boldsymbol{u}_r), \boldsymbol{\varphi}_i\right) \approx \left(\boldsymbol{g}(\boldsymbol{u}_r), \boldsymbol{\varphi}_i\right), \quad i = 1, \ldots, r.} \tag{20}$$

5.2 Least Squares Problem

To determine the coefficients/parameters in $\boldsymbol{g}$ used in (20), in the offline stage, we solve the following low-dimensional *least squares problem*:

$$\boxed{\begin{aligned}\min_{\boldsymbol{g} \text{ parameters}} \sum_{j=1}^{M} \Big\| \Big[\big(\boldsymbol{f}(\boldsymbol{u}_R^{FOM}(t_j))\,,\boldsymbol{\varphi}_i \big) - \big(\boldsymbol{f}(\boldsymbol{u}_r^{FOM}(t_j))\,,\boldsymbol{\varphi}_i \big) \Big] \\ - \big(\boldsymbol{g}(\boldsymbol{u}_r^{FOM}(t_j))\,,\boldsymbol{\varphi}_i \big) \Big\|^2 ,\end{aligned}} \tag{21}$$

where $\boldsymbol{u}_R^{FOM}$ and $\boldsymbol{u}_r^{FOM}$ are obtained from the FOM data, and M is the number of snapshots. Once $\boldsymbol{g}$ is determined, the I-ROM (17) with the I-ROM closure term replaced by $\boldsymbol{g}$ yields the *data-driven VMS-ROM (D2-VMS-ROM)*:

$$\boxed{\big(\boldsymbol{u}_{r,t}\,,\boldsymbol{\varphi}_i\big) = \big(\boldsymbol{f}(\boldsymbol{u}_r)\,,\boldsymbol{\varphi}_i\big) + \big(\boldsymbol{g}(\boldsymbol{u}_r)\,,\boldsymbol{\varphi}_i\big),} \qquad i = 1, \ldots, r. \tag{22}$$

We emphasize that we have a lot of *flexibility* in choosing the model form ansatz (20) in the D2-VMS-ROM. For example, for the NSE, we can choose the following model form: $\forall\, i = 1, \ldots, r$,

$$\big(\boldsymbol{g}(\boldsymbol{u}_r)\,,\boldsymbol{\varphi}_i\big) = \big(\tilde{A}\,\boldsymbol{a} + \boldsymbol{a}^\top \tilde{B}\,\boldsymbol{a}\big)_i\,, \tag{23}$$

where, for computational efficiency, we assume that the structures of $\boldsymbol{g}$ and $\boldsymbol{f}$ are similar. Thus, in the least squares problem (21), we solve for all the entries in the $r \times r$ matrix $\tilde{A}$ and the $r \times r \times r$ tensor $\tilde{B}$.

The least squares problem (21) is *low-dimensional* since there are only $(r^2 + r^3)$ entries in $\tilde{A}$ and $\tilde{B}$ to be optimized, and r is small. Thus, (21) can be efficiently solved in the offline stage. For the NSE, the D2-VMS-ROM (22) takes the form

$$\boxed{\overset{\bullet}{\boldsymbol{a}} = (A + \tilde{A})\boldsymbol{a} + \boldsymbol{a}^\top (B + \tilde{B})\boldsymbol{a}\,,} \tag{24}$$

where A and B are the G-ROM operators in (11), and $\tilde{A}$ and $\tilde{B}$ are the VMS-ROM closure operators in (23).

Remark 6 (Physical Constraints) To improve the D2-VMS-ROM accuracy, one can use physical constraints when solving the least squares problem (21) to find the entries of the VMS-ROM closure operators $\tilde{A}$ and $\tilde{B}$. Numerical experiments have shown that imposing physical constraints can indeed increase the D2-VMS-ROM accuracy [35].

In Algorithm 2, we list the main steps in the construction of ROMs equipped with data-driven closure models.

6 ROM Closures in Action: Numerical Results

In the previous sections, we tried to convince the reader that ROM closures are important since they significantly increase the ROM accuracy in the under-

Algorithm 2 Data-driven ROM closure algorithm

```
1: Use numerical or experimental data to construct modes $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_R\}$, which represent the recurrent spatial structures in the system.
2: Choose the dominant modes $\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$, $r \leq R$, as ROM basis functions.
3: Use a Galerkin expansion $\boldsymbol{u}_R(\boldsymbol{x}, t) = \sum_{j=1}^{R} a_j(t)\, \boldsymbol{\varphi}_j(\boldsymbol{x})$.
4: Replace $\boldsymbol{u}$ with $\boldsymbol{u}_R$ in (6).
5: Use a Galerkin projection of the PDE obtained in step 4 onto the space of resolved ROM scales $\mathbf{X}^r := \text{span}\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_r\}$ to obtain the ideal ROM (I-ROM):

       $\overset{\bullet}{\boldsymbol{a}} = \boldsymbol{F}(\boldsymbol{a}) + \boldsymbol{\tau}^{I-ROM},$        (25)

   where $\boldsymbol{a}(t) = (a_i(t))_{i=1,\ldots,r}$ is the vector of coefficients in the Galerkin expansion in step 3, $\boldsymbol{F}$ comprises the G-ROM operators, and $\boldsymbol{\tau}^{I-ROM}$ is the ideal ROM closure term defined in (17).
6: In the offline stage:
     • Compute the G-ROM operators (e.g., vectors, matrices, and tensors), which are preassembled from the ROM basis.
     • Choose a model form $\boldsymbol{g}$ for $\boldsymbol{\tau}^{I-ROM}$ in (25).
     • Solve the least squares problem (21) to find the parameters in the model form.
     • Compute $\boldsymbol{G}(\boldsymbol{a})$, which comprises the ROM closure operators corresponding to the model form $\boldsymbol{g}$ for $\boldsymbol{\tau}^{I-ROM}$.
     • Replace the I-ROM (25) with the data-driven ROM closure model

       $\overset{\bullet}{\boldsymbol{a}} = \boldsymbol{F}(\boldsymbol{a}) + \boldsymbol{G}(\boldsymbol{a}).$        (26)

7: In the online stage, repeatedly use the data-driven ROM closure (26) for various parameter settings and/or longer time intervals.
```

resolved regime. We note, however, that all our arguments have been *mathematical* arguments. Thus, we can ask the following natural question:

? Q6

Do ROM closures work in practice?

The answer to Q6 is simple:

> A6

Yes!

Table 1 2D flow past a circular cylinder. L^2 norm of errors for G-ROM, D2-VMS-ROM, and I-ROM for different r values

r	G-ROM	I-ROM	D2-VMS-ROM
2	1.509e+00	5.987e−02	1.504e−02
3	8.595e−01	5.072e−01	8.024e−02
4	6.583e−01	3.415e−02	2.538e−02
5	7.095e−01	4.197e−01	5.156e−01
6	5.562e−01	2.371e−01	3.132e−02
7	4.760e−01	2.324e−01	6.482e−02
8	2.692e−01	2.122e−01	1.691e−02

The answer A6 is elaborated in the survey in [1], which presents a plethora of examples of under-resolved ROM simulations of complex dynamics (e.g., turbulent flows) in which ROM closures significantly increase the accuracy at a modest computational overhead.

In this section, for clarity of presentation, we illustrate how a specific ROM closure model (i.e., the D2-VMS-ROM outlined in Sect. 5) increases the ROM accuracy for the 2D flow past a circular cylinder [37], which is a simple test problem commonly used in the ROM community. (We note, however, that the D2-VMS-ROM was successfully used for challenging test problems, e.g., turbulent channel flow [36] and the quasi-geostrophic equations [38].) In our numerical investigation, we use a Reynolds number $Re = 1000$ and four ROM basis functions (i.e., $r = 4$). Details of the computational setting can be found in [37].

In Table 1, we list the L^2 norm of the error, $\boldsymbol{u}^{FOM} - \boldsymbol{u}^{ROM}$, for G-ROM (11) (second column), I-ROM (17) (third column), and D2-VMS-ROM (22) (fourth column). We note that the G-ROM error is relatively large, whereas both the D2-VMS-ROM and I-ROM error are much smaller than G-ROM. In particular, the D2-VMS-ROM error is one and even two orders of magnitude smaller than the G-ROM error for some r values. In Fig. 5, we present plots of the Euclidian norm of the error at each mesh point at the final time, for G-ROM (11) (top), I-ROM (17) (middle), and D2-VMS-ROM (22) (bottom). We note that the G-ROM error is relatively large, whereas the D2-VMS-ROM error is almost negligible. These two plots clearly show that adding the data-driven closure model to the classical G-ROM (i.e., using the D2-VMS-ROM) significantly increases the G-ROM accuracy. Although the I-ROM cannot be used in practical computations (since it is not closed), we included I-ROM results for comparison purposes. Table 1 and Fig. 5 show that the D2-VMS-ROM is not only more accurate than the standard G-ROM, but it is almost as accurate as the I-ROM (which includes an ideal closure model). Thus, for this test problem, the D2-VMS-ROM error almost reaches the theoretical lower bound given by the I-ROM error. Overall, Fig. 5 clearly shows that closure models can significantly increase the ROM accuracy in under-resolved simulations.

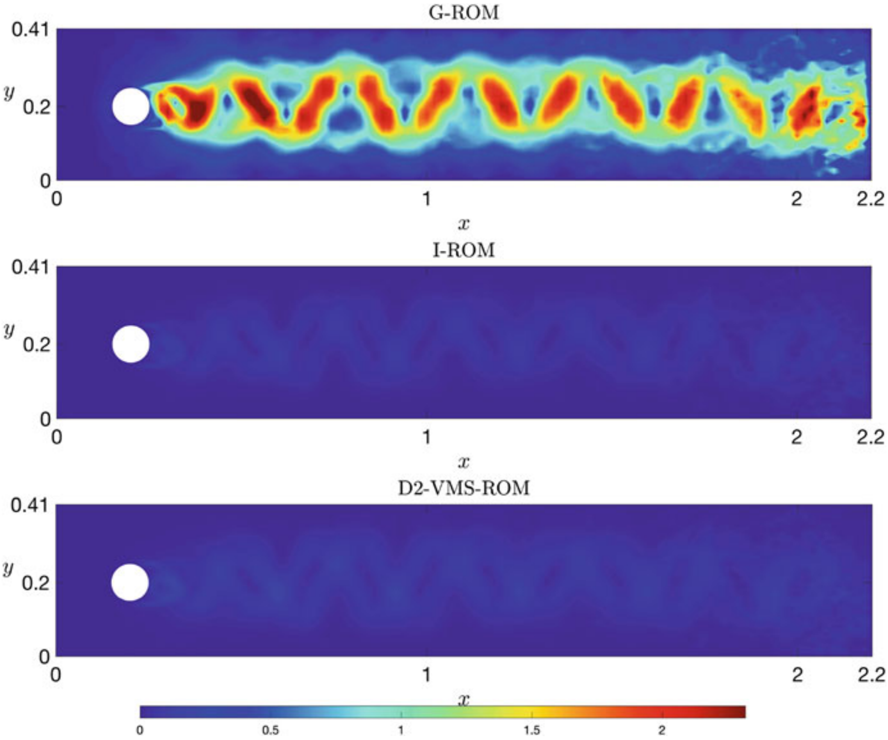


Fig. 5 2D flow past a circular cylinder. The Euclidian norm of the error, $\boldsymbol{u}^{FOM} - \boldsymbol{u}^{ROM}$, at mesh points for G-ROM (11) (top), I-ROM (17) (middle), and D2-VMS-ROM (22) (bottom). The D2-VMS-ROM error is significantly lower than the G-ROM error, which illustrates the benefit of ROM closure modeling. Also note that, in this case, the D2-VMS-ROM error almost reaches the theoretical lower bound given by the I-ROM error

7 Mathematical Foundations of ROM Closures

In Sects. 4 and 5, we discussed the *mathematical modeling* of ROM closures. In Sect. 6, we discussed the *numerical simulation* of ROM closures. The following is a natural question:

? Q7

What can we prove about ROM closures?

The answer to Q7 is simple:

A7

Not so much. Yet.

In this section, we briefly summarize some relevant theoretical aspects associated with ROM closure modeling. Compared with the analysis of classical numerical schemes [6, 27, 44], the theoretical foundations for ROM closures are much less developed. We emphasize, however, that recently there have been significant advancements in this exciting and important research area.

The theoretical investigations of ROM closure modeling generally aim at proving error bounds for ROM closures of the form

$$\|\boldsymbol{u}^{FOM} - \boldsymbol{u}^{ROM}\| \leq C\,(\text{space error} + \text{time error} + \text{ROM error})\,, \tag{27}$$

where $\boldsymbol{u}^{FOM}$ is the FOM solution, $\boldsymbol{u}^{ROM}$ is the ROM solution, $\|\cdot\|$ is a given norm, the space error is the error that results from the spatial approximation, the time error is the error that results from the time approximation, the ROM error is the error that results from the ROM approximation, and C is a generic constant that does not depend on the discretization parameters. We note that the first two terms on the right-hand side of (27) appear in error bounds for classical numerical discretizations, e.g., the FEM [27]. The third term, however, does not appear in these bounds.

The main purpose of the error bound (27) is to show the *convergence* of the ROM solution to the FOM solution. For example, as the spatial mesh size and the time step go to zero, the space error and time error in (27), respectively, are expected to go to zero (at a rate that depends on the particular spatial and time discretizations used). Furthermore, as the number of ROM basis functions goes to the rank of the snapshot matrix, the ROM error in (27) is also expected to go to zero. Thus, as the right-hand side of (27) goes to zero, so does the error on the left-hand side of (27), which proves the convergence of the ROM solution to the FOM solution.

For the G-ROM (11), the numerical analysis started two decades ago with the pioneering work of Kunisch and Volkwein, who proved the first error bounds for the POD of parabolic equations, e.g., the heat equation [31] and the Navier-Stokes equations [32]. More than a decade later, Singler improved Kunisch and Volkwein's results, by proving sharper error bounds [47]. Recently, optimal pointwise in time error bounds were proved in [30]. These results finally bring the G-ROM numerical analysis to a level comparable to (although not as developed as) the level of the numerical analysis of the FEM.

For the ROM closure models, the numerical analysis is relatively scarce. The numerical analysis for ROM closures aims at proving a modified form of the G-ROM error bound (27):

$$\|\boldsymbol{u}^{FOM} - \boldsymbol{u}^{ROM}\| \leq C \, (\text{space error} + \text{time error} + \text{ROM error} + \text{closure error}) \, , \tag{28}$$

where the closure error is the error that results from the approximation of the closure term $\boldsymbol{\tau}^{I-ROM}$ in the I-ROM (17) with a closure model.

As mentioned in [1], the first numerical analysis of ROM closures was performed in [7], where error bounds for the time discretization of the Smagorinsky model (i.e., a ROM closure model developed on phenomenological arguments) were proven. Error bounds for the time and space discretizations of the Smagorinsky model were later proven in [43] in an RBM context. Error bounds for VMS closure models were proved in [19, 25, 26, 45] (see also [4, 46] for related work). Finally, error bounds for the D2-VMS-ROM (22) were proved in [29] (see also [28] for related work).

8 Conclusions and Outlook

In this paper, we presented a brief tutorial for reduced order model (ROM) closures. In the first part of our tutorial, we motivated the ROM closures. We note that ROM closure modeling is often misunderstood in the ROM community. Thus, we started our tutorial by explaining the need for ROM closure modeling (i.e., the ROM closure problem) in realistic applications, and then we carefully described the ROM closure model. Specifically, we first outlined the main steps used to construct the Galerkin ROM (G-ROM), which is based on leveraging a data-driven basis in the classical Galerkin framework. Next, we noted that, although G-ROM can decrease the computational cost of standard numerical discretizations by orders of magnitude, it yields inaccurate results in under-resolved ROM simulations, i.e., when the number of basis functions is not enough to capture the underlying system's dynamics. To address the G-ROM's inaccuracy in under-resolved simulations, we introduced the ROM closure model. We motivated the need for ROM closure by presenting a mathematical extension of the classical Galerkin framework to include not only the space of resolved scales, but also the space of unresolved scales. In this extended variational multiscale framework, we showed that the correct ROM dynamics include an additional term (i.e., the closure term), which represents the effect of the unresolved scales. Furthermore, we showed that this mathematical framework, which we named the ideal ROM (I-ROM), yields numerical results that are significantly more accurate than the G-ROM results. Thus, we concluded that a ROM closure model, which is a practical model for the I-ROM closure term, should be added to the G-ROM to increase its accuracy in realistic, under-resolved simulations.

In the second part of our tutorial, we outlined the main steps in the construction of ROM closure models. To simplify our presentation, we focused on one particular type of ROM closure modeling, i.e., data-driven modeling. Furthermore, we illustrated this construction for one specific data-driven ROM closure model, i.e., the data-driven variational multiscale ROM (D2-VMS-ROM). In our construction, we started with the closure term in the I-ROM, and we simply posed the closure problem as leveraging the available FOM data to find the "best" ROM closure model. To this end, we first postulated a model form for the ROM closure model. Then, we solved a least squares problem to find the parameters in the model form that yield the ROM closure model that is the closest to the ideal ROM closure model. Finally, we also included numerical results for the two-dimensional flow past a circular cylinder, which showed that the D2-VMS-ROM was significantly more accurate than the standard G-ROM, and almost as accurate as the I-ROM. These numerical results illustrated the significant benefit of ROM closure modeling in under-resolved simulations.

We hope that this brief tutorial offers a glimpse into the exciting research field of ROM closure modeling, which has witnessed a significant development over the past two decades. This research area is currently experiencing a dynamic development in several directions. One of the most active research directions is the use of machine learning tools to construct more accurate and more efficient ROM closure models. Recently, deep learning models have been shown to be quite effective and computationally efficient in capturing the relationship between resolved and unresolved scales [2]. However, these models often need large amounts of training data and their generalization, expressivity, and analysis still remain mostly challenging.

Another important research direction is the development of ROM closures for problems in solid mechanics. Although most ROM closure modeling has been performed in computational fluid dynamics [1], there has been recent work done in solid mechanics. For example, approximations of the mechanical behavior of soft tissue showed substantial improvement in accuracy over G-ROM with the addition of ROM closure terms at a modest computational overhead [48]. The ability of ROM closure to capture the nonlinearities of soft tissue behavior is especially promising for its application in biomechanics.

Depending on the applications, one can also couple ROMs with additional parameterization schemes or surrogate models for some of the unresolved scales in order to recover more dynamical features of the original system, especially when the ROMs are constructed for under-resolved dynamical regimes. For instance, in the context of data assimilation, when observations are only available for the (large-scale) low-frequency modes, one can design computationally efficient strategies within the conditional Gaussian framework [12–14] to approximate the dynamics of the high-frequency (unresolved) modes with quantified uncertainties by a suitable dynamical model for the unresolved modes.

Finally, providing mathematical support for ROM closures is also an important research direction. We note that significant mathematical support has been provided for closures in classical computational fluid dynamics [6, 27, 44]. For ROM closures, however, only the first steps have been taken and much more remains to be done.

References

1. S.E. Ahmed, S. Pawar, O. San, A. Rasheed, T. Iliescu, B.R. Noack, On closures for reduced order models—a spectrum of first-principle to machine-learned avenues. Phys. Fluids **33**(9), 091301 (2021)
2. S.E. Ahmed, O. San, A. Rasheed, T. Iliescu, A long short-term memory embedding for hybrid uplifted reduced order models. Phys. D 132471 (2020)
3. S.E. Ahmed, O. San, A. Rasheed, T. Iliescu, A. Veneziani, Physics guided machine learning for variational multiscale reduced order modeling (2022). in preparation
4. M. Azaïez, T.C. Rebollo, S. Rubino, A cure for instabilities due to advection-dominance in POD solution to advection-diffusion-reaction equations. J. Comput. Phys. **425**, 109916 (2021)
5. J. Berner, U. Achatz, L. Batté, L. Bengtsson, A. de la Cámara, H.M. Christensen, M. Colangeli, D.R.B. Coleman, D. Crommelin, S.I. Dolaptchiev, C.L.E. Franzke, P. Friederichs, P. Imkeller, H. Järvinen, S. Juricke, V. Kitsios, F. Lott, V. Lucarini, S. Mahajan, ..., and J.-I. Yano, Stochastic parameterization toward a new view of weather and climate models. Bull. Am. Meteorol. Soc. **98**(3), 565–588 (2017)
6. L.C. Berselli, T. Iliescu, W.J. Layton, *Mathematics of Large Eddy Simulation of Turbulent Flows*. Scientific Computation. (Springer, Berlin, 2006)
7. J. Borggaard, T. Iliescu, Z. Wang, Artificial viscosity proper orthogonal decomposition. Math. Comput. Model. **53**(1–2), 269–279 (2011)
8. S.L. Brunton, J.N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, Cambridge, 2019)
9. J. Burkardt, M. Gunzburger, H.C. Lee, POD and CVT-based reduced-order modeling of Navier–Stokes flows. Comput. Methods Appl. Mech. Eng. **196**(1–3), 337–355 (2006)
10. M.D. Chekroun, H. Liu, J.C. McWilliams, Variational approach to closure of nonlinear dynamical systems: autonomous case. J. Stat. Phys. **179**, 1073–1160 (2020)
11. M.D. Chekroun, H. Liu, S. Wang, *Stochastic parameterizing manifolds and non-Markovian reduced equations: stochastic manifolds for nonlinear SPDEs II*. Springer Briefs in Mathematics (Springer, Berlin, 2015)
12. N. Chen, Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics. J. Comput. Phys. **418**, 109635 (2020)
13. N. Chen, Y. Li, H. Liu, Conditional Gaussian nonlinear system: a fast preconditioner and a cheap surrogate model for complex nonlinear systems (2021). arXiv preprint arXiv:2112.05226
14. N. Chen, A.J. Majda, Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. Entropy **20**(7), 509 (2018)
15. F. Chinesta, P. Ladeveze, E. Cueto, A short review on model order reduction based on proper generalized decomposition. Arch. Comput. Methods Eng. **18**(4), 395–404 (2011)
16. A.J. Chorin, F. Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics. Proc. Natl. Acad. Sci. U.S.A. **112**(32), 9804–9809 (2015)
17. D. Crommelin, E. Vanden-Eijnden, Subgrid-scale parameterization with conditional Markov chains. J. Atmos. Sci. **65**(8), 2661–2675 (2008)
18. V. DeCaria, T. Iliescu, W. Layton, M. McLaughlin, M. Schneier, An artificial compression reduced order model. SIAM J. Numer. Anal. **58**(1), 565–589 (2020)
19. F.G. Eroglu, S. Kaya, L.G. Rebholz, A modular regularized variational multiscale proper orthogonal decomposition for incompressible flows. Comput. Methods Appl. Mech. Eng. **325**, 350–368 (2017)
20. C. Foiaş, O. Manley, R. Rosa, R. Temam, *Navier–Stokes Equations and Turbulence* (Cambridge University Press, Cambridge, 2001)
21. J.-L. Guermond, Stabilization of Galerkin approximations of transport equations by subgrid modeling. M2AN Math. Model. Numer. Anal. **33**(6), 1293–1316 (1999)
22. J.S. Hesthaven, G. Rozza, B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations* (Springer, Berlin, 2015)

23. P. Holmes, J.L. Lumley, G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge, 1996)
24. T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. Comput. Methods Appl. Mech. Eng. **127**(1–4), 387–401 (1995)
25. T. Iliescu, Z. Wang, Variational multiscale proper orthogonal decomposition: convection-dominated convection-diffusion-reaction equations. Math. Comput. **82**(283), 1357–1378 (2013)
26. T. Iliescu, Z. Wang, Variational multiscale proper orthogonal decomposition: Navier-Stokes equations. Num. Methods P.D.E.s **30**(2), 641–663 (2014)
27. V. John, *Finite Element Methods for Incompressible Low Problems* (Springer, Berlin, 2016)
28. B. Koc, M. Mohebujjaman, C. Mou, T. Iliescu, Commutation error in reduced order modeling of fluid flows. Adv. Comput. Math. **45**(5–6), 2587–2621 (2019)
29. B. Koc, C. Mou, H. Liu, Z. Wang, G. Rozza, T. Iliescu, Verifiability of the data-driven variational multiscale reduced order model (2021). http://arxiv.org/abs/2108.04982
30. B. Koc, S. Rubino, M. Schneier, J.R. Singler, T. Iliescu, On optimal pointwise in time error bounds and difference quotients for the proper orthogonal decomposition. SIAM J. Numer. Anal. **59**(4), 2163–2196 (2021)
31. K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for parabolic problems. Numer. Math. **90**(1), 117–148 (2001)
32. K. Kunisch, S. Volkwein, Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. SIAM J. Numer. Anal. **40**(2), 492–515 (electronic) (2002)
33. W.J. Layton, A connection between subgrid scale eddy viscosity and mixed methods. Appl. Math. Comput. **133**, 147–157 (2002)
34. A.J. Majda, I. Timofeyev, E. Vanden-Eijnden, A mathematical framework for stochastic climate models. Commun. Pure Appl. Math. **54**, 891–974 (2001)
35. M. Mohebujjaman, L.G. Rebholz, T. Iliescu, Physically-constrained data-driven correction for reduced order modeling of fluid flows. Int. J. Num. Meth. Fluids **89**(3), 103–122 (2019)
36. C. Mou, *Data-Driven Variational Multiscale Reduced Order Modeling of Turbulent Flows*. Ph.D. Thesis, Virginia Tech, 2021
37. C. Mou, B. Koc, O. San, L.G. Rebholz, T. Iliescu, Data-driven variational multiscale reduced order models. Comput. Methods Appl. Mech. Eng. **373**, 113470 (2021)
38. C. Mou, H. Liu, D.R. Wells, T. Iliescu, Data-driven correction reduced order models for the quasi-geostrophic equations: a numerical investigation. Int. J. Comput. Fluid Dyn. **34**, 147–159 (2020)
39. B.R. Noack, M. Morzynski, G. Tadmor, *Reduced-Order Modelling for Flow Control*, vol. 528 (Springer, Berlin, 2011)
40. S. Pawar, S.E. Ahmed, O. San, A. Rasheed, Data-driven recovery of hidden physics in reduced order modeling of fluid flows. Phys. Fluids **32**(3), 036602 (2020)
41. S. Pawar, S.E. Ahmed, O. San, A. Rasheed, An evolve-then-correct reduced order model for hidden fluid dynamics. Mathematics **8**(4), 570 (2020)
42. A. Quarteroni, A. Manzoni, F. Negri, *Reduced Basis Methods for Partial Differential Equations: An Introduction*, vol. 92 (Springer, Berlin, 2015)
43. T.C. Rebollo, E.D. Ávila, M.G. Mármol, F. Ballarin, G. Rozza, On a certified Smagorinsky reduced basis turbulence model. SIAM J. Numer. Anal. **55**(6), 3047–3067 (2017)
44. T. Chacón Rebollo, R. Lewandowski, *Mathematical and Numerical Foundations of Turbulence Models and Applications* (Springer, Berlin, 2014)
45. J.P. Roop, A proper-orthogonal decomposition variational multiscale approximation method for a generalized Oseen problem. Adv. Numer. Anal. (2013)
46. S. Rubino, Numerical analysis of a projection-based stabilized POD-ROM for incompressible flows. SIAM J. Numer. Anal. **58**(4), 2019–2058 (2020)
47. J.R. Singler, New POD error expressions, error bounds, and asymptotic results for reduced order models of parabolic PDEs. SIAM J. Numer. Anal. **52**(2), 852–876 (2014)

48. W. Snyder, J.A. McGuire, C. Mou, D. A. Dillard, T. Iliescu, R. De Vita, Data-driven variational multiscale reduced order modeling of vaginal tissue (2022). in preparation
49. K. Taira, M.S. Hemati, S.L. Brunton, Y. Sun, K. Duraisamy, S. Bagheri, S.T. M. Dawson, C.-A. Yeh, Modal analysis of fluid flows: applications and outlook. AIAA J. **58**(3), 998–1022 (2020)
50. R. Temam, *Navier-Stokes equations: Theory and Numerical Analysis*, vol. 2 (American Mathematical Society, Providence, 2001)
51. S. Volkwein, Proper orthogonal decomposition: theory and reduced-order modelling. Lecture Notes, University of Konstanz (2013). http://www.math.uni-konstanz.de/numerik/personen/volkwein/teaching/POD-Book.pdf
52. X. Xie, M. Mohebujjaman, L.G. Rebholz, T. Iliescu, Data-driven filtered reduced order modeling of fluid flows. SIAM J. Sci. Comput. **40**(3), B834–B857 (2018)
53. L. Zanna, P. Porta Mana, J. Anstey, T. David, T. Bolton, Scale-aware deterministic and stochastic parametrizations of eddy-mean flow interaction. Ocean Model. **111**, 66–80 (2017)

Artificial Stress Diffusion in Numerical Simulations of Viscoelastic Fluid Flows

Marília Pires and Tomáš Bodnár

1 Introduction

Viscoelastic fluids are quite common in many areas of industrial, environmental, and biomedical fluid mechanics. There exist a number of models describing specific sub-classes of these fluids, capturing their various distinct properties. Most of these models are rather complex, relating the stress tensor with the fluid rate of deformation tensor and its history. This is why the mathematical modeling and numerical simulations of viscoelastic fluid flows are some of the most challenging problems of contemporary computational fluid dynamics.

1.1 Motivation

The motivation for present work comes from the biomedical fluid mechanics, where the viscoelastic fluid models are often used to describe specific behavior of blood, synovial fluids, and various other gel-like bio materials [8]. The flow of such fluids is

M. Pires
CIMA-UE and Mathematics Department, Technology Sciences School, University of Évora, Rua Romão Ramalho, Évora, Portugal

CEMAT, Instituto Superior Técnico, Lisbon, Portugal
e-mail: marilia@uevora.pt

T. Bodnár (✉)
Institute of Mathematics, Czech Academy of Sciences, Prague 1, Czech Republic

Department of Technical Mathematics, Faculty of Mechanical Engineering, Czech Technical University in Prague, Prague 2, Czech Republic
e-mail: Tomas.Bodnar@fs.cvut.cz

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_9

described by a set of coupled partial differential equations, representing the balance laws for mass and linear momentum, complemented by suitable rheological model for the stress tensor. One of the classical and most common models for viscoelastic fluid flows is the Oldroyd-B model (described further in Sect. 2.2). This model shares the structure and many properties with a whole class of other more complex rate type models for viscoelastic liquids. This is why the Oldroyd-B model is used in this work as a prototype of a viscoelastic fluid model, showing some of its most distinct properties.

It was found by many authors in the past, that while solving the governing equations of the Oldroyd-B model, the numerical methods often fail to converge in certain regimes, due to instabilities encountered in the solution process. The critical regime at which the numerical instabilities occur is related to the characteristic Weissenberg number of the solved problem. The loss of numerical stability at high Weissenberg numbers was addressed in a number of works in the past decades, developing various specific schemes and algorithms to fight the numerical instabilities [6, 33].

1.2 Artificial Diffusion Concept

The concept of numerical and artificial diffusion is well known in computational fluid dynamics. It starts with the presence and/or absence of diffusive terms in the mathematical models of fluid flows. The Euler equations of fluid dynamics (for both compressible and incompressible fluids) can be understood as the limit (singular) model arising from the original (viscous) Navier-Stokes equations for vanishing viscosity. It can be shown that limit solutions emanating from a sequence of viscous model solutions for successively decreasing viscosity lead to physically relevant solutions of the inviscid model [7, 14]. The vanishing viscosity is thus not only relevant in developing the inviscid model, but it also plays an important role in obtaining the successive approximate solutions to that model, possibly leading to unique physically realistic solution of the problem [3, 4].

At the discrete level, most of the numerical methods are associated with certain level of numerical diffusion or dispersion. It can be shown that the leading order term in the discretization error can either have diffusive or dispersive character. It is well known that certain amount of numerical diffusion is necessary for numerical methods to be stable and robust. Such numerical diffusion can either be directly embedded in the numerical scheme as a part (or side effect) of the discrete approximation, or it can be added artificially as a special term or step in the algorithm.

The embedded *numerical diffusion* is typically introduced by some kind of upwinding or intentional use of highly diffusive numerical methods. It can, for example, be shown that the (first order) upwind scheme can be rewritten as central scheme with added numerical diffusion of specific form. The Lax-Friedrichs scheme can be used as an example of highly diffusive scheme, which again can be rewritten into the form of standard (non-diffusive) central scheme and additional diffusive term. Such highly diffusive methods are quite robust, but due to excessive diffusion their accuracy is reduced to first order only. The detailed analysis based on the modified equation approach can be found in [13, 19] or [2]. When the level of diffusivity can't easily be adjusted in these methods, they are often combined with higher order (but less diffusive and less robust) methods to achieve better accuracy with only minor sacrifices regarding the overall robustness of the combined method. This is, for example, the case of so-called Total Variation Diminishing (TVD) schemes [11, 20], where the (less diffusive) higher order discretization is used where the solution is smooth enough, while dropping to some first order (more diffusive) method in the proximity of high solution gradients or shocks. Such spatial blending of numerical methods with various levels of numerical diffusion can lead to desired robustness and increased accuracy of the numerical scheme. Similar effect can be achieved by the so-called composite schemes [21, 22], where most of the time-steps (or iterations) are performed using higher order (less diffusive) method, while some (smoothing, stabilizing) steps are performed by a diffusive method with lower accuracy. The ratio higher versus lower order steps can be adjusted, so the total level of numerical diffusion introduced to the numerical solution can be kept under control.

The *artificial diffusion* approach is similar in principle, except that the numerical diffusion is not an intrinsic part of the numerical discretization (as it is, e.g., in the Lax-Friedrichs or upwind scheme), but is expressed separately, either as an extra term in the scheme or extra step in the numerical method. Typically some standard less-diffusive scheme (of possibly higher order) is used, and the extra numerical diffusion is added artificially at the next stage. Typically such added artificial diffusion mimics the physical diffusion terms being proportional to Laplacian or bi-Laplacian of the corresponding quantity. Such artificial diffusion can either be seen as a separate smoothing (post-processing) step, or as an operator splitting method applied at the discretization level. This allows full control over the form, behavior, and amount of the numerical diffusion added by the numerical method. Such numerical schemes involving an artificial diffusion were extensively studied and used in past decades. For some practical examples of artificial diffusion terms, see, for example, [15, 16]. Some hints concerning the effects of numerical diffusion on the stability of numerical schemes can be found in [10].

The fact that the diffusion added to mathematical models and numerical methods has some physical motivation can be seen as an important advantage over many other purely artificial and algorithmic stabilization approaches. In many cases the artificially added numerical diffusion just substitutes the physical diffusion that was dropped out from the mathematical model as a consequence of its (over-) simplification.

1.3 Tensorial Stress Diffusion in Oldroyd-B-Like Models

As mentioned before, the added diffusion terms are important for the well posedness of mathematical models as well as for stabilization of numerical methods. This concept proved to be useful for numerical solution of many fluid mechanics problems, which was the main motivation for testing this approach also for the stabilization of numerical models of viscoelastic fluid flows.

In the context of this paper, the viscoelastic fluids are modeled by a constitutive relation, linking together the flow kinematics with the stress in the fluid, represented by a stress tensor. The corresponding governing equations describing the spatio-temporal evolution of this stress tensor can be seen as a specific type of (tensorial) transport equations, with non-linear source terms depending on the flow field. In the Oldroyd-B model (and many other models with similar structure), only the advection and source terms are present, but there are no diffusion terms (see Sect. 2.2). It is an interesting open question, whether from the physical point of view, the presence (or absence) of tensorial diffusion in the viscoelastic constitutive relations can be justified (is necessary) [4]. There exist viscoelastic fluid models containing some physical diffusion [23, 32], mostly based on arguments related to microstructure of the fluid. These models however form a specific category of rheological constitutive laws and are not subject of the present investigation. Here such diffusive viscoelastic models only serve as motivation for the use of artificial diffusion in numerical simulations of otherwise non-diffusive viscoelastic fluid flow models.

In order to introduce and analyze the tensorial stress diffusion in the Oldroyd-B model, the constitutive relation (20) is modified by adding an extra term $\boldsymbol{E}$ containing the continuous version of the additional diffusive terms aimed to stabilize the numerical simulations at the discrete level. The artificially extended constitutive model has the form:

$$\boldsymbol{\tau} + \mathcal{W}e\left(\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \nabla\boldsymbol{u}^T\cdot\boldsymbol{\tau} - \boldsymbol{\tau}\cdot\nabla\boldsymbol{u}\right) = 2\eta\mathbf{D} + \boldsymbol{E} \quad . \tag{1}$$

Different variants of the added artificial diffusion term $\boldsymbol{E}$ were used in our previous works [26, 27, 29] and [28]. Here all these variants are summarized and discussed.

(a) Constant diffusive term proportional to the Laplacian of elastic stress:[1]

$$\boldsymbol{E} = \alpha\cdot\Delta\,\boldsymbol{\tau} \approx \alpha\cdot\Delta\,\boldsymbol{\tau}^n \quad . \tag{2}$$

[1] The temporal index n corresponds to pseudo-time, used in the time-marching iterative procedure (see the description of numerical algorithm, Sect. 3, for more details).

This simplest and most primitive form of the artificial diffusion term is closest to the physical diffusion which sometimes appears in the truly diffusive constitutive laws. In general, this extra term is always present in the model and does not vanish even when the solution reaches the steady state, so $\boldsymbol{E} = \alpha \cdot \Delta \boldsymbol{\tau}^n \nrightarrow 0$ when $\boldsymbol{\tau}^n \xrightarrow{n \to \infty} \boldsymbol{\tau}$. This means that rather than just stabilizing the numerical solution of the non-diffusive model, the whole model is modified by the added diffusive term and the results may (will) depend on the values of the parameter α. In numerical simulations this artificial diffusion coefficient α should be kept, small, at least of order $O(h^2)$ to preserve the consistency of the numerical method with the original non-diffusive problem. This however only guarantees that for $h \to 0$ the whole artificial diffusive term will asymptotically vanish; however on any finite size grid the term remains to be present, affecting the final solution.

(b) *Time-dependent* diffusive term is just a minor modification of the above-mentioned constant diffusive term, where instead of the constant diffusion coefficient α, it is made time-dependent, i.e., $\alpha = \alpha(t)$, leading to:

$$\boldsymbol{E} = \alpha(t) \cdot \Delta\, \boldsymbol{\tau} \approx \alpha(t) \cdot \Delta\, \boldsymbol{\tau}^n \quad . \tag{3}$$

The purpose of the variable in time artificial diffusion coefficient $\alpha(t)$ is to make it decay in time, to make it eventually vanish in the limit for $t \to \infty$. The idea is to keep the artificial diffusion term active just during the initial stage of the iterative process, while letting it to vanish later when it's no more needed. So it will help to overcome the initial solution instabilities, but it will not affect the final solution. The diffusion coefficient function $\alpha(t)$ should be tuned accordingly, to be strong enough at the beginning of the simulation and decay fast enough to vanish at the end of the iterative process.

(c) *Time-derivative dependent* variant of the diffusive term takes the form

$$\boldsymbol{E} = \alpha(\phi_t) \cdot \Delta\, \boldsymbol{\tau} \approx \alpha(\phi_t) \cdot \Delta\, \boldsymbol{\tau}^n \quad , \tag{4}$$

where ϕ is suitably chosen flow quantity and t stands for iterative (pseudo) time. The goal is to avoid the need of tricky manual adjustments of constant diffusion coefficient α or variable $\alpha(t)$ and rather make the whole process automatic, by choosing the diffusion coefficient $\alpha(\phi_t)$ such that it will decay with the time-derivative $\|\phi_t\| \to 0$, meaning that the added artificial diffusion automatically vanishes when the steady solution is reached. In this case there is no need to a-priori estimate the number of iterations till the steady state (as for time-dependent $\alpha(t)$) to adjust the appropriate diffusion decay, and it is guaranteed that the artificial diffusion will automatically vanish for the steady solution (when $\phi_t = 0$).

(d) *Residual* diffusive term is made proportional to the Laplacian of the (pseudo) time derivative of the stress:

$$\boldsymbol{E} = \alpha \cdot \Delta \boldsymbol{\tau}_t \approx \alpha \cdot \Delta \left(\boldsymbol{\tau}^n - \boldsymbol{\tau}^{n-1} \right) \quad . \tag{5}$$

The temporal index n corresponds to pseudo-time, used in the time-marching iterative procedure. In steady problem iterative solution, the difference $\left(\boldsymbol{\tau}^n - \boldsymbol{\tau}^{n-1}\right)$ can be considered as a steady residual of the problem that should converge to zero, i.e., $\boldsymbol{\tau}^n \rightarrow \boldsymbol{\tau}$ meaning that $\left(\boldsymbol{\tau}^n - \boldsymbol{\tau}^{n-1}\right) \rightarrow \mathbf{0}$. This also implies that the extra term $\boldsymbol{E}$ in the form (5) will vanish when the numerical solution converges to steady state. Due to this property, the added diffusivity will only act during the transitional stage of (pseudo) time stepping. As a consequence, the solution of the original (non-diffusive) model is recovered, and the final results should not depend on the choice of the parameter α. The whole stabilization process in this case can be seen as residual smoothing, rather than the stress tensor smoothing used in the previous three variants of artificial stabilization involving just the Laplacian of the elastic stress tensor $\Delta\,\boldsymbol{\tau}$.

1.4 Structure and Aim of This Work

This paper is meant as a summary and overview of the various methods of artificial stress diffusion applicable to Oldroyd-B model (introduced in Sect. 2) of viscoelastic fluids (and related models). The description of numerical method and its implementation details are given in full length in Sects. 3 and 4, supplementing the partial presentations in our previous papers [26, 27, 29] and [28]. Some additional numerical simulations and their results are presented in Sect. 5 documenting the main conclusions of this work. The final Sect. 6 is fully dedicated to extended discussion of the obtained results and practical experience.

2 Mathematical Model

From the physical point of view, fluid is a substance that does not resist to deformation under the action of an external force and shear (tangential) stress. The deformation of fluid tends to recover the hydrostatic balance (hydrostatic stress-free condition) where the shear stress forces are null. Typically the fluids are characterized by their viscosity and density.

The (dynamic) viscosity $\mu > 0$ is a physical property (which may depend on the temperature) that characterizes the resistance of the fluid to flow. As the temperature increases, the viscosity of the liquid decreases, and consequently the velocity of the fluid may increase. In the absence of external forces, only the friction forces generated due to viscosity act on the fluid in motion and may eventually force it to come to rest.

The density $\rho > 0$ is the fluid mass per unit of volume. In incompressible fluids the mass of certain fluid volume is preserved, which in case of invariant volume leads to preservation of density. In the particular case of homogeneous incompressible fluids, it results in constant density.

2.1 Equations of Motion

Let ρ be the mass density, $\mu > 0$ the dynamic viscosity, $\boldsymbol{u}$ the velocity vector field, p the pressure, and $\mathbf{T}$ the stress tensor field of an unsteady, incompressible, homogeneous, isothermal fluid flow in a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ during the interval of time $\left[0, T_f\right]$ $(T_f > 0)$.

The governing equations are defined by two fundamental principles:[2]

They arise from the balance laws for mass and linear momentum of the fluid.

- *Mass balance—continuity equation*. The balance of mass (in absence of sources/sinks) is written in differential form as

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{u}) = 0 \quad . \tag{6}$$

This equation can be rewritten using the material derivative as

$$\frac{D\rho}{Dt} = -\rho \nabla \cdot \boldsymbol{u} \quad , \tag{7}$$

where the material derivative $\dfrac{D}{Dt} = \dfrac{\partial}{\partial t} + \boldsymbol{u} \cdot \nabla$ describes the time rate of change of the quantity being transported by the fluid at velocity $\boldsymbol{u}$.

The physical requirement of incompressibility of a fluid implies that its volume remains constant (independently of applied force and deformation). This property can be expressed by the necessary condition on the velocity field stating that it must be divergence-free, i.e., $\nabla \cdot \boldsymbol{u} = 0$. If in addition the fluid is assumed (required) to be homogeneous, i.e., having constant density (both in space and time) $\rho = const$, the continuity equation (6) is satisfied automatically. Therefore in the models of incompressible (homogeneous) fluid flows, only the incompressibility constraint

$$\nabla \cdot \boldsymbol{u} = 0 \tag{8}$$

is considered in place of the continuity equation.

[2] Assuming the fluid as a continuum media. This means, the properties of the fluid vary continuously in space. With this assumption, it is possible to use differential calculus in fluid mechanics problems, although there may be jumps or discontinuities in the properties of the fluids.

- *Momentum balance* describes the inertial effects in fluid flows. It is based on Newton's second law stating that the momentum rate of change of fluid is equal to the net force acting on it.

$$\rho \frac{D\boldsymbol{u}}{Dt} = \nabla \cdot \mathbf{T} - \nabla p \quad . \tag{9}$$

  Here just the pressure gradient force is considered, together with the action of the stress tensor $\mathbf{T}$ expressed via its divergence.

The system of equations of motion represented by the continuity and momentum equations (8) and (9) must further be complemented (closed) by suitable constitutive relation for the stress tensor $\mathbf{T}$.

2.2 Constitutive Relation: Oldroyd-B Model

Constitutive relation postulates in mathematical form the mechanical behavior of fluid in terms of the tension states related to the velocity gradient (rate of deformation). Such relations can have either explicit form $\mathbf{T} = f(\mathbf{D})$ or more general implicit form $f(\mathbf{T}, \mathbf{D}) = 0$ [30]. For the fluids of Oldroyd-B type the constitutive law is defined by

$$\mathbf{T} + \lambda_1 \overset{\nabla}{\mathbf{T}} = 2\left(\mu_s \mathbf{D} + \lambda_2 \overset{\nabla}{\mathbf{D}}\right) \quad , \tag{10}$$

where the parameter $\lambda_1 > 0$ corresponds to the relaxation time scale indicating the time during which the fluid remembers the history of stress. The retardation time scale parameter $\lambda_2 > 0$ follows from the response time of the fluid to sudden application of tension, μ_s is the solvent (dynamic) viscosity, and $\mathbf{D} = \frac{1}{2}\left(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\right)$ is the rate of deformation tensor. The upper convected derivative of any tensor $\mathbf{M}$ defined by $\overset{\nabla}{\mathbf{M}} = \frac{D\mathbf{M}}{Dt} - \nabla \boldsymbol{u}\mathbf{M} - \mathbf{M}\left(\nabla \boldsymbol{u}\right)^T$ describes the rate of change of the tensor $\mathbf{M}$ in coordinate system which stretches and rotates with the fluid. This derivative can be seen as generalization of the material time derivative assuring that the constitutive model is objective (meaning that the laws of motion are the same independently of the inertial frame) [1, 17, 24].

The stress tensor $\mathbf{T}$ can further be decomposed into viscoelastic part $\boldsymbol{\tau}$ (so-called extra stress) and the purely viscous (Newtonian or solvent) stress component.

$$\mathbf{T} = 2\frac{\lambda_2}{\lambda_1}\mu \mathbf{D} + \boldsymbol{\tau} \quad . \tag{11}$$

Taking into account that $\lambda_1 = \frac{\mu_e}{G}$ and $\lambda_2 = \lambda_1 \frac{\mu_s}{\mu}$, where μ_e, μ_s and $\mu = \mu_s + \mu_e$ are the elastic, solvent, and total viscosities, respectively, and G is the Young modulus, considering the decomposition (11), the momentum equations (9) and the constitutive law (10) can be rewritten as

$$\rho \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} \right) = 2\mu_s \nabla \cdot \mathbf{D} + \nabla \cdot \boldsymbol{\tau} - \nabla p \quad , \tag{12}$$

$$\boldsymbol{\tau} + \lambda_1 \left(\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\tau} - \nabla \boldsymbol{u} \boldsymbol{\tau} - \boldsymbol{\tau} \left(\nabla \boldsymbol{u} \right)^T \right) = 2\mu_e \mathbf{D} \quad . \tag{13}$$

The complete set of governing equations describing the Oldroyd-B fluid flow is represented by Eqs. (8), (12), and (13).

2.3 Dimensionless Form of Equations

Let U and L be the characteristic velocity of the fluid and the characteristic length scale of the domain, respectively. This also determines a time scale $T = \frac{L}{U}$. Using these scales a non-dimensional coordinate system is defined by

$$x = \frac{\tilde{x}}{L} , \qquad\qquad t = \frac{U\tilde{t}}{L} , \tag{14}$$

where the tilde symbol $\sim$ is used to denote the original dimensional parameters. For viscoelastic fluid flows, the Reynolds and Weissenberg numbers $\mathcal{Re}$ and $\mathcal{We}$ can be used to characterize the flow.

The *Reynolds number* is defined as the ratio between the inertial forces and the viscous forces expressed by

$$\mathcal{Re} = \frac{UL}{\nu} \quad , \tag{15}$$

where $\nu = \frac{\mu}{\rho}$ is the kinematic viscosity.

The *Weissenberg number* characterizes the relative importance of the elasticity of the fluid, defined as the rate of two characteristic time scales. The first one, denoted λ_1, represents the relaxation time scale (the memory) of the fluid, while the second convection/advection time scale represents the time needed by the fluid to pass the distance L at the velocity U.

$$\mathcal{We} = \frac{\lambda_1 U}{L} \quad . \tag{16}$$

The dimensionless contribution of the polymer viscosity μ_e is defined by

$$\frac{\mu_e}{\mu} = \frac{\mu - \mu_s}{\mu} = 1 - \frac{\mu_s}{\mu} = \eta \in [0, 1] \; . \tag{17}$$

Scaling the velocity vector field, the pressure, and the extra stress tensor of the fluid by

$$\boldsymbol{u} = \frac{\tilde{\boldsymbol{u}}}{U} \quad , \quad p = \frac{\tilde{p}L}{\mu U} \quad , \quad \boldsymbol{\tau} = \frac{\tilde{\boldsymbol{\tau}}L}{\mu U} \quad , \tag{18}$$

and taking into account the definitions (15), (16), and (17), the dimensionless momentum equations can be written as

$$\mathcal{R}e\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = 2(1-\eta)\nabla\cdot\mathbf{D} + \nabla\cdot\boldsymbol{\tau} - \nabla p \quad , \tag{19}$$

and the dimensionless constitutive equation takes the form

$$\boldsymbol{\tau} + \mathcal{W}e\left(\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \nabla\boldsymbol{u}\boldsymbol{\tau} - \boldsymbol{\tau}\,(\nabla\boldsymbol{u})^T\right) = 2\eta\mathbf{D} \quad . \tag{20}$$

It has been proved that Eqs. (19), (20), and (8) are stable in the sense of Hadamard [25]. The complete dimensionless system for the Oldroyd-B model is written below for future reference.

$$\left\{\begin{array}{l} \nabla\cdot\boldsymbol{u} = 0 \quad , \\ \mathcal{R}e\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) + \nabla p = 2(1-\eta)\nabla\cdot\mathbf{D} + \nabla\cdot\boldsymbol{\tau} \quad , \\ \boldsymbol{\tau} + \mathcal{W}e\left(\dfrac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau} - \nabla\boldsymbol{u}^T\boldsymbol{\tau} - \boldsymbol{\tau}\nabla\boldsymbol{u}\right) = 2\eta\mathbf{D} \quad . \end{array}\right. \tag{21}$$

This stress tensor splitting allows to decouple the kinematics and non-Newtonian viscoelastic stress even though the divergence of $\boldsymbol{\tau}$ is included in the momentum equation as a pseudo-body force and the constitutive equation contains a contribution from the Newtonian part.

This governing system should be supplemented by appropriate initial and boundary conditions.

2.4 Boundary Conditions

In practical applications the flow in bounded domain during limited time is of interest. Therefore some initial and boundary conditions should be specified. The mathematical need for boundary conditions hinges on their role in proving well posedness of the problem, i.e., showing that there exists a unique solution that depends continuously on the initial data.

For the considered problem of flow in a closed channel (with tube-like geometry), the bounded domain Ω has boundary $\partial\Omega = \Gamma_{in} \cup \Gamma_w \cup \Gamma_{out}$. The following set of conditions can be used:

- The *inlet* boundary—the Dirichlet boundary conditions are imposed for both the velocity and extra stress, assuming, for example, that the analytical solution is known from a Poiseuille-like flow for $\boldsymbol{u}$ and $\boldsymbol{\tau}$.

$$\boldsymbol{u} = u_{in} \text{ on } \Gamma_{in} \ ,$$

$$\boldsymbol{\tau} = \boldsymbol{\tau}_{in} \text{ on } \Gamma_{in} \ .$$

- The (rigid) *walls* of the domain—the no-slip boundary conditions are imposed on the fluid velocity $\boldsymbol{u}$, i.e.,

$$\boldsymbol{u} = 0 \text{ on } \Gamma_w \ .$$

- The *outlet* boundary—the homogeneous Neumann boundary conditions are used for $\boldsymbol{u}$, i.e.,

$$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma_{out} \ ,$$

where $\boldsymbol{n}$ is the outward unit normal vector to the Γ_{out}.

2.5 Variational Formulation

Considering a bounded domain Ω, whose boundary is $\partial\Omega = \Gamma_{in} \cup \Gamma_w \cup \Gamma_{out}$, the dimensionless strong formulation of the Oldroyd-B fluid flow problem (without additional body forces) can be written as

$$\begin{cases} \nabla \cdot \boldsymbol{u} = 0 \ , & \text{in } \Omega \\ \mathcal{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u}\right) + \nabla p = 2(1-\eta)\nabla \cdot \mathbf{D} + \nabla \cdot \boldsymbol{\tau} \ , & \text{in } \Omega \\ \boldsymbol{\tau} + \mathcal{We}\left(\dfrac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\tau} - \nabla \boldsymbol{u}^T \boldsymbol{\tau} - \boldsymbol{\tau} \nabla \boldsymbol{u}\right) = 2\eta \mathbf{D} \ , & \text{in } \Omega \\ \boldsymbol{u} = 0 \ , & \text{on } \Gamma_w \\ \boldsymbol{u} = u_{in}, & \text{on } \Gamma_{in} \\ \boldsymbol{\tau} = \boldsymbol{\tau}_{in} \ , & \text{on } \Gamma_{in} \\ \boldsymbol{u}|_{t=0} = u_0 \quad . & \end{cases} \tag{22}$$

The function spaces for bi-dimensional case ($\Omega \subset \mathbb{R}^d$) are chosen in the following way:

$$\mathcal{V} = \left\{ \boldsymbol{u} \in \mathbf{H}^1(\Omega) : \boldsymbol{u} = 0 \text{ on } \Gamma_w \text{ and } \boldsymbol{u} = u_{in} \text{ on } \Gamma_{in} \right\} \quad , \tag{23}$$

$$L_0^2(\Omega) = \left\{ p \in L^2(\Omega) : \int_\Omega p \; d\Omega = 0 \right\} \quad , \tag{24}$$

$$\mathcal{S} = \left\{ \boldsymbol{S} \in [L^2(\Omega)]^{d \times d} : \boldsymbol{S}^T = \boldsymbol{S} \right\} \quad . \tag{25}$$

The variational formulation or the weak problem corresponding to (22), given $u_0 \in \mathcal{V}$ such that $\nabla \cdot u_0 = 0$, is:

Find $(\boldsymbol{u}, p, \boldsymbol{\tau}) \in L^2(0, T; \mathcal{V}) \times L^2\left(0, T; L_0^2(\Omega)\right) \times L^2(0, T; \mathcal{S})$ such that

$$\begin{cases} \displaystyle\int_\Omega (\nabla \cdot \boldsymbol{u})\, q = 0\,, \\ \displaystyle\int_\Omega 2(1-\eta)\mathbf{D} : \nabla \boldsymbol{v} + \mathcal{R}e \int_\Omega \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} \right) \cdot \boldsymbol{v} - \int_\Omega p \nabla \cdot \boldsymbol{v} = - \int_\Omega \boldsymbol{\tau} : \nabla \boldsymbol{v}, \\ \displaystyle\int_\Omega \left[\boldsymbol{\tau} + \mathcal{W}e \left(\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\tau} \right) \right] : \boldsymbol{S} = \int_\Omega \left[2\eta \mathbf{D} + \mathcal{W}e \left(\nabla \boldsymbol{u}^T \boldsymbol{\tau} + \boldsymbol{\tau} \nabla \boldsymbol{u} \right) \right] : \boldsymbol{S}. \end{cases} \tag{26}$$

holds , $\forall \boldsymbol{v} \in \mathbf{H}_0^1(\Omega)$, $\forall q \in L_0^2(\Omega)$ and $\forall \boldsymbol{S} \in \mathcal{S}$.

3 Numerical Approximation

This section describes the details of finite element discretization of the governing system. The focus is on obtaining a steady solution by a time-marching algorithm as a limit for $t \to \infty$ of the unsteady system solved with stationary boundary conditions. The effects of various artificial stress diffusion techniques are evaluated for a two-dimensional case of flow in a symmetric, smoothly corrugated channel.

The discrete model is obtained using the Rothe method in which the temporal and spatial discretizations are performed separately. First, the discretization in time (of the material time derivative) is realized by characteristic Galerkin method associated with the implicit Euler method. The spatial discretization is then based on a finite element approximation of the variational formulation.

3.1 Discretization in Time: Convective Term

The Characteristic Galerkin Method [31] evaluates time derivative[3] of a field ϕ (scalar, vector, or tensor) on Lagrangian frame using characteristic lines (trajectories) of material particles driven at velocity $\boldsymbol{u}$. For the instant t and the interval of time Δt, it requires that

$$\phi(t+\Delta t, \boldsymbol{x}) = \phi(t, \boldsymbol{x} - \boldsymbol{u}\Delta t) \quad .$$

In a time interval $[0, T_f]$ is defined a set of points $t^n = n\Delta t = n\frac{T_f}{N}, n = 0, \ldots, N$.

Denote ϕ^n the approximation of function ϕ at the instant of time $t^n = n\Delta t$, i.e., $\phi^n \approx \phi(t^n, \boldsymbol{x}),\ t^n = n\Delta t, \boldsymbol{x} \in \Omega$.

The material derivative of ϕ which represents the rate of change of ϕ along the trajectory is approximated by the backward Euler scheme as

$$\begin{aligned}\frac{D\phi}{Dt}(t^{n+1}, \boldsymbol{x}) = \frac{\partial \phi}{\partial t}(t^{n+1}, \boldsymbol{x}) + \boldsymbol{u}(t^{n+1}, \boldsymbol{x}) \cdot \nabla\phi(t^{n+1}, \boldsymbol{x}) \approx \\ \approx \frac{\phi(t^{n+1}, \boldsymbol{x}) - \phi(t^n, \boldsymbol{x} - \boldsymbol{u}(t^n, \boldsymbol{x})\Delta t)}{\Delta t} = \\ = \frac{\phi(t^n, \boldsymbol{x}) - \phi(t^{n-1}, \boldsymbol{x}_\star)}{\Delta t} \quad , \end{aligned} \tag{27}$$

where $\boldsymbol{x}_\star = \boldsymbol{x} - \boldsymbol{u}(t^n, \boldsymbol{x})\Delta t$ is the position at time t^{n-1} of the particle located at $\boldsymbol{x}$ at time t^n.

Figure 1 shows the characteristic path ξ in time and space of a point of the fluid that is at the position $\boldsymbol{x}_i$ of the domain's grid at the instant of time t^n, which was at

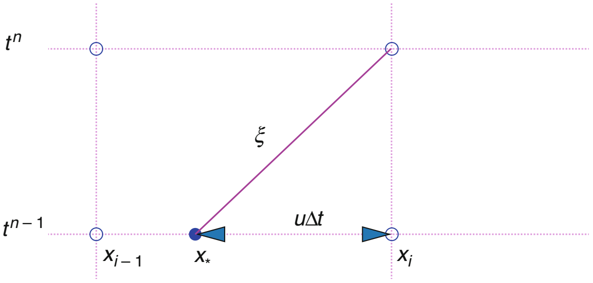


Fig. 1 Simplified scheme of the advection characteristic

[3] The finite-difference approximation (in time) used in this method [5] is quite popular among researchers, because it allows to achieve second-order of accuracy in time (at least in the case of uniform velocity field) and only requires one level of memory storage for the values from previous time step. In more general case of a non-uniform velocity field or multidimensional flows, the scheme is considered only first-order accurate in time.

the node $\boldsymbol{x}_\star = \boldsymbol{x}_i - \boldsymbol{u}(t^n, \boldsymbol{x}_i)\Delta t$ at the previous time instant t^{n-1}. Since at instant t^{n-1} only the values of ϕ are known in the grid nodes $\boldsymbol{x}_{i-1}$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$, the value of $\phi(t^{n-1}, \boldsymbol{x}_\star)$ is evaluated by interpolation and set $\phi(t^{n-1}, \boldsymbol{x}_\star) = \phi(t^n, \boldsymbol{x}_i)$.

Taking into account (27), the semi-discretized Oldroyd-B problem is defined $\forall \boldsymbol{v} \in \mathbf{H}_0^1(\Omega)$, $\forall q \in L_0^2(\Omega)$ and $\forall \boldsymbol{S} \in \mathcal{S}$ by

$$\begin{cases} \boldsymbol{u}^0 = u_0 \, , \\ \displaystyle\int_\Omega \left(\nabla \cdot \boldsymbol{u}^n\right) q = 0 \, , \\ \displaystyle\int_\Omega 2(1-\eta)\mathbf{D}^n : \nabla \boldsymbol{v} + \mathcal{Re}\int_\Omega \frac{\boldsymbol{u}^n - \boldsymbol{u}_\star^{n-1}}{\Delta t} \cdot \boldsymbol{v} - \int_\Omega p^n \nabla \cdot \boldsymbol{v} = -\int_\Omega \boldsymbol{\tau} : \nabla \boldsymbol{v} \, , \\ \displaystyle\int_\Omega \left(\boldsymbol{\tau}^n + \mathcal{We}\frac{\boldsymbol{\tau}^n - \boldsymbol{\tau}_\star^{n-1}}{\Delta t}\right) : \boldsymbol{S} = \int_\Omega \left[2\eta\mathbf{D} + \mathcal{We}\left(\nabla \boldsymbol{u}^T \cdot \boldsymbol{\tau}^n + \boldsymbol{\tau}^n \cdot \nabla \boldsymbol{u}\right)\right] : \boldsymbol{S} \, . \end{cases} \tag{28}$$

3.2 Discretization in Space

The Finite Element Method (FEM) was adopted to discretize the considered problem (28) in space. The domain Ω was decomposed into finite number $N_\mathcal{T}$ of triangles $\mathcal{T}$ whose union constitute a non-degenerated mesh $\mathbb{T}_h$, meaning that:

- The interior of each $\mathcal{T}_i$ is non-empty ($\mathcal{T}_i^\circ \neq \emptyset, \; i = 1, \dots N_\mathcal{T}$).
- The interior of two distinct triangles are disjoints ($\mathcal{T}_i^\circ \cap \mathcal{T}_j^\circ = \emptyset, \; i \neq j, \; i, j = 0, \dots, N_\mathcal{T}$) .
- Every boundary of $\mathcal{T}_i, i = 1, \dots N_\mathcal{T}$ is a boundary of another triangle (the triangles are adjacent or part of boundary $\partial\mathbb{T}_h$).
- $\Omega = \bigcup\limits_{i=1}^{N_\mathcal{T}} \mathcal{T}_i = \overline{\mathbb{T}}_h$.

The parameter $h = \max\limits_{\mathcal{T} \in \mathbb{T}_h} h_\mathcal{T}$ defines the diameter of the triangulation $\mathbb{T}_h$, where $h_\mathcal{T}$ is the diameter of the circumscribed circle into $\mathcal{T}$. The mesh $\mathbb{T}_h$ is a uniform regular mesh, where all the triangles have approximately the same size. This means that there exist positive constants C_1, C_2 independent of h and $\mathcal{T}$ such that:

- $C_1 h \leq h_\mathcal{T}, \quad \forall \mathcal{T} \in \mathbb{T}_h$,
- $\dfrac{h_\mathcal{T}}{\rho_\mathcal{T}} \leq C_2, \quad \forall \mathcal{T} \in \mathbb{T}_h$,

where $\rho_\mathcal{T}$ is the diameter of the inscribed circle into $\mathcal{T}$.

The discretization elements are chosen to guarantee the compatibility condition known as the discrete LBB (Ladyzheskaya, Babuška and Brezzi) or inf-sup condition, which requires that there exists $\gamma > 0$ (independent of h) such that

$$\inf_{q_h \in \mathcal{L}_h \backslash \{0\}} \sup_{\boldsymbol{v}_h \in \mathbf{m} X_h \backslash \{0\}} \frac{|(q_h, \nabla \cdot \boldsymbol{v}_h)|}{\|\boldsymbol{v}_h\|_{X_h} \, \|q_h\|_{\mathcal{L}_h}} \geq \gamma \quad .$$

where the finite dimensional function spaces are defined by

$$\boldsymbol{X}_h = \left\{ \boldsymbol{v}_h \in C(\overline{\Omega}) \cap \mathcal{V} : \boldsymbol{v}_h|_{\mathcal{T}} \in \mathbb{P}_2(\mathcal{T}), \forall \mathcal{T} \in \mathbb{T}_h \right\}, \tag{29}$$

$$\mathcal{L}_h = \left\{ q_h \in C(\overline{\Omega}) \cap L^2_0(\Omega) : q_h|_{\mathcal{T}} \in \mathbb{P}_1(\mathcal{T}), \forall \mathcal{T} \in \mathbb{T}_h \right\}, \tag{30}$$

being the $\mathbb{P}_n$ be the space of polynomials of degree $n > 0$ defined on triangles.

The momentum equations are discretized with the mixed finite element known as Hood-Taylor elements $P_2 - P_1$ which are associated to the approximation of saddle point problems. The constitutive equation for extra stress tensor is also discretized by quadratics finite elements[4]. Defining the finite dimensional function space

$$\mathcal{S}_h = \left\{ \boldsymbol{S}_h \in \mathbf{C}(\overline{\Omega}) \cap \mathcal{S} : \boldsymbol{S}_{h,ij}|_{\mathcal{T}} \in \mathbb{P}_2(\mathcal{T}), \forall \mathcal{T} \in \mathbb{T}_h \right\}, \tag{31}$$

the approximate finite element problem based on (28) can be written for each $t \in [0, T_f]$, $h > 0$, $\boldsymbol{u}^0_h \in \boldsymbol{X}_h$ and $\boldsymbol{\tau}^0_h \in \mathcal{S}_h$, leading to:

Find $(\boldsymbol{u}_h, p_h, \boldsymbol{\tau}_h) \equiv (\boldsymbol{u}_h(t, \cdot), p_h(t, \cdot), \boldsymbol{\tau}_h(t, \cdot)) \in \boldsymbol{X}_h \times \mathcal{L}_h \times \mathcal{S}_h$ such that

$$\begin{cases} \displaystyle\int_\Omega \left(\nabla \cdot \boldsymbol{u}^n_h\right) q_h = 0, \\ \displaystyle\int_\Omega 2(1-\eta)\mathbf{D}^n_h : \nabla \boldsymbol{v}_h + \mathcal{R}e \int_\Omega \frac{\boldsymbol{u}^n_h - \boldsymbol{u}^{n-1}_{\star h}}{\Delta t} \cdot \boldsymbol{v}_h - \int_\Omega p^n_h \nabla \cdot \boldsymbol{v}_h = - \int_\Omega \boldsymbol{\tau}_h : \nabla \boldsymbol{v}_h, \\ \displaystyle\int_\Omega \left(\boldsymbol{\tau}^n_h + \mathcal{W}e \frac{\boldsymbol{\tau}^n_h - \boldsymbol{\tau}^{n-1}_{\star h}}{\Delta t} \right) : \boldsymbol{S}_h = \int_\Omega \left[2\eta \mathbf{D}_h + \mathcal{W}e \left(\nabla \boldsymbol{u}^T_h \cdot \boldsymbol{\tau}^n_h + \boldsymbol{\tau}^n_h \cdot \nabla \boldsymbol{u}_h \right) \right] : \boldsymbol{S}_h, \end{cases} \tag{32}$$

holds for all $(\boldsymbol{v}_h, q_h, \boldsymbol{S}_h) \in \boldsymbol{X}_h \times \mathcal{L}_h \times \mathcal{S}_h$.

Further details about the properties of the finite element method and about the rigorous convergence analysis of spatial discretization of the Navier-Stokes problem can be found in [9].

As Eq. $(32)_3$ is verified $\forall \boldsymbol{S}_h \in \mathcal{S}_h$, it is also verified for the symmetric tensor $\boldsymbol{M}_h \in \mathcal{S}_h$ such that for some fixed $i, j = 1, 2$ the corresponding component $\boldsymbol{M}_{h,ij}$

[4] There is no general mathematical theory to guarantee the stability of finite element method for a given choice of element type, in the case of viscoelastic problems.

belongs to the space

$$\mathcal{M}_h = \left\{ M_h \in C(\overline{\Omega}) \cap L^2(\Omega) : M_h|_{\mathcal{T}} \in \mathbb{P}_2(\mathcal{T}), \forall \mathcal{T} \in \mathbb{T}_h \right\}. \tag{33}$$

and the other components of the tensor are null. Hence, the tensorial equation (32)$_3$ can be decoupled into the set of scalar equations. This means that the tensorial equation (32)$_3$ can be replaced in the problem (32) by the system of scalar equations for tensor components:

$$\begin{cases} \displaystyle\int_\Omega \left(\boldsymbol{\tau}^n_{h,11} + \mathcal{W}e \frac{\boldsymbol{\tau}^n_{h,11} - \boldsymbol{\tau}^{n-1}_{\star h,11}}{\Delta t} \right) : M_h = \\ \qquad \displaystyle = \int_\Omega \left[2\eta \frac{\partial u_1}{\partial x_1} + 2\mathcal{W}e \left(\frac{\partial u_1}{\partial x_1} \boldsymbol{\tau}^n_{h,11} + \frac{\partial u_1}{\partial x_2} \boldsymbol{\tau}^{n-1}_{h,12} \right) \right] : M_h \ , \\ \displaystyle\int_\Omega \left(\boldsymbol{\tau}^n_{h,12} + \mathcal{W}e \frac{\boldsymbol{\tau}^n_{h,12} - \boldsymbol{\tau}^{n-1}_{\star h,12}}{\Delta t} \right) : M_h = \\ \qquad \displaystyle = \int_\Omega \left[\eta \left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} \right) + \mathcal{W}e \left(\frac{\partial u_2}{\partial x_1} \boldsymbol{\tau}^n_{h,11} + \frac{\partial u_1}{\partial x_2} \boldsymbol{\tau}^{n-1}_{h,22} \right) \right] : M_h \ , \\ \displaystyle\int_\Omega \left(\boldsymbol{\tau}^n_{h,22} + \mathcal{W}e \frac{\boldsymbol{\tau}^n_{h,22} - \boldsymbol{\tau}^{n-1}_{\star h,22}}{\Delta t} \right) : M_h = \\ \qquad \displaystyle = \int_\Omega \left[2\eta \frac{\partial u_2}{\partial x_2} + 2\mathcal{W}e \left(\frac{\partial u_2}{\partial x_1} \boldsymbol{\tau}^n_{h,12} + \frac{\partial u_2}{\partial x_2} \boldsymbol{\tau}^n_{h,22} \right) \right] : M_h \ . \end{cases} \tag{34}$$

4 Artificial Stress Diffusion Implementation

The four different types of artificial stress diffusion terms $\boldsymbol{E}$ introduced in Sect. 1.3 share a common form involving the Laplacian of the extra stress $\boldsymbol{\tau}$. Therefore also most of the details of their implementation in the finite-element framework are shared. All cases lead to the following modified constitutive relation to be solved:

$$\boldsymbol{\tau} + \mathcal{W}e \left(\frac{\partial \boldsymbol{\tau}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{\tau} - \nabla \boldsymbol{u}^T \cdot \boldsymbol{\tau} - \boldsymbol{\tau} \cdot \nabla \boldsymbol{u} \right) = 2\eta \mathbf{D} + \alpha(\cdot) \Delta \boldsymbol{\sigma} \ , \tag{35}$$

where the artificial diffusion coefficient α can either be constant or variable as $\alpha(\cdot)$, and the tensor $\boldsymbol{\sigma}$ either corresponds to the elastic stress tensor, i.e., $\boldsymbol{\sigma} = \boldsymbol{\tau}$ or to the time derivative (time-difference, steady residual) of that tensor, i.e., $\boldsymbol{\sigma} = \boldsymbol{\tau}_t$. So, the weak formulation of (35) is defined $\forall \boldsymbol{S} \in \mathcal{S}$ by

$$
\begin{aligned}
&\int_{\Omega}\left[\boldsymbol{\tau}+\mathcal{W}e\left(\frac{\partial\boldsymbol{\tau}}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{\tau}\right)\right]:\boldsymbol{S}= \\
&\qquad=\int_{\Omega}\left[2\eta\mathbf{D}+\mathcal{W}e\left(\nabla\boldsymbol{u}^{T}\cdot\boldsymbol{\tau}+\boldsymbol{\tau}\cdot\nabla\boldsymbol{u}\right)\right]:\boldsymbol{S}-\alpha(\cdot)\nabla\boldsymbol{\sigma}:\nabla\boldsymbol{S},
\end{aligned}
\tag{36}
$$

where $\mathcal{S}=\left\{\boldsymbol{S}\in[L^{2}(\Omega)]^{d\times d}:\boldsymbol{S}^{T}=\boldsymbol{S}\right\}$.

When a variable artificial diffusion coefficient is used, the following types of functional dependencies can be (were) used:

- *Time-dependent function* $\alpha(t)$—Diffusion coefficient α is monotonically decaying with (pseudo) time t and vanishing for the limit case of $t \longrightarrow \infty$, where the steady-state solution should be reached. Different shapes of such monotonically decaying functions were considered and tested, searching for optimal initial values of the diffusion coefficient and suitable decay rate. The function $\alpha(t)$ used in the presented simulation has the form

$$
\alpha(t)=\alpha_0\cdot\frac{1}{1+\varepsilon\cdot t}\quad, \tag{37}
$$

  where $\alpha_0=\alpha(0)$ is the initial value for the diffusion parameter α, while the adjustable parameter ε affects the rate of decay of the function (by scaling the time variable).
- *Time-derivative dependent function* $\alpha(\phi_t)$—Diffusion coefficient α is made proportional to (dependent on) the time-derivative of some solved quantity ϕ. The dependence is such that the diffusion coefficient $\alpha \longrightarrow 0$ for the steady-state solution where $\phi_t \longrightarrow 0$. Proposed and tested were different functional dependencies of α on the time derivative ϕ_t as well as different choices of the indicator variable ϕ for evaluation of the time derivative (e.g., pressure p, tensor components τ_{ij}, or tensor norm $\|\boldsymbol{\tau}\|$). The general form of the functional dependency of $\alpha(\phi_t)$ is the following:

$$
\alpha(\phi_t)=\alpha_0\cdot\frac{\varepsilon}{\varepsilon+(1-\varepsilon)\cdot(\|\phi_t\|)^{m}}\quad, \tag{38}
$$

  where again $\alpha_0=\alpha(0)$ is the initial value for the diffusion parameter α, and the adjustable parameters ε and m affect the rate of decay of the function.

These functional dependencies of $\alpha(\cdot)$ for *time-dependent* and *time-derivative-dependent* diffusion coefficients were extensively tested and discussed in [28], being compared to each other, to the *constant* diffusion term and also to the original non-diffusive numerical method without any added artificial diffusion.

5 Numerical Results

The model was implemented in FreeFem++, which is a finite element solver and simulation software for the solution of partial differential equations (see [12] for more details about FreeFem++).

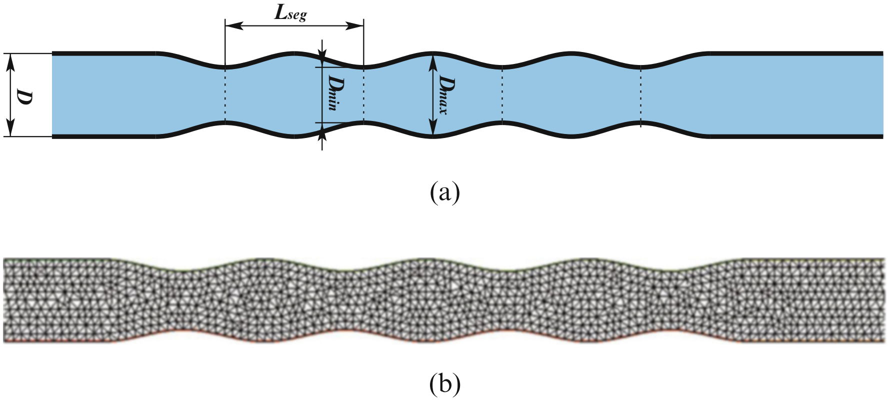
Numerical simulations were performed in a domain having the shape of a 2D corrugated channel consisting of three smoothly connected identical sinusoidally shaped segments. The straight inlet and outlet parts of the channel are sufficiently long to guarantee a fully developed Poiseuille flow upstream and downstream from the corrugated part.

The computational grid consisting of triangular finite elements was generated using the FreeFem++ [12] by Delaunay-Voronoi algorithm, considering 10 elements along each unit of length of wall, without any special treatment to symmetrize the grid with respect to x axis.

Figure 2 shows some details about the computational domain, already used in our previous publications [26–29], where many of the numerical simulations and their results were already presented and discussed. Here the focus is on some additional results supporting the observations regarding the applicability and efficiency of individual stress diffusion terms. All tests illustrated here were obtained with a fixed Reynolds number $\mathcal{Re} = 1000$.

5.1 Constant Diffusion Coefficient

This section, and, namely, Fig. 3, shows the effects of the diffusion term based on Laplacian of the extra stress tensor when the artificial diffusion coefficient



(a)

(b)

Fig. 2 Geometrical configuration of the test case—corrugated tube (2D channel). (**a**) Geometry sketch of the channel. (**b**) Grid for finite element approximation

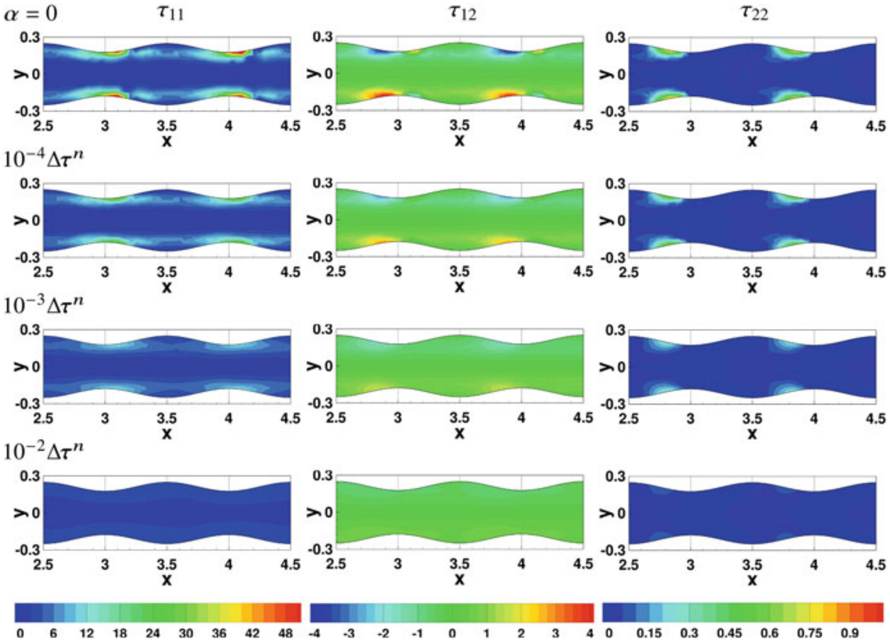


Fig. 3 Components of the elastic stress tensor $\boldsymbol{\tau}$, obtained using the stress diffusion (2) with different values of α

$\alpha = const$ is used (i.e., the diffusion term in the form (2). It should be noted that the solution obtained with setting $\alpha = 0$ corresponds to the original non-diffusive system, so it might be considered as the reference solution which we should obtain.

The case or relatively low Weissenberg number $\mathcal{We} = 0.4$ was chosen here to allow for mutual comparison of the simulation results obtained with and without the added stabilization term. Evidently such comparison will be not possible in case of higher $\mathcal{We}$, when the stabilization can't be removed (switched off) to obtain numerical solution.

In this series of simulations the results were obtained for $\alpha \in \left\{10^{-4}, 10^{-3}, 10^{-2}\right\}$ in the stabilization term $\alpha \cdot \Delta\boldsymbol{\tau}^n$. The contours of the extra stress tensor $\boldsymbol{\tau}$ components shown in Fig. 3 clearly demonstrate the smoothing effect of this kind of added diffusive term. Even for the smallest chosen value of $\alpha = 10^{-4}$, the results differ visibly from the reference (non-diffusive) solution obtained for $\alpha = 0$. With increasing the value of α, the problem of the solution (over-)smoothing only gets more apparent.

These test with the standard artificial diffusion term with constant coefficient α confirm that the use of such added term may lead to unacceptable errors on the solution which no more corresponds to the original non-diffusive Oldroyd-B problem. Therefore the artificial diffusion coefficient α should be kept as small as possible. Evidently no diffusion is needed for low $\mathcal{We}$, but with increasing values of $\mathcal{We}$ the need for stabilization grows.

This suggests a possibility to use the artificial diffusion coefficient that will depend on the Weissenberg number, i.e., $\alpha = \alpha(\mathcal{We})$, while still being constant in space and time. This will allow to almost automatically adjust the level of artificial diffusion in dependence on its anticipated need. This possibility was explained and tested in [29] using the expression $\alpha(\mathcal{We}) \propto \; h^2 \cdot \mathrm{atan}(\varepsilon \mathcal{We})$ which increases monotonically in dependence on $\mathcal{We}$, reaching some finite asymptotic value α_∞ for high Weissenberg numbers $\mathcal{We} \to \infty$. Use of such case dependent ($\mathcal{We}$ dependent) artificial diffusion coefficient makes the whole method much more safe to use, avoiding to large extent the risk of over-smoothing the solution.

5.2 Time-Dependent Diffusion Coefficient

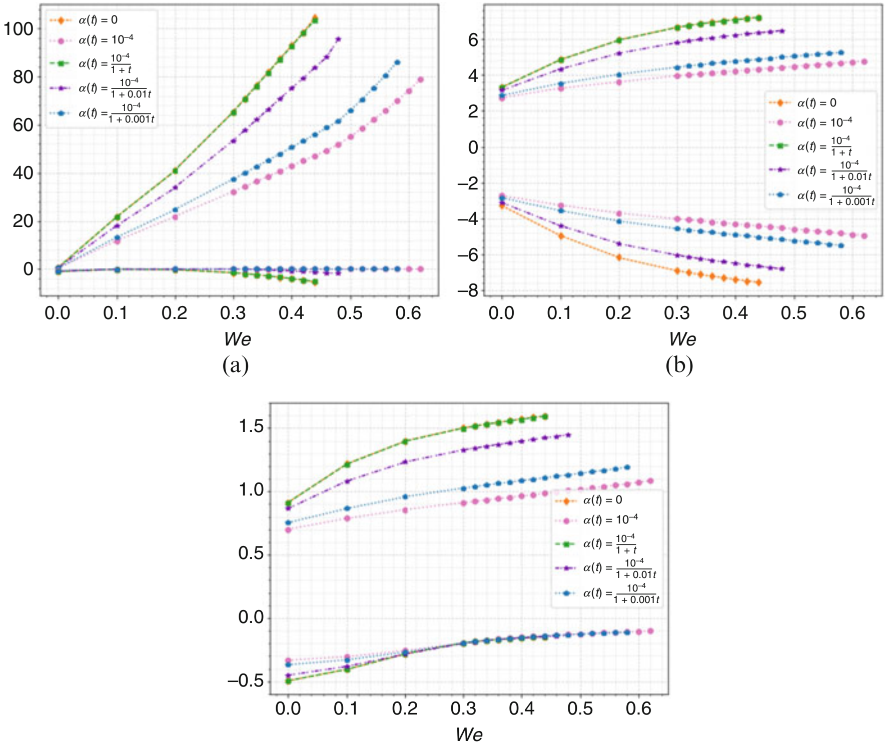
In this case the artificial stress diffusion has the form (3), i.e., the diffusion coefficient α depends on (iterative) time in such a way that $\alpha(t)$ monotonically approaches to zero as $t \to \infty$. The dependence for $\alpha(t)$ is described by the relation (37), where the parameter ε can be chosen to allow for adjustment of the decay of the function $\alpha(t)$.

In this setup it makes not much sense to show the final converged solutions for different values of ε, because in ideal case $\alpha \to 0$ and all solutions will be virtually identical. To demonstrate the behavior of this kind of variable (decaying) in time diffusion coefficient $\alpha(t)$ the results are presented after fixed number of time steps, allowing to compare the effects of the stabilization term before it vanishes. Moreover in this case it was pushed to the highest attainable Weissenberg number, i.e., for each artificial diffusion setting the critical $\mathcal{We}$ was found experimentally.

The results are in this case presented in the form of graphs of extremal values (minima and maxima) of the elastic stress tensor $\boldsymbol{\tau}$ components, depending on the Weissenberg number. In Fig. 4 the individual curves correspond to different settings of the decay parameter ε in (37). The reference (non-diffusive) solution corresponds to $\alpha(t) = 0$. In the presented graphs the vertical discrepancies with respect to the reference solution curve correspond to smoothing effects resulting in the cut-off of the solution local extrema. The horizontal extent of each curve corresponds to the maximum attainable Weissenberg number $\mathcal{We}$ for chosen setting decay of $\alpha(t)$.

From Fig. 4 it is evident that the fast decay of $\alpha(t)$ allows to recover the non-diffusive reference solution (after finite number of pseudo-time steps). On the other hand, slow decay of $\alpha(t)$ results into non-negligible value of α at the time the simulation was stopped, which leads to more stable numerical method (with higher critical $\mathcal{We}$), but at the price of higher effective smoothing, leading to possibly too diffusive solutions that do not correspond to the original (non-diffusive) Oldroyd-B problem.

Similar observations can be made in the graphs of the extrema of the stress tension (see Fig. 5) along the channel wall. Also here the stability of the method due to added diffusion seems to come at the price of non-negligible smoothing of the solution. The underestimation (due to excessive smoothing) of tension extrema



(a)

(b)

(c)

Fig. 4 Maximum and minimum values of the elastic stress tensor $\boldsymbol{\tau}$ components, for different settings of the parameter ε depending on the Weissenberg number $\mathcal{W}e$. (**a**) Component τ_{11}. (**b**) Component τ_{12}. (**c**) Component τ_{22}

on the wall may have important consequences in practical use of artificial diffusion methods in engineering or biomedicine.

5.3 Time-Derivative-Dependent Diffusion Coefficient

The choice of time-derivative dependent diffusion coefficient attempts to avoid the need to adjust the decay of $\alpha(t)$ based on some a-priori estimate of the number of iterations needed to reach the steady-state solution. Here at the same time the (pseudo) time difference of solution approximations $(\phi^n - \phi^{n-1}) \propto \phi_t$ is used to adjust the artificial diffusion coefficient $\alpha(\phi_t)$ and also to stop the time-marching simulation when $(\phi^n - \phi^{n-1}) \to 0$, i.e., stopping the simulation when $\|\phi^n - \phi^{n-1}\| < tol$.

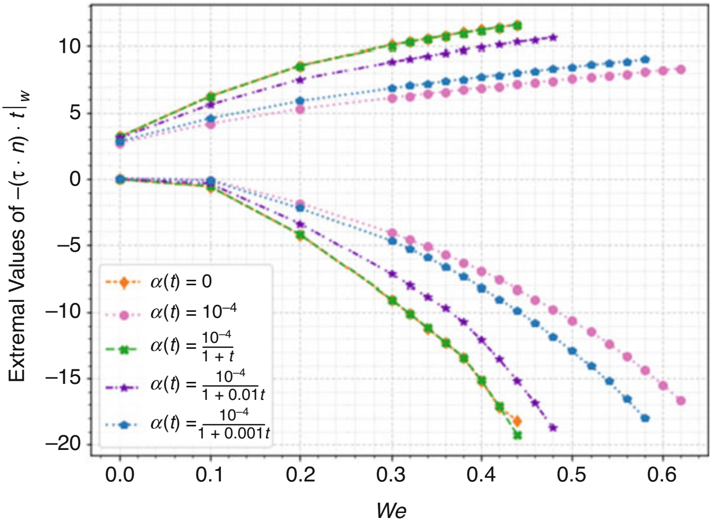


Fig. 5 Extremal values of the stress tension on the wall $-(\tau \cdot n) \cdot t|_w$ along the channel wall for different values of parameter ε in the diffusion coefficient $\alpha(t) = \frac{10^{-4}}{1+\varepsilon t}$, depending of the Weissenberg number

The choice of the artificial diffusion coefficient being solution-dependent brings an extra non-linearity into the extended (stabilized, diffusive) problem. The parameter α is reset every iteration, which may in the worst case lead to de-stabilization of the solution. In order to face this problem (avoid rapid in time changes of α), the dependence (38) can be modified to take into account some recent history of the coefficient α and use some kind of floating average value rather than the actual one just based on the latest iteration. The modified formula has the form

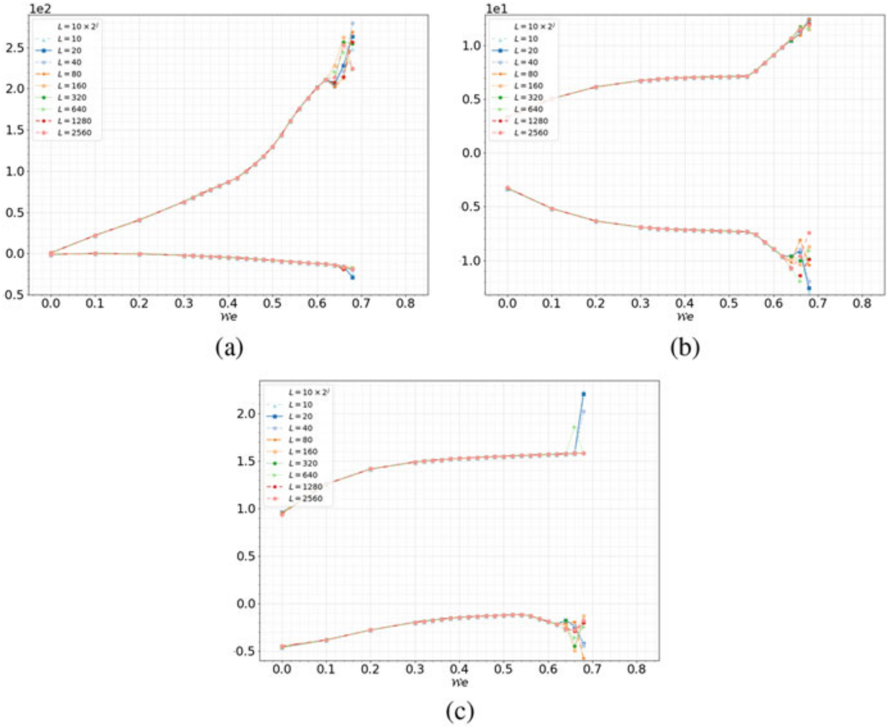
$$\alpha(\phi_t) = \frac{1}{L} \sum_{k=n-L}^{n} \alpha(\phi_t^k) = \frac{\alpha_0}{L} \sum_{k=n-L}^{n} \frac{\varepsilon}{\varepsilon + (1-\varepsilon) \cdot \left(\|\phi_t^k\|\right)^m} \quad , \tag{39}$$

where L corresponds to the number of previous time steps considered to evaluate the new diffusion coefficient α for the current iteration.

The numerical tests based on the formula (39) were performed for different averaging lengths L, using either the elastic stress tensor $\boldsymbol{\tau}$ or pressure p to govern the diffusion coefficient $\alpha = \alpha(\phi_t)$, i.e., either $\alpha \propto \|\boldsymbol{\tau}_t\|$ or $\alpha \propto \|p_t\|$.

When using the tensor-dependent diffusion coefficient $\alpha = \alpha(\boldsymbol{\tau}_t)$, the results shown in the Fig. 6 seem to be quite insensitive to the averaging length L for whole range of Weissenberg numbers $\mathcal{W}e$, which is also confirmed in the graphs of tension on the wall shown on Fig. 7.

Another technically relevant output characterizing globally the flow in the considered channel is the pressure drop between the inlet and outlet boundary. This



(a)

(b)

(c)

Fig. 6 Maximum and minimum values of the elastic stress tensor $\boldsymbol{\tau}$ components, for different lengths of the averaging history L with α defined by (39) and $\phi = \boldsymbol{\tau}$ depending on the Weissenberg number $\mathcal{W}e$. (**a**) Component τ_{11}. (**b**) Component τ_{12}. (**c**) Component τ_{22}

quantity is plotted for all solved cased in Fig. 8 showing again the robustness of the tensor driven diffusion coefficient $\alpha(\boldsymbol{\tau}_t)$ with respect to the choice of averaging length L.

The situation changes dramatically when trying to use $\phi = p$, i.e., pressure driven diffusion parameter $\alpha = \alpha(p_t)$. Figure 9 shows that for higher Weissenberg numbers the results are quite sensitive to the choice of the averaging length L. For shorter L the results are rather randomly affected and inconsistent, while for large L the high diffusive coefficient is preserved which leads to over-smoothed results with underestimated extremal values.

The quick look at the tension on the wall (Fig. 10) and pressure drop (Fig. 11) just confirm the fact that the choice of $\alpha = \alpha(p_t)$ is not suitable for practical use, leading to large variability of the results depending on L.

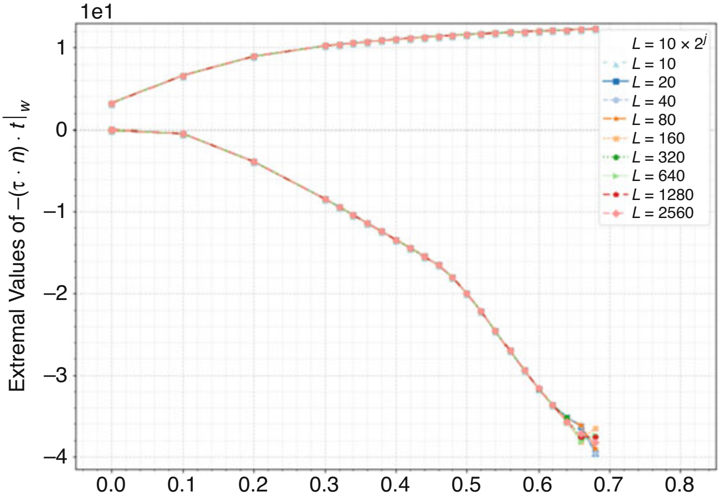


Fig. 7 Extremes values of the stress tension on the wall $-(\tau \cdot n) \cdot t|_w$ along the channel wall for different values of lengths of the averaging history L, depending on the Weissenberg numbers

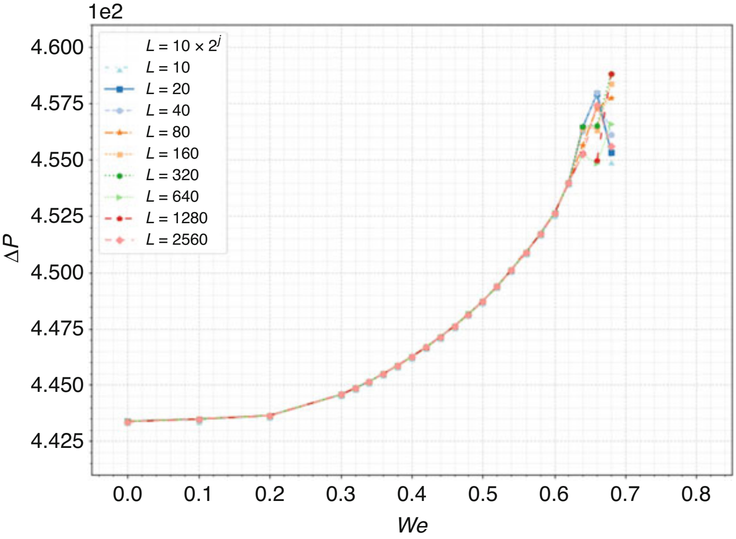


Fig. 8 Pressure drop depending on $\mathcal{W}e$ for different values of lengths of the averaging history L

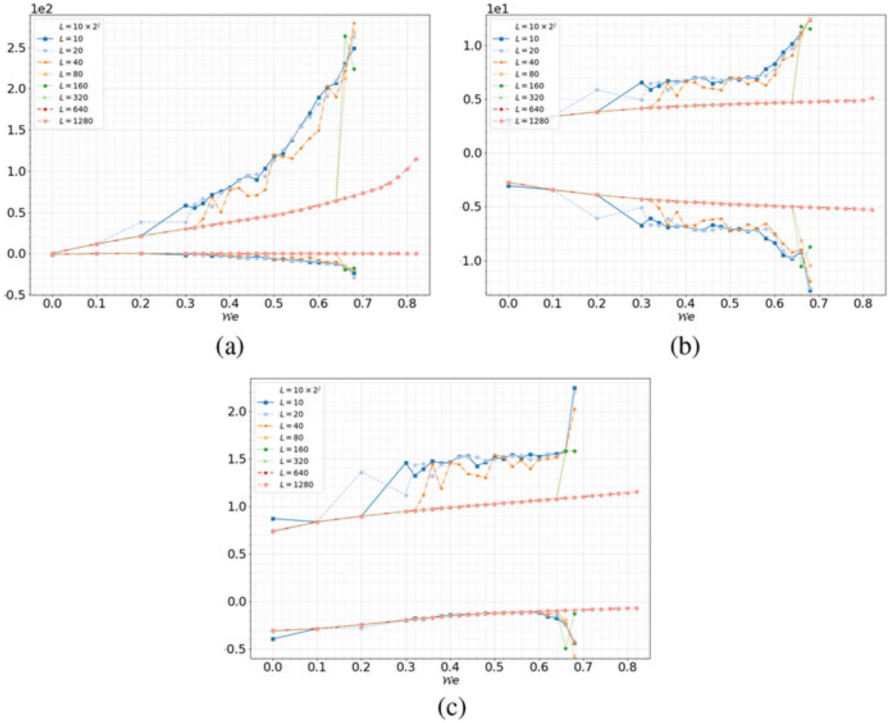


Fig. 9 Maximum and minimum values of the elastic stress tensor $\boldsymbol{\tau}$ components, for different lengths of the averaging history L with α defined by (39) and $\phi = p$ depending on the Weissenberg number $\mathcal{We}$. (**a**) Component τ_{11}. (**b**) Component τ_{12}. (**c**) Component τ_{22}

5.4 Residual Diffusive Term

In order to avoid the dilemma with the choice of the functional dependence of $\alpha(\cdot)$, the stabilization by added Laplacian can rather be applied to steady residual (time-derivative) of stress tensor. This indirect smoothing applied on residual seems to be very robust and yet quite easy to implement choice. For the numerical simulations using this residual stabilization defined by the diffusive term $\alpha \cdot \Delta \boldsymbol{\tau}_t$, the same values of the parameter α described in Sect. 5.1 were used, i.e., $\alpha \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. The contour fields of the components of the elastic stress tensor are shown in Fig. 12. The final fully converged results are almost identical for all choices of parameter α, including the reference non-diffusive solution with $\alpha = 0$. This is significant improvement over the standard artificial diffusion shown in Sect. 5.1 where the results were heavily affected by the choice of parameter α, leading to solutions that differ from the one of the original non-diffusive model.

Further details regarding the extremal stress values shown in Fig. 13 document the extension of the range of attainable Weissenberg numbers due to applied residual

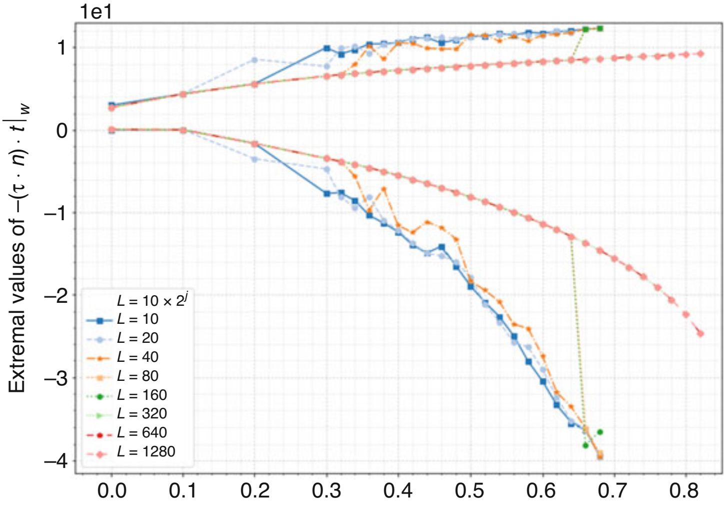


Fig. 10 Extremes values of the stress tension on the wall $-(\tau \cdot n) \cdot t|_w$ along the channel wall for different values of lengths of the averaging history L, depending on the Weissenberg number

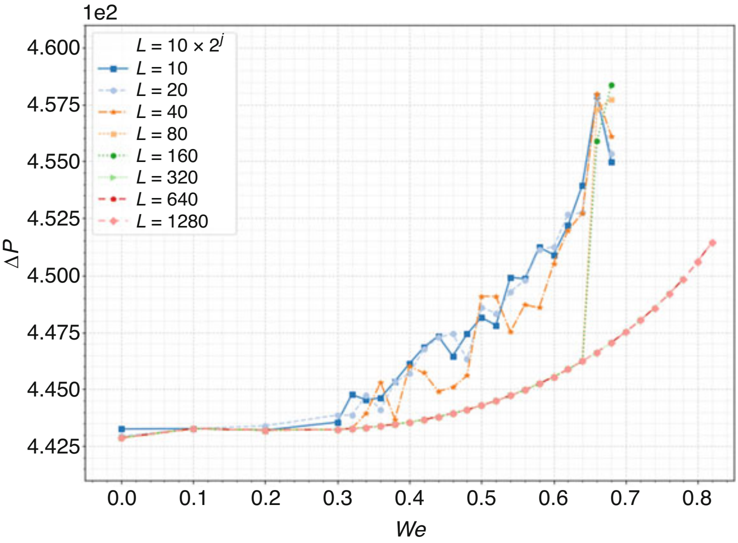


Fig. 11 Pressure drop dependence on $\mathcal{We}$ for different values of lengths of the averaging history L

stabilization. By choosing sufficiently high parameter α, the critical Weissenberg number was increased by about 50%. Figures 14 and 15 confirm this trend, showing that with $\alpha = 10^{-1}$ the critical Weissenberg number was raised up to $\mathcal{We} = 0.68$, while only $\mathcal{We} = 0.44$ could have been achieved without the stabilization.

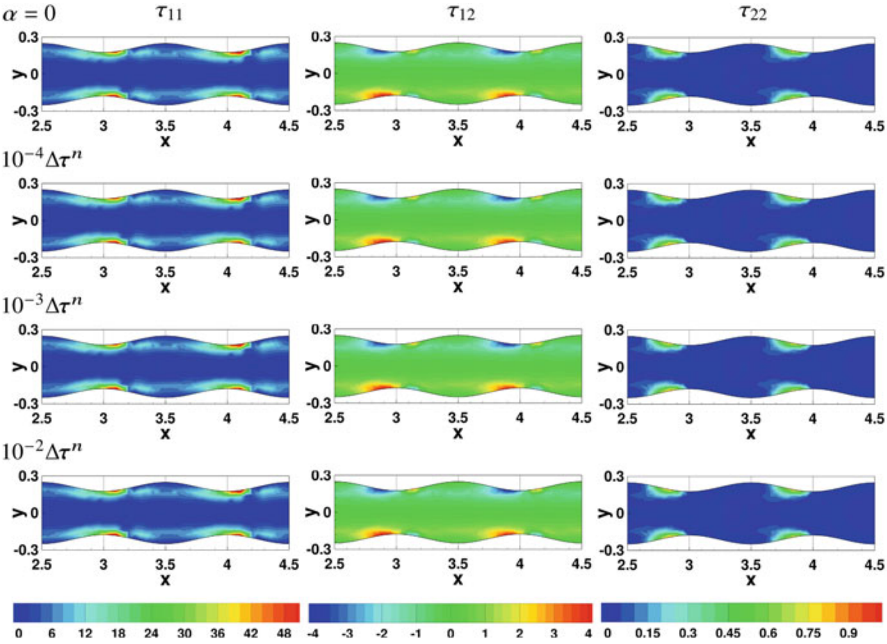


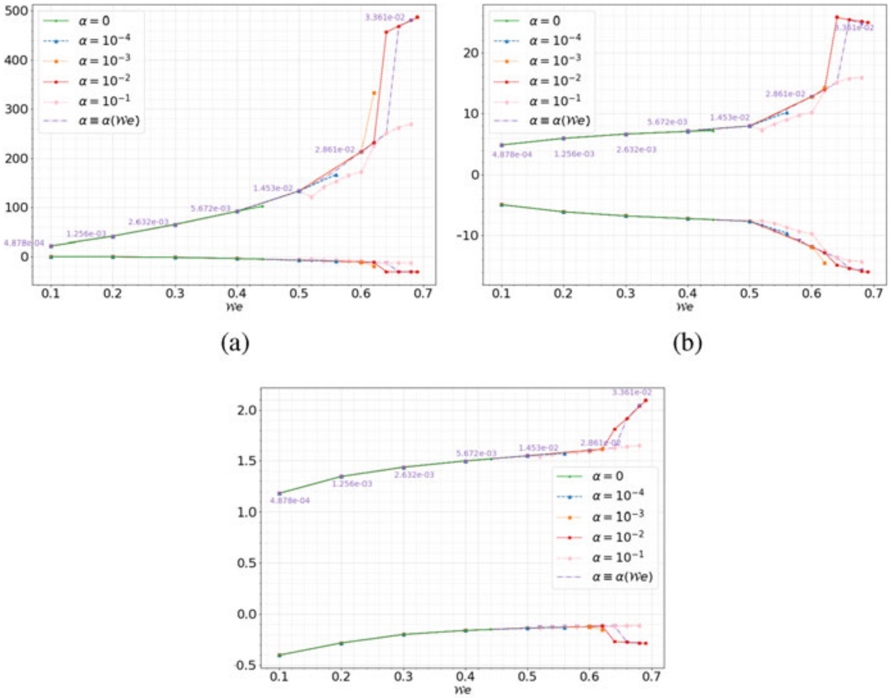
Fig. 12 Components of the elastic stress tensor $\boldsymbol{\tau}$, obtained using the residual stress diffusion (5)

Important is that this form of vanishing residual stabilization is safe in the sense that it doesn't affect the final steady solution.

6 Conclusions and Remarks

The series of numerical tests revealed several characteristic features and characteristics associated with the use of the described variants of the artificial stress diffusion. The main conclusions for all variants are summarized here, while for more details supporting these conclusions we refer to our previous works, where individual stress diffusion terms were studied separately.

(a) *Constant* diffusive term in the form (2) - This is the simplest generic version of the artificial stress diffusion. It was shown that such added term allows to extend the robustness and working range of the numerical method to higher Weissenberg numbers, but it comes at the price of reduced accuracy of the method. The simulations performed using this term for the cases at moderate Weissenberg numbers (where the solution can also be obtained without the use of any stabilization) have shown that the solutions obtained without and with this kind of diffusive term differ significantly, with progressive deterioration of the diffusive solution for higher values of the diffusion coefficient α. This is



(a)

(b)

(c)

Fig. 13 Maximum and minimum values of the elastic stress tensor $\boldsymbol{\tau}$ components, obtained using the residual diffusion term for different values of the parameter α in (5) depending on the Weissenberg number $\mathcal{W}e$. (**a**) Component τ_{11}. (**b**) Component τ_{12}. (**c**) Component τ_{22}

why this primitive form of the artificial diffusion should be used with extreme caution, with the level of diffusion being kept as small as possible. An attempt has been made to make an automatic adjustment of the diffusion coefficient with respect to Weissenberg number, i.e., $\alpha = \alpha(\mathcal{W}e)$ with values close to zero for low $\mathcal{W}e$ (when the method is stable even without any stabilization) and setting substantially larger values of α as the $\mathcal{W}e$ reaches certain critical threshold. The threshold value of $\mathcal{W}e$ is case dependent and should be properly adjusted together with the asymptotic value of α for the highest range of $\mathcal{W}e$. This automatic setup of $\alpha = \alpha(\mathcal{W}e)$ was studied in detail in [29], showing that this choice is safer for practical use, preventing excessive artificial diffusion to spoil the numerical results. In general however this constant (in time and space) choice of the diffusion coefficient α leads to results that are always to some extent affected by the artificial (non-physical) added diffusion. The advantage of this choice might be seen in simplicity of the formulation that allows for obtaining some theoretical results concerning the mathematical well-posedness

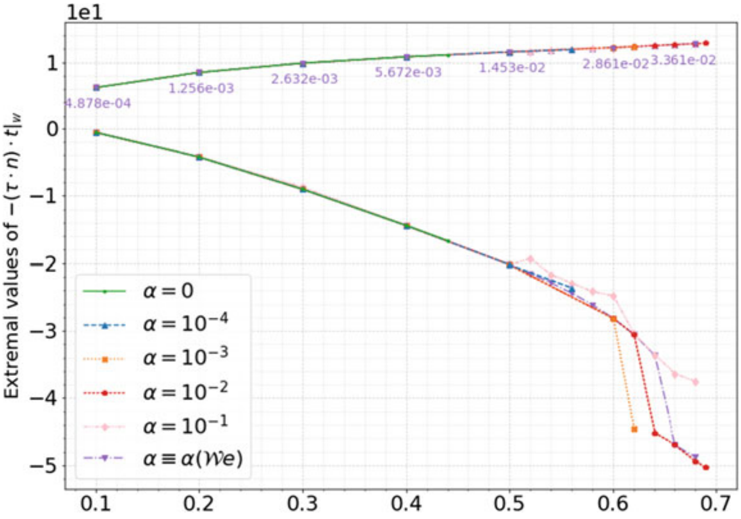


Fig. 14 Extreme values of the stress tension on the wall $-(\tau \cdot n) \cdot t|_w$ along the channel wall for different values of α, depending on the Weissenberg number

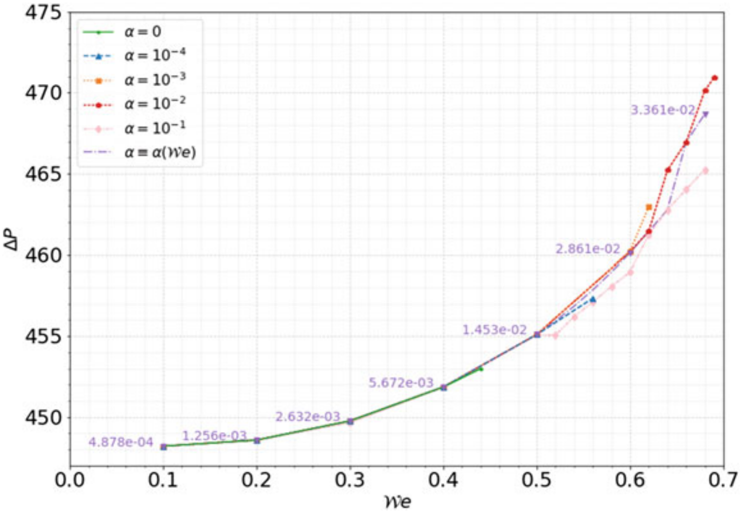


Fig. 15 Pressure drop depending on $\mathcal{W}e$ for different values of α

of the problem, including the stabilization term. Such theoretical results are often not available for the more complicated stabilization techniques.

(b) Time-dependent diffusive term in the form (3) - This is the first step in the improvement of the original constant diffusion technique, trying to remove its main weakness which is the presence of the added artificial term effects in the steady-state problem solution. The principle is based on practical observations

from numerical simulations showing that the method tends to be mostly unstable at the beginning of the iterative process, starting from some ad-hoc chosen initial condition. The added time-decay of the diffusion coefficient $\alpha = \alpha(t)$ helps to stabilize the initial phase of iterations, while it is significantly reduced (or completely vanished) towards the end of the time-marching iterative process, where the final (steady) solution should be recovered. This setting was successfully tested in [28], showing that if properly fine-tuned, this approach is very safe to use leading to solutions that are almost free of the artifacts of over-smoothing due to presence of the artificial diffusive term. The disadvantage of this approach can mainly be seen in the need to manually adjust the decay rate for the $\alpha(t)$, allowing it to reduce sufficiently before the iterative process is stopped. For this it is necessary to have some a-priori estimate of iterations needed to reach the steady state. As an advantage of this particular method can be named the fact that from the point of view the mathematical analysis, this version of the model is not any different as the previous case with constant diffusion coefficient α, provided that the choice of $\alpha(t)$ guarantees that it is a bounded, positive, smooth, monotonically decaying function of (iterative) time t.

(c) *Time-derivative-dependent* diffusive term in the form (4) - The problem of the previous (time-dependent) artificial diffusion method, where some estimate of the number of iterations till the steady state was needed, is resolved now by choosing the diffusion coefficient being a function of time-derivative of the solution $\alpha \propto \phi_t$. In this way it is possible to make the artificial diffusion term to completely vanish at the moment the steady solution is reached for which $\phi_t = 0$. This behavior was documented by simulations and discussed in detail in [28]. This time-derivative dependent setup showed to be quite robust and insensitive to the choice of the starting value of α_0 in (38). The main problem is to choose properly the flow variable ϕ on which the function $\alpha(\phi_t)$ will depend. The numerical experiments have shown that the choice of $\phi = \|\boldsymbol{\tau}\|$ is better than $\phi = p$, because pressure p tends to fluctuate more, which may de-stabilize the solution by rapidly varying coefficient α. The norm of the extra stress tensor $\|\boldsymbol{\tau}\|$ work visibly better, probably also because it combines the changes of all components of the tensor, avoiding some very fast variation typical for individual components or other scalar quantities. From the numerical implementation point of view this version of the added stabilization term is still almost identical to previous two simpler variants, but mathematical analysis of the underlying model is already significantly more complicated as some assumptions on temporal behavior of the solutions are needed to control the diffusion coefficient $\alpha(\phi_t)$.

(d) *Residual* diffusive term in the form (5) - This method proved to be most efficient and robust from the four artificial diffusion terms used in the present study. Its design was motivated by the idea of making the stabilization term proportional to the time derivative of the solution as in the case of $\alpha(\phi_t)$, but removing the need of finding suitable form of the function $\alpha(\phi_t)$. In the present form, the implementation of the method with $\Delta(\boldsymbol{\tau}^n - \boldsymbol{\tau}^{n-1}) = \Delta\boldsymbol{\tau}^n - \Delta\boldsymbol{\tau}^{n-1}$ is

very straightforward, because it only requires to remember and use again the values of $\Delta\boldsymbol{\tau}^{n-1}$ from the previous time step. The only adjustable parameter is the constant proportionality coefficient α in $\boldsymbol{E} = \alpha \cdot \Delta\boldsymbol{\tau}_t$. The numerical experiments have shown that the method is quite robust with respect to values of α, leading (obviously) to the same final steady solution, independently of the choice of α. It is interesting to note that although the stabilization term contains the third order mixed partial derivatives, the term is linear in contrast to the previously used time-derivative-dependent stabilization term in the form (4). This may simplify the rigorous mathematical analysis of the extended model and numerical solver. As mentioned earlier, the physical or mathematical interpretation of this stabilization term proportional to $\Delta(\boldsymbol{\tau}_t)$ (or $(\Delta\boldsymbol{\tau})_t$) is not so obvious as in the previous cases. Probably the best interpretation can be in the context of residual smoothing approach, considering that the difference $\boldsymbol{\tau}^n - \boldsymbol{\tau}^{n-1}$ corresponds to the steady residual for the stationary problem; hence the Laplacian is applied to smooth this residual between the iterations.

The stabilization methods presented here were considered in the context of iterative solution of steady problems; however the same approach can possibly be used to stabilize sub-iterative process within the unsteady time-stepping in physically non-stationary problems.

The numerical simulations performed in the framework of this paper have shown the applicability of various versions of the artificial stress diffusion for stabilization of numerical methods for simulation of viscoelastic Oldroyd-B fluid flows at moderate Weissenberg numbers. The artificial stress diffusion proved to be very simple and effective tool in improving the robustness of existing numerical solvers. The main advantage comes from the ease of implementation and use with existing standard codes. On the other hand although the artificial diffusion helps to enlarge the region of applicability of the basic numerical method, it is evident that this approach has some limitations, making it efficient just for moderate Weissenberg numbers. For extremely high Weissenberg numbers flows some other specialized methods and solvers should be recommended, using specific discretization techniques (e.g., based on the log-conformation tensor reformulation [6, 18]).

Acknowledgments This research was funded by *Fundação para a Ciência e a Tecnologia - FCT* under the *Grant Ref. SFRH/BSAB/150464/2019* and Project FCT - *UIDB/04674/2020* and partly by the *Praemium Academiae* of Š. Nečasová and RVO:67985840.

References

1. R.B. Bird, R.C. Armstrong, O. Hassager, *Dynamics of Polymeric Liquids*, vol. 1, 2nd edn. (John Willey & Sons, New York, 1987)
2. T. Bodnár, Ph. Fraunié, K. Kozel, Modified equation for a class of explicit and implicit schemes solving one-dimensional advection problem. Acta Polytechnica **61**(SI), 49–58 (2021)

3. A. Bressan, Viscosity solutions for nonlinear hyperbolic systems, in *Hyperbolic Problems: Theory, Numerics, Applications*, ed. by T.Y. Hou, E. Tadmor (Springer, Berlin, 2003), pp. 19–41
4. L. Chupin, S. Martin, Stationary Oldroyd model with diffusive stress: mathematical analysis of the model and vanishing diffusion process. J. Non-Newtonian Fluid Mech. **218**, 27–39 (2015)
5. W.P. Crowley, Numerical advection experiments. Monthly Weather Rev. **96**(1), 1–11 (1968)
6. H. Damanik, J. Hron, A. Ouazzi, S. Turek, A monolithic FEM approach for the log-conformation reformulation (LCR) of viscoelastic flow problems. J. Non-Newtonian Fluid Mech. **165**(19–20), 1105–1113 (2010)
7. R.J. DiPerna, Convergence of the viscosity method for isentropic gas dynamics. Commun. Math. Phys. **91**, 1–30 (1983)
8. G.P. Galdi, R. Rannacher, A.M. Robertson, S. Turek (eds.), *Hemodynamical Flows—Modeling, Analysis and Simulation*, vol. 37 of Oberwolfach Seminars (Birkäuser, Basel, 2008)
9. V. Girault, P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations—Theory and Algorithms*, volume 5 of *Springer Series in Computational Mathematics* (Springer, Berlin Heidelberg, 1996)
10. B. Guo, D. Bian, F. Li, X. Xi, *Vanishing Viscosity Method: Solutions to Nonlinear Systems* (De Gruyter, Berlin, 2016)
11. A. Harten, High resolution schemes for hyperbolic conservation laws. J. Comput. Phys. **49**(3), 357–393 (1983)
12. F. Hecht, New development in FreeFem++. J. Numer. Math. **20**(3–4), 251–265 (2012)
13. C. Hirsch, *Numerical Computation of Internal and External Flows*, vols. 1, 2 (John Willey & Sons, New York, 1988)
14. F. Huang, Z. Wang, Convergence of viscosity solutions for isothermal gas dynamics. SIAM J. Math. Anal. **34**(3), 595–610 (2002)
15. A. Jameson, Time dependent calculations using multigrid, with applications to unsteady flows past airfoils and wings, in *AIAA 10th Computational Fluid Dynamics Conference, Honolulu* (1991). AIAA Paper 91-1596
16. A. Jameson, W. Schmidt, E. Turkel, Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes, in *AIAA 14th Fluid and Plasma Dynamic Conference, Palo Alto* (1981). AIAA paper 81-1259
17. D.D. Joseph, *Fluid Dynamics of Viscoelasic Liquids*, volume 84 of *Applied Mathematical Sciences* (Springer, Berlin, 1990)
18. R. Kupferman, Simulation of viscoelastic fluids: Couette–taylor flow. J. Comput. Phys. **147**, 22–59 (1998)
19. R.J. LeVeque, *Numerical Methods for Conservation Laws*. Lectures in Mathematics (Birkhäuser Verlag, Basel, 1990)
20. R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. (Cambridge University Press, Cambridge, 2002)
21. R. Liska, B. Wendroff, Composite schemes for conservation laws. SIAM J. Numer. Anal. **35**(6), 2250–2271 (1998)
22. R. Liska, B. Wendroff, Composite centered schemes for multidimensional conservation laws, in *Hyperbolic Problems: Theory, Numerics, Applications*, ed. by R. Jeltsch, M. Fey (Birkhäuser, Basel, 1999), pp. 661–670
23. M. Lukáčová-Medvid'ová, H. Notsu, B. She, Energy dissipative characteristic schemes for the diffusive Oldroyd-B viscoelastic fluid. Int. J. Numer. Methods Fluids **81**, 523–557 (2016)
24. J.G. Oldroyd, Non-Newtonian effects in steady motion of some idealized elastico-viscous liquids. Proc. R. Soc. Lond. Ser. A **245**, 278–297 (1958)
25. R.G. Owens, T.N. Phillips, *Computational Rheology* (Imperial College Press, 2002)
26. M. Pires, T. Bodnár, On the influence of diffusion stabilization in Oldroyd-B fluid flow simulations, in *Topical Problems of Fluid Mechanics 2020* (Institute of Thermomechanics CAS, Prague, 2020), pp. 176–183

27. M. Pires, T. Bodnár, Numerical tests of vanishing diffusion stabilization in Oldroyd-B fluid flow simulations, in *Topical Problems of Fluid Mechanics 2021* (Institute of Thermomechanics CAS, Prague, 2021), pp. 102–109
28. M. Pires, T. Bodnár, Application of vanishing diffusion stabilization in Oldroyd-B fluid flow simulations. (2022) (to appear)
29. M. Pires, T. Bodnár, Temporal artificial stress diffusion for numerical simulations of Oldroyd-B fluid flow. Mathematics **10**(3), 404 (2022)
30. V. Průša, K.R. Rajagopal, Implicit type constitutive relations for elastic solids and their use in the development of mathematical models for viscoelastic fluids. Fluids **6**(3), 131 (2021)
31. A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations*, volume 23 of Springer Series in Computational Mathematics, 2nd edn. (Springer, Berlin, 2008)
32. O. Radulescu, P.D. Olmsted, Matched asymptotic solutions for the steady banded flow of the diffusive Johnson-Segalman model in various geometries. J. Non-Newtonian Fluid Mech. **91**, 143–164 (2000)
33. D. Trebotich, P. Colella, G.H. Miller, A stable and convergent scheme for viscoelastic flow in contraction channels. J. Comput. Phys. **205**, 315–342 (2005)

Cellular Automata Describing Non-equilibrium Fluids with Non-mixing Substances

Carlos Ramos, Fernando Carapau, and Paulo Correia

1 Introduction

Cellular automata (CA) are discrete dynamical systems introduced by Von Neumann and Ulam, in the late 1940s [1]. Since then many applications in natural sciences and mathematics were developed, for example, in fluid dynamics with lattice Boltzmann methods [2, 3], fluids in heterogeneous porous media [4], in genetics [7], dune dynamics [5], spatial pattern formation [6], and many others [8]. Cellular automata can be seen as an idealization of a physical system in which space, time, and certain physical quantities take a finite set of values. Cellular automata provide simple models of complex systems showing that collective complex behavior can emerge from the composition or interaction of simple components. Even if the local interactions are perfectly described in a direct manner, it is possible that the global behavior of a system obeys unexpected patterns. This fact makes CA suitable to model and simulate non-equilibrium systems. In the 1980s, Wolfram [9, 10] gave a classification of cellular automata which produces an intuitive way to distinguish the dynamical behavior of cellular automata in four distinct classes, accordingly: Class 1: almost every initial conditions produce an eventually fixed point behavior. Class 2: almost every initial conditions produce an eventually periodic behavior. Class 3: almost every initial conditions produce a pseudo-random behavior. Class 4: almost every initial conditions produce a complex behavior articulating regular patterns with structured non-periodic geometric patterns.

C. Ramos (✉) · F. Carapau · P. Correia
Departamento de Matemática and CIMA-UE, Universidade de Évora, Évora, Portugal
e-mail: ccr@uevora.pt; flc@uevora.pt; pcorreia@uevora.pt

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_10

A very detailed comparison of the CA methods, for practical fluid-dynamics problems, with conventional methods from numerical analysis is explained in [11], in particular, with details from a computational point of view, considering memory usage, computational time, and other characteristics.

Nevertheless, most of the interest in the use of cellular automata focuses on non-equilibrium fluids or fluids composed with different phases, in which the differential equations are hard to implement.

In [12] were introduced and discussed several techniques to explore evolutionary dynamics of the automata space, using biologically motivated concepts. In particular, specific genetic algorithms and techniques such as mutation, assembly, and recombination of CA. In that context the code rule of a CA was called the genotype, and the diverse characteristics of generic CA realizations were called phenotype. Here, these denominations are changed and adapted to the present context. The advantage of evolutionary methods is to efficiently obtain CA rules with specific characteristics. Previous work on evolutionary search over cellular automata can be found also in [13, 14].

In this paper are presented techniques for modeling systems, seen as idealized fluids, where may coexist distinct substances in diverse phases. These techniques, using cellular automata, are suitable to simulate transient, non-equilibrium behavior in fluid mechanics or other phenomena, such as fracture dynamics on heterogeneous materials.

Our main result is the development of the assembly method, introduced in [12], to determine CA code rules of increasing complex behavior. This means that the systems present an increasing number of distinct behavior and spatial-temporal patterns. A canonical process of assembling two CA rules is defined. This method allows the study of the singular perturbation of a complex fluid and the study of the interaction between two similar fluids subject to instabilities, leading to global phase transitions.

In Sect. 2 the notions and concepts used in the paper regarding cellular automata are introduced, in particular those notions from [12], such as the singular perturbation, assembly, and the canonical assembly. The basic CA rule $3E6IGS58S$, which is used in the simulations, is also defined. In Sect. 3, the computation of the canonical assembly of the CA rule $3E6IGS58S$, its variations, and the simulations of its perturbations are presented.

2 Preliminaries and Definitions

Some notions regarding one dimension cellular automata are here introduced. Let $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$, $n > 0$, be the *local state space*. Let $\phi : \mathbb{Z}_n^m \to \mathbb{Z}_n$ be a map, which determines the local dynamics of the system and is called *local map* or *CA rule*. An element in $\mathbb{Z}_n^m$, *i.e.*, a word or a block of size m in the alphabet $\mathbb{Z}_n$, is called a *local configuration*. The map ϕ induces a block map ϕ_k, $k \in \mathbb{N}$, which transforms words in $\mathbb{Z}_n$, of size $m + k$, into words of size k, through

$$\phi_k : \mathbb{Z}_n^{m+k} \to \mathbb{Z}_n^k,$$

$$\phi_k\left(x_1 \dots x_{m+k}\right) := \phi\left(x_1 \dots x_m\right) \phi\left(x_1 \dots x_m\right) \dots \phi\left(x_1 \dots x_m\right).$$

There is a natural identification of ϕ_1 with ϕ. To simplify the exposition, consider m to be an odd number so that $m = 2r + 1$, for a certain natural number r. The global map is then defined by

$$\Phi : \mathbb{Z}_n^{\mathbb{I}} \to \mathbb{Z}_n^{\mathbb{I}},$$

$$\Phi\left(x\right) := \left(\phi\left(x_{[j-r,j+r]}\right)\right)_{j \in \mathbb{I}},$$

where $\mathbb{I}$ can be $\mathbb{Z}$, $\mathbb{N}$ or a finite set $\mathbb{Z}_L = \{1, 2, \dots, L\}$. A *cellular automaton* is the specification of the number of local states n, the size of the local configuration m, the local map or CA rule ϕ, the configuration space or global state space $\mathbb{I}$, and if needed, the boundary conditions which depend naturally on $\mathbb{I}$ and m. The time evolution of the system is given by the iteration of the map Φ, given an initial condition $x\left(0\right) = \left(x_i\left(0\right)\right)_{i \in \mathbb{I}} \in \mathbb{Z}_n^{\mathbb{I}}$,

$$\begin{gathered} x\left(t+1\right) = \Phi\left(x\left(t\right)\right), t \geq 0, \\ x\left(0\right) = \left(x_i\left(0\right)\right)_{i \in \mathbb{I}} \in \mathbb{Z}_n^{\mathbb{I}}. \end{gathered}$$

The parameter $m = 2r + 1$ gives the dependence of each state, in the next time instant on the states of the neighbor cells, r cells to the left and r cells to the right. In the case $\mathbb{I}$ is $\mathbb{N}$ or a finite set, it is necessary to specify boundary conditions, on the left in the first case and both left and right in the second case. For convenience, $[j]_n$ denote the n-expansion of the natural number j, that is, j in base n. By convention, the number of digits in $[j]_n$ is fixed and equal to n^m. That is, if $[j]_n = j_1 \dots j_r$ then

$$j = j_1 \times n^{r-1} + j_2 \times n^{r-2} + \dots + j_{r-1} \times n^1 + j_r \times n^0.$$

On the other hand, a word $j_1 \dots j_r$, in $\mathbb{Z}_n^r$ with $r \geq 1$, can be seen as a representation of a natural number $j \in \mathbb{N}$, in base n, denoted by $\langle j_1 \dots j_r \rangle_n \in \mathbb{N}$. With this notation

$$\begin{aligned} j_1 \dots j_r \in \mathbb{Z}_n^r &\to \langle j_1 \dots j_r \rangle_n = j \in \mathbb{N}, \\ j \in \mathbb{N} &\to [j]_n = j_1 \dots j_r \in \mathbb{Z}_n^r. \end{aligned}$$

Once fixed the value m and the configuration space $\mathbb{I}$ (and eventually the boundary condition), a cellular automaton is completely characterized specifying a sequence $\alpha = \left(\alpha_1, \dots, \alpha_{n^m}\right) \in \mathbb{Z}_n^{n^m}$ corresponding to the sequence of the images of every local configuration under ϕ. This sequence is called *CA code rule* and is a *functional representation* of the CA, that is, a particular symbol in a certain position in the referred sequence has a functional meaning. The position j in the sequence α

gives a configuration which is the n-expansion of the integer $(j-1)$ and the value α_j is the image of that configuration under the rule ϕ, that is, $\alpha_j = \phi\,[j-1]_n$. Therefore, the CA code rule is

$$\alpha = \left(\phi\,[j-1]_n\right)_{j=1}^{n^m}.$$

A more compact way to give a particular CA code rule is to use the *Wolfram numbering*. The CA code rule is seen as the expansion in base n of a certain number which when converted to decimal is designated as the Wolfram number of the CA code rule. If $n < 10$ the number of digits of the Wolfram number is less than the number of digits corresponding to the original CA code rule; therefore it is a more compact way of specifying the CA rule. An even more compact form is to use hexadecimal base (if the number of states is less than 16), or a larger base number. Since we deal with very large CA-code rules, we will use base 32-expansion to represent the CA code rules in compact way. The base 32, similarly to base 16, uses the digit set

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V\}.$$

As an illustrative example consider the CA code rule 01110110 which determines the local map

$$\begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 \end{array}$$

The sequence 01110110 represents a number in binary. The corresponding number in decimal base is $110 = 0\times2^0+1\times2^1+1\times2^2+1\times2^3+0\times2^4+1\times2^5+1\times2^6+0\times2^7$ (note the reversed order). In hexadecimal the rule 110 is designated by $6E$, and in base 32 is $3E$. See Figs. 1 and 2.

To resume, a CA code rule will be a sequence $\alpha = \alpha_1\alpha_2\ldots\alpha_{n^m} \in \mathbb{Z}_n^{n^m}$, with $n \in \mathbb{N}$. The space of the CA code rules is denoted by $\mathcal{G}$. The space of CA code rules which have n different symbols is denoted by $\mathcal{G}_n$, and the space of CA code rules which have n different symbols and with neighbor number equal to m is denoted by $\mathcal{G}_{n,m}$.

The cellular automaton which is central in this work is a 3-state rule, with $m = 3$ and $\mathbb{I} = \mathbb{Z}_L$, for a certain natural L. The CA code rule is

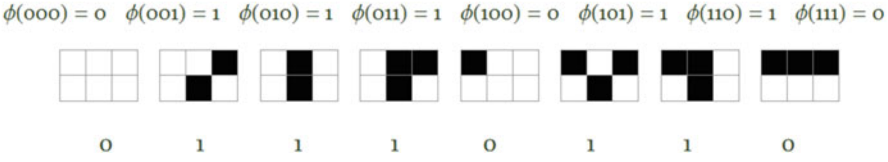


Fig. 1 Explicit CA code rule 110, in Wolfram numbering

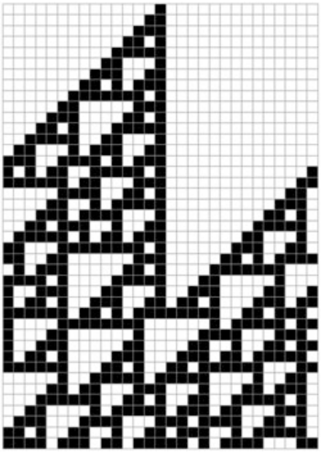


Fig. 2 Example of a realization of the automaton rule 110

$$\alpha = 202000211011010222222101111.$$

The local map in $\mathbb{Z}_3 = \{0, 1, 2\}$, is then defined by

$$\begin{array}{ccccccccc}
000 & 001 & 002 & 010 & 011 & 012 & 020 & 021 & 022 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
2 & 0 & 2 & 0 & 0 & 0 & 2 & 1 & 1
\end{array}$$

$$\begin{array}{ccccccccc}
100 & 101 & 102 & 110 & 111 & 112 & 120 & 121 & 122 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
0 & 1 & 1 & 0 & 1 & 0 & 2 & 2 & 2
\end{array}$$

$$\begin{array}{ccccccccc}
200 & 201 & 202 & 210 & 211 & 212 & 220 & 221 & 222 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
2 & 2 & 2 & 1 & 0 & 1 & 1 & 1 & 1
\end{array}$$

The corresponding natural number (in decimal base), Wolfram number, is

$$\begin{aligned}
3786635351324 &= 2 \times 3^0 + 0 \times 3^1 + 2 \times 3^2 + 0 \times 3^3 + 0 \times 3^4 + 0 \times 3^5 \\
&+ 2 \times 3^6 + 1 \times 3^7 + 1 \times 3^8 + 0 \times 3^9 + 1 \times 3^{10} + 1 \times 3^{11} \\
&+ 0 \times 3^{12} + 1 \times 3^{13} + 0 \times 3^{14} + 2 \times 3^{15} + 2 \times 3^{16} + 2 \times 3^{17} \\
&+ 2 \times 3^{18} + 2 \times 3^{19} + 2 \times 3^{20} + 1 \times 3^{21} + 0 \times 3^{22} + 1 \times 3^{23} \\
&+ 1 \times 3^{24} + 1 \times 3^{25} + 1 \times 3^{26}.
\end{aligned}$$

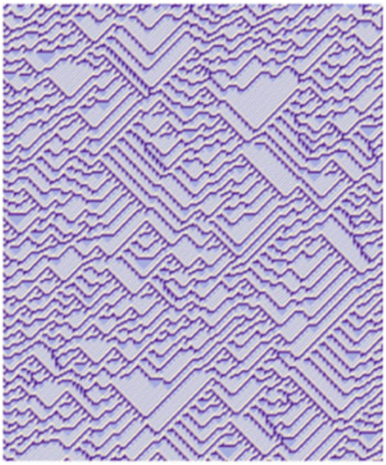


Fig. 3 Realization of CA rule $3E6IGS58S$, with random initial conditions

Note that the same number in base 2 is the CA code rule in reversed order, that is,

$$1111012222220101101112000202_{binary} = 3786635351324_{decimal}.$$

In a more compact description, its hexadecimal representation is $371A50E151C$ and in 32-base is $3E6IGS58S$. This last representation will be chosen to refer the CA rule, since it is shorter. In Fig. 3 we present an example of a realization with initial global state given by a random vector $x_0 \in \mathbb{Z}_3^{150}$.

2.1 Singular Perturbation and Pattern Stability

A singular perturbation of the CA rule is a transformation in a single symbol of the CA code rule, and it is the simplest possible transformation defined on the rule space $\mathcal{G}_{n,m}$. This perturbation can be generated randomly or generated by a deterministic process. To give a singular perturbation, it is necessary to specify the position in the CA code rule where the mutation is to occur and how it occurs. Recall that a position j in the sequence α gives a configuration which is the $n-$expansion of the integer $(j-1)$, that is, $[j-1]_n$, $(00\ldots00$ is the configuration of the position 1), and the value α_j is the image of the configuration under the automaton ϕ, that is, $\alpha_j = \phi\,[j-1]_n$, $j = 1, \ldots, n^m$.

Now, consider the stability of the patterns produced by time evolution of an initial condition, with respect to singular perturbation. There are several cellular automata which are very robust under singular perturbation, regarding the geometric structure of the patterns produced, and others very sensitive. However, some CA are robust to singular perturbation in some positions and in other positions are strongly sensitive. As an example of this phenomena, see the Figs. 4, 5, 6, 7, and 8. The same CA code

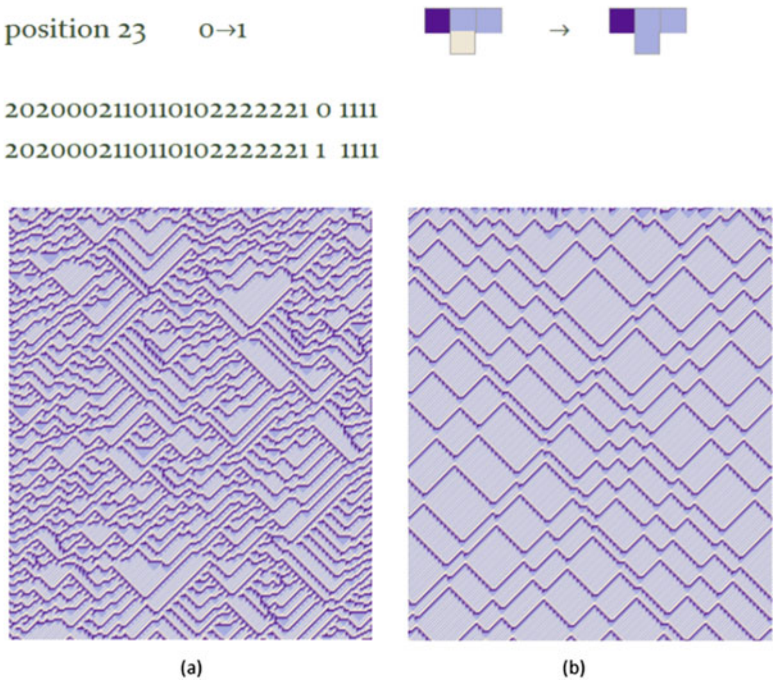


Fig. 4 (**a**) Realization of the CA rule $3E6IGS58S$. (**b**) Realization of a perturbation of the CA rule $3E6IGS58S$ at position 23

rule $3E6IGS58S$ is singularly perturbed in different positions, and for each case a realization of the CA is obtained, with random initial conditions. The realizations show the similarity of some of the mutated CA codes and the drastic changes in others.

2.2 Assembly of CA Code Rules

Next, it is described the assembly technique which produces CA rules obtained from two given CA rules. The assembled CA rule inherits several characteristics from the original rules; in particular, it maintains the original CA as subcases for special initial conditions. Let $\alpha = \alpha_1 \ldots \alpha_{p^m} \in \mathcal{G}_{p,m}$ and $\beta = \beta_1 \ldots \beta_{q^m} \in \mathcal{G}_{q,m}$ be two CA code rules with $p, q, m \in \mathbb{N}$. The alphabet underlying $\mathcal{G}_{p,m}$ is, as usual $\mathbb{Z}_p$, and for $\mathcal{G}_{q,m}$ is $\mathbb{Z}_q$. The assembly of α with β is a general procedure which gives a class of CA code rules in the space $\mathcal{G}_{n,m}$, where $n = p + q$.

The first p symbols of $\mathbb{Z}_n$ are reserved to codify the rule α and the last q symbols of $\mathbb{Z}_n$ to codify the rule β, using the correspondence

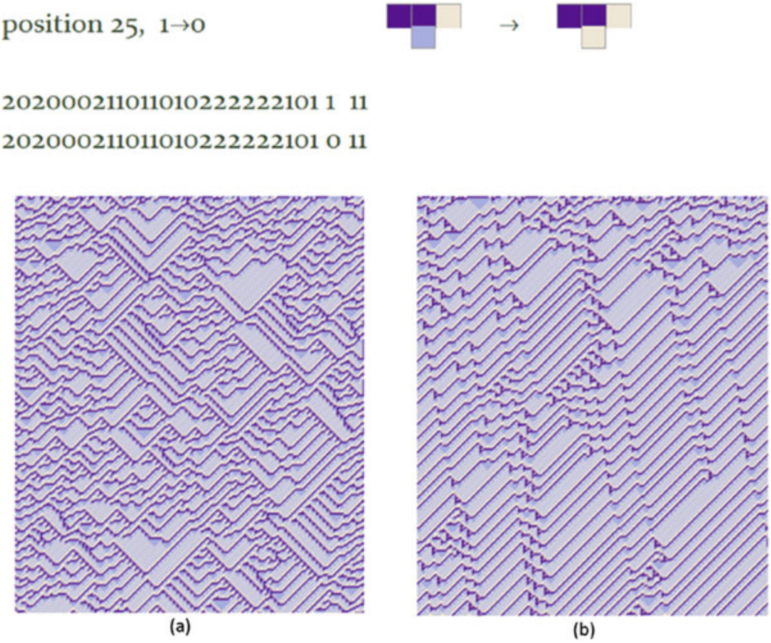


Fig. 5 (**a**) Realization of the CA rule $3E6IGS58S$. (**b**) Realization of a perturbation of the CA rule $3E6IGS58S$ at position 23, with random initial conditions

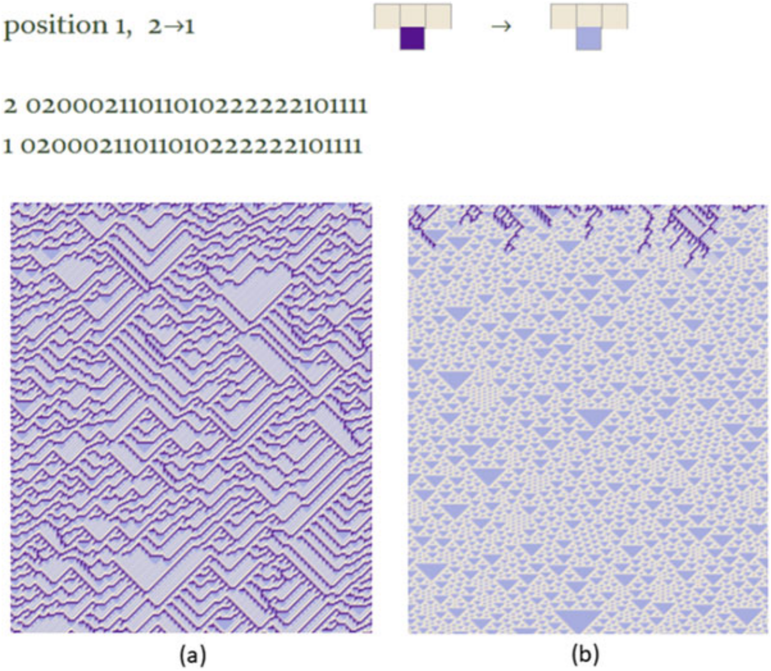
$$\begin{array}{ccccc} \mathbb{Z}_p & 0 & \dots & p-1 \\ \downarrow & \downarrow & & \downarrow \\ \mathbb{Z}_n & 0 & \dots & p-1 \end{array} \quad \text{and} \quad \begin{array}{ccccc} \mathbb{Z}_q & 0 & \dots & q-1 \\ \downarrow & \downarrow & & \downarrow \\ \mathbb{Z}_n & p & \dots & p+q \end{array}$$

Denote the correspondence$\widehat{\ }$: $\mathbb{Z}_q \to \{p, \dots, p+q\} \subset \mathbb{Z}_n$ and the reversed correspondence$\widetilde{\ }$: $\{p, \dots, p+q\} \to \mathbb{Z}_q$. Note that$\widehat{\ }$can be seen as adding p to each symbol if each symbol is seen as a natural number. For each local configuration $i_1 \dots i_m \in \mathbb{Z}_p^m$, corresponding to the CA code rule $\alpha \in \mathcal{G}_{p,m}$, it is associated the same configuration (with the same symbols) in $\mathbb{Z}_n^m$. To each configuration $j_1 \dots j_m \in \mathbb{Z}_q^m$, corresponding to the CA code rule $\beta \in \mathcal{G}_{q,m}$ it is associated the configuration $\widehat{j_1} \dots \widehat{j_m}$ in $\mathbb{Z}_n^m$. This gives a large number of degrees of freedom to choose the image of the local map associated with configurations which mix symbols from $\{0, \dots, p-1\}$ and $\{p, \dots, p+q\}$. This means that there are many different CA code rules arising from assembly of two specific CA code rules α, β,.

Let ϕ_α,ϕ_β, ϕ_γ denote the local rules for each CA code rule α, β, γ. Then γ is a CA code rule assembly of α, β if the following property is Satisfied:

$$x_1 \dots x_m \in \{0, \dots, p-1\}^m \Rightarrow \phi_\gamma\left(x_1 \dots x_m\right) = \phi_\alpha\left(x_1 \dots x_m\right),$$

$$x_1 \dots x_m \in \{p, \dots, p+q\}^m \Rightarrow \phi_\gamma\left(x_1 \dots x_m\right) = \phi_\beta\left(\widetilde{x}_1 \dots \widetilde{x}_m\right).$$



position 1, 2→1

2 020002110110102222222101111

1 020002110110102222222101111

Fig. 6 (**a**) Realization of the CA rule $3E6IGS58S$. (**b**) Realization of a perturbation of the CA rule $3E6IGS58S$ at position 23

The local configurations for ϕ_γ with digits exclusively from $\{0, \ldots, p-1\}$ or exclusively $\{p, \ldots, p+q\}$ are called *pure local configurations*; the local configurations mixing digits from $\{0, \ldots, p-1\}$ and $\{p, \ldots, p+q\}$ are called *mixed local configurations*. The images under ϕ_γ of pure local configurations in $\{0, \ldots, p-1\}$ are determined by ϕ_α, and the images under ϕ_γ of pure local configurations in $\{p, \ldots, p+q\}$ are determined by ϕ_β. The images under ϕ_γ of the mixed configurations are not determined by α, β. Therefore, must be as external parameters or degrees of freedom. As an example, consider the CA code rules $\alpha \in \mathcal{G}_{2,3}$, rule 18, and $\beta \in \mathcal{G}_{2,3}$, rule 110, given by

$$\alpha = 01001000 \quad \text{and} \quad \beta = 01110110.$$

The second CA code rule, β, is transformed *via* $0 \to \widehat{0} = 2$ and $1 \to \widehat{1} = 3$ into

$$\widehat{\beta} = 23332332.$$

Note that the cellular automata β and $\widehat{\beta}$ are equivalent, although the symbols are distinct; therefore, the two automata are identified $\beta \longleftrightarrow \widehat{\beta}$.

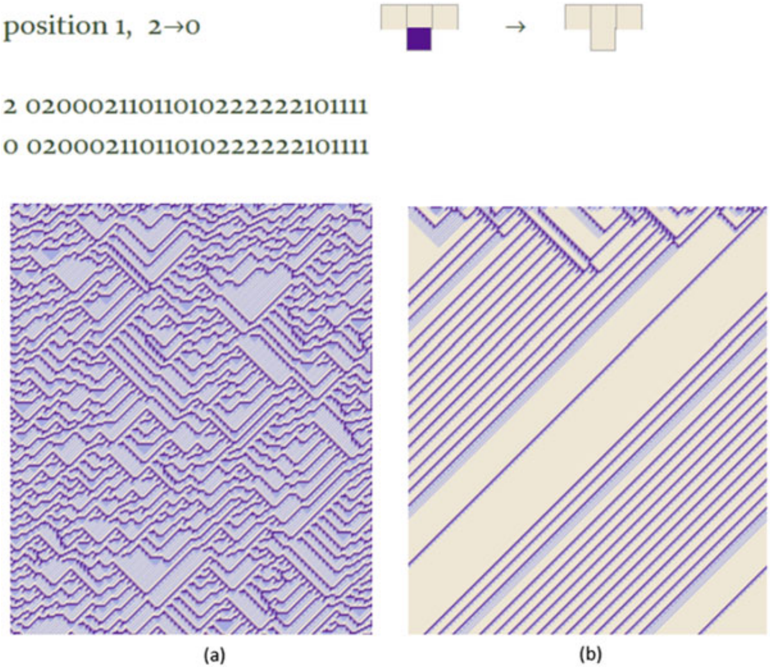
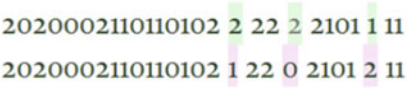


Fig. 7 (**a**) Realization of the CA rule $3E6IGS58S$. (**b**) Realization of a perturbation of the CA rule $3E6IGS58S$ at position 23

Consider a CA code rule $\gamma \in \mathcal{G}_{4,3}$ obtained by assembly of α and β. Therefore, corresponds to a cellular automaton that when restricted to initial conditions (and boundary conditions) with states $0, 1$ will reproduce the exact patterns of α and when restricted to initial conditions (and eventual boundary conditions) with states $3, 4$ will reproduce the patterns of β (up to the transformation $0 \to 3$, $1 \to 4$). A CA code rule γ satisfying this property is called the assembly of α and β. There are many different CA code rules arising from assembly. The local map in $\mathbb{Z}_4 = \{0, 1, 2, 3\}$, for a rule $\gamma \in \mathcal{G}_{4,3}$, assembly of α and β has the following structure:

$$\begin{array}{cccccccc}
000 & 001 & 002 & 003 & 010 & 011 & 012 & 013 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\gamma_1 = \alpha_1 & \gamma_2 = \alpha_2 & \gamma_3 & \gamma_4 & \gamma_5 = \alpha_3 & \gamma_6 = \alpha_4 & \gamma_7 & \gamma_8
\end{array}$$

$$\begin{array}{cccccccc}
020 & 021 & 022 & 023 & 030 & 031 & 032 & 033 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\gamma_9 & \gamma_{10} & \gamma_{11} & \gamma_{12} & \gamma_{13} & \gamma_{14} & \gamma_{15} & \gamma_{16}
\end{array}$$



2020002110110102 2 22 2 2101 1 11

2020002110110102 1 22 0 2101 2 11

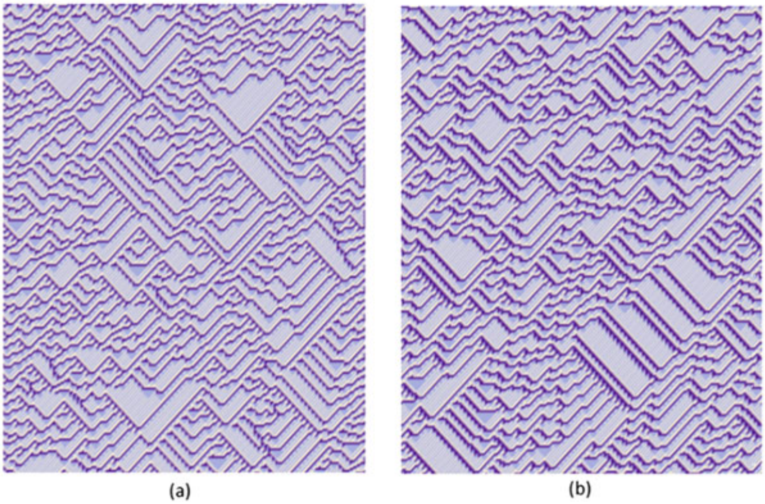


Fig. 8 (**a**) Realization of the CA rule $3E6IGS58S$. (**b**) Realization of a perturbation of the CA rule $3E6IGS58S$ at position 23

$$\begin{array}{cccccccc} 100 & 101 & 102 & 103 & 110 & 111 & 112 & 113 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \gamma_{17}=\alpha_5 & \gamma_{18}=\alpha_6 & \gamma_{19} & \gamma_{20} & \gamma_{21}=\alpha_7 & \gamma_{22}=\alpha_8 & \gamma_{23} & \gamma_{24} \end{array}$$

$$\begin{array}{cccccccc} 120 & 121 & 122 & 123 & 130 & 131 & 132 & 133 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \gamma_{25} & \gamma_{26} & \gamma_{27} & \gamma_{28} & \gamma_{29} & \gamma_{30} & \gamma_{31} & \gamma_{32} \end{array}$$

$$\begin{array}{cccccccc} 200 & 201 & 202 & 203 & 210 & 211 & 212 & 213 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \gamma_{33} & \gamma_{34} & \gamma_{35} & \gamma_{36} & \gamma_{371} & \gamma_{38} & \gamma_{39} & \gamma_{40} \end{array}$$

$$\begin{array}{cccccccc} 220 & 221 & 222 & 223 & 230 & 231 & 232 & 233 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \gamma_{41} & \gamma_{42} & \gamma_{43}=\widehat{\beta_1} & \gamma_{44}=\widehat{\beta_2} & \gamma_{45} & \gamma_{46} & \gamma_{47}=\widehat{\beta_3} & \gamma_{48}=\widehat{\beta_4} \end{array}$$

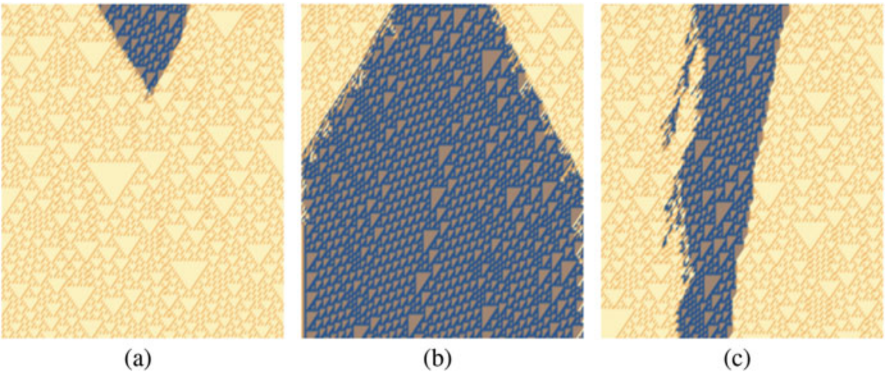


Fig. 9 Distinct assembly of rule 18 with rule 110: (**a**) Realization for $\gamma^{(1)}$. (**b**) Realization for $\gamma^{(2)}$. (**c**) Realization for $\gamma^{(3)}$

$$\begin{array}{cccccccc}
300 & 301 & 302 & 303 & 310 & 311 & 312 & 313 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\gamma_{49} & \gamma_{50} & \gamma_{51} & \gamma_{52} & \gamma_{53} & \gamma_{54} & \gamma_{55} & \gamma_{56}
\end{array}$$

$$\begin{array}{cccccccc}
320 & 321 & 322 & 323 & 330 & 331 & 332 & 333 \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\gamma_{57} & \gamma_{58} & \gamma_{59} = \widehat{\beta}_5 & \gamma_{60} = \widehat{\beta}_6 & \gamma_{61} & \gamma_{62} & \gamma_{63} = \widehat{\beta}_7 & \gamma_{64} = \widehat{\beta}_8
\end{array}$$

In Fig. 9 are shown three realizations of distinct CA-rules in $\mathcal{G}_{4,3}$ arising from the assembly of α and β. From left to right are denoted by $\gamma^{(1)}$, $\gamma^{(2)}$, and $\gamma^{(3)}$. The initial conditions are composed by two segments with random initial conditions from $\{0, 1\}$, and the middle segment is generated randomly from $\{2, 3\}$. The difference between the rules $\gamma^{(1)}$, $\gamma^{(2)}$ and $\gamma^{(3)}$ are that for $\gamma^{(1)}$ the values of the rule for mixed local configurations are taken randomly only from $\{0, 1\}$ which means that the patterns arising from α dominate. For $\gamma^{(2)}$ the values of the rule for mixed local configurations are taken randomly only from $\{2, 3\}$ which means that the patterns arising from β dominate. Finally, for $\gamma^{(3)}$ the values of the rule for mixed local configurations are taken randomly only from $\{0, 1, 2, 3\}$ with equal probability, which means that the initial patterns arising from α and β mix and interact along the time flow.

The examples shown in Fig. 9, with $\alpha = 01001000$ and $\beta = 01110110$, correspond to

$$\gamma^{(1)} = 0112001120110101101100200111012000101000022311331001101202232132,$$

with a majority of states for mixed configurations taken randomly from $\{0, 1\}$,

$$\gamma^{(2)} = 0122001322123301103200332232312232321203022331331203321232232132,$$

with a majority of states for mixed configurations taken randomly from {2, 3},

$$\gamma^{(3)} = 0132003130120123101100230113312000201000022313331023103232232132,$$

with an equilibrium of states for mixed configurations taken randomly with probability 1/2 from {0, 1} and {2, 3}.

2.3 Canonical Assembly of a CA Rule

Consider now a process to assembly two copies of the same CA rule. In this case, it is possible to turn the assembly uniquely determined, that is, not depending on externally given parameters. Therefore, this process is called *canonical assembly*. The canonical assembly can be viewed as the embedding of a particular system in a larger one containing two copies of the original system. This process is particularly important if it is necessary to model a system in non-equilibrium which is transforming and exhibiting new patterns of behavior although maintaining others. This can be achieved allowing singular perturbations after a canonical assembly, as it is seen in the next section.

Let $\alpha \in \mathcal{G}_{p,m}$ and $n = 2p$. Consider the state transformations

$$\widehat{\ }: \mathbb{Z}_n \to \mathbb{Z}_n \ \text{ and } \widetilde{\ }: \mathbb{Z}_n \to \mathbb{Z}_n,$$

with

$$\widehat{0} = p, \widehat{1} = p+1, \ldots, \widehat{p-1} = 2p-1, \quad \widehat{p} = p, \widehat{p+1} = p+1, \ldots, \widehat{2p-1} = 2p-1,$$

and

$$\widetilde{0} = 0, \widetilde{1} = 1, \ldots, \widetilde{p-1} = p-1, \quad \widetilde{p} = 0, \widetilde{p+1} = 1, \ldots, \widetilde{2p-1} = p-1.$$

Note that $\widehat{\mathbb{Z}_n} = \{p, p+1, \ldots, 2p-1\}$ and $\widetilde{\mathbb{Z}_n} = \mathbb{Z}_p = \{0, 1, \ldots, p-1\}$. Moreover, $\widetilde{\widehat{x}} = x$ and $\widehat{\widetilde{x}} = x$.

Let $N_X\,(i_1 i_2 \ldots i_m)$ be the number of digits in $i_1 i_2 \ldots i_m$ belonging to a certain subset $X \subset \mathbb{Z}_n$. Let

$$\chi\,(i_1 i_2 \ldots i_m) = \begin{cases} 0 \text{ if } N_{\mathbb{Z}_p}\,(i_1 i_2 \ldots i_m) > r, \\ 1 \text{ if } N_{\mathbb{Z}_p}\,(i_1 i_2 \ldots i_m) \leq r. \end{cases}$$

Recall that $r = m/2 - 1$. The condition above simply determines if the number of digits in $i_1 i_2 \ldots i_m$ belonging to $\mathbb{Z}_p$ is larger than the number of digits in $\{p, p+1, \ldots, n-1\}$, with $n = 2p$. Then the canonical assembly of α produces a CA code rule $\gamma = (\gamma_k)_{k=1,\ldots n^m}$ with

$$\gamma_{\langle i_1 \dots i_m \rangle_n} = \begin{cases} \alpha_{\langle \widetilde{i_1} \dots \widetilde{i_m} \rangle_p} \text{ if } \chi\left(i_1 i_2 \dots i_m\right) = 0, \\ \beta_{\langle \widetilde{i_1} \dots \widetilde{i_m} \rangle_p} \text{ if } \chi\left(i_1 i_2 \dots i_m\right) = 1. \end{cases}$$

where $\beta = \widehat{\alpha} = (\widehat{\alpha}_k)_{k=1,\dots,p^m}$. Recall that $\langle i_1 \dots i_m \rangle_n$ is the position number associated to the local configuration $i_1 \dots i_m$ in base n. The number $\left\langle \widetilde{i_1} \dots \widetilde{i_{mp}} \right\rangle$ is the position number associated with the local configuration $\widetilde{i_1} \dots \widetilde{i_m}$ in base p, since the transformation$\widetilde{\ }$sends $\mathbb{Z}_n$ to $\mathbb{Z}_p$. The canonical assembly produces a CA rule which in practical terms, reproduces two copies of the same CA with the duplication of the number of states.

3 Case Study: Rule 3*E*6*IGS*58*S*

The CA rule $3E6IGS58S$ (see Fig. 3) is seen as an idealized fluid where two substances which do not mix easily and two different phases of one of the substances are in unstable equilibrium. The states 0, 1 (lighter colors) are seen as the same substance in a different phase, and the state 2 (darker color) is a different substance. This phenomenon reflects on the persistency of the local state 2 in refined geometric structures and on the interaction between states 0, 1 which interchanges in a complex way.

Through the general process of the canonical assembly applied to the CA rule $3E6IGS58S$, it is obtained a CA rule which models a system composed of two fluids of the same type. Moreover, the perturbation of the CA rule leads to complex behavior, where the realizations of the CA rule present the patterns of the original fluids and the patterns arising from the perturbation. In particular, the singular perturbations considered in Sect. 2.1, for the individual fluid, are applied.

3.1 Canonical Assembly of Rule 3*E*6*IGS*58*S*

Let $p = 3$, consider the 3-state CA rule $3E6IGS58S$, given by

$$\alpha = 202000211011010222222101111,$$

and let

$$\beta = \widehat{\alpha} = 535333544344343555555434444,$$

which is obtained from α adding 3 to each symbol, as explained in the assembly section. Now, consider the canonical assembly of α, with

$$\widehat{0} = 2, \widehat{1} = 3, \widehat{2} = 5, \widehat{3} = 3, \widehat{4} = 4, \widehat{5} = 5,$$

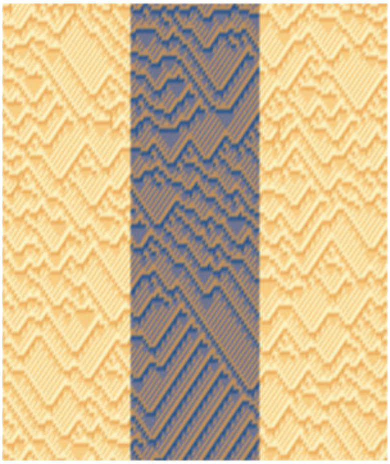


Fig. 10 Realization for the canonical assembly of the rule with structured initial conditions: three segments randomly generated from {0, 1, 2}, from {3, 4, 5}, and again from {0, 1, 2}. Random initial conditions

and

$$\widetilde{0} = 0, \widetilde{1} = 1, \widetilde{2} = 2, \widetilde{3} = 0, \widetilde{4} = 1, \widetilde{5} = 2.$$

The CA code rule is, represented in base 32,

8598OE44A81JS1KVBGVUQ5KDRJ7UP0JUBPL9CBFFBDIPFH1669SA00CU2

...MLT0B3AI26QSATJCN6LO7PKRTSIC2QFB180IQJTCQAFIUC6CEOSMSRI.

In Fig. 10 is shown a realization of the canonical assembly of the rule, exhibiting the coexistence of the two fluids in similar regimes.

3.2 *Perturbations of the Canonical Assembly*

In Fig. 11 it is shown the singular perturbations of the canonical assembly of the CA rule $3E6IGS58S$ in which there is a singular perturbation in the local configurations: 333, 5 $\mapsto$ 4, 544, 3 $\mapsto$ 4, 554, 4 $\mapsto$ 3, showing the coexistence of the patterns of the original fluid and the patterns arising from the perturbed CA, from Sect. 2.1, Figs. 4, 5, and 6.

Finally, consider the assembly of two CA code rules with randomly chosen values for the mixed configurations. In this case, the original patterns are maintained, as long as the initial condition is restricted to pure local configuration states. If the initial conditions mix states, then there is a complex interaction between the two fluids and the patterns arising from the original CA.

This method produces CA rules which have unstable equilibrium between the two coexisting fluids as is shown in Fig. 12.

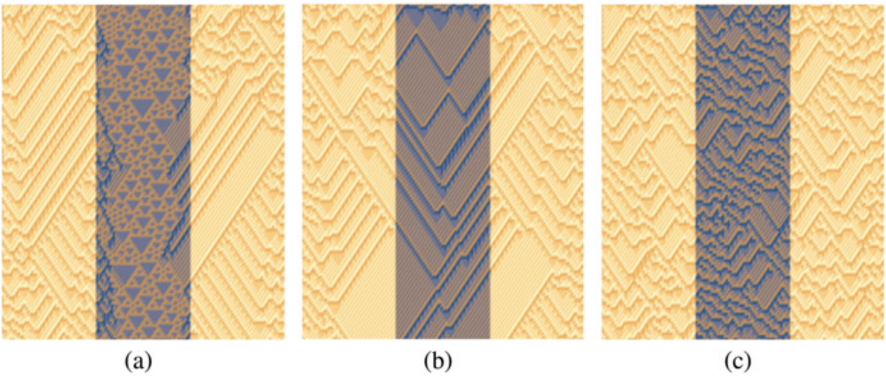


Fig. 11 (**a**) Perturbation of the canonical assembly of rule on the local configuration 333 as in the figure (in this case local configuration 000). (**b**) Perturbation of the canonical assembly of rule on the local configuration 544 as in the figure (in this case local configuration 211). (**c**) Perturbation of the canonical assembly of rule on the local configuration 554 as in the figure (in this case local configuration 221

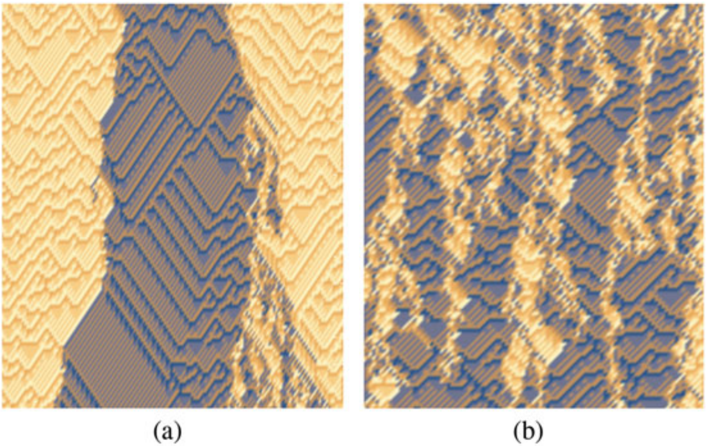


Fig. 12 (**a**) Realization for γ_4, with structured initial conditions: three segments randomly generated from {0, 1, 2}, from {3, 4, 5}, and again from {0, 1, 2}. (**b**) Realization for γ_5, random initial conditions

4 Conclusions and Further Developments

In the present paper techniques are developed for modeling idealized fluids where coexist distinct substances in different phases. These techniques, based on cellular automata, are appropriate to simulate transient and non-equilibrium behavior in fluid mechanics. The main result is the development of the canonical assembly method which allows the determination of CA code rules with complex behavior, obtained from given initial CA rules. The systems subject to assembly present an increasing number of distinct behavior and spatial-temporal patterns, maintaining, for certain initial conditions, the original patterns. Several families of CA, associated with idealized fluid substances, are considered. The canonical assembly method allows the study of small perturbations of a complex fluid and the study of the interaction between two similar fluids subject to instabilities. It is clear that the instabilities depend on the particular rules. The considered rules are sensitive to certain singular perturbations and not to other. The systematic study of the perturbations of these rules will be considered in future work, aiming a complete classification of its behavior.

Acknowledgments This work has been partially supported by Centro de Investigação em Matemática e AplicaçÕes da Universidade de É-vora (CIMA-UÉ), through the grand UIDB/04-674/2020 of FCT-Fundação para a Ciência e a Tecnologia, Portugal.

References

1. J. von Neumann, *Theory of Self-reproducing Automata* (University of Illinois Press, Urbana, 1966)
2. B. Chopard, A. Masselot, Cellular automata and lattice Boltzmann methods: a new approach to computational fluid dynamics and particle transport. Future Gener. Comput. Syst. **16**, 249–257 (1999)
3. G. Doolen, *Lattice Gas Methods For Partial Differential Equations* (CRC Press, Boca Raton, 2019)
4. D. Kristanto, W. Paradhita, Simulation study of fluid flow and estimation of a heterogeneous porous media properties using lattice gas automata method. J. Pet. Geotherm. Technol. **1**(2), 71–82 (2020)
5. C. Narteau, D. Zhang, O. Rozier, P. Claudin, Setting the length and time scales of a cellular automaton dune model from the analysis of superimposed bed forms. J. Geophys. Res. **114**, F03006 (2009)
6. T. Suzudo, Spatial pattern formation in asyncronous cellular automata with mass conservation. Phys. A **343**, 185–200 (2004)
7. G.Ch. Sirakoulisa, I. Karafyllidisb, Ch. Mizasa, V. Mardirisa, A. Thanailakisb, Ph. Tsalidesb, A cellular automaton model for the study of DNA sequence evolution. Comput. Biol. Med. **33**, 439–453 (2003)
8. T. Bossomaier, D. Green, *Complex Systems* (Cambridge University Press, Cambridge, 2000)
9. S. Wolfram, Statistical mechanics of cellular automata. Rev. Mod. Phys. **55**(3), 601–644 (1983)
10. S. Wolfram, *Cellular Automata and Complexity* (Addison-Wesley, New York, 1994)
11. H. Abarbanel et.al, *Cellular Automata and Parallel Processing for Practical Fluid-Dynamics Problems*, Report JSR-86-303 (1990)
12. C. Ramos, M. Riera, *Evolutionary Dynamics and the Generation of Cellular Automata*. Iteration theory (ECIT '08), Grazer Math. Ber., 354 (Institut für Mathematik, Karl-Franzens-Universität Graz, Graz, 2009), pp. 219–236
13. M. Mitchell, P. Hraber, J. Crutchfield, Revisiting the edge of chaos: evolving Cellular automata to perform computations. Complex Syst. **7**, 89–130 (1993)
14. M. Mitchell, P. Hraber, J. Crutchfield, Evolving Cellular automata to perform computations: mechanisms and impediments. Phys. D **75**, 361–391 (1994)

Part III
Experiments

Circular Causality and Function in Self-Organized Systems with Solid-Fluid Interactions

Benjamin De Bari and James A. Dixon

1 Introduction

The concept of self-organization has far-reaching applications, including non-equilibrium physics such as the Benard convections [1] and lasers [2], and even biological phenomena such as collective behavior [3] and motor control [4, 5]. One way of characterizing these self-organized phenomena is as the emergence of mutual constraint among constituent elements in a complex system. These constituents could be the light particles in a laser [2] or individual amoeba in a colony of bacteria [6]. In each case, the nonlinear interactions between constituents drive the emergence of macroscopic organization of the collective ensemble and possibly new properties of the system. Often these macroscopic dynamics exert an influence back on the microscopic dynamics, causing the constituents to become co-constrained and leading to a reduction in the total degrees of freedom of the system. For example, in the case of Benard convections, the emergence of convective rolls imposes constraints on the trajectories of the particles [1]. Self-organization in these systems is driven by dissipative entropy-producing processes, and they are thus called *dissipative structures*. Notably dissipative structures obtain in both living and

B. De Bari (✉)
Department of Psychology, Lehigh University, Bethlehem, PA, USA

Center for the Ecological Study of Perception and Action, University of Connecticut, Storrs, CT, USA
e-mail: benjamin.de_bari@uconn.edu

J. A. Dixon
Center for the Ecological Study of Perception and Action, University of Connecticut, Storrs, CT, USA

Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_11

nonliving systems and thus have been identified as a promising phenomenon for bridging the life and natural sciences.

We review two nonliving dissipative structures in which fluids play a pivotal role for supporting self-organization and the emergence of mutual constraint among solid elements. In each case, individual solid constituents are embedded in a fluid milieu (oil and water, respectively) and subject to non-equilibrium forcing (electrical and chemical, respectively). These solid elements have dynamics influenced by the fluid, and those dynamics also feedback and alter properties of the fluid. Because the solid elements all alter a shared fluid, the reciprocal interactions between an individual and the fluid lead to mutual constraint among *all* solid elements in the system and to the self-organization of structures and dynamics. A host of interesting phenomena are observed due to this mutual constraint, including formation of structures, oscillatory dynamics, coordination among multiple structures, and even emergent sensitivity to weak magnetic fields. While the two systems are quite distinct instantiations, we argue that the self-organized dynamics in each system derive from a shared circular causality that stems from the reciprocal solid-fluid interactions.

These nonlinear interactions are also subject to the thermodynamic contingencies within each system, as dissipative self-organization is maintained by entropy-producing processes. Researchers have, for some time now, been engaged in identifying overarching principles that predict the time-evolution of nonequilibrium systems. One candidate hypothesis is that nonequilibrium systems will evolve towards states (i.e., configurations, processes) that produce entropy at the fastest possible rate given the boundary conditions and constraints, often called the maximum entropy production principle (MEPP) [7–10]. We have found evidence that the dynamics of both dissipative structures discussed herein can be well explained by a MEPP. As we discuss, the joint influence of this MEPP and the nonlinear solid-fluid interactions leads to interesting and sometimes surprising dynamics in these systems.

2 Exemplary Dissipative Structures

2.1 The Electrical Self-Organized Foraging Implementation: E-SOFI

Our most studied system is an electrically driven dissipative structure. Metal beads sit in a 6-cm^2 dish with a shallow bath of oil. Approximately 5 cm above the dish, separated by an air gap, is a fixed-source electrode that delivers positive charges to the system. A metal ring surrounds the beads and is connected to a grounding electrode. Charges accumulate on the surface of the oil and on the beads, which become dipoles that are attracted to the grounding metal ring. After some time, the beads tend to spontaneously aggregate into branching strings of beads called “trees”

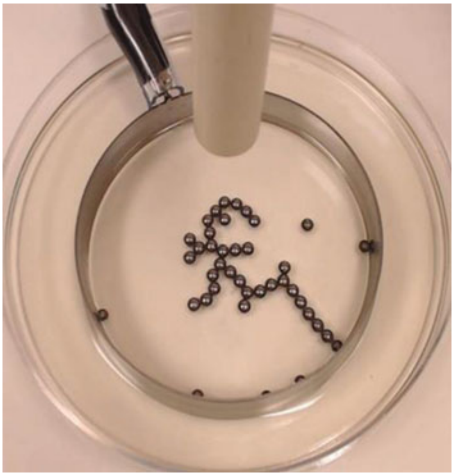


Fig. 1 Characteristic example of a tree-structure in the E-SOFI. The white column above the dish houses the source electrode. The grounding electrode is attached near the top-left portion of the metal ring. Trees like this one will tend to move around in the dish and change orientation and shape while the system is running

(Fig. 1). These trees maintain contact with the grounding ring, serving as pathways for the conduction of charge. The trees also exhibit dynamic motion, translating along the interior edge of the ring as well as flexing and swaying through the oil.

The primary dissipative process measured is the electrical current through the grounding ring. We calculate the rate of entropy production Σ as a function of the applied voltage *V*, electrical current *I*, and system temperature *T* according to

$$\Sigma = \frac{VI(x,t)}{T}$$

where current *I* is a function of time *t* and the position *x* of the tree. One key finding from studying this system is that the trees, and the system as a whole, appear to abide by a variational principle to maximize Σ [11–13], that is, to maximize the rate of entropy production. A wealth of evidence supports that the morphology of the structures and their dynamics tend to emerge such that Σ is maximized. We thus posit that the system is rudimentarily end-directed to maximize Σ. Crucially, the structures are also increasingly stable with greater Σ, and thus by maximizing the flow of charges, the trees are end-directed to maintain themselves. We have argued elsewhere that the end-directedness of dissipative structures may be analogous to the goal-directed behavior seen in biological systems [11–14]. Nonliving physical systems governed by variational principles tend to demonstrate equifinality, converging on a given end-state independent of initial conditions. For example, an isolated system with a small thermal gradient will evolve toward a state of maximal entropy. End-directedness may be generally construed as a system's evolution being determined by the optimization of a physical quantity, such as entropy, energy, or Σ. We have suggested that an end-directedness of this form is common to living and nonliving dissipative structures [11–14].

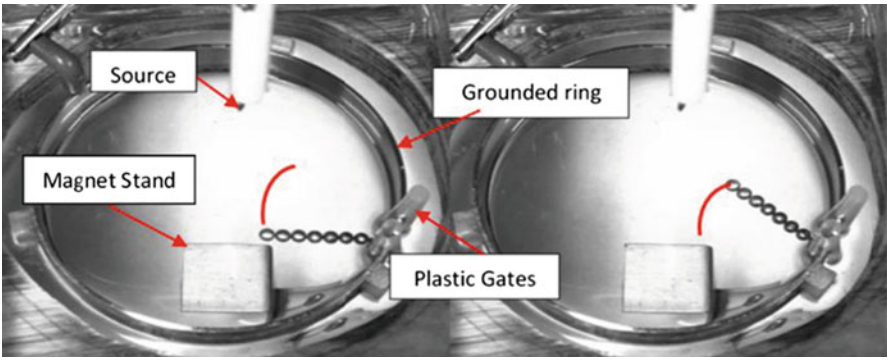


Fig. 2 E-SOFI setup for single-tree oscillations. The tip-bead trajectory is depicted as a red arc

The trees exhibit a variety of dynamics and are understood to move in order to collect charges and increase the rate of entropy production. The tree will tend to translate along the interior edge of the grounding ring, as well as swaying and bending the trunk and branches through the oil. Because these dynamics contribute to collecting charges, we consider them to be *foraging* behaviors, and thus we call the system the Electrical Self-Organized Foraging Implementation (E-SOFI). These foraging dynamics are robust. Even when a single-branched tree is fixed at the base so that it cannot translate along the interior edge of the ring, it will tend to oscillate, pivoting on its base bead (Fig. 2). The tip bead of the tree oscillates along a short arcing path, centered nearly on the minimum distance from the source electrode (Fig. 2). Interestingly, during this oscillatory cycle, the tip bead makes departures from the charge-rich region near the source electrode, and we even observe Σ *decreasing* during this time. This was initially mysterious; given that the system seems to be trying to maximize Σ, why would the tree move *away from* the source in a way that decreases Σ?

The answer lies in the interaction between the tree and the embedding fluid milieu. Charges accumulate on the oil surface, and the tree moves up increasing gradients of charge. The tree conducts charges to the ground, depleting their local concentration, while elsewhere charges accumulate. If we imagine a distribution of charges along the tip bead's arc trajectory, charges will tend to be depleted near the tree and accumulate elsewhere. The tree is thus continually reshaping the distribution of charges on the oil surface and changing the gradients that it follows. The charge distribution directs the tree's motion, while the motion of the tree in turn changes that same distribution. There is thus a reciprocal interaction between the solid (tree) and fluid (oil) elements of the system that leads to these self-organized dynamics.

This hypothesized interaction between the charge distribution and the tree was investigated with a computational model of the system [15]. The model represents the tip bead of a single tree moving along an approximately one-dimensional arc. Electrical charges are distributed across that one-dimensional space. The model is instantiated as a set of coupled differential equations representing the forces on the tip bead and the concentration of charges at each location in space (for details, see [15]). The model readily reproduces the dynamics of the tree, with the simulated bead oscillating around the virtual source electrode. In the model, it is evident that the bead is continually following increasing gradients of the charge distribution, always moving to more charge-rich regions. However, the tree's conduction of charges is continually reshaping that distribution, and the feedback between the distribution and the bead drives the emergence of the oscillatory cycle.

While these results did suggest that the trees are following increasing gradients and would thus tend to increase Σ, it remained a curious fact that the system would prefer this nonstationary oscillatory state in which entropy periodically decreases. To evaluate whether the oscillations produced greater rates of entropy production than steady states, we conducted a parallel set of empirical and simulated experiments [15]. Experiments focused on two conditions: (1) the tree (bead) was allowed to oscillate freely, and (2) the tree (bead) was fixed at a point that minimized the distance to the source electrode. The electrical current (proportional to the rate of entropy production) was measured and averaged within trials. Comparisons of these two conditions revealed that the average Σ was higher when the tree was able to oscillate than when it was fixed at a minimum distance from the source electrode [15]. This result held in both the physical and simulated systems. Thus, while Σ decreases within a given oscillatory cycle, the entire oscillatory process produces entropy at a faster rate than if the tree were static and Σ was stationary. Here we observed that a variational principle might direct this far-from-equilibrium system, but the complex reciprocal solid-fluid interactions drive nonobvious emergent dynamics to satisfy that end.

The interaction between the tree and charge distribution leads not only to the oscillatory and motive dynamics of a single tree but even enables coupling between multiple trees [16–18]. This coupling has remarkable dynamical and functional consequences for the trees' activities. To investigate some of these consequences, the system was expanded to have two separate grounding electrodes that could support separate trees. The current through each grounding electrode and individual tree was measured separately. These grounds were small metal brackets with insulating material on the sides of the front faces. The base bead of a tree was situated in-between these insulating constraints so the tree would form and only exhibit swaying dynamics without any translation on the grounding electrode (Fig. 3). Two grounds were situated approximately 5 cm from each other in parallel, with identical six-bead trees on each. The trees were thus constrained to share a region of the charge distribution.

Each tree individually demonstrated oscillatory dynamics like those observed in [15]. The two trees very quickly became synchronized, oscillating back and forth together. This synchronization occurs due to the joint modulation of the charge distribution. Each tree modifies the charge distribution, which modifies the forces on *both* trees. These reciprocal effects thus couple the trees and drive the emergence of a global (i.e., two-tree) steady state that here manifests as synchronization. The computational model was extended to include two beads in a one-dimensional

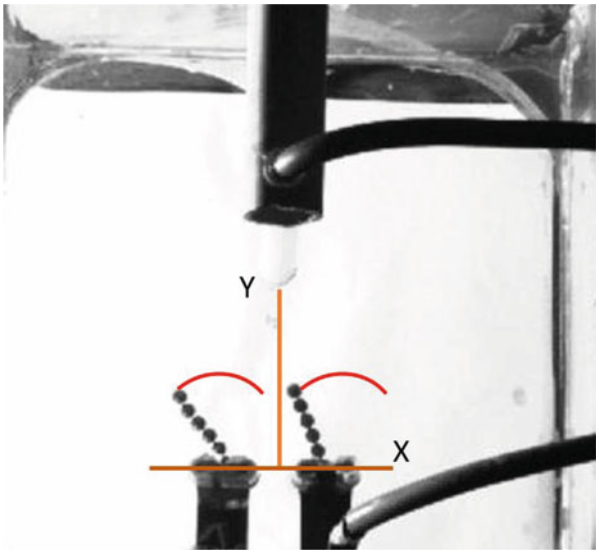


Fig. 3 E-SOFI setup for two-tree oscillation. X and Y coordinates used as frame of reference for image processing. Tip beads are largely constrained to motion in the X-dimension

space, analogous to the laboratory experiments. The same synchronization effect is observed in computer simulations, further supporting the charge distribution and fluid as a mediator for this emergent organization [17]. Synchronization is quite a common phenomenon in both biotic and abiotic oscillatory systems [19]. Here the synchronization emerges due to the reciprocal interactions between the solid and fluid elements in the system.

In addition to self-organized dynamical states, this fluid-mediated-coupling enables *functional* interactions between trees [18]. Recall that the system appears to abide by the variational principle to maximize Σ. This principle is analogous to a “goal” of the system, and we treat the system and the trees themselves as being rudimentarily end-directed to maximize Σ. Dynamics that contribute to increasing Σ thus are construed as *functional* in that they serve an implicit end for this system.

A pair of trees was again placed on separate grounding electrodes and allowed to oscillate. The tip bead of Tree 1 (Fig. 4a) was replaced with a magnetically sensitive chrome bead. All other beads in both trees were composed of nonmagnetic aluminum. A magnet on a moveable arm was positioned below the dish. It was initially placed far below the dish, exerting negligible force on the chrome bead such that Tree 1’s dynamics were entirely unaffected. The magnet could be mechanically raised so that it held the tip bead of Tree 1 in a fixed position. When the tip bead of Tree 1 was locked down by the magnet, it constrained the entire tree so that it was bent away from the charge-rich region of oil nearer the source electrode (Fig. 4b). We intentionally locked Tree 1 in a region that was further from the source electrode and therefore would not draw as much charge; under this manipulation the current

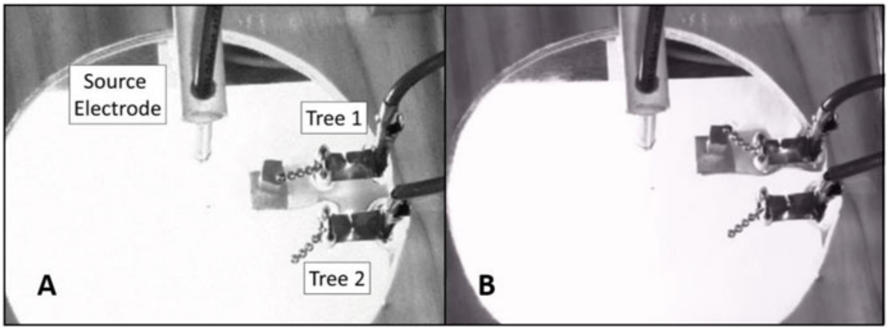


Fig. 4 (**a**) Pair of E-SOFI trees co-oscillating. Trees are approximately equidistant from the source electrode. This is the "Unlocked Phase" of trials. (**b**) The magnet has been raised near the dish to constrain Tree 1, locking it away from the charge-rich region between the two trees. Tree 2 oscillates freely. This is the "Locked Phase" of trials

measured from its grounding electrode decreases. The magnetic constraint is thus a functional perturbation to Tree 1 and to the system, as it decreases the current and consequently the rate of entropy production.

Experiments were conducted as two 10-min phases: (1) both trees oscillated freely for 10 min (the Unlocked phase), and (2) Tree 1 was magnetically constrained while Tree 2 oscillated freely (the Locked phase). The electrical current through each grounding electrode was collected, as well as the position of the tip bead of each tree, during all phases of the trial. Data from the Unlocked and Locked phases were compared to evaluate if Tree 2's dynamics and current change due to the functional perturbation to Tree 1.

It was observed that when Tree 1 is constrained, the current it conducts is dramatically decreased compared with the Unlocked phase. The magnetic constraint thus worked as a functional impairment. Crucially, the current conducted by Tree 2 *increased* in the Locked phase, indicating that it was in some way compensating for the loss of entropy production by Tree 1 [18]. This change in Tree 2's current was accompanied by a change in its motive dynamics. Tree 2's oscillation amplitude was averaged within each trial phase. Comparing between Unlocked and Locked phases, we observed that Tree 2's oscillation amplitude increased during the Locked Phase [18]. Together, the results suggest that Tree 2 exhibited a change in dynamics that compensated for the functional impairment to Tree 1. Here we see that the reciprocal effects mediated by solid-fluid interactions enable *functional* coupling between elements in the system. The two trees appear to be *coordinated* in maintaining Σ by maintaining the electrical current, with their dynamics entangled due to the reciprocal interactions.

2.2 *The Chemical Self-Organized Foraging Implementation: C-SOFI*

Another remarkable system we have investigated is a chemical dissipative structure called the Chemical-Self Organized Foraging Implementation (C-SOFI), a simple system that displays quite complex dynamics [20, 21, 27]. Thin, fragmented pellets of benzoquinone (BQ) float at the air-water interface in a small petri dish. These pellets dissolve into the aqueous milieu, thus altering the surface-tension gradients on the water surface. These surface-tension gradients pull the pellets across the water while the pellets continue to dissolve. There is a reciprocal interaction between the pellet and the aqueous environment, reminiscent of that discussed in the E-SOFI. The pellets dissolve and alter the surface-tension gradients, and the surface-tension gradients move the pellets and alter how and where they dissolve. The interaction between a single pellet and the aqueous milieu again leads to fluid-mediated coupling between the pellets; one pellet's alteration of the surface tension field changes the forces on *all* pellets embedded in that field. When there are a large number (i.e., more than 12) of pellets, these reciprocal effects drive the pellets to aggregate, forming into a dynamic collective that tends to move through the dish as a single entity (Fig. 5). This "flock," as it is called, is a dissipative structure emerging from the entropy-producing dissolution and motive processes.

Interestingly, the emergence of a flock depends on the geometry of the pellets. Irregularly shaped pellet fragments, like those in Fig. 5, will readily form flocks. However, circularly shaped pellets of approximately the same average size will not. While the details of this phenomenon have not been fully explained, we speculate that the nature of the reciprocal effects between the pellet and the aqueous

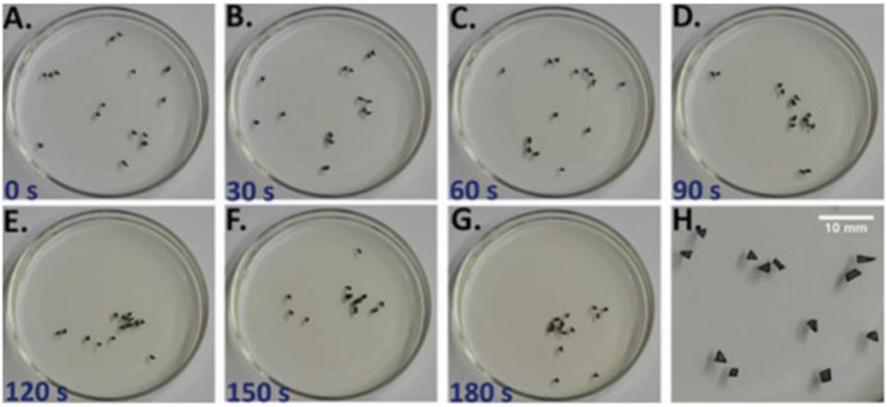


Fig. 5 Selected frames from a video documenting the emergence of a flock in the C-SOFI. After about 90 s, the flock has formed. The flock varies in organization and particle number, while retaining a tendency to aggregate. (Reprinted with permission from Ref. [21]. Copyright 2022, American Chemical Society)

environment change with different pellet shapes, leading to different coupling relationships and consequently either flocking or not flocking. Some evidence suggests that a thermodynamic variational principle may be at work. For example, estimates of the entropy production in flocking irregular pellets and nonflocking circular pellets demonstrated that the flocking pellets produce more entropy than the nonflocking counterparts [21]. It is thus possible that the flock, a dissipative structure, emerges as an opportunistic pathway for dissipation, consistent with the maximum entropy production principle (MEPP). Computational modeling of the system investigated the stability and likelihood of different flock sizes for each of the irregular and circular systems [22]. The model, derived from the empirical data, reproduces the tendency for irregular particles to produce large-pellet-number flocks and the tendency for the circular pellets not to flock. Moreover, thermodynamic analysis of the Gibbs energy of different flock sizes revealed that the system's preferred (i.e., most stable) state was also the minimum Gibbs energy state [22]. This suggests that a Minimum Gibbs Energy Principle may be driving the emergence of flocks of different sizes and accounts for the dynamics of both irregular and circular pellets.

One of the most surprising phenomena displayed by the C-SOFI is a self-organized sensitivity to weak magnetic fields [21]. To demonstrate this, a single irregular pellet was created with some embedded ferrous material, making it magnetically sensitive. This pellet is referred to as the "sensor" pellet. The sensor pellet is placed alone in the dish where it swims across the water surface. A magnet is positioned above the dish, and its height is adjusted such that it has only a very weak interaction with the sensor pellet. At this height, the pellet is slightly biased by the magnet but is not completely captured and still swims throughout the entire dish. In a subsequent experiment, a single sensor pellet is placed into the dish with 14 other nonferrous irregular pellets. These pellets form a flock, incorporating the sensor. When the magnet is positioned above the dish, at the height previously shown to be too great to constrain the single sensor pellet, the *entire flock* moves under the magnet and remains there (Fig. 6). While the magnetic field (at the set height) was not able to capture the single sensor pellet, the entire flock nevertheless is constrained by interaction between the sensor and the magnet. Control experiments revealed that the magnet does not constrain a flock of irregular pellets with no sensor, nor are circular pellets with a circular sensor among them constrained by the magnet [21].

We are currently pursuing a full explanation of these phenomena, but we speculate again that the fluid-mediated coupling among particles is playing a crucial role. It has been documented that some nonequilibrium systems will demonstrate emergent sensitivity to weak energy fields through cooperative self-organizing processes [23–26]. In such systems, the very weak microscopic interaction between the system and field is amplified by the self-organizing processes and biases the macroscopic activity of the system. Similarly, we hypothesize that in the C-SOFI the reciprocal coupling between pellets leads to self-organizing processes that amplify the effect of the magnet on the sensor pellet, biasing the entire flock to orient relative to the magnet.

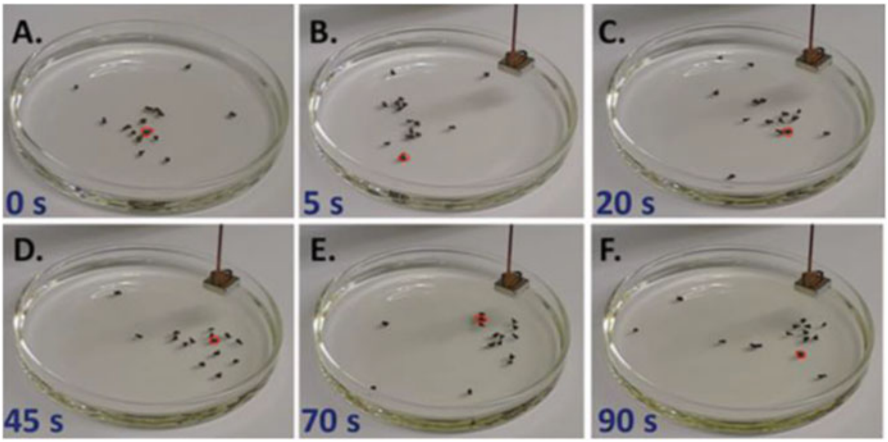


Fig. 6 Selected frames from a video documenting the magnet sensitivity of a C-SOFI flock. The sensor pellet is highlighted in red. After about 20 s, the flock has moved under the magnet and remains in the area for the duration of the trial. (Reprinted with permission from Ref. [21]. Copyright 2022, American Chemical Society)

One of the clearest demonstrations of the feedback between the pellets' dissolution and the properties of the aqueous milieu is detailed in this final experiment from Chen and colleagues [27]. A thin hydrophobic plastic divider is placed in the petri dish, bisecting it into roughly two equal compartments of water. The divider has a small 3-cm gap – referred to as a "gate" – through which the water surface extends and bridges the two baths. When a batch of irregular pellets is placed in the dish on one side, they initially spread out and distribute themselves throughout both compartments. Over time, they tend to aggregate near the center of the dish under the gate. Shortly after, the flock, now consisting of nearly all particles in the dish, selects one compartment of the dish to move into, breaking the symmetry of the system. In some trials, the flock will later make a subsequent transition to the other side.

Discrete samples of the surface tension on either side of the dish reveal that just before the initial symmetry-breaking event, there is a large difference in the surface tension on either side of the dish [27]. For example, in Fig. 7 the flock ultimately selects side B. Immediately prior to this transition, the surface tension on side B was much higher than on side A. It was repeatedly observed that the flock tended to make a transition to whichever side has greater surface tension [27]. The imbalance in surface tension emerges in part due to the dissolution of the pellets, as they decrease the surface tension locally. Thus, the flocks modulate the properties of the aqueous environment, which in turn feedback to constrain the activity of the flock. In trials where repeated transitions were observed (i.e., the flock moved from A to B then back to A), this occurred due to repeated flipping of the relative surface tension on either side. The flock lowered the surface tension on one side, made a transition to the other side, and then lowered the surface tension there until an imbalance

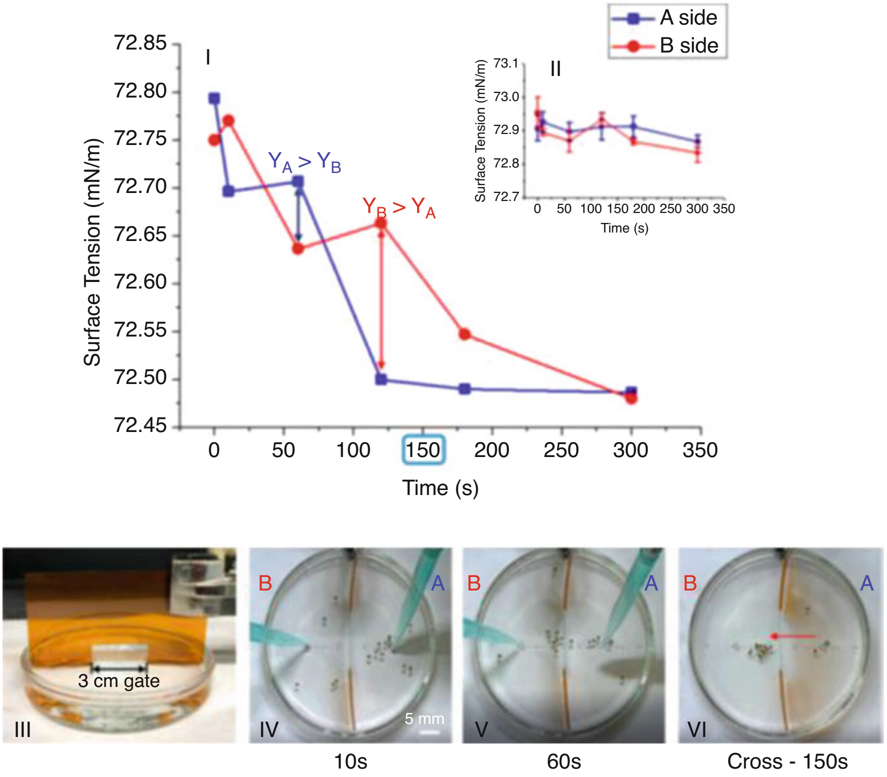


Fig. 7 (I) Samples of the surface tension on either side of the dish. (II) Samples of surface tension from a control trial with no pellets. (III) Image of the gate. (IV–VI) Selected frames demonstrating the flock's transition from side A to B. (Reprinted with permission from Ref. [27]. Copyright 2022, American Chemical Society)

prompted the transition. While experiments could not be sustained long enough to observe long timescale oscillation, the reciprocal feedback between the structure and environment leading to oscillatory symmetry breaking is strikingly like that observed in the E-SOFI. Again, we observe a variety of striking emergent dynamics stemming from complex reciprocal solid-fluid interactions.

3 Agent-Environment Reciprocities as a General Framework for Self-Organization

Herein we have reviewed some interesting dynamics in nonequilibrium systems that all appear to derive from the dissipative processes and reciprocal interactions between solid elements in a fluid milieu. We generalize this scheme by thinking of these systems in terms of agent-environment reciprocities. In each system, the agent is the dissipative structure, while the environment is the embedding fluid and its properties. More specifically, in the E-SOFI the agent is the bead structure while the environment is the oil and the charge distribution. Similarly in the C-SOFI the agent is the flock and the environment is the aqueous milieu with its surface tension and concentration gradients. Notice that the system's behavior is best understood at the scale of the agent-environment system, not by appealing to one or another alone (i.e., the dynamics of the bead structure can't be explained by its own internal properties and processes alone).

We can observe analogues of the reciprocal interactions displayed by the E- and C-SOFIs in the case of a foraging bacterium. An amoeba, while navigating sugar gradients, is all-the-while consuming sugar, altering the embedding distribution much like the E-SOFI. An individual bacterium might not exert a dramatic effect on the sugar gradients, but colonies of bacteria have been shown to do so. For example, pioneering work by Adler [28] looked at the collective activity of *E. coli* confined to a narrow glass tube. The tube was full of sugar with initially uniform distribution. The *E. coli* colony was introduced to the tube at one end, where it began to eat the sugar. Over time, the entire colony moved along the length of the tube, consuming the sugar within. Recent modeling work demonstrated that the colony's consumption generates gradients of the sugar that then stimulate chemotaxis, leading to a feedback process that drives the colony along the length of the tube [29]. This consumption-induced motion is remarkably similar to that observed in the E-SOFI. The same modeling work identified that the collective chemotaxis observed is a dynamic that maximizes entropy production compared with other possible modes of propagation through the tube [29]. Similar modeling projects have further supported that bacterial foraging occurs due to reciprocal interactions between metabolism, consumption, and embedding distributions of metabolizable resources [30–32].

Such reciprocal effects are not limited to single-celled organisms. Researchers studying chemotaxis of eukaryotic cells in multicellular organisms demonstrated that some may be unable to individually sense and follow very shallow concentration gradients but will collectively detect and navigate those gradients [33]. The researchers identified that the cell-to-cell interactions are moderated by the local concentrations of chemoattractant around individual cells, leading to an emergent anisotropy. This emergent sensitivity to weak chemical gradients is compellingly similar to the emergent magneto-sensitivity demonstrated by the C-SOFI.

At yet a larger scale, some problems in motor control can be characterized in terms of the emergence of mutual constraint among coupled physiological elements. To grip an object and keep it in hand, one must coordinate the forces on the object from each finger. If one finger is perturbed such that it exerts less force, another finger may exert more force to compensate. This phenomenon is called "reciprocal compensation" and has been observed in the control and coordination of speech effectors [34], force production by fingers [35, 36], and even interpersonal coordination tasks [37, 38]. Such examples are strikingly like the coordinated dynamics exhibited by the E-SOFI, wherein a tree will change its dynamics to compensate for the perturbations to its partner, thereby maintaining the system's REP. Reciprocal compensation requires that physiological elements are coupled, and it is generally understood that this coupling is a complex process that includes neural activity, mechanical forces, and perceptual information. An enticing yet ambitious possibility is that we may use the E-SOFI as a minimal model of these broader examples of reciprocal compensation, with the reciprocal fluid-solid interactions as a minimal model of the more diverse instantiations of coupling in biology.

The language of agent-environment interactions comes from well-established theory within the field of ecological psychology [39–41]. An agent is very broadly any organism, and the environment is the pocket of the world that it is engaged with, with many salient features such as food sources, shelter, and other organisms. As an example, consider an amoeba such as *Escherichia coli* embedded in an aqueous medium with dissolved metabolizable sugars. *E. coli* are attracted to these food resources and will tend to swim up increasing concentration gradients of sugar [28, 42]. When these bacteria detect a sugar gradient, they will orient and rotate their flagella to propel themselves along the gradient and will continue swimming if the concentration of sugar is increasing. This is known as "running" behavior. If, alternatively, the bacteria detect that they are not following an increasing gradient (e.g., the gradient is zero or decreasing in the direction of travel), they will throw out their flagella at all angles and rotate them such that their trajectory is pseudo-random. This is known as the "tumbling" behavior. While running and tumbling are activities of the *E. coli* cell itself, they cannot be fully understood without reference to the embedding context. Running or tumbling is necessarily understood *relative to* properties of the embedding sugar gradients. Much like in the E- and C-SOFIs, the proper unit of analysis is the agent-environment system. Moreover, in each case the agent is a dissipative structure, whether the bead-structure, pellet flock, or amoeba, and thus a thermodynamics-based framework of explanation may be available for all such systems.

Here, we lay the foundations for such a framework. The variational MEPP is analogous to the end-directed nature of biological behavior; the present state of the system is constrained by a future state. In biology an organism may have an end, such as finding metabolizable resources, but its behavior is also constrained by its embedding context and its own capabilities. Behavior can be explained by three factors: (1) the agent and its properties, (2) the environment, and (3) the intentions or ends. Analogues of each of these factors can be identified in these nonliving dissipative structures. As discussed, the agent is the structure itself, and so its properties include its configuration, its motion, its electrical or chemical potential, and likely other factors. The environment of each system is most easily identified with the embedding milieu with such salient factors as charge density, aqueous concentration, and surface tension. Constraining the interaction of this reciprocal pair is the intention, which we identify with the variational MEPP.

The agent and environment interact dynamically in the context of the system-level end for entropy to be maximized. For example, a tree structure (agent) in the E-SOFI interacts with the charge distribution (environment) in nonlinear ways and selects an oscillatory dynamic (behavior) that maximizes entropy production (intention). If one of the three factors changes so does the behavior. We discovered that oscillations did not occur if the voltage was low, and consequently the charge density in the oil was lower [15]; a change in the environment while maintaining the same agent properties and intentional state leads to a different behavior. This is directly analogous to the transition between running and tumbling behavior in *E. coli* driven by the change in ambient concentration gradients. If you decrease the sugar gradient while maintaining the same properties of the bacterium and its goal to forage, that triggers a transition to a different behavioral mode. The whole range of the interesting life-like dynamics of the E- and C-SOFIs can similarly be cast in terms of this tripartite formalization.

4 Conclusions

We have aimed to illustrate that reciprocal interactions are essential to a host of self-organizing events in both living and nonliving dissipative systems. In the E- and C-SOFI, fluids played a key role in mediating those interactions. While the role of fluids was evaluated only qualitatively herein, we hope that subsequent work will more significantly evaluate the fluid dynamics in each system. In particular, we seek to find a mapping of the tripartite formalization of agent, environment, and intention in the theory of fluid dynamics. It is possible that mathematical features of the fluid dynamics can be generalized to use as a common framework for all the self-organized phenomena outlined herein and more.

References

1. G. Nicolis, Physics of far-from equilibrium systems and self-organization, in *New Physics*, ed. by P. Davies, (Cambridge Press, Cambridge, UK, 1989)
2. H. Haken, *The Science of Structure: Synergetics* (Deutsche Verlags-Anstalt, Stuttgart, 1981)
3. I.D. Couzin, J. Krause, Self-organization and collective behavior in vertebrates. Adv Study Behav ***32***, 10–1016 (2003)
4. P.N. Kugler, J.A.S. Kelso, M.T. Turvey, On the concept of coordinative structures as dissipative structures: I. theoretical lines of convergence, in *Tutorials in Motor Behavior*, ed. by G. E. Stelmach, J. Requin, (North-Holland Publishing Co., New York, 1980)
5. P.N. Kugler, M.T. Turvey, *Information, Natural Law, and the Self-Assembly of Rhythmic Movement* (L. Erlbaum Associates, Hillsdale, 1987)
6. H. Salman, A. Libchaber, A concentration-dependent switch in the bacterial response to temperature. Nat. Cell Biol. **9**, 1098–1100 (2007)
7. R.G. Endres, Entropy production selects nonequilibrium states in multistable systems. Sci. Rep. **7** (2018)
8. L.M. Martyushev, V.D. Seleznev, Maximum entropy production principle in physics, chemistry, and biology. Phys. Rep. **1**, 1–45 (2006)
9. R. Swenson, M.T. Turvey, Thermodynamic reasons for perception—Action cycles. Ecol. Psychol. **3**, 317–348 (1991)

10. R. Swenson, Emergent attractors and the law of maximum entropy production: Foundations to a theory of general evolution. Syst. Res. **6**, 187–197 (1989)
11. J.A. Dixon, D. Kondepudi, T.J. Davis, End-directedness and context in nonliving dissipative structures, in *Contextuality from Quantum Physics to Psychology*, (World Scientific, Singapore, 2016)
12. D. Kondepudi, B.A. Kay, J.A. Dixon, End-directed evolution and the emergence of energy-seeking behavior in a complex system. Phys. Rev. E **91** (2015)
13. D. Kondepudi, B.A. Kay, J.A. Dixon, Dissipative structures, machines, and organisms: a perspective. Chaos **27**(10) (2017)
14. J.A. Dixon, B. De Bari, D. Kondepudi, B.A. Kay, Anticipation and dissipative structures. Mind and Matter **19**, 167–187 (2022)
15. B. De Bari, J.A. Dixon, D. Kondepudi, B.A. Kay, Oscillatory dynamics of an electrically driven dissipative structure. PLoS One **14** (2019)
16. T.J. Davis, B.A. Kay, D. Kondepudi, J.A. Dixon, Spontaneous interentity coordination in a dissipative structure. Ecol. Psychol. **28**, 23–36 (2016)
17. B. De Bari, D. Kondepudi, B.A. Kay, J.A. Dixon, Collective dissipative structures, force-flow reciprocity, and the foundations of perception-action mutuality. Ecol. Psychol. (2020)
18. B. De Bari, A. Paxton, D. Kondepudi, B.A. Kay, J.A. Dixon, Functional interdependence in coupled dissipative structures: Physical foundations of biological coordination. Entropy **23** (2021)
19. A. Pikovsky, M. Rosenblum, K. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences* (Cambridge University Press, Cambridge, 2001)
20. J.E. Satterwhite-Warden, D.K. Kondepudi, J.A. Dixon, J.F. Rusling, Co-operative motion of multiple benzoquinone disks at the air–water interface. Physical Chemistry Chemical Physics **17** (2015)
21. J.E. Satterwhite-Warden, D.K. Kondepudi, J.A. Dixon, J.F. Rusling, Thermal- and magnetic-sensitive particle flocking motion at the air-water interface. J. Phys. Chem. B **123** (2019)
22. B. De Bari, J. Dixon, J. Pateras, J. Rusling, J. Satterwhite-Warden, A. Vaidya, A thermodynamic analysis of end-directed particle flocking in chemical systems. Commun. Nonlinear Sci. Numer. Simul. **106** (2022)
23. D.K. Kondepudi, G.W. Nelson, Weak neutral currents and the origin of biomolecular chirality. Nature **314**, 438–441 (1985)
24. D.K. Kondepudi, G.W. Nelson, Chiral symmetry breaking states and their sensitivity in nonequilibrium systems. Physica. A **125**, 465–496 (1984)
25. D.K. Kondepudi, I. Prigogine, Sensitivity of nonequilibrium systems. Physica. A **107**, 1–24 (1981)
26. D.K. Kondepudi, Sensitivity of chemical dissipative structures to external fields: Formation of propagating bands. Physica. A **115**, 552–566 (1982)
27. T. Chen, D.K. Kondepudi, J.A. Dixon, J.F. Rusling, Particle flock motion at air-water interface driven by interfacial free energy foraging. Langmuir **35**, 11066–11070 (2019)
28. J. Adler, Chemoreceptors in bacteria. Science **166**, 1588–1597 (1969)
29. P. Županović, M. Brumen, M. Jagodič, D. Juretić, Bacterial chemotaxis and entropy production. Philos. Trans. R. Soc., B **365**(1545) (2013)
30. M.D. Egbert, X.E. Barandiaran, E.A. Di Paolo, A minimal model of metabolism-based chemotaxis. PLoS Comput. Biol. **6**(12) (2010)
31. M. D. Egbert, Dissertation, University of Sussex, 2012
32. M.D. Egbert, Bacterial chemotaxis: introverted or extroverted? A comparison of the advantages and disadvantages of metabolism-based and metabolism-independent behavior using a computational model. PLoS One **8**(5) (2013)
33. B. Camley, J. Zimmermann, H. Levine, W. Rappel, Emergent collective chemotaxis without single-cell gradient sensing. Phys. Rev. Lett. **116**(9) (2016)
34. J.A.S. Kelso, B. Tuller, E. Vatikiotis-Bateson, C.A. Fowler, Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. J. Exp. Psychol. Hum. Percept. Perform **10**, 812–832 (1984)

35. M.L. Latash, Z.M. Li, V.M. Zatsiorsky, A principle of error compensation studied within a task of force production by a redundant set of fingers. Exp. Brain Res. **122**, 131–138 (1998)
36. M.L. Latash, J.F. Scholz, F. Danion, G. Schoner, Structure of motor variability in marginally redundant multi-finger force production tasks. Exp. Brain Res. **141**, 153–165 (2001)
37. D.P. Black, M.A. Riley, C.K. McCord, Synergies in intra- and interpersonal interlimb rhythmic coordination. Hum. Kinet. **11**, 348–373 (2007)
38. M.A. Riley, M. Richardson, K. Shockley, V.C. Ramenzoni, Interpersonal synergies. Front. Psychol. **2** (2011)
39. J.J. Gibson, L. Carmichael, *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston, 1966)
40. J.J. Gibson, *The Ecological Approach to Visual Perception* (Psychology Press, Sussex, 1986)
41. M.T. Turvey, R.E. Shaw, Ecological foundations of cognition. I: Symmetry and specificity of animal-environment systems. Int. J. Consum. Stud. **6**(11–12) (1999)
42. M.H. Bickhard, The interactivist model. Synthese **166** (2009)

Hydrokinetic Energy Harvesting Potential of Triangular Prims and Cross Cylinders

Rachmadian Wulandana and Fairooz Haque

1 Introduction

The undertaking project is motivated by extensive studies reported by Chung, Vaidya, et al. on vortex-induced autorotation of symmetric bodies [1–3]. In particular, the current study builds upon the characteristics of hinged short Delrin cylinders exposed to water flow confined in transparent observation chamber [3]. The tests were performed using a commercial closed loop flow tank furnished with a centrifugal pump capable of delivering maximum of 60 cm/s average flow speed. The vortex structure was visualized using hydrogen bubble, and the images were analyzed for quantitative results [4]. The Delrin cylindrical samples with its long axis perpendicular to the flow demonstrate four (4) distinct responses: stagnation or no motion, random oscillation, periodic oscillation, and autorotation. Parameters that dictate the motions include the flow speed and non-dimensional inertia, defined as $I^* = \frac{I}{(\rho_f d^5)}$, where the I, ρ_f, and d refer to moment of inertia, fluid density, and the diameter of the cylinder model, respectively. The autorotation for the short solid cylinder was indicated to occur at Reynolds numbers around 3000 and 4500 for bodies with I^* of approximately a little less than 0.20. The cylindrical bodies demonstrate predominantly random oscillations, and the oscillation frequency tends to linearly increase with Reynolds number. Bodies with Aspect Ratio (AR, ratio of length to diameter) of unity show larger propensity for rotation than that of bodies with AR of two.

R. Wulandana (✉) · F. Haque
SUNY New Paltz, New Paltz, NY, USA
e-mail: wulandar@newpaltz.edu; haquef1@hawkmail.newpaltz.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_12

The importance of upstream blockage on the autorotation of Delrin short cylinder was highlighted in our works [5]. The rotation frequency or rpm of the suspended Delrin cylinder was significantly increased when the incoming water flow was perturbed by an object. A video of such phenomenon can be viewed in this website [2]. The blockage also causes the autorotation to occur at lower flow speed than that without the obstacle. When cylinders with AR of less than and larger than unity were tested, it was found that the upstream obstacle increases the autorotation potential of these cylinders. Past experiments with such cylinders in normal flow did not show autorotation [3]. The upstream distance and size of the blockage from the cylinder also show effects on the rpm of the autorotation. The effect from the distance diminishes as the obstacle moves away from the cylinder. In the current paper, we will report effects on the upstream distance on the power production of our turbine models.

The utilization of upstream blockage to enhance autorotation is not novel. A similar asymmetric blocking technique was employed by Skews to enhance the vortex shedding behind the body and increase the rotational speed of polygonal prisms exposed to air flow [6, 7]. Armandei and Fernandes use a similar technique called buffeting for marine energy harvesting [8]. Here, a blunt object is placed directly in front of a power extracting device to generate an oscillating wake. The vibration of the device occurs when the frequency of the wake matches with the natural frequency of the device. Upstream deflectors have been extensively studied by researchers interested in the hydrokinetic energy harvesting. In particular, effects on the placement and geometry of deflectors on the performance of vertical axis turbines, such as Darrieus and Savonius types, have been investigated both experimentally and numerically. The deflector is a stationary thick plate placed upstream relative to the turbine and partially blocks the incoming flow. The reduction in the cross-section area increases the incoming flow speed that impinges into the advancing blades. Deflectors that make obtuse angle to the incoming flow guide the flow toward the advancing blades and prevent the flow to stream into the returning blades [9]. Experimental works by Zhang et al. found that deflectors that make 120 degree angle to the incoming flow best improve the power coefficient of a 2-bladed Savonius-like vertical axis turbine [9]. Golecha et al. studied the effects of deflector's angle relative to the flow direction and its position relative to the turbine on multi-stage Savonius turbines [10]. They concluded that if the deflector optimally placed relative to the turbine, it can double the power coefficient. Similar conclusion was drawn when a pair of deflectors was utilized [11]. Patel et al. conducted a similar test on a three-bladed Darrieus turbine using a deflector that is normal to the flow direction [12]. Patel and Patel further utilized a similar deflector concept to improve the performance of dual rotor Savonius turbine [13]. Jeeva et al. performed experimentation on a pair of angled deflectors placed upstream an inclined three-bladed Savonius turbine [14]. They concluded that the inclination of the shaft improved the power coefficient of the turbine. Mosbahi et al. designed and studied a combination of a deflector and narrowing channel to improve the performance of Savonius turbines with twisted blades (helical Savonius rotor) [15, 16]. Salleh et al. compared the power enhancement by upstream deflectors on the power coefficients of conventional Savonius turbines placed in air and water flows [17]. They concluded that the effects of the deflectors on the power production by the two fluids are very similar.

Computer simulation studies on such deflectors were performed to understand complex flow characteristics and to estimate improvement of torques and potential power output. Nimvari et al. studied the effects of placing a porous deflector on the performance of Savonius wind turbine, and they discovered that the flows through the porosity caused breakdown of the wake behind the deflector that leads to an incoming flow with fewer fluctuations [18]. Using computer simulation, Alizadeh et al. studied the effects of placing a simple barrier in front of Savonius turbine and concluded that the power can be increased by 18% at an optimal distance [19]. Mosbahi et al. performed three-dimensional numerical studies of hydrokinetic helical Savonius turbines with upstream deflectors and concluded an increase of 17.4% in power generation [16]. Pulijala and Singh [20] provided the computer model of the experiment setup by Golecha et al. [10] and found reasonable validation of the improvement in power coefficient. Kerikous and Thevenin conducted optimization studies for the shape and position of a thick deflector placed upstream a two-bladed Savonius turbine and concluded that the optimum configuration increased the power coefficient by 15% [21]. Patel and Patel recently performed numerical study to investigate the effects of diverging and converging deflectors on the performance of dual-rotor Savonius turbines [13]. Both experiment and computer simulation studies showed significant effects of the deflector on the turbine power output. The local increase of incoming flow velocities due to the partial reduction in the cross-section area of the channel provides extra momentum on the turbine blades. In addition to that, the blockage also prevents negative torque caused by the incoming flow on the returning blades.

Renewable energy is appealing since it is a clean and ecologically acceptable alternative to traditional power generation methods that may be employed in remote societies without causing major environmental degradation. A recent review on hydrokinetic energy harvesting technologies pointed out the safety and sustainability of hydrokinetic energy systems, particularly for applications in remote places that are difficult to reach via transmission lines [22]. The hydrokinetic system is attractive because it does not require massive and expensive infrastructure unlike the hydropower system. The power capacity from hydrokinetic is known to be small but that may be appropriate for local needs. The power production of hydrokinetic systems in remote areas has been investigated by numerous researchers from around the globe. In Indonesia, for example, the potential power production from micro hydro systems has been estimated to reach 144 MW [23]. Susilowati et al. argued that micro hydrokinetic plants in seven locations along the Mahakam river in East Kalimantan province of Indonesia can substitute the existing diesel-fueled generators operated by national government [24]. In Malaysia, the power generation from the hydrokinetic system is estimated to total about 500 MW [25]. Abundant river debris, limited technology, and low current speeds put challenges on the utilization of hydrokinetic energy systems and, hence, careful design and selection of turbine types are crucial [25–27]. A small hydrokinetic system powered by a

Pelton wheel turbine has been constructed as example in Sarawak, Malaysia [27]. The small plant was able to provide electricity for 15 families. In Thailand, the 77-km U-Tapao river was estimated to produce 72 MW of hydrokinetic power from its 38 sub-basin small rivers [28]. In Tunisia, Africa, specifically near the Hama city, an attempt to extract hydrokinetic energy from irrigation channels was facilitated using Savonius turbines with twisted blades [15, 16]. A similar attempt to install a 5-kW underwater axial turbine in an irrigation canal was reported in Northern Cape Province of South Africa [29]. In North Central Nigeria, the hydrokinetic potential of rivers in the Lower Niger Basin was estimated to be about 826 MW [30]. In India, the hydrokinetic power generation along the 195-km eastern Yamuna canal, situated in Saharanpur district of Uttar Pradesh, was estimated to reach 27 MW for an average of 2.5 m/s flow speed. In Brazil, a complex analysis utilizing river average speed, change of elevation, and river depth along with 58 possible sites of the plants estimates the hydrokinetic power generation along Amazon River in Brazil to be about 910 MW [31].

Based on the mechanism of the kinetic energy conversion, the hydrokinetic harvesting technologies can be classified as turbines, which constitute devices with rotary motions, and non-turbines, which constitute devices that do not have rotary motions [32, 33]. Turbines can be further classified according to the orientation of the rotation axis with respect to the water flow direction, such as axial, vertical, horizontal, and cross-flow turbines [22]. The axis of vertical turbines is perpendicular to the flow direction, and the rotational plane is parallel to the water surface. The vertical axis orientation allows the turbine to rotate despite of the flow direction; hence it is attractive for applications in oceans [34]. Interest of applications of vertical axis turbines, such as Savonius, Gorlov, and Darrieus turbines, for applications in rivers was found to be increasing [35]. The turbine category is populated by devices equipped with bladed rotors that rotate due to the combination of drag and lift forces by the water flow. On the other hand, the non-turbine category mostly constitutes devices that convert flow-induced motion (FIM) and oscillation as well as vortex-induced vibrations (VIV) of the harvesting objects exposed to water flow into useful electrical energy [22]. Various energy harvesting mechanisms stem from different modes of vortex-induced vibration (VIV) that can be classified as fluttering, galloping, vortex-induced vibration, and autorotation [36]. Summaries and reviews of various academic and industrial research on vortex-induced energy harvesting technologies were provided by Rostami and Armandei [37] and recently by Wang et al. [38]. While investigation in this area has been dominated by academic research and small-scale experimentation and computer modeling, several products have attracted industries for large-scale development. For example, Vortex Induced Vibration Aquatic Clean Energy (VIVACE) converts the vortex-induced lateral oscillation of a bar exposed to water flow into electricity by means of electromagnetic induction [39, 40]. On the other hand, Festo developed DualWing Generator that exploits the flow-induced flutter mechanism by means of a pair of NACA 0014 wings that oscillate synchronously in the opposite directions [41]. The generator by Vortex Bladeless [42], a company based in Spain, exploits the vortex-induced vibration of a blunt cantilever body exposed to wind [43]. The lateral vibration of the pole is converted to electricity by means of a patented alternator technology.

In this project, three distinct vortex-induced autorotating bladeless turbine designs are evaluated for their performance in terms of energy generation. Among the many modes of vortex-induced vibration previously discussed, the autorotation mode has received the least attention for energy harvesting. The Vertical Axis Autorotating Current Turbine or VAACT, which constitutes a rectangular plate hinged at its symmetric axis [44, 45], perhaps serves the best example for an energy harvester that exploits the rotating mode of vortex-induced motion. Such autorotation of symmetric rectangular plate under air flow was studied by Skews [46]. The work, however, was not aimed for power generation. One of the two triangular designs presented in this paper has straight sides while the other has curved sides. The triangular prism designs are selected in this study due to its well-studied autorotation characteristics when they are exposed to air flow [6, 7]. The curved sides of our design are expected to increase surface area needed to generate torque fluid shear stresses. In terms of energy harvesting, an early work by Vaidya et al. estimated possible power production from vortex-induced oscillation of short cylinders [47]. Recently, our collaborative work with Chung and Vaidya reported the autorotation potentials of various 3D-printed symmetric bodies under water flow [5, 48]. The work has discovered that Cross Cylinder turbine models demonstrated autorotation and power generation. This innovative model resembles a merge of two short cylinders; each possesses an AR of unity, in orthogonal manner. Other models studied in this experiment that did not show any rotation can be viewed as combinations of symmetric bodies such as rectangular prisms, cubes, polygonal prisms, star, ellipsoids, etc. Our works on the 3D-printed Cross Cylinder models also revealed the minimum effects of turbine density on the power generation and rotation-per-minute (rpm) of the turbine and the significant effects of upstream obstacle on the power production [48].

In this paper, we will first discuss the turbine model, specification of the flow tank, and tools utilized in this project, delving into the specifics of the operating conditions and methods for experiment. Results on the effects of upstream obstacle and tandem arrangement on the performance of the Cross Cylinder model will be discussed. Moreover, the effectiveness of upstream obstacle on the rotation of such vortex-induced turbines is investigated using triangular prism models.

2 Methods

The power generation potential of three 3D-printed bladeless turbine models (shown in Fig. 1), triangular prism with straight sides (panel a), triangular prism with curved sides (panel b), and Cross Cylinder turbine model (panel c), is investigated using a custom-made water flow tank shown in Fig. 2 panel a. During the experiment, the turbine model is placed in the middle between the two walls of the observation chamber and is exposed to either free stream or perturbed flow caused by an

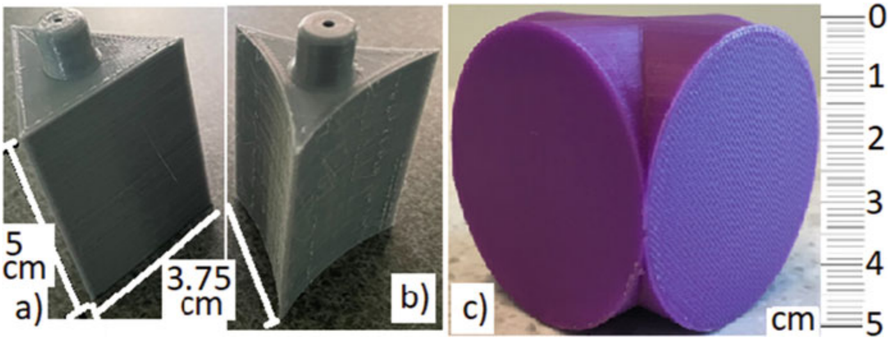


Fig. 1 3D-printed turbine models used in this project: (**a**) a triangular prism with straight sides, (**b**) a triangular prism with curved sides, and (**c**) a Cross Cylinder model

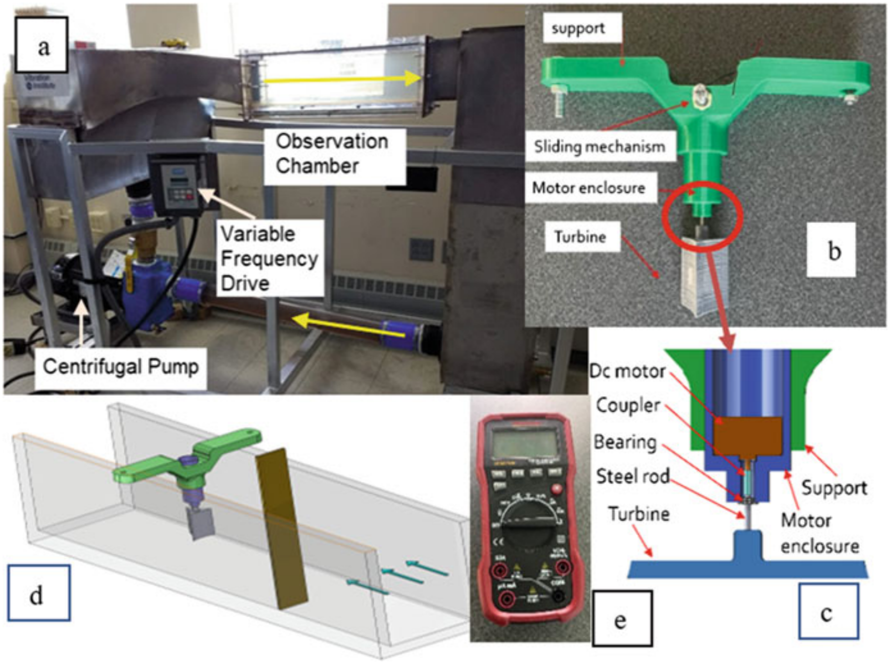


Fig. 2 Tools utilized in the project: (**a**) the custom-made open water flume furnished with centrifugal pump, variable frequency drive, and observation chamber. The yellow arrows indicate flow direction. The telescoping mechanism for turbine is shown in (**b**), and figure (**c**) shows the detail of the connection between the turbine and motor. (**d**) shows the schematic when the turbine is exposed to upstream blockage (the arrows indicate flow direction), and (**e**) shows the digital multimeter used in the experiment

Table 1 Properties of bladeless turbine models used in this project

Turbine type	Mass (gram)	Volume (cm^3)	Side surface area (cm^2)	Moment of inertia ($g.mm^2$)
Cross Cylinder	50.73	113.69	7854	57593
Triangular prism with straight sides	13.65	33.60	5850	1700
Triangular prism with curved sides	12.7	21.90	5941	1220

obstacle. The model was placed about 7 or 8 cm below the water surface that occupies a little less than 15 cm of chamber depth. The width and length of the transparent observation chamber are 15 and 60 cm, respectively. The panel d of Fig. 2 shows the sketch of a turbine model exposed to such perturbed flow. The asymmetric stream obstacle is provided by a 5-cm wide wooden ruler placed perpendicular to the flow direction at either 10-cm, for close obstacle, or 20-cm, for far obstacle, in front of the single turbine. In addition to the perturbed flow mentioned above, the Cross Cylinder model is also tested under tandem configuration. In the tandem arrangement, another identical Cross Cylinder is placed approximately 10-cm behind the front turbine. The perturbed flow is not applied to the tandem arrangement due to length limitation of the observation channel. In the current report, only the power production from the main front turbine in the tandem configuration will be reported. The second turbine was particularly "blocked" by the main turbine, and there was very minimal rotational motion and power that can be observed. Both triangular prism turbines can be considered as extruded planar triangular shapes. One of them with straight sides, while the other with curved sides. The heights of triangular turbines are 5 cm and the corner-to-corner distance is about 3 cm. The Cross Cylinder model represents a merge of two short cylinders; each has equal diameter and length of 5 cm. These turbines were made of Polylactic Acid (PLA) with 20% infill printing parameter (Table 1).

The flow tank is equipped with a 3-hp centrifugal pump capable of delivering up to 60 cm/s of average water speed in the observation chamber. The experimentation and observation of the turbine performance are made available through the $15 \times 15 \times 60\,cm^3$ transparent chamber made of 1/2″ thick plexiglass. The Variable Frequency Drive (FVD) allows the water flow to be controlled either manually or automatically at frequencies ranging from 20 Hz to 60 Hz. The maximum average flow speed that can be achieved is approximately 60 cm/s. A converging chamber was designed at the entrance of the observation chamber to reduce the complexity of the flow and to create uniform flow. The custom made closed-loop flow tank was sponsored by the Vibration Institute, and it was constructed as a senior design project within the Division of Engineering of SUNY New Paltz [49]. A 3D-printed hollow cylindrical casing is designed to hold and waterproof a 0.5-V DC motor that is used to generate continuous voltage and current data for this study. The wire connections from the DC motor were connected to a Dawson Digital Multimeter shown in Fig. 2 panel d. The digital multimeter is equipped with a USB connection and a data acquisition software that allows for the collected date to be processed using Excel. The multimeter and software record measurements at a sampling rate of 3 Hz. Each measurement set was taken over 180 data points culminating in a total of approximately 60 s. Because the multimeter cannot measure voltage, current, and power generated at the same time, tests conducted with this multimeter first measured voltage, then switched to measure current while the flow tank continued to run. The collected data was imported into Microsoft Excel for further process.

Stainless steel shafts of 2-mm in diameter are press-fitted into the turbine plastic models. The coupling between the metal shaft and the motor shaft is facilitated by a 2-cm-long rigid plastic tubing with 2-mm internal diameter. The shaft connection is stabilized using 2-mm ball bearings properly secured at the base of the casings. The waterproof 3D-printed casing for the DC generator allows the DC motor to be lowered into the water stream by means of telescoping mechanism. The vertical axis orientation is preferred than the horizontal due to its practicality for the installation of the DC motor. This results in a short connecting metal coupling shaft that is less prone to large bending caused by the water flow pushing the turbine models. Figure 2 panel b displays the 3D-printed suspension frame with its telescoping mechanism that allows the turbine casing to be lowered into the water. Panel c depicts various components involved in the coupling between the turbine model and the DC motor.

The turbine performance will be measured using Average Power production, possible Maximum Power production, Efficiency, Number of "Flips," and Maximum Rotation Time (Table 2). The Average Power is a product of the time-averaged of the current and voltage absolute data. On the other hand, the possible Maximum Power is a product of the maximum values of the current and voltage absolute data. The Efficiency of the turbine is defined as the ratio of the Average Power to the possible Maximum Power. The recorded current and voltage continuous data demonstrate changes of signs from positive to negative due to the oscillation and change of rotational movement. The total numbers of sign changes from the current and voltage data during the 60-s observation period are combined, and this will be reported as Number of Flips. On the other hand, the data collection also allows the counting of time period in between two consecutive flips, when the turbine would rotate in a single direction. The longest time in between these two sign changes of each current and voltage data is combined and reported here as Maximum Rotation Time. The unit used for Maximum Rotation Time is second.

Table 2 Definition of output by the turbine used for analysis

Term	Definition
Average Power	Product of the average voltage and current absolute data
Maximum Power	Product of the maximum voltage and current absolute data
Efficiency	Ratio between the Average Power and Maximum Power
Number of Flips	Total number of sign changes of the current and voltage data during the observation period
Maximum Rotation Time	The sum of the longest times between two consecutive sign changes in the current and voltage data

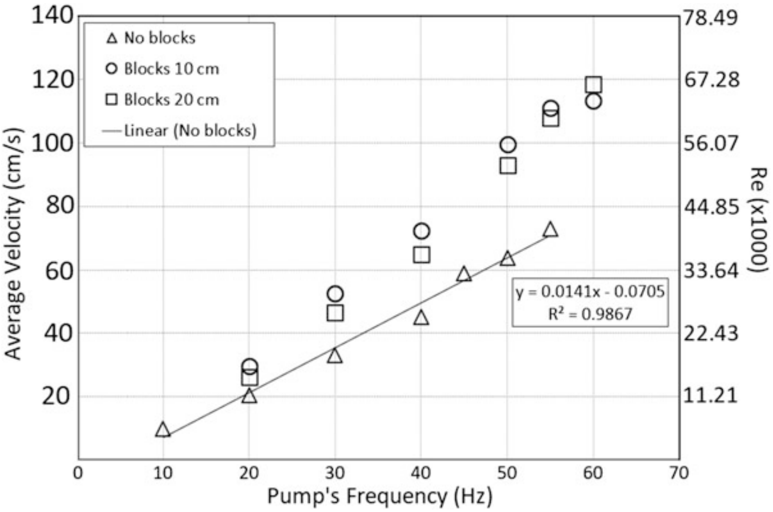


Fig. 3 This figure shows the linear equation shows the relationship between the pump frequency (x axis in Hz) and the average speed (y axis in m/s) for the case of normal flow without obstacles and flows with close and far upstream obstacles

Prior to collecting the power data, the mean velocity of the water stream in the observation channel was verified using a propeller flow meter from Vernier [50]. The flow meter was placed in the middle between the two walls of the observation chamber, which would be the location of the turbine. The measurement was performed for the free stream condition and the perturbed conditions with close (10-cm upstream) and far (20-cm upstream) obstacle. The speed data was collected for 30 s, and the time-averaged mean velocity was calculated for a range of pump's frequencies from 10 to 60 Hz. The outcomes for the free stream and perturbed flows are presented in Fig. 3. The mean velocity data for free stream are presented using triangular markers. A linear regression of the data revealed useful relationship between the pump's frequency (x, in Hz) and estimated mean velocity (y, in m/s) in the observation channel:

$$y = 0.0141x - 0.0705 \tag{1}$$

This linear relationship is comparable to one obtained previously [49]. In this graph, the circle and square markers represent the mean velocities detected by the flow meter for the perturbed cases due to close and far obstacles, respectively. This data represents local speed that would be experienced by the turbine due to the presence of upstream partial blockage. The consistent data among the two cases suggests that the obstacle certainly increases the local speed at the turbine's location, but its distances from the location do not have effect on the local speed. The increase in speed is expected due to the decrease in channel width caused by the upstream blockage.

In this paper, the experiment results will be presented with respect to the Reynolds numbers (Re) defined below:

$$Re = \frac{\rho V d}{\mu} \tag{2}$$

In this formulation, the V is mean velocity of the main incoming stream, not local velocity experienced by the turbine. This mean velocity is estimated from the pump frequency using Eq. 1 described above. The characteristic length, d, for the triangular prism models is taken as the edge-to-edge distance of the triangular shapes (3.75 cm). For the Cross Cylinder model, this length is taken as the diameter of the cylinder (5 cm) that defines the model. The density, $\rho = 998\,\mathrm{kg/m^3}$, and viscosity, $\mu = 0.89\,\mathrm{cP}$, are for the water taken at 20 °C.

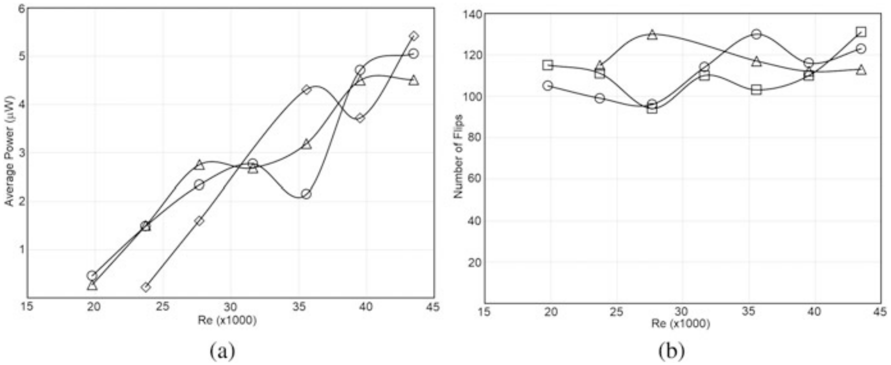
3 Results

Results presented in this report are organized as follows:

- Data comparison from the Cross Cylinder turbine model
- Results on the performance of the Cross Cylinder turbine model
- Results on the performance of the triangular prism with straight sides
- Results on the performance of the triangular prism with curved sides

3.1 Data Comparison from the Cross Cylinder Turbine Model

In this section, it will be first shown the consistency of data among a series of same experiments. Figure 4 panels a and b display three (3) data sets of Average Power and Number of Flips, respectively, from a single Cross Cylinder turbine model exposed to obstacle-free water stream. Figure 4 panel 1 demonstrates a consistent linear relationship between the Average Power, in the y axis, versus the Reynolds number, in the x axis, among all three data sets. The average power production by the Cross Cylinder turbine here is slightly higher (less than double) than that reported previously [5, 48] that were calculated from discrete, instead of continuous, measurement of current and voltage. The linear relationship between the power production and flow speed is duplicated here, but for larger range of Reynolds numbers. The discrepancy between the current and past data can be caused by many sources, particularly the differences in friction of the motor shaft and physical resistance of the motor bearing used in these two experiments. The current work also revealed consistent linear increase of the Maximum Power with respect to the increase in flow speeds for the three (3) data sets. This data is not displayed, but, subsequently, the calculated Efficiency of the Cross Cylinder exposed to free stream is approximately constant, about 5%.



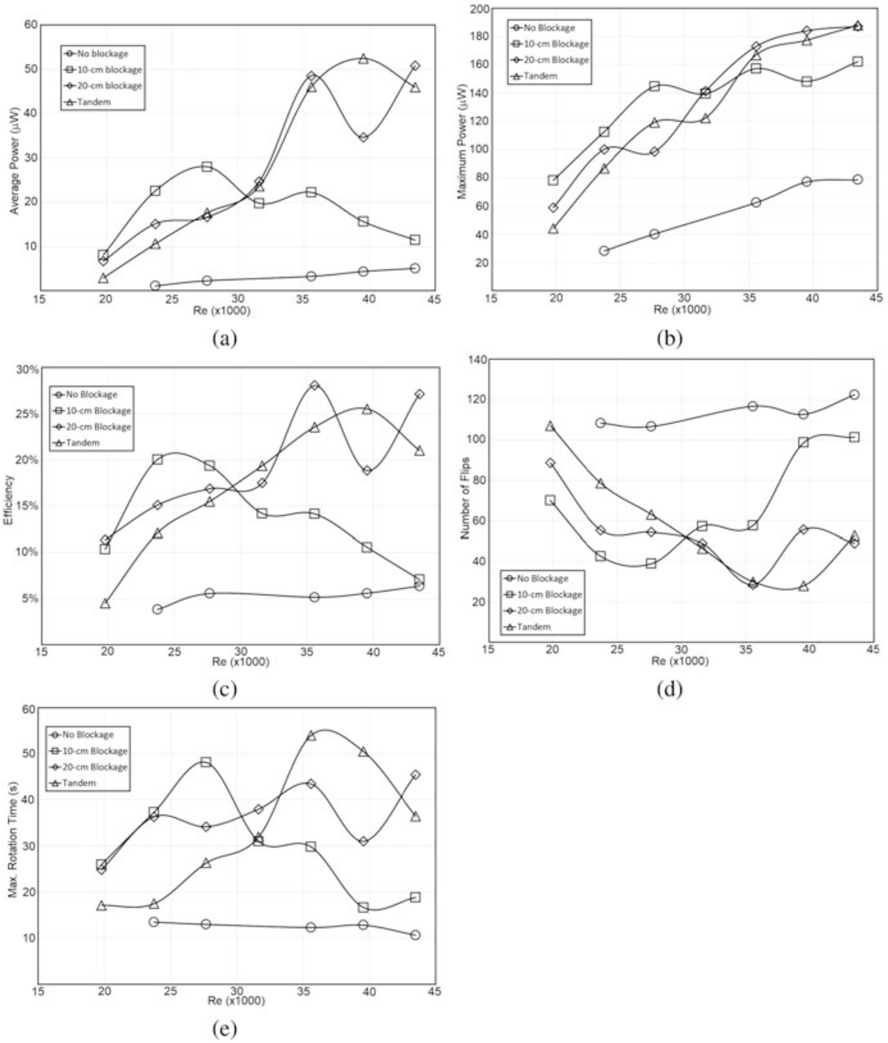
(a) (b)

Fig. 4 The two panels are data from the Cross Cylinder turbine exposed to free stream at various flow speeds. Panel (**a**) shows Average Power generation versus Reynolds number (×1000), and the data indicates linear increase in power with respect to the increase in flow speed. Panel (**b**) shows the Number of Flips versus Reynolds number (×1000). Here, data indicates unchanged number of flips with increasing flow speed for all three data sets. The two panels show that the data are consistent within the three separate experiments

Figure 4 panel b displays the Number of Flips for increasing Reynolds number from three (3) sets of data. The graph shows no consistent trend of Number of Flips with respect to the increase of flow speeds. The same inconsistency was shown in our previous work on a similar Cross Cylinder model but having high shaft friction. Nevertheless, while no specific tendency is shown, the data variation indicates that the Number of Flips is relatively the same for the given range of Reynolds numbers. Similar trend to this is also observed for the Maximum Rotation Time given by the three (3) sets of experiment data. Data on maximum rotation time will be displayed later in comparison with other experiment data.

3.2 *Results from the Cross Cylinder Turbine Models*

The effects of the blockage distance and tandem configuration on the Cross Cylinder turbine's performance will be presented in the following sections. The results represent average data from three (3) separate experiments. Figure 5 panels a to e show the experiment results for the Cross Cylinder turbine model exposed to free stream (circle marker), perturbed flow due to close (10-cm) obstacle (rectangular marker), perturbed flow due to far (20-cm) obstacle (diamond marker), and perturbed flow due to far obstacle in tandem arrangement (triangle marker). Panel a shows the Average Power versus Reynolds number. This data shows the multiplying effects of perturbed flow in the power production. As expected, the power production in the tandem arrangement is consistent with the perturbed flow due to the far obstacle as the distance of the upstream obstacle from the turbine is essentially identical. Generally, the close upstream obstacle results in larger multiplying effects



(a)

(b)

(c)

(d)

(e)

Fig. 5 Results from the Cross Cylinder model are depicted relative to Reynolds number: (**a**) Average power, (**b**) Maximum possible power, (**c**) Efficiency, (**d**) Number of flips, and (**e**) Maximum rotation time

than the far obstacle when the Reynolds number is below 30,000. The multiplying effects of the close obstacle however are diminishing, from about 13 times to only 2 times, with the increase of Reynolds number. On the other hand, the graph shows that the average power due to the far obstacle and tandem configuration are consistently increasing with Reynolds number. The maximum multiplying factor due to far obstacle is about 15 times. This is less than the maximum multiplying factor presented by the close obstacle of about 21 times at low Reynolds number.

Presented in panel b of Fig. 5 are the potential Maximum Power production of the Cross Cylinder model for all flow cases studied in this project. The graph clearly indicates the strong multiplying effects of perturbed flow on power production of the turbine exposed to free stream. The turbine under free stream—without blockage—shows the lowest possible maximum power for all Reynolds numbers and also low gradient of the linear relationship. And, again, the possible Maximum Power for the far obstacle (diamond) and tandem configuration (triangle) are essentially the same for all Reynolds numbers observed here. Nevertheless, the maximum power for the perturbed flow cause by 10-cm obstacle does not show large deviation from than that by the other two cases. Regression analysis on each data set strongly indicates linear relationship as the $R^2 \approx 0.9$ for all data set, except for the perturbed flow case due to close obstacle. This case only shows $R^2 \approx 0.7$, indicating low linear preference. Note that, as it is indicated in Fig. 3, the local speeds for the perturbed case are essentially the same. Hence, the trend of data shown here may indicate that the maximum power solely depends on the local speed experienced by the turbine. Another important note that we can draw here is that the second turbine located downstream in the tandem configuration does not affect the power production. It can be seen that both the Average Power and Maximum Power data are consistent between the tandem configuration and far obstacle configuration.

The ratio of the Average Power to the Maximum possible Power is defined as the Efficiency of the turbine. Panel c of Fig. 5 depicts the Efficiency of the Cross Cylinder turbine for all cases studied in this project. When the turbine is exposed to the free stream, without obstacle, the Efficiency is increasing with the Reynolds number, but the value is very small—only about 5%. The Efficiency is multiplied when the turbine is exposed to perturbed flow. However, the effects from each are not the same. The close obstacle—10-cm blockage—increases the efficiency up to about 20%, but then the Efficiency decreases with Reynolds numbers down to 5% when the flow speed is maximum. The tandem configuration and far obstacle setup both increase the Efficiency up to about 28%. The Efficiency shown by these two configurations is shown to be steadily increasing with the Reynolds numbers. The tandem configuration seems to be better as it shows less up and down. So, this might be the effect by the turbine placed on the back of the main turbine. The graph of the Efficiency clearly reflects the power generation by the turbine exposed to these four cases.

The effects of the flow perturbation and tandem arrangement on the Number of Flips are displayed in panel d of Fig. 5. Clearly, the graph shows that the perturbed flow reduced the number of flips. The effects of the far obstacle and tandem configuration, again, are very similar. The close obstacle reduces the number of flips, but the effects are reduced after Reynolds number around 30,000. Interestingly, the entire graph resembles a mirror of the Average Power where the effect of close obstacle also shows a local maximum around the same Reynolds number and the effects of far obstacle and tandem are similar.

Panel e of Fig. 5 depicts the Maximum Rotation Time recorded for this turbine. The rotation time also reflects some of the power production by the turbine. During the two 60-s observation time, the turbine demonstrates oscillation and rotation. In the free stream, the longest rotation time shown by the turbine was about 11 or 12 min (combined from the current and voltage measurement). Note that this is not the sum of all rotation times, but only the maximum rotation time demonstrated by the turbine for the two 60-s observation period. This Maximum Rotation Time for the free stream case also seems to be unchanged with the Reynolds numbers, unlike the Average Power and Maximum Power. The low rotation time is reflected in a very low power production by the no-blockage case. Other curves represent results from all perturbed flow cases. The flow perturbation clearly increases the maximum rotation time. However, the effects of the perturbed flow on the maximum rotation time cannot be easily comprehended as there is no specific trend that can be observed. The effects of the close obstacle and tandem configuration show local peaks on different Reynolds numbers. On the other hand, the effects of the 20-cm blockage seem to increase with the Reynolds numbers with a slight local deficit at Reynolds number around 40,000. The tandem configuration shows the most consistent increase of Maximum Rotation Time with respect to the increase in Reynolds numbers, with the exception of a drastic drop when the flow speed is near maximum.

3.3 Results from Triangular Prism with Straight Sides

Figures 6 panels a–d, and e show the outcomes from the triangular prism turbine with straight sides: Average Power, Maximum Power, Efficiency, Number of Flips, and Maximum Rotation Time, respectively. Note that the range of Reynolds numbers for the triangular turbines is lower than one used for the Cross Cylinder because the characteristic length used here is shorter. Panel a of Fig. 6 depicts the Average Power versus the Reynolds number for the turbine exposed to free stream (circle), 10-cm upstream blockage or close obstacle (rectangle), and 20-cm upstream blockage or far obstacle (diamond). Here, it can be seen that the power production by the turbine exposed to free stream is linearly increasing with the Reynolds numbers. This reflects results by the Cross Cylinder model, but the values here are much higher. The data also shows that the upstream blockage generally increases the power production. Effects from the close obstacle, however, are seen to be more consistent than the effects by the far obstacle. Based on previous results from the Cross Cylinder turbine, the far obstacle is expected to provide the largest multiplying effects on the power production. Here, these results are not consistently shown by the far obstacle as the power is not always the largest among the observed flow speeds.

The possible Maximum Power, depicted in panel b of Fig. 6, also indicates consistent increase of maximum power with respect to the Reynolds number. The multiplying effects of the close obstacle look constant across the Reynolds numbers. The effects of the far obstacle, again, do not seem to be consistent across the range of Reynolds numbers.

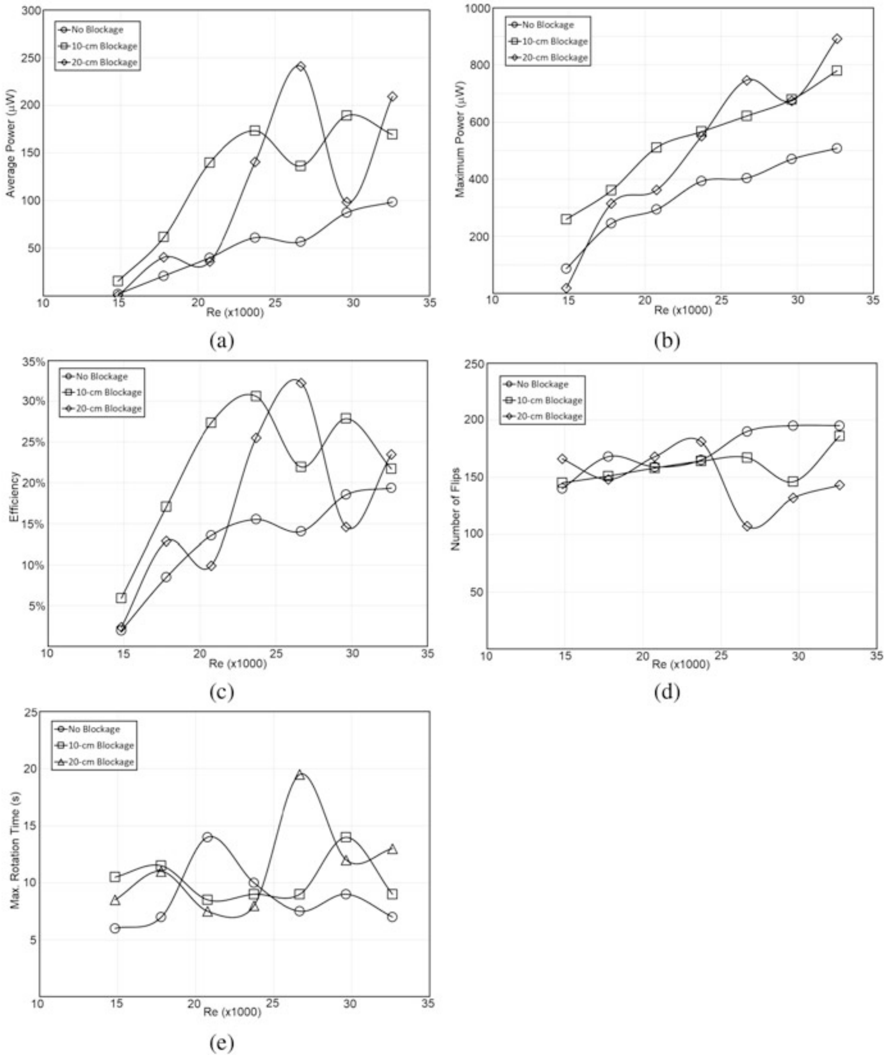


Fig. 6 Results from the triangular prisms with straight sides are depicted relative to Reynolds number: (**a**) Average power, (**b**) Maximum possible power, (**c**) Efficiency, (**d**) Number of flips, and (**e**) Maximum rotation time

The ratio between the Average Power to the possible Maximum Power results in the Efficiency that is depicted in panel c of Fig. 6. Data indicates that the Efficiency is increasing with the Reynolds numbers. The Efficiency can reach around 33% when the turbine is exposed to perturbed flow. When the turbine is exposed to free stream, the Efficiency can reach 20%. This is much better than that of Cross Cylinder model which offers around 5–7%.

Panels d and e of Fig. 6 show the Number of Flips and Maximum Rotation Time produced by the triangular prism turbine with straight sides, respectively. The graphs in panel d of Fig. 6 demonstrate an interesting finding. Here, up to Reynolds number about 25,000, the perturbed flow does not provide any effect on the number of flips. However, beyond this Reynolds number, there is a reduction effect by the perturbed flow, particularly when the obstacle is far. This is in contrast with the data obtained from the Cross Cylinder model which show significant reduction of the number of flips by the perturbed flow across the range of observed Reynolds numbers. Also, it should be pointed out that the Number of Flips presented here is quite high, about 150. This is much higher than that of the Cross Cylinder exposed to perturbed flow than can be reduced to about 40 times.

Data shown on the Maximum Rotation Time shows no specific trend with respect to the increase in Reynolds numbers. All data seem to populate in between 5 and 15 min, with one exception for almost 20 min. This data seems to indicate that there is minimum effect of the perturbed flow on the Maximum Rotation Time of the triangular prism with straight sides.

3.4 Results from Triangular Prism with Curved Sides

Figure 7 panels a to e show result from the experiment with the triangular prism with curved sides. Panel a) displays the relation between Average Power and Reynolds numbers. Unlike the power data of the Cross Cylinder and triangular prism with straight sides, here, the average power shows a non-linear trend with local maximums. The average power is steadily increasing to maximum values when the Reynolds number is about 27,000, and then the average power decreases. When the turbine is exposed to free stream, the average power reaches a little less than 200 μW. The perturbed flows increase the maximum achievement to about 400 μW, double that produced by the free stream case. The similarity of data from the experiment using close and far obstacles indicates that the distance of the upstream blockage does not show significant effect on the power production.

Panel b shows the Maximum Power production in relation with the Reynolds number. The Maximum Power produced by this turbine shows a non-linear trend with local peaks around Reynolds number about 27,000, similar to the trend shown by the Average Power. The maximum power production by this turbine exposed to free stream can go to up to almost 600 μW, while that produced by the perturbed flow can go up to about 900 μW. Similar to the case with the Average Power, the blockage distance does not significantly affect the maximum power production. In fact, at high Reynolds number, the maximum production is lower than the free stream case. Generally speaking, the power production by this turbine is much larger than that by the Cross Cylinder model but is comparable to the triangular prism model with straight sides.

Panel c shows the Efficiency of this turbine. Because both the Average Power and Maximum Power data are non-linear, it is not surprising to see that the Efficiency

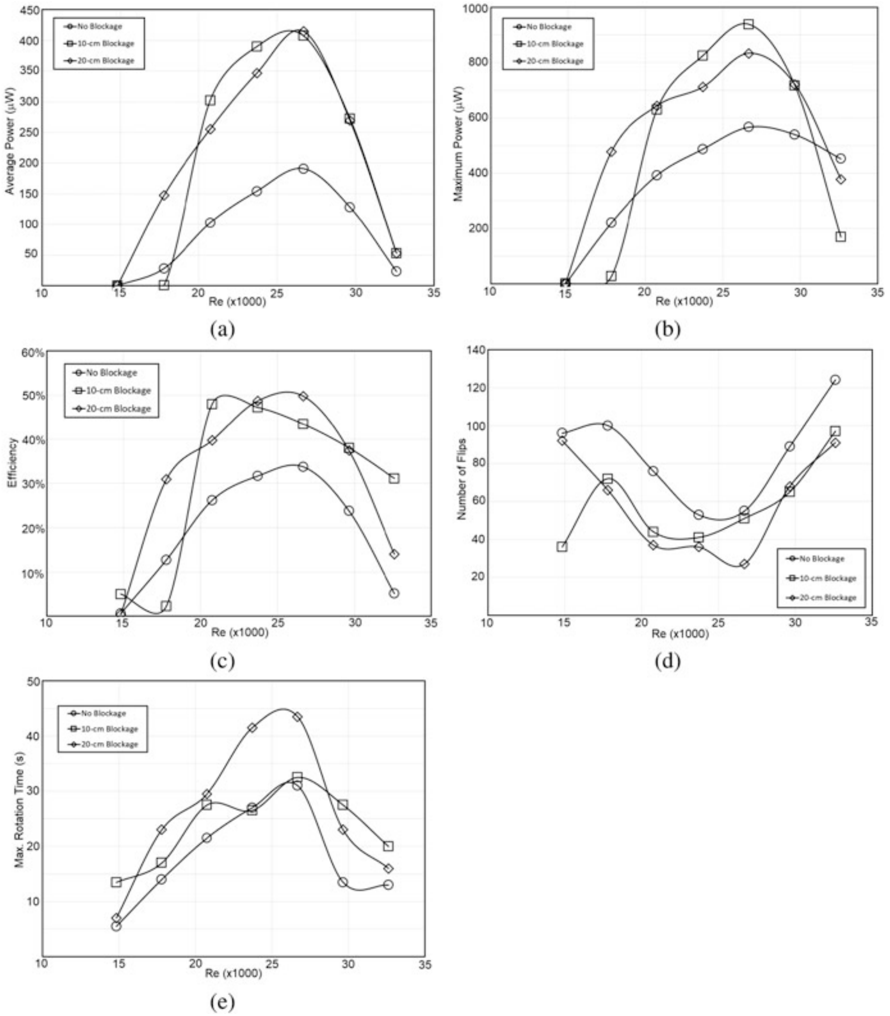


Fig. 7 Results from the triangular prisms with curved sides are depicted relative to Reynolds number: (**a**) Average power, (**b**) Maximum possible power, (**c**) Efficiency, (**d**) Number of flips, and (**e**) Maximum rotation time

shows non-linear trend with local peaks as well. The Efficiency of this turbine can reach up to 50% for the perturbed flow and about 33% for the free stream case. The close and far positions of the upstream obstacles do not show clear differentiation on their effects on the Efficiency. Nevertheless, the 20-cm far blockage shows more consistent trend compare to the 10-cm close blockage. The Efficiency of this turbine is the highest among the three turbines studied here.

Panel d displays the Number of Flips for this turbine exposed to the three cases of flows. Here, the graph also shows non-linear trend with local minimums near the

Reynolds number of 27,000, corresponding to the peaks for average and maximum powers. Nevertheless, the flow perturbation does not seem to affect the Number of Flips obtained by the free stream case as there is not much reduction of flips that can be seen from the graph. Comparing to the other two turbines, this turbine shows the least number of flips (much less than 120), indicating best autorotation performance among the three turbine designs studied here. Nevertheless, the non-linear trend and the specific Reynolds number corresponding to the maximum power prompt the importance of further investigation on the effects of flow speed and turbine's geometry on the autorotation.

Consequently, the Maximum Rotation Time, shown in panel e also, displays a non-linear trend, but in opposite direction to the Number of Flips. The trend shows local peaks of maximum rotation time at around Reynolds number of 27,000. The graph shows that the close obstacle does not affect the maximum rotation time of the free stream case. On the other hand, the far obstacle shows significant improvement over the Maximum Rotation Time around Reynolds number of 25,000. The maximum rotation time obtained here is slightly better than that obtained by the triangular prism with straight sides but less than the Cross Cylinder model.

3.5 Summary and Discussion

The three bladeless turbine designs discussed in this paper demonstrate distinctive performance. The Cross Cylinder turbine and triangular prism design with straight sides shows linear relationship between Average and Maximum Power production with the increase in Reynolds numbers. On the other hand, the triangular design with curved sides shows non-linear trend of Average Power and Maximum Power versus Reynolds numbers. Table 3 shows that the Cross Cylinder model results in the lowest Average Power among the three turbine models for all flow cases studied here. The triangular prism with curved sides produces 191 μW when it is exposed to free stream. The Average Power by this turbine can reach more than 400 μW when the flow is perturbed by the upstream obstacle. Table 4 summarizes the possible Maximum Power produced by the three turbine designs exposed to the three flow cases. Consistent with the Average Power, the Cross Cylinder model shows the lowest possible Maximum Power, of around 200 μW, among the three models. The two triangular prism models show similar possible maximum power of around 500 μW when they are exposed to free stream.

Tables 3 and 4 also signify the effects of upstream obstacles on the Average Power and Maximum Power. As is shown in Fig. 3, the upstream blockage increases the local velocity experienced by the turbine. However, the distances, far or close, do not affect the amount of velocity. For the Cross Cylinder model, the flow perturbation can increase the Average Power by tenfold. For the triangular turbine models, the flow obstruction can double or even triple the Average Power. However, the multiplying effects by the blockages on the Maximum Power can be seen to be less than on the Average Power. The Average Power by the triangular prism

Table 3 Summary of Average Power (μW) by the three turbines exposed to the three flow conditions. Only the maximum values are presented in this table

Turbine type	No blockage μW	10-cm blockage μW	20-cm blockage μW
Cross Cylinder	4.99	27.97	50.79
Triangular prism with straight sides	98.40	189.29	241.17
Triangular prism with curved sides	191.00	407.35	414.67

Table 4 Summary of possible Maximum Power (μW) by the three turbines exposed to the three flow conditions. Only the maximum values are presented in this table

Turbine type	No blockage μW	10-cm blockage μW	20-cm blockage μW
Cross Cylinder	78.59	162.05	187.18
Triangular prism with straight sides	508.27	780.35	891.55
Triangular prism with curved sides	566.78	938.70	833.85

model with curved sides can exceed 191 μW in the free stream and 414 μW, when the flow is perturbed. Possible maximum power of this turbine can reach 930 μW. The blockage distances, far and close, do not show clear distinctive effects on the turbine's performance.

Effects of the perturbed flow on power production by the Cross Cylinder model have been indicated in our past works with Chung [5, 48]. In these past studies, the current and voltage data were obtained discretely, not continuously, and the power production reached only about 1.5 μW at Reynolds number around 34,000. These studies were conducted between Reynolds numbers 27,000 and 34,000. The linear relationship between the power and flow speed and the multiplying effects of the upstream obstacle were observed. Effects of the blockage distance were not studied in the past. Instead, effects of the turbine mass density were investigated, and it was found that the effects of the turbine's density are minimal.

Considering the small size of the turbines tested in this project, the power production presented here is quite large. The large-scale power production of turbine prototypes may be estimated using the Power Coefficient (C_p) formulation defined as

$$C_p = \frac{W}{\rho n^3 D^5} \tag{3}$$

where ρ is the fluid density, n is the revolution per minute, and D is the diameter of the turbine [51]. Assuming that the C_p, ρ, and n for the turbine model and prototype are equal, the large-scale output of the prototype can be calculated as follows:

$$W_p = (\frac{D_p}{D_m})^5 W_m \tag{4}$$

Table 5 Summary of the Efficiency (%) by the three turbines exposed to the three flow conditions. Only the maximum values are presented in this table

Turbine type	No blockage %	10-cm blockage %	20-cm blockage %
Cross Cylinder	6.36	20.00	28.05
Triangular prism with straight sides	19.36	30.60	32.28
Triangular prism with curved sides	33.70	47.90	49.73

Table 6 Summary of the Maximum Rotation Time (second) by the three turbines exposed to the three flow conditions. Only the maximum values are presented in this table

Turbine type	No blockage sec.	10-cm blockage sec.	20-cm blockage sec.
Cross Cylinder	13.5	48.17	45.50
Triangular prism with straight sides	14.00	14.00	19.50
Triangular prism with curved sides	31.00	32.50	19.50

where W_p and W_m are the power by the prototype and model, respectively, and D_p and D_m are the diameters of the prototype and model, respectively. The formulation indicates that a-20 times scale-up would multiply the model's power by 3.2e6 times larger. A Cross Cylinder turbine prototype with 1.00 diameter and length potentially can produce 15 W of power under free symmetric stream and more than 150 W when it is exposed to perturbed flow. On the other hand, a 60-cm wide triangular prism turbine with curved sides potentially can produce 600 W of power under free stream and 1324 W of power when it gets exposed to perturbed flow.

Table 5 displays the Efficiency of the three turbines exposed to the three flow conditions. Only the maximum values are presented in this table. The Efficiency is defined as the ratio of the Average Power to the possible Maximum Power that can be achieved by the turbine. The Cross Cylinder model shows the lowest possible efficiency among the three turbine designs. The triangular prism model with curved sides shows the best efficiency among the three models studied here. The Efficiency by the triangular prism model with straight sides shows slightly less values than the triangular model curved sides. The perturbed flow can be seen to significantly increase the efficiency of the three turbines. The efficiency can go up to 50% for the triangular prism model with curved sides.

Lastly, Table 6 summarizes the Maximum Rotation Time that was recorded during the two 60-s observation period. Only the largest times are included in this table. Note that this parameter does not represent the rotation per minute or angular speed of the turbines. The Cross Cylinder model exposed to the free stream shows the least amount of rotation time. The triangular prism with straight sides shows similar maximum rotation time with the Cross Cylinder. The rotation time by the triangular prism with curved sides double the times by the Cross Cylinder model.

The upstream blockage shows varying effects. It significantly increases the maximum rotation time of the Cross Cylinder model (from 13.5 to 45.50 s), but it does not provide any significant effect on the rotation time of the triangular prism with straight sides. The effects on the triangular prism with curved sides are also minimal. The maximum rotation time can reach 45 s for the Cross Cylinder model and the triangular prism with curved sides.

Low Number of Flips and high Rotation Time indicate high power production. When the turbines are exposed to free stream, the triangular model with curved sides shows the lowest number of flips among the designs discussed in this report. Combined with its long rotation time, it generates large power production for this turbine. The upstream blockage significantly reduces the Number of Flips of the Cross Cylinder model. Interestingly, the flow perturbation does not affect the flip frequency and rotation time demonstrated by both triangular models.

Data presented here indicates that the power production is not proportional to the moment inertia and side surface area of the turbine models (shown in Table 1). The moment inertia and the side surface area, needed to generate shear stress and torque, of the Cross Cylinder turbine are the largest among the three models, but its power production is the least. As the flow is provided by the same pump, the available kinetic energy is the same for all the models. Hence, it should be expected that the angular speed of the Cross Cylinder would be the least. Nevertheless, as neither the torque nor the angular velocity data were measured in this experiment, it is difficult to relate the power production with the geometry and mass properties of the turbines. The triangular model with curved sides demonstrates better performance than its counterpart model with straight sides. The two triangular models have similar side surface area and the moment of inertia. The current data, however, are not sufficient to support any conclusion regarding the role of the curvatures of the triangular model. The roles of the curvature and sharp edges on the autorotation and power production warrant further investigation.

4 Conclusion

Three distinctive bladeless turbine designs have been tested for their hydrokinetic power production potential. The Average Power and possible Maximum Power production by the triangular models are superior to the Cross Cylinder model. In particular, the power produced by the triangular model with curved sides can exceed 60 times greater than that of the Cross Cylinder model. The upstream asymmetric blockage significantly improves the power production of all turbines, but the greatest effect is on the Cross Cylinder model. The Cross Cylinder and triangular model with straight sides show linear relationship between the power production and Reynolds numbers. Interestingly, the triangular model with curved sides shows non-linear trend with local peaks at around Reynolds number of 27,000. Certainly, this is interesting to note as this turbine promises the best performance, but this feature is limited for a certain range of flow speeds. This certainly warrants further investigation.

The upstream blockage disrupts the symmetric flow presented by the observation channel. Hence, the blockage may provide a near-realistic situation of natural rivers where the turbines would have been placed. In our experiments, the flow perturbation significantly increases the power production of the turbines. The effects on the performance of Cross Cylinder model are more significant than on the performance of triangular turbines. However, its effects on the Number of Flips and Rotation Time remain to be investigated. It certainly improves the autorotation of the Cross Cylinder model, but it does not seem to show clear benefit for the reduction of flip frequency and maximization of the rotation time of the triangular prism turbines. The upstream distance of the blockage from the turbine affects the power production of the Cross Cylinder, but not the Efficiency and rotation time. The tandem configuration does not improve the performance of the front turbine.

References

1. R. Camassa, B.J. Chung, P. Howard, R. McLaughlin, A. Vaidya, Vortex induced oscillations of cylinders at low and intermediate Reynolds numbers, in *Advances in Mathematical Fluid Mechanics—Dedicated to Giovanni Paolo Galdi on the Occasion of His 60th Birthday*. (Springer, 2010), pp. 135–145
2. R. Camassa, B.J. Chung, G. Gipson, R. McLaughlin, A. Vaidya, Vortex induced oscillations of cylinders (2008)
3. B. Chung, M. Cohrs, W. Ernst, G.P. Galdi, A. Vaidya, Wake–cylinder interactions of a hinged cylinder at low and intermediate Reynolds numbers. Arch. Appl. Mech. **86**(4), 627–641 (2016)
4. B.J. Chung, G. Gipson, A. Shenoy, A. Vaidya, Image analysis of wake structure past cylinders of finite lengths. Int. J. Imaging **A10**(4), 18–32 (2010)
5. J. Araneo, B.J. Chung, M. Cristaldi, J. Pateras, A. Vaidya, R. Wulandana, Experimental control from wake induced autorotation with applications to energy harvesting. Int. J. Green Energy **16**(15), 1400–1413 (2019)
6. B.W. Skews, Autorotation of many-sided bodies in an airstream. Nature **352**(6335), 512–513 (1991)
7. B.W. Skews, Autorotation of polygonal prisms with an upstream vane. J. Wind Eng. Ind. Aerodyn. **73**(2), 145–158 (1998)
8. M. Armandei, A.C. Fernandes, Marine current energy extraction through buffeting. Int. J. Mar. Energy **14**, 52–67 (2016)
9. Y. Zhang, C. Kang, H. Zhao, H.B. Kim, Effects of the deflector plate on performance and flow characteristics of a drag-type hydrokinetic rotor. Ocean Eng. **238**, 109760 (2021)
10. K. Golecha, T.I. Eldho, S.V. Prabhu, Influence of the deflector plate on the performance of modified Savonius water turbine. Appl. Energy **88**(9), 3207–3217 (2011)
11. G. Kailash, T.I. Eldho, S.V. Prabhu, Performance study of modified Savonius water turbine with two deflector plates. Int. J. Rotating Mach. **2012** (2012)
12. V. Patel, T.I. Eldho, S.V. Prabhu, Performance enhancement of a Darrieus hydrokinetic turbine with the blocking of a specific flow region for optimum use of hydropower. Renewable Energy **135**, 1144–1156 (2019)
13. V.K. Patel, R.S. Patel, Optimization of an angle between the deflector plates and its orientation to enhance the energy efficiency of Savonius hydrokinetic turbine for dual rotor configuration. Int. J. Green Energy **19**(5), 476–89 (2022)
14. B. Jeeva, S. Jai Sandeep, N. Ramsundram, M. Prasanth, B. Praveen, Experimental investigation of three bladed inclined Savonius hydrokinetic turbine by using deflector plate. IOP Conf. Series: Mater. Sci. Eng. **1146**(1), 012009 (2021)

15. M. Mosbahi, A. Ayadi, Y. Chouaibi, Z. Driss, T. Tucciarelli, Performance study of a Helical Savonius hydrokinetic turbine with a new deflector system design. Energy Convers. Manag. **194**, 55–74 (2019)
16. M. Mosbahi, S. Elgasri, M. Lajnef, B. Mosbahi, Z. Driss, Performance enhancement of a twisted Savonius hydrokinetic turbine with an upstream deflector. Int. J. Green Energy **18**(1), 51–65 (2021)
17. M.B. Salleh, N.M. Kamaruddin, P.H. Tion, Z. Mohamed-Kassim, Comparison of the power performance of a conventional Savonius turbine with various deflector configurations in wind and water. Energy Convers. Manag. **247**, 114726 (2021)
18. M.E. Nimvari, H. Fatahian, E. Fatahian, Performance improvement of a Savonius vertical axis wind turbine using a porous deflector. Energy Convers. Manag. **220**, 113062 (2020)
19. H. Alizadeh, M.H. Jahangir, R. Ghasempour, CFD-based improvement of Savonius type hydrokinetic turbine using optimized barrier at the low-speed flows. Ocean Eng. **202**, (107178) (2020)
20. P.K. Pulijala, R.K. Singh, *Performance Analysis of Savonius Hydrokinetic Turbine with Stationary Deflector Plates Using CFD*. Lecture Notes in Mechanical Engineering. (Springer Science and Business Media Deutschland GmbH, 2021), pp. 541–552
21. E. Kerikous, D. Thévenin, Optimal shape and position of a thick deflector plate in front of a hydraulic Savonius turbine. Energy **189**, 116157 (2019)
22. W.I. Ibrahim, M.R. Mohamed, R.M.T.R. Ismail, P.K. Leung, W.W. Xing, A.A. Shah, Hydrokinetic energy harnessing technologies: a review. Energy Rep. **7**, 2021–2042 (2021)
23. E. Erinofiardi, A. Akbarzadeh, P. Gokhale, A. Date, P. Bismantolo, A.F. Suryono, A.K. Mainil, A. Nuramal, A review on micro hydropower in Indonesia. Energy Procedia 316–321 (2017)
24. Y. Susilowati, P. Irasari, A. Susatyo, Study of hydroelectric power plant potential of Mahakam River Basin East Kalimantan Indonesia, in *Proceeding—2019 International Conference on Sustainable Energy Engineering and Application: Innovative Technology Toward Energy Resilience, ICSEEA 2019* (2019), pp. 207–213
25. M.B. Salleh, N.M. Kamaruddin, Z. Mohamed-Kassim, Micro-hydrokinetic turbine potential for sustainable power generation in Malaysia. IOP Conf. Series: Mater. Sci. Eng. **370**(1), 012053 (2018)
26. F. Behrouzi, M. Nakisa, A. Maimun, Y.M. Ahmed, Renewable energy potential in Malaysia: hydrokinetic river/marine technology (2016)
27. M. Anyi, B. Kirke, S. Ali, Remote community electrification in Sarawak, Malaysia. Renewable Energy **35**(7), 1609–1613 (2010)
28. F. Ali, C. Srisuwan, K. Techato, A. Bennui, T. Suepa, D. Niammuad, Theoretical hydrokinetic power potential assessment of the U-Tapao River Basin using GIS. Energies **13**(7), 1749 (2020)
29. C.M. Niebuhr, M. van Dijk, V.S. Neary, J.N. Bhagwan, A review of hydrokinetic turbines and enhancement techniques for canal installations: technology, applicability and potential (2019)
30. L. Ladokun, K.R. Ajao, B. Sule, Regional scale assessment of the gross hydrokinetic energy potentials of some rivers in lower Niger River Basin, Nigeria. Niger. J. Technol. **34**, 421 (2015)
31. C.H. da Costa Oliveira, M. de Lourdes Cavalcanti Barros, D.A.C. Branco, R. Soria, P.C.C. Rosman, Evaluation of the hydraulic potential with hydrokinetic turbines for isolated systems in locations of the Amazon region. Sustain. Energy Technol. Assess. **45**, 101079 (2021)
32. M.J. Khan, T. Iqbal, J.E. Quaicoe, M.T. Iqbal, J.E. Quaicoe, River current energy conversion systems: Progress, prospects and challenges. Renewable Sustain. Energy Rev. **12**(8), 2177–2193 (2008)
33. L.I. Lago, F.L. Ponta, L. Chen, Advances and trends in hydrokinetic turbine systems. Energy Sustain. Dev. **14**(4), 287–296 (2010)
34. S.S. Khalid, Z. Liang, N. Shah, Harnessing tidal energy using vertical axis tidal turbine. Res. J. Appl. Sci. Eng. Technol. **5**, 239–252 (2013)
35. M.J. Khan, G. Bhuyan, T. Iqbal, J.E. Quaicoe, M.T. Iqbal, J.E. Quaicoe, Hydrokinetic energy conversion systems and assessment of horizontal and vertical axis turbines for river and tidal applications: a technology status review. Appl. Energy **86**(10), 1823–1835 (2009)

36. M. Armandei, A.C. Fernandes, A. Bakhshandeh Rostami, Hydroelastic buffeting assessment over a vertically hinged flat plate. Exp. Tech. **40**(2), 833–839 (2016)
37. A. Bakhshandeh Rostami, M. Armandei, Renewable energy harvesting by vortex-induced motions: review and benchmarking of technologies (2017)
38. J. Wang, L. Geng, L. Ding, H. Zhu, D. Yurchenko, The state-of-the-art review on energy harvesting from flow-induced vibrations (2020)
39. M.M. Bernitsas, K. Raghavan, Y. Ben-Simon, E.M.H. Garcia, VIVACE (Vortex induced vibration aquatic clean energy): a new concept in generation of clean and renewable energy from fluid flow. ASME. J. Offshore Mech. Arct. Eng. **130**(4), 41101–41115 (2008)
40. W. Xu, M. Yang, E. Wang, H. Sun, Performance of single-cylinder VIVACE converter for hydrokinetic energy harvesting from flow-induced vibration near a free surface. Ocean Eng. **218**, 108168 (2020)
41. A.G. Festo, K.G. Co, Festo dualwing. https://www.festo.com/group/en/cms/10222.htm, 20
42. Vortex Bladeless, Vortex Bladeless Turbine - Reinventing wind energy! https://vortexbladeless.com/
43. D. Villarreal, VIV Resonant Wind Generators. Technical report, Vortex Bladeless, Madrid, Spain, 2018
44. A.C. Fernandes, A. Bakhshandeh Rostami, Hydrokinetic energy harvesting by an innovative vertical axis current turbine. Renewable Energy **81**, 694–706 (2015)
45. A. Bakhshandeh Rostami, A.C. Fernandes, The effect of inertia and flap on autorotation applied for hydrokinetic energy harvesting. Appl. Energy **143**, 312–323 (2015)
46. B.W. Skews, Autorotation of rectangular plates. J. Fluid Mech. **217**, 33–40 (1990)
47. M. Cohrs, W. Ernst, A. Vaidya, Potential for energy harvesting from vortex induced oscillations. Int. J. Ecol. Dev. **26**, 1–9 (2013)
48. R. Wulandana, D. Foote, A. Vaidya, B.J. Chung, Vortex-induced autorotation potentials of bladeless turbine models. Int. J. Green Energy **19**(2), 190–200 (2022)
49. R. Wulandana, Open water flume for fluid mechanics lab. Fluids **6**(7), 242 (2021)
50. Vernier, Flow rate sensor. https://www.vernier.com/product/flow-rate-sensor/
51. Y.A. Cengel, J.M. Cimbala, *Fluid Mechanics Fundamental and Applications*, 1st edn. (McGraw Hill, New York, 2006)

Part IV
Applications

Fickian and Non-Fickian Transports in Ultrasound Enhanced Drug Delivery: Modeling and Numerical Simulation

Ebrahim Azhdari, Aram Emami, and José Augusto Ferreira

1 Introduction

In the *World health statistics 2021: Monitoring Health for the SGDs*, from the World Health Organization, it is reported that cancers are among the leading causes of morbidity and mortality worldwide with approximately 8.7 million cancer-related deaths in 2016 and a projection of over 13 million deaths in 2030.

The classical approach to treat cancer is the chemotherapy administered with different procedures depending on the cancer type. Traditionally, the cytotoxic drugs are systemically administered and transported to the target by the blood stream leading to severe side effects. Only a small part of the administered drug reaches the target. The drug dose-limiting toxicity restricts the amount drug administered in each chemotherapy protocol [37].

The cancer microenvironment is the major barrier to the drug delivery. The connective tissue in the body is composed by interstitial fluid and extracellular matrix (ECM) that is composed by proteins, glycoproteins, proteoglycans, and polysaccharides. The change of the ECM properties is one of the main features of cancer. Tumor progression is accompanied by the tumor fibrosis (desmoplasia)

E. Azhdari
Department of Mathematics, Salman Farsi University of Kazerun, Kazerun, Iran
e-mail: e.azhdari@kazerunsfu.ac.ir

A. Emami
Department of Mathematics, Salman Farsi University of Kazerun, Kazerun, Iran

Department of Mathematics, Faculty of Sciences, Fasa University, Fasa, Iran
e-mail: emami@fasau.ac.ir

J. A. Ferreira (✉)
Department of Mathematics, University of Coimbra, CMUC, Coimbra, Portugal
e-mail: ferreira@mat.uc.pt

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_13

characterized by excessive collagen depositions in the surroundings of the tumor, often crosslinked, that leads to an increase in the tissue stiffness [30, 34].

Many solid tumors are characterized by abnormal vasculature with intercellular gaps and endothelial fenestrae that lead to vascular leakage; some regions present a reduced penetration of blood vessels and consequently reduced blood flow. The leaky vasculature and the non-existence of an efficient lymphatic drainage system lead to the fluid accumulation and an elevated interstitial fluid pressure. In the tumor periphery is observed lower interstitial fluid pressure due to the action of the functioning lymphatic system. The balance between the irregular and leaky vasculature system and the inefficient drainage lead to enhanced permeability and retention effects [2, 31].

Novel tools and technological approaches have captured the attention of researchers in drug delivery in order to improve the performance of conventional therapeutics and patient compliance for cancer therapy. Physical (also called exogenous, external, or extrinsic) stimuli-responsive drug delivery systems (SRDDS) are promising approaches to control and target drug delivery for cancer treatments [7, 13, 15, 33, 46]. In this case, the drugs are entrapped in nanocarriers (liposomes, dendrimers, micelles, polymeric nanoparticles, carbon nanotubes,. . .) that can be systemically administered and transported to the target. The application of external stimuli (ultrasound, temperature, electric fields, magnetic fields, light,. . .) activates the drug release at appropriate rate, specific time, and desired site and eventually changing the properties of the target tissue that leads to the increase of the drug transport. In this paper we will be focused in the use of ultrasound as enhancer of the drug delivery.

Ultrasound has a number of attractive characteristics as a trigger for drug delivery. It is promising because of its non-invasiveness, the absence of ionizing radiations, and the facile regulation of tissue penetration depth by tuning frequency, duty cycles, and time of exposure. When ultrasound is applied, the drug release can be triggered through the thermal and mechanical effects generated by hyperthermia, cavitations, and radiation forces. This phenomenon promotes the fusion of the drug carrier and the heating of the cancer and the surrounding tissues, resulting in change in the thermal and mechanical properties of the tissue. These changes are experimentally observed in [14] and [16].

Ultrasound is mechanical longitudinal wave propagating in a medium through changes in pressure, at frequencies higher than the audible ones for the human ear (20 kHz). As the ultrasound wave propagates, it induces changes in the pressure of the surrounding medium with a succession of compression and decompression events. Ultrasound can be modulated by varying different parameters such as frequency and intensity that are of utmost importance. Thermal and mechanical effects are among the most important biological effects which are induced by ultrasound [7, 11, 43].

When an ultrasonic wave propagates through the body, it is attenuated by the contact with different tissues by absorption and scattering. As a consequence of the energy absorption, an increase of tissue temperature is observed that leads to

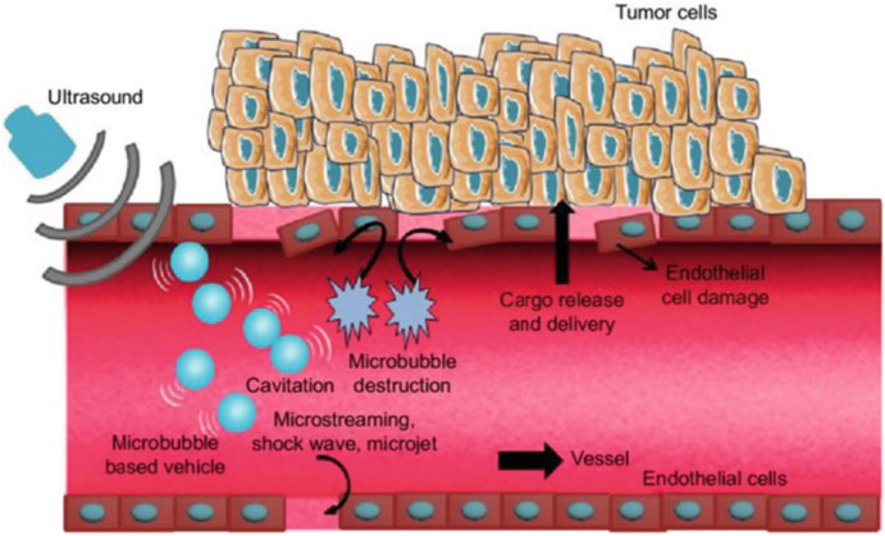


Fig. 1 Schematic representation of US enhanced drug delivery

biological properties change, an increase of the blood flow due to the dilation of blood vessels, and an increase in permeability of the normal vascular walls.

Cavitation is recognized as a major cause of ultrasound-induced mechanical effects. It is defined as the creation or motion of very small gas bubbles that are produced in tissue due to the alternating expansion and compression of tissue as acoustic pressure waves propagate through it. Once the cavitation bubbles are produced, they may undergo oscillations during many cycles of the acoustic wave, called non-inertial (stable) cavitation. When low pressure acoustic ultrasound is applied, stable cavitation bubbles oscillate in size but do not collapse. This oscillating motion causes the rapid movement of fluid near the cavitation bubble, a phenomenon which is called micro-streaming. Stable cavitation can generate mechanical stress on blood vessels to enhance vascular permeability of the tissue. Inertial cavitation occurs due to violent oscillation, rapid bubble growth during the rarefaction cycle of the acoustic wave, and then violent collapse and destruction of the bubble. Bubbles that collapse close to a cell wall or solid surface produce a very high-speed liquid jet that drives into the surface and results in pitting of the surface or cell wall (Fig. 1).

The use of ultrasound has been shown to enhance drug delivery to solid tumors through iterations with ECM and interstitial fluid pressure. The increasing on the temperature leads to an increasing of the blood flow and to the modification of the ECM structure. In fact, the collagen heating induces an increase in the interfibrillar space. The unfolding of the dense collagen matrix is accompanied by an increase in hydraulic conductivity that can enhance the fluid flow and consequently can lead to a reduction of the interstitial fluid pressure. It should be pointed out that ultrasound can also induce pore formation at cellular level resulting from the displacement of the soft tissue due to the pressure waves propagation. Cavitation can lead to the disruption of the collagen matrix. In both cases, an enhancing of the fluid transport is observed inducing a reduction of the interstitial fluid pressure. At a macroscopic level, ultrasound enhances the diffusion transport due to the temperature increasing and convective transport due to the reduction of the interstitial fluid pressure and to the pressure waves propagation [2, 31].

Viscoelastic materials are materials that present viscous properties (that means that they deform subject to a force) and present elastic properties, that is, they return to their initial form when the deformation force stops [9]. Biological materials like extracellular matrix scaffold, cancer cells and tissues are considered viscoelastic materials (see, for instance, [1, 10, 38]). It should be pointed out that the viscoelastic properties of the tissues are mainly determined by the ECM viscoelasticity that depends on the types and strength of the matrix crosslink bonds and the molecular weight of the matrix. The tumor progression leads to a collagen accumulation, often crosslinked, that often results in a stiffness increasing. The viscoelastic behavior is traditionally described by stress-strain relations—Maxwell models, Kelvin-Voigt models, Zener models, or generalized Maxwell models [9, 45].

ECM viscoelasticity has an important role in tissue dynamics. For instance, ECM stiffness and viscoelasticity are key factors on cell dynamics [21, 32].

Transport in viscoelastic material has been object of intense research during the last decade due to the fact that the diffusion transport violates the classical Fick law. Several approaches have been proposed to model the pathological behavior observed in this kind of material (see, for instance, [12, 17, 19, 20, 29, 40, 42]). As mentioned before, the stress-strain relation depends on the material. For instance, in biological tissues, Maxwell, Kelvin-Voigt, Zener, or generalized Maxwell models were considered. To include the viscoelastic effect in the drug transport, it was assumed that the strain depends on the drug concentration, and consequently, a stress-drug concentration relation is established. The mass flux is decomposed into two parts, one of Fickian type and the second one is given in function of the gradient of the stress and, considering mass conservation, integro-differential equations were proposed to replace the traditional parabolic equations for the concentration (see, for instance, [4, 5, 8, 25–27]).

As mentioned above, experiments show that application of ultrasound alters the mechanical properties of the target tissue. It has been observed that the diffusion and convection transports increase with the ultrasound intensity. Then, if we consider a drug initially distributed in a neighborhood of a cancer tissue, when ultrasound is used as enhancer, then the coupling between the pressure waves propagation, the structural change in the tissues, the drug transport, the viscoelastic behavior in the two target need to be considered. A multiphysics and multidomain approach should be adopted to describe accurately the drug transport in this scenario.

There are numerous contributions on computational modeling of ultrasound enhanced drug transport. Without being exhaustive we mention the following papers: [6] where the acoustic pressure propagation is described by the Khokhlov-Zabolotskaya-Kuznetsov equation; [18] that deals with a Helmholtz equation to describe the acoustic pressure propagation, [23] considers a wave equation for the acoustic pressure coupled with a convection-diffusion equation for the drug concentration; [24] where the mathematical model introduced in the previous paper is modified introducing the heat effect; [39] uses a modified Westervelt equation for the pressure waves description; and [47] takes an explicit expression for the pressure intensity. However, up to now no computational methodology is provided that combines the propagation of the pressure waves induced by ultrasound, the change of the target tissues, and the drug transport in the two neighboring tissues: healthy and diseased tissues (cancer) and their different viscoelastic properties.

J.A. Ferreira et al. in [23] studied a system of partial differential equations defined by a hyperbolic equation (wave equation) and a parabolic equation (convection-diffusion-reaction equation) that can be used to describe the drug transport in a target tissue enhanced by ultrasound. In this paper is proposed the coupling between the acoustic pressure wave propagation and the drug transport considering that the convective velocity depends on the acoustic pressure intensity and eventually on its gradient. In [24] the drug transport enhanced by ultrasound is also considered but introducing the heat effect resulting from the acoustic pressure waves propagation. In these papers the authors propose numerical methods to compute second-order accurate approximations for the acoustic pressure intensity and for the drug concentration.

In this paper we consider the scenario the approach described in [23] and [24], that is, we consider a healthy and a cancer tissues, a drug initially distributed in the healthy tissue, the intensity of the acoustic pressure waves described by wave equations, with attenuation terms due to the energy absorption by the targets, the drug transport is described by convection-diffusion equations where the convective velocities and the diffusion coefficients depend on the pressure waves intensity to take into account the structural change in the targets and the reduction of the interstitial fluid pressures in both tissues. To simplify, the viscoelastic target effects on the drug transport are only considered in the cancer tissue. We assume that the viscoelastic behavior of the last target tissue is described by a Zener model [45]. The paper is organized as follows. Section 2 is devoted to the introduction of the system of partial differential equations that will be considered in what follows. Taking into account phenomenological information, the behavior of the drug mass in the system is studied in Sect. 3. In Sect. 4 we present a variational formulation, and we establish a stability result for the continuous model that leads to the uniqueness of solution of the differential problem. Numerical simulations are presented in Sect. 5. Finally, in Sect. 6 we present some conclusions. In the near future, we intend to study the existence of solution of the differential problem considered here as well as propose efficient and accurate numerical methods.

2 Coupling Acoustic Pressure with Drug Transport

We consider a healthy tissue Ω_1 where a drug is initially dispersed. This tissue is in contact with a solid cancer tissue represented by Ω_2 (see Fig. 2). Let Γ_i, $i = 1, 2, 3, 4$, be the boundary of Ω_1, and let $\Gamma_i, i = 4, 5, 6, 7$, be the boundary of Ω_2 being Γ_4 the interface between the two domains that represents the interface between the healthy and cancer tissue.

Let p_i be the acoustic wave pressure intensity in the tissue $\Omega_i, i = 1, 2$, that we consider be described by the following telegraph equation:

$$\frac{\partial^2 p_i}{\partial t^2} + 2\alpha_i\beta_i\frac{\partial p_i}{\partial t} = \beta_i^2\Delta p_i \ \texttt{in} \ \Omega_i \times (0, T_f], \tag{1}$$

for $i = 1, 2$, where β_i is the sound speed, α_i is the attenuation coefficient, and T_f is the final time (see [18, 23, 36]). In (1), $\Delta p_i(t)$ denotes the Laplacian of p_i with respect to the spatial variables.

In the healthy tissue we consider that the drug transport is described by the following convection-diffusion equation

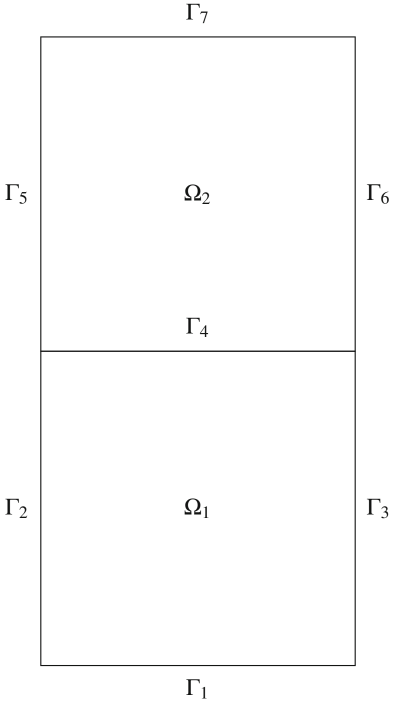


Fig. 2 Spatial domain: healthy and cancer tissues

$$\frac{\partial c_1}{\partial t} + \nabla.(v_1 c_1) - \nabla.(D_1 \nabla c_1) = 0 \texttt{ in } \Omega_1 \times (0, T_f], \qquad (2)$$

where c_1 is a drug concentration, v_1 denotes the convective velocity due to the ultrasound effect, and D_1 is diffusion coefficient. In (2), $\nabla.(v_1 c_1)$ represents the divergence of $v_1 c_1$, ∇c_1 denotes the gradient of c_1 with respect to the spatial variables, and $\nabla.(D_1 \nabla c_1)$ represents the divergence of $D_1 \nabla c_1$.

If a low pressure acoustic ultrasound is applied, stable cavitation bubbles oscillate in size but do not collapse. This oscillating motion causes the rapid movement of fluid near the cavitation bubble, a phenomenon which is called micro-streaming. To take into account this effect, we assume that v_1 is defined as follows [44]:

$$v_1 = v_{1,0} + v_1^*, \qquad (3)$$

where $v_{1,0}$ is the steady-state fluid velocity and v_1^* is the enhanced velocity due to the ultrasound defined as follows:

$$v_1^* = \phi_1 p_1^2, \qquad (4)$$

where ϕ_1 is a constant. Also in the presence of ultrasound, the diffusion transport increases being D_1 defined by

$$D_1 = D_{1,0} + \psi_1 v_1,$$

where ψ_1 is a constant and $D_{1,0}$ is the molecular diffusion coefficient [44].

In the cancer tissue we take into account the viscoelastic effect in the drug transport, and then the drug transport is defined by

$$\frac{\partial c_2}{\partial t} + \nabla.(v_2 c_2) = \nabla.(D_2 \nabla c_2) + \nabla.(D_v \nabla \sigma) - \lambda c_2 \texttt{ in } \Omega_2 \times (0, T_f], \qquad (5)$$

where c_2 represents the drug concentration in the non-healthy tissue Ω_2, σ denotes the scalar stress due to the viscoelastic characteristics of the target, D_2 is the drug diffusion coefficient, D_v is the viscoelastic diffusion coefficient, and λ represents the drug consumption rate by the cancer cells. In (5), v_2 represents the ultrasound enhanced convective velocity that is defined as (3), but it should take into account the reduction of the interstitial fluid pressure due to structural modification of the ECM induced by ultrasound, and is defined by

$$v_2 = v_{2,0} + v_2^*,$$

where $v_{2,0}$ is the steady-state fluid velocity and

$$v_2^*(t) = \phi_2 p_2^2(t) + \phi_3 \|\nabla p_2(t)\|^2,$$

with $\phi_i = 2, 3$, constants and $\|\nabla p_2(t)\|^2 = \sum_{i=1,2} \|\frac{\partial p_2}{\partial x_i}(t)\|^2_{L^2(\Omega_i)}$. The diffusion coefficient D_2 is given by

$$D_2(t) = D_{2,0} + \psi_2 p_2^2(t),$$

where $D_{2,0}$ is the molecular diffusion coefficient and ψ_2 represents a positive constant.

Regarding the diffusive component of the flux, it is well-known that Fick's law does not represent an accurate description of the diffusion phenomenon due to the viscoelastic effects as pointed out in the introduction. One of the simplest rheological models to characterize stress-strain relaxation in biological tissues is the so-called Zener model

$$\frac{\partial \sigma}{\partial t} + \frac{E_2}{\mu}\sigma = -(E_1 + E_2)\frac{\partial \varepsilon}{\partial t} - \frac{E_1 E_2}{\mu}\varepsilon \text{ in } \Omega_2 \times (0, T_f], \tag{6}$$

where σ and ε denote the stress and strain, respectively, E_1 denotes the stiffness of the single spring, and E_2 and μ the stiffness and damping coefficient of the spring-dashpot couple, respectively (see [9, 35, 45]).

It should be remarked that in Eq. (6), we assumed that the relation between strain and concentration of drug in the solid cancer tissue is a linear version of the $\varepsilon = f(c_2)$ adopted, for instance, in [22], that is, $\varepsilon = \gamma c_2$, where γ is a positive constant. Then (6) is replaced by

$$\frac{\partial \sigma}{\partial t} + \frac{E_2}{\mu}\sigma = -(E_1 + E_2)\gamma\frac{\partial c_2}{\partial t} - \frac{E_1 E_2}{\mu}\gamma c_2 \text{ in } \Omega_2 \times (0, T_f]. \tag{7}$$

We also remark that the minus sign in the right hand side of Eq. (7) means that the solid cancer tissue acts as a barrier to the drug transport.

There is a well-established theory for diffusion in linear viscoelastic media presented in [3] and [41], considering the stress tensor T_S, that takes into account the stress supported by the diffusion substance, and a diffusive force vector that takes into account the momentum change between the diffusion species and the medium.

From (7) we easily obtain, by assuming that the coefficients are constant and the initial drug concentration in the tumor, $c_2(0)$, is zero,

$$\sigma(t) = \sigma_0 e^{-\frac{E_2}{\mu}t} - (E_1 + E_2)\gamma c_2(t) + \frac{E_2^2\gamma}{\mu}\int_0^t e^{-\frac{E_2}{\mu}(t-s)} c_2(s)ds.$$

Consequently the mass flux J_2 admits the following representation

$$J_2(t) = -D_2^*\nabla c_2 + v_2 c_2 - \frac{D_v E_2^2\gamma}{\mu}\int_0^t e^{-\frac{E_2}{\mu}(t-s)}\nabla c_2(s)ds,$$

where $D_2^* = D_2 - D_v(E_1 + E_2)\gamma$. Assuming that D_v and the initial stress, σ_0, are constants, Eq. (5) is equivalent to

$$\frac{\partial c_2}{\partial t} + \nabla.(v_2 c_2) = \nabla.(D_2^* \nabla c_2) + \frac{D_v E_2^2 \gamma}{\mu} \int_0^t e^{-\frac{E_2}{\mu}(t-s)} \nabla^2 c_2(s) ds - \lambda c_2. \tag{8}$$

For the initial acoustic pressures and drug concentrations, we assume the following conditions

$$\begin{cases} p_i(0) = \frac{\partial p_i}{\partial t}(0) = 0 \;\; \texttt{in} \; \Omega_i, \; i = 1, 2, \\ c_1(0) = c_{1,0} \;\texttt{in}\; \Omega_1, \\ c_2(0) = 0, \sigma(0) = \sigma_0 \;\texttt{in}\; \Omega_2. \end{cases} \tag{9}$$

For $i = 1, 2$, let J_i be the drug flux in Ω_i defined by

$$J_1 = -D_1 \nabla c_1 + v_1 c_1, \;\; J_2 = -D_2 \nabla c_2 + v_2 c_2 - D_v \nabla \sigma.$$

We observe that J_1 has two contributions: a Fickian and a convective one, and J_2 presents two contributions analogous to the ones of J_1 and a contribution due to the viscoelastic effect of the target tissue on the drug transport defined by $-D_v \nabla \sigma$. We assume that the boundaries Γ_j, $j = 1, 2, 3$, are isolated; that means that no mass flux crosses it,

$$J_1.\eta = 0 \texttt{ on } \left(\cup_{j=1,2,3} \Gamma_j \right) \times (0, T_f], \tag{10}$$

where η denotes the exterior unit normal to Ω_1.

We assume that the boundaries Γ_j, $j = 5, 6, 7$, are not isolated, that is, the drug cross these boundaries is defined by

$$J_2.\eta = A_1 c_2 \texttt{ on } \left(\cup_{j=5,6,7} \Gamma_j \right) \times (0, T_f], \tag{11}$$

where A_1 is a permeability coefficient. Equation (11) means that the amount of drug that crosses $\cup_{j=5,6,7}\Gamma_j$ depends on the amount of drug that reaches this boundary and on its permeability.

On the interface boundary Γ_4 , the continuity of the drug mass fluxes are assumed

$$\begin{cases} J_1.\eta = A_2(c_1 - c_2) \\ J_1.\eta = -J_2.\nu \qquad \texttt{on } \Gamma_4 \times (0, T_f] \end{cases}, \tag{12}$$

where A_2 is a partition coefficient, η is the unit exterior normal to Ω_1 on Γ_4 and $\eta = -\nu$. The first equation of (12) means that the amount of drug that crosses Γ_4 is proportional to the difference between the drug concentration that reaches Γ_4 through Ω_1 and the drug concentration that is in Γ_4 from Ω_2. The continuity of the drug mass flux through Γ_4 is represented by the second equation of (12).

The acoustic pressure is assumed to be known on the boundary Γ_1, that is,

$$p_1 = p_{\Gamma_1} \text{ on } \Gamma_1 \times (0, T_f]. \tag{13}$$

The boundaries Γ_j, $j = 2, 3, 5, 6, 7$, do not interfere with the pressure wave propagation; that is, a homogeneous Neumann boundary condition is prescribed

$$\nabla p_1.\eta = 0 \text{ on } \left(\cup_{j=2,3} \Gamma_j\right) \times (0, T_f], \tag{14}$$

and

$$\nabla p_2.\eta = 0 \text{ on } \left(\cup_{j=5,6,7} \Gamma_j\right) \times (0, T_f]. \tag{15}$$

On the interface boundary Γ_4, we assume continuity of the acoustic pressure

$$p_1 = p_2 \text{ on } \Gamma_4 \times (0, T_f], \tag{16}$$

and

$$\beta_1 \nabla p_1.\eta + \beta_2 \nabla p_2.\nu = 0 \text{ on } \Gamma_4 \times (0, T_f]. \tag{17}$$

In (17), η and ν are the unitary normals on Γ_4 exterior to Ω_1 and to Ω_2, respectively.

The boundary and interface conditions are summarized in what follows

$$\begin{cases} J_1.\eta = 0 \text{ on } (\Gamma_1 \cup \Gamma_2 \cup \Gamma_3) \times (0, T_f], \\ J_1.\eta = A_2(c_1 - c_2) \text{ on } \Gamma_4 \times (0, T_f], \\ J_2.\eta = -J_1.\nu \text{ on } \Gamma_4 \times (0, T_f], \\ J_2.\eta = A_1 c_2 \text{ on } \left(\cup_{j=5,6,7} \Gamma_j\right) \times (0, T_f], \\ p_1 = p_{\Gamma_1} \text{ on } \Gamma_1 \times (0, T_f], \\ \nabla p_i.\eta = 0 \text{ on } \left(\cup_{j=2,3,5,6,7} \Gamma_j\right) \times (0, T_f],\ i = 1, 2, \\ \beta_1 \nabla p_1.\eta + \beta_2 \nabla p_2.\nu = 0 \text{ on } \Gamma_4 \times (0, T_f]. \end{cases} \tag{18}$$

The meaning and units of all variables and parameters used in the model are presented in Table 1.

3 Qualitative Behavior of the Total Mass

In what follows we analyze the time behavior of the total mass of drug,

$$\mathcal{M}(t) = \sum_{i=1,2} \int_{\Omega_i} c_i(t) dxi,$$

Table 1 Values for the parameters in the healthy and cancer tissues

Variable/parameter	Definition	Value
$c_i, i = 1, 2$	Concentration of the drug	–
A_1	Permeability coefficient	10^{-8} [m/s]
$D_{1,0}$	Diffusion coefficient	10^{-10} [m^2/s]
$c_{1,0}$	Initial concentration of the drug	10^{-2} [mol/m^3]
$\alpha_i, i = 1, 2$	Attenuation coefficient	8.3×10^{-3} [Np/m]
$\beta_i, i = 1, 2$	Sound speed	1500 [m/s]
$\phi_i, i = 1, 2, 3$	Positive constant	2×10^{-4} [m/(Pa.s)]
$v_{i,0}, i = 1, 2$	Conductive velocity	2.06×10^{-3} [m/s]
$\psi_i, i = 1, 2$	Positive constant	110^{-4} [m]
E_1	Stiffness coefficient of the single spring	1.2294×10^{-5} [Pa]
E_2	Stiffness coefficient of the spring	1.7239×10^{-5} [Pa]
μ	Stiffness coefficient of the dashpot	17.7432×10^{-4}
D_v	Viscoelastic diffusion coefficient	10^{-12} [mol/(m^3. s. Pa)]
A_2	Partition coefficient	10^{-8} [m/s]
$D_{2,0}$	Diffusion coefficient of the drug	3.6×10^{-10} [m^2/s]
$c_{2,0}$	Initial concentration of the drug	0 [mol/m^3]
σ_0	Initial stress	10^{-3} [Pa]
ρ	Density	1000 [kg/m^3]
λ	Drug consumption rate	10^{-5} [1/s]

where Ω_1 and Ω_2 stand for the healthy tissue and the solid cancer tissue domains, respectively. As we have

$$\mathcal{M}'(t) = \sum_{i=1,2} \int_{\Omega_i} \frac{\partial c_i}{\partial t}(t)dx,$$

for c_1 and c_2 regular enough, considering (2) and (5) in the equivalent form $\dfrac{\partial c_1}{\partial t}(t) = -\nabla.J_1(t)$, $\dfrac{\partial c_2}{\partial t}(t) = -\nabla.J_2(t) - \lambda_2 c_2(t)$, we obtain

$$\mathcal{M}'(t) = \int_{\Gamma^*} -J_1(t).\eta ds + \int_{\Gamma^{**}} -J_2(t).\eta ds - \lambda \int_{\Omega_2} c_2 dx,$$

where $\Gamma^* = \bigcup_{i=1}^{4} \Gamma_i$ and $\Gamma^{**} = \bigcup_{i=4}^{7} \Gamma_i$. Taking into account the boundary conditions (10)–(12), we get

$$\mathcal{M}'(t) = -A_1 \int_{\Gamma_5 \cup \Gamma_6 \cup \Gamma_7} c_2(t)ds - \lambda \int_{\Omega_2} c_2 dx,$$

that leads to

$$\mathcal{M}(t)=\mathcal{M}(0)-\int_0^t\int_{\Gamma_5\cup\Gamma_6\cup\Gamma_7}A_1c_2(\tau)dsd\tau-\int_0^t\int_{\Omega_2}\lambda c_2dxd\tau,\, t\in[0,T_f]. \tag{19}$$

Mathematically, we conclude that the drug mass in the system at time t, $\mathcal{M}(t)$, is equal to the drug mass at initial time minus the drug mass that passes through the boundary of the target tissue and the drug consumed until the time t which agree with the behavior of the physical system.

4 Stability Analysis

In this section, we study the stability of the coupled problems (1), (2), (5), and (7) (see [28]). We start by introducing some notations. Let Ω be a bounded domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. By $L^2(\Omega)$, $H^1(\Omega)$ and $L^2(\partial\Omega)$ we denote the usual Sobolev spaces endowed with the usual inner products $(.,.)$, $(.,.)_1$ and $(.,.)_{\partial\Omega}$, respectively, and norms $\|.\|_{L^2(\Omega)}$, $\|.\|_{H^1(\Omega)}$ and $\|.\|_{L^2(\partial\Omega)}$, respectively. The usual inner product in $[L^2(\Omega)]^2$ is denoted by $((.,.))$. By $L^2(0,T_f;H^1(\Omega))$ and $L^2(0,T_f;L^2(\Omega))$ we represent, respectively, the space of functions $u:(0,T_f)\to H^1(\Omega)$ and $u:(0,T_f)\to L^2(\Omega)$ such that

$$\int_0^{T_f}\|u(t)\|^2_{H^1(\Omega)}dt<+\infty,\quad \int_0^{T_f}\|u(t)\|^2_{L^2(\Omega)}dt<+\infty.$$

We also introduce the space $H^1_{\Gamma_1}(\Omega_1)=\{w_1\in H^1(\Omega_1):w_1=0 \text{ on } \Gamma_1\}$ and the spaces $H^2(0,T_f,L^2(\Omega_i)), i=1,2,$ given by the space of function $w\in L^2(0,T_f,L^2(\Omega_i))$ such that the weak derivatives $w^{(j)}\in L^2(0,T_f,L^2(\Omega_i)), j=1,2, i=1,2.$

In what follows we consider the weak solution of the initial boundary value problem (IBVP) (1), (2), (8), and (18) with general initial conditions defined by the following: for $i=1,2$, $p_i\in H^2(0,T_f,L^2(\Omega_i))\cap L^2(0,T_f,H^1(\Omega_i))$, and $p_1(t)=p_{\Gamma_1}$ on $(0,T_f)\times\Gamma_1$,

$$\sum_{i=1,2}\Big(\frac{\partial^2 p_i}{\partial t^2}(t),w_i\Big)+\sum_{i=1,2}\Big(2\alpha_i\beta_i\frac{\partial p_i}{\partial t}(t)),w_i\Big)=-\sum_{i=1,2}\Big(\big(\beta_i^2\nabla p_i(t),\nabla w_i\big)\Big), \tag{20}$$

for all $w_1\in H^1_{\Gamma_1}(\Omega_1)$ and for all $w_2\in H^1(\Omega_2)$,

$$(p_i(0),w_i)=(p_{0,i},w_i),\forall w_i\in L^2(\Omega_i),\ (p_i'(0),w_i)=(p_{d,i},w_i),\forall w_i\in L^2(\Omega_i), \tag{21}$$

for $i = 1, 2$, $c_i \in H^1(0, T_f, L^2(\Omega_i)) \cap L^2(0, T_f, H^1(\Omega_i))$ and

$$\sum_{i=1,2} \Big(\frac{\partial c_i}{\partial t}(t), w_{i+2}\Big) = \sum_{i=1,2} \Big(\big(v_i(t)c_i(t), \nabla w_{i+2}\big)\Big) - \sum_{i=1,2} \Big(\big(D_i \nabla c_i(t), \nabla w_{i+2}\big)\Big)$$
$$-\big(A_2(c_1(t) - c_2(t)), w_3 - w_4\big)_{\Gamma_4} - \big(A_1 c_2(t), w_4\big)_{\Gamma_5\cup\Gamma_6\cup\Gamma_7} - \lambda(c_2(t), w_4)$$
$$-\Big(\frac{D_v E_2^2 \gamma}{\mu} \int_0^t e^{-\frac{E_2}{\mu}(t-s)} \nabla c_2(s)ds, \nabla w_4(t)\Big), \tag{22}$$

for all $w_{i+2} \in H^1(\Omega_i)$, $i = 1, 2$,

$$(c_1(0), w_3) = (c_{1,0}, w_3),\ \forall w_3 \in L^2(\Omega_1),\ \ (c_2(0), w_4) = (c_{2,0}, w_4),\ w_4 \in L^2(\Omega_2). \tag{23}$$

In (22), to simplify the notation, D_2^* was represented only by D_2.

To simplify the analysis, in what follows we assume the convective velocities depend only on the acoustic pressure, that is, $v_i = (v_{1,i}(p_i), v_{2,i}(p_i))$, $i = 1, 2$, and

$$|v_{j,i}(x)| \leq \beta_0 |x|, x \in \mathbb{R}, j = 1, 2, i = 1, 2,$$
$$D_1 \in C_b^1(\mathbb{R}) \text{ and } D_1 \geq \chi_1 > 0 \text{ in } \mathbb{R},$$
$$D_2 \in C_b^1(\mathbb{R}) \text{ and } D_2 \geq \chi_2 > 0 \text{ in } \mathbb{R},$$

where $C_b^1(\mathbb{R})$ denotes the space of bounded functions with bounded first order derivatives in $\mathbb{R}$, χ_1 and χ_2 are positive constants. The previous assumptions will be used to obtain an upper bound for p_1, p_2, c_1 and c_2.

1. Energy estimates for the acoustic pressure: We assume that $p_{\Gamma_1} = 0$ and p_i are such that $\nabla \frac{\partial p_i}{\partial t}(t) = \frac{\partial}{\partial t}(\nabla p_i(t))$ almost everywhere in Ω_i and $p_{0,i} \in H^1(\Omega_i)$, $p_{d,i} \in L^2(\Omega_i)$ for $i = 1, 2$.

   Taking in (20) $w_i = \frac{\partial p_i}{\partial t}(t)$, we get

$$\frac{1}{2}\sum_{i=1,2} \frac{d}{dt}\Big\|\frac{\partial p_i}{\partial t}(t)\Big\|^2 + \sum_{i=1,2} 2\alpha_i\beta_i \Big\|\frac{\partial p_i}{\partial t}(t)\Big\|^2 = -\sum_{i=1,2}\Big(\big(\beta_i^2 \nabla p_i(t), \nabla \frac{\partial p_i}{\partial t}(t)\big)\Big),$$

that can be written in the following equivalent form

$$\frac{1}{2}\sum_{i=1,2} \frac{d}{dt}\Big(\Big\|\frac{\partial p_i}{\partial t}(t)\Big\|^2 + \beta_i^2 \|\nabla p_i(t)\|^2\Big) + \sum_{i=1,2} 2\alpha_i\beta_i \Big\|\frac{\partial p_i}{\partial t}(t)\Big\|^2 = 0.$$

As $p_i \in H^2(0, T_f, L^2(\Omega_i))$, then

$$\sum_{i=1,2} \big(\|\frac{\partial p_i}{\partial t}(t)\|^2 + \beta_i^2 \|\nabla p_i(t)\|^2\big) + \sum_{i=1,2} 4\alpha_i \beta_i \int_0^t \|\frac{\partial p_i}{\partial t}(s)\|^2 ds$$
$$= \sum_{i=1,2} \big(\|p_{d,i}\|^2 + \beta_i^2 \|\nabla p_{0,i}\|^2\big), \qquad (24)$$

for $t \in [0, T_f]$.

We observe that as $p_i(t) \in H^1_{\Gamma_1}(\Omega_i)$, holds the Poincaré- Friedrichs inequality, that is, there exists a positive constant C_P such that

$$\|p_i(t)\| \leq C_P \|\nabla p_i(t)\|,$$

and consequently, the conservation relation (24) allows us to obtain the following upper bound

$$\sum_{i=1,2} \big(\|\frac{\partial p_i}{\partial t}(t)\|^2 + \beta_i^2 \|p_i(t)\|^2_{H^1}\big) + \sum_{i=1,2} 4\alpha_i \beta_i \int_0^t \|\frac{\partial p_i}{\partial t}(s)\|^2 ds \qquad (25)$$
$$\leq C \sum_{i=1,2} \big(\|p_{d,i}\|^2 + \beta_i^2 \|\nabla p_{0,i}\|^2\big),$$

for a positive constant C, t independent.

We observe that to obtain an estimate for $\|p_i(t)\|_{L^\infty}$ we need to assume some regularity. While for the one dimensional case, $H^1(\Omega_i)$ is embedded in $C(\overline{\Omega}_i)$ and, consequently, the upper bound (25) leads to

$$\sum_{i=1,2} \big(\|\frac{\partial p_i}{\partial t}(t)\|^2 + \beta_i^2 \|p_i(t)\|^2_{L^\infty}\big) + \sum_{i=1,2} 4\alpha_i \beta_i \int_0^t \|\frac{\partial p_i}{\partial t}(s)\|^2 ds$$
$$\leq C \sum_{i=1,2} \big(\|p_{d,i}\|^2 + \beta_i^2 \|\nabla p_{0,i}\|^2\big),$$

in our situation we need to increase the regularity of our data. In fact, assuming that $p_{0,i} \in H^2(\Omega_i)$, $p_{d,i} \in H^1(\Omega_i)$ for $i = 1, 2$, and considering the acoustic pressure problem for p_i replaced by the corresponding problem for ∇p_i, following the proof of (24), it can be shown that

$$\sum_{i=1,2} \big(\|\frac{\partial}{\partial t}(\nabla p_i)(t)\|^2 + \beta_i^2 |p_i(t)|^2_{H^2}\big) + \sum_{i=1,2} 4\alpha_i \beta_i \int_0^t \|\frac{\partial}{\partial t}\nabla p_i(s)\|^2 ds$$
$$= \sum_{i=1,2} \big(\|\nabla p_{d,i}\|^2 + \beta_i^2 |p_{0,i}|^2_{H^2}\big), (26)$$

where $|p_i(t)|_{H^2}$ denotes the semi-norm in $H^2(\Omega)$, that is

$$|p_i(t)|^2_{H^2} = \sum_{|\omega|=2} \|\frac{\partial^{|\alpha|} p_i}{\partial x_1^{\omega_1}\partial^{\omega_2}}(t)\|^2,\ \omega = (\omega_1, \omega_2), \omega_j \in \mathbb{N}_0,\ j = 1, 2, |\omega| = \omega_1 + \omega_2.$$

From (24) and (26) we get

$$\sum_{i=1,2} \big(\|\frac{\partial pc_i}{\partial t}(t)\|^2_{H^1} + \beta_i^2\|p_i(t)\|^2_{H^2}\big) + \sum_{i=1,2} 4\alpha_i\beta_i \int_0^t \|\frac{\partial p_i}{\partial t}(s)\|^2_{H^1} ds$$
$$\leq C \sum_{i=1,2} \big(\|p_{d,i}\|^2_{H^1} + \beta_i^2\|p_{0,i}\|^2_{H^2}\big).$$

Taking now into account that $H^2(\Omega_i)$ is embedded in $C(\overline{\Omega}_i)$, we conclude

$$\sum_{i=1,2} \big(\|\frac{\partial p_i}{\partial t}(t)\|^2_{H^1} + \beta_i^2\|p_i(t)\|^2_{L^\infty}\big) + \sum_{i=1,2} 4\alpha_i\beta_i \int_0^t \|\frac{\partial p_i}{\partial t}(s)\|^2_{H^1} ds$$
$$\leq C \sum_{i=1,2} \big(\|p_{d,i}\|^2_{H^1} + \beta_i^2\|p_{0,i}\|^2_{H^2}\big). \quad (27)$$

From inequality (27) we conclude the existence of positive constant C, t and p_i independent, such that

$$\max_{i=1,2} \|p_i(t)\|^2_{L^\infty} \leq C \sum_{i=1,2} \big(\|p_{d,i}\|^2_{H^1} + \beta_i^2\|p_{0,i}\|^2_{H^2}\big). \quad (28)$$

2. Energy estimates for the concentrations: Taking in (22) $w_3 = c_1(t)$ and $w_4 = c_2(t)$ we get

$$\frac{1}{2}\sum_{i=1,2} \frac{d}{dt}\|c_i(t)\|^2 \leq -\chi_1\|\nabla c_1(t)\|^2 - \chi_2\|\nabla c_2(t)\|^2 - A_2\|c_1(t) - c_2(t)\|^2_{\Gamma_4}$$
$$-\Big(\big(\frac{D_v E_2^2\gamma}{\mu}\int_0^t e^{-\frac{E_2}{\mu}(t-s)}\nabla c_2(s)ds, \nabla c_2(t)\big)\Big) \quad (29)$$
$$+\sum_{i=1,2} \big((v_i(t)c_i(t), \nabla c_i(t))\big) - A_1\|c_2(t)\|^2_{\Gamma_5\cup\Gamma_6\cup\Gamma_7} - \lambda\|c_2(t)\|^2.$$

It is easy to show the following estimates:

$$|((v_i(p_i(t))c_i(t), \nabla c_i(t)))| \leq \beta_0\|p_i(t)\|_{L^\infty}\|c_i(t)\|\|\nabla c_i(t)\|$$
$$\leq \frac{1}{4\varepsilon_i^2}\beta_0^2\|p_i(t)\|^2_{L^\infty}\|c_i(t)\|^2 + \varepsilon_i^2\|\nabla c_i(t)\|^2,\ i = 1, 2, \quad (30)$$

and

$$
\begin{aligned}
-\big(\tfrac{D_v E_2^2\gamma}{\mu}\textstyle\int_0^t e^{-\frac{E_2}{\mu}(t-s)}\nabla c_2(s)ds, \nabla c_2(t)\big) \leq \varepsilon_3^2\big\|\nabla c_2(t)\big\|^2 \\
+\frac{D_v^2E_2^3\gamma}{8\varepsilon_3^2\mu^2}\int_0^t \big\|\nabla c_2(s)\big\|^2 ds,
\end{aligned}
\tag{31}
$$

where $\varepsilon_i \neq 0, i = 1, 2, 3$, are arbitrary constants.

Then, from (29), (30), and (31), we obtain

$$
\begin{aligned}
&\frac{d}{dt}\sum_{i=1,2}\big\|c_i(t)\big\|^2 + \sum_{i=1,2} 2(\chi_i - \varepsilon_i^2)\|\nabla c_i(t)\|^2 - 2\varepsilon_3^2\|\nabla c_2(t)\|^2 \\
&+ 2A_2\|c_1(t) - c_2(t)\|_{\Gamma_4}^2 + 2A_1\big\|c_2(t)\big\|_{\Gamma_5\cup\Gamma_6\cup\Gamma_7}^2 \\
&\leq \frac{D_v^2E_2^3\gamma}{4\varepsilon_3^2\mu^2}\int_0^t \big\|\nabla c_2(s)\big\|^2 ds + \sum_{i=1,2}\frac{1}{2\varepsilon_i^2}\beta_0^2\|p_i(t)\|_{L^\infty}^2\|c_i(s)\|^2 - 2\lambda\|c_2(t)\|^2.
\end{aligned}
$$

Then, with $\varepsilon_1^2 = \frac{\chi_1}{2}$ and $\varepsilon_2^2 = \varepsilon_3^2 = \frac{\chi_2}{4}$, we have

$$
\begin{aligned}
&\sum_{i=1,2}\big\|c_i(t)\big\|^2 + \int_0^t \Big(\sum_{i=1,2}\chi_i\|\nabla c_i(s)\|^2\Big) \\
&\quad + \int_0^t \Big(2A_2\|c_1(s) - c_2(s)\|_{\Gamma_4}^2 + 2A_1\big\|c_2(s)\big\|_{\Gamma_5\cup\Gamma_6\cup\Gamma_7}^2\Big) ds \\
&\leq \sum_{i=1,2}\big\|c_i(0)\big\|^2 + \frac{D_v^2E_2^3\gamma}{\chi_2\mu}\int_0^t\int_0^s \big\|\nabla c_2(\theta)\big\|^2 d\theta ds \\
&\quad + \int_0^t \Big(\max\{\frac{\beta_0^2}{\chi_1}\|p_1(s)\|_{L^\infty}^2, 2\frac{\beta_0^2}{\chi_2}\|p_2(s)\|_{L^\infty}^2 - 2\lambda\}\sum_{i=1,2}\big\|c_i(s)\big\|^2\Big) ds,
\end{aligned}
$$

that is, with

$$
\begin{aligned}
E(t) = \sum_{i=1,2}\big\|c_i(t)\big\|^2 + \sum_{i=1,2}\chi_i\int_0^t \|\nabla c_i(s)\|^2 ds \\
+2\int_0^t \big(A_2\|c_1(s) - c_2(s)\|_{\Gamma_4}^2 + A_1\big\|c_2(s)\big\|_{\Gamma_5\cup\Gamma_6\cup\Gamma_7}^2\big) ds,
\end{aligned}
$$

and

$$
h(t) = \max\{\frac{D_v^2E_2^3\gamma}{\chi_2\mu}, \max\{\frac{\beta_0^2}{\chi_1}\|p_1(t)\|_{L^\infty}^2, 2\frac{\beta_0^2}{\chi_2}\|p_2(t)\|_{L^\infty}^2 - 2\lambda\}\},
$$

we have

$$E(t) \leq E(0) + \int_0^t h(s) \int_0^s E(\theta) d\theta ds.$$

Considering the Gronwall lemma to the last inequality, we obtain

$$E(t) \leq E(0) e^{\int_0^t h(\theta) d\theta}, \ t \in [0, T_f]. \tag{32}$$

We remark that combining (32) with (28) we get the upper bound for $E(t)$ in function of the parameters and the initial conditions for the acoustic pressure and for the initial concentrations.

5 Numerical Simulations

In this section we illustrate the behavior of drug concentration in the healthy and cancer tissues. The numerical results were obtained using the commercial software package COMSOL Multiphysics 5.3. The numerical solutions were obtained following the MOL (Method of Lines) approach: spatial discretization that leads to ordinary differential systems (for the pressure and for the concentration) followed by the time integration. In the spatial discretization of the governing equations, we use the piecewise quadratic finite element method defined on the mesh illustrated in Fig. 3.

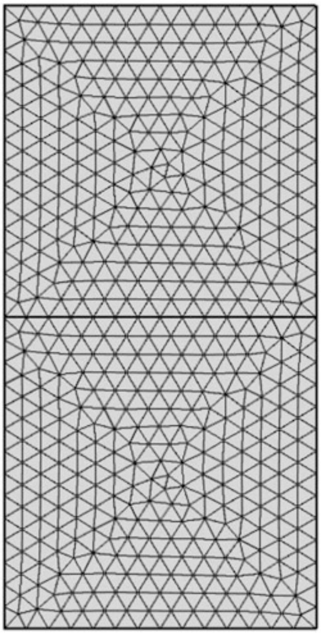


Fig. 3 Computational meshes in the spatial domain

In the time integration of the first order ordinary differential systems, an adaptive Backward Differentiation Formula with order between 1 and 2, with adaptive time step, has been used. The computational time for the reference simulation performed on an Intel (R) Core (TM) i3-4170 3.70 GHz processor and 8.0 GB RAM is around half an hour. Different mesh sizes were used for simulation to verify that the solution is convergent and mesh independent.

The parameters used in the computation of the numerical approximations that we present in what follows are included in Table 1, and they have been extracted from [7, 18], and [36].

The numerical results that we present in what follows are grouped into two sets. In the first set we illustrate the influence of the viscoelastic nature of the tissue on the drug transport, the influence of the stiffness of the cancer tissue that is consequence of the tumor fibrosis characterized by excessive collagen depositions, often crosslinked, as well as the opposition to the transport of drug due to high interstitial fluid pressure. In the second group of results, we intend to illustrate the influence of ultrasound in the rupture of the microenvironment cancer barrier to the drug delivery because ultrasound promotes the convective and diffusive transport.

Drug Transport in Viscoelastic Tissue

Figures 4 and 5 illustrate the drug distribution in the healthy and cancer tissues during 1000s. As time increases, the drug concentration decreases in the healthy tissue and increases in the cancer tissue. We observe that in average, the drug concentration in the healthy tissue decreases and in the cancer tissue increases until $t = 100$s and after decreases. This behavior is consequence of the drug consumption as well as due to the drug transport through the boundary $\Gamma_5 \cup \Gamma_b \cup \Gamma_7$ defined by the condition (11) (Fig. 6).

The parameter D_v is used to take into account the increasing of the opposition to the drug transport in viscoelastic materials due to stiffness. In Fig. 7 we plot the drug mass in the cancer tissue (a) and in the line $x_1 = 0.5$, $x_2 \in (1, 2)$ (b) at $t = 50$ s. As D_v increases, it increases the opposition to the drug transport, and consequently lower values of the drug mass are observed. Fig. 7b also illustrates this effect. In fact, for lower value of D_v, we observe higher values for the drug concentration.

The parameter E_2 is related with the existence of collagen crosslinks in the target tissue. The effect of the increasing of the amount of bounds in the collagen fibers of the target tissue is illustrated in Fig. 8. As the crosslinks increase, increases ECM stiffness and consequently the resistance of the collagen fibers to the drug transport. These facts lead to an increasing of the drug mass in the target tissue for lower values of E_2 (Fig. 8a, b). Lower values of the drug concentration are also observed for higher values of E_2 as illustrated in Fig. 8c.

The effect of the drug consumption in the target tissue in the drug mass is illustrated in Fig. 9. As increases the drug consumption, lower values of drug mass are accumulated in the target tissue—Fig. 9. This effect is also observed in the concentration for $x_1 = 0.5$ and $x_2 \in (1, 2)$. For higher consumption, lower values for the concentration are observed.

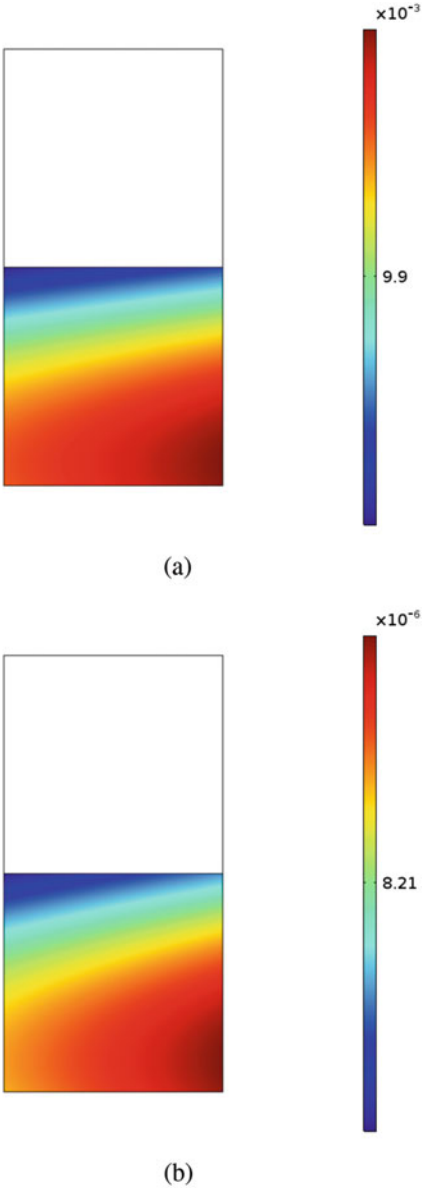
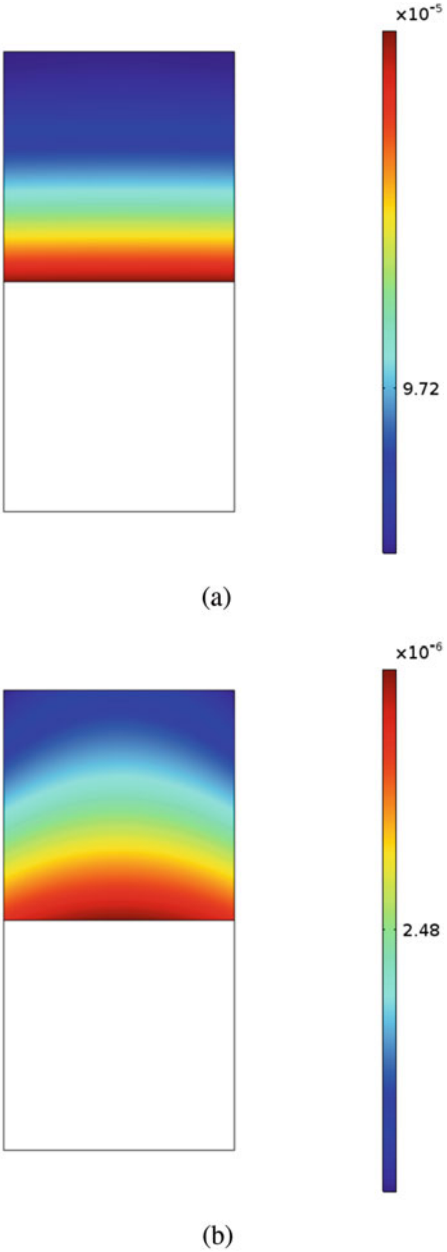


Fig. 4 Drug distribution in the healthy tissue. (**a**) Drug distribution, 1 s. (**b**) Drug distribution, 1000 s

Drug Transport and Ultrasound

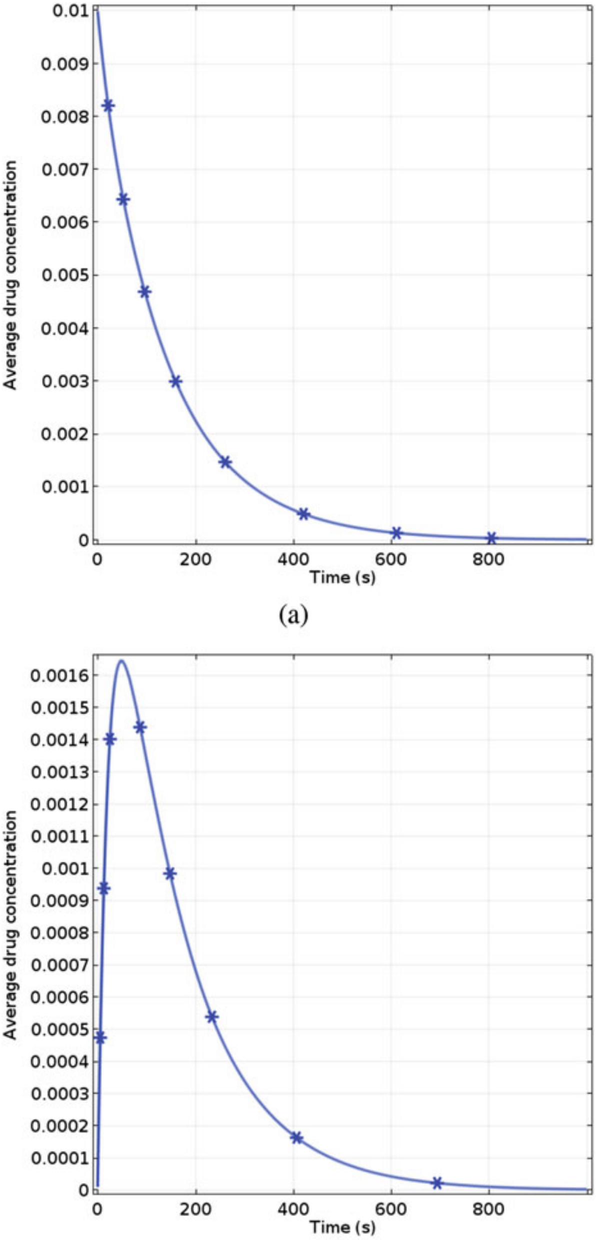
In what follows we plot the drug masses in the target tissue and the drug concentrations for $x_1 = 0.5, x_2 \in (1, 2)$ at $t = 50$ s considering the scenarios defined in the first part but considering the ultrasound effect defined by $p_{\Gamma_1} = 10^{-4}$ Pa. In Fig. 10 we consider different values of D_v. Comparing Figs. 7 and 10, we observe

Fig. 5 Drug distribution in the cancer tissue. (**a**) Drug release, 1 s. (**b**) Drug release, 1000 s



higher drug mass peaks when ultrasound is used as well as higher values for the concentrations for $x_1 = 0.5$, $x_2 \in (1, 2)$ at $t = 50$s.

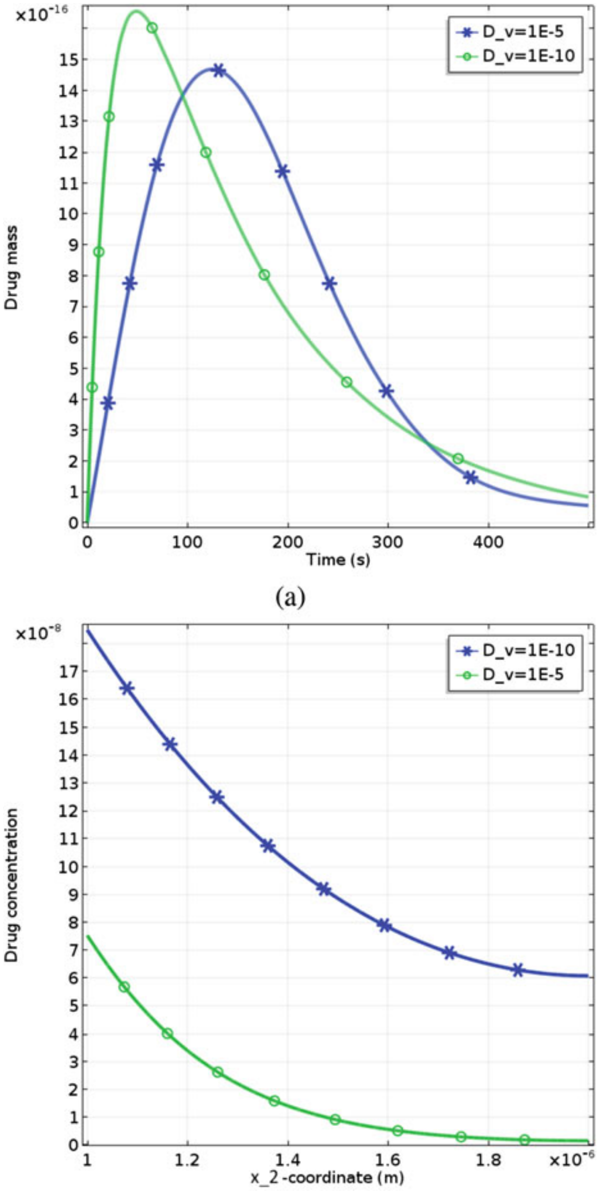
The effectiveness of ultrasound in the promotion of the drug transport through the cancer tissue is also illustrated in Fig. 11 when we consider different values of the coefficient E_2 related with the ECM collagen crosslinks. In fact, when we



(a)

(b)

Fig. 6 Average drug concentration in the healthy tissue (**a**) and in the cancer tissue (**b**)

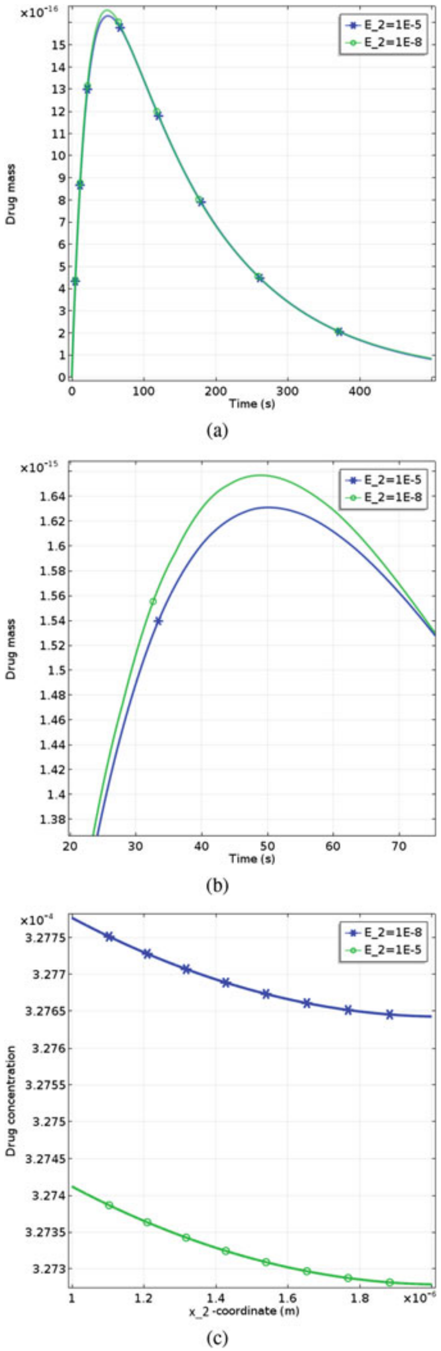


(a)

(b)

Fig. 7 Drug mass in target (**a**) and drug concentrations for $x_1 = 0.5$ and $x_2 \in (1, 2)$ at $t = 50\,\text{s}$ (**b**) for different values of D_v

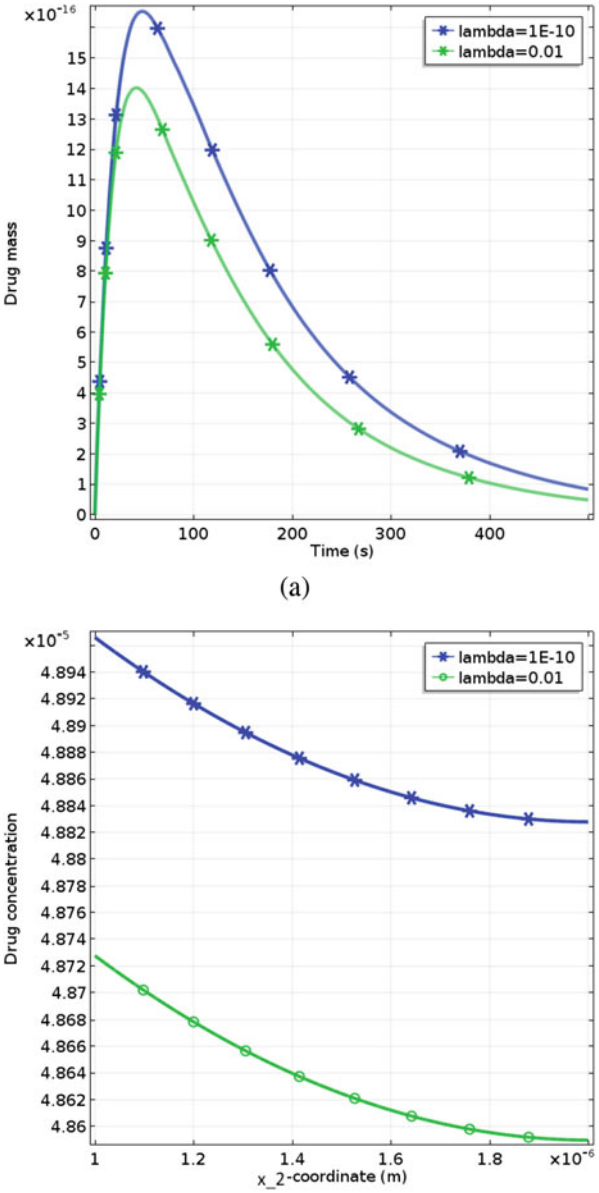
Fig. 8 Drug mass in the target tissue (**a**) and its zoom (**b**), drug concentrations for $x_1 = 0.5$ and $x_2 \in (1, 2)$ at $t = 50\,\text{s}$ (**c**) for different values of E_2



(a)

(b)

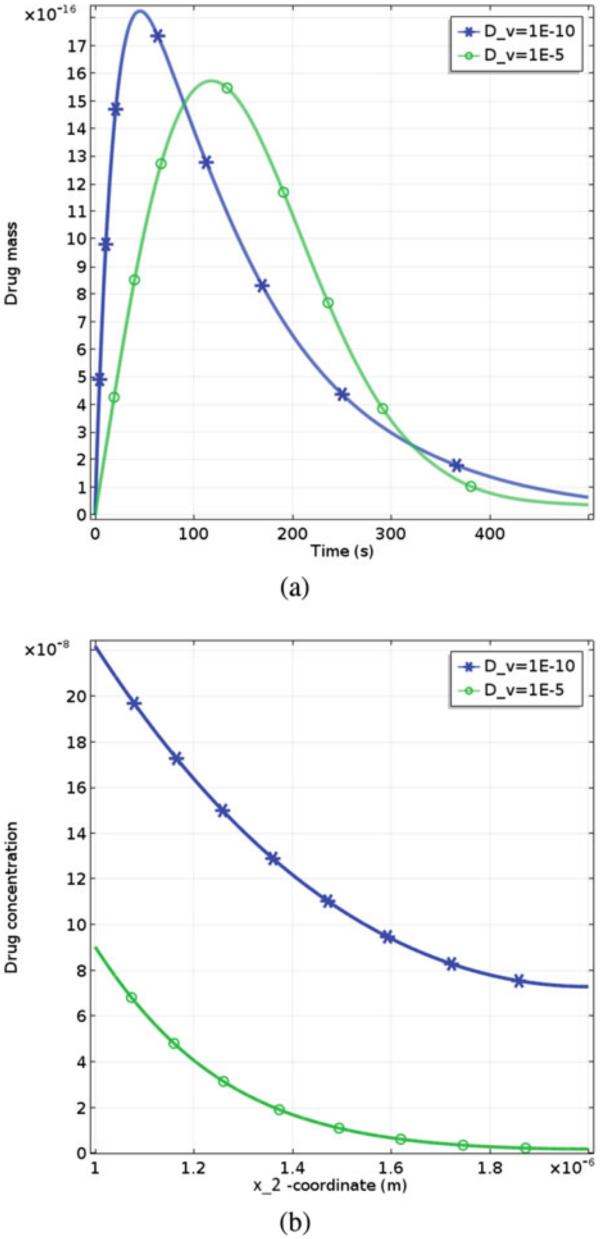
(c)



(a)

(b)

Fig. 9 Drug mass in the target tissue (**a**) and drug concentrations for $x_1 = 0.5$ and $x_2 \in (1, 2)$ at $t = 50\,\text{s}$ (**b**) for different values of λ

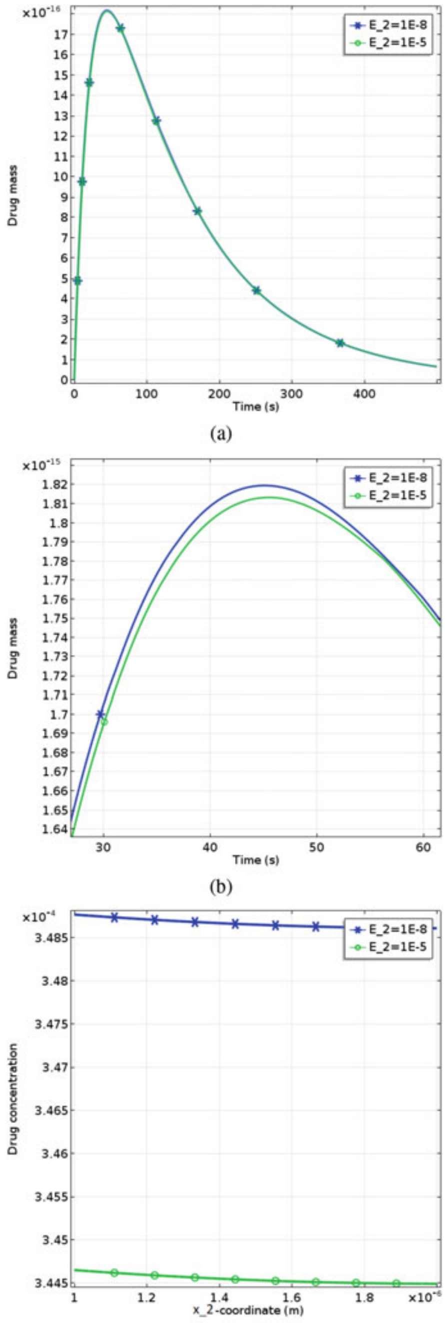


(a)

(b)

Fig. 10 Drug mass in the target tissue (**a**) and drug concentration for $x_1 = 0.5$ and $x_2 \in (1, 2)$ at $t = 50\,\mathrm{s}$ (**b**) for different values of D_v

Fig. 11 Drug mass (**a**), zoom of drug mass (**b**), and drug concentrations for $x = 0.5$ and $x_2 \in (1, 2)$ (**c**) for different values of E_2



(a)

(b)

(c)

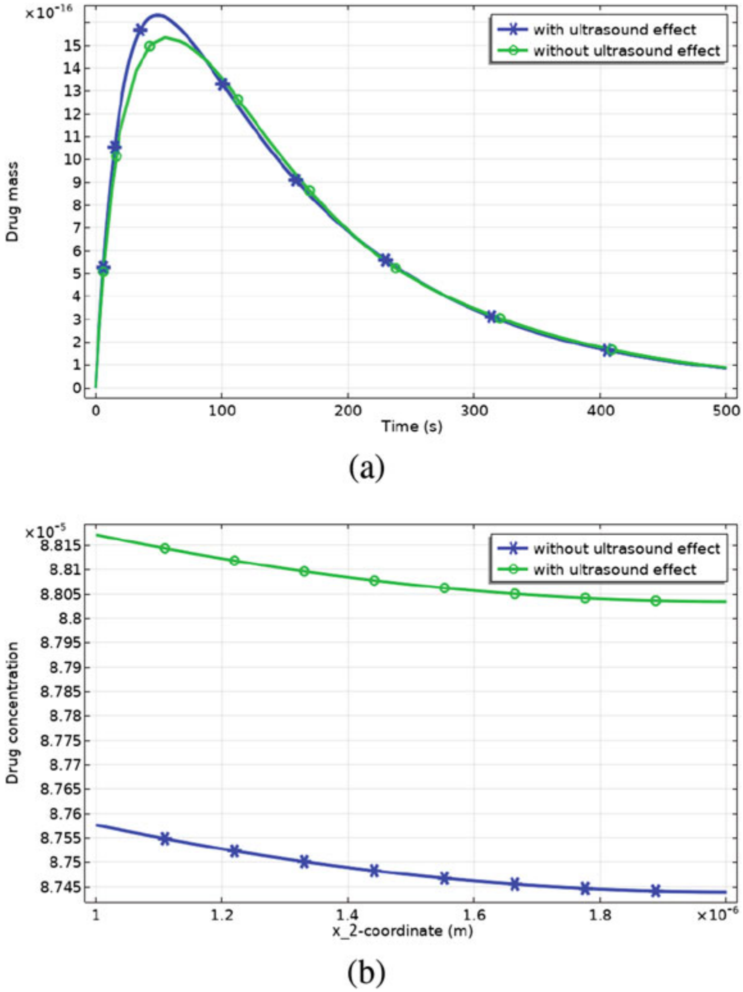


Fig. 12 Drug mass in the target (**a**) and drug concentrations for $x_1 = 0.5$ and $x_2 \in (1, 2)$ (**b**) with and without ultrasound

compare the results in Fig. 11 with the corresponding ones in Fig. 8, we observe higher values for the drug masses and for the drug concentrations when the drug transport is enhanced by ultrasound.

Finally we present in the same figures the drug masses and the drug concentrations obtained with and without ultrasound. From the results presented in Fig. 12, the drug mass peak is higher when ultrasound is used. In what concerns the concentration, the same behavior is observed: in presence of ultrasound, higher concentrations are observed in the target tissue.

6 Conclusions

In this paper a multiphysics and multidomain problem, mathematically written as system of partial differential equations complemented with boundary, interface, and initial conditions defined in (1)–(11), is studied from analytical and numerical point of views. This system describes the drug transport enhanced by ultrasound in healthy and cancer tissues. The mathematical problem was defined taking into account the physiological characteristics of the target, namely, in what concerns the ECM alterations during the cancer evolution, as well as their implications in the drug transport: Fickian description in the healthy tissue and non-Fickian description in the cancer tissue were considered. As the healthy tissue is also a viscoelastic material, a non-Fickian equation could also be considered. However, to take into account the significant differences between both tissues, two different approaches were considered in each domain. Numerical results illustrating the behavior of the unknowns of the mathematical problem are presented. From these results we conclude that system (1)–(11) describes accurately at least qualitatively the drug delivery in a cancer tissue when the drug is administered in the neighboring healthy tissue and the drug transport is enhanced by ultrasound.

Energy estimates where established for the acoustic pressure and for the drug concentration that allow us to conclude that the model will have good stability properties.

From the numerical results presented in the paper, we conclude the following:

1. Due to physiological modifications of the cancer ECM as cancer evolves, the resistance to the drug transport increases. In the mathematical problem, such resistance is associated with the parameters D_v linked with the stiffness and E_2 connected with collagen crosslinks. From Figs. 7 and 8, as these values increase, we conclude that the drug masses in the target decrease.
2. Ultrasound has been used to enhance the drug transport modifying the cancer ECM. In fact, ultrasound induces an increase in the interfibrillar space that is accompanied by an increase in hydraulic conductivity that promotes the fluid flow reducing the interstitial fluid pressure. Ultrasound can also induce pore formation at cellular level. From Figs. 10, 11, and 12, we conclude that ultrasound can be an efficient enhancer of the drug transport in cancer.

Acknowledgments This work was partially supported by the Centre for Mathematics of the University of Coimbra—UIDB/00324/2020, funded by the Portuguese Government through FCT/MCTES.

References

1. Y. Abidine, A. Giannett, J. Revilloud, V. Laurent, C. Verdier, Viscoelastic properties in cancer: from cells to spheroids. Cells **10**, 170 (2021)

2. H. Abyaneh, M. Regenold, T. McKee, C. Allen, M. Gauthier, Towards extracellular matrix normalization for improved treatment of solid tumors. Theranostic **10**, 1060–1980 (2020)
3. E.C. Aifantis, On the problem of diffusion in solids. Acta Mech. **37**, 265–296 (1980)
4. A. Ebrahim, J.A. Ferreira, P. de Oliveira, P. da Silva, Diffusion, viscoelasticity and erosion: analytical study and medical applications. J. Comput. Appl. Math. **275**, 489–501 (2015)
5. E. Azhdari, P. de Oliveira, P.M. da Silva, *Numerical and analytical study of drug release from a biodegradable viscoelastic platform*. Mathematical Methods in Applied Sciences, vol. 39 (2016), pp. 4688–4699
6. M. Bakhtiari-Nejad, S. Shahab, Effects of nonlinear propagation of focused ultrasound on the stable cavitation of a single subble. Acoustics **1**, 14–34 (2019)
7. T. Boissenot, A. Bordat, E. Fattal, N. Tsapis, Ultrasound-triggered drug delivery for cancer treatment using drug delivery systems: from theoretical considerations to practical applications. J. Controlled Release **40**, 2037–2052 (2016)
8. D. Borrmann, A. Danzer, G. Sadowski, Generalized diffusion-relaxation model for solvent sorption in polymers. Ind. Eng. Chem. Res. **60**, 15766–15781 (2021)
9. H.F. Brinson, L.C. Brinson, *Polymer Engineering Science and Viscoelasticity: An Introduction* (Springer, New York, 2010)
10. O. Chaudhuri, J. Cooper-White, P. Janmey, D. Mooney, Effects of extracellular matrix viscoelasticity on cellular behaviour. Nature **584**, 535–546 (2020)
11. C.-H. Fan, C-Y. Lin, H.-L. Liu, Ultrasound targeted CNS gene delivery for Parkinson's disease treatment. J. Control Release **10**, 246–262 (2017)
12. D. Cohen, A. White, Sharp fronts due to diffusion and viscoelastic relaxation in polymers. SIAM J. Appl. Math. **51**, 472–483 (1991)
13. O. Couture, J. Foley, N.F. Kassell, B. Larrat, J.F. Aubry, Review of ultrasound mediated drug delivery for cancer treatment: updates from pre-clinical studies. Transl. Cancer Res. **3**, 494–511 (2014)
14. P.C. Chu, W.Y. Chai, C.H. Tsai, S.T. Kang, C.K. Yeh, H.L. Liu, Focused ultrasound-induced blood-brain barrier opening: association with mechanical index and cavitation index analyzed by dynamic contrast-enhanced magnetic-resonance imaging. Sci. Rep. **6**, 33264 (2016)
15. P.R. Chandran, N. Sandhyarani, An electric field responsive drug delivery system based on chitosan-gold nanocomposites for site specific and controlled delivery of 5 fluorouracil. RSC Adv. **4**, 44922–44929 (2014)
16. D. Dalecki, Mechanical bioeffects of ultrasound. Ann. Rev. Biomed. Eng. **6**, 229–248 (2004)
17. D. De Kee, Q. Liu, J. Hinestroza, Viscoelastic (Non-Fickian) diffusion. Can. J. Chem. Eng. **83**, 913–929 (2008)
18. F.J. Dehkordi, A. Shakeri-Zadeh, S. Khoei, H. Ghadiri, M.B. Shiran, Thermal distribution of ultrasound waves in prostate tumor: comparison of computational modeling with in vivo experiments. ISRN Biomath. **2013**, 428659 (2013)
19. D. Edwards, Constant front speed in weakly diffusive non-Fickian systems. SIAM J. Appl. Math. **55**, 1039–1058 (1995)
20. D. Edwards, Non-Fickian diffusion in thin polymer fielms. J. Poly. Sci. B Poly. Phys. **34**, 981–997 (1996)
21. A. Elosegui-Artola, The extracellular matrix viscoelasticity as a regulator of cell and tissue dynamics. Curr. Opin. Cell Biol. **72**, 10–18 (2021)
22. J.A. Ferreira, M. Grassi, E. Gudino, P. de Oliveira, A 3D model for mechanistic control drug release. SIAM J. Appl. Math. **74**, 620–633 (2014)
23. J.A. Ferreira, D. Jordão, L. Pinto, Approximating coupled hyperbolic—parabolic systems arising in enhanced drug delivery. Comput. Math. Appl. **76**, 81–97 (2018)
24. J.A. Ferreira, D. Jordão, L. Pinto, Drug delivery enhanced by ultrasound: Mathematical modeling and simulation. Comput. Math. Appl. **107**, 57–69 (2022)
25. J.A. Ferreira, M. Grassi, E. Gudiño, P. de Oliveira, A 3D model for mechanistic control drug release. SIAM J. Appl. Math. **74**, 620–633 (2014)

26. J.A. Ferreira, P. de Oliveira, P.M. da Silva, L. Simon, Molecular transport in viscoelastic materials: mechanistic properties and chemical affinities. SIAM J. Appl. Math. **74**, 1598–1614 (2014)
27. J.A. Ferreira, M. Grassi, E. Gudiño, P. de Oliveira, A new look to non-Fickian diffusion. Appl. Math. Model. **39**, 194–204 (2015)
28. J.A. Ferreira, P. de Oliveira, E. Silveira, Drug release enhanced by temperature: an accurate discrete model for solutions in H^3. Comput. Math. Appl. **79**, 852–875 (2020)
29. M. Grassi, G. Grassi, Mathematical modeling and controlled drug delivery: matrix systems. Curr. Drug Delivery **2**, 97–116 (2005)
30. J. Huang, L. Zhang, D. Wang, S. Zheng, S. Lin, Y. Qiao, Extracellular matrix and its therapeutics potential for cancer treatment. Signal Transduction Targeted Ther. **6**, 153 (2021)
31. S. Keller, M. Averkiou, The role of ultrasound in modulating interstetial fluid pressure in solid tumors for improved drug delivery. Bioconjugate Chem. (2021)
32. C. Mierke, Viscoelasticity acts as a marker for tumor extracellular matrix characteristics. Front. Cell Dev. Biol. **9**, 785139 (2021)
33. J. Motoyama, T. Hakata, R. Kato, N. Yamashita, T. Morino, T. Kobayashi, H. Honda, Size dependent heat generation of magnetite nanoparticles under AC magnetic field for cancer therapy. BioMagn. Res. Technol. **6**, 4–13 (2008)
34. N. Nissen, M. Karsdal, N. Willumsen, Collagens and Cancer associated fibroblasts in the reactive stroma and its relation to Cancer biology. J. Exp. Clin. Cancer Res. **38**, 115 (2019)
35. J. Palacio-Torralba, S. Hammer, D.W. Good, S.A. McNeill, G.D. Stewart, R.L. Reuben, Y. Chen, Quantitative diagnostics of soft tissue through viscoelastic characterization using time-based instrumented palpation. J. Mech. Behav. Biomed. Mater. **41**, 149–160 (2015)
36. A. Pulkkinen, B. Werner, E. Martin, K. Hynynen, Numerical simulations of clinical focused ultrasound functional neurosurgery. Phys. Med. Biol. **59**(7), 1679–1700 (2014)
37. M. Rahim, N. Jan, S. Khan, H. Shah, A. Madni, A. Khan, A. Jabar, S. Khan, A. Elhissi, Z. Hussain, H. Aziz, M. Sohail, M. Khan, H. Thu, Recent advancements in stimuli responsive drug delivery platforms for active and passive cancer targeting. Caners **13**, 670 (2021)
38. C. Rianna, M. Radmacher, Comparison of viscoelastic properties of cancer and normal thyroid cells on different stiffness substrates. Eur. Biophys. **43**, 309–324 (2017)
39. M. Rezaeian, A. Sedaghatkish, M. Soltani, Numerical modeling of high-intensity focused ultrasound-mediated intraperitoneal delivery of thermosensitive liposomal doxorubicin for cancer chemotherapy. Drug Delivery **26**, 898–917 (2019)
40. M. Sanopoulou, J. Petropoulos, Systematic analysis and model interpretation of micromolecular non-Fickian sorption kinetics in polymer films. Macromolecules **34**, 1400–1410 (2001)
41. P. Taylor, E.C. Aifantis, On the theory of diffusion in linear viscoelastic media. Acta Mech. **44**, 259–298 (1982)
42. N. Thomas, A. Windle, A theory of case II diffusion. Polymer **23**, 529–542 (1982)
43. J. Tu, H. Zhang, J. Yu, C. Liufu, Z. Chen, Ultrasound-mediated microbubble destruction: a new method in cancer immunotherapy. OncoTargets Ther. **11**, 5763–5775 (2018)
44. E.T. Vogler, C.V. Chrysikopoulos, Experimental investigation of acoustically enhanced solute transport in porous media. Geophys. Res. Lett. **29**, 1710 (2002)
45. W. Zhang, A. Capilnasiu, D. Nordsletten, Comparative analysis of nonlinear viscoelastic models across common biomechanical experiments. J. Elast. **145**, 117–152 (2021)
46. P. Wust, B. Hildebrandt, G. Sreenivasa, B. Rau, J. Gellermann, H. Riess, R. Felix, P.M. Schlag, Hyperthermia in combined treatment of cancer. Lancet Oncol. **3**, 487–497 (2002)
47. W. Zhan, W. Gedroyc, X.Y. Xu, Towards a multiphysics modelling framework for thermosensitive liposomal drug delivery to solid tumour combined with focused ultrasound hyperthermia. Biophys. Rep. **5**, 43–59 (2019)

Computational Analysis to Study the Efficiency of Shear-Activated Nano-Therapeutics in the Treatment of Atherosclerosis

Nicholas Jefopoulos and Bong Jae Chung

1 Introduction

Every year in the United States 795,000 people suffer from a stroke, which is a governing cause of long-term disability and the fifth leading cause of death in the country [1, 2]. Approximately 85% of all strokes are ischemic (blockage in blood flow), and intracranial atherosclerosis is a leading cause of ischemic stroke [1, 3]. This study seeks to gain insight into a novel shear-activated nano-therapeutic to treat atherosclerosis and prevent stroke in at-risk patients. Approximately 50% of ischemic strokes occur within the middle cerebral artery (MCA) region [1]. The MCA is positioned within a connection of several arteries located in the brain's inferior region known as the Circle of Willis (CoW) [4]. Plaque formation within the CoW is primarily consigned to its large arteries which includes the MCA [3].

In addition to the prevalence of strokes, a connection between strokes and other diseases necessitates research into medical prevention measures. Intracranial atherosclerosis and ischemic stroke are also risk factors to the development of dementia [3]. A link has been established between atherosclerosis within the CoW and Alzheimer's disease [5]. Hypoperfusion due to CoW plaques could be the contributing factor, as considerable widespread pathologic hemodynamic changes in the brain have been observed in Alzheimer's disease patients [5]. This seems reasonable considering the CoW supplies 80% of the oxygenated blood to the cerebrum, whose functions include reasoning and problem-solving [1].

Antithrombotic therapy, risk factor modification, and lipid-lowering treatments, along with more invasive stenting and bypass surgeries, are all currently being used to treat intracranial atherosclerosis [6]. Apart from healthy lifestyle changes, all of

N. Jefopoulos (✉) · B. J. Chung
Department of Applied Mathematics, Montclair State University, Montclair, NJ, USA
e-mail: jefopoulosn1@mail.montclair.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_14

these treatments are not without their risks, and most do not attempt to remove plaque from arteries. Among the noninvasive treatments, antithrombotic therapy, the use of an antiplatelet or anticoagulant to reduce clotting, comes at the risk of increased bleeding [7]. Lipid-lowering treatments use statins to lower overall cholesterol to slow down the buildup of plaques with possible risk of liver damage and development of type II diabetes [8].

A therapy that is effective in dissolving plaque from arteries is necessary for treating patients at risk of experiencing a stroke. Targeted nano-therapeutics have increasingly been developed and used to dissolve malignant tumors [9]. Specifically targeted nano-therapeutics that take advantage of mechanical forces may be a novel method to attack atherosclerosis in the future. Thrombosed vasculature displays mechanical characteristics which differ from normal blood vessels. In a thrombosed vasculature, the local fluid shear stress (caused mainly by friction) may increase greatly, from under 70 dyne/cm^2 to greater than 1000 dyne/cm^2 (1 dyne $= 1 \times 10^{-5}$ N) [10].

The high fluid shear stress in these locally stenosed regions activates platelets which quickly adhere to the vessel, causing narrowing. Activation of platelets through high fluid shear stress is a major contributing factor to the development of atherosclerotic plaques. Korin et al., as described in their 2012 paper, developed a shear stress activated nano-therapeutic (SA-NT) inspired by platelet shear stress activation to target atherosclerotic plaques [10]. The therapeutic consists of particles that are approximately the size of platelets, between one and five micrometers in diameter. Each particle is an aggregate consisting of smaller nanoparticles. The therapeutic remains intact during normal flow conditions but breaks up into their smaller components when exposed to higher levels of fluid shear stress. These smaller nanoparticles will experience lower drag forces and consequently have greater adherence to the stenosis allowing the therapeutic to be locally targeted and dissolve the atherosclerosis (Fig. 1).

These SA-NTs are constructed by spray-drying solutions of poly-lactic-co-glycolic acid (PLGA) to form a micrometer-sized aggregate composed of smaller nanoparticles. Most other current therapeutics work to stop plaque growth instead of dissolving it, as the SA-NT is designed to do. The great benefit of using targeted SA-NTs is the ability to use a much smaller dosage without compromising effectiveness. It was shown that to clear a pulmonary embolism within mice, this method used ~1/100 the normal dose [10]. SA-NTs, in conjunction with temporary endovascular bypass, have been shown to achieve high rates of re-canalization without the dangers of vascular trauma seen in stent-retriever thrombectomies [11].

While targeting atherosclerosis with high dosage therapeutics is desirable, studies must be conducted to ensure that unwanted side effects are minimized through effective aggregate breakup. Korin et al. determined a shear stress threshold of 100 dyne/cm^2 [10]. Nano-particles breaking off from the aggregate at this, or higher, shear stress intensity or higher were detected at levels that are an 8–12-fold increase as compared to the levels detected under normal shear stress conditions. Using computational fluid dynamics (CFD), they equated this shear stress level to a 60% obstructed vessel. Normal vessels experience a typical level of shear stress of

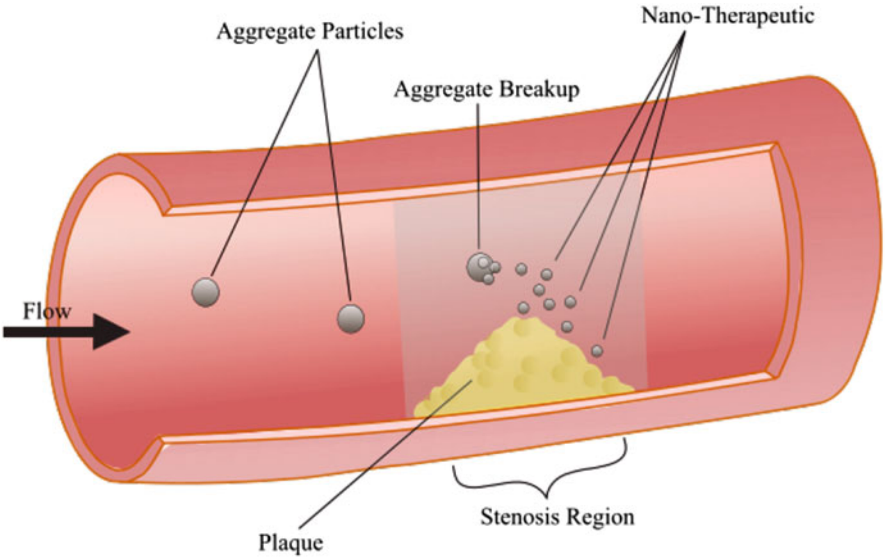


Fig. 1 Aggregate particle breakup into nanoparticle components

approximately 10–30 dyne/cm^2 [10]. Aggregate particle parameters that will allow targeting of vessels less than 60% obstructed are a practical pursuit, as narrowing of 50–69% is considered moderate and may require aggressive treatment, especially if the patient is showing symptoms of the disease [12, 13].

Several numerical studies have been conducted analyzing different aspects of SA-NTs. A study by Qiao et al. was conducted to determine aggregate particle injection sites for stenosed vascular. An idealized curved geometry with three supra-aortic branches was created with a 75% occlusion after the aortic arch. A breakup threshold for the aggregate particle was determined by area, averaging the shear stress rate of the entire aortic wall during one cardiac cycle. This was determined to be 975 s^{-1}. A shear stress rate above the average (1000 s^{-1}) was chosen to be the shear stress rate threshold for the particle. At first, the center point of a radial section after the aortic arch and before the stenosed region was chosen as an injection site. This location was chosen due to its relatively low shear stress rate in order to not cause premature breakup. It was discovered that the aggregate particles would only be broken up at the most severe narrowing, and no nano-therapeutics were discovered in the center of the stenosis during this test. The injection site was then moved to 1 mm away from the aortic wall; this resulted in aggregate particle breakup and nano-therapeutics in the center of the stenosis [14].

A numerical study of topological flow structures formed by atherosclerosis in vessels and its effects on SA-NTs was conducted by Meschi et al. The study focused on a Lagrangian Coherent Structure (LCS) formed by flow separation after the center of the stenosis. The LCS acted as a transport barrier causing a high shear stress rate aiding in aggregate particle breakup. The transport barrier also led to nano-therapeutic accumulation in a post-stenosis region which could result in higher drug absorption.

Numerical studies have been conducted to determine the effect of certain parameters of particles in the bloodstream, which aid in drug development. These parameters were primarily studied to give insight to particle binding which may contribute to the retention of large amounts of toxic particles. Doig et al. studied the influence of particle size compared to average particle residence time in a bifurcated carotid artery using numerical methods [9]. Using an arterial geometry with a diameter of approximately 0.34 cm, the conclusion was that particle size and mean residence time are positively correlated, with the maximum residence time dropping sharply with a reduction in particle size. However, as the particle diameter decreased, the number of particles experiencing wall interactions increased. The test was also run for an arteriole geometry with a diameter of approximately 0.0034 cm. The smaller size allowed Brownian motion to be a larger factor, and the residence time increased by 3% when reducing the particle size from 500 nm to 50 nm [9].

Studies concentrating on SA-NTs and how different parameters influence their breakup have not yet been fully conducted. Additionally the applicability of SA-NTs in the treatment of atherosclerosis in the CoW is not fully understood [10]. The roles that particle density, particle diameter, vessel geometry, stenosis shape, and breakup threshold (shear rate) play in the effectiveness of SA-NTs have not been studied extensively. This study seeks to investigate these parameters and their influence on aggregate breakup position and rate using numerical modeling techniques. For SA-NTs to work as intended, enough of the aggregate must break up at the stenosis. Breakup before or after will not be effective in treating atherosclerosis and could have potentially harmful effects. We will explore the effect of the parameters on breakup position and rate using several idealized arterial geometries. Each geometry will have one of three curvatures and either a concentric or eccentric stenosis.

2 Methods

The computational method involves a number of steps, which we now broadly describe. Details of each step are described in the subsections which follow. The numerical simulations begin by creating idealized arterial geometries. Flow data, in the form of a velocity field, is then calculated using the Navier-Stokes equation for each geometry. A force balance equation is solved using a combination of investigated parameters to determine particle trajectories. The aggregate particle model consists of twenty-five nanoparticles attached to the surface of each aggregate particle. Each nanoparticle has a 1.8×10^{-5} cm diameter which matches the diameter of the nanoparticles in Korin et al.'s 2012 paper [10]. Once a breakup threshold has been met, in this case once the particle reaches a certain angular velocity, the aggregate particle will break up, and its components are tracked as they disperse through the flow. Figure 2 outlines the computation method.

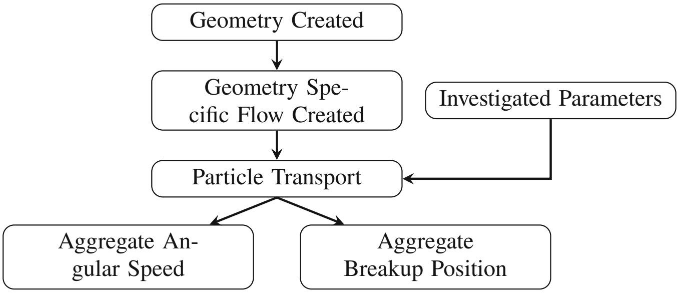


Fig. 2 Overview of the computational method. Idealized geometries are used to create flow data. Flow data is then used in particle transport model along with investigated particle parameters to output aggregate breakup position and angular speed

A total of seven arterial geometries were evaluated. These seven are idealized geometries, with each possessing one of three curvatures. All of the idealized geometries represent a vessel with a diameter of 0.5 cm and a length of 7.0 cm. The flow enters each geometry in a single inlet and exits using a single outlet. Geometry curvatures are either a straight pipe (R_1), a 7.0 cm segment of a 44.56 mm radius torus, i.e., a quarter of a torus (R_2), or a 7.0 cm segment of a 22.28 mm radius torus, i.e., a half of a torus (R_3). Concentric or eccentric (off-center with respect to width) stenoses were created within the center (with respect to length) of each geometry. For the R_1 and R_3 vessels, a unique geometry with 50% occlusion (0.25 cm opening) was created for each combination of stenosis characteristics (concentric or eccentric; and occlusion). In the eccentric R_3 case, superior, inferior, and ventral/dorsal locations also are studied. Figure 3 shows a selection of the arterial geometries, and Table 1 gives a listing of all idealized geometries tested. The inclusion of concentric and eccentric stenoses is due to their dual prevalence in the CoW. One study of 1,220 CoW segments found that 79% of advanced plaques were eccentric and 19% were concentric. The other 2% were completely occluded plaques [3]. All seven idealized geometries were created using FreeCAD software version 0.18 and used a .stl file for CFD analysis to assess blood flow data [15].

2.1 *Computational Fluid Dynamics (CFD) Analysis*

We triangulate the idealized geometries using in-house software for segmentation/model construction (ZMD). Each model is then used as a surface to generate a finite element grid based on an advancing front method. The method uses in-house software (GEN3D) to re-triangulate the surface and generate tetrahedral elements [16, 17].

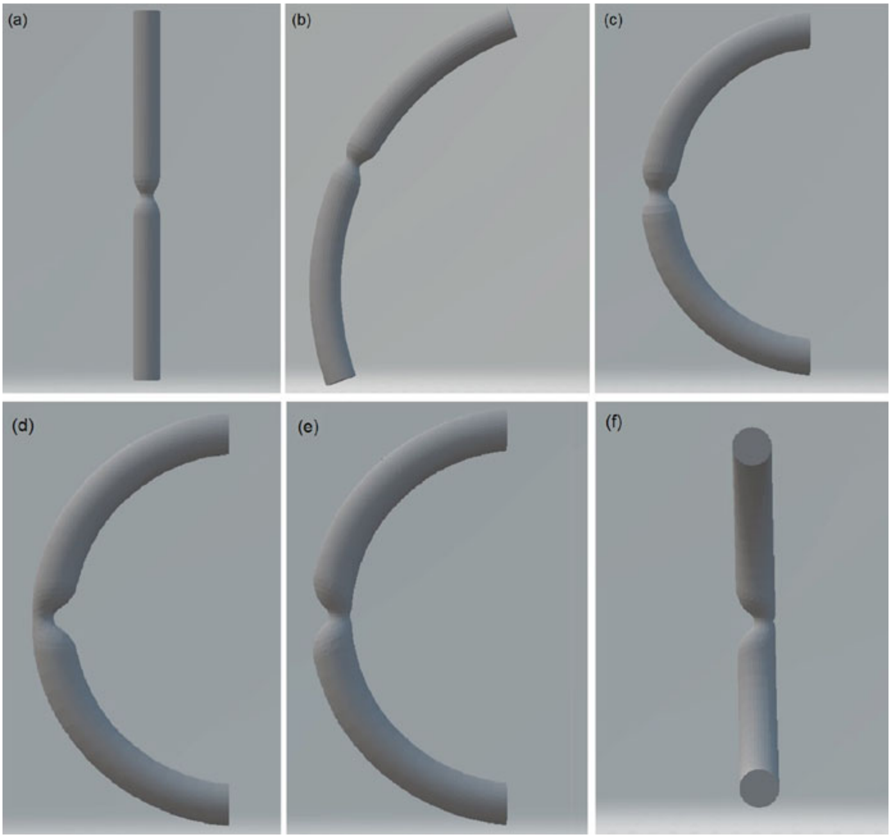


Fig. 3 Idealized geometries. (**a**) R_1 geometry with concentric 50% occlusion. (**b**) R_2 geometry with concentric 50% occlusion. (**c**) R_3 geometry with concentric 50% occlusion. The remaining (**d**–**f**) images show the R_3 geometry with eccentric stenosis in the (**d**) superior, (**e**) inferior, and (**f**) ventral/dorsal locations

Continuity and unsteady Navier-Stokes equations are used to model blood flow as an incompressible Newtonian fluid (density, $\rho = 1.105\,\text{g/cm}^3$ and viscosity, $\mu = 0.04\,\text{Poise}$). The equations are as follows:

$$\nabla \cdot \vec{v} = 0, \tag{1}$$

$$\rho(\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v}) = -\nabla P + \mu \nabla^2 \vec{v}, \tag{2}$$

where $\vec{v}$ is the flow velocity and P is the mechanical pressure. The unsteady flow equations are solved with in-house software that utilizes a fully implicit scheme and efficient solution algorithms (FEFLO) [18, 19]. A parabolic inlet velocity profile and traction-free boundary condition at the outlet is implemented. Vessel wall

Table 1 Overview of the vessel geometries, stenosis shapes, and stenosis locations that were tested

Case number	Vessel geometry	Stenosis shape	Stenosis location
1	R_1	Concentric	–
2	R_1	Eccentric	–
3	R_2	Concentric	–
4	R_3	Concentric	–
5	R_3	Eccentric	Inferior
6	R_3	Eccentric	Superior
7	R_3	Eccentric	Ventral/dorsal

compliance (ability to distend) is neglected for this study, and thus, no-slip boundary conditions are imposed at the walls.

2.2 Particle Trajectories

To attain aggregate and nanoparticle position as well as translational and rotational (angular) velocities, a force balance equation was used. Using Newton's 2nd law $ma = F$, one has

$$S\frac{d\vec{v}_p}{dt} = \vec{F}_{AM} + \vec{F}_B + \vec{F}_D + \vec{F}_L, \tag{3}$$

where S is the specific density of the particle (ratio of particle density and fluid density) , $\frac{d\vec{v}_p}{dt}$ is the acceleration of the particle, $\vec{F}_{AM}$ is the force of added mass, $\vec{F}_B$ is the Bassett force, $\vec{F}_D$ is the drag force, and $\vec{F}_L$ is the lift force [20]. Substitution of the specific forms of these forces leads to the following force balance equation:

$$S\frac{d\vec{v}_p}{dt} = \frac{D\vec{u}}{Dt} + \frac{1}{2}\left(\frac{D\vec{u}}{Dt} - \frac{d\vec{v}_p}{dt}\right) + \frac{3}{4}\frac{C_D}{D}|\vec{u} - \vec{v}_p|(\vec{u} - \vec{v}_p) + \vec{f}_{\text{lift}}, \tag{4}$$

$$\vec{f}_{\text{lift}} = \vec{f}_{\text{lift: shear}} + \vec{f}_{\text{lift: rotational}}, \tag{5}$$

where $\frac{D}{Dt}$ is the material derivative, $\vec{u}$ is the velocity field, C_D is the coefficient of drag, D is the diameter of the particle, $| \cdot |$ denotes the magnitude of the vector, $\vec{f}_{\text{lift: shear}}$ is the shear-induced lift force, and $\vec{f}_{\text{lift: rotational}}$ is the rotation-induced lift or "Magnus force."

This force balance equation is solved using the second-order Runge-Kutta method (midpoint method). This technique approximates the solution of the second-order Taylor expansion without needing to compute derivatives of $f(t, y)$. After aggregate breakup, each nanoparticle is also governed by Brownian motion due to its small size, and thus, we add motion in the form of a scaled pseudo-random vector to the position of each nanoparticle.

2.3 Breakup Criterion

Aggregate particles are designed to be broken up into their nano-therapeutic components when they reach a region of high fluid shear stress [10]. An aggregate particle will experience a fluid shear force on the surface of the particle in that region, which depends on the particle's radial position from the center of the vessel due to its parabolic velocity profile. For instance a particle can have a higher shear force near the wall and have a lower shear rate near the center of the lumen where the particle is not largely influenced by fluid shear force [14]. A fluid shear force on each particle determines the rotational force of the particle. We assume that the threshold for breakup of aggregate particle depends on the magnitude of angular velocity (angular speed); therefore angular speed is used as a numerical criterion for the breakup of each aggregate particle.

2.4 Interpolation of Flow Data

The CFD analysis produces flow data at every 0.01 s. It generates 100 snapshots of flow data including flow velocities and pressures for each cardiac cycle of 1 s. In order to reduce our computational costs for the CFD analysis, the flow data is interpolated into smaller time step sizes.

The binary output files are quite large, ranging from approximately 10–40 MB. This file size made it prudent to develop code to interpolate the flow data within the model rather than creating larger output files. The time step size for the simulation of particle trajectories is determined by a convergence test, which will be discussed in the results section.

The subroutine interpolates the data linearly using

$$t = t_0 + i\frac{t_1 - t_0}{s}, \tag{6}$$

$$\vec{u} = \vec{u}_0 + (t - t_0)\frac{\vec{u}_1 - \vec{u}_0}{t_1 - t_0}, \tag{7}$$

where $\vec{u} = (u, v, w)$ and t are the interpolated velocity and time, respectively, $\vec{u}_0 = (u_0, v_0, w_0)$, and t_0 are the velocities and time from an output file, and $\vec{u}_1 = (u_1, v_1, w_1)$, and t_1 are the velocities and time from the preceding output file, s is the number of interpolated data points between two original data points, and $\{i \in \mathbb{Z} | 1 \le i \le s\}$. The appropriate time step was computed from the convergence test to be 0.002 of the original 0.01 s. This algorithm does not store any of the interpolated data after each iteration in order to minimize computer memory usage.

2.5 Initial Conditions

Aggregate particles are allocated every 50 elements on a plane 0.05 cm from the geometry inlet. The number of elements between aggregate particles was chosen arbitrarily. The number of aggregate particles in the test and their position is dependent on the number of triangular elements that make up the inlet. This method produces between 10 and 20 particles for each geometry. After aggregate particles break up into their nanoparticle components, the aggregate particle is still tracked in the flow as if it had not broken up for the possibility of gaining further insight.

2.6 Particle Ricochet Assumption

The computational method allows particles to exit the geometry at any point in a cycle. Therefore, a method had to be devised to prevent a particle, aggregate or nano-, from leaving the domain through the arterial vessel wall. We assume that a particle hitting the vessel walls is bounced back to the luminal region so that (1) the particle motion obeys the linear momentum conservation law by considering the walls are rigid and (2) there are no biochemical reactions between the walls and particle. To ensure that particles do not leave the geometric domain prematurely, we have developed a ricochet method that enables particles to exit the outlet.

If $\vec{v}$ is the incident vector, $\hat{n}$ is the normal vector to the surface at the point at which $\vec{v}$ hits, then the reflected vector $\vec{w}$ is described by

$$\vec{w} = 2\hat{n}(\vec{v} \cdot \hat{n}) - \vec{v}. \tag{8}$$

The reflected vector has an angle of reflection that is the same as the angle of incidence (Fig. 4).

If a particle leaves the geometry domain but did not exit through the geometry outlet during a time step iteration, the algorithm creates a line between the particle's position at the previous time step and the particle's current position outside the domain. This line consists of 100,000 equally spaced points. The geometric boundaries used in our study are comprised of triangular elements. The centroid (geometric center) of each triangular element and the distance between every point on the created line and every element's centroid is calculated, and the minimum distance is determined. The centroid with the minimum distance from the line is then used to create a unit normal from the surface. Vector $\vec{v}$ is then calculated by finding the component wise distance from the original position (x_1, y_1, z_1) to the centroid and normalized by its magnitude. The reflection vector is scaled by the distance from the geometry surface to the particle's current position outside the domain (x_2, y_2, z_2). Vector $\vec{w}$ is calculated using Eq. 8, scaled, and added to the centroid vector to determine the new position of the particle.

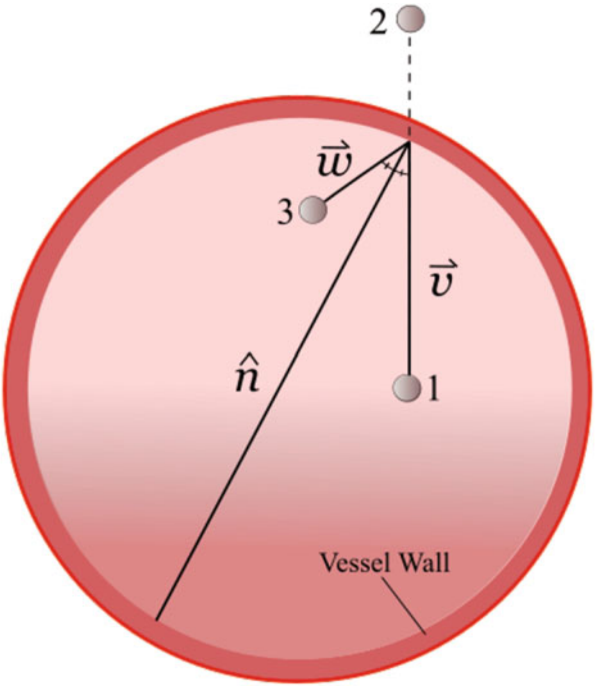


Fig. 4 Geometry of the particle ricochet subroutine. The particle (position 1) has moved outside the domain to position 2. This routine moves the particle back into the domain to a point (position 3) of reflection off the boundary

2.7 Convergence Test

A convergence test was performed to determine a time step at which the discretization error is minimal. A straight arterial vessel with diameter of 0.5 cm and length of 7.0 cm was used as the geometry for this test. It contained a concentric stenosis at its center with a maximum narrowing of 0.25 cm. Eighteen particles were simultaneously tracked, each having a different arbitrary starting position. This test was performed eleven times, starting with the original time step from the binary flow data of 0.01 s and dividing that time step in half for each study thereafter. For each particle, the radial position data was collected from each study. Figure 5 shows the convergence test of a single aggregate particle moving through the R_1 geometry with 50% occlusion.

A comparison of particle position for each particle every 0.01 s was made between each consecutive study. The maximum difference between position data was calculated and normalized using the geometry's radius of 0.5 cm. Using an error threshold of 0.001, it was determined that the difference between study 10

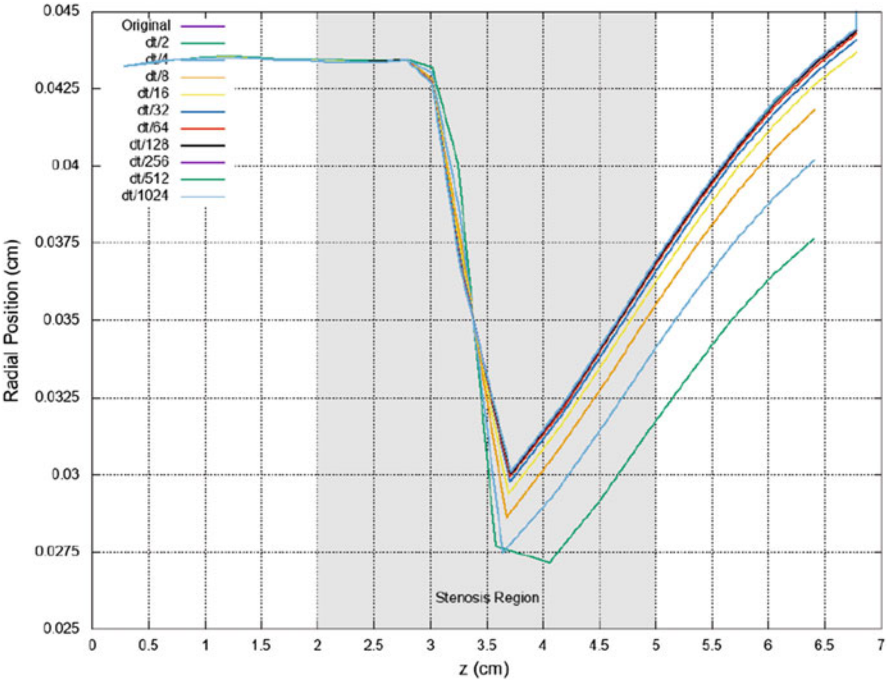


Fig. 5 Convergence test of a single aggregate particle moving through R_1 geometry with concentric 50% occlusion

(0.01 s/512) and study 11 (0.01 s/1024) met the convergence threshold requirements for all particles except one (MP_8). It was observed that the majority of maximum position differences were located near the outlet of the geometry. Since this study is concerned with the region around the stenosis, it was sensible to define a stenosis region and restrict the data analysis to it. The stenosis region is defined as being 1.5 cm before and after the center of the stenosis at 3.5 cm from the inlet. Tables 2 and 3 compare the maximum position with all data and the maximum position contained around the stenosis.

From 2.0–5.0 cm from the inlet, all particles met the threshold requirement between study 10 and 11. Therefore 0.01/512 was determined to be the time step for this model. We use the time step for the rest of our simulations. In conducting the convergence study an interesting correlation was found between initial radial position of the aggregate particle and the normalized maximum difference between position data of the dt/512 versus dt/1024 case. It appears that particles positioned farther away from the stenosis center had a greater gap between the dt/512 and dt/1024 cases, as seen in Fig. 6.

Table 2 Convergence test: normalized Δr by study (all data). Cells highlighted in blue meet the threshold requirement

Study	1	2	3	4	5	6	7	8	9	10
Data Set 1	Original	DT/2	DT/4	DT/8	DT/16	DT/32	DT/64	DT/128	DT/256	DT/512
Data Set 2	DT/2	DT/4	DT/8	DT/16	DT/32	DT/64	DT/128	DT/256	DT/512	DT/1024
Particle	Normalized Δr									
1	9.03E-02	2.71E-02	2.24E-02	1.29E-02	1.00E-02	4.90E-03	2.39E-03	1.29E-03	6.07E-04	9.83E-04
2	6.85E-02	3.05E-02	2.49E-02	1.48E-02	7.73E-03	4.90E-03	2.72E-03	1.90E-03	9.65E-04	4.91E-04
3	1.81E-01	4.36E-02	3.05E-02	1.70E-02	9.13E-03	6.19E-03	3.32E-03	2.56E-03	1.61E-03	8.13E-04
4	5.33E-02	5.68E-02	3.19E-02	1.79E-02	8.26E-03	3.80E-03	1.83E-03	8.99E-04	6.46E-04	3.65E-04
5	2.05E-02	1.12E-02	6.62E-03	3.42E-03	1.79E-03	9.27E-04	4.68E-04	2.37E-04	1.21E-04	5.99E-05
6	1.39E-02	1.03E-02	5.78E-03	2.95E-03	1.47E-03	7.34E-04	3.77E-04	1.90E-04	9.48E-05	4.59E-05
7	5.15E-02	6.04E-02	3.37E-02	1.68E-02	7.86E-03	3.87E-03	1.88E-03	9.16E-04	4.62E-04	2.34E-04
8	1.58E-01	3.13E-02	4.02E-02	2.12E-02	4.42E-02	3.56E-02	1.33E-02	5.65E-03	3.83E-03	1.54E-03
9	6.12E-02	5.74E-02	3.68E-02	1.88E-02	9.39E-03	4.60E-03	2.31E-03	1.19E-03	6.52E-04	2.98E-04
10	2.40E-02	1.38E-02	7.30E-03	4.10E-03	2.17E-03	1.06E-03	5.26E-04	2.67E-04	1.36E-04	7.45E-05
11	1.03E-02	5.12E-03	3.22E-03	1.63E-03	8.10E-04	4.15E-04	2.08E-04	1.05E-04	5.16E-05	2.64E-05
12	2.24E-02	1.14E-02	6.30E-03	3.41E-03	1.73E-03	8.89E-04	4.46E-04	2.24E-04	1.12E-04	5.73E-05
13	7.69E-02	7.06E-02	3.51E-02	2.05E-02	9.92E-03	5.15E-03	2.46E-03	1.24E-03	6.08E-04	3.08E-04
14	1.07E-01	6.20E-02	3.07E-02	3.52E-02	1.87E-02	8.85E-03	3.87E-03	1.88E-03	8.86E-04	4.34E-04
15	8.02E-02	2.71E-02	3.72E-02	2.42E-02	1.55E-02	9.58E-03	4.74E-03	2.90E-03	1.43E-03	7.35E-04
16	5.71E-02	3.49E-02	1.98E-02	1.26E-02	5.73E-03	3.02E-03	1.49E-03	7.51E-04	3.77E-04	1.86E-04
17	9.08E-02	5.80E-02	2.98E-02	1.57E-02	7.56E-03	3.66E-03	1.82E-03	8.96E-04	4.48E-04	2.32E-04
18	1.50E-01	2.46E-02	2.81E-02	2.54E-02	1.17E-02	7.01E-03	3.50E-03	1.72E-03	8.56E-04	4.25E-04

Table 3 Convergence test: normalized Δr by study (stenosis region). Cells highlighted in blue meet the threshold requirement. The stenosis region is defined as being 1.5 cm before and after the center of the stenosis at 3.5 cm from the inlet

Study	1	2	3	4	5	6	7	8	9	10
Data Set 1	Original	DT/2	DT/4	DT/8	DT/16	DT/32	DT/64	DT/128	DT/256	DT/512
Data Set 2	DT/2	DT/4	DT/8	DT/16	DT/32	DT/64	DT/128	DT/256	DT/512	DT/1024
Particle	Normalized Δr									
1	9.03E-02	2.71E-02	2.20E-02	9.04E-03	4.79E-03	2.25E-03	1.01E-03	5.06E-04	5.88E-04	9.83E-04
2	6.85E-02	2.08E-02	2.48E-02	1.48E-02	7.73E-03	4.23E-03	2.07E-03	1.06E-03	5.26E-04	2.64E-04
3	1.81E-01	4.29E-02	3.05E-02	1.70E-02	9.13E-03	4.73E-03	2.52E-03	1.28E-03	6.45E-04	3.17E-04
4	5.33E-02	4.23E-02	2.32E-02	1.37E-02	6.57E-03	3.17E-03	1.55E-03	7.75E-04	3.86E-04	1.90E-04
5	2.05E-02	1.05E-02	6.16E-03	3.38E-03	1.79E-03	9.27E-04	4.68E-04	2.37E-04	1.21E-04	5.35E-05
6	1.39E-02	1.03E-02	5.78E-03	2.95E-03	1.30E-03	6.38E-04	3.28E-04	1.63E-04	8.14E-05	3.96E-05
7	5.15E-02	4.66E-02	2.56E-02	1.38E-02	6.70E-03	3.34E-03	1.65E-03	8.13E-04	4.09E-04	2.10E-04
8	1.58E-01	3.13E-02	4.02E-02	2.12E-02	1.16E-02	5.63E-03	2.82E-03	1.40E-03	7.04E-04	3.71E-04
9	6.12E-02	4.88E-02	3.24E-02	1.79E-02	9.12E-03	4.60E-03	2.31E-03	1.14E-03	5.73E-04	2.92E-04
10	2.40E-02	1.35E-02	7.03E-03	3.91E-03	2.02E-03	9.88E-04	4.92E-04	2.49E-04	1.27E-04	7.45E-05
11	1.03E-02	5.12E-03	3.00E-03	1.57E-03	7.61E-04	3.90E-04	1.96E-04	9.88E-05	4.86E-05	2.46E-05
12	2.24E-02	1.14E-02	6.30E-03	3.34E-03	1.69E-03	8.64E-04	4.34E-04	2.18E-04	1.09E-04	5.43E-05
13	7.69E-02	6.59E-02	3.31E-02	1.99E-02	9.79E-03	5.15E-03	2.46E-03	1.24E-03	6.08E-04	3.08E-04
14	9.47E-02	6.20E-02	3.07E-02	2.06E-02	1.35E-02	6.28E-03	2.98E-03	1.43E-03	7.33E-04	3.33E-04
15	8.02E-02	2.71E-02	3.72E-02	2.11E-02	1.17E-02	5.71E-03	2.79E-03	1.39E-03	7.37E-04	3.67E-04
16	5.30E-02	2.95E-02	1.59E-02	9.21E-03	4.24E-03	2.31E-03	1.15E-03	5.79E-04	2.89E-04	1.42E-04
17	8.21E-02	4.88E-02	2.51E-02	1.38E-02	6.83E-03	3.39E-03	1.70E-03	8.36E-04	4.19E-04	2.20E-04
18	1.45E-01	2.46E-02	2.81E-02	2.02E-02	1.08E-02	5.43E-03	2.68E-03	1.30E-03	5.65E-04	2.83E-04

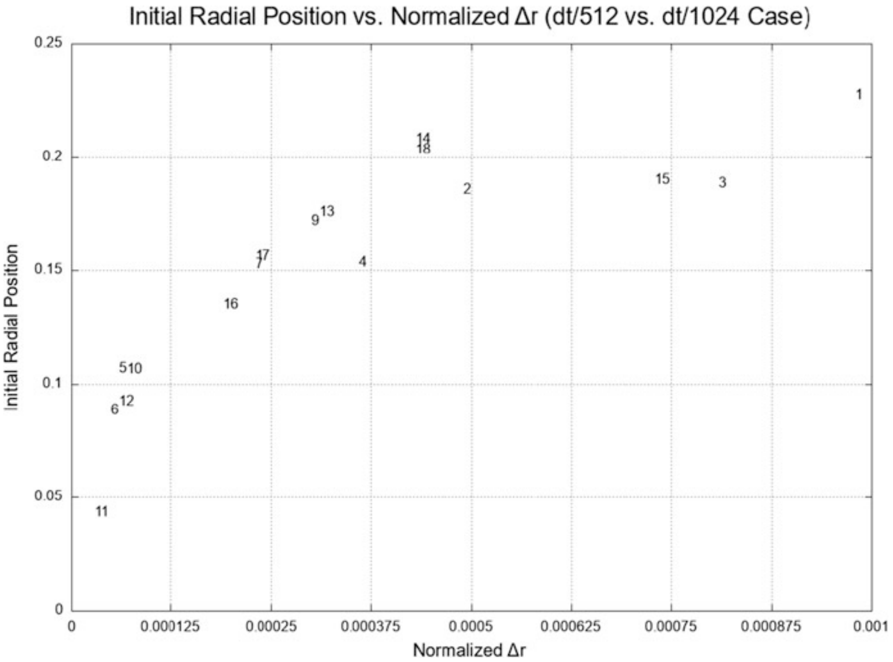


Fig. 6 Convergence test showing the positive correlation between initial radial particle position and normalized maximum difference between position data of dt/512 vs dt/1025 cases

3 Results

3.1 Optimal Breakup Threshold

Angular speed, ω, was used as a breakup threshold for aggregate particles. When the aggregate particles reach an angular speed threshold, they break up into their nanoparticle components. Once an aggregate particle is broken, the twenty-five nanoparticles will break off with the same velocity and angular velocity as the aggregate particle had just prior to breakup as shown in Fig. 1. To determine a breakup threshold, the angular speed at time step, $t = 1$, was calculated for each aggregate particle. The minimum angular speed, ω_{min}, was determined, and different breakup thresholds were generated by scaling ω_{min} to a variety of magnitudes. To determine the optimal breakup threshold for each study, two criteria were measured: the percentage of particles that broke up within the stenosis region and the average distance the aggregate particle broke up from the center of the stenosis. This average includes all aggregate particles, including those that did not break up within the stenosis region as well as those that broke up before and after the stenosis.

For the R_1 and R_3 geometries, we chose the stenosis region to be 1 cm before and 1 cm after the center of the stenosis along the z-axis. This a slightly smaller region than the region used in our original convergence test. The stenosis region on the R_2 geometry consisted of the area spanning 0.5 cm before and 0.5 cm after the center of the stenosis. The shorter region is due to the stenosis lying on a slant relative to the z-axis. The average breakup distance from the center of the stenosis was ascertained by capturing the position of the first appearance of a nanoparticle from each aggregate particle. Due to time and post processing limitations, data from every fifth time step was captured. No biochemical nanoparticle binding components were used in this study to determine if the nanoparticles will adhere to the stenosis after breakup.

The R_1, R_2 and R_3 geometries with a concentric stenosis and 50% occlusion were tested. The specific density and aggregate particle diameter were kept constant at 1.0 and 3.8 μm, respectively. An aggregate particle diameter of 3.8 μm was chosen for consistency with Ref [10].

For the R_1 geometry, fifteen aggregate particles were inserted into the flow to study the effect of different breakup thresholds. Thirteen different breakup thresholds, determined by scaling $\omega_{min} = 14.89$ rad/s for this R_1 geometry, were tested to determine the optimal value. All aggregate particles broke up within the stenosis region when the threshold value was between $3\omega_{min}$ (44.66 rad/s) and $5\omega_{min}$ (74.44 rad/s). At thresholds less than $3\omega_{min}$ (44.66 rad/s) particles broke up prematurely, while at thresholds greater than $5\omega_{min}$ (74.44 rad/s) particles did not break up. By comparing the average distance of particle breakup from the stenosis center, the optimal breakup threshold was determined to be $3\omega_{min}$ (44.66 rad/s) as seen in Fig. 7. This breakup threshold yielded the closest average distance from the stenosis center (0.29 cm), with all particles breaking up after the stenosis center. Particle breakup after the stenosis center may be beneficial in nano-therapeutic residence time due to the development of a LCS which acts as a barrier post-stenosis [21]. Table 4 shows the results for all thirteen breakup thresholds that were tested.

It was observed that the initial particle position was correlated with the proximity of the breakup position to the stenosis center. To ensure this observance was not an error caused by recording the data at every 5th time step (as opposed to recording the data at every time step), the simulation for the R_1 geometry was repeated; this time recording the data at every time step. The least squares linear regression analysis was performed to determine the relationship between the initial radial position of the particle and the proximity of the breakup to the stenosis center. For this analysis the optimal threshold (44.66 rad/s) was used with aggregate particles possessing a 3.8 μm diameter and specific density of 1. The line of best fit that emerged from the analysis is given as $\hat{y} = 0.96 - 9.49r_0$, with an R^2 value of 0.90, and where $\hat{y}$ is the predicted breakup proximity (cm) to the stenosis center, and r_0 is the initial radial particle position. The initial radial position of the aggregate particles is negatively correlated with distance from the breakup position to the stenosis center. As particles are placed into the flow, the farther their initial radial distance is from the geometry center, the closer the breakup occurs to the stenosis center as seen in

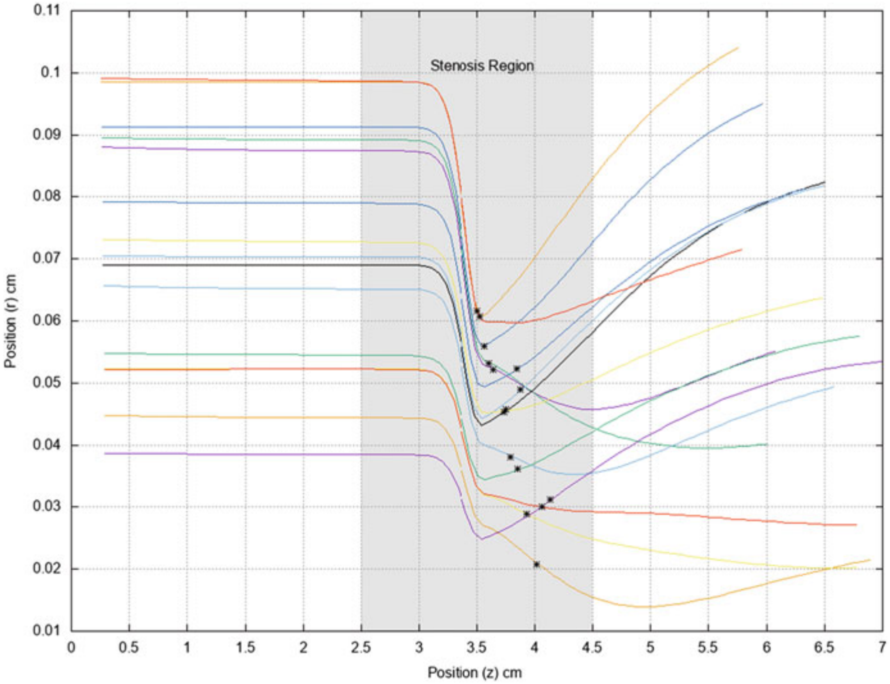


Fig. 7 Radial position as a function of position along the z-axis for aggregate particles flowing through the concentric R_1 vessel geometry with an optimal breakup threshold. Each line represents an individual aggregate particle path, and the stars (*) indicate particle breakup position. Note that aggregate particles continue to be tracked through the flow as though they never broke apart even if they do breakup

Table 4 Breakup threshold data for the R_1 geometry. Note that the shear threshold values are computed using multiples of $\omega_{min} = 14.8879598$ and then rounded to two decimals

Shear Threshold (ω)	Total AP	Total Broken AP	Total Broken in Stenosis Region	Avg Distance from Center (cm)
37.22	15	15	4	1.777
40.32	15	15	12	0.915
43.18	15	15	13	0.701
44.66	15	15	15	0.292
46.15	15	15	15	0.294
46.52	15	15	15	0.304
55.83	15	15	15	0.347
65.13	15	15	15	0.413
74.44	15	15	15	0.466
93.05	15	13	13	0.498
111.66	15	11	9	0.564
148.88	15	6	5	0.673
223.32	15	0	0	-

Fig. 8. This agrees with the optimal aggregate particle injection site proposed by Qiao et al. [14].

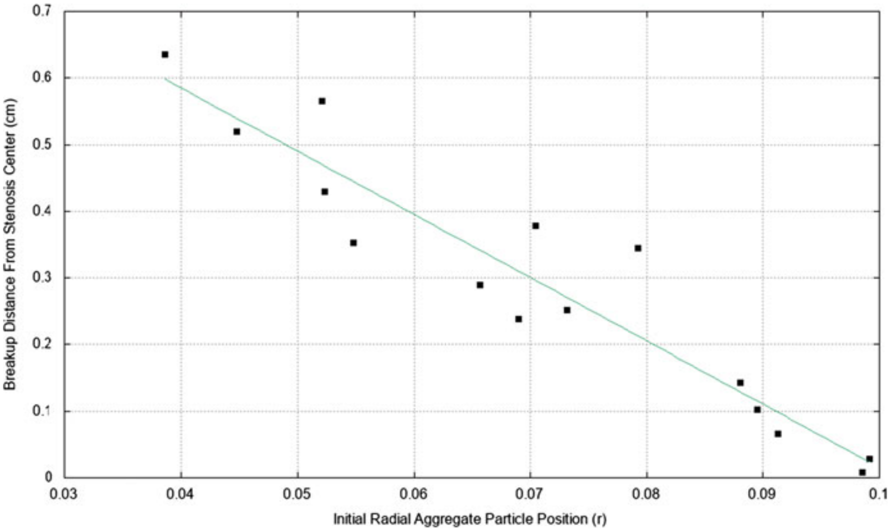


Fig. 8 Breakup distance from stenosis center versus initial radial position of the aggregate particle for the concentric R_1 geometry. The line of best fit shows a negative correlation, and is represented by $\hat{y} = 0.96 - 9.49r_0$ with an R^2 value of 0.90

Table 5 Breakup threshold data for the R_2 geometry. Breakup threshold data for the R_2 geometry. Note that the shear threshold values are computed using multiples of $\omega_{min} = 5.9040012962963$ and then rounded to two decimals

Shear Threshold (ω)	Total AP	Total Broken AP	Total Broken in Stenosis Region	Avg Distance from Center (cm)
44.28	17	17	1	0.992
59.04	17	17	4	0.848
118.08	17	17	9	0.489
121.03	17	17	15	0.221
123.98	17	17	16	0.211
129.89	17	17	16	0.184
135.79	17	17	17	0.150
141.70	17	17	17	0.145
177.12	17	17	17	0.129
283.39	17	17	17	0.104
289.30	17	17	17	0.103
295.20	17	16	16	0.097
324.72	17	16	16	0.097

For the R_2 geometry, seventeen aggregate particles were inserted into the flow to study the effect of different breakup thresholds. Thirteen different breakup thresholds, determined by scaling $\omega_{min} = 5.90$ rad/s for this R_2 geometry, were tested to determine the optimal value. All aggregate particles broke up within the stenosis region when the threshold value was between $23\omega_{min}$ (135.79 rad/s) and $49\omega_{min}$ (289.1 rad/s). The average distance from the stenosis center was minimized for 289.1 rad/s at 0.10 cm from the center. All but one particle broke up before the stenosis center, which is a stark difference to the R_1 geometry in which all particles broke up after the stenosis center. A similar regression analysis of initial

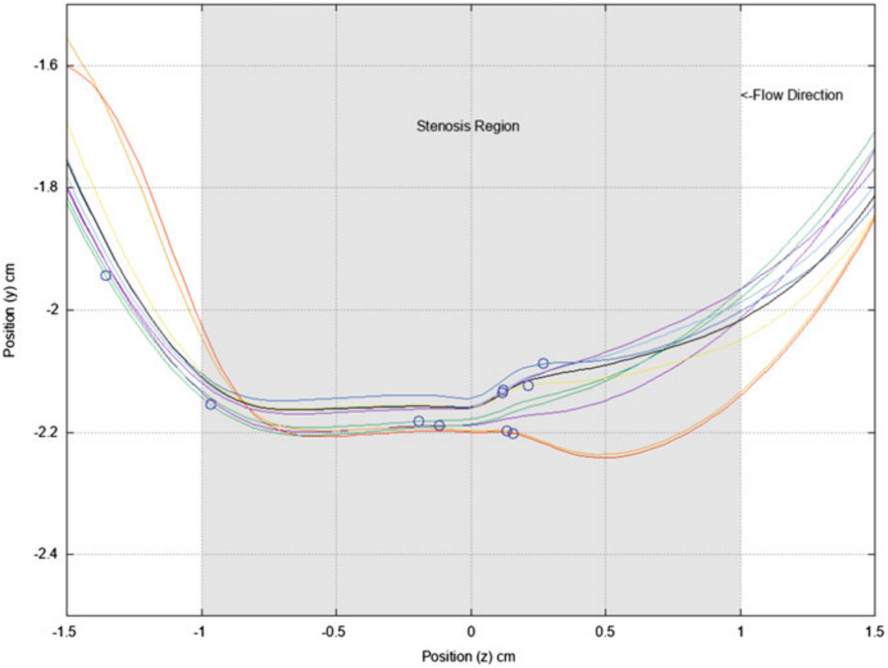


Fig. 9 Position of aggregate particles flowing through the concentric R_3 vessel geometry with an optimal breakup threshold. Each line represents the path of an aggregate particle, and the circles indicate particle breakup position. Note that aggregate particles continue to be tracked through the flow as though they never broke apart even if they do breakup

radial position and breakup position that was completed for the R_1 geometry was conducted. However, unlike the R_1 geometry, no correlation was discovered in the curved R_2 geometry. However, this could be due to the initial radial position being too far away from the stenosis and may hold for particle position when entering the stenosis region. Table 5 shows the results for all thirteen breakup thresholds that were tested.

For the R_3 geometry, ten aggregate particles were inserted into the flow to study the effect of different breakup thresholds. Thirteen different breakup thresholds, determined by scaling $\omega_{min} = 16.38$ rad/s for this R_3 geometry, were tested to determine the optimal value. The maximum number of aggregate particles which broke up within the stenosis region was nine, and occurred when the threshold value was between $10\omega_{min}$ (163.84 rad/s) and $12\omega_{min}$ (196.61 rad/s). No distinct characteristics of the tenth particle were observed. The average distance from the stenosis center was minimal at a threshold value of $12\omega_{min}$ (196.61 rad/s) as shown in Fig. 9. A similar regression analysis to that described previously of initial radial position and breakup position was conducted. No correlation was discovered in the curved R_3 geometry from this analysis. In contrast to the R_1 and R_2 geometries at their optimal thresholds, a slight majority (60%) of aggregate particles broke

Table 6 Breakup threshold data for the R_3 geometry. Breakup threshold data for the R_3 geometry. Note that the shear threshold values are computed using multiples of $\omega_{min} = 16.3839233$ and then rounded to two decimals

Shear Threshold (ω)	Total AP	Total Broken AP	Total Broken in Stenosis Region	Avg Distance from Center (cm)
40.96	10	10	0	2.108
122.88	10	10	4	1.347
147.46	10	10	7	0.584
163.84	10	10	9	0.445
165.48	10	10	9	0.445
175.31	10	10	9	0.482
176.95	10	10	9	0.482
178.58	10	10	9	0.481
180.22	10	10	9	0.488
196.61	10	10	9	0.364
212.99	10	8	8	0.157
245.76	10	6	6	0.143
327.68	10	6	6	0.121

up before the stenosis center in the R_3 geometry. Table 6 shows the results for all thirteen breakup thresholds that were tested.

In summary, the curvature of vessel geometry greatly affects the optimal aggregate particle breakup threshold. The optimal breakup threshold of the R_2 was approximately 6.3 times greater than R_1, and the R_3 curved geometry was approximately 4.5 times greater than that of the R_1 straight geometry. A clear correlation between vessel curvature and optimal breakup threshold was not observed, but it can be said that curvature creates greater complexity, which this study cannot examine fully. The optimal breakup thresholds for both curved cases were at the highest end of the optimal range of thresholds, while the straight case threshold was found at the lower end of its range. It was discovered that for the straight case, a negative correlation exists between initial aggregate particle radial position and average breakup distance from the stenosis; this pattern was not seen in the R_2 or R_3 curved cases.

3.2 Specific Density

Specific density was tested on both the R_1 and R_3 geometries. Aggregate particle diameter and breakup threshold were kept constant at 3.8 µm and $10\omega_{min}$, respectively. Specific densities from 1 to 1.3 were tested. There was no change in the results for these cases. It was determined that in order for specific density to make any noticeable change, the specific density would have to be set to an unrealistic value of 10^4 or above.

Table 7 Aggregate particle diameter data

Diameter (μm)	R_1 Average distance from stenosis center (cm)	R_3 average distance from stenosis center (cm)
1	0.29160212	0.449844683
2	0.291686948	0.449477008
3	0.291774464	0.447882889
3.8	0.29184599	0.445322032
4	0.291864522	0.443923137
5	0.291957333	0.438509313

3.3 Particle Diameter

The effect of particle diameter on breakup threshold was studied using the R_1 and R_3 geometries, both with a 50% concentric occlusion. The specific density was kept constant at 1. Optimal breakup threshold values of 44.66 rad/s for the R_1 geometry and 196.61 rad/s for the R_3 geometry were used. Three particle diameters (1.0, 3.8 and 5.0 μm) were studied for each geometry. These diameters were chosen based on the diameter of natural platelets which lies between 1.0 and 5.0 μm, while 3.8 μm was included because it is the average diameter of fabricated SA-NT aggregate particles [10].

A positive correlation between aggregate particle diameter and average breakup distance from stenosis center was discovered for the R_1 geometry. Using regression analysis, the relationship can be described by the equation $\hat{y} = 0.292 + 0.0000887d$, with $R^2 = 0.99$, and where d is the aggregate particle diameter (μm), and $\hat{y}$ is the predicted breakup proximity (cm) to the stenosis center. Similar analysis was performed for the R_3 geometry, and a negative correlation was found with $\hat{y} = 0.454 - 0.00272d$, with $R^2 = 0.85$.

This demonstrates that curvature matters when choosing an optimal aggregate particle diameter. Smaller particles may be ideal for straight vessels and larger particles for a vessel with greater curvature. Overall the ranges of breakup distance for the R_1 and R_3 geometries respectively were 3.55×10^{-4} and 1.13×10^{-2}, so the benefit may be marginal. The positive correlation between particle size and average particle residence time found in Doig et al.'s study may warrant a larger aggregate particle diameter in order to gain the binding benefit of larger nanoparticles [9]. Table 7 shows the average distance from stenosis center for the R_1 and R_3 diameter cases that were simulated.

3.4 Stenosis Shape and Location

The R_1 geometry was tested with a 50% occluded eccentric stenosis inserted into it in order to compare it to the R_1 concentric case. The minimum angular velocity for this case was 11.66 rad/s. In order to attain a similar threshold value to the R_1 concentric case, $3.82\omega_{min}$ was used to attain a threshold of 44.66 rad/s. Sixteen aggregate particles were introduced into the flow. In this case, all particles broke up within the stenosis region. The eccentric R_1 case had an average minimum distance from stenosis center of 0.29 cm, which is identical to the concentric case to two significant digits.

The R_3 geometry was tested with 50% occluded eccentric stenoses in three positions: superior, inferior, and ventral/dorsal. The superior case had a minimum angular velocity of 1.23 rad/s. To meet the concentric case threshold, the minimum was multiplied by 159.98. Twenty particles were tested, and all particles broke up within the stenosis region. The average distance from the stenosis center is 0.19 cm, which was closer than for the concentric case with the same parameters.

The inferior case had a minimum angular velocity of 2.10 rad/s. To meet the R_3 concentric case threshold, the minimum was multiplied by 93.52. Twenty particles were tested. All particles broke up within the stenosis region. The average distance from stenosis center is 0.07 cm, which is closer than for the concentric case.

The ventral/dorsal case had a minimum angular velocity of 2.74 rad/s, and in order to match the concentric case threshold of 196.61, the minimum was multiplied

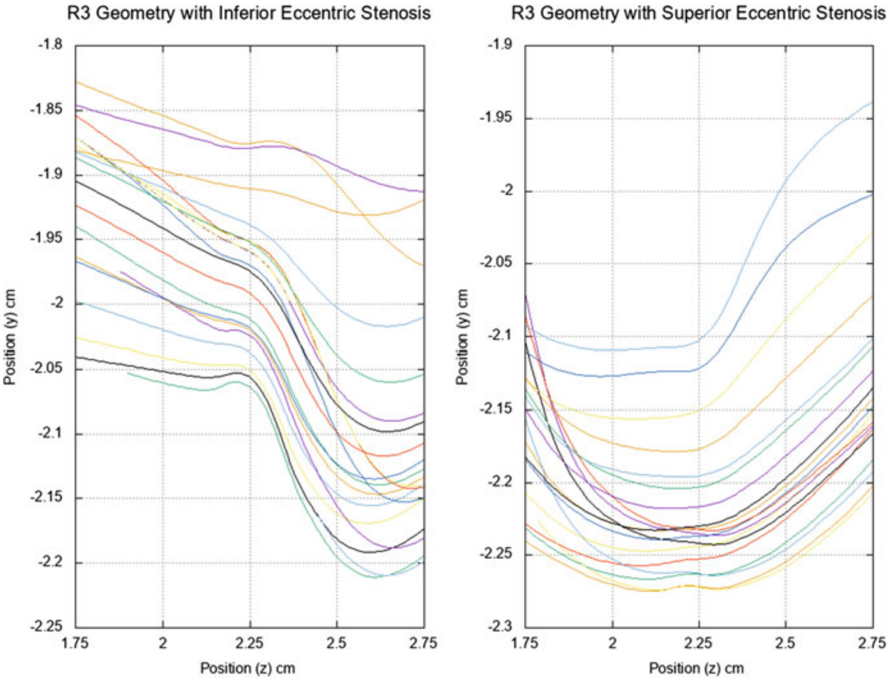


Fig. 10 Comparison of aggregate particle flow of the R_3 inferior eccentric case (left) and the R_3 superior eccentric case (right). Each line traces an individual aggregate particle flowing through the geometry

by 71.73 to attain the threshold. Fourteen aggregate particles passed through the flow. All particles broke up within the stenosis region, and this case had an average distance from stenosis center of 0.20 cm, which again is closer than for the concentric case.

In all R_3 eccentric cases, the particles broke up before the stenosis center, which agrees with the concentric case. All three eccentric cases fared better in both particle breakup within the stenosis region and average breakup distance from stenosis center. The inferior eccentric case performed the best due to its position relative to the particle trajectory, reminiscent of the correlation between initial particle position and breakup position shown in the R_1 concentric case. A comparison of the inferior and superior case is shown in Fig. 10 for a visual of stenosis position and its effect on particle trajectory. It should be noted that all eccentric cases had a better average particle breakup position, and all also had greater occlusion in the radial center than the concentric case. This greater center occlusion did not seem to affect the straight cases' average breakup distance. Curvature, which in part drives particle trajectory, combined with stenosis location, has a sizeable effect on optimal breakup distance.

4 Summary and Conclusion

This study used computational methods to better understand how aggregate particle breakup threshold, diameter and specific density, as well as vessel curvature and stenosis shape affect the efficiency of shear-activated nano-therapeutics in the treatment of atherosclerosis. Different idealized geometries were used to test and analyze these parameters. Optimal angular velocity breakup thresholds were discovered for both straight and curved geometry cases. Geometry curvature was a sizeable factor in breakup threshold, as the curved vessel cases (R_2, R_3) required 6.3 and 4.5 times the angular velocity of the straight vessel, respectively. No clear pattern was shown relating vessel curvature and optimal breakup threshold.

In the straight geometry cases, a correlation was found between initial particle position and particle breakup proximity to the stenosis center. As particles are positioned farther away from the vessel center, their breakup proximity from center is decreased. A similar correlation was not found in the curved vessel cases, but could hold true if a position nearer to the stenosis region was used. This finding in the straight geometry case corresponds with Qiao et al.'s study and further iterates their proposal for an injection site close to the aortic wall [14]. This correlation may hold true for the R_2 and R_3 if analysis was calculated using particle position when entering the stenosis region and not initial particle position.

Aggregate particle diameter was also explored. Diameters from 1.0 to 5.0 μm were used for both straight and curved geometries. For the straight vessel case, small decreases in average particle breakup distance from stenosis center were seen as particle diameters decreased. The opposite happened with the curved vessel case. As the particle diameter increased, a slight decrease in average particle breakup distance from stenosis was achieved. This indicates that as vessel curvature increases, so should the diameter of the aggregate particle for optimal results. This result could be useful in tailoring the aggregate particle to disease site.

Different specific densities were tested for both straight and curved geometries. It was determined that specific density will not play a role in efficiency of shear-activated nano-therapeutics.

Stenosis shape and location were tested in both the straight vessel and one of the curved vessel geometries. Curvature in conjunction with stenosis location had a great effect on average breakup distance from the stenosis center. This is a similar observation to the initial particle position correlation found in the R_1 cases. Both indicate that it is optimal to have the stenosis in the aggregate particle path.

Further studies need to be conducted to find optimal parameters. Effects of greater occlusions and more complicated geometries should give us more insight and a better understanding of the effectiveness of shear activated nano-therapeutics. There is evidence that blood viscosity is a risk factor for atherosclerosis; therefore, studies should be conducted using greater blood viscosity than used in this study to determine its effect on particle breakup [22].

References

1. T.E. Nogles, M.A. Galuska, *Middle Cerebral Artery Stroke*, in *StatPearls, 2021* [Internet] (StatPearls Publishing, Treasure Island, 2022). PMID: 32310592
2. S.S. Virani, A. Alonso, E.J. Benjamin, M.S. Bittencourt, C.W. Callaway et al.. American heart association council on epidemiology and prevention statistics committee and stroke statistics subcommittee. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. Circulation **141**(9), e139–e596 (2020). https://doi.org/10.1161/CIR.0000000000000757. PMID: 31992061
3. N.P. Denswil, A.C. van der Wal, K. Ritz, O.J. de Boer, E. Aronica, D. Troost, M.J.A.P. Daemen, Atherosclerosis in the circle of Willis: spatial differences in composition and in distribution of plaques. Atherosclerosis **251**, 78–84 (2016). https://doi.org/10.1016/j.atherosclerosis.2016.05.047. PMID: 27288902
4. Circle of Willis, *Medlineplus Medical Encyclopedia*. MedlinePlus, U.S. National Library of Medicine. Accessed 29 March 2021
5. A.E. Roher, C. Esh, T.A. Kokjohn, W. Kalback, D.C. Luehrs, J.D. Seward, L.I. Sue, T. G. Beach, Circle of Willis atherosclerosis is a risk factor for sporadic Alzheimer's disease. Arteriosclerosis Thrombosis Vasc. Biol. **23**(11), 2055–2062 (2003). PMID: 14617615
6. B. Flusty, A. de Havenon, S. Prabhakaran, D.S. Liebeskind, S. Yaghi, Intracranial atherosclerosis treatment: past, present, and future. Stroke **51**(3), e49–e53 (2020). https://doi.org/10.1161/strokeaha.119.028528. PMID: 32078441
7. R. Olie, P. van der Meijden, H.M.H. Spronk, H.T. Cate, Antithrombotic therapy: prevention and treatment of atherosclerosis and atherothrombosis, in *Handbook of Experimental Pharmacology* (2020)
8. B.B. Adhyaru, T.A. Jacobson, Safety and efficacy of statin therapy. Nat. Rev. Cardiol. **15**(12), 757–769 (2018). https://doi.org/10.1038/s41569-018-0098-5. PMID: 30375494
9. G. Doig, G.H. Yeoh, V. Timchenko, G. Rosengarten, T.J. Barber, S.C. Cheung, Simulation of blood flow and nanoparticle transport in a stenosed carotid bifurcation and pseudo-arteriole. J. Comput. Multiphase Flows **4**(1), 85–101 (2012)
10. N. Korin, M. Kanapathipillai, B.D. Matthews, M. Crescente, A. Brill, T. Mammoto, K. Ghosh, S. Jurek, S.A. Bencherif, D. Bhatta, A.U. Coskun, C.L. Feldman, D.D. Wagner, D.E.

Ingber, Shear-activated nanotherapeutics for drug targeting to obstructed blood vessels. Science **337**(6095), 738–42 (2012). https://doi.org/10.1126/science.1217815. Erratum in: Science **337**(6101), 1453 (2012). PMID: 22767894

11. M.G. Marosfoi, N. Korin, M.J. Gounis, O. Uzun, S. Veantham, E.T. Langan, A.l. Papa, O.W. Brooks, C. Johnson, A.S. Puri, D. Bhatta, M. Kanapathip-illai, B.R. Bronstein, J.Y. Chueh, D.E. Ingber, A.K. Wakhloo, Shear-activated nanoparticle aggregates combined with temporary endovascular bypass to treat large vessel occlusion. Stroke **46**(12), 3507–3513 (2015). PMID: 26493676
12. The crucial, controversial carotid artery part I: The artery in health and disease. Harvard Health (2011). Accessed 29 March 2021
13. H.J. Barnett, D.W. Taylor, M. Eliasziw, A.J. Fox, G.G. Ferguson, R.B. Haynes, R.N. Rankin, G.P. Clagett, V.C. Hachinski, D.L. Sackett, K.E. Thorpe, H.E. Meldrum, J.D. Spence, Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators. N. Engl. J. Med. **339**(20), 1415–25 (1998). https://doi.org/10.1056/NEJM199811123392002. PMID: 9811916
14. Y. Qiao, Y. Wang, Y. Chen, K. Luo, J. Fan, Mathematical modeling of shear-activated targeted nanoparticle drug delivery for the treatment of aortic diseases. Biomech. Model. Mechanobiol. **21**(1), 221–230 (2022). https://doi.org/10.1007/s10237-021-01530-9. PMID: 34748063
15. Freecad. https://www.freecadweb.org/
16. R. Lohner, Automatic unstructured grid generators. Finite Ele. Anal. Design **25**, 111–134 (1997)
17. R. Lohner, Renumbering strategies for unstructured grid solvers operating onshared memory, cached based parallel machines. Comput. Methods Appl. Mech. Eng. **163**, 95–109 (1998)
18. J.R. Cebral, R. Lohner, S. Appananboyina, C.M., Putman, Image-based computational hemodynamics methods and their application for the analysis of blood flow past endovascular devices, in *Biomechanical Systems Technology: Computational Methods* (WorldScientific, Singapore, 2007), pp. 29–85
19. F. Mut, Fast numerical solutions of patient-specific blood flows in 3D arterialsystems. Int. J. Num. Methods Biomed. Eng. **26**(1), 73–85 (2010)
20. B.J. Chung, D. Platt, A. Vaidya, The mechanics of clearance in a non-Newtonian lubrication layer. Int. J. Non-Linear Mech. **86**, 133–145 (2016)
21. S. Meschi, A. Farghadan, A. Arzani, Flow topology and targeted drug delivery in cardiovascular disease. J. Biomech. **119**, 110307 (2021). https://doi.org/10.1016/j.jbiomech.2021.110307
22. R.C. Becker, The role of blood viscosity in the development and progression of coronary artery disease. Cleveland Clin. J. Med. **60**(5), 53–358 (1993)

Compressed CO_2 Refrigeration for Energy Storage and CO_2 Utilization

Tran X. Phuoc and Mehrdad Massoudi

1 Introduction

Carbon dioxide (CO_2) has been considered as the most environmentally friendly refrigerant used in industrial and marine refrigeration [1–5]. This is because CO_2 is an inert gas, where the ozone depletion associated with conventional refrigerants, such as perfluorocarbons (PFCs) and hydrofluorocarbons (HFCs), does not exist. Carbon dioxide, which is available as a byproduct from many processes and plants (e.g., power plants, ammonia and beer production units etc.), is also more economical in comparison with conventional synthetic refrigerants. Like other refrigerants, the conventional use of CO_2 for refrigeration or air conditioning applications is achieved through vapor-to-liquid compression and liquid-to-vapor expansion processes. To achieve sufficient refrigeration, subcritical cycle and transcritical cycle have been commonly used, and high operation pressures (7–12 MPa) are required. Refrigeration systems with either subcritical or transcritical operations are more complex, leading to higher costs in components and installation. High operating pressure is more hazardous and increases the potential for leaks, and specially designed components are required.

When the critical temperature of CO_2 is around 31 °C, the heat released by CO_2 condensation cannot be discharged into the surrounding atmosphere above this temperature; in this work we report a simple analysis on a CO_2 cooling system that can be achieved simply based on its natural Joule-Thomson cooling capability. The **Joule-Thomson effect** is the change in a fluid temperature when it expands without

T. X. Phuoc · M. Massoudi (✉)
U.S. Department of Energy, National Energy Technology Laboratory (NETL), Pittsburgh, PA, USA
e-mail: Phuoc.Tran@netl.doe.gov; mehrdad.massoudi@netl.doe.gov

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_15

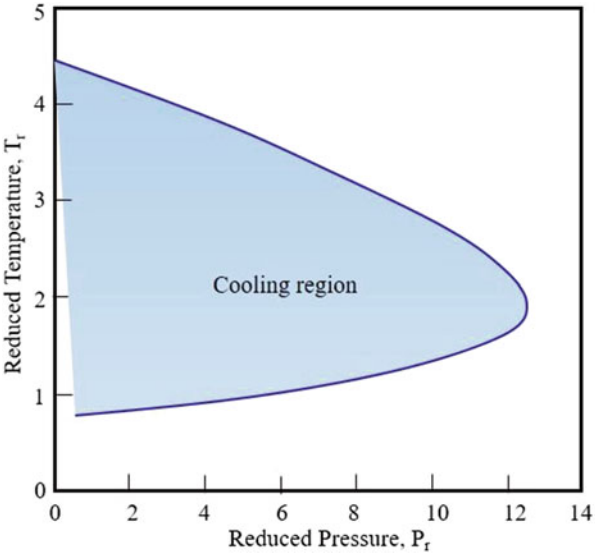


Fig. 1 Joule-Thomson inversion curves for CO_2

work and heat transfer. The Joule-Thomson inversion curve for CO_2 [6, 7] (the locus of states where the fluid temperature is invariant upon isenthalpic expansion) as shown in Fig. 1 indicates that in the reduced pressure up to about 12, and the reduced temperature between 1 and 4.5, the Joule-Thomson coefficient of CO_2 is positive, that is, if it is expanded isenthalpically, it will be cooled down. For CO_2 injection into a depleted natural gas reservoir, early studies [8–14] have reported that depending on the reservoir and the injection conditions, such effects could reduce reservoir temperature so significantly that thermal fracturing, formation of hydrates, and freezing of residual water could be induced.

The new cooling concept is presented in Fig. 2. It simply consists of four main components: a high-pressure piston-cylinder storage tank, a solar-powered compressor, a multipath heat exchanger, and a low-pressure storage tank. Excess solar (or wind) energy is used to compress CO_2 into the piston-cylinder tank that is set at a constant pressure by a moving piston (of about 1 MPa to 5 MPa). The compression process is carried out slowly so that the temperature of CO_2 in the compressed tank can be in equilibrium with the surrounding air temperature. When it is needed, the compressed CO_2 is allowed to expand at constant enthalpy via a plug valve into the low-pressure heat exchanger ($\leq$0.1 MPa). Such an expansion reduces CO_2 temperature significantly when it enters the heat exchanger where it is heated by absorbing heat from the hot air flowing through the heat exchanger. When the CO_2 exits the heat exchanger, it is pumped and stored into a low-pressure tank ($\leq$0.1 MPa). When excess solar or wind energy is available, it is recompressed and stored back into the high-pressure tank. Cooling this way is based strictly only on the Joule-Thomson cooling effect of CO_2. We calculated the CO_2

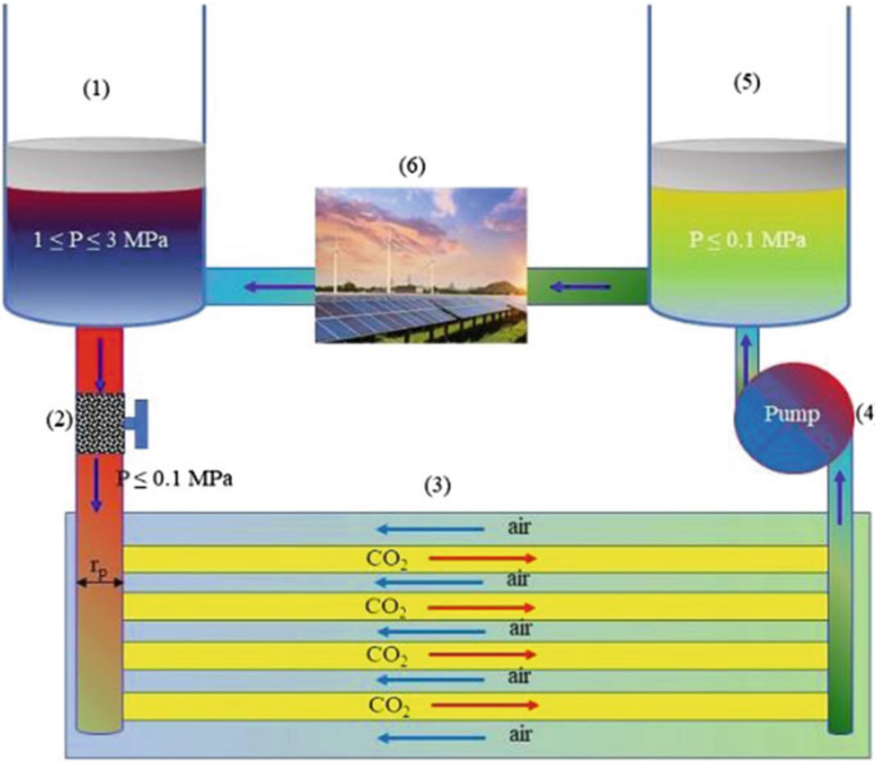


Fig. 2 Compressed CO_2 for energy storage and cooling utilization. (1) High-pressure tank, (2) plug valve, (3) multipath heat exchanger, (4) pump, (5) low-pressure tank, (6) solar-powered compressor

Joule-Thomson coefficient using the NIST database [15] and also from the Wagner-proposed equation of state for CO_2 [16]. The values of the CO_2 Joule-Thomson coefficient at 35 °C increase from 10.082 K/MPa at 1 MPa to 10.11 K/MPa at 3 MPa and then start to decrease as the pressure increases further. For 40 °C, for the same range of pressure, it is constant at 9.67 K/MPa. The temperatures of 35 °C and 40 °C used here are common environment temperatures. Thus, considering the environment temperature of about 40 °C, the most effective range of pressures used is about 1–3 MPa. The proposed refrigeration cycle requires only a few no-moving part components, and the working fluid remains in its single vapor phase; thus, it is simple and cost-effective in design, components, and installation. The proposed cooling system will serve as cooling and an energy storage system for excess solar (or wind) energy at the same time. In the following sections, the performance of a parallel-flow heat exchanger and counter-flow heat exchanger will be analyzed.

2 Heat Transfer Analysis

When compressed CO_2 is released at a constant enthalpy from the high-pressure tank through the plug valve as shown in Fig. 2, a cold-stream CO_2 at the heat exchanger pressure is generated and enters the heat exchanger at x = 0 and exits the heat exchanger at x = L, which is the length of the heat exchanger. The CO_2 temperature and velocity at the heat exchanger inlet are related to the tank pressure and temperature and the pressure of the heat exchanger; they can be calculated as follows:

$$\mu_{JT} = \left(\frac{\partial T}{\partial P}\right)_h \approx \frac{\left(T_{CO_2,tnk} - T_{CO_2,in}\right)}{\left(P_{tank} - P_{in}\right)} \tag{1}$$

$$T_{CO_2,in} = T_{CO_2,tnk} - \mu_{JT}\left(P_{tank} - P_{in}\right) \tag{2}$$

$$P_{tnk} = P_{in} + \rho_{CO_2,in} V_{in}^2 \tag{3}$$

Thus, the mass flow rate of the cold CO_2 stream before entering the multiple flow paths of the heat exchanger is

$$\dot{m}_{CO_2} = \pi r_p^2 \left[\rho_{CO_2,in}\left(P_{tank} - P_{in}\right)\right]^{\frac{1}{2}} \tag{4}$$

where $T_{CO_2,in}$ is the CO_2 inlet temperature,$T_{CO_2,tnk}$ is the temperature of CO_2 stored in high-pressure tank, μ_{JT} is the Joule-Thomson coefficient of CO_2 (at $T_{CO_2,tnk}$, P_{tnk}), P_{in} (about 0.1 MPa) is the heat-exchanger pressure, $\rho_{CO_2,in}$ is the density of CO_2 at the heat-exchanger inlet, V_{in} is the CO_2 velocity, and r_p is the radius of the release pipe.

This analysis is given to a single flow path of a heat exchanger as represented in Fig. 3. Two types of flow configurations are used: a parallel flow and a counter flow. The flow path has a width b, which is also the width of the heat exchanger, and a thickness a. For the parallel-flow heat configuration, the hot air stream enters the flow path at x = 0 and exits at x = L. For the counter-flow configuration, it enters at x = L and exits at x = 0. The cold CO_2 stream inlet temperature, $T_{CO_2,in}$, is determined using Eq. (2). Since the heat exchanger has N identical flow paths, the mass flow rate of the CO_2 stream per path is $\dot{m}_{CO_2,p} = \dot{m}_{CO_2}/N$, where $\dot{m}_{CO_2}$is determined using Eq. (4). Keeping the dimensions of the releasing pipe ($r_p = 5$ mm) and of the flow path (a = 2 cm, b = 10 cm, L = 1 m) unchanged, the goal is to calculate the air and CO_2 outlet temperatures using the number of the flow path N, the compressed tank pressure, at P_{tnk}, and temperature, at $T_{CO_2,tnk}$, and the air mass flow rate, $\dot{m}_{air,p}$, as parameters.

The heat capacity of the air stream and CO_2 stream are

$$\dot{W}_{air} = \dot{m}_{air,p} c_{p,air} \tag{5}$$

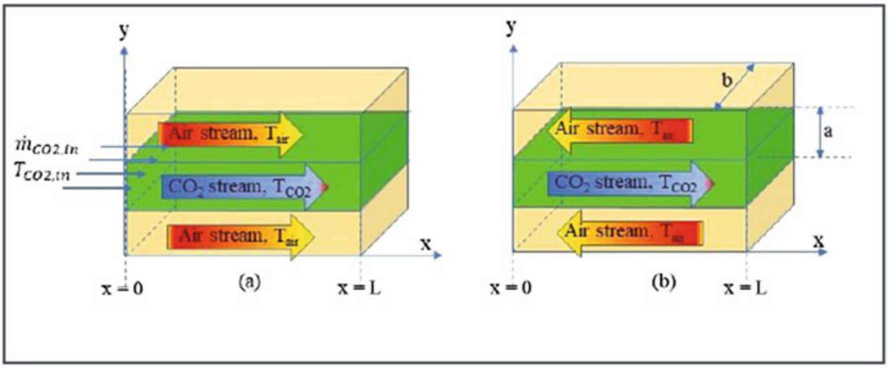


Fig. 3 Heat exchanger: (**a**) parallel flow, (**b**) counter flow

$$\dot{\mathrm{W}}_{CO_2} = \dot{m}_{CO_2,p} c_{p,CO_2} \tag{6}$$

The total heat transfer between the two streams is

$$\dot{Q} = \dot{W}_{air} \left(T_{air,in} - T_{air,out} \right) = \dot{W}_{CO_2} \left(T_{CO_2,out} - T_{CO_2,in} \right) \tag{7}$$

Or

$$\dot{Q} = \frac{\left(T_{air,in} - T_{CO_2,in} \right) - \left(T_{air,out} - T_{CO_2,out} \right)}{\left(\frac{1}{\dot{W}_{air}} + \frac{1}{\dot{W}_{CO_2}} \right)} \tag{8}$$

where $\dot{m}_{CO_2,p}$ is the mass flow rate of the CO2 stream, $\dot{m}_{air,p}$ is the mass flow rate of the air stream, c_{p,CO_2} is the specific heat at constant pressure of CO2, $c_{p,\,air}$ is the specific heat at constant pressure of air, and $\dot{q}''$ is the heat transfer rate per unit area between the two fluid stream at any location along the heat exchanger expressed as follows:

$$\dot{q''} = h_o \left(T_{air} - T_{CO_2} \right) \tag{9}$$

If the inside tube wall is thin and its thermal resistance is neglected, the heat transfer coefficient h_o calculated from the heat transfer coefficient of the air stream, $h_{c,\,air}$, and the heat transfer coefficient of the CO2 stream, h_{c,CO_2}, as

$$\frac{1}{h_o} = \frac{1}{h_{c,air}} + \frac{1}{h_{c,CO_2}} \tag{10}$$

2.1 Parallel-Flow Heat Exchanger

Referring to Fig. 3a, the mean temperatures of the air stream, T_{air}, and CO2 stream, T_{CO_2}, in the heat exchanger are

$$\frac{dT_{CO_2}}{dx} = \frac{4\left(a+b\right)\dot{q}''}{\dot{W}_{CO_2}} = \frac{4\left(a+b\right)h_o}{\dot{W}_{CO_2}}\left(T_{air} - T_{CO_2}\right) \tag{11}$$

$$\frac{dT_{air}}{dx} = -\frac{4\left(a+b\right)\dot{q}''}{\dot{W}_{air}} = -\frac{4\left(a+b\right)h_o}{\dot{W}_{air}}\left(T_{air} - T_{CO_2}\right) \tag{12}$$

Subtract Eq. (11) from Eq. (12)

$$\frac{d\left(T_{air} - T_{CO_2}\right)}{\left(T_{air} - T_{CO_2}\right)} = -4\left(a+b\right)h_o\left(\frac{1}{\dot{W}_{air}} + \frac{1}{\dot{W}_{CO_2}}\right)dx \tag{13}$$

Integrating Eq. (13) from x = 0, where $T_{air,} = T_{air,in}$ *and* $T_{CO_2} = T_{CO_2,in}$ to x = L where $T_{air} = T_{air,out}$ *and* $T_{CO_2} = T_{CO_2,out}$, we obtain the outlet temperatures of both streams and the total heat transfer rate, $\dot{Q}$, as

$$T_{air,out} - T_{CO_2,out} = \left(T_{air,in} - T_{CO_2,in}\right)e^{-\beta L} \tag{14}$$

where

$$\beta = 4\left(a+b\right)h_o\left(\frac{1}{\dot{W}_{air}} + \frac{1}{\dot{W}_{CO_2}}\right) \tag{15}$$

From Eq. (14)

$$\left(\frac{1}{\dot{W}_{air}} + \frac{1}{\dot{W}_{CO_2}}\right) = -\frac{1}{4\left(a+b\right)h_oL}Ln\left(\frac{T_{air,out} - T_{CO_2,out}}{T_{air,in} - T_{CO_2,in}}\right) \tag{16}$$

The total heat transfer, given by Eq. (8), becomes

$$\dot{Q} = -4\left(a+b\right)h_oL\left[\frac{\left(T_{air,in} - T_{CO_2,in}\right) - \left(T_{air,out} - T_{CO_2,out}\right)}{Ln\left(\frac{T_{air,out} - T_{CO_2,out}}{T_{air,in} - T_{CO_2,in}}\right)}\right] \tag{17}$$

From Eq. (7), the CO2 outlet temperature is expressed as follows:

$$T_{CO_2,out} = T_{CO_2,in} + \frac{W_{air}}{W_{CO_2}}\left(T_{air,in} - T_{air,out}\right) \tag{18}$$

From Eqs. (14) and (18), the air outlet temperature is

$$T_{air,out} = T_{air,in} - \frac{\left(T_{air,in} - T_{CO_2CO_2,in}\right)\left(1 - e^{-\beta L}\right)}{\left(1 + \frac{\dot{W}_{air}}{\dot{W}_{CO_2}}\right)} \tag{19}$$

2.2 Counter-Flow Heat Exchanger

Referring to Fig. 3b, the temperatures of the CO2 stream in the heat exchanger is described by Eq. (11); temperatures of the air stream, T_{air}, in the heat exchanger is expressed in Eq. (20):

$$\frac{dT_{air}}{dx} = \frac{4\,(a+b)\,h_o}{\dot{W}_{air}}\left(T_{air} - T_{CO_2}\right) \tag{20}$$

Combining Eqs. (11) and (20)

$$\frac{d\left(T_{air} - T_{CO_2}\right)}{\left(T_{air} - T_{CO_2}\right)} = \alpha dx \tag{21}$$

where

$$\alpha = 4\,(a+b)\,h_o\left(\frac{1}{\dot{W}_{air}} - \frac{1}{\dot{W}_{CO_2}}\right) \tag{22}$$

And integrating Eq. (21) from x = 0, where $T_{air,} = T_{air,out}, and\ T_{CO_2} = T_{CO_2,in}$ to x = L, where $T_{air} = T_{air,in}\ and\ T_{CO_2} = T_{CO_2,out}$ we obtain the outlet temperatures of both streams and the total heat transfer rate, $\dot{Q}$, as follows:

$$Ln\frac{\left(T_{air,out} - T_{CO_2,in}\right)}{\left(T_{air,in} - T_{CO_2,out}\right)} = -\alpha L \tag{23}$$

And

$$\left(\frac{1}{\dot{W}_{air}} - \frac{1}{\dot{W}_{CO_2}}\right) = -\frac{1}{4\,(a+b)\,h_o L} Ln\frac{\left(T_{air,out} - T_{CO_2,in}\right)}{\left(T_{air,in} - T_{CO_2,out}\right)} \tag{24}$$

From Eq. (7), the total heat transfer and CO2 outlet temperature are

$$\dot{Q} = -4\,(a+b)\,h_o L\left[\frac{\left(T_{air,in} - T_{air,out}\right) - \left(T_{CO_2,out} - T_{CO_2,in}\right)}{Ln\frac{\left(T_{air,out} - T_{CO_2,in}\right)}{\left(T_{air,in} - T_{CO_2,out}\right)}}\right] \tag{25}$$

$$T_{CO_2,out} = T_{CO_2,in} + \frac{\dot{W}_{air}}{\dot{W}_{CO_2}}\left(T_{air,in} - T_{air,out}\right) \tag{26}$$

From Eq. (23), the air outlet temperature is

$$T_{air,out} = T_{air,in} + \left(T_{CO_2,in} - T_{air,in}\right)\frac{\left(1 - e^{-\alpha L}\right)}{\left(1 - \frac{\dot{W}_{air}}{\dot{W}_{CO_2}}e^{-\alpha L}\right)} \tag{27}$$

3 Results and Discussions

The following results are calculated for a heat exchanger that has 10 or 20 flow paths with a = 2 cm, b = 10 cm, and a length L = 1 m. The diameter of the CO_2 release pipe is set to 1 cm.

Figure 4 shows the exit temperatures of air and CO_2 streams as they pass through the heat exchanger having 20 flow paths and a length of 1 m. The temperature drop of the air stream shown here does not seem to depend on the compressed tank temperature but depends significantly on the air mass flow rate. The counter-flow configuration also results in a larger temperature drop than the parallel-flow does. For the parallel-flow configuration, the exit temperature of the air stream is about 8–15 °C lower than its initial temperature for the range of the mass flow rate ratio from 0.25 to 1. For the same range of the mass flow rate ratio, the counter-flow results in the air temperature drop from about 9–19 °C.

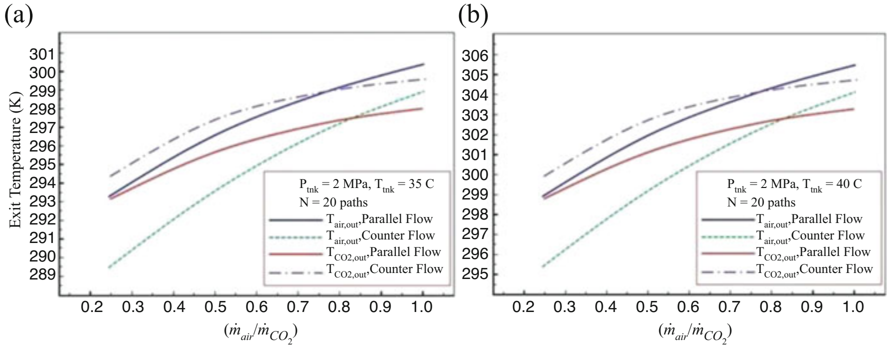


Fig. 4 Air and CO_2 outlet temperatures as a function of the air and CO_2 mass flow rates ratio (P_{tnk} = 2 MPa, 20 flow paths; (**a**) $T_{air,in} = T_{tnk}$ = 35 °C (308 K), $T_{CO_2,in}$ = 288 K, $\dot{m}_{CO_2} = 7.4\ g/s$; (**b**) $T_{air,in} = T_{tnk}$ = 40 °C (313 K), $T_{CO_2,in}$ = 294 K, $\dot{m}_{CO_2} = 7.3\ g/s$)

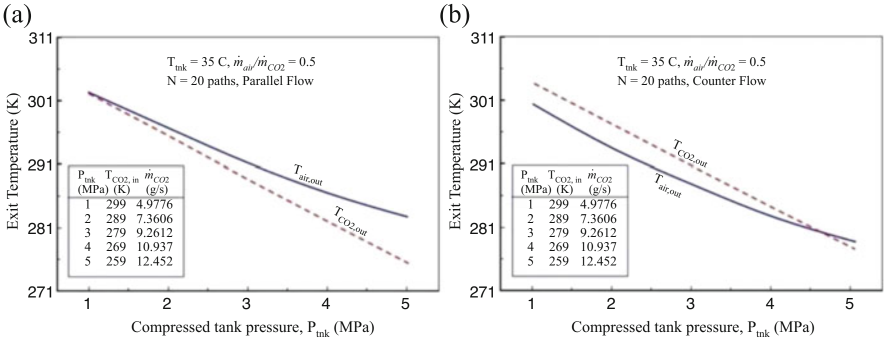


Fig. 5 Exit temperature as a function of the compressed tank pressure (ratio of air mass flow rate to the CO_2 mass flow rate: 50%, N = 20; (**a**) parallel flow, (**b**) counter flow)

Figure 5 shows the exit temperatures of air and CO_2 streams as a function of the compressed tank pressure when the initial air temperature of 35 °C (308 K) and the mass flow rate ratio of 0.5 are kept unchanged. The temperatures and the mass flow rates of CO_2 as a function of the compressed tank pressure before entering the heat exchanger are also included in the figure for easy reference. The results indicate that the tank pressure has a significant effect on the exit temperatures of both streams. For the conditions used here, a decrease of about 25–30 °C in the air stream temperature is achieved when the compressed tank pressure is at 5 MPa. For example, with $P_{tnk} = 5$ MPa and $T_{tnk} = 35$ °C (308 K), the air exit temperature is 10 °C (283 K) with the parallel-flow configuration and 6 °C (279 K) with the counter-flow configuration. Such a significant cooling is due to the fact that as seen from Eqs. (2) and (4), where the CO_2 temperature and mass flow rate are calculated, colder and higher mass flow rate of CO_2 can be released entering the heat exchanger as the tank pressure is higher.

Figure 6 shows the total heat transfer for a heat exchanger having 20 flow paths. The results shown here indicate that the total heat transfer increases with the mass flow rate ratio and with the increase in the compressed tank pressure. The heat transfer is also higher with the counter-flow configuration. As seen from Fig. 5, increasing the compressed tank pressure would result in a CO_2 stream that enters the heat exchanger with higher mass flow rate and lower temperature. Thus, more heat from the surrounding hot air is absorbed. For example, for the conditions used here, with the mass flow ratio of 50%, a total heat transfer rate in the range from 284 W to 3174 W (parallel flow) and from 386 W to 3592 W (counter flow) is exchanged between the two streams as the tank pressure increased from 1 to 5 MPa.

The effects of the number of the flow paths on air and CO_2 exit temperatures and the total heat transfer rate are presented in Figs. 7 and 8. With the conditions used here, a total of 3–4 kW of heat exchange can be delivered. Increasing the number of the flow paths, the surface heat transfer is increased, and the CO_2 mass flow rate per path is decreased; both of these effects have a significant impact on the total heat

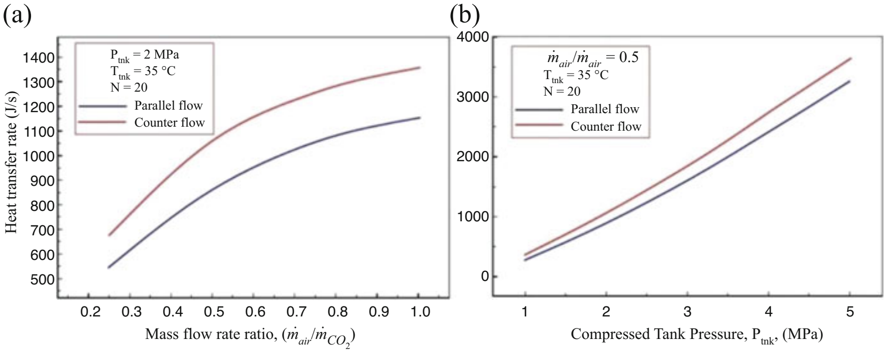
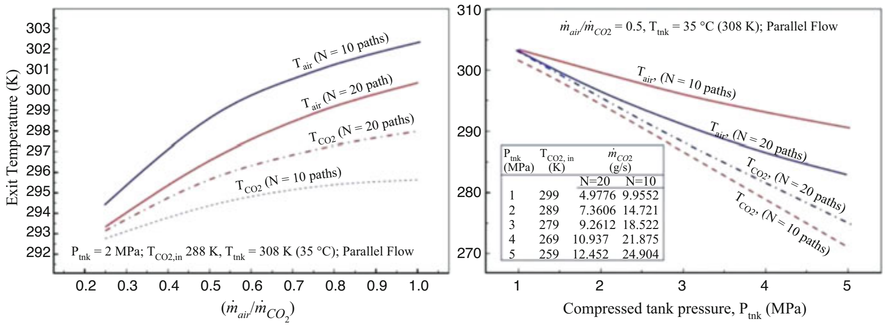


Fig. 6 Total heat transfer for a single path: (**a**) as a function of the flow rate ratio, (**b**) as a function of the compressed tank pressure



P_{tnk} (MPa)	$T_{CO2, in}$ (K)	$\dot{m}_{CO2}$ (g/s)	
		N=20	N=10
1	299	4.9776	9.9552
2	289	7.3606	14.721
3	279	9.2612	18.522
4	269	10.937	21.875
5	259	12.452	24.904

Fig. 7 Exit temperature as a function of the flow rate ratio and of the compressed tank pressure: effect of the number of the flow paths

transfer rates but less significant on the air and CO_2 exit temperatures. For example, for $P_{tnk} = 2$ MPa and 35 °C (308 K), a CO_2 stream of 147.2 g/s and 288 K is released before entering the heat exchanger. With 10 flow paths, a stream of 14.7 g/s of CO_2 enters each path, resulting in the exit air temperature in the range from 294 K to 302 K and a total heat transfer rate from 504 W to 851 W when the ratio of the mass flow rate increases from 0.25 to 1. For the same condition, a heat exchanger with 20 flow path can deliver a heat transfer rate from 545 W to 1144 W, resulting in the air exit temperature from 293 K to 300 K. However, it is noticed that with fewer flow paths, the difference in the air and CO_2 exit temperature becomes wider. This means that the cooling capacity of a heat exchanger with fewer flow paths can be improved if it is longer.

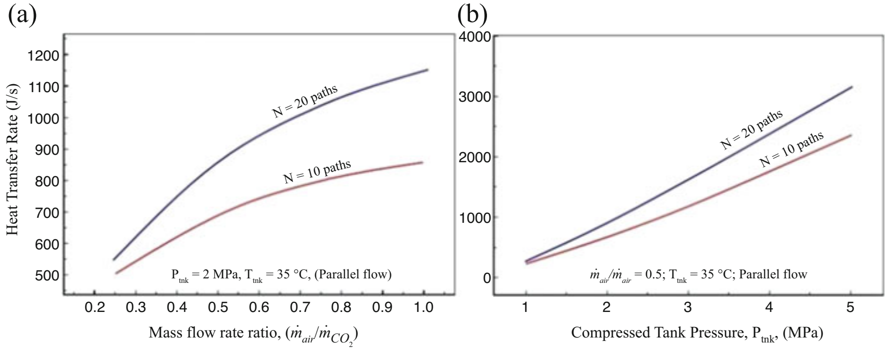


Fig. 8 Total heat transfer ratio as a function of the flow rate ratio and of the compressed tank pressure. Effect of the number of the flow paths

4 Conclusions

We have performed a simple analysis to explore the possibility of using compressed CO_2 for air-cooling applications. The goal is to develop a compressed CO_2 system for both excess solar/wind energy storage and CO_2 utilization. The cooling capacity of the gaseous CO_2 is achieved naturally using the Joule-Thomson cooling capability of the expanding CO_2 from a high-pressure compressed tank to a lower-pressure heat exchanger. Keeping the heat-exchanger dimension fixed, the analysis was aimed at the exit air temperature and the total heat transfer using the compressed tank pressure, air mass flow rate, and number of the flow paths within the heat exchanger as parameters. For a heat exchanger that has 10 or 20 flow paths with a = 2 cm, b = 10 cm, and a length L = 1 m and a pressure of 0.1 MPa, our results indicate that the Joule-Thomson cooling capability of the gaseous CO_2 that expands from a compressed tank at 5 MPa into such a heat exchanger could generate a 3–4 kW of cooling power, and a stream of 124 g/s of hot air flowing through it could have a temperature drop from 25 to 30 °C. The pressure of 5 MPa used here is very much lower than the pressure required in conventional vapor-compression refrigeration cycles (from 7–12 MPa).

Disclaimer This paper was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

1. P.K. Bansal, S. Jain, Cascade systems: past, present, and future. ASHRAE Trans. **113**, 245–252 (2007)
2. H.M. Getu, P.K. Bansal, Thermodynamic analysis of an R717- R744 cascade refrigeration system, Int. J. Refrig. 31 (2008) 45–54
3. S. Sawalha, Using CO2 in supermarket refrigeration. ASHRAE J. **47**, 26–30 (2005)
4. P. Bansal, A review e Status of CO2 as a low temperature refrigerant: Fundamentals and R&D opportunities. Appl. Therm. Eng. **41**, 18–29 (2012)
5. J.M. Belman-Flores, V. Pérez-García, J.F. Ituna-Yudonago, J.L. Rodríguez-Muñoz, J. de Jesús Ramírez-Minguela, General aspects of carbon dioxide as a refrigerant. J. Energy South. Afr. **25**, 96–106 (2014)
6. C.M. Colina, M. L'ısal, F.R. Siperstein, K.E. Gubbins, Accurate CO2 Joule–Thomson inversion curve by molecular simulations. Fluid Phase Equilib. **202**, 253–262 (2002)
7. B. Haghighi, M.R. Bozorgmehr, Joule-Thomson inversion curves calculation by using equation of state. Asian J. Chem. **24**(2), 533–537 (2012)
8. C.M. Oldenburg, Joule–Thomson cooling due to CO2 injection into natural gas reservoirs. Energy Convers. Manag. **48**, 1808–1815 (2007)
9. S.A. Mathias, J.G. Gluyas, C.M. Oldenburg, C.F. Tsang, Analytical solution for Joule–Thomson cooling during CO2 geo-sequestration in depleted oil and gas reservoirs. Int. J. Greenhouse Gas Control **4**, 806–810 (2010)
10. A.S. Ramazanov, V.M. Nagimov, Analytical model for the calculation of temperature distribution in the oil reservoir during unsteady fluid inflow. Oil Gas Bus. **2007**, 10–20 (2007)
11. Z. Ziabakhsh-Ganji, H. Kooi, Sensitivity of Joule–Thomson cooling to impure CO2 injection in depleted gas reservoirs. Appl. Energy **113**, 434–451 (2014)
12. A.K. Singh, U.J. Goerke, O. Kolditz, Numerical simulation of non-isothermal compositional gas flow: Application to carbon dioxide injection into gas reservoirs. Energy **36**, 3446–3458 (2011)
13. A.K. Singh, G. Baumann, J. Henninges, U.J. Goerke, O. Kolditz, Numerical analysis of thermal effects during carbon dioxide injection with enhanced gas recovery: A theoretical case study for the Altmark gas field. Environ. Earth Sci. **67**, 497–509 (2012)
14. R. Middleton, H. Viswanathan, R. Currier, R. Gupta, CO2 as a fracturing fluid: Potential for commercial-scale shale gas production and CO2 sequestration. Energy Procedia **63**, 7780–7784 (2014)
15. https://webbook.nist.gov/chemistry/
16. R. Span, W. Wagner, New equation of state for carbon dioxide covering the fluid region from the triple-point temperature to 1100 K at pressures up to 800 MPa. J. Phys. Chem. Ref. Data **25**, 1509–1596 (1996)

A Two-Phase Model for Mucosal Aggregation and Clearance in the Human Tear Film

Bong Jae Chung, Brandon Martinez, and Ashwin Vaidya

1 Introduction

Approximately 7% of population of the USA, especially women, suffer from aqueous tear deficiency or dry eye disease [1] while nearly 60% of glaucoma patients have symptoms of dry eyes [2]. Dry eye syndrome mainly occurs due to inadequate lacrimal layer production, or meibomian gland dysfunction at the rim of the eyelids, which can cause excessive evaporation of the tear film [3]. Ocular mucin are known to regulate the function of tear film, especially to protect ocular surface from the evaporation of tear film (maintaining water), associated with dry eye syndrome [4]. Additionally, they are also known to serve as mucosal barriers to wrap and remove debris from the tear film [5, 6]. Therefore mucin, glycosylated proteins, plays a key role in lubricating and protecting the ocular mucosa in general. The tear film contains three types of mucin: (i) the large gel-forming mucin MUC5AC expressed by conjunctival goblet cells, playing a role of removal of debris, (ii) the small soluble mucin MUC7 secreted by the lacrimal gland acini, and (iii) the membrane-associated mucin MUCs 1, 4, and 16 expressed by the corneal and conjunctival epithelia, preventing pathogen penetration [4, 7, 8]. A number of earlier studies show that mucin properties and its distribution are altered by ocular surface diseases such as dry eye [4, 7, 9, 10].

B. J. Chung · B. Martinez
Department of Applied Mathematics and Statistics, Montclair State University, Montclair, NJ, USA
e-mail: chungb@montclair.edu; martinezb11@montclair.edu

A. Vaidya (✉)
Department of Mathematics, Department of Physics and Astronomy, Complex Fluids Laboratory, Montclair State University, Montclair, NJ, USA
e-mail: vaidyaa@montclair.edu

F. Carapau, A. Vaidya (eds.), *Recent Advances in Mechanics and Fluid-Structure Interaction with Applications*, Advances in Mathematical Fluid Mechanics,
https://doi.org/10.1007/978-3-031-14324-3_16

Simply stated, the human tear film is essentially composed of a very thin lipid layer, an aqueous layer, and a mucus layer. However modern theories suggest a less distinct demarcation of its composition [7, 11–13]. The bulk of the tear film is composed of the aqueous layer [14], which is enriched saline and serves to moisten the eye and provide nutrients [15]. The lipid layer serves mostly to reduce evaporation of the aqueous layer and resides mostly between the two edges of the eyelid [16]. The aqueous layer behaves strictly as a Newtonian fluid [17] while the mucus layer, which lies between the aqueous layer and the surface of the eyeball displays non-Newtonian characteristics; its molecular composition gives it a shear dependent and possibly elastic character [18], although the properties of mucus that we know about come from sources other than the eye due to the difficulty in isolating the very small amounts present in the eye. Details about the physiological properties of the tear film can be found extensively in the literature [7, 12, 13, 19–22]. Following the recent theories, the tear film can be assumed (as is done in this paper) that the aqueous and mucus layers form a single system containing mucin. In the current paper, mucin is considered to be discrete entities embedded in the background fluid with a prescribed distribution profile depending upon possible clinical conditions.

The tear film is subjected to the blinking motion of human eyelid, with blinking classified as (i) voluntary, (ii) regular, and (iii) responsive; the opening period is around twice as large as the closing period [23, 24]. Consequently, the nature of the flow of the layers is oscillatory. The embedded mucin in the layers is also governed by this periodic motion of the background fluid compounded by their own inertial motion. In the current paper, a wrapping mechanism of foreign body capture by free floating mucin is modeled. The process of mucin wrapping in the removal of foreign debris has been well documented in the gastrointestinal tract of the human body [25]; ([26, section4.2] polystyrene particle interaction with mucin in humans); mucin aggregation has also been observed in other mammalian species besides humans. Figure 1 shows a clear image of a mucus coated particle taken from the intestine of a rat [6].

Even in other mammalian species apart from humans, it has been found that India Ink particles injected in the intestines of cats are surrounded by mucus, indicating that any embedded foreign object must pass a mucus barrier which prevents it from reaching the epithelial cells [5]. Several studies on drug delivery in human subjects [5, 25, 26] using nanoparticles (NPs) have revealed the inner workings of the mucin wrapping mechanism which could be a hindrance in targeted drug delivery in organs containing mucus. These studies indirectly provide evidence for the possible purpose and mechanics of mucin—foreign particle interaction in the eye. In an article pertaining to the impact of mucus on drug delivery mechanisms, Wongsakorn [27] points to a possible explanation for this interaction which: "…might be the electrostatic interaction of negatively charged mucus that wraps NPs, thereby changing their physico-chemical properties." The existence of a wrapping mechanism for foreign body capture within the human tear film of the eye is also supported by the findings in [28], which describes changes in the tear film and ocular surface stemming from dry eye syndrome. The idea that mucus binds

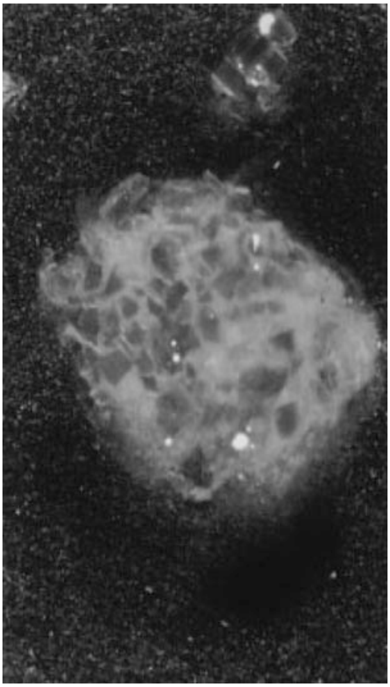


Fig. 1 The image shows mucus coated particles taken from the small intestine of a rat, immediately following discharge. Reprinted from Journal of Pharmaceutical Sciences, Vol. 87, Issue 4, Boaz Tirosh,Abraham Rubinstein, 'Migration of Adhesive and Nonadhesive Particles in the Rat Intestine under Altered Mucus Secretion Conditions', Pages 453–456, Copyright (1998), with permission from Elsevier

to foreign particles for their elimination is consistent with the notion of a wrapping mechanism and expressed thus [28]: "Bladder mucosa exhibits nonspecific anti-adherence to bacteria, attributed to electrochemical repulsion by its negatively charged residues [29] although at the ocular surface, mucus more commonly binds with potentially harmful tear contaminants and acts as a debris removal system." It is evident that the mucin wrapping mechanism exists in the gastrointestinal tract of mammalian species besides humans; it is even noted in one of the studies considered above that mucus has the function of adhering to and clearing foreign debris entering the ocular surface of the human eye [28]. We therefore hypothesize that the mechanism of mucin wrapping exists in the tear film of the human eye. Modeling attempts begin with a definition of adhesion between mucin and a foreign object.

Adhesion between mucin and the object occurs when they are sufficiently close, i.e., within a predefined activation range. The inter- and intra-binding forces between mucin and mucin-bacteria or mucin-protein systems have been extensively studied [30–34]. While the impact of fluid flow and properties of the tear film on pathologies such as glaucoma [35, 36] and of the role of lipid layers related to dry eye disease [37] have been investigated, several fundamental questions about the underlying mechanics still need to be addressed, to the best of the authors' knowledge. The goals of the current paper are to specifically model the wrapping mechanics and clearance rates of "ocular debris" under various conditions, such as

(i) mucin distribution, (ii) mucin population (which contributes to the bulk viscosity of the tear film), and (iii) adhesion force between mucin and debris. Such a study could lead to better understanding of the underlying physics of clearance with potential remedies for ocular diseases such as glaucoma and dry eye disease.

The outline of the rest of the paper is as follows: the next section develops the various parts of the theoretical model, namely, the fluid flow induced by blinking, the mucin distribution profile, and biochemical forces at play. This is followed by a section outlining the computational strategies involved in the study. Finally, we discuss the results of our calculations followed by a discussion of the possible biological implications of this study.

2 Theoretical Model

The model in our study (see Fig. 2) consists of three parts which include the (a) background, aqueous fluid component of the tear film, (b) the mucin proteins, and (c) the foreign body that penetrates the tear film. The governing equations are chosen so that each component of the model is physically reasonable but also to optimize computational time.

2.1 The Fluid Model

Whereas the literature on tear film recognizes the presence of aqueous and mucus layers in the eye, it is difficult to demarcate such zones clearly in the tear film since mucins are found to be present throughout the tear film. There is controversy even in the estimation of the exact depth of the net ocular tear film [38] with depths ranging between 3 µm and 40 µm. However, there have been attempts to identify regions dominated by each. We rely on recent reports [39] which suggest that the thickness of the mucus layer ranges between 0.02 and 0.05 µm while the aqueous layer is estimated to be between 6 and 9 µm in thickness. Therefore, according to these estimates, the mucus layer occupies less than 1% of the entire film.

Therefore, in the model employed here we treat the aqueous and mucus layers as a single, combined medium. Since the inhomogeneity of the tear film system comes from the presence of mucin "particles" in the background fluid, we take the model tear film as a fluid, whose material characteristics (viscosity) are spatially varying due to changing mucin distribution. In many practical cases where complex fluids play an important role, the shear viscosity is often modeled as a function of one or more of the following: time (t), shear rate ($\dot{\gamma}$), concentration(ϕ), temperature(θ), pressure(p), electric field ($\mathbf{E}$), and magnetic field ($\mathbf{M}$). So in general, we can write

$$\mu = \mu\left(t, \dot{\gamma}, \phi, \theta, p, |\mathbf{E}|, |\mathbf{M}|\right). \tag{1}$$

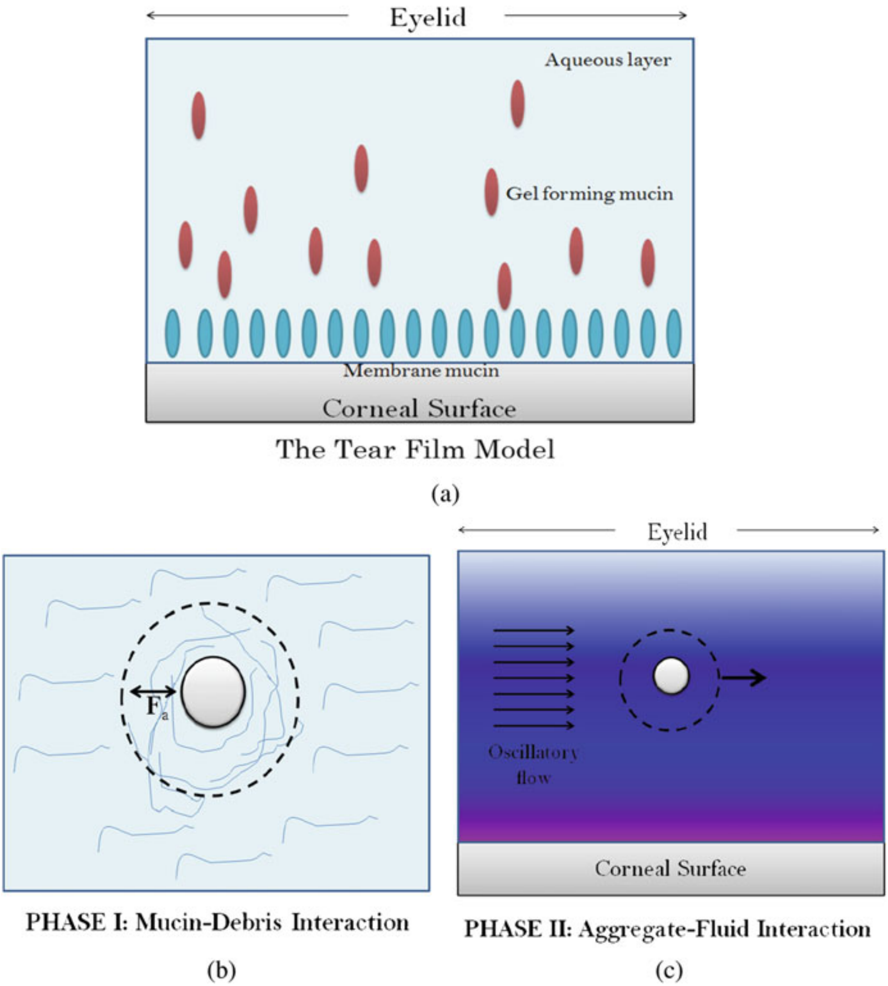


Fig. 2 This figure outlines our overall modeling strategy. The problem of clearance of a foreign body in the tear film is broken down to two interaction phases, at various scales. Panel (**a**) shows the model assumption regarding the tear film makeup. Panel (**b**) shows a schematic of Phase I which involves the mucoadhesion process between the foreign body and mucin at the small scale, and finally, panel (**c**) shows Phase II which includes the large scale transport of the foreign body due to the coupling with the background flow

In this particular study, the dependence on concentration (i.e., ϕ) alone is relevant. Our hypothesis is that the tear film layer acquires its complex characteristics due to this inhomogeneous mixture resulting in an effective viscosity of the system which changes with depth. Our assumption stems from the fact that viscosity (μ) is well known to depend upon concentration (ϕ) [40]. Therefore, for a concentration changing with depth, i.e., $\phi = \phi(z)$, it implies $\mu = \mu(z)$. The particular advantage of using this model lies in the fact that it allows for an analytical solution for the flow velocity under oscillatory (blinking) conditions, which captures the essential characteristics of the tear film. We realize that the model ignores some aspects of a complex fluid such as the shear rate dependence of viscosity. However, the current approximation can be justified by the facts that the net tear film is predominantly Newtonian with about 99% of its volume occupied by the aqueous layer. Also, since the exact distribution of free-floating mucin in the aqueous layer is unknown and may be impossible to ascertain, an assumption allowing material properties of the fluid to vary spatially alone may be reasonable for its computational affordances; such an assumption allows for determination of a completely analytical solution to the time-dependent flow equations. Yet another point to be noted is that local shear rate effects in the bulk fluid (with aqueous and mucin components) are accounted for in the way it impacts the motion of the immersed body, which will be described in greater detail in Sect. 2.3.

The fluid model used here is based on earlier work [41] where the fluid stress tensor is given by

$$\mathbf{T} = -p\mathbf{I} + 2\mu(\mathbf{x})\,\mathbf{D}, \tag{2}$$

$$\mathbf{D} = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla^T\mathbf{u}\right). \tag{3}$$

The linear momentum equation representing the incompressible flow of the bulk tear film system is given by

$$\rho\frac{\partial \mathbf{u}}{\partial t} = \text{div}\,\mathbf{T}, \tag{4}$$

$$\text{div}\,\mathbf{u} = 0. \tag{5}$$

The velocity profile for such an inhomogeneous fluid under oscillatory shearing motion of the boundary [41] is solved using the separation of variables $\mathbf{u} = u(z)\omega(t)\mathbf{e}_x$, where $u(z)$ is given by

$$u(z) = U\frac{e^{B_0 z} - 1}{e^{B_0} - 1}, \tag{6}$$

and the viscosity varies with depth according to the expression

$$\mu(z) = A_0 e^{-B_0 z}. \tag{7}$$

In these expressions, U is the eyelid blinking speed, and the coefficients A_0 and B_0 determine the mucin distribution magnitude and profile, respectively. The total period for a blinking cycle is about 0.1–0.4 s, and the maximum displacement of the eyelid is around 0.08 m [42]. Also, the speed of blinking is known to be 0.1 m/s. While there is asymmetry in the closing and opening speeds of the eyelid, during a complete blink, for the present study, we apply a sinusoidal input. Specifically, we take this function to be as follows:

$$\omega(t) = cos\left(\frac{2\pi t}{T}\right), \tag{8}$$

where T is the period of blinking motion assumed to be 0.2 s. As stated earlier, the above velocity profile (Eq. 6) can be assumed to model the flow corresponding to the tear film mixture with its viscosity varying according to the distribution profile of mucins in the layer (as depicted by Eq. 7).

We define the dimensionless parameters as follows:

$$Re = \frac{\rho U H}{\bar{\mu}},\ x^* = \frac{x}{H},\ u^* = \frac{u}{U},\ t^* = \frac{t}{H/U},\ F^* = \frac{F}{U^2/H}, \tag{9}$$

where U, H, and F correspond, respectively, to the maximum speed of eyelid (0.1 m/s), the thickness of aqueous and mucus layer (10^{-5} m), and binding force (which is discussed in the following section). Also $\bar{\mu}$ is the average dynamic viscosity over the tear film domain. From Eq. (7), $\bar{\mu}$ can be written in terms of the function of A_0, B_0,

$$\bar{\mu} = \frac{1}{Z}\int_0^Z \mu(z)dz = \frac{1}{Z}\int_0^Z A_0 e^{-B_0 z}dz = -\frac{1}{Z}\left(\frac{A_0}{B_0}\right)(e^{-B_0 Z} - 1), \tag{10}$$

where Z is the height of the fluid domain. The Reynolds number, Re in the layer induced by the blinking motion of the eyelid is quite low; it is reported [43] that the viscosity of mucin layer is at least 100 times higher than of water. As a result, Re for the mucin layer is around 10^{-2}. However, it is not surprising to estimate the average viscosity of the combined tear film (mucus and aqueous layers) would be much less than that, and in turn, Re in the whole layer should be slightly larger. Note that henceforth, the variable without * will be used thereafter.

2.2 Phase I: Mucoadhesion and Wrapping Mechanics

This portion of our model accounts for the microscopic interactions between the foreign body and the embedded mucins in the background fluid and in turn impacts the Phase II interaction between the fluid and the particle. This part of the paper can be summarized by the following sequence of reactions:

$$B + M \to A_1, \tag{11}$$

$$A_1 + (n-1)M \to A_n, \tag{12}$$

where B refers to the foreign body, M to a critical amount[1] of mucin strands which increases the dimension of the debris, A_1 is a single aggregate formed by the interaction of B and M through Van der Waal's forces, and A_n is a larger aggregate of size n. The chemistry of this interaction is discussed below and accounted for with care in our simulations.

2.2.1 Adhesion Mechanics

Adhesion force between mucins and proteins was measured in a study by Efremova et al. [32]. We follow their results of tethered polyethylene glycol (PEG) chains interacting with adsorbed mucin in the presence of soluble mucin. The adhesive force of mucin measured between the PEG bilayer and an adsorbed mucin layer bathed in a 0.2 mg/mL mucin solution (pH 7.2), which is relevant to the properties of tear film [44] was -0.3 ± 0.1 mN/m with adhesive contact from 593 ± 30 Å. We assume that adhesion activation begins as mucin and particle approach within 600 Å (the distance between the center of the mucin and the surface of the debris is referred to as the activation zone, denoted L_c) and the magnitude of the adhesion force, F_a is 0.3×10^{-8} milli-Newton. The mass of mucin is around 0.2–200 MDa as a larger aggregate [45]; in our computations we consider the mass to be 0.2 MDa. The adhesion force, which is activated in the activation zone, has a linear profile with respect to distance, given by the equations:

$$F_a(i) = -\frac{F_a}{L_c} d(i) + F_a, \quad \text{if} \quad d(i) \leq L_c, \tag{13}$$

$$F_a(i) = 0 \quad \text{otherwise}, \tag{14}$$

where $i = 1, \ldots, N$, the index number of mucin, and $d(i)$ is the distance between the centers of mucin and debris. L_c is the activation length, which is approximately $600\,\text{Å} + R_{debris}$. Therefore, $d(i) \leq L_c$ helps determine if the debris is within the activation zone or not (see Fig. 2b).

Mucins within L_c adhere to the surface of the debris simultaneously each time, and thus, the adhesion force of each mucin, $F_a(i)$, acts on the debris in the direction of approach. The force vector of each mucin can be expressed as

$$\mathbf{F}_a(i) = F_a(i)\hat{\mathbf{r}}(i). \tag{15}$$

The total force acting on the debris by mucins within L_c, therefore, is

$$\mathbf{F}_a = \sum_i^N \mathbf{F}_a(i) \quad \text{for} \quad i = 1, \cdots N. \tag{16}$$

[1] So if m refers to a single mucin strand, $M = n_{wrap} m$ where n_{wrap} is a critical number which depends upon the dimension of mucin. See Sect. 2.2.2 for more discussion on this point.

2.2.2 Wrapping and Size of Aggregate

We assume that the mucin-particle aggregate is a spherical homogeneous polychain as shown in Fig. 4 and its diameter and density change in accordance with the rate of adhesion. The radius of the aggregate changes such that

$$r_a = r_p + d_{k,m}, \tag{17}$$

where r_a is the radius of the mucin-particle aggregate, r_p is the radius of the foreign particle, and $d_{k,m}$ stands for the thickness of the mucin layer adhering to the debris (see Fig. 4). As a result, the number of mucin required to completely wrap the foreign body, n_{wrap}, is given by

$$n_{wrap} = \frac{SA_p}{PA_m} = \lceil \left(\frac{4\pi r_p^2}{\frac{\pi}{4} d_m^2} \right) = \lceil \left(\frac{16 r_p^2}{d_m^2} \right), \tag{18}$$

where SA_p is the surface area of the particle and PA_m refers to the projected area of the mucin and d_m is the thickness (or diameter) of each mucin particle which is assumed to be a sphere. Once the body binds with sufficiently many mucin spheres, denoted n_{wrap}, the diameter of the aggregate (i.e., debris and mucin) increase by the amount, d_m. As a result, for a particle coming into contact with k mucins, the increase in radius of the aggregate can be written as

$$d_{k,m} = \gamma \, d_m, \tag{19}$$

where $\gamma = \lfloor \left(\frac{k}{n_{wrap}} \right)$. Consequently, the specific density of a k-aggregate (i.e., particle interaction with k mucins) also changes after binding and is determined by the equation

$$s_a^{(k)} = \frac{s_p V_p + s_m V_m}{V_a^{(k)}} = \frac{s_p \frac{4\pi}{3} r_p^3 + s_m \frac{\pi}{3} d_{k,m}^3}{\frac{4\pi}{3} r_a^3} = \frac{4 s_p r_d^3 + s_m d_{k,m}^3}{4 r_a^3}. \tag{20}$$

In this expression s_p is the original specific density of debris, V_p is its volume, and r_p represents its radius. Similarly s_m and V_m represent the specific density and volume of the bound mucins.

We choose the initial non-dimensional diameter of the debris particle to be $r_p = 0.025$. The specific density, s_p, for the debris is chosen 1.03, while s_m for the freely floating mucin is taken to be 1 since the mucin is considered to be completely buoyant and as having negligible inertia. The original size (thickness) of mucin of the order of 10^{-8} m [46, 47]. In this study, we use the smallest size, 10^{-8} m, corresponding to a non-dimensional thickness value of $d_m = 10^{-3}$ in the computational domain (see Eqs. (9)). The projected area of each mucin (PA_m) is then around 10^{-6} and in turn, Eq. (18) yields $n_{wrap} \approx 10^4$, which can entirely wrap the debris particle.

It is worth noting that the change in thickness of the aggregate as given in Eqs. (17) and (19) refers to a discrete increment in the dimension of the aggregate, given by γ, which requires that the debris be completely wrapped before increasing further in size. In our computations, this condition is modified for computational convenience[2] so that $\gamma = \left(\frac{k}{n_{wrap}}\right)$

$$d_{k,m} = \left(\frac{k\,d_m}{n_{wrap}}\right) \approx \left(\frac{10^{-3}k}{10^4}\right) = 10^{-7}k, \tag{21}$$

which can be thought of as an average increment in the aggregate size for interaction with k mucins. The mucins that adhere to the surface are removed for the next time step from the environment, but immediately replaced by same number of mucins in the same positions accounting for the physiologic secretion process.

2.3 Phase II: The Particle Model

Phase two of our model considers the macroscopic transport properties of the foreign particle, based on the interaction forces between the surrounding fluid and the body. For computational convenience, mucins and a foreign object (i.e., ocular debris) immersed in the tear film are represented by immersed spherical particles, and their motions are governed by the force balance equation. We assume that the immersed objects do not disturb the flow through one way coupling of the flow and particle equations. The link between the particles and fluid is made through the flow viscosity. Therefore,

$$S\frac{d\mathbf{v}_p}{dt} = \mathbf{F}_{AM} + \mathbf{F}_B + \mathbf{F}_D + \mathbf{F}_L + \mathbf{F}_a, \tag{22}$$

where $S = \rho_p/\rho$ represents the specific gravity ($\rho = 1000\,\text{kg/m}^3$ for water), or ratio of particle to fluid density, and $\mathbf{v}_p$ is the particle velocity.

A more explicit form of this equation can be given as

$$\begin{aligned} S\frac{d\mathbf{v}_p}{dt} &= \frac{D\mathbf{u}}{Dt} + \frac{1}{2}(\frac{D\mathbf{u}}{Dt} - \frac{d\mathbf{v_p}}{dt}) + \frac{3}{4}\frac{C_D}{D}|\mathbf{u} - \mathbf{v}_p|(\mathbf{u} - \mathbf{v}_p) \\ &+ \frac{3}{4}\frac{C_L}{D}(|\mathbf{u} - \mathbf{v}_p|^2_{\text{top}} - |\mathbf{u} - \mathbf{v}_p|^2_{\text{bot}})\mathbf{n} + \mathbf{F}_a, \end{aligned} \tag{23}$$

[2] This assumption is made to side-step the issue of having as many as 10^4 mucin particles in the simulation to see an increase in aggregate size, as modeled by Eq. (19).

where the first term in the equation is $\mathbf{F_{AM}}$, the force of added mass, resulting from an accelerating object moving through a fluid having to move the surrounding fluid out of the way of the particle, or, along with the particle. This is the result of the fact that a particle and the fluid cannot both occupy the same space. The second term, $\mathbf{F_B}$, is the Bassett force, representing the fluid's history, due to a particle moving through a fluid and moving faster than the fluid can recover. As a particle accelerates in a fluid, the front half of the particle will push the fluid it encounters along or aside. The fluid on the rear side of the particle tends to fill in the space vacated by the particle. If the particle is moving too quickly, or accelerates too sharply, there is a gap between the fluid and the particle, which influences the particle's motion. The third term, $\mathbf{F_D}$, is the drag force, representing the force of retardation due to the dissipative nature of the fluid. The expression for drag force relies on a constant termed the drag coefficient. This coefficient, denoted C_D, is given in terms of the particle Reynolds number, Re_p as

$$C_D = \begin{cases} \frac{24}{Re_p}, & \text{if } Re_p \leq 1; \\ \frac{24}{Re_p}(1 + 0.15Re_p^{0.687}) & \text{if } Re_p > 1. \end{cases}$$

where Re_p is the Reynolds number of the particle, $Re_p = (\rho D_p U)/\mu$ (D_p is the diameter of the particle). When an object is surrounded by a fluid with an overall shearing motion, the velocity differential in the fluid upon the different parts of the object creates a lift force. The immersed object experiences a reduction in pressure on the side that is experiencing a larger velocity difference resulting in a net force in the direction of lower pressure. Since our fluid experiences different velocities at different levels above the moving floor, there will be a lift force generated by the particle, represented by the last term, $\mathbf{F_L}$, the lift force, where C_L is the lift coefficient and $\mathbf{n}$ represents the normal vector to the flow direction. As described earlier in Eqs. (13) and (14), $\mathbf{F_a}$ is the force of adherence acting on the debris particle by neighboring mucins.

3 Computational Methods

Figure 5 describes the overall computational scheme utilized in our study. There are several aspects to our computational approach, each of which is described below.

3.1 The Fluid Model

The fluid flow equation (6) is discretized in the three dimensional computational domain, containing rectangular blocks (194 × 130 × 290) and its unitless dimension ($L \times W \times H = 5 \times 0.3 \times 1$) units. The maximum blinking speed at the top (i.e.,

$Z = 1$) is $U(Z) = 1$. The window allocating the mucin and debris particles in the computational domain has the dimension of $L/10 \times W/3 \times H$ and is located in the middle of the domain to reduce computational cost.

Based on our model (Eq. (7)), mucins are assumed to be initially distributed based on the exponential viscosity profile such that the number of mucins is a function of z (and independent of x), given by

$$N(z) = A_0 e^{-B_0 z}. \tag{24}$$

At time $t = 0$, mucins are evenly distributed in x and y, and the count is simply determined by $N(z)$. The mucins are equally spaced in the chosen domain, and the spacing is given by $W/60$ in y and $H/100$ in the z directions. Therefore, in the entire volume, mucins are distributed more densely near the bottom, i.e., at the corneal surface.

To maintain consistency in the study, $N(z)$ and therefore the average viscosity, $\bar{\mu}$ must be held constant. Consequently, the constants A_0 and B_0, which prescribe the viscosity profile, are not independently determined but must be written in terms of the average viscosity of the tear film over the domain. Using Eq. (10), we can write the relationship between A_0, B_0, and $\bar{\mu}$ as

$$A_0 = \frac{B_0 \bar{\mu} Z}{1 - e^{-B_0 Z}}. \tag{25}$$

As illustrated in the curves on the left bottom in the figure, increasing B_0 stretches out the curve towards the origin so that higher mucin population proximal to corneal surface will be allocated for a larger B_0. In our computations, we explore various cases of B_0 with several different values of $\bar{\mu}$ (see Table 1).

3.2 The FSI Particle Model

A second order Runge-Kutta scheme is used to solve the non-dimensional Laplacian particle equations (22); detailed methods are reported in our earlier work [10]. As discussed in the earlier section, the adhesion force is activated as mucin and debris begin to approach each other and are within the activation range, L_c. In the computation, once within the range, the adhesion process is assumed to be completed. The specific density and diameter of debris is then immediately updated for the next time. We assume that mucin is constantly secreted, so for every mucin that leaves the domain after adhering to a debris which enters the activation range, a new mucin is generated to replace the old one. This is in keeping with the computational assumption that the number of mucin particles (i.e., $N(z)$) is a constant throughout the computation. This is also consistent with the healthy or normal physiological case of a constant mucin secretion rate.

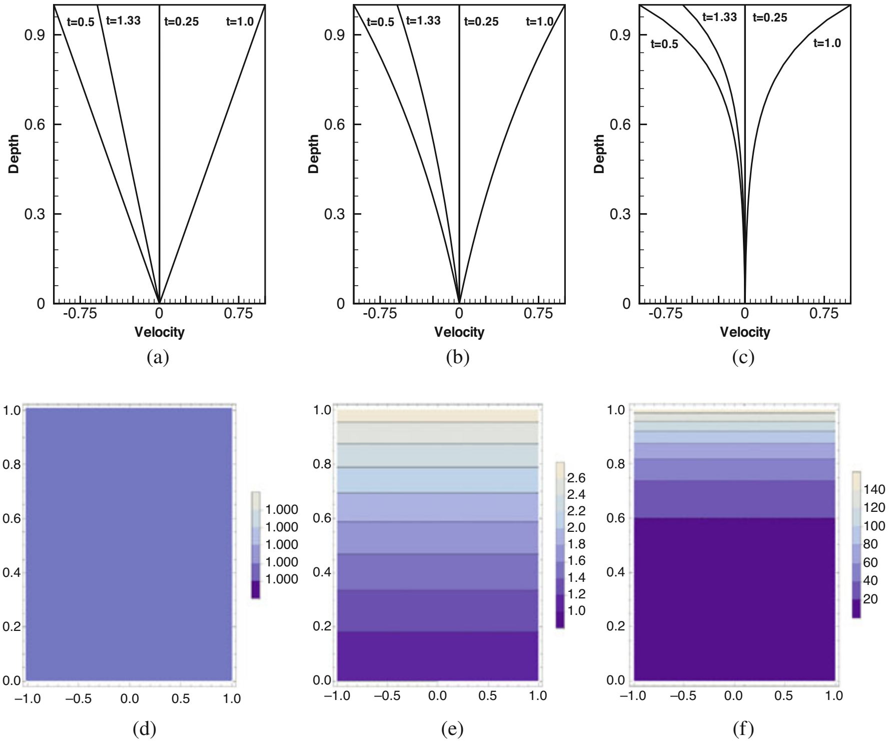


Fig. 3 This figure shows the velocity flow profile as a function of depth under time-dependent oscillation of the one of the walls (bottom). The different panels indicate the impact of changing viscosity dependence on depth, corresponding to varying mucin distributions; panel (**a**) corresponds to homogeneous fluid with no mucin (B_0) while panel (**b**) corresponds to the case $B_0 = 1$ and (**c**) to $B_0 = 5$. The last two panels indicate a more rapid change in the distribution of mucin, with progressively increasing mucin concentration towards the top. The figures (**d**–**f**) show the respective changes in viscosity as a function of depth

Initial mucin distribution is shown in Fig. 3. In our computations, mucins are confined within the smaller domain in the xy plane, near the object, to reduce computational cost. Therefore the foreign object should reside within the smaller domain. Mucin particles have periodic boundary conditions on both the lateral sides in y as well as the axial sides in x. In the z direction, mucins are not allowed to move out of the domain, which is enforced by setting the particles which wander outside the boundary, back into the domain to conserve the total number of mucin. The computation is terminated when the foreign object lifts up to 95% of the eye lid, along the z direction. Clearance of the foreign object is assumed to occur when it reaches approximately 95% of the thickness of the layer which is composed of mostly water, i.e., when the aggregate (debris+mucin) reach a non-dimensional height of $Z^* = 0.95$.

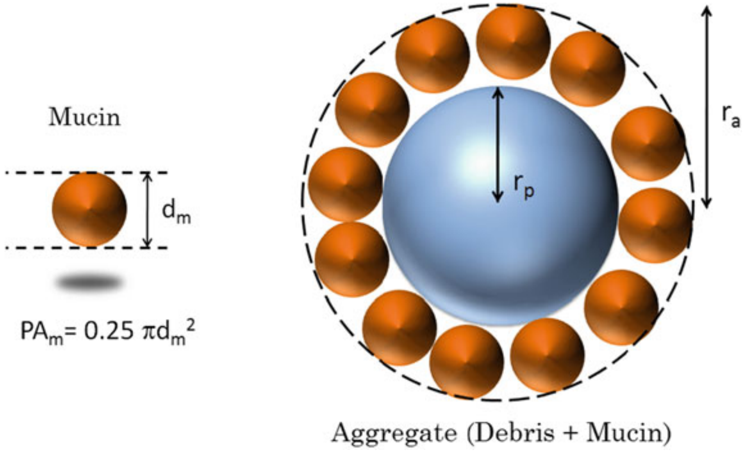


Fig. 4 The figure shows a schematic of the assumed physical properties of mucin in the model and the wrapping mechanism. A single strand of mucin is assumed to be spherical as shown on the left side of the figure. The figure on the right shows a cross section binding of the aggregate particle after binding, i.e., the debris and mucin. Once the debris, which is also treated as a sphere, is completely bound by n_{wrap} mucins, the size of the debris increases from $r_d \to r_a = r_d + d_m$

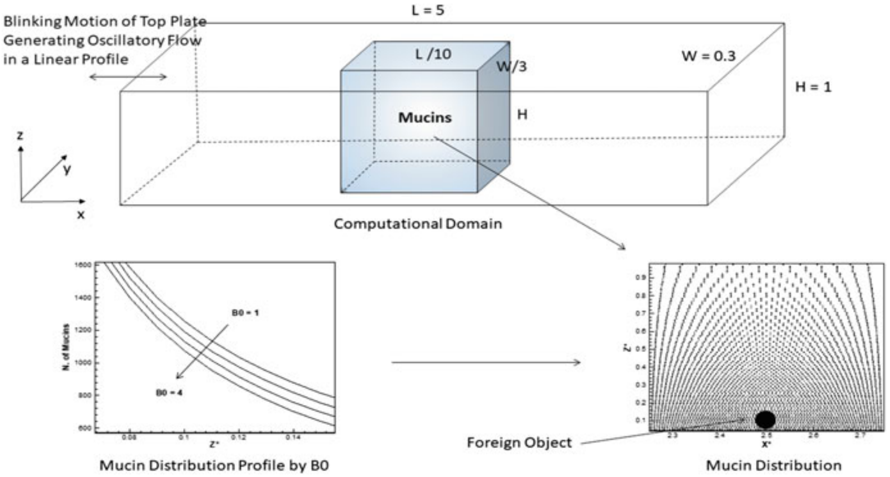


Fig. 5 This figure shows the computational domain in unitless dimensions with the allocated window for mucin. The bottom left panel shows mucin distribution with a viscosity profile for the cases of $B_0 = 1$ to $B_0 = 4$. The bottom right panel shows the resulting mucin distribution with a foreign object placed near the cornea for the specific case of $B_0 = 1$

4 Results

The present study aims to identify the role of mucin in the tear film subject. Based on the literature and our own contributions, we note that responsive blinking motion of the eyelid due to the entrance of foreign objects into the tear film activates mucin binding and the collective transport of aggregates. However, the effectiveness of this clearance mechanism can depend on the chemical and mechanical properties of mucin and also its secretion (or distribution) in the tear film. Our calculations reveal that the time required for the debris to move towards the eyelid (clearance time) can change depending on (i) viscosity distribution (i.e., by changing B_0 and A_0 in Eq. 25), (ii) viscosity, $\bar{\mu}$, and (iii) the adhesion force, $\mathbf{F}_a$.

In our computation, the foreign object is initially positioned at the location (L/2, W/2, H/10), i.e., near the bottom (see Fig. 5). The studies are performed by changing the values of parameters shown in Table 1.

As shown in the table, we explore several values of B_0, $\bar{\mu}$, and two values of F_a corresponding to the "normal" and "abnormal" cases. Equations (13) and (14) are based on known estimates [32] of the binding force, and any significant deviation from this value is considered to be "unhealthy" in our case. We collect the vertical displacement of the foreign particle in the z direction as a function of time shown in Fig. 6. Note that Z^* close to 1 represents the foreign particle proximal to the eyelid.

Note that higher B_0 corresponds to a more rapid decay in mucin distribution from the corneal surface to the eyelid. As seen in the figure, the time taken to lift up the particle to the eyelid reduces as $\bar{\mu}$ increases, for any B_0. Furthermore, B_0 increases (from (a) to (d) in the figure), the rate of displacement changes more rapidly. We also note that the clearance time is shorter at the middle range of B_0 explored in this study which is indicative of an optimal distribution profile for mucin. We also explore the case of $B_0 = 5$, which is shown in the inset of Fig. 6d. This inset panel figure is restricted to the case of the lowest value of $\bar{\mu}$ (namely, 108). In this particular case, the foreign object is not lifted up to required clearance height of 95% of total height because of lowered viscosity. As will be discussed later, a viscosity over a threshold value may be required for each B_0. Since B_0 determines the number distribution of mucin profile along the vertical direction, a higher B_0 results a greater number of mucins proximal to the cornea. Our study reveals that heavier mucin distribution near the cornea is, in fact, not optimal in the case when $B_0 = 5$. In order to help the role of protection of cornea by mucins, an appropriate mucin distribution is required.

Table 1 Parameter values corresponding to viscosity $\bar{\mu}$ and adhesion force (F_a) used in the computations

B_0	1	2	3	4	5	–	–
F_a	$F_a = F_{normal}$	$F_{normal}/3$	–	–	–	–	–
$\bar{\mu}$	$\mu_1 = 108$	$\mu_2 = 129$	$\mu_3 = 172$	$\mu_4 = 215$	$\mu_5 = 237$	$\mu_6 = 258$	$\mu_7 = 301$

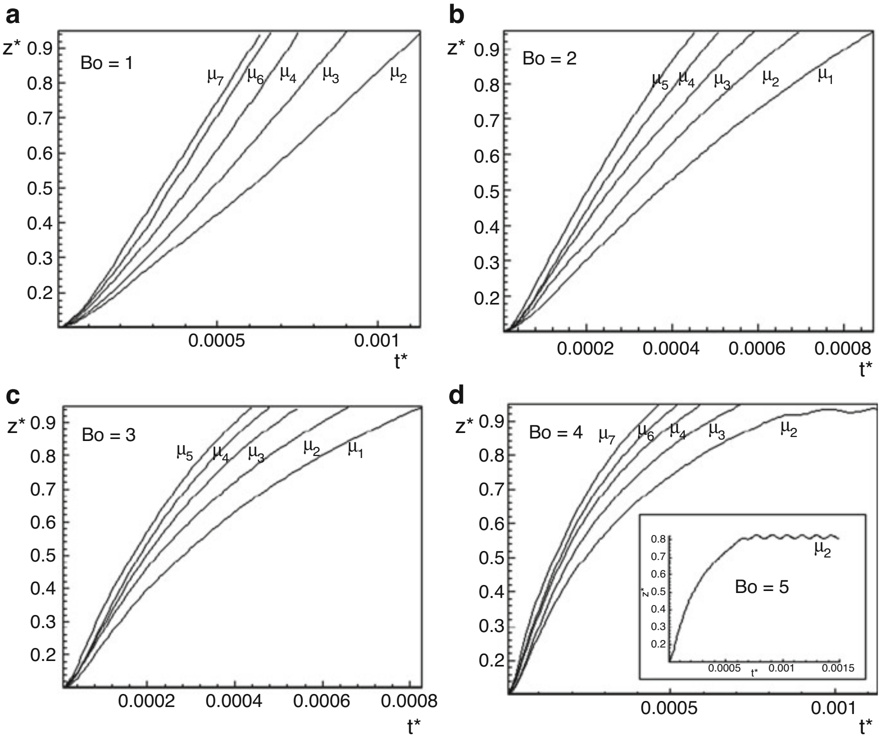


Fig. 6 This figure shows the vertical displacement (non-dimensional height, Z^*) of the debris as a function of time for varying viscosity ($\bar{\mu}$) at normal F_a. Each panel in the figure assumes a different value of B_0, corresponding to a different mucin distribution profile: (**a**) $B_0 = 1$, (**b**) $B_0 = 2$, (**c**) $B_0 = 3$ and (**d**) $B_0 = 4$,. The inset figure in (**d**) also showcases a single case of $B_0 = 5$. The optimal clearance time is defined as the minimum t^* when $Z^* = 0.95$ and the optimal physical parameters of the system correspond to the values of μ_i and B_0, which correspond to the minimum clearance time

5 Discussion

The parametric study of the effect of mechanical properties of mucin such as its profile B_0, adhesion force, and density (total number of mucins in the system estimated through the average viscosity), on the lift force of the particle produces valuable data to quantitatively explore the protection mechanism of mucins in the ocular tear film. The results of our computation suggest optimal mechanical properties of mucins. The study shows that an appropriate denser population of mucins proximal to mucosa as well as denser distribution in the entire layer ($\bar{\mu}$) in the tear film helps the cause of protection through accelerating the clearance rate of foreign immersed bodies.

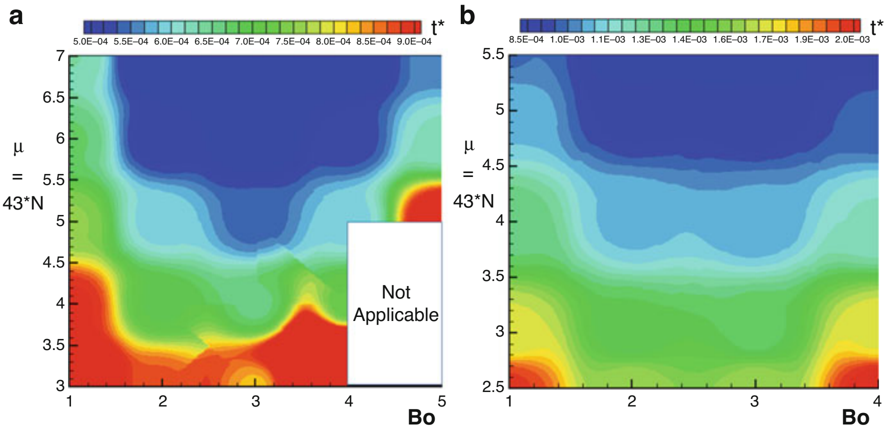


Fig. 7 This figure summarizes the results of our model computations using a contour plot, which identify the clearance lift-time of the debris in terms of control parameters B_0 and μ. Panel (**a**) corresponds to the case of a normal adhesion force $F_a = F_{normal}$ as defined in Table 1 while panel (**b**) corresponds to the case when $F_a = F_{normal}/3$. The blank area in panel (**a**) identified as "Not applicable" corresponds to the range where the aggregate (debris) never clears the tear film, i.e., always stays at $Z^* < 0.95$ within the computational time

The band plots showing the time of lift in Fig. 7 as a function of B_0 and $\bar{\mu}$ plane clearly show the "safe zone" where the fastest clearance is guaranteed. The figure illustrates that both the normal and abnormal (i.e., $F_a = F_{normal}/3$ which is lower than the normal value determined by experiments) adhesion forces can have small lift times (indicated by the color blue). The lift times are however a function of B_0 and $\bar{\mu}$; the optimal zones being identified with the region of the medium B_0 and higher viscosity, $\bar{\mu}$. The lift time is about two times slower for the abnormal adhesion force of mucin compared with normal cases, which seems to be qualitatively reasonable. For the normal adhesion force, an additional $B_0 = 5$ case was explored for three higher viscosity cases since two lower viscosities did not lift the object to the top near the eyelid. As seen in the inset in Fig. 6d, the object was not fully lifted due to the lower density of mucin, resulting in a blank region in the figure. In summary, the tear film can be seen as a complex system serving a protective role, which depends on physical (B_0, $\bar{\mu}$) and chemical factors (F_a). Optimal conditions for mucin to serve its recognized role of protection of the cornea are explored and understood via its ability to wrap and transport debris away from the cornea quickly and shown in the Fig. 7.

No quantified data on the mucin mechanical properties such as eye mucin adhesion force and its distribution under physiologic and pathological conditions would limit our analysis. Nevertheless from our study with systematic assumptions on the property data, the altered mechanical properties of mucins due to ocular surface disease may downregulate the role of mucins with regard to their protective capacity resulting in viral, fungal, or bacterial eye infections such as pink eye, or conjunctivitis. Understanding the normal behavioral conditions and mechanics of mucins-tear film complex therefore can be beneficial to the treatment of the ocular surface diseases. Our study can suggest that the treatment of ocular disease using synthetic or artificial tears should be associated with the appropriate mucin properties, as suggested in this work. Patient-specific data on mucin properties under healthy and pathological conditions would help further theoretical and computational studies to unveil the detailed mechanical roles of mucins.

Acknowledgments AV dedicates this work to his dear friend and co-author Bong Jae Chung who was an essential part of this project. Bong Jae was working hard on finding collaborators who could help provide experimental support for this work at the time of his untimely passing. I fondly remember all the wonderful discussions about this project and his excitement about all the questions that lay ahead that we planned to work on. BM, a former graduate student of Dr. Chung, remembers countless lessons he instilled in those who knew him, his immeasurable dedication to teaching, and his ever willingness to give his time to students and foster the growth of new ideas.

References

1. J.L. Gayton, Etiology, prevalence, and treatment of dry eye disease. Clin. Ophthalmol. **3**, 405 (2009)
2. H. Liang, C. Baudouin, P. Daull, J. Garrigue, F. Brignole-Baudouin, Effects of prostaglandin analogues anti-glaucoma therapies on ocular surface mucins. ARVO Annual Meeting Abstract (2012)
3. C.F. Cerretani, The Role of the Tear-Film Lipid Layer in Tear Dynamics and in Dry Eye. Ph.D. Dissertation, Chemical Engineering, University of California, Berkeley, 2013
4. I.K. Gipson,Y. Hori, P. Argeso, Character of ocular surface mucins and their alteration in dry eye disease. Ocul Surf. **2**(2), 131–148 (2004)
5. S.K. Lai, Y.Y. Wang, J. Hanes, Mucus-penetrating nanoparticles for drug and gene delivery to mucosal tissues. Adv. Drug Delivery Rev. **61**(2), 158–171 (2009)
6. B. Tirosh, A. Rubinstein, Migration of adhesive and nonadhesive particles in the rat intestine under altered mucus secretion conditions. J. Pharm. Sci. **87**(4), 453–456 (1998)
7. B. Govindarajan, I.K. Gipson, Membrane-tethered mucins have multiple functions on the ocular surface. Exp. Eye Res. **90**, 655–663 (2010)
8. S.K. Linden, P. Sutton, N.G. Karlsson, V. Korolik, M.A. McGuckin, Mucins in the mucosal barrier to infection. Mucosal Immunol. **1**, 183–197 (2008)
9. Y. Danjo, H. Watanabe, A.S. Tisdale, M. George, T. Tsumura, M.B. Abelson, I.K. Gipson, Alteration of mucin in human conjunctival epithelia in dry eye. Invest. Ophthalmol. Vis. Sci. **39**, 2602–2609 (1998)
10. B.J. Chung, D. Platt, A. Vaidya, The mechanics of clearance in a non-Newtonian lubrication layer. Int. J. Non-Linear Mech. **86**, 133–145 (2016)
11. Y. Danjo, M. Hakamura, Hamano, Measurement of the precorneal tear film thickness with a non-contact optical interferometry film thickness measurement system. Jpn. J. Ophthalmol. **38**, 260 (1994)
12. P.E. King-Smith, B.A. Fink, R.M. Hill, K.W. Koelling, J.M. Tiffany, The thickness of the tear film. Curr. Eye Res. **29**, 357 (2004)
13. B.A. Nichols, M.L. Chiappino, C.R. Dawson, Demonstration of the mucous layer of the tear film by electron microscopy. Invest. Ophthalmol. Visual Sci. **26**, 464 (1985)
14. A. Mircheff, Lacrimal fluid and electrolyte secretion: a review. Curr. Eye Res. **8**, 607 (1989)

15. S. Mishima, A. Gasset, S.D. Klyce, J.L. Baum, Determination of tear volume and tear flow. Invest. Ophthalmol. **5**, 264 (1966)
16. A.J. Bron, J.M. Tiffany, S.M. Gouveia, N. Yokoi, L.W. Voon, Functional aspects of the tear film lipid layer. Exp. Eye Res. **78**, 347 (2004)
17. F. Holly, B.S. Hong, Biochemical and surface characteristics of human tear proteins. Am. J. Optom. Physiol. Opt. **59**, 43 (1982)
18. J. Moore, J. Tiffany, Human ocular mucus, chemical studies. Exp. Eye Res. **33**, 203 (1981)
19. S.H. Choi, K.S. Park, M.W. Sung, K.H. Kim, Dynamic and quantitative evaluation of eyelid motion using image analysis. Med. Biol. Eng. Comput. **41**, 146 (2003)
20. R.J. Braun, Dynamics of the tear film. Annu. Rev. Fluid Mech. **44**, 267 (2012)
21. J.M. Tiffany, The viscosity of human tears. Int. Ophthalmol. **15**, 371 (1991)
22. Y.L. Zhang, O.K. Matar, R.V. Craster, Analysis of tear film rupture: the effects of non-Newtonian rheology. J. Colloid Interface Sci. **262**, 130–148 (2003)
23. J.D. Rodriguez, K.J. Lane, G.W. Ousler III, E. Angjeli, L.M. Smith, M.B. Abelson, Blink: characteristics, controls, and relation to dry eyes. Curr. Eye Res. **43**(1), 52–66 (2018)
24. F.C. Volkmann, L.A. Riggs, A.G. Ellicott, R.K. Moore, Measurements of visual suppression during opening, closing and blinking of the eyes. Vision Res. **22**(8), 991–996 (1982)
25. L.M. Ensign, C. Richard, H. Justin, Oral drug delivery with polymeric nanoparticles: the gastrointestinal mucus barriers. Adv. Drug Delivery Rev. **64**(6), 557–570 (2012)
26. R.A. Cone, Barrier properties of mucus. Adv. Drug Delivery Rev. **61**(2), 75–85 (2009)
27. W. Suchaoin et al., Development and in vitro evaluation of zeta potential changing self-emulsifying drug delivery systems for enhanced mucus permeation. Int. J. Pharm. **510**(1), 255–262 (2016)
28. M.E. Johnson, P.J. Murphy, Changes in the tear film and ocular surface from dry eye syndrome. Progr. Retinal Eye Res. **23**(4), 449–474 (2004)
29. Parsons, C.L. et al., Bladder surface mucin. Examination of possible mechanisms for its antibacterial effect. Invest. Urol. **16**(3), 196–200 (1978)
30. E. Perez, J.E. Proust, Forces between mica surfaces covered with adsorbed mucinacross aqueous solution. J. Colloid Interface Sci. **118**(1), 182–191 (1987)
31. J. Lukic, I. Strahinic, B. Jovcic, B. Filipic, L. Topisirovic, M. Kojic, J. Begovic, Different roles for lactococcal aggregation factor and mucin binding protein in adhesion to gastrointestinal mucosa. Appl. Environ. Microbiol. **78**(22), 7993–8000 (2012)
32. N.V. Efremova, Y. Huang, N.A. Peppas, D.E. Leckband, Direct measurement of interactions between tethered poly(ethylene glycol) chains and adsorbed mucin layers. Langmuir **18**(3), 836–845 (2002). https://doi.org/10.1021/la011303p
33. D.T.L. Le, Y.G. rardel, P. Loubière, M. Mercier-Bonin, and E. Dague, Measuring kinetic dissociation/association constants between lactococcus lactis bacteria and mucins using living cell probes. Biophys. J. **101**, 2843 (2011)
34. R. Tareb, M. Bernardeau, M. Gueguen, J. Vernoux, In vitro characterization of aggregation and adhesion properties of viable and heat-killed forms of two probiotic Lactobacillus strains and interaction with foodborne zoonotic bacteria, especially Campylobacter jejuni. J. Med. Microbiol. **62**, 637 (2013)
35. J.H. Siggers, C. Ross Ethier, Fluid mechanics of the eye. Annu. Rev. Fluid Mech. **44**, 347 (2012)
36. A.D. Fitt, G. Gonzalez, Fluid mechanics of the human eye: aqueous humour flow in the anterior chamber. Bull. Math. Biol. **68**, 53 (2006)
37. J. Telenius, Properties of the human tear film lipid layer. Aalto University Publication Series, Doctoral Dissertation 210/2013
38. P.E. King-Smith et al., The thickness of the human precorneal tear film: evidence from reflection spectra. Invest. Ophthal. Visual Sci. **41**(11), 3348–3359 (2000)
39. H. Zhu, Tear Dynamics. Diss. University of Florida, 2007
40. M. Massoudi, A note on the meaning of mixture viscosity using the classical continuum theories of mixtures. Int. J. Eng. Sci. **46**(7), 677–689 (2008)

41. M. Massoudi, A. Vaidya, Analytical solutions to Stokes-type flows of inhomogeneous fluids. Appl. Math. Comput. **218**(11), 6314–6329 (2012)
42. Y. Hashimoto, Y. Yotsumoto, The amount of time dilation for visual flickers corresponds to the amount of neural entrainments measured by EEG. Front. Comput. Neurosci. **12**, 30 (2018)
43. J. Leal, H.D. Smyth, D. Ghosh, Physicochemical properties of mucus and their impact on transmucosal drug delivery. Int. J. Pharm. **532**(1), 555–572 (2017)
44. F.H. Fischer, M. Wiederholt, Human precorneal tear film pH measured by microelectrodes. Graefes Arch Clin. Exp. Ophthalmol. **218**(3), 168–70 (1982)
45. M. Kesimer, J.K. Sheehan, Mass spectrometric analysis of mucin core proteins. Methods Mol. Biol. 842, 67–79 (2012)
46. C.J. Roberts et al., Topographical investigations of human ovarian-carcinoma polymorphic epithelial mucin by scanning tunnelling microscopy. Biochem. J. **283**(1), 181–185 (1992)
47. M. Kesimer et al., Molecular organization of the mucins and glycocalyx underlying mucus transport over mucosal surfaces of the airways. Mucosal Immunol. **6**(2), 379–392 (2013)